# Ordinary differential equations and Dynamical Systems

Gerald Teschl

Gerald Teschl
Institut für Mathematik
Nordbergstraße 15
Universität Wien
1090 Wien, Austria

*E-mail:* Gerald.Teschl@univie.ac.at
*URL:* http://www.mat.univie.ac.at/~gerald/

*1991 Mathematics subject classification.* 34-01

**Abstract.** This manuscript provides an introduction to ordinary differential equations and dynamical systems. We start with some simple examples of explicitly solvable equations. Then we prove the fundamental results concerning the initial value problem: existence, uniqueness, extensibility, dependence on initial conditions. Furthermore we consider linear equations, the Floquet theorem, and the autonomous linear flow.

Then we establish the Frobenius method for linear equations in the complex domain and investigates Sturm–Liouville type boundary value problems including oscillation theory.

Next we introduce the concept of a dynamical system and discuss stability including the stable manifold and the Hartman–Grobman theorem for both continuous and discrete systems.

We prove the Poincaré–Bendixson theorem and investigate several examples of planar systems from classical mechanics, ecology, and electrical engineering. Moreover, attractors, Hamiltonian systems, the KAM theorem, and periodic solutions are discussed as well.

Finally, there is an introduction to chaos. Beginning with the basics for iterated interval maps and ending with the Smale–Birkhoff theorem and the Melnikov method for homoclinic orbits.

*Keywords and phrases.* Ordinary differential equations, dynamical systems, Sturm-Liouville equations.

Typeset by $\mathcal{AMS}$-LaTeX and Makeindex.
Version: October 21, 2004

# Contents

# Preface

The present manuscript constitutes the lecture notes for my courses *Ordinary Differential Equations* and *Dynamical Systems and Chaos* held at the University of Vienna in Summer 2000 (5hrs.) and Winter 2000/01 (3hrs), respectively.

It is supposed to give a self contained introduction to the field of ordinary differential equations with emphasize on the view point of dynamical systems. It only requires some basic knowledge from calculus, complex functions, and linear algebra which should be covered in the usual courses. I tried to show how a computer system, *Mathematica*, can help with the investigation of differential equations. However, any other program can be used as well.

The manuscript is available from

`http://www.mat.univie.ac.at/~gerald/ftp/book-ode/`

*Acknowledgments*

I wish to thank my students P. Capka and F. Wisser who have pointed out several typos and made useful suggestions for improvements.

Gerald Teschl

Vienna, Austria
May, 2001

*Part 1*

# Classical theory

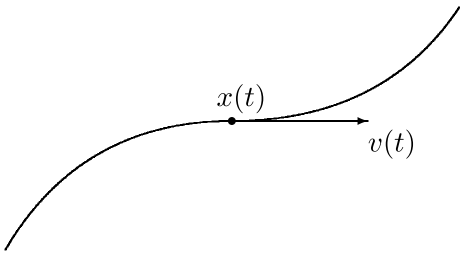
## Chapter 1

# Introduction

### 1.1. Newton's equations

Let us begin with an example from physics. In classical mechanics a particle is described by a point in space whose location is given by a function

$$x:\ \mathbb{R} \to \mathbb{R}^3. \tag{1.1}$$




The derivative of this function with respect to time is the velocity

$$v = \dot{x}:\ \mathbb{R} \to \mathbb{R}^3 \tag{1.2}$$

of the particle, and the derivative of the velocity is called acceleration

$$a = \dot{v}:\ \mathbb{R} \to \mathbb{R}^3. \tag{1.3}$$

In such a model the particle is usually moving in an external force field

$$F:\ \mathbb{R}^3 \to \mathbb{R}^3 \tag{1.4}$$

describing the force $F(x)$ acting on the particle at $x$. The basic law of Newton states that at each point $x$ in space the force acting on the particle must be equal to the acceleration times the mass $m > 0$ of the particle, that is,

$$m\,\ddot{x}(t) = F(x(t)), \qquad \text{for all } t \in \mathbb{R}. \tag{1.5}$$

Such a relation between a function $x(t)$ and its derivatives is called a **differential equation**. Equation (1.5) is called of second order since the highest derivative is of second degree. More precisely, we even have a system of differential equations since there is one for each coordinate direction.

In our case $x$ is called the dependent and $t$ is called the independent variable. It is also possible to increase the number of dependent variables by considering $(x, v)$. The advantage is that we now have a first order system

$$\begin{aligned} \dot{x}(t) &= v(t) \\ \dot{v}(t) &= \frac{1}{m} F(x(t)). \end{aligned} \tag{1.6}$$

This form is often better suited for theoretical investigations.

For given force $F$ one wants to find solutions, that is functions $x(t)$ which satisfy (1.5) (respectively (1.6)). To become more specific, let us look at the motion of a stone falling towards the earth. In the vicinity of the surface of the earth, the gravitational force acting on the stone is approximately constant and given by

$$F(x) = -m\,g \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}. \tag{1.7}$$

Here $g$ is a positive constant and the $x_3$ direction is assumed to be normal to the surface. Hence our system of differential equations reads

$$\begin{aligned} m\,\ddot{x}_1 &= 0, \\ m\,\ddot{x}_2 &= 0, \\ m\,\ddot{x}_3 &= -m\,g. \end{aligned} \tag{1.8}$$

The first equation can be integrated with respect to $t$ twice, resulting in $x_1(t) = C_1 + C_2 t$, where $C_1$, $C_2$ are the integration constants. Computing the values of $x_1$, $\dot{x}_1$ at $t = 0$ shows $C_1 = x_1(0)$, $C_2 = v_1(0)$, respectively. Proceeding analogously with the remaining two equations we end up with

$$x(t) = x(0) + v(0)\,t - \frac{g}{2} \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} t^2. \tag{1.9}$$

Hence the entire fate (past and future) of our particle is uniquely determined by specifying the initial location $x(0)$ together with the initial velocity $v(0)$.

From this example you might get the impression, that solutions of differential equations can always be found by straightforward integration. However, this is not the case in general. The reason why it worked here is, that the force is independent of $x$. If we refine our model and take the real gravitational force

$$F(x) = -\gamma\, m\, M \frac{x}{|x|^3}, \qquad \gamma, M > 0, \tag{1.10}$$

our differential equation reads

$$\begin{aligned} m\,\ddot{x}_1 &= -\frac{\gamma\, m\, M\, x_1}{(x_1^2 + x_2^2 + x_3^2)^{3/2}}, \\ m\,\ddot{x}_2 &= -\frac{\gamma\, m\, M\, x_2}{(x_1^2 + x_2^2 + x_3^2)^{3/2}}, \\ m\,\ddot{x}_3 &= -\frac{\gamma\, m\, M\, x_3}{(x_1^2 + x_2^2 + x_3^2)^{3/2}} \end{aligned} \tag{1.11}$$

and it is no longer clear how to solve it. Moreover, it is even unclear whether solutions exist at all! (We will return to this problem in Section 9.5.)

**Problem 1.1.** *Consider the case of a stone dropped from the height $h$. Denote by $r$ the distance of the stone from the surface. The initial condition reads $r(0) = h$, $\dot{r}(0) = 0$. The equation of motion reads*

$$\ddot{r} = -\frac{\gamma M}{(R+r)^2} \qquad \textit{(exact model)}$$

*respectively*

$$\ddot{r} = -g \qquad \textit{(approximate model)},$$

*where $g = \gamma M/R^2$ and $R$, $M$ are the radius, mass of the earth, respectively.*

(i) *Transform both equations into a first order system.*

(ii) *Compute the solution to the approximate system corresponding to the given initial condition. Compute the time it takes for the stone to hit the surface ($r = 0$).*

(iii) *Assume that the exact equation has also a unique solution corresponding to the given initial condition. What can you say about the time it takes for the stone to hit the surface in comparison to the approximate model? Will it be longer or shorter? Estimate the difference between the solutions in the exact and in the approximate case. (Hints: You should not compute the solution to the exact equation! Look at the minimum, maximum of the force.)*

(iv) *Grab your physics book from high school and give numerical values for the case $h = 10m$.*

**Problem 1.2.** *Consider again the exact model from the previous problem and write*

$$\ddot{r} = -\frac{\gamma M \varepsilon^2}{(1+\varepsilon r)^2}, \qquad \varepsilon = \frac{1}{R}.$$

*It can be shown that the solution $r(t) = r(t,\varepsilon)$ to the above initial conditions is $C^\infty$ (with respect to both $t$ and $\varepsilon$). Show that*

$$r(t) = h - g(1 - \frac{h}{R})\frac{t^2}{2} + O(\frac{1}{R^4}).$$

*(Hint: Insert $r(t,\varepsilon) = r_0(t) + r_1(t)\varepsilon + r_2(t)\varepsilon^2 + r_3(t)\varepsilon^3 + O(\varepsilon^4)$ into the differential equation and collect powers of $\varepsilon$.)*

## 1.2. Classification of differential equations

Let $U \subseteq \mathbb{R}^m$, $V \subseteq \mathbb{R}^n$ and $k \in \mathbb{N}_0$. Then $C^k(U,V)$ denotes the set of functions $U \to V$ having continuous derivatives up to order $k$. In addition, we will abbreviate $C(U,V) = C^0(U,V)$ and $C^k(U) = C^k(U,\mathbb{R})$.

A classical **ordinary differential equation** (ODE) is a relation of the form

$$F(t, x, x^{(1)}, \dots, x^{(k)}) = 0 \tag{1.12}$$

for the unknown function $x \in C^k(J)$, $J \subseteq \mathbb{R}$. Here $F \in C(U)$ with $U$ an open subset of $\mathbb{R}^{k+2}$ and

$$x^{(k)}(t) = \frac{d^k x(t)}{dt^k}, \qquad k \in \mathbb{N}_0, \tag{1.13}$$

are the *ordinary* derivatives of $x$. One frequently calls $t$ the **independent** and $x$ the **dependent variable**. The highest derivative appearing in $F$ is called the **order** of the differential equation. A **solution** of the ODE (1.12) is a function $\phi \in C^k(I)$, where $I \subseteq J$ is an interval, such that

$$F(t, \phi(t), \phi^{(1)}(t), \dots, \phi^{(k)}(t)) = 0, \qquad \text{for all } t \in I. \tag{1.14}$$

This implicitly implies $(t, \phi(t), \phi^{(1)}(t), \dots, \phi^{(k)}(t)) \in U$ for all $t \in I$.

Unfortunately there is not too much one can say about differential equations in the above form (1.12). Hence we will assume that one can solve $F$ for the highest derivative resulting in a differential equation of the form

$$x^{(k)} = f(t, x, x^{(1)}, \dots, x^{(k-1)}). \tag{1.15}$$

By the implicit function theorem this can be done at least locally near some point $(t,y) \in U$ if the partial derivative with respect to the highest derivative does not vanish at that point, $\frac{\partial F(t,y)}{\partial y_k}(t,y) \neq 0$. This is the type of differential equations we will from now on look at.

We have seen in the previous section that the case of real-valued functions is not enough and we should admit the case $x : \mathbb{R} \to \mathbb{R}^n$. This leads us to **systems of ordinary differential equations**

$$\begin{aligned} x_1^{(k)} &= f_1(t, x, x^{(1)}, \dots, x^{(k-1)}), \\ &\vdots \\ x_n^{(k)} &= f_n(t, x, x^{(1)}, \dots, x^{(k-1)}). \end{aligned} \tag{1.16}$$

Such a system is said to be **linear**, if it is of the form

$$x_i^{(k)} = g_i(t) + \sum_{l=1}^{n} \sum_{j=0}^{k-1} f_{i,j,l}(t) x_l^{(j)}. \tag{1.17}$$

It is called **homogeneous**, if $g_i(t) = 0$.

Moreover, any system can always be reduced to a first order system by changing to the new set of independent variables $y = (x, x^{(1)}, \dots, x^{(k-1)})$. This yields the new **first order system**

$$\begin{aligned} \dot{y}_1 &= y_2, \\ &\vdots \\ \dot{y}_{k-1} &= y_k, \\ \dot{y}_k &= f(t, y). \end{aligned} \tag{1.18}$$

We can even add $t$ to the independent variables $z = (t, y)$, making the right hand side independent of $t$

$$\begin{aligned} \dot{z}_1 &= 1, \\ \dot{z}_2 &= z_3, \\ &\vdots \\ \dot{z}_k &= z_{k+1}, \\ \dot{z}_{k+1} &= f(z). \end{aligned} \tag{1.19}$$

Such a system, where $f$ does not depend on $t$, is called **autonomous**. In particular, it suffices to consider the case of autonomous first order systems which we will frequently do.

Of course, we could also look at the case $t \in \mathbb{R}^m$ implying that we have to deal with partial derivatives. We then enter the realm of **partial differential equations** (PDE). However, this case is much more complicated and is not part of this manuscript.

Finally note that we could admit complex values for the dependent variables. It will make no difference in the sequel whether we use real or complex dependent variables. However, we will state most results only for the real case and leave the obvious changes to the reader. On the other hand, the case where the independent variable $t$ is complex requires more than obvious modifications and will be considered in Chapter 4.

**Problem 1.3.** *Classify the following differential equations.*

(i) $y'(x) + y(x) = 0$.
(ii) $\frac{d^2}{dt^2}u(t) = \sin(u(t))$.
(iii) $y(t)^2 + 2y(t) = 0$.
(iv) $\frac{\partial^2}{\partial x^2}u(x,y) + \frac{\partial^2}{\partial y^2}u(x,y) = 0$.
(v) $\dot{x} = -y$, $\dot{y} = x$.

**Problem 1.4.** *Which of the following differential equations are linear?*

(i) $y'(x) = \sin(x)y + \cos(y)$.
(ii) $y'(x) = \sin(y)x + \cos(x)$.
(iii) $y'(x) = \sin(x)y + \cos(x)$.

**Problem 1.5.** *Find the most general form of a second order linear equation.*

**Problem 1.6.** *Transform the following differential equations into first order systems.*

(i) $\ddot{x} + t\sin(\dot{x}) = x$.
(ii) $\ddot{x} = -y$, $\ddot{y} = x$.

*The last system is linear. Is the corresponding first order system also linear? Is this always the case?*

**Problem 1.7.** *Transform the following differential equations into autonomous first order systems.*

(i) $\ddot{x} + t\sin(\dot{x}) = x$.
(ii) $\ddot{x} = -\cos(t)x$.

*The last equation is linear. Is the corresponding autonomous system also linear?*

## 1.3. First order autonomous equations

Let us look at the simplest (nontrivial) case of a first order autonomous equation

$$\dot{x} = f(x), \quad x(0) = x_0, \qquad f \in C(\mathbb{R}). \tag{1.20}$$

This equation can be solved using a small ruse. If $f(x_0) \neq 0$, we can divide both sides by $f(x)$ and integrate both sides with respect to $t$:

$$\int_0^t \frac{\dot{x}(s)ds}{f(x(s))} = t. \tag{1.21}$$

Abbreviating $F(x) = \int_{x_0}^{x} f(y)^{-1} dy$ we see that every solution $x(t)$ of (1.20) must satisfy $F(x(t)) = t$. Since $F(x)$ is strictly monotone near $x_0$, it can be inverted and we obtain a unique solution

$$\phi(t) = F^{-1}(t), \qquad \phi(0) = F^{-1}(0) = x_0, \tag{1.22}$$

of our initial value problem.

Now let us look at the maximal interval of existence. If $f(x) > 0$ for $x \in (x_1, x_2)$ (the case $f(x) < 0$ follows by replacing $x \to -x$), we can define

$$T_+ = F(x_2) \in (0, \infty], \quad \text{respectively} \quad T_- = F(x_1) \in [-\infty, 0). \tag{1.23}$$

Then $\phi \in C^1((T_-, T_+))$ and

$$\lim_{t \uparrow T_+} \phi(t) = x_2, \quad \text{respectively} \quad \lim_{t \downarrow T_-} \phi(t) = x_1. \tag{1.24}$$

In particular, $\phi$ exists for all $t > 0$ (resp. $t < 0$) if and only if $1/f(x)$ is *not* integrable near $x_2$ (resp. $x_1$). Now let us look at some examples. If $f(x) = x$ we have $(x_1, x_2) = (0, \infty)$ and

$$F(x) = \ln(\frac{x}{x_0}). \tag{1.25}$$

Hence $T_\pm = \pm\infty$ and

$$\phi(t) = x_0 \mathrm{e}^t. \tag{1.26}$$

Thus the solution is globally defined for all $t \in \mathbb{R}$. Next, let $f(x) = x^2$. We have $(x_1, x_2) = (0, \infty)$ and

$$F(x) = \frac{1}{x_0} - \frac{1}{x}. \tag{1.27}$$

Hence $T_+ = 1/x_0$, $T_- = -\infty$ and

$$\phi(t) = \frac{x_0}{1 - x_0 t}. \tag{1.28}$$

In particular, the solution is no longer defined for all $t \in \mathbb{R}$. Moreover, since $\lim_{t \uparrow 1/x_0} \phi(t) = \infty$, there is no way we can possibly extend this solution for $t \geq T_+$.

Now what is so special about the zeros of $f(x)$? Clearly, if $f(x_1) = 0$, there is a trivial solution

$$\phi(t) = x_1 \tag{1.29}$$

to the initial condition $x(0) = x_1$. But is this the only one? If we have

$$\int_{x_1}^{x_0} \frac{dy}{f(y)} < \infty, \tag{1.30}$$

then there is another solution

$$\varphi(t) = F^{-1}(t), \qquad F(x) = \int_{x_1}^{x} \frac{dy}{f(y)} \tag{1.31}$$

with $\varphi(0) = x_1$ which is different from $\phi(t)$!

For example, consider $f(x) = \sqrt{|x|}$, then $(x_1, x_2) = (0, \infty)$,

$$F(x) = 2(\sqrt{x} - \sqrt{x_0}). \tag{1.32}$$

and

$$\varphi(t) = (\sqrt{x_0} + \frac{t}{2})^2, \quad -2\sqrt{x_0} < t < \infty. \tag{1.33}$$

So for $x_0 = 0$ there are several solutions which can be obtained by patching the trivial solution $\phi(t) = 0$ with the above one as follows

$$\tilde{\phi}(t) = \begin{cases} -\frac{(t-t_0)^2}{4}, & t \le t_0 \\ 0, & t_0 \le t \le t_1 \\ \frac{(t-t_1)^2}{4}, & t_1 \le t \end{cases}. \tag{1.34}$$

As a conclusion of the previous examples we have.

- Solutions might only exist locally, even for perfectly nice $f$.
- Solutions might not be unique. Note however, that $f$ is not differentiable at the point which causes the problems.

Note that the same ruse can be used to solve so-called **separable** equations

$$\dot{x} = f(x)g(t) \tag{1.35}$$

(see Problem 1.9).

**Problem 1.8.** *Solve the following differential equations:*

(i) $\dot{x} = x^3$.

(ii) $\dot{x} = x(1 - x)$.

(iii) $\dot{x} = x(1 - x) - c$.

**Problem 1.9** (Separable equations)**.** *Show that the equation*

$$\dot{x} = f(x)g(t), \qquad x(t_0) = x_0,$$

*has locally a unique solution if $f(x_0) \neq 0$. Give an implicit formula for the solution.*

**Problem 1.10.** *Solve the following differential equations:*

(i) $\dot{x} = \sin(t)x$.

(ii) $\dot{x} = g(t)\tan(x)$.

**Problem 1.11** (Linear homogeneous equation)**.** *Show that the solution of $\dot{x} = q(t)x$, where $q \in C(\mathbb{R})$, is given by*

$$\phi(t) = x_0 \exp\left(\int_{t_0}^{t} q(s)ds\right).$$

**Problem 1.12** (Growth of bacteria)**.** *A certain species of bacteria grows according to*

$$\dot{N}(t) = \kappa N(t), \qquad N(0) = N_0,$$

*where $N(t)$ is the amount of bacteria at time $t$ and $N_0$ is the initial amount. If there is only space for $N_{\max}$ bacteria, this has to be modified according to*

$$\dot{N}(t) = \kappa(1 - \frac{N(t)}{N_{\max}})N(t), \qquad N(0) = N_0.$$

*Solve both equations, assuming $0 < N_0 < N_{\max}$ and discuss the solutions. What is the behavior of $N(t)$ as $t \to \infty$?*

**Problem 1.13** (Optimal harvest)**.** *Take the same setting as in the previous problem. Now suppose that you harvest bacteria at a certain rate $H > 0$. Then the situation is modeled by*

$$\dot{N}(t) = \kappa(1 - \frac{N(t)}{N_{\max}})N(t) - H, \qquad N(0) = N_0.$$

*Make a scaling*

$$x(\tau) = \frac{N(t)}{N_{\max}}, \qquad \tau = \kappa t$$

*and show that the equation transforms into*

$$\dot{x}(\tau) = (1 - x(\tau))x(\tau) - h, \qquad h = \frac{H}{\kappa N_{\max}}.$$

*Visualize the region where $f(x,h) = (1-x)x - h$, $(x,h) \in U = (0,1) \times (0,\infty)$, is positive respectively negative. For given $(x_0,h) \in U$, what is the behavior of the solution as $t \to \infty$? How is it connected to the regions plotted above? What is the maximal harvest rate you would suggest?*

**Problem 1.14** (Parachutist)**.** *Consider the free fall with air resistance modeled by*

$$\ddot{x} = -\eta\dot{x} - g, \qquad \eta > 0.$$

*Solve this equation (Hint: Introduce the velocity $v = \dot{x}$ as new independent variable). Is there a limit to the speed the object can attain? If yes, find it. Consider the case of a parachutist. Suppose the chute is opened at a certain time $t_0 > 0$. Model this situation by assuming $\eta = \eta_1$ for $0 < t < t_0$ and $\eta = \eta_2 > \eta_1$ for $t > t_0$. What does the solution look like?*

## 1.4. Finding explicit solutions

We have seen in the previous section, that some differential equations can be solved explicitly. Unfortunately, there is no general recipe for solving a given differential equation. Moreover, finding explicit solutions is in general impossible unless the equation is of a particular form. In this section I will show you some classes of first order equations which are explicitly solvable.

The general idea is to find a suitable change of variables which transforms the given equation into a solvable form. Hence we want to review this concept first. Given the point $(t, x)$, we transform it to the new one $(s, y)$ given by

$$s = \sigma(t, x), \qquad y = \eta(t, x). \tag{1.36}$$

Since we do not want to loose information, we require this transformation to be invertible. A given function $\phi(t)$ will be transformed into a function $\psi(s)$ which has to be obtained by eliminating $t$ from

$$s = \sigma(t, \phi(t)), \qquad \psi = \eta(t, \phi(t)). \tag{1.37}$$

Unfortunately this will not always be possible (e.g., if we rotate the graph of a function in $\mathbb{R}^2$, the result might not be the graph of a function). To avoid this problem we restrict our attention to the special case of **fiber preserving transformations**

$$s = \sigma(t), \qquad y = \eta(t, x) \tag{1.38}$$

(which map the fibers $t = const$ to the fibers $s = const$). Denoting the inverse transform by

$$t = \tau(s), \qquad x = \xi(s, y), \tag{1.39}$$

a straightforward application of the chain rule shows that $\phi(t)$ satisfies

$$\dot{x} = f(t, x) \tag{1.40}$$

if and only if $\psi(s) = \eta(\tau(s), \phi(\tau(s)))$ satisfies

$$\dot{y} = \dot{\tau}\left(\frac{\partial \eta}{\partial t}(\tau, \xi) + \frac{\partial \eta}{\partial x}(\tau, \xi)\, f(\tau, \xi)\right), \tag{1.41}$$

where $\tau = \tau(s)$ and $\xi = \xi(s, y)$. Similarly, we could work out formulas for higher order equations. However, these formulas are usually of little help for practical computations and it is better to use the simpler (but ambiguous) notation

$$\frac{dy}{ds} = \frac{dy(t(s), x(t(s))}{ds} = \frac{\partial y}{\partial t}\frac{dt}{ds} + \frac{\partial y}{\partial x}\frac{dx}{dt}\frac{dt}{ds}. \tag{1.42}$$

But now let us see how transformations can be used to solve differential equations.

A (nonlinear) differential equation is called **homogeneous** if it is of the form

$$\dot{x} = f(\frac{x}{t}). \tag{1.43}$$

This special form suggests the change of variables $y = \frac{x}{t}$ ($t \neq 0$), which transforms our equation into

$$\dot{y} = \frac{f(y) - y}{t}. \tag{1.44}$$

This equation is separable.

More generally, consider the differential equation

$$\dot{x} = f(\frac{ax + bt + c}{\alpha x + \beta t + \gamma}). \tag{1.45}$$

Two cases can occur. If $a\beta - \alpha b = 0$, our differential equation is of the form

$$\dot{x} = \tilde{f}(ax + bt), \tag{1.46}$$

which transforms into

$$\dot{y} = a\tilde{f}(y) + b \tag{1.47}$$

if we set $y = ax + bt$. If $a\beta - \alpha b \neq 0$, we can use $y = x - x_0$ and $s = t - t_0$ which transforms (1.45) to the homogeneous equation

$$\dot{y} = \hat{f}(\frac{ay + bs}{\alpha y + \beta s}) \tag{1.48}$$

if $(x_0, t_0)$ is the unique solution of the linear system $ax + bt + c = 0$, $\alpha x + \beta t + \gamma = 0$.

A differential equation is of **Bernoulli** type if it is of the form

$$\dot{x} = f(t)x + g(t)x^n, \qquad n \neq 1. \tag{1.49}$$

The transformation

$$y = x^{1-n} \tag{1.50}$$

gives the linear equation

$$\dot{y} = (1 - n)f(t)y + (1 - n)g(t). \tag{1.51}$$

We will show how to solve this equation in Section 3.3 (or see Problem 1.18).

A differential equation is of **Riccati** type if it is of the form

$$\dot{x} = f(t)x + g(t)x^2 + h(t). \tag{1.52}$$

Solving this equation is only possible if a particular solution $x_p(t)$ is known. Then the transformation

$$y = \frac{1}{x - x_p(t)} \tag{1.53}$$

yields the linear equation

$$\dot{y} = (2x_p(t)g(t) + f(t))y + g(t). \tag{1.54}$$

These are only a few of the most important equations which can be explicitly solved using some clever transformation. In fact, there are reference books like the one by Kamke [**17**], where you can look up a given equation and find out if it is known to be solvable. As a rule of thumb one has that for a first order equation there is a realistic chance that it is explicitly solvable. But already for second order equations explicitly solvable ones are rare.

Alternatively, we can also ask a symbolic computer program like *Mathematica* to solve differential equations for us. For example, to solve

$$\dot{x} = \sin(t)x \tag{1.55}$$

you would use the command

*In[1]:=* `DSolve[x'[t] == x[t]Sin[t], x[t], t]`

*Out[1]=* $\{\{\mathtt{x[t]} \to \mathrm{e}^{-\mathtt{Cos[t]}}\mathtt{C[1]}\}\}$

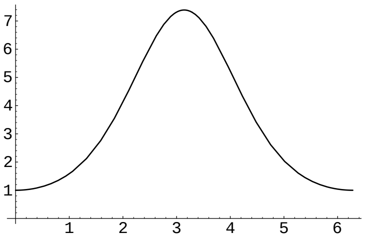
Here the constant `C[1]` introduced by *Mathematica* can be chosen arbitrarily (e.g. to satisfy an initial condition). We can also solve the corresponding initial value problem using

*In[2]:=* `DSolve[{x'[t] == x[t]Sin[t], x[0] == 1}, x[t], t]`

*Out[2]=* $\{\{\mathtt{x[t]} \to \mathrm{e}^{1-\mathtt{Cos[t]}}\}\}$

and plot it using

*In[3]:=* `Plot[x[t] /. %, {t, 0, 2π}];`




So it almost looks like *Mathematica* can do everything for us and all we have to do is type in the equation, press enter, and wait for the solution. However, as always, life is not that easy. Since, as mentioned earlier, only very few differential equations can be solved explicitly, the `DSolve` command can only help us in very few cases. The other cases, that is those which cannot be explicitly solved, will the the subject of the remainder of this book!

Let me close this section with a warning. Solving one of our previous examples using *Mathematica* produces

*In[4]:=* `DSolve[{x'[t] == √x[t], x[0] == 0}, x[t], t]`

*Out[4]=* $\{\{\mathtt{x[t]} \to \frac{\mathtt{t}^2}{4}\}\}$

However, our investigations of the previous section show that this is not the only solution to the posed problem! *Mathematica* expects you to know that there are other solutions and how to get them.

**Problem 1.15.** *Try to find solutions of the following differential equations:*

(i) $\dot{x} = \frac{3x-2t}{t}$.

(ii) $\dot{x} = \frac{x-t+2}{2x+t+1} + 5$.

(iii) $y' = y^2 - \frac{y}{x} - \frac{1}{x^2}$.

(iv) $y' = \frac{y}{x} - \tan(\frac{y}{x})$.

**Problem 1.16** (Euler equation). *Transform the differential equation*

$$t^2\ddot{x} + 3t\dot{x} + x = \frac{2}{t}$$

*to the new coordinates $y = x$, $s = \ln(t)$. (Hint: You are* not *asked to solve it.)*

**Problem 1.17.** *Pick some differential equations from the previous problems and solve them using your favorite mathematical software. Plot the solutions.*

**Problem 1.18** (Linear inhomogeneous equation). *Verify that the solution of $\dot{x} = q(t)x + p(t)$, where $p, q \in C(\mathbb{R})$, is given by*

$$\phi(t) = x_0 \exp\left(\int_{t_0}^{t} q(s)ds\right) + \int_{t_0}^{t} \exp\left(\int_{s}^{t} q(r)dr\right) p(s)\, ds.$$

**Problem 1.19** (Exact equations). *Consider the equation*

$$F(x, y) = 0,$$

*where $F \in C^2(\mathbb{R}^2, \mathbb{R})$. Suppose $y(x)$ solves this equation. Show that $y(x)$ satisfies*

$$p(x, y)y' + q(x, y) = 0,$$

*where*

$$p(x, y) = \frac{\partial F(x, y)}{\partial y} \quad \text{and} \quad q(x, y) = \frac{\partial F(x, y)}{\partial x}.$$

*Show that we have*

$$\frac{\partial p(x, y)}{\partial x} = \frac{\partial q(x, y)}{\partial y}.$$

*Conversely, a first order differential equation as above (with arbitrary coefficients $p(x, y)$ and $q(x, y)$) satisfying this last condition is called* **exact**. *Show that if the equation is exact, then there is a corresponding function $F$ as above. Find an explicit formula for $F$ in terms of $p$ and $q$. Is $F$ uniquely determined by $p$ and $q$?*

*Show that*

$$(4bxy + 3x + 5)y' + 3x^2 + 8ax + 2by^2 + 3y = 0$$

*is exact. Find $F$ and find the solution.*

**Problem 1.20** (Integrating factor). *Consider*

$$p(x,y)y' + q(x,y) = 0.$$

*A function $\mu(x,y)$ is called* **integrating factor** *if*

$$\mu(x,y)p(x,y)y' + \mu(x,y)q(x,y) = 0$$

*is exact.*

*Finding an integrating factor is in general as hard as solving the original equation. However, in some cases making an ansatz for the form of $\mu$ works.*

*Consider*

$$xy' + 3x - 2y = 0$$

*and look for an integrating factor $\mu(x)$ depending only on $x$. Solve the equation.*

**Problem 1.21** (Focusing of waves). *Suppose you have an incoming electromagnetic wave along the $y$-axis which should be focused on a receiver sitting at the origin $(0,0)$. What is the optimal shape for the mirror?*

*(Hint: An incoming ray, hitting the mirror at $(x,y)$ is given by*

$$R_{\text{in}}(t) = \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} 0 \\ 1 \end{pmatrix} t, \quad t \in (-\infty, 0].$$

*At $(x,y)$ it is reflected and moves along*

$$R_{\text{rfl}}(t) = \begin{pmatrix} x \\ y \end{pmatrix} (1-t), \quad t \in [0,1].$$

*The laws of physics require that the angle between the tangent of the mirror and the incoming respectively reflected ray must be equal. Considering the scalar products of the vectors with the tangent vector this yields*

$$\frac{1}{\sqrt{1+u^2}} \begin{pmatrix} 1 \\ u \end{pmatrix} \begin{pmatrix} 1 \\ y' \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \begin{pmatrix} 1 \\ y' \end{pmatrix}, \quad u = \frac{y}{x},$$

*which is the differential equation for $y = y(x)$ you have to solve.)*

**Problem 1.22** (Nonlinear boundary value problem). *Show that the nonlinear boundary value problem*

$$y''(x) + y(x)^2 = 0, \qquad y(0) = y(1) = 0,$$

*has a unique nontrivial solution. (Note that instead of initial conditions $y(0)$, $y'(0)$ the values at the boundary $y(0)$, $y(1)$ are given.) Show that the solution is symmetric with respect to $x = \frac{1}{2}$ and compute $y'(0)$ respectively $y(\frac{1}{2})$.*

*(Hint: Set $y' = p(y)$ and derive a first order equation for $p(y)$. Solve this equation for $p(y)$ and then solve the equation $y' = p(y)$. Note that this works for any equation of the type $y'' = f(y)$.)*

## 1.5. Qualitative analysis of first order equations

As already noted in the previous section, only very few ordinary differential equations are explicitly solvable. Fortunately, in many situations a solution is not needed and only some qualitative aspects of the solutions are of interest. For example, does it stay within a certain region, what does it look like for large $t$, etc..

In this section I want to investigate the differential equation

$$\dot{x} = x^2 - t^2 \tag{1.56}$$

as a prototypical example. It is of Riccati type and according to the previous section, it cannot be solved unless a particular solution can be found. But there does not seem to be a solution which can be easily guessed. (We will show later, in Problem 4.7, that it is explicitly solvable in terms of special functions.)

So let us try to analyze this equation without knowing the solution. Well, first of all we should make sure that solutions exist at all! Since we will attack this in full generality in the next chapter, let me just state that if $f(t,x) \in C^1(\mathbb{R}^2, \mathbb{R})$, then for every $(t_0, x_0) \in \mathbb{R}^2$ there exists a unique solution of the initial value problem

$$\dot{x} = f(t,x), \qquad x(t_0) = x_0 \tag{1.57}$$

defined in a neighborhood of $t_0$ (Theorem 2.3). However, as we already know from Section 1.3, solutions might not exist for all $t$ even though the differential equation is defined for all $(t,x) \in \mathbb{R}^2$. However, we will show that a solution must converge to $\pm\infty$ if it does not exist for all $t$ (Corollary 2.11).

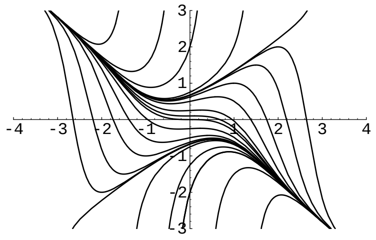
In order to get some feeling of what we should expect, a good starting point is a numerical investigation. Using the command

```
In[5]:= NDSolve[{x'[t] == x[t]^2 - t^2, x[0] == 1}, x[t], {t, -2, 2}]

    NDSolve::ndsz: At t == 1.0374678967709798‘, step size is
      effectively zero; singularity suspected.

Out[5]= {{x[t] → InterpolatingFunction[{{-2., 1.03747}}, <>][t]}}
```

we can compute a numerical solution on the interval $(-2, 2)$. Numerically solving an ordinary differential equations means computing a sequence of points $(t_j, x_j)$ which are hopefully close to the graph of the real solution (we will briefly discuss numerical methods in Section 2.5). Instead of this list of points, *Mathematica* returns an interpolation function which – as you might have already guessed from the name – interpolates between these points and hence can be used as any other function.

Note, that in our particular example, *Mathematica* complained about the step size (i.e., the difference $t_j - t_{j-1}$) getting too small and stopped at $t = 1.037\dots$. Hence the result is only defined on the interval $(-2, 1.03747)$ even tough we have requested the solution on $(-2, 2)$. This indicates that the solution only exist for finite time.

Combining the solutions for different initial conditions into one plot we get the following picture:




First of all we note the symmetry with respect to the transformation $(t, x) \to (-t, -x)$. Hence it suffices to consider $t \geq 0$. Moreover, observe that different solutions do never cross, which is a consequence of uniqueness.

According to our picture, there seem to be two cases. Either the solution escapes to $+\infty$ in finite time or it converges to the line $x = -t$. But is this really the correct behavior? There could be some numerical errors accumulating. Maybe there are also solutions which converge to the line $x = t$ (we could have missed the corresponding initial conditions in our picture)? Moreover, we could have missed some important things by restricting ourselves to the interval $t \in (-2, 2)$! So let us try to prove that our picture is indeed correct and that we have not missed anything.

We begin by splitting the plane into regions according to the sign of $f(t, x) = x^2 - t^2$. Since it suffices to consider $t \geq 0$ there are only three regions: I: $x > t$, II: $-t < x < t$, and III: $x < -t$. In region I and III the solution is increasing, in region II it is decreasing. Furthermore, on the line $x = t$ each solution has a horizontal tangent and hence solutions can only get from region I to II but not the other way round. Similarly, solutions can only get from III to II but not from II to III.

More generally, let $x(t)$ be a solution of $\dot{x} = f(t, x)$ and assume that it is defined on $[t_0, T)$, $T > t_0$. A function $x_+(t)$ satisfying

$$\dot{x}_+(t) > f(t, x_+(t)), \qquad t \in (t_0, T), \tag{1.58}$$

is called a **super solution** of our equation. Every super solution satisfies

$$x(t) < x_+(t), \quad t \in (t_0, T), \quad \text{whenever} \quad x(t_0) \leq x_+(t_0). \tag{1.59}$$

In fact, consider $\Delta(t) = x_+(t) - x(t)$. Then we have $\Delta(t_0) \ge 0$ and $\dot\Delta(t) > 0$ whenever $\Delta(t) = 0$. Hence $\Delta(t)$ can cross 0 only from below.

Similarly, a function $x_-(t)$ satisfying

$$\dot x_-(t) < f(t, x_-(t)), \qquad t \in (t_0, T), \tag{1.60}$$

is called a **sub solution**. Every sub solution satisfies

$$x_-(t) < x(t), \quad t \in (t_0, T), \quad \text{whenever} \quad x(t_0) \ge x_-(t_0). \tag{1.61}$$

Similar results hold for $t < t_0$. The details are left to the reader (Problem 1.23).

Using this notation, $x_+(t) = t$ is a super solution and $x_-(t) = -t$ is a sub solution for $t \ge 0$. This already has important consequences for the solutions:

- For solutions starting in region I there are two cases; either the solution stays in I for all time and hence must converge to $+\infty$ (maybe in finite time) or it enters region II.
- A solution starting in region II (or entering region II) will stay there for all time and hence must converge to $-\infty$. Since it must stay above $x = -t$ this cannot happen in finite time.
- A solution starting in III will eventually hit $x = -t$ and enter region II.

Hence there are two remaining questions: Do the solutions in region I which converge to $+\infty$ reach $+\infty$ in finite time, or are there also solutions which converge to $+\infty$, e.g., along the line $x = t$? Do the other solutions all converge to the line $x = -t$ as our numerical solutions indicate?

To answer both questions, we will again resort to super/sub solutions. For example, let us look at the **isoclines** $f(x,t) = const.$ Considering $x^2 - t^2 = -2$ the corresponding curve is

$$y_+(t) = -\sqrt{t^2 - 2}, \qquad t > \sqrt{2}, \tag{1.62}$$

which is easily seen to be a super solution

$$\dot y_+(t) = -\frac{t}{\sqrt{t^2-2}} > -2 = f(t, y_+(t)) \tag{1.63}$$

for $t > \frac{4}{\sqrt{3}}$. Thus, as soon as a solution $x(t)$ enters the region between $y_+(t)$ and $x_-(t)$ it must stay there and hence converge to the line $x = -t$ since $y_+(t)$ does.

But will every solution in region II eventually end up between $y_+(t)$ and $x_-(t)$? The answer is yes, since above $y_+(t)$ we have $\dot x(t) < -2$. Hence a solution starting at a point $(t_0, x_0)$ above $y_+(t)$ stays below $x_0 - 2(t - t_0)$.

Hence every solution which is in region II at some time will converge to the line $x = -t$.

Finally note that there is nothing special about $-2$, any value smaller than $-1$ would have worked as well.

Now let us turn to the other question. This time we take an isocline $x^2 - t^2 = 2$ to obtain a corresponding sub solution

$$y_-(t) = -\sqrt{2+t^2}, \qquad t > 0. \tag{1.64}$$

At first sight this does not seem to help much because the sub solution $y_-(t)$ lies *above* the super solution $x_+(t)$. Hence solutions are able to leave the region between $y_-(t)$ and $x_+(t)$ but cannot come back. However, let us look at the solutions which stay inside at least for some finite time $t \in [0,T]$. By following the solutions with initial conditions $(T, x_+(T))$ and $(T, y_-(T))$ we see that they hit the line $t = 0$ at some points $a(T)$ and $b(T)$, respectively. Since different solutions can never cross, the solutions which stay inside for (at least) $t \in [0,T]$ are precisely those starting at $t = 0$ in the interval $[a(T), b(T)]$! Taking $T \to \infty$ we see that all solutions starting in the interval $[a(\infty), b(\infty)]$ (which might be just one point) at $t = 0$, stay inside for all $t > 0$. Furthermore, since $f(t,.)$ is increasing in region I, we see that the distance between two solutions

$$x_1(t) - x_0(t) = x_1(t_0) - x_0(t_0) + \int_{t_0}^{t} f(s, x_1(s)) - f(s, x_0(s)) ds \tag{1.65}$$

must increase as well. Thus there can be at most one solution $x_0(t)$ which stays between $x_+(t)$ and $y_-(t)$ for all $t > 0$ (i.e., $a(\infty) = b(\infty)$). All solutions below $x_0(t)$ will eventually enter region II and converge to $-\infty$ along $x = -t$. All solutions above $x_0$ will eventually be above $y_-(t)$ and converge to $+\infty$. To show that they escape to $+\infty$ in finite time we use that

$$\dot{x}(t) = x(t)^2 - t^2 \geq 2 \tag{1.66}$$

for every solutions above $y_-(t)$. Hence $x(t) \geq x_0 + 2(t - t_0)$ and thus there is an $\varepsilon > 0$ such that

$$x(t) \geq \frac{t}{\sqrt{1-\varepsilon}}. \tag{1.67}$$

This implies

$$\dot{x}(t) = x(t)^2 - t^2 \geq x(t)^2 - (1-\varepsilon)x(t)^2 = \varepsilon x(t)^2 \tag{1.68}$$

and every solution $x(t)$ is a super solution to a corresponding solution of

$$\dot{x}(t) = \varepsilon x(t)^2. \tag{1.69}$$

But we already know that the solutions of the last equation escape to $+\infty$ in finite time and so the same must be true for our equation.

In summary, we have shown the following

- There is a unique solution $x_0(t)$ which converges to the line $x = t$.
- All solutions above $x_0(t)$ will eventually converge to $+\infty$ in finite time.
- All solutions below $x_0(t)$ converge to the line $x = -t$.

It is clear that similar considerations can be applied to any first order equation $\dot{x} = f(t, x)$ and one usually can obtain a quite complete picture of the solutions. However, it is important to point out that the reason for our success was the fact hat our equation lives in two dimensions $(t, x) \in \mathbb{R}^2$. If we consider higher order equations or systems of equations, we need more dimensions. At first sight this seems only to imply that we can no longer plot everything, but there is another more severe difference: In $\mathbb{R}^2$ a curve splits our space into two regions: one above and one below the curve. The only way to get from one region into the other is by crossing the curve. In more than two dimensions this is no longer true and this allows for much more complicated behavior of solutions. In fact, equations in three (or more) dimensions will often exhibit *chaotic* behavior which makes a simple description of solutions impossible!

**Problem 1.23.** *Generalize the concept of sub and super solutions to the interval* $(T, t_0)$*, where* $T < t_0$.

**Problem 1.24.** *Discuss the following equations:*

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

(ii) $\dot{x} = x^2 - t$.

Chapter 2

# Initial value problems

Our main task in this section will be to prove the basic existence and uniqueness result for ordinary differential equations. The key ingredient will be the contraction principle (Banach fixed point theorem), which we will derive first.

## 2.1. Fixed point theorems

Let $X$ be a real **vector space**. A **norm** on $X$ is a map $\|.\| : X \to [0, \infty)$ satisfying the following requirements:

(i) $\|0\| = 0$, $\|x\| > 0$ for $x \in X \backslash \{0\}$.

(ii) $\|\lambda x\| = |\lambda| \, \|x\|$ for $\lambda \in \mathbb{R}$ and $x \in X$.

(iii) $\|x + y\| \le \|x\| + \|y\|$ for $x, y \in X$ (**triangle inequality**).

The pair $(X, \|.\|)$ is called a **normed vector space**. Given a normed vector space $X$, we have the concept of convergence and of a **Cauchy sequence** in this space. The normed vector space is called **complete** if every Cauchy sequence converges. A complete normed vector space is called a **Banach space**.

As an example, let $I$ be a compact interval and consider the continuous functions $C(I)$ over this set. They form a vector space if all operations are defined pointwise. Moreover, $C(I)$ becomes a normed space if we define

$$\|x\| = \sup_{t \in I} |x(t)|. \tag{2.1}$$

I leave it as an exercise to check the three requirements from above. Now what about convergence in this space? A sequence of functions $x_n(t)$ converges to $x$ if and only if

$$\lim_{n\to\infty} \|x - x_n\| = \lim_{n\to\infty} \sup_{t\in I} |x_n(t) - x(t)| = 0. \tag{2.2}$$

That is, in the language of real analysis, $x_n$ converges uniformly to $x$. Now let us look at the case where $x_n$ is only a Cauchy sequence. Then $x_n(t)$ is clearly a Cauchy sequence of real numbers for any fixed $t \in I$. In particular, by completeness of $\mathbb{R}$, there is a limit $x(t)$ for each $t$. Thus we get a limiting function $x(t)$. Moreover, letting $m \to \infty$ in

$$|x_n(t) - x_m(t)| \le \varepsilon \qquad \forall n, m > N_\varepsilon,\ t \in I \tag{2.3}$$

we see

$$|x_n(t) - x(t)| \le \varepsilon \qquad \forall n > N_\varepsilon,\ t \in I, \tag{2.4}$$

that is, $x_n(t)$ converges uniformly to $x(t)$. However, up to this point we don't know whether it is in our vector space $C(I)$ or not, that is, whether it is continuous or not. Fortunately, there is a well-known result from real analysis which tells us that the uniform limit of continuous functions is again continuous. Hence $x(t) \in C(I)$ and thus every Cauchy sequence in $C(I)$ converges. Or, in other words, $C(I)$ is a Banach space.

You will certainly ask how all these considerations should help us with our investigation of differential equations? Well, you will see in the next section that it will allow us to give an easy and transparent proof of our basic existence and uniqueness theorem based on the following results of this section.

A **fixed point** of a mapping $K : C \subseteq X \to C$ is an element $x \in C$ such that $K(x) = x$. Moreover, $K$ is called a contraction if there is a contraction constant $\theta \in [0, 1)$ such that

$$\|K(x) - K(y)\| \le \theta \|x - y\|, \qquad x, y \in C. \tag{2.5}$$

We also recall the notation $K^n(x) = K(K^{n-1}(x))$, $K^0(x) = x$.

**Theorem 2.1** (Contraction principle)**.** *Let $C$ be a (nonempty) closed subset of a Banach space $X$ and let $K : C \to C$ be a contraction, then $K$ has a unique fixed point $\overline{x} \in C$ such that*

$$\|K^n(x) - \overline{x}\| \le \frac{\theta^n}{1-\theta} \|K(x) - x\|, \qquad x \in C. \tag{2.6}$$

**Proof.** If $x = K(x)$ and $\tilde{x} = K(\tilde{x})$, then $\|x - \tilde{x}\| = \|K(x) - K(\tilde{x})\| \le \theta \|x - \tilde{x}\|$ shows that there can be at most one fixed point.

Concerning existence, fix $x_0 \in U$ and consider the sequence $x_n = K^n(x_0)$. We have

$$\|x_{n+1} - x_n\| \le \theta \|x_n - x_{n-1}\| \le \cdots \le \theta^n \|x_1 - x_0\| \tag{2.7}$$

and hence by the triangle inequality (for $n > m$)

$$\begin{aligned} \|x_n - x_m\| &\le \sum_{j=m+1}^{n} \|x_j - x_{j-1}\| \le \theta^m \sum_{j=0}^{n-m-1} \theta^j \|x_1 - x_0\| \\ &\le \frac{\theta^m}{1-\theta} \|x_1 - x_0\|. \end{aligned} \tag{2.8}$$

Thus $x_n$ is Cauchy and tends to a limit $\overline{x}$. Moreover,

$$\|K(\overline{x}) - \overline{x}\| = \lim_{n\to\infty} \|x_{n+1} - x_n\| = 0 \tag{2.9}$$

shows that $\overline{x}$ is a fixed point and the estimate (2.6) follows after taking the limit $n \to \infty$ in (2.8). □

Note that the same proof works if we replace $\theta^n$ by any other summable sequence $\theta_n$ (Problem 2.3).

**Theorem 2.2** (Weissinger). *Suppose $K : C \subseteq X \to C$ satisfies*

$$\|K^n(x) - K^n(y)\| \le \theta_n \|x - y\|, \qquad x, y \in C, \tag{2.10}$$

*with $\sum_{n=1}^{\infty} \theta_n < \infty$. Then $K$ has a unique fixed point $\overline{x}$ such that*

$$\|K^n(x) - \overline{x}\| \le \left( \sum_{j=n}^{\infty} \theta_n \right) \|K(x) - x\|, \qquad x \in C. \tag{2.11}$$

**Problem 2.1.** *Show that the space $C(I, \mathbb{R}^n)$ together with the sup norm (2.1) is a Banach space.*

**Problem 2.2.** *Derive Newton's method for finding the zeros of a function $f(x)$,*

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)},$$

*from the contraction principle. What is the advantage/disadvantage of using*

$$x_{n+1} = x_n - \theta \frac{f(x_n)}{f'(x_n)}, \quad \theta > 0,$$

*instead?*

**Problem 2.3.** *Prove Theorem 2.2. Moreover, suppose $K : C \to C$ and that $K^n$ is a contraction. Show that the fixed point of $K^n$ is also one of $K$ (Hint: Use uniqueness). Hence Theorem 2.2 (except for the estimate) can also be considered as a special case of Theorem 2.1 since the assumption implies that $K^n$ is a contraction for $n$ sufficiently large.*

## 2.2. The basic existence and uniqueness result

Now we want to use the preparations of the previous section to show existence and uniqueness of solutions for the following **initial value problem** (IVP)

$$\dot{x} = f(t, x), \quad x(t_0) = x_0. \tag{2.12}$$

We suppose $f \in C(U, \mathbb{R}^n)$, where $U$ is an open subset of $\mathbb{R}^{n+1}$, and $(t_0, x_0) \in U$.

First of all note that integrating both sides with respect to $t$ shows that (2.12) is equivalent to the following **integral equation**

$$x(t) = x_0 + \int_{t_0}^{t} f(s, x(s))\, ds. \tag{2.13}$$

At first sight this does not seem to help much. However, note that $x_0(t) = x_0$ is an approximating solution at least for small $t$. Plugging $x_0(t)$ into our integral equation we get another approximating solution

$$x_1(t) = x_0 + \int_{t_0}^{t} f(s, x_0(s))\, ds. \tag{2.14}$$

Iterating this procedure we get a sequence of approximating solutions

$$x_n(t) = K^n(x_0)(t), \qquad K(x)(t) = x_0 + \int_{t_0}^{t} f(s, x(s))\, ds. \tag{2.15}$$

Now this observation begs us to apply the contraction principle from the previous section to the fixed point equation $x = K(x)$, which is precisely our integral equation (2.13).

To apply the contraction principle, we need to estimate

$$|K(x)(t) - K(y)(t)| \le \int_{t_0}^{t} |f(s, x(s)) - f(s, y(s))| ds. \tag{2.16}$$

Clearly, since $f$ is continuous, we know that $|f(s, x(s)) - f(s, y(s))|$ is small if $|x(s) - y(s)|$ is. However, this is not good enough to estimate the integral above. For this we need the following stronger condition. Suppose $f$ is locally **Lipschitz continuous** in the second argument. That is, for every compact set $V \subset U$ the following number

$$L = \sup_{(t,x) \neq (t,y) \in V} \frac{|f(t, x) - f(t, y)|}{|x - y|} < \infty \tag{2.17}$$

(which depends on $V$) is finite. Now let us choose $V = [t_0 - T, t_0 + T] \times B_\delta(x_0)$, $B_\delta(x_0) = \{x \in \mathbb{R}^n | \, |x - x_0| \le \delta\}$, and abbreviate

$$T_0 = \min(T, \frac{\delta}{M}), \qquad M = \sup_{(t,x) \in V} |f(t, x)|. \tag{2.18}$$

Furthermore, we will set $t_0 = 0$ and $x_0 = 0$ (which can always be achieved by a shift of the coordinate axes) for notational simplicity in the following calculation. Then,

$$\begin{aligned}\left|\int_0^t |f(s,x(s)) - f(s,y(s))|ds\right| &\leq \left|L\int_0^t |x(s)-y(s)|ds\right| \\ &\leq L|t| \sup_{|s|\leq t} |x(s)-y(s)| \qquad (2.19)\end{aligned}$$

provided the graphs of both $x(t)$ and $y(t)$ lie in $V$. Moreover, if the graph of $x(t)$ lies in $V$, the same is true for $K(x)(t)$ since

$$|K(x)(t) - x_0| \leq |t|M \leq \delta \qquad (2.20)$$

for all $|t| \leq T_0$. That is, $K$ maps $C([-T_0, T_0], B_\delta(x_0))$ into itself. Moreover, choosing $T_0 < L^{-1}$ it is even a contraction and existence of a unique solution follows from the contraction principle. However, we can do even a little better. We claim

$$|K^n(x)(t) - K^n(y)(t)| \leq \frac{(L|t|)^n}{n!} \sup_{|s|\leq|t|} |x(s)-y(s)|. \qquad (2.21)$$

Using (2.19)

$$\begin{aligned}|K^{n+1}(x)(t) - K^{n+1}(y)(t)| &\leq \left|\int_0^t |f(s,K^n(x)(s)) - f(s,K^n(y)(s))|ds\right| \\ &\leq L\int_0^{|t|} \frac{(L|s|)^n}{n!} \sup_{|r|\leq|s|} |x(r)-y(r)|ds \\ &\leq \frac{L^{n+1}}{n!} \sup_{|r|\leq|t|} |x(r)-y(r)| \int_0^{|t|} s^n ds \\ &= \frac{(L|t|)^{n+1}}{n!} \sup_{|r|\leq|t|} |x(r)-y(r)| \qquad (2.22)\end{aligned}$$

the claim follows from induction. Hence $K$ satisfies the assumptions of Theorem 2.2 which finally yields

**Theorem 2.3** (Picard-Lindelöf)**.** *Suppose $f \in C(U, \mathbb{R}^n)$, where $U$ is an open subset of $\mathbb{R}^{n+1}$, and $(t_0, x_0) \in U$. If $f$ is locally Lipschitz continuous in the second argument, then there exists a unique local solution $\overline{x}(t)$ of the IVP (2.12).*

*Moreover, let $L$, $T_0$ be defined as before, then*

$$\overline{x} = \lim_{n\to\infty} K^n(x_0) \in C^1([t_0 - T_0, t_0 + T_0], B_\delta(x_0)) \qquad (2.23)$$

*satisfies the estimate*

$$\sup_{|t-t_0|\leq T_0} |\overline{x}(t) - K^n(x_0)(t)| \leq \frac{(LT_0)^n}{n!} \mathrm{e}^{LT_0} \int_{-T_0}^{T_0} |f(t_0+s, x_0)|ds. \qquad (2.24)$$

The procedure to find the solution is called **Picard iteration**. Unfortunately, it is not suitable for actually finding the solution since computing the integrals in each iteration step will not be possible in general. Even for numerical computations it is of no great help, since evaluating the integrals is too time consuming. However, at least we know that there is a unique solution to the initial value problem.

If $f$ is differentiable, we can say even more. In particular, note that $f \in C^1(U, \mathbb{R}^n)$ implies that $f$ is Lipschitz continuous (see the problems below).

**Lemma 2.4.** *Suppose $f \in C^k(U, \mathbb{R}^n)$, $k \geq 1$, where $U$ is an open subset of $\mathbb{R}^{n+1}$, and $(t_0, x_0) \in U$. Then the local solution $\overline{x}$ of the IVP (2.12) is $C^{k+1}$.*

**Proof.** Let $k = 1$. Then $\overline{x}(t) \in C^1$ by the above theorem. Moreover, using $\dot{\overline{x}}(t) = f(t, \overline{x}(t)) \in C^1$ we infer $\overline{x}(t) \in C^2$. The rest follows from induction. □

Finally, let me remark that the requirement that $f$ is continuous in Theorem 2.3 is already more then we actually needed in its proof. In fact, all one needs to require is that

$$L(t) = \sup_{x \neq y \in B_\delta(x_0)} \frac{|f(t,x) - f(t,y)|}{|x - y|} \tag{2.25}$$

is locally integrable (i.e., $\int_I L(t) dt < \infty$ for any compact interval $I$). Choosing $T_0$ so small that $|\int_{t_0}^{t_0 \pm T_0} L(s) ds| < 1$ we have that $K$ is a contraction and the result follows as above.

However, then the solution of the integral equation is only absolutely continuous and might fail to be continuously differentiable. In particular, when going back from the integral to the differential equation, the differentiation has to be understood in a generalized sense. I do not want to go into further details here, but rather give you an example. Consider

$$\dot{x} = \mathrm{sgn}(t) x, \qquad x(0) = 1. \tag{2.26}$$

Then $x(t) = \exp(|t|)$ might be considered a solution even though it is not differentiable at $t = 0$.

**Problem 2.4.** *Are the following functions Lipschitz continuous at 0? If yes, find a Lipschitz constant for some interval containing 0.*

(i) $f(x) = \frac{1}{1-x^2}$.
(ii) $f(x) = |x|^{1/2}$.
(iii) $f(x) = x^2 \sin(\frac{1}{x})$.

**Problem 2.5.** *Show that $f \in C^1(\mathbb{R})$ is locally Lipschitz continuous. In fact, show that*

$$|f(y) - f(x)| \leq \sup_{\varepsilon \in [0,1]} |f'(x + \varepsilon(y-x))||x-y|.$$

*Generalize this result to $f \in C^1(\mathbb{R}^m, \mathbb{R}^n)$.*

**Problem 2.6.** *Apply the Picard iteration to the first order linear equation*

$$\dot{x} = x, \qquad x(0) = 1.$$

**Problem 2.7.** *Investigate uniqueness of the differential equation*

$$\dot{x} = \begin{cases} -t\sqrt{|x|}, & x \geq 0 \\ t\sqrt{|x|}, & x \leq 0 \end{cases}.$$

## 2.3. Dependence on the initial condition

Usually, in applications several data are only known approximately. If the problem is **well-posed**, one expects that small changes in the data will result in small changes of the solution. This will be shown in our next theorem. As a preparation we need **Gronwall's inequality**.

**Lemma 2.5** (Gronwall's inequality)**.** *Suppose $\psi(t) \geq 0$ satisfies*

$$\psi(t) \leq \alpha + \int_0^t \beta(s)\psi(s)ds \tag{2.27}$$

*with $\alpha, \beta(s) \geq 0$. Then*

$$\psi(t) \leq \alpha \exp(\int_0^t \beta(s)ds). \tag{2.28}$$

**Proof.** It suffices to prove the case $\alpha > 0$, since the case $\alpha = 0$ then follows by taking the limit. Now observe

$$\frac{d}{dt} \ln\left(\alpha + \int_0^t \beta(s)\psi(s)ds\right) = \frac{\beta(t)\psi(t)}{\alpha + \int_0^t \beta(s)\psi(s)ds} \leq \beta(t) \tag{2.29}$$

and integrate this inequality with respect to $t$. □

Now we can show that our IVP is well posed.

**Theorem 2.6.** *Suppose $f, g \in C(U, \mathbb{R}^n)$ and let $f$ be Lipschitz continuous with constant $L$. If $x(t)$ and $y(t)$ are the respective solutions of the IVPs*

$$\begin{array}{c} \dot{x} = f(t,x) \\ x(t_0) = x_0 \end{array} \quad \text{and} \quad \begin{array}{c} \dot{y} = g(t,y) \\ y(t_0) = y_0 \end{array}, \tag{2.30}$$

*then*

$$|x(t) - y(t)| \leq |x_0 - y_0| \, \mathrm{e}^{L|t-t_0|} + \frac{M}{L}(\mathrm{e}^{L|t-t_0|} - 1), \tag{2.31}$$

*where*

$$M = \sup_{(t,x)\in U} |f(t,x) - g(t,x)|. \tag{2.32}$$

**Proof.** Without restriction we set $t_0 = 0$. Then we have

$$|x(t) - y(t)| \le |x_0 - y_0| + \int_0^t |f(s,x(s)) - g(s,y(s))|ds. \tag{2.33}$$

Estimating the integrand shows

$$\begin{aligned} &|f(s,x(s)) - g(s,y(s))| \\ &\quad \le |f(s,x(s)) - f(s,y(s))| + |f(s,y(s)) - g(s,y(s))| \\ &\quad \le L|x(s) - y(s)| + M. \end{aligned} \tag{2.34}$$

Setting

$$\psi(t) = |x(t) - y(t)| + \frac{M}{L} \tag{2.35}$$

and applying Gronwall's inequality finishes the proof. □

In particular, denote the solution of the IVP (2.12) by

$$\phi(t, x_0) \tag{2.36}$$

to emphasize the dependence on the initial condition. Then our theorem, in the special case $f = g$,

$$|\phi(t,x_0) - \phi(t,x_1)| \le |x_0 - x_1| \, \mathrm{e}^{L|t|}, \tag{2.37}$$

shows that $\phi$ depends continuously on the initial value. However, in many cases this is not good enough and we need to be able to differentiate with respect to the initial condition.

We first suppose that $\phi(t,x)$ is differentiable with respect to $x$. Then, by differentiating (2.12), its derivative

$$\frac{\partial\phi(t,x)}{\partial x} \tag{2.38}$$

necessarily satisfies the **first variational equation**

$$\dot{y} = A(t,x)y, \qquad A(t,x) = \frac{\partial f(t,\phi(t,x))}{\partial x}, \tag{2.39}$$

which is linear. The corresponding integral equation reads

$$y(t) = \mathbb{I} + \int_{t_0}^t A(s,x)y(s)ds, \tag{2.40}$$

where we have used $\phi(t_0,x) = x$ and hence $\frac{\partial\phi(t_0,x)}{\partial x} = \mathbb{I}$. Applying similar fixed point techniques as before, one can show that the first variational equation has a solution which is indeed the derivative of $\phi(t,x)$ with respect to $x$. The details are deferred to Section 2.6 at the end of this chapter and we only state the final result (see Corollary 2.21).

**Theorem 2.7.** *Suppose $f \in C(U, \mathbb{R}^n)$, is Lipschitz continuous. Around each point $(t_0, x_0) \in U$ we can find an open set $I \times V \subseteq U$ such that $\phi(t, x) \in C(I \times V, \mathbb{R}^n)$.*

*Moreover, if $f \in C^k(U, \mathbb{R}^n)$, $k \geq 1$, then $\phi(t, x) \in C^k(I \times V, \mathbb{R}^n)$.*

In fact, we can also handle the dependence on parameters. Suppose $f$ depends on some parameters $\lambda \in \Lambda \subseteq \mathbb{R}^p$ and consider the IVP

$$\dot{x}(t) = f(t, x, \lambda), \qquad x(t_0) = x_0, \tag{2.41}$$

with corresponding solution

$$\phi(t, x_0, \lambda). \tag{2.42}$$

**Theorem 2.8.** *Suppose $f \in C^k(U \times \Lambda, \mathbb{R}^n)$, $x_0 \in C^k(\Lambda, V)$, $k \geq 1$. Around each point $(t_0, x_0, \lambda_0) \in U \times \Lambda$ we can find an open set $I_0 \times V_0 \times \Lambda_0 \subseteq U \times \Lambda$ such that $\phi(t, x, \lambda) \in C^k(I_0 \times V_0 \times \Lambda_0, \mathbb{R}^n)$.*

**Proof.** This follows from the previous result by adding the parameters $\lambda$ to the dependent variables and requiring $\dot{\lambda} = 0$. Details are left to the reader. (It also follows directly from Corollary 2.21.) □

**Problem 2.8** (Generalized Gronwall). *Suppose $\psi(t)$ satisfies*

$$\psi(t) \leq \alpha(t) + \int_0^t \beta(s)\psi(s)ds$$

*with $\beta(t) \geq 0$ and that $\psi(t) - \alpha(t)$ is continuous. Show that*

$$\psi(t) \leq \alpha(t) + \int_0^t \alpha(s)\beta(s) \exp\left(\int_s^t \beta(r)dr\right) ds.$$

*Moreover, if $\alpha(s) \leq \alpha(t)$ for $s \leq t$, then*

$$\psi(t) \leq \alpha(t) \exp\left(\int_0^t \beta(s)ds\right).$$

*Hint: Denote the right hand side of the above inequality by $\phi(t)$ and show that it satisfies*

$$\phi(t) = \alpha(t) + \int_0^t \beta(s)\phi(s)ds.$$

*Then consider $\Delta(t) = \psi(t) - \phi(t)$ which is continuous and satisfies*

$$\Delta(t) \leq \int_0^t \beta(s)\Delta(s)ds.$$

**Problem 2.9.** *In which case does the inequality in (2.31) become an equality?*

## 2.4. Extensibility of solutions

We have already seen that solutions might not exist for all $t \in \mathbb{R}$ even though the differential equation is defined for all $t \in \mathbb{R}$. This raises the question about the maximal interval on which a solution can be defined.

Suppose that solutions of the IVP (2.12) exist locally and are unique (e.g., $f$ is Lipschitz). Let $\phi_1$, $\phi_2$ be two solutions of the IVP (2.12) defined on the open intervals $I_1$, $I_2$, respectively. Let $I = I_1 \cap I_2 = (T_-, T_+)$ and let $(t_-, t_+)$ be the maximal open interval on which both solutions coincide. I claim that $(t_-, t_+) = (T_-, T_+)$. In fact, if $t_+ < T_+$, both solutions would also coincide at $t_+$ by continuity. Next, considering the IVP $x(t_+) = \phi_1(t_+) = \phi_2(t_+)$ shows that both solutions coincide in a neighborhood of $t_+$ by Theorem 2.3. This contradicts maximality of $t_+$ and hence $t_+ = T_+$. Similarly, $t_- = T_-$.

In summary, uniqueness is equivalent to saying that two solution curves $t \mapsto (t, x_j(t))$, $j = 1, 2$, always coincide on their common domain of definition.

Moreover, we get a solution

$$\phi(t) = \begin{cases} \phi_1(t), & t \in I_1 \\ \phi_2(t), & t \in I_2 \end{cases} \tag{2.43}$$

defined on $I_1 \cup I_2$. In fact, this even extends to an arbitrary number of solutions and in this way we get a (unique) solution defined on some maximal interval $I_{(t_0,x_0)}$.

If we drop uniqueness of solutions, two given solutions of the IVP (2.12) can still be glued together at $t_0$ (if necessary) to obtain a solution defined on $I_1 \cup I_2$. Furthermore, Zorn's lemma even ensures existence of maximal solutions in this case.

Now let us look at how we can tell from a given solution whether an extension exists or not.

**Lemma 2.9.** *Let $\phi(t)$ be a solution of (2.12) defined on the interval $(t_-, t_+)$. Then there exists an extension to the interval $(t_-, t_+ + \varepsilon)$ for some $\varepsilon > 0$ if and only if*

$$\lim_{t \uparrow t_+} (t, \phi(t)) = (t_+, y) \in U \tag{2.44}$$

*exists. Similarly for $t_-$.*

**Proof.** Clearly, if there is an extension, the limit (2.44) exists. Conversely, suppose (2.44) exists. Then, by Theorem 2.14 below there is a solution $\tilde{\phi}(t)$ of the IVP $x(t_+) = y$ defined on the interval $(t_+ - \varepsilon, t_+ + \varepsilon)$. As before, we can glue $\phi(t)$ and $\tilde{\phi}(t)$ at $t_+$ to obtain a solution defined on $(t_-, t_+ + \varepsilon)$. $\square$

Our final goal is to show that solutions exist for all $t \in \mathbb{R}$ if $f(t,x)$ grows at most linearly with respect to $x$. But first we need a better criterion which does not require a complete knowledge of the solution.

**Lemma 2.10.** *Let $\phi(t)$ be a solution of (2.12) defined on the interval $(t_-, t_+)$. Suppose there is a compact set $[t_0, t_+] \times C \subset U$ such that $\phi(t) \in C$ for all $t \in [t_0, t_+)$, then there exists an extension to the interval $(t_-, t_+ + \varepsilon)$ for some $\varepsilon > 0$.*

*In particular, if there is such a compact set $C$ for every $t_+ > t_0$ ($C$ might depend on $t_+$), then the solution exists for all $t > t_0$.*

*Similarly for $t_-$.*

**Proof.** Let $t_n \to t_+$. It suffices to show that $\phi(t_n)$ is Cauchy. This follows from

$$|\phi(t_n) - \phi(t_m)| \leq \left| \int_{t_m}^{t_n} f(s, \phi(s)) ds \right| \leq M |t_n - t_m|, \tag{2.45}$$

where $M = \sup_{[t_0,t_+]\times C} f(t,x) < \infty$. □

Note that this result says that

**Corollary 2.11.** *If $T_+ < \infty$, then the solution must leave every compact set $C$ with $[t_0, T_+) \times C \subset U$ as $t$ approaches $T_+$. In particular, if $U = \mathbb{R} \times \mathbb{R}^n$, the solution must tend to infinity as $t$ approaches $T_+$.*

Now we come to the proof of our anticipated result.

**Theorem 2.12.** *Suppose $U = \mathbb{R} \times \mathbb{R}^n$ and for every $T > 0$ there are constants $M(T)$, $L(T)$ such that*

$$|f(t,x)| \leq M(T) + L(T)|x|, \qquad (t,x) \in [-T,T] \times \mathbb{R}^n. \tag{2.46}$$

*Then all solutions of the IVP (2.12) are defined for all $t \in \mathbb{R}$.*

**Proof.** Using the above estimate for $f$ we have ($t_0 = 0$ without loss of generality)

$$|\phi(t)| \leq |x_0| + \int_0^t (M + L|\phi(s)|) ds, \quad t \in [0,T] \cap I. \tag{2.47}$$

Setting $\psi(t) = \frac{M}{L} + |\phi(t)|$ and applying Gronwall's inequality shows

$$|\phi(t)| \leq |x_0| \mathrm{e}^{LT} + \frac{M}{L}(\mathrm{e}^{LT} - 1). \tag{2.48}$$

Thus $\phi$ lies in a compact ball and the result follows by the previous lemma. □

Again, let me remark that it suffices to assume

$$|f(t,x)| \le M(t) + L(t)|x|, \qquad x \in \mathbb{R}^n, \tag{2.49}$$

where $M(t)$, $L(t)$ are locally integrable (however, for the proof you now need the generalized Gronwall inequality from Problem 2.8).

**Problem 2.10.** *Show that Theorem 2.12 is false (in general) if the estimate is replaced by*

$$|f(t,x)| \le M(T) + L(T)|x|^\alpha$$

*with $\alpha > 1$.*

**Problem 2.11.** *Consider a first order autonomous system with $f(x)$ Lipschitz. Show that $x(t)$ is a solution if and only if $x(t - t_0)$ is. Use this and uniqueness to show that for two maximal solutions $x_j(t)$, $j = 1, 2$, the images $\gamma_j = \{x_j(t) | t \in I_j\}$ either coincide or are disjoint.*

**Problem 2.12.** *Consider a first order autonomous system in $\mathbb{R}^1$ with $f(x)$ Lipschitz. Suppose $f(0) = f(1) = 0$. Show that solutions starting in $[0, 1]$ cannot leave this interval. What is the maximal interval of definition for solutions starting in $[0, 1]$?*

**Problem 2.13.** *Consider a first order system in $\mathbb{R}^1$ with $f(t,x)$ defined on $\mathbb{R} \times \mathbb{R}$. Suppose $xf(t,x) < 0$ for $|x| > R$. Show that all solutions exists for all $t \in \mathbb{R}$.*

## 2.5. Euler's method and the Peano theorem

In this section we want to show that continuity of $f(t,x)$ is sufficient for existence of at least one solution of the initial value problem (2.12). If $\phi(t)$ is a solution, then by Taylor's theorem we have

$$\phi(t_0 + h) = x_0 + \dot{\phi}(t_0)h + o(h) = x_0 + f(t_0, x_0)h + o(h). \tag{2.50}$$

This suggests to define an approximate solution by omitting the error term and applying the procedure iteratively. That is, we set

$$x_h(t_{n+1}) = x_h(t_n) + f(t_n, x_h(t_n))h, \qquad t_n = t_0 + nh, \tag{2.51}$$

and use linear interpolation in between. This procedure is known as **Euler method**.

We expect that $x_h(t)$ converges to a solution as $h \downarrow 0$. But how should we prove this? Well, the key observation is that, since $f$ is continuous, it is bounded by a constant on each compact interval. Hence the derivative of $x_h(t)$ is bounded by the same constant. Since this constant is independent of $h$, the functions $x_h(t)$ form an equicontinuous family of functions which converges uniformly after maybe passing to a subsequence by the Arzelà-Ascoli theorem.

**Theorem 2.13** (Arzelà-Ascoli)**.** *Suppose the sequence of functions $f_n(x)$, $n \in \mathbb{N}$, on a compact interval is (uniformly) equicontinuous, that is, for every $\varepsilon > 0$ there is a $\delta > 0$ (independent of $n$) such that*

$$|f_n(x) - f_n(y)| \leq \varepsilon \quad \textit{if} \quad |x - y| < \delta. \tag{2.52}$$

*If the sequence $f_n$ is bounded, then there is a uniformly convergent subsequence.*

**Proof.** Let $\{x_j\}_{j=1}^{\infty}$ be a dense subset of our interval (e.g., all rational in this set). Since $f_n(x_j)$ is bounded, we can choose a convergent subsequence $f_n^{(j)}(x_j)$ (Bolzano-Weierstraß). The diagonal subsequence $\tilde{f}_n = f_n^{(n)}(x)$ will hence converge for all $x = x_j$. We will show that it converges uniformly for all $x$:

Fix $\varepsilon > 0$ and chose $\delta$ such that $|f_n(x) - f_n(y)| \leq \frac{\varepsilon}{3}$ for $|x - y| < \delta$. The balls $B_\delta(x_j)$ cover our interval and by compactness even finitely many, say $1 \leq j \leq p$ suffice. Furthermore, choose $N_\varepsilon$ such that $|\tilde{f}_m(x_j) - \tilde{f}_n(x_j)| \leq \varepsilon$ for $n, m \geq N_\varepsilon$ and $1 \leq j \leq p$.

Now pick $x$ and note that $x \in B_\delta(x_j)$ for some $j$. Thus

$$|\tilde{f}_m(x) - \tilde{f}_n(x)| \leq |\tilde{f}_m(x) - \tilde{f}_m(x_j)| + |\tilde{f}_m(x_j) - \tilde{f}_n(x_j)| + |\tilde{f}_n(x_j) - \tilde{f}_n(x)| \leq \varepsilon \tag{2.53}$$

for $n, m \geq N_\varepsilon$, which shows that $\tilde{f}_n$ is Cauchy with respect to the maximum norm. □

More precisely, pick $\delta, T > 0$ such that $\tilde{U} = [t_0, t_0 + T] \times B_\delta(x_0) \subset U$ and let

$$M = \max_{(t,x) \in \tilde{U}} |f(t, x)|. \tag{2.54}$$

Then $x_h(t) \in B_\delta(x_0)$ for $t \in [t_0, t_0 + T_0]$, where $T_0 = \min\{T, \frac{\delta}{M}\}$, and

$$|x_h(t) - x_h(s)| \leq M|t - s|. \tag{2.55}$$

Hence the family $x_{1/n}(t)$ is equicontinuous and there is a uniformly convergent subsequence $\phi_n(t) \to \phi(t)$. It remains to show that the limit $\phi(t)$ solves our initial value problem (2.12). We will show this by verifying that the corresponding integral equation holds. Using that $f$ is uniformly continuous on $\tilde{U}$, we can find $\delta(h) \to 0$ as $h \to 0$ such that

$$|f(s, y) - f(t, x)| \leq \delta(h) \quad \text{for} \quad |y - x| \leq Mh, \ |s - t| \leq h. \tag{2.56}$$

Writing

$$x_h(t) = x_0 + \sum_{j=0}^{n-1} \int_{t_j}^{t_{j+1}} \chi(s) f(t_j, x_h(t_j)) ds, \tag{2.57}$$

where $\chi(s) = 1$ for $s \in [t_0, t]$ and $\chi(s) = 0$ else, we obtain

$$\begin{aligned}
&\left| x_h(t) - x_0 - \int_{t_0}^{t} f(s, x_h(s)) ds \right| \\
&\quad \le \sum_{j=0}^{n-1} \int_{t_j}^{t_{j+1}} \chi(s) |f(s, x_h(s)) - f(t_j, x_h(t_j))| ds \\
&\quad \le \delta(h) \sum_{j=0}^{n-1} \int_{t_j}^{t_{j+1}} \chi(s) ds = |t - t_0| \delta(h).
\end{aligned} \tag{2.58}$$

From this it follows that $\phi$ is indeed a solution

$$\phi(t) = \lim_{n\to\infty} \phi_n(t) = x_0 + \lim_{n\to\infty} \int_{t_0}^{t} f(s, \phi_n(s)) ds = x_0 + \int_{t_0}^{t} f(s, \phi(s)) ds \tag{2.59}$$

since we can interchange limit and integral by uniform convergence.

Hence we have proven **Peano's theorem**.

**Theorem 2.14** (Peano)**.** *Suppose $f$ is continuous on $\tilde{U} = [t_0, t_0 + T] \times B_\delta(x_0)$ and denote its maximum by $M$. Then there exists at least one solution of the initial value problem (2.12) for $t \in [t_0, t_0 + T_0]$, where $T_0 = \min\{T, \frac{\delta}{M}\}$. The analogous result holds for the interval $[t_0 - T, t_0]$.*

Finally, let me remark that the Euler algorithm is well suited for the numerical computation of an approximate solution since it only requires the evaluation of $f$ at certain points. On the other hand, it is not clear how to find the converging subsequence, and so let us show that $x_h(t)$ converges uniformly if $f$ is Lipschitz. In this case we can choose $\delta(h) = LMh$ and our above estimate reads

$$\|x_h - K(x_h)\| \le T_0 LMh, \qquad t \in [t_0, t_0 + T_0], \tag{2.60}$$

using the same notation as in the proof of Theorem 2.3. By (2.21) this yields

$$\begin{aligned}
\|x_h - K^n(x_h)\| &\le \sum_{j=0}^{n-1} \|K^j(x_h) - K^{j+1}(x_h)\| \\
&\le \|x_h - K(x_h)\| \sum_{j=0}^{n-1} \frac{(LT_0)^j}{j!}.
\end{aligned} \tag{2.61}$$

and taking $n \to \infty$ we finally obtain

$$\|x_h - \phi\| \le T_0 LM \mathrm{e}^{LT_0} h, \qquad t \in [t_0, t_0 + T_0]. \tag{2.62}$$

Thus we have a simple numerical method for computing solutions plus an error estimate. However, in practical computations one usually uses some heuristic error estimates, e.g., by performing each step using two step sizes

$h$ and $\frac{h}{2}$. If the difference between the two results becomes too big, the step size is reduced and the last step is repeated.

Of course the Euler algorithm is not the most effective one available today. Usually one takes more terms in the Taylor expansion and approximates all differentials by their difference quotients. The resulting algorithm will converge faster, but it will also involve more calculations in each step. A good compromise is usually a method, where one approximates $\phi(t_0 + h)$ up to the fourth order in $h$. The resulting algorithm

$$x(t_{n+1}) = x(t_n) + \frac{h}{6}(k_{1,n} + 2k_{2,n} + 2k_{3,n} + k_{4,n}),$$

where

$$\begin{array}{ll} k_{1,n} = f(t_n, x(t_n)) & k_{2,n} = f(t_n + \frac{h}{2}, x(t_n) + \frac{k_{1,n}}{2}) \\ k_{3,n} = f(t_n + \frac{h}{2}, x(t_n) + \frac{k_{2,n}}{2}) & k_{4,n} = f(t_{n+1}, x(t_n) + k_{3,n}) \end{array}, \qquad (2.63)$$

is called **Runge-Kutta algorithm**. For even better methods see the literature on numerical methods for ordinary differential equations.

**Problem 2.14.** *Compute the solution of the initial value problem $\dot{x} = x$, $x(0) = 1$, using the Euler and Runge-Kutta algorithm with step size $h = 10^{-1}$. Compare the results with the exact solution.*

## 2.6. Appendix: Volterra integral equations

I hope that, after the previous sections, you are by now convinced that integral equations are an important tool in the investigation of differential equations. Moreover, the proof of Theorem 2.7 requires a result from the theory of Volterra integral equations which we will show in this section. The results are somewhat technical and can be omitted.

The main ingredient will again be fixed point theorems. But now we need the case where our fixed point equation depends on additional parameters $\lambda \in \Lambda$, where $\Lambda$ is a subset of some Banach space.

**Theorem 2.15** (Uniform contraction principle). *Suppose $K_\lambda : C \to C$ is a uniform contraction, that is,*

$$\|K_\lambda(x) - K_\lambda(y)\| \le \theta\|x - y\|, \qquad x, y \in C, 0 \le \theta < 1, \lambda \in \Lambda, \qquad (2.64)$$

*and $K_\lambda(x)$ is continuous with respect to $\lambda$ for every $x \in C$. Then the unique fixed point $\overline{x}(\lambda)$ is continuous with respect to $\lambda$.*

*Moreover, if $\lambda_n \to \lambda$, then*

$$x_{n+1} = K_{\lambda_n}(x_n) \to \overline{x}(\lambda). \qquad (2.65)$$

**Proof.** We first show that $\overline{x}(\lambda)$ is continuous. By the triangle inequality we have

$$\begin{aligned}\|\overline{x}(\lambda)-\overline{x}(\eta)\| &= \|K_\lambda(\overline{x}(\lambda))-K_\eta(\overline{x}(\eta))\|\\ &\le \theta\|\overline{x}(\lambda)-\overline{x}(\eta)\|+\|K_\lambda(\overline{x}(\eta))-K_\eta(\overline{x}(\eta))\| \quad (2.66)\end{aligned}$$

and hence

$$\|\overline{x}(\lambda)-\overline{x}(\eta)\| \le \frac{1}{1-\theta}\|K_\lambda(\overline{x}(\eta))-K_\eta(\overline{x}(\eta))\|. \quad (2.67)$$

Since the right hand side converges to zero as $\lambda \to \eta$ so does the left hand side and thus $\overline{x}(\lambda)$ is continuous.

Abbreviate $\Delta_n = \|x_n - \overline{x}(\lambda)\|$, $\varepsilon_n = \|\overline{x}(\lambda_n)-\overline{x}(\lambda)\|$ and observe

$$\begin{aligned}\Delta_{n+1} &\le \|x_{n+1}-\overline{x}(\lambda_n)\|+\|\overline{x}(\lambda_n)-\overline{x}(\lambda)\| \le \theta\|x_n-\overline{x}(\lambda_n)\|+\varepsilon_n\\ &\le \theta\Delta_n+(1+\theta)\varepsilon_n. \quad (2.68)\end{aligned}$$

Hence

$$\Delta_n \le \theta^n\Delta_0+(1+\theta)\sum_{j=1}^{n}\theta^{n-j}\varepsilon_{j-1} \quad (2.69)$$

which converges to 0 since $\varepsilon_n$ does (show this). □

There is also a uniform version of Theorem 2.2.

**Theorem 2.16.** *Suppose $K_\lambda : C \to C$ is continuous with respect to $\lambda$ for every $x \in C$ and satisfies*

$$\|K_{\lambda_n}\circ\cdots\circ K_{\lambda_1}(x)-K_{\lambda_n}\circ\cdots\circ K_{\lambda_1}(y)\| \le \theta_n\|x-y\|, \quad x,y\in C, \lambda_j\in\Lambda, \quad (2.70)$$

*with $\sum_{n=1}^\infty \theta_n < \infty$. Then the unique fixed point $\overline{x}(\lambda)$ is continuous with respect to $\lambda$.*

*Moreover, if $\lambda_n \to \lambda$, then*

$$x_{n+1} = K_{\lambda_n}(x_n) \to \overline{x}(\lambda). \quad (2.71)$$

**Proof.** We first show that $K_{\underline{\lambda}} = K_{\lambda_n}\circ\cdots\circ K_{\lambda_1}$, $\underline{\lambda} = (\lambda_1,\dots,\lambda_n)$, is continuous with respect to $\underline{\lambda}\in\Lambda^n$. The claim holds for $n=1$ by assumption. It remains to show it holds for $n$ provided it holds for $n-1$. But this follows from

$$\begin{aligned}&\|K_{\lambda_n}\circ K_{\underline{\lambda}}(x)-K_{\eta_n}\circ K_{\underline{\eta}}(x)\|\\ &\quad\le \|K_{\lambda_n}\circ K_{\underline{\lambda}}(x)-K_{\lambda_n}\circ K_{\underline{\eta}}(x)\|+\|K_{\lambda_n}\circ K_{\underline{\eta}}(x)-K_{\eta_n}\circ K_{\underline{\eta}}(x)\|\\ &\quad\le \theta_1\|K_{\underline{\lambda}}(x)-K_{\underline{\eta}}(x)\|+\|K_{\lambda_n}\circ K_{\underline{\eta}}(x)-K_{\eta_n}\circ K_{\underline{\eta}}(x)\|. \quad (2.72)\end{aligned}$$

Now observe that for $n$ sufficiently large we have $\theta_n < 1$ and hence $K_{\underline{\lambda}}$ is a uniform contraction to which we can apply Theorem 2.15. In particular, choosing $\underline{\lambda}_j = (\lambda_j, \dots, \lambda_{j+n-1})$ we have that $x_{n(j+1)+l} = K_{\underline{\lambda}_j}(x_{nj+l})$ converges to the unique fixed point of $K_{(\lambda,\dots,\lambda)}$ which is precisely $\overline{x}(\lambda)$. Hence $\lim_{j\to\infty} x_{nj+l} = \overline{x}(\lambda)$ for every $0 \le l \le n-1$ implying $\lim_{j\to\infty} x_j = \overline{x}(\lambda)$. □

Now we are ready to apply these results to integral equations. However, the proofs require some results from integration theory which I state first.

**Theorem 2.17** (Dominated convergence)**.** *Suppose $f_n(x)$ is a sequence of integrable functions converging pointwise to an integrable function $f(x)$. If there is a* **dominating function** *$g(x)$, that is, $g(x)$ is integrable and satisfies*

$$|f_n(x)| \le g(x), \tag{2.73}$$

*then*

$$\lim_{n\to\infty} \int f_n(x)dx = \int f(x)dx. \tag{2.74}$$

For a proof see any book on real analysis or measure theory.

This result has two immediate consequences which we will need below.

**Corollary 2.18.** *Suppose $f_n(x) \to f(x)$ pointwise and $df_n(x) \to g(x)$ pointwise. If there is (locally) a dominating function for $df_n(x)$, then $f(x)$ is differentiable and $df(x) = g(x)$.*

**Proof.** It suffices to prove the case where $f$ is one dimensional. Using

$$f_n(x) = f_n(x_0) + \int_{x_0}^{x} f_n'(t)dt \tag{2.75}$$

the result follows after taking the limit on both sides. □

**Corollary 2.19.** *Suppose $f(x,\lambda)$ is integrable with respect to $x$ for any $\lambda$ and continuously differentiable with respect to $\lambda$ for any $x$. If there is a dominating function $g(x)$ such that*

$$|\frac{\partial f}{\partial\lambda}(x,\lambda)| \le g(x), \tag{2.76}$$

*then the function*

$$F(\lambda) = \int f(x,\lambda)dx \tag{2.77}$$

*is continuously differentiable with derivative given by*

$$\frac{\partial F}{\partial\lambda}(\lambda) = \int \frac{\partial f}{\partial\lambda}(x,\lambda)dx. \tag{2.78}$$

**Proof.** Again it suffices to consider one dimension. Since

$$f(x,\lambda+\varepsilon) - f(x,\lambda) = \varepsilon \int_0^1 f'(x,\lambda+\varepsilon t)dt \tag{2.79}$$

we have

$$\frac{F(\lambda+\varepsilon)-F(\lambda)}{\varepsilon} = \iint_0^1 f'(x,\lambda+\varepsilon t) dt\, dx. \tag{2.80}$$

Moreover, by $|f'(x,\lambda+\varepsilon t)| \le g(x)$ we have

$$\lim_{\varepsilon\to 0} \int_0^1 f'(x,\lambda+\varepsilon t) dt = f'(x,\lambda) \tag{2.81}$$

by the dominated convergence theorem. Applying dominated convergence again, note $|\int_0^1 f'(x,\lambda+\varepsilon t)dt| \le g(x)$, the claim follows. □

Now let us turn to integral equations. Suppose $U$ is an open subset of $\mathbb{R}^n$ and consider the following (nonlinear) **Volterra integral equation**

$$K_\lambda(x)(t) = k(t,\lambda) + \int_0^t K(s,x(s),\lambda) ds, \tag{2.82}$$

where

$$k \in C(I\times\Lambda, U), \qquad K \in C(I\times U\times\Lambda, \mathbb{R}^n), \tag{2.83}$$

with $I = [-T,T]$ and $\Lambda \subset \mathbb{R}^n$ compact. We will require that there is a constant $L$ (independent of $t$ and $\lambda$) such that

$$|K(t,x,\lambda) - K(t,y,\lambda)| \le L|x-y|, \qquad x,y\in U. \tag{2.84}$$

By the results of the previous section we know that there is a unique solution $\overline{x}(t,\lambda)$ for fixed $\lambda$. The following result shows that it is even continuous and also differentiable if $k$ and $K$ are.

**Theorem 2.20.** *Let $K_\lambda$ satisfy the requirements from above and let $T_0 = \min(T, \frac{\delta}{M})$, where $\delta > 0$ is such that*

$$C_\delta = \{B_\delta(k(t,\lambda))\,|(t,\lambda)\in[T,T]\times\Lambda\} \subset U \tag{2.85}$$

*and*

$$M = \sup_{(t,x,\lambda)\in[-T,T]\times B_\delta(0)\times\Lambda} |K(t,k(t,\lambda)+x,\lambda)|. \tag{2.86}$$

*Then the integral equation $K_\lambda(x) = x$ has a unique solution $\overline{x}(t,\lambda) \in C([-T_0,T_0]\times\Lambda, U)$ satisfying*

$$|\overline{x}(t,\lambda) - k(t,\lambda)| \le \mathrm{e}^{LT_0} \sup_{\lambda\in\Lambda} \int_{-T_0}^{T_0} |K(s,k(s,\lambda),\lambda)| ds. \tag{2.87}$$

*Moreover, if in addition all partial derivatives of order up to $r$ with respect to $\lambda$ and $x$ of $k(t,\lambda)$ and $K(t,x,\lambda)$ are continuous, then all partial derivatives of order up to $r$ with respect to $\lambda$ of $\overline{x}(t,\lambda)$ are continuous as well.*

**Proof.** First observe that it is no restriction to assume $k(t,\lambda) \equiv 0$ by changing $K(t,x,\lambda)$ and $U$. Then existence and the bound follows as in the previous section from Theorem 2.2. By the dominated convergence theorem $K_\lambda(x)$ is continuous with respect to $\lambda$ for fixed $x(t)$. Hence the second term in

$$|\overline{x}(t,\lambda) - \overline{x}(s,\eta)| \le |\overline{x}(t,\lambda) - \overline{x}(s,\lambda)| + |\overline{x}(s,\lambda) - \overline{x}(s,\eta)| \tag{2.88}$$

converges to zero as $(t,\lambda) \to (s,\eta)$ and so does the first since

$$|\overline{x}(t,\lambda) - \overline{x}(s,\lambda)| \le |\int_s^t K(r,\overline{x}(r,\lambda),\lambda)dr| \le M|t-s|. \tag{2.89}$$

Now let us turn to the second claim. Suppose that $\overline{x}(t,\lambda) \in C^1$, then $\overline{y}(t,\lambda) = \frac{\partial}{\partial\lambda}\overline{x}(t,\lambda)$ is a solution of the fixed point equation $\tilde{K}_\lambda(\overline{x}(\lambda),y) = y$. Here

$$\tilde{K}_\lambda(x,y)(t) = \int_0^t K_\lambda(s,x(s),\lambda)ds + \int_0^t K_x(s,x(s),\lambda)y(s)ds, \tag{2.90}$$

where the subscripts denote partial derivatives. This integral operator is linear with respect to $y$ and by the mean value theorem and (2.84) we have

$$\|K_x(t,x,\lambda)\| \le L. \tag{2.91}$$

Hence the first part implies existence of a continuous solution $\overline{y}(t,\lambda)$ of $\tilde{K}_\lambda(\overline{x}(\lambda),y) = y$. It remains to show that this is indeed the derivative of $\overline{x}(\lambda)$.

Fix $\lambda$. Starting with $(x_0(t),y_0(t)) = (0,0)$ we get a sequence $(x_{n+1},y_{n+1}) = (K_\lambda(x_n),\tilde{K}_\lambda(x_n,y_n))$ such that $y_n(t) = \frac{\partial}{\partial\lambda}x_n(t)$. Since $\tilde{K}_\lambda$ is continuous with respect to $x$ (Problem 2.16), Theorem 2.16 implies $(x_n,y_n) \to (\overline{x}(\lambda),\overline{y}(\lambda))$. Moreover, since $(x_n,y_n)$ is uniformly bounded with respect to $\lambda$, we conclude by Corollary 2.18 that $\overline{y}(\lambda)$ is indeed the derivative of $\overline{x}(\lambda)$.

This settles the $r=1$ case. Now suppose the claim holds for $r-1$. Since the equation for $y$ is of the same type as the one for $x$ and since $k_\lambda, K_\lambda, K_x \in C^{r-1}$ we can conclude $y \in C^{r-1}$ and hence $x \in C^r$. □

**Corollary 2.21.** *Let $K_\lambda$ satisfy the requirements from above. If in addition $k \in C^r(I \times \Lambda, V)$ and $K \in C^r(I \times V \times \Lambda, \mathbb{R}^n)$ then $\overline{x}(t,\lambda) \in C^r(I \times \Lambda, V)$.*

**Proof.** The case $r=0$ follows from the above theorem. Now let $r=1$. Differentiating the fixed point equation with respect to $t$ we see that

$$\dot{\overline{x}}(t,\lambda) = \dot{k}(t,\lambda) + K(t,\overline{x}(t,\lambda),\lambda) \tag{2.92}$$

is continuous. Hence, together with the result from above, all partial derivatives exist and are continuous, implying $\overline{x} \in C^1$. The case for general $r$ now follows by induction as in the proof of the above theorem. □

**Problem 2.15.** *Suppose* $K : C \subseteq X \to C$ *is a contraction and*

$$x_{n+1} = K(x_n) + y_n, \qquad \|y_n\| \le \alpha_n + \beta_n \|x_n\|, \tag{2.93}$$

*with* $\lim_{n\to\infty} \alpha_n = \lim_{n\to\infty} \beta_n = 0$. *Then* $\lim_{n\to\infty} x_n = \overline{x}$.

**Problem 2.16.** *Suppose* $K(t, x, y)$ *is a continuous function. Show that the map*

$$K_x(y)(t) = \int_0^t K(s, x(s), y(s)) ds$$

*is continuous with respect to* $x \in C(I, \mathbb{R}^n)$. *Conclude that (2.90) is continuous with respect to* $x \in C(I, \mathbb{R}^n)$. *(Hint: Use the dominated convergence theorem.)*

*Chapter 3*

# Linear equations

## 3.1. Preliminaries from linear algebra

This chapter requires several advanced concepts from linear algebra. In particular, the exponential of a matrix and the Jordan canonical form. Hence I review some necessary facts first. If you feel familiar with these topics, you can move on directly to the next section.

We will use $\mathbb{C}^n$ rather than $\mathbb{R}^n$ as underlying vector space since $\mathbb{C}$ is algebraically closed. Let $A$ be a complex matrix acting on $\mathbb{C}^n$. Introducing the **matrix norm**

$$\|A\| = \sup_{x\colon |x|=1} |Ax| \tag{3.1}$$

it is not hard to see that the space of $n$ by $n$ matrices becomes a Banach space.

The most important object we will need in the study of linear autonomous differential equations is the **matrix exponential** of $A$. It is given by

$$\exp(A) = \sum_{j=0}^{\infty} \frac{1}{j!} A^j \tag{3.2}$$

and, as in the case $n = 1$, one can show that this series converges for all $t \in \mathbb{R}$. However, note that in general

$$\exp(A + B) \neq \exp(A)\exp(B) \tag{3.3}$$

unless $A$ and $B$ commute, that is, unless the **commutator**

$$[A, B] = AB - BA \tag{3.4}$$

vanishes.

In order to understand the structure of $\exp(A)$, we need the Jordan canonical form which we recall next.

Consider a decomposition $\mathbb{C}^n = V_1 \oplus V_2$. Such a decomposition is said to **reduce** $A$ if both subspaces $V_1$ and $V_2$ are **invariant** under $A$, that is, $AV_j \subseteq V_j$, $j = 1, 2$. Changing to a new basis $u_1, \dots, u_n$ such that $u_1, \dots, u_m$ is a basis for $V_1$ and $u_{m+1}, \dots, u_n$ is a basis for $V_2$, implies that $A$ is transformed to the block form

$$U^{-1} A U = \begin{pmatrix} A_1 & 0 \\ 0 & A_2 \end{pmatrix} \tag{3.5}$$

in these new coordinates. Moreover, we even have

$$U^{-1} \exp(A) U = \exp(U^{-1} A U) = \begin{pmatrix} \exp(A_1) & 0 \\ 0 & \exp(A_2) \end{pmatrix}. \tag{3.6}$$

Hence we need to find some invariant subspaces which reduce $A$. If we look at one-dimensional subspaces we must have

$$Ax = \alpha x, \qquad x \neq 0, \tag{3.7}$$

for some $\alpha \in \mathbb{C}$. If (3.7) holds, $\alpha$ is called an **eigenvalue** of $A$ and $x$ is called **eigenvector**. In particular, $\alpha$ is an eigenvalue if and only if $\operatorname{Ker}(A - \alpha) \neq \{0\}$ and hence $\operatorname{Ker}(A - \alpha)$ is called the **eigenspace** of $\alpha$ in this case. Since $\operatorname{Ker}(A - \alpha) \neq \{0\}$ implies that $A - \alpha$ is *not* invertible, the eigenvalues are the zeros of the **characteristic polynomial** of $A$,

$$\chi_A(z) = \prod_{j=1}^{m} (z - \alpha_j)^{a_j} = \det(z\mathbb{I} - A), \tag{3.8}$$

where $\alpha_i \neq \alpha_j$. The number $a_j$ is called **algebraic multiplicity** of $\alpha_j$ and $g_j = \dim \operatorname{Ker}(A - \alpha_j)$ is called **geometric multiplicity** of $\alpha_j$.

The set of all eigenvalues of $A$ is called the **spectrum** of $A$,

$$\sigma(A) = \{\alpha \in \mathbb{C} | \operatorname{Ker}(A - \alpha) \neq \{0\}\}. \tag{3.9}$$

If the algebraic and geometric multiplicities of all eigenvalues happen to be the same, we can find a basis consisting only of eigenvectors and $U^{-1}AU$ is a diagonal matrix with the eigenvalues as diagonal entries. Moreover, $U^{-1}\exp(A)U$ is again diagonal with the exponentials of the eigenvalues as diagonal entries.

However, life is not that simple and we only have $g_j \leq a_j$ in general. It turns out that the right objects to look at are the **generalized eigenspaces**

$$V_j = \operatorname{Ker}(A - \alpha_j)^{a_j}. \tag{3.10}$$

**Lemma 3.1.** *Let $A$ be an $n$ by $n$ matrix and let $V_j = \mathrm{Ker}(A - \alpha_j)^{a_j}$. Then the $V_j$'s are invariant subspaces and $\mathbb{C}^n$ can be written as a direct sum*

$$\mathbb{C}^n = V_1 \oplus \cdots \oplus V_m. \tag{3.11}$$

As a consequence we obtain

**Theorem 3.2** (Cayley–Hamilton)**.** *Every matrix satisfies its own characteristic equation*

$$\chi_A(A) = 0. \tag{3.12}$$

So, if we choose a basis $u_j$ of generalized eigenvectors, the matrix $U = (u_1, \dots, u_n)$ transforms $A$ to a block structure

$$U^{-1}AU = \begin{pmatrix} A_1 & & \\ & \ddots & \\ & & A_m \end{pmatrix}, \tag{3.13}$$

where each matrix $A_j$ has only the eigenvalue $\alpha_j$. Hence it suffices to restrict our attention to this case.

A vector $u \in \mathbb{C}^n$ is called a **cyclic vector** for $A$ if the vectors $A^j u$, $0 \le j \le n-1$ span $\mathbb{C}^n$, that is,

$$\mathbb{C}^n = \{\sum_{j=0}^{n-1} a_j A^j u | a_j \in \mathbb{C}\}. \tag{3.14}$$

The case where $A$ has only one eigenvalue and where there exists a cyclic vector $u$ is quite simple. Take

$$U = (u, (A-\alpha)u, \dots, (A-\alpha)^{n-1}u), \tag{3.15}$$

then $U$ transforms $A$ to

$$J = U^{-1}AU = \begin{pmatrix} \alpha & 1 & & & \\ & \alpha & 1 & & \\ & & \alpha & \ddots & \\ & & & \ddots & 1 \\ & & & & \alpha \end{pmatrix}, \tag{3.16}$$

since $\chi_A(A) = (A-\alpha)^n = 0$ by the Cayley–Hamilton theorem. The matrix (3.16) is called a **Jordan block**. It is of the form $\alpha\mathbb{I} + N$, where $N$ is **nilpotent**, that is, $N^n = 0$.

Hence, we need to find a decomposition of the spaces $V_j$ into a direct sum of spaces $V_{jk}$, each of which has a cyclic vector $u_{jk}$.

We again restrict our attention to the case where $A$ has only one eigenvalue $\alpha$ and set

$$K_j = \mathrm{Ker}(A-\alpha)^j. \tag{3.17}$$

In the cyclic case we have $K_j = \oplus_{k=1}^{j} \mathrm{span}\{(A-\alpha)^{n-k}\}$. In the general case, using $K_j \subseteq K_{j+1}$, we can find $L_k$ such that

$$K_j = \bigoplus_{k=1}^{j} L_k. \tag{3.18}$$

In the cyclic case $L_n = \mathrm{span}\{u\}$ and we would work our way down to $L_1$ by applying $A - \alpha$ recursively. Mimicking this, we set $M_n = L_n$ and since $(A-\alpha)L_{j+1} \subseteq L_j$ we have $L_{n-1} = (A-\alpha)L_n \oplus M_{n-1}$. Proceeding like this we can find $M_l$ such that

$$L_k = \bigoplus_{l=k}^{n} (A-\alpha)^{n-l} M_l. \tag{3.19}$$

Now choose a basis $u_j$ for $M_1 \oplus \cdots \oplus M_n$, where each $u_j$ lies in some $M_l$. Let $V_j$ be the subspace generated by $(A-\alpha)^l u_j$, then $V = V_1 \oplus \cdots \oplus V_m$ by construction of the sets $M_k$ and each $V_j$ has a cyclic vector $u_j$. In summary, we get

**Theorem 3.3 (Jordan canonical form).** *Let $A$ be an $n$ by $n$ matrix. Then there exists a basis for $\mathbb{C}^n$, such that $A$ is of block form with each block as in (3.16).*

It is often useful to split $\mathbb{C}^n$ according to the subspaces on which $A$ is contracting, expanding, respectively unitary. We set

$$\begin{aligned} E^{\pm}(A) &= \bigoplus_{|\alpha_j|^{\pm 1} > 1} \mathrm{Ker}(A-\alpha_j)^{a_j}, \\ E^{0}(A) &= \bigoplus_{|\alpha_j| = 1} \mathrm{Ker}(A-\alpha_j)^{a_j}. \end{aligned} \tag{3.20}$$

The subspaces $E^-(A)$, $E^+(A)$, $E^0(A)$ are called **contracting**, **expanding**, **unitary subspace** of $A$, respectively. The restriction of $A$ to these subspaces is denoted by $A_+$, $A_-$, $A_0$, respectively.

Now it remains to show how to compute the exponential of a Jordan block $J = \alpha\mathbb{I} + N$. Since $\alpha\mathbb{I}$ commutes with $N$ we infer that

$$\exp(J) = \exp(\alpha\mathbb{I})\exp(N) = \mathrm{e}^{\alpha} \sum_{j=0}^{n-1} \frac{1}{j!} N^j. \tag{3.21}$$

Next, it is not hard to see that $N^j$ is a matrix with ones in the $j$-th diagonal above the main diagonal and hence $\exp(J)$ explicitly reads

$$\exp(J) = \mathrm{e}^{\alpha} \begin{pmatrix} 1 & 1 & \frac{1}{2!} & \cdots & \frac{1}{(n-1)!} \\ & 1 & 1 & \ddots & \vdots \\ & & 1 & \ddots & \frac{1}{2!} \\ & & & \ddots & 1 \\ & & & & 1 \end{pmatrix}. \tag{3.22}$$

Note that if $A$ is in Jordan canonical form, then it is not hard to see that

$$\det(\exp(A)) = \exp(\mathrm{tr}(A)). \tag{3.23}$$

Since both the determinant and the trace are invariant under linear transformations, the formula also holds in the general case.

In addition, to the matrix exponential we will also need its inverse. That is, given a matrix $A$ we want to find a matrix $B$ such that

$$A = \exp(B). \tag{3.24}$$

Clearly, by (3.23) this can only work if $\det(A) \neq 0$. Hence suppose that $\det(A) \neq 0$. It is no restriction to assume that $A$ is in Jordan canonical form and to consider the case of only one Jordan block, $A = \alpha\mathbb{I} + N$.

Motivated by the power series for the logarithm,

$$\ln(1+x) = \sum_{j=1}^{\infty} \frac{(-1)^{j+1}}{j} x^j, \qquad |x| < 1, \tag{3.25}$$

we set

$$\begin{aligned} B &= \ln(\alpha)\mathbb{I} + \sum_{j=1}^{n-1} \frac{(-1)^{j+1}}{j\alpha^j} N^j \\ &= \begin{pmatrix} \ln(\alpha) & \frac{1}{\alpha} & \frac{-1}{2\alpha^2} & \cdots & \frac{(-1)^n}{(n-1)\alpha^{n-1}} \\ & \ln(\alpha) & \frac{1}{\alpha} & \ddots & \vdots \\ & & \ln(\alpha) & \ddots & \frac{-1}{2\alpha^2} \\ & & & \ddots & \frac{1}{\alpha} \\ & & & & \ln(\alpha) \end{pmatrix}. \end{aligned} \tag{3.26}$$

By construction we have $\exp(B) = A$.

Similarly, note that the resolvent $(A-z)^{-1}$ can also be easily computed in Jordan canonical form, in particular for a Jordan block we have

$$(J-z)^{-1} = \frac{1}{\alpha - z} \sum_{j=0}^{n-1} \frac{1}{(z-\alpha)^j} N^j. \tag{3.27}$$

In particular, note that the resolvent has a pole at each eigenvalue with the residue being the projector onto the corresponding generalized eigenspace.

Let me emphasize, that both the eigenvalues and generalized eigenvectors can be complex even if the matrix $A$ has only real entries. Since in many applications only real solutions are of interest, one likes to have a canonical form involving only real matrices. This form is called **real Jordan canonical form** and it can be obtained as follows.

Suppose the matrix $A$ has only real entries. Let $\alpha_i$ be its eigenvalues and let $u_j$ be a basis in which $A$ has Jordan canonical form. Look at the complex conjugate of the equation

$$Au_j = \alpha_i u_j, \tag{3.28}$$

it is not hard to conclude the following for a given Jordan block $J = \alpha\mathbb{I} + N$:

If $\alpha$ is real, the corresponding generalized eigenvectors can assumed to be real. Hence there is nothing to be done in this case.

If $\alpha$ is nonreal, there must be a corresponding block $\tilde{J} = \alpha^*\mathbb{I} + N$ and the corresponding generalized eigenvectors can be assumed to be the complex conjugates of our original ones. Therefore we can replace the pairs $u_j$, $u_j^*$ in our basis by $\mathrm{Re}(u_j)$ and $\mathrm{Im}(u_j)$. In this new basis the block $J \oplus \tilde{J}$ is replaced by

$$\begin{pmatrix} R & \mathbb{I} & & & \\ & R & \mathbb{I} & & \\ & & R & \ddots & \\ & & & \ddots & \mathbb{I} \\ & & & & R \end{pmatrix}, \tag{3.29}$$

where

$$R = \begin{pmatrix} \mathrm{Re}(\alpha) & \mathrm{Im}(\alpha) \\ -\mathrm{Im}(\alpha) & \mathrm{Re}(\alpha) \end{pmatrix} \quad \text{and} \quad \mathbb{I} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}. \tag{3.30}$$

Since the matrices

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \tag{3.31}$$

commute, the exponential is given by

$$\begin{pmatrix} \exp(R) & \exp(R) & \exp(R)\frac{1}{2!} & \dots & \exp(R)\frac{1}{(n-1)!} \\ & \exp(R) & \exp(R) & \ddots & \vdots \\ & & \exp(R) & \ddots & \exp(R)\frac{1}{2!} \\ & & & \ddots & \exp(R) \\ & & & & \exp(R) \end{pmatrix}, \tag{3.32}$$

where

$$\exp(R) = \mathrm{e}^{\mathrm{Re}(\alpha)} \begin{pmatrix} \cos(\mathrm{Im}(\alpha)) & -\sin(\mathrm{Im}(\alpha)) \\ \sin(\mathrm{Im}(\alpha)) & \cos(\mathrm{Im}(\alpha)) \end{pmatrix}. \tag{3.33}$$

Finally, let me remark that a matrix $A(t)$ is called differentiable with respect to $t$ if all coefficients are. In this case we will denote by $\frac{d}{dt}A(t) \equiv \dot{A}(t)$ the matrix, whose coefficients are the derivatives of the coefficients of $A(t)$. The usual rules of calculus hold in this case as long as one takes noncommutativity of matrices into account. For example we have the product rule

$$\frac{d}{dt}A(t)B(t) = \dot{A}(t)B(t) + A(t)\dot{B}(t) \tag{3.34}$$

(Problem 3.1).

**Problem 3.1** (Differential calculus for matrices.)**.** *Suppose $A(t)$ and $B(t)$ are differentiable. Prove (3.34) (note that the order is important!). Suppose* $\det(A(t)) \neq 0$*, show*

$$\frac{d}{dt}A(t)^{-1} = -A(t)^{-1}\dot{A}(t)A(t)^{-1}$$

*(Hint: $AA^{-1} = \mathbb{I}$.)*

**Problem 3.2.** *(i) Compute* $\exp(A)$ *for*

$$A = \begin{pmatrix} a+d & b \\ c & a-d \end{pmatrix}.$$

*(ii) Is there a real matrix $A$ such that*

$$\exp(A) = \begin{pmatrix} -\alpha & 0 \\ 0 & -\beta \end{pmatrix}, \quad \alpha, \beta > 0?$$

**Problem 3.3.** *Denote by* $r(A) = \max_j\{|\alpha_j|\}$ *the* **spectral radius** *of $A$. Show that for every $\varepsilon > 0$ there is a norm $\|.\|_\varepsilon$ such that*

$$\|A\|_\varepsilon = \sup_{x\colon \|x\|_\varepsilon = 1} \|Ax\|_\varepsilon \leq r(A) + \varepsilon.$$

*(Hint: It suffices to prove the claim for a Jordan block $J = \alpha\mathbb{I} + N$ (why?). Now choose a diagonal matrix $Q = \mathrm{diag}(1, \varepsilon, \dots, \varepsilon^n)$ and observe $Q^{-1}JQ = \alpha\mathbb{I} + \varepsilon N$.)*

**Problem 3.4.** *Suppose $A(\lambda)$ is $C^k$ and has no unitary subspace. Then the projectors $P^{\pm}(A(\lambda))$ onto the contracting, expanding subspace are $C^k$. (Hint: Use the formulas*

$$P^{+}(A(\lambda)) = \frac{1}{2\pi\mathrm{i}} \int_{|z|=1} \frac{dz}{z - A(\lambda)}, \quad P^{-}(A(\lambda)) = \mathbb{I} - P^{+}(A(\lambda)).)$$

## 3.2. Linear autonomous first order systems

We begin with the study of autonomous linear first order system

$$\dot{x}(t) = Ax(t), \qquad x(0) = x_0, \tag{3.35}$$

where $A$ is an $n$ by $n$ matrix. In this case the Picard iteration can be computed explicitly, producing

$$x_n(t) = \sum_{j=0}^{n} \frac{t^j}{j!} A^j x_0. \tag{3.36}$$

The limit as $n \to \infty$ is given by

$$x(t) = \lim_{n\to\infty} x_n(t) = \exp(tA)x_0. \tag{3.37}$$

Hence in order to understand the dynamics of the system (3.35), we need to understand the properties of the function $\exp(tA)$. This is best done by using a linear change of coordinates

$$y = Ux \tag{3.38}$$

which transforms $A$ into a simpler form $UAU^{-1}$. The form most suitable for computing the exponential is the Jordan canonical form, discussed in Section 3.1. In fact, if $A$ is in Jordan canonical form, it is not hard to compute $\exp(tA)$. It even suffices to consider the case of one Jordan block $J$ (see (3.16)), where it is not hard to see that

$$\exp(tJ) = \mathrm{e}^{\alpha t} \begin{pmatrix} 1 & t & \frac{t^2}{2!} & \dots & \frac{t^{n-1}}{(n-1)!} \\ & 1 & t & \ddots & \vdots \\ & & 1 & \ddots & \frac{t^2}{2!} \\ & & & \ddots & t \\ & & & & 1 \end{pmatrix}. \tag{3.39}$$

On the other hand, the procedure of finding the Jordan canonical form is quite cumbersome and hence we will use *Mathematica* to do the calculations for us. For example, let

$$\texttt{In[1]:= A} = \begin{pmatrix} -11 & -35 & -24 \\ -1 & -1 & -2 \\ 8 & 22 & 17 \end{pmatrix};$$

Then the command

```
In[2]:= {U, J} = JordanDecomposition[A];
```

gives us the transformation matrix $U$ plus the Jordan canonical form $J = U^{-1}AU$.

```
In[3]:= J // MatrixForm
```

Out[3]//MatrixForm=

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 1 \\ 0 & 0 & 2 \end{pmatrix}$$

If you don't trust me (or *Mathematica*), you can also check it:

```
In[4]:= A == U.J.Inverse[U]
Out[4]= True
```

Furthermore, *Mathematica* can even compute the exponential for us

```
In[5]:= MatrixExp[tJ] // MatrixForm
```

Out[5]//MatrixForm=

$$\begin{pmatrix} \mathrm{e}^t & 0 & 0 \\ 0 & \mathrm{e}^{2t} & t\mathrm{e}^{2t} \\ 0 & 0 & \mathrm{e}^{2t} \end{pmatrix}$$

Let me emphasize again, that both the eigenvalues and generalized eigenvectors can be complex even if the matrix $A$ has only real entries. Since in many applications only real solutions are of interest, one has to use the real Jordan canonical form instead.

To understand the qualitative behavior for large $t$, we need to understand the function $\exp(\alpha t)$. Using Euler's formula

$$\mathrm{e}^{\mathrm{i}\omega} = \cos(\omega) + \mathrm{i}\sin(\omega) \tag{3.40}$$

we see

$$\mathrm{e}^{\alpha t} = \mathrm{e}^{\alpha t}\left(\cos(\omega t) + \mathrm{i}\sin(\omega t)\right), \qquad \alpha = \lambda + \mathrm{i}\omega. \tag{3.41}$$

Thus $\exp(\alpha t)$ will converge to 0 as $t \to \infty$ if $\lambda = \mathrm{Re}(\alpha) < 0$. If $\lambda = 0$, it will remain at least bounded.

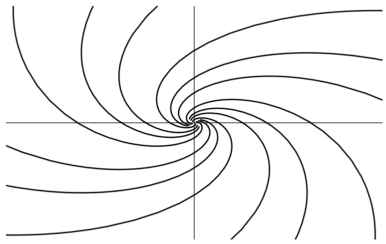
**Theorem 3.4.** *A solution of the linear system (3.35) converges to* 0 *as* $t \to \infty$ *if the initial condition* $x_0$ *lies in the subspace spanned by the generalized eigenspaces corresponding to eigenvalues with negative real part.*

*It will remain bounded if* $x_0$ *lies in the subspace spanned by the generalized eigenspaces corresponding to eigenvalues with negative real part plus the eigenspaces corresponding to eigenvalues with vanishing real part.*

Note that to get the behavior as $t \to -\infty$, you just need to replace *negative* by *posiitve*.

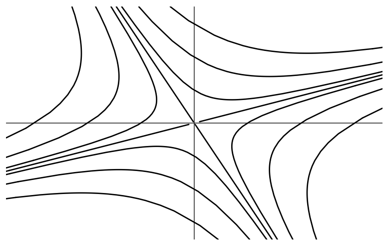
To get a better understanding let us look at the possible cases in $\mathbb{R}^2$:

Firstly, suppose both eigenvalues have positive real part. Then all solutions grow exponentially as $t \to \infty$ and decay exponentially as $t \to -\infty$. The origin is called a **source** in this case and the typical phase portrait is depicted below.



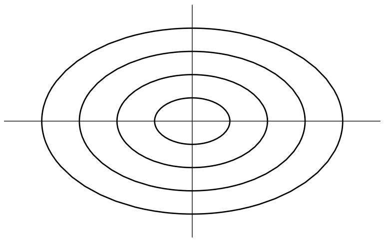
Similarly, if both eigenvalues have negative real part, the situation can be reduced to the previous one by replacing $t \to -t$. The phase portrait stays the same except that the orbits are traversed in the opposite direction. The origin is called a **sink** in this case.

If one eigenvalue has positive and one eigenvalue has negative real part, the phase portrait looks as follows



and the origin is called a **saddle**. The long-time behavior now depends on the initial condition $x_0$. If $x_0$ lies in the eigenspace corresponding to the eigenvalue having negative real part, the solution will decay exponentially as $t \to \infty$ and grow exponentially as $t \to -\infty$. If $x_0$ lies in the eigenspace corresponding to the other eigenvalue it is the other way round. If $x_0$ has components in both eigenspaces, it will grow exponentially as $t \to \pm\infty$.

Finally, if both eigenvalues are purely imaginary, the solutions will be periodic and the phase portrait looks as follows



The solution of the inhomogeneous equation

$$\dot{x}(t) = Ax(t) + g(t), \qquad x(0) = x_0 \tag{3.42}$$

is given by

$$x(t) = \exp(tA)\left(x_0 + \int_0^t \exp(-sA)g(s)ds\right), \tag{3.43}$$

which can be verified by a straightforward computation.

At the end of this section, let us look at the case of the $n$-th order equation

$$x^{(n)} + c_{n-1}x^{(n-1)} + \cdots + c_1\dot{x} + c_0 x = 0. \tag{3.44}$$

The corresponding system is given by

$$A = \begin{pmatrix} 0 & 1 & & & \\ & 0 & 1 & & \\ & & \ddots & \ddots & \\ & & & \ddots & 1 \\ -c_0 & -c_1 & \cdots & \cdots & -c_{n-1} \end{pmatrix} \tag{3.45}$$

and hence all our considerations apply: The characteristic polynomial is

$$\det(z - A) = z^n + c_{n-1}z^{n-1} + \cdots + c_1 z + c_0 \tag{3.46}$$

and since the geometric multiplicity is always one (Hint: Can you find a cyclic vector? Why does this help you?) we obtain:

**Theorem 3.5.** *Let $\alpha_j$, $1 \le j \le m$ be the eigenvalues of $A$ with corresponding algebraic multiplicites $a_j$, then the functions*

$$x_{j,k}(t) = t^k \exp(\alpha_j t), \qquad 0 \le k < a_j, \quad 1 \le j \le m \tag{3.47}$$

*are $n$ linearly independent solutions of (3.44).*

*In particular, any other solution can be written as a linear combination of these solutions.*

Note that if the coefficients $c_j$ are real, and if we are interested in real solutions, all we have to do is to take real and imaginary part. That is, for $\alpha_j = \lambda_j + \mathrm{i}\omega_j$ take

$$t^k \mathrm{e}^{\lambda_j t} \cos(\omega_j t), \qquad t^k \mathrm{e}^{\lambda_j t} \sin(\omega_j t). \tag{3.48}$$

By (3.43) the solution of the inhomogeneous equation

$$x^{(n)} + c_{n-1} x^{(n-1)} + \cdots + c_1 \dot{x} + c_0 x = g(t) \tag{3.49}$$

is given by

$$x(t) = \int_0^t u(t-s) g(s) ds, \tag{3.50}$$

where $u(t)$ is the solution of the homogeneous equation corresponding to the initial condition $u(0) = \dot{u}(0) = \cdots = u^{(n-1)}(0) = 0$ and $u^{(n)}(0) = 1$ (Problem 3.13).

**Problem 3.5.** *Solve the following equations.*

(i) $\dot{x} = 3x$.

(ii) $\dot{x} = \frac{\gamma}{t} x$, $\gamma \in \mathbb{R}$.

(iii) $\dot{x} = x + \sin(t)$.

**Problem 3.6.** *Solve the systems corresponding to the following matrices:*

$$(i).\ A = \begin{pmatrix} 2 & 1 \\ 0 & 2 \end{pmatrix}, \qquad (ii).\ A = \begin{pmatrix} -1 & 1 \\ 0 & 1 \end{pmatrix}.$$

**Problem 3.7.** *Solve*

$$\dot{x} = -y - t, \quad \dot{y} = x + t, \quad x(0) = 1, y(0) = 0.$$

**Problem 3.8.** *Find a two by two matrix such that $x(t) = (\sinh(t), \mathrm{e}^t)$ is a solution.*

**Problem 3.9.** *Which of the following functions*

(i) $x(t) = 3\mathrm{e}^t + \mathrm{e}^{-t}$, $y(t) = \mathrm{e}^{2t}$.

(ii) $x(t) = 3\mathrm{e}^t + \mathrm{e}^{-t}$, $y(t) = \mathrm{e}^t$.

(iii) $x(t) = 3\mathrm{e}^t + \mathrm{e}^{-t}$, $y(t) = t\mathrm{e}^t$.

(iv) $x(t) = 3\mathrm{e}^t$, $y(t) = t^2\mathrm{e}^t$.

*can be solutions of a first order autonomous homogeneous system?*

**Problem 3.10.** *Look at the second order equation*

$$\ddot{x} + c_1 \dot{x} + c_0 x = 0.$$

*Transform it into a system and discuss the possible solutions in terms of $c_1$, $c_0$. Find a formula for the Wronskian $W(x, y) = x\dot{y} - \dot{x}y$ of two solutions.*

*Suppose $c_0, c_1 \in \mathbb{R}$, show that real and imaginary part of a solution is again a solution. Discuss the real form of the solution in this case.*

**Problem 3.11** (Euler equation)**.** *Show that the equation*

$$\ddot{x} + \frac{c_1}{t}\dot{x} + \frac{c_0}{t^2}x = 0, \qquad t > 0,$$

*can be solved by introducing the new dependent variable $\tau = \ln(t)$. Discuss the possible solutions for $c_0, c_1 \in \mathbb{R}$.*

**Problem 3.12.** *Prove Theorem 3.5 directly (Hint:$\left(\frac{d}{dt} - \alpha\right) t^k \exp(\alpha t) =?$).*

**Problem 3.13.** *Prove (3.50).*

**Problem 3.14** (Laplace transform)**.** *Consider the* **Laplace transform**

$$\mathcal{L}(x)(s) = \int_0^\infty \mathrm{e}^{-st} x(t) dt.$$

*Show that the initial value problem*

$$\dot{x} = Ax + f(t), \qquad x(0) = x_0$$

*is transformed into a linear system of equations by the Laplace transform.*

## 3.3. General linear first order systems

We now turn to the general linear first order system

$$\dot{x}(t) = A(t)x(t), \tag{3.51}$$

where $A \in C(I, \mathbb{R}^{n^2})$. Clearly, our basic existence and uniqueness result (Theorem 2.3) applies to this system. Moreover, if $I = \mathbb{R}$, solutions exist for all $t \in \mathbb{R}$ by Theorem 2.12.

Now observe that linear combinations of solutions are again solutions. Hence the set of all solutions forms a vector space. This is often referred to as **superposition principle**. In particular, the solution corresponding to the initial condition $x(t_0) = x_0$ can be written as

$$\phi(t, t_0, x_0) = \sum_{j=1}^{n} \phi(t, t_0, \delta_j) x_{0,j}, \tag{3.52}$$

where $\delta_j$ are the canonical basis vectors, (i.e, $\delta_{j,k} = 1$ if $j = k$ and $\delta_{j,k} = 0$ if $j \neq k$) and $x_{0,j}$ are the components of $x_0$ (i.e., $x_0 = \sum_{j=1}^n \delta_j x_{0,j}$). Using the solutions $\phi(t, t_0, \delta_j)$ as columns of a matrix

$$\Pi(t, t_0) = (\phi(t, t_0, \delta_1), \dots, \phi(t, t_0, \delta_n)). \tag{3.53}$$

we see that there is a linear mapping $x_0 \mapsto \phi(t, t_0, x_0)$ given by

$$\phi(t, t_0, x_0) = \Pi(t, t_0) x_0. \tag{3.54}$$

The matrix $\Pi(t, t_0)$ is called **principal matrix solution** and it solves the matrix valued initial value problem

$$\dot{\Pi}(t, t_0) = A(t)\Pi(t, t_0), \qquad \Pi(t_0, t_0) = \mathbb{I}. \tag{3.55}$$

Furthermore, it satisfies

$$\Pi(t, t_1)\Pi(t_1, t_0) = \Pi(t, t_0) \tag{3.56}$$

since both sides solve $\dot{\Pi} = A(t)\Pi$ and coincide for $t = t_1$. In particular, $\Pi(t, t_0)$ is an isomorphism with inverse $\Pi(t, t_0)^{-1} = \Pi(t_0, t)$.

Let us summarize the most important findings in the following theorem.

**Theorem 3.6.** *The solutions of the system (3.51) form an $n$ dimensional vector space. Moreover, there exists a matrix-valued solution $\Pi(t, t_0)$ such that the solution of the IVP $x(t_0) = x_0$ is given by $\Pi(t, t_0)x_0$.*

More generally, taking $n$ solutions $\phi_1, \dots, \phi_n$ we obtain a matrix solution $U(t) = (\phi_1(t), \dots, \phi_n(t))$. The determinant of $U(t)$ is called **Wronski determinant**

$$W(t) = \det(\phi_1(t), \dots, \phi_n(t)). \tag{3.57}$$

If $\det U(t) \neq 0$, the matrix solution $U(t)$ is called a **fundamental matrix solution**. Moreover, if $U(t)$ is a matrix solution, so is $U(t)C$, where $C$ is a constant matrix. Hence, given two fundamental matrix solutions $U(t)$ and $V(t)$ we always have $V(t) = U(t)U(t_0)^{-1}V(t_0)$ since a matrix solution is uniquely determined by an initial condition. In particular, the principal matrix solution can be obtained from any fundamental matrix solution via $\Pi(t, t_0) = U(t)U(t_0)^{-1}$.

The following lemma shows that it suffices to check $\det U(t) \neq 0$ for one $t \in \mathbb{R}$.

**Lemma 3.7** (Liouville)**.** *The Wronski determinant of $n$ solutions satisfies*

$$W(t) = W(t_0) \exp\left(\int_{t_0}^{t} \operatorname{tr}(A(s))\, ds\right). \tag{3.58}$$

*This is known as* **Liouville's formula**.

**Proof.** Using $U(t + \varepsilon) = \Pi(t + \varepsilon, t)U(t)$ and

$$\Pi(t + \varepsilon, t) = \mathbb{I} + A(t)\varepsilon + o(\varepsilon) \tag{3.59}$$

we obtain

$$W(t + \varepsilon) = \det(\mathbb{I} + A(t)\varepsilon + o(\varepsilon))W(t) = (1 + \operatorname{tr}(A(t))\varepsilon + o(\varepsilon))W(t) \tag{3.60}$$

(this is easily seen by induction on $n$) implying

$$\frac{d}{dt}W(t) = \operatorname{tr}(A(t))\, W(t). \tag{3.61}$$

This equation is separable and the solution is given by (3.58). □

Now let us turn to the inhomogeneous system

$$\dot{x} = A(t)x + g(t), \qquad x(t_0) = x_0, \tag{3.62}$$

where $A \in C(I, \mathbb{R}^n \times \mathbb{R}^n)$ and $g \in C(I, \mathbb{R}^n)$. Since the difference of two solutions of the inhomogeneous system (3.62) satisfies the corresponding homogeneous system (3.51), it suffices to find one particular solution. This can be done using the following ansatz

$$x(t) = \Pi(t, t_0)c(t), \qquad c(t_0) = x_0, \tag{3.63}$$

which is known as **variation of constants**. Differentiating this ansatz we see

$$\dot{x}(t) = A(t)x(t) + \Pi(t, t_0)\dot{c}(t) \tag{3.64}$$

and comparison with (3.62) yields

$$\dot{c}(t) = \Pi(t_0, t)g(t), \tag{3.65}$$

Integrating this equation shows

$$c(t) = x_0 + \int_{t_0}^{t} \Pi(t_0, s)g(s)ds \tag{3.66}$$

and we obtain (using (3.56))

**Theorem 3.8.** *The solution of the inhomogeneous system corresponding to the initial condition $x(t_0) = x_0$ is given by*

$$x(t) = \Pi(t, t_0)x_0 + \int_{t_0}^{t} \Pi(t, s)g(s)ds, \tag{3.67}$$

*where $\Pi(t, t_0)$ is the principal matrix solution of the corresponding homogeneous system.*

**Problem 3.15.** *Compute $\Pi(t, t_0)$ for the system*

$$A(t) = \begin{pmatrix} t & 0 \\ 1 & t \end{pmatrix}.$$

**Problem 3.16.** *Consider the equation $\ddot{x} = q(t)x + g(t)$.*

*(i) Show that the Wronski determinant*

$$W(u, v) = u(t)v'(t) - u'(t)v(t)$$

*of two solutions $u$, $v$ of the homogeneous equation is independent of $t$.*

*(ii) Show that the fundamental matrix of the associated system is given by*

$$\Pi(t, s) = \frac{1}{W(u, v)} \begin{pmatrix} u(t)v'(s) - v(t)u'(s) & v(t)u(s) - v(s)u(t) \\ v'(s)u'(t) - v'(t)u'(s) & u(s)v'(t) - v(s)u'(t) \end{pmatrix}$$

*and use the variation of constants formula to show that*

$$x(t) = \frac{u(t)}{W(u,v)} \int^t v(s)g(s)ds - \frac{v(t)}{W(u,v)} \int^t u(s)g(s)ds$$

*is a solutions of the inhomogeneous equation.*

*(iii) Given one solution $u(t)$ of the homogeneous equation, make a variation of constants ansatz $v(t) = c(t)u(t)$ and show that a second solution is given by*

$$v(t) = u(t) \int^t \frac{1}{u(s)^2} ds.$$

*(iv) Show that if $u$ is a solution of the homogeneous equation, then $\phi = u'/u$ satisfies the Riccati equation*

$$\phi' + \phi^2 = q(t).$$

**Problem 3.17** (Reduction of order (d'Alembert)). *Look at the $n$-th order equation*

$$x^{(n)} + q_{n-1}(t)x^{(n-1)} + \cdots + q_1(t)\dot{x} + q_0(t)x = 0.$$

*Show that if one solutions $x_1(t)$ is known, the variation of constants ansatz $x(t) = c(t)x_1(t)$ gives a $(n-1)$-th order equation for $\dot{c}$. Hence the order can be reduced by one.*

**Problem 3.18** (Quantum Mechanics). *A quantum mechanical system which can only attain finitely many states is described by a complex valued vector $\psi(t) \in \mathbb{C}^n$. The square of the absolute values of the components $|\psi_j|^2$ are interpreted as the probability of finding the system in the $j$-th state at time $t$. Since there are only $n$ possible states, these probabilities must add up to one, that is, $\psi(t)$ must be normalized, $|\psi| = 1$. The time evolution of the system is governed by the* **Schrödinger equation**

$$\mathrm{i}\dot{\psi}(t) = H(t)\psi(t), \quad \psi(t_0) = \psi_0,$$

*where $H(t)$, is a self-adjoint matrix, that is, $H(t)^* = H(t)$. Here $H(t)$ is called the Hamiltonian and describes the interaction. Show that the solution is given by*

$$\psi(t) = U(t, t_0)\psi_0,$$

*where $U(t, t_0)$ is unitary, that is, $U(t, t_0)^{-1} = U(t, t_0)^*$ (Hint: Problem 3.1). Conclude that $\psi(t)$ remains normalized for all $t$ if $\psi_0$ is.*

*Each observable (quantity you can measure) corresponds to a self-adjoint matrix, say $L_0$. The expectation value for a measurement of $L_0$ if the system is in the state $\psi(t)$ is given by*

$$\langle \psi(t), L_0\psi(t) \rangle,$$

*where $\langle \varphi, \psi \rangle = \varphi^* \psi$ is the scalar product in $\mathbb{C}^n$. Show that*

$$\frac{d}{dt}\langle \psi(t), L_0 \psi(t) \rangle = \mathrm{i}\langle \psi(t), [H(t), L_0]\psi(t) \rangle$$

*and conclude that the solution of the* **Heisenberg equation**

$$\dot{L}(t) = \mathrm{i}[H(t), L(t)] + \dot{H}(t), \quad L(t_0) = L_0,$$

*where $[H, L] = HL - LH$ is the commutator, is given by*

$$L(t) = U(t_0, t) L_0 U(t, t_0).$$

## 3.4. Periodic linear systems

In this section we want to consider (3.51) in the special case where $A(t)$ is periodic,

$$A(t + T) = A(t), \qquad T > 0. \tag{3.68}$$

This periodicity condition implies that $x(t+T)$ is again a solution if $x(t)$ is. Hence it suggests itself to investigate what happens if we move on by one period, that is, to look at the **monodromy matrix**

$$M(t_0) = \Pi(t_0 + T, t_0). \tag{3.69}$$

A first naive guess would be that all initial conditions return to their starting values after one period (i.e., $M(t_0) = \mathbb{I}$) and hence all solutions are periodic. However, this is too much to hope for since it already fails in one dimension with $A(t)$ a constant.

On the other hand, since it does not matter whether we start our period at $t_0$, at $t_0 + T$, or even $t_0 + \ell T$, $\ell \in \mathbb{Z}$, we infer that $M(t_0)$ is periodic, that is, $M(t_0 + T) = M(t_0)$. Moreover, we even have

$$\begin{aligned} \Pi(t_0 + \ell T, t_0) &= \Pi(t_0 + \ell T, t_0 + (\ell - 1)T)\Pi(t_0 + (\ell - 1)T, t_0) \\ &= M(t_0 + (\ell - 1)T)\Pi(t_0 + (\ell - 1)T, t_0) \\ &= M(t_0)\Pi(t_0 + (\ell - 1)T, t_0) \\ &= M(t_0)^\ell \Pi(t_0, t_0) = M(t_0)^\ell. \end{aligned} \tag{3.70}$$

Thus $\Pi(t, t_0)$ exhibits an exponential behavior if we move on by one period in each step. If we factor out this exponential term, the remainder should be periodic.

For this purpose we rewrite $M(t_0)$ a little bit. By Liouville's formula (3.58) the determinant of the monodromy matrix

$$\det(M(t_0)) = \exp\left( \int_{t_0}^{t_0 + T} \mathrm{tr}(A(s)) ds \right) = \exp\left( \int_0^T \mathrm{tr}(A(s)) ds \right) \tag{3.71}$$

is independent of $t_0$ and positive. Hence there is a matrix $Q(t_0)$ (which is not unique) such that

$$M(t_0) = \exp(TQ(t_0)), \qquad Q(t_0 + T) = Q(t_0). \tag{3.72}$$

Writing

$$\Pi(t, t_0) = P(t, t_0) \exp((t - t_0)Q(t_0)) \tag{3.73}$$

a straightforward computation shows that

$$\begin{aligned} P(t + T, t_0) &= \Pi(t + T, t_0) M(t_0)^{-1} \mathrm{e}^{-(t-t_0)Q(t_0)} \\ &= \Pi(t + T, t_0 + T) \mathrm{e}^{-(t-t_0)Q(t_0)} \\ &= \Pi(t, t_0) \mathrm{e}^{-(t-t_0)Q(t_0)} = P(t, t_0) \end{aligned} \tag{3.74}$$

as anticipated. In summary we have proven **Floquet's theorem**.

**Theorem 3.9** (Floquet)**.** *Suppose $A(t)$ is periodic, then the principal matrix solution of the corresponding linear system has the form*

$$\Pi(t, t_0) = P(t, t_0) \exp((t - t_0)Q(t_0)), \tag{3.75}$$

*where $P(., t_0)$ has the same period as $A(.)$ and $P(t_0, t_0) = \mathbb{I}$.*

Note that any fundamental matrix solution can be written in this form (Problem 3.19).

Hence to understand the behavior of solutions one needs to understand the Jordan canonical form of the monodromy matrix. Moreover, we can choose any $t_0$ since $M(t_1)$ and $M(t_0)$ are similar matrices by virtue of

$$M(t_1) = \Pi(t_1, t_0) M(t_0) \Pi(t_1, t_0)^{-1}. \tag{3.76}$$

Thus the eigenvalues and the Jordan structure is independent of $t_0$ (hence the same also follows for $Q(t_0)$).

Before I show how this result is used in a concrete example, let me note another consequence of Theorem 3.9. The proof is left as an exercise (Problem 3.20).

**Corollary 3.10.** *The transformation $y(t) = P(t, t_0)^{-1} x(t)$ renders the system into one with constant coefficients,*

$$\dot{y}(t) = Q(t_0) y(t). \tag{3.77}$$

Note also that we have $P(t, t_0)^{-1} = \exp((t - t_0)Q(t_0)) P(t_0, t) \exp(-(t - t_0)Q(t_0))$ by virtue of $\Pi(t, t_0)^{-1} = \Pi(t_0, t)$.

One of the most prominent examples is **Hill's equation**

$$\ddot{x} + q(t)x = 0, \qquad q(t + T) = q(t). \tag{3.78}$$

In this case

$$\Pi(t, t_0) = \begin{pmatrix} c(t, t_0) & s(t, t_0) \\ \dot{c}(t, t_0) & \dot{s}(t, t_0) \end{pmatrix}, \tag{3.79}$$

where $c(t, t_0)$ is the solution corresponding to the initial condition $c(t_0, t_0) = 1$, $\dot{c}(t_0, t_0) = 0$ and similarly for $s(t, t_0)$. Liouville's formula (3.58) shows

$$\det \Pi(t, t_0) = 1 \tag{3.80}$$

and hence the characteristic equation for $M(t)$ is given by

$$\mu^2 - 2\Delta\mu + 1 = 0, \tag{3.81}$$

where

$$\Delta = \frac{\operatorname{tr}(M(t))}{2} = \frac{c(t+T, t) + \dot{s}(t+T, t)}{2}. \tag{3.82}$$

If $\Delta^2 > 1$ we have two different real eigenvalues

$$\mu_\pm = \Delta \pm \sqrt{\Delta^2 - 1} = \sigma\, \mathrm{e}^{\pm T\gamma}, \tag{3.83}$$

with corresponding eigenvectors

$$u_\pm(t_0) = \begin{pmatrix} 1 \\ \frac{\mu_\pm - c(t_0+T, t_0)}{s(t_0+T, t_0)} \end{pmatrix} = \begin{pmatrix} 1 \\ \frac{\dot{s}(t_0+T, t_0)}{\mu_\pm - \dot{c}(t_0+T, t_0)} \end{pmatrix}. \tag{3.84}$$

Note that $u_\pm(t_0)$ are also eigenvectors of $Q(t_0)$ corresponding to the eigenvalues $\gamma_\pm = \frac{1}{T}\ln(\mu_\pm)$ (compare (3.26)). From $\mu_+\mu_- = 1$ we obtain $\gamma_+ + \gamma_- = 0$ and it is no restriction to assume $|\mu_+| > 1$ respectively $\mathrm{Re}(\gamma_+) > 0$.

Considering

$$\begin{aligned} \Pi(t, t_0) u_\pm(t_0) &= P(t, t_0) \exp((t - t_0) Q(t_0)) u_\pm(t_0) \\ &= \mathrm{e}^{\gamma_\pm (t - t_0)} P(t, t_0) u_\pm(t_0), \end{aligned} \tag{3.85}$$

we see that there are two solutions of the form

$$\mathrm{e}^{\pm\gamma t} p_\pm(t), \qquad p_\pm(t+T) = \sigma\, p_\pm(t),\ \sigma^2 = 1, \gamma > 0, \tag{3.86}$$

where $\sigma = \mathrm{sgn}(\Delta)$ and $\gamma = \mathrm{Re}(\gamma_+)$. Similarly, if $\Delta^2 < 1$ we have two different purely complex eigenvalues and hence two solutions

$$\mathrm{e}^{\pm\mathrm{i}\gamma t} p_\pm(t), \qquad p_\pm(t+T) = p_\pm(t)\ \gamma > 0, \tag{3.87}$$

where $\gamma = \mathrm{Im}(\gamma_+)$. If $\Delta^2 = 1$ we have $\mu_\pm = \Delta$ and either two solutions

$$p_\pm(t), \qquad p_\pm(t+T) = \sigma\, p_\pm(t), \tag{3.88}$$

or two solutions

$$p_+(t), \quad p_-(t) + t p_+(t), \qquad p_\pm(t+T) = \sigma\, p_\pm(t), \tag{3.89}$$

where $\sigma = \mathrm{sgn}(\Delta)$.

A periodic equation is called **stable** if all solutions are bounded. Thus we have shown

**Theorem 3.11.** *Hills equation is stable if $|\Delta| < 1$ and unstable if $|\Delta| > 1$.*

This result is of high practical importance in applications. For example, the potential of a charged particle moving in the electric field of a quadrupole is given by

$$U(x) = e\frac{V}{a^2}(x^2 - y^2).$$

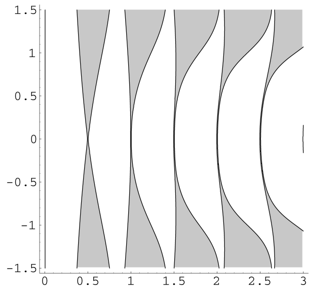
If we set for the voltage $V = V_0 + V_1\cos(t)$, one gets the following equations of motion (neglecting the induced magnetic filed)

$$\begin{aligned}\ddot{x} &= -\frac{2e}{ma^2}(V_0 + V_1\cos(t))x,\\ \ddot{y} &= +\frac{2e}{ma^2}(V_0 + V_1\cos(t))y,\\ \ddot{z} &= 0.\end{aligned} \tag{3.90}$$

The equation for the $x$ and $y$ coordinates is the **Mathieu equation**

$$\ddot{x} = \omega^2(1 + \varepsilon\cos(t))x. \tag{3.91}$$

A numerically computed stability diagram for $0 \le \omega \le 3$ and $-1.5 \le \varepsilon \le 1.5$ is depicted below.




The shaded regions are the ones where $\Delta(\omega,\varepsilon)^2 > 1$, that is, where the equation is unstable. Observe that these unstable regions emerge from the points $2\omega \in \mathbb{N}_0$ where $\Delta(\omega, 0) = \cos(2\pi\omega) = \pm 1$.

Varying the voltages $V_0$ and $V_1$ one can achieve that the equation is only stable (in the $x$ or $y$ direction) if the mass of the particle lies in a certain region. This can be used to filter charged particles according to their mass.

**Problem 3.19.** *Show that any fundamental matrix solution $U(t)$ of a periodic linear system can be written as $U(t) = V(t)\exp(tR)$, where $V(t)$ is periodic and $R$ is similar to $Q(t_0)$.*

**Problem 3.20.** *Prove Corollary 3.10.*

**Problem 3.21.** *Consider the inhomogeneous equation*

$$\dot{x}(t) = A(t)x(t) + g(t),$$

*where both $A(t)$ and $g(t)$ are periodic of period $T$. Show that this equation has a periodic solution of period $T$ if 1 is not an eigenvalue of the monodromy* *matrix* $M(t_0)$. *(Hint: Note that* $x(t)$ *is periodic if and only if* $x(T) = x(0)$ *and use the variation of constants formula (3.67).)*

**Problem 3.22** (Reflection symmetry). *Suppose* $q$ *is periodic* $q(t+T) = q(t)$ *and symmetric* $q(-t) = q(t)$. *Prove*

(i) $c(-t) = c(t)$ *and* $s(-t) = -s(t)$,

(ii) $c(t \pm T) = c(T)c(t) \pm \dot{c}(T)s(t)$ *and* $s(t \pm T) = \pm s(T)c(t) + \dot{s}(T)s(t)$,

(iii) $c(T) = \dot{s}(T)$,

*where* $c(t,0) = c(t)$, $s(t,0) = s(t)$.

**Problem 3.23** (Resonance). *Solve the equation*

$$\ddot{x} + \omega^2 x = \cos(\alpha t), \quad \omega, \alpha > 0.$$

*Discuss the behavior of solutions as* $t \to \infty$.

**Problem 3.24.** *A simple quantum mechanical model for an electron in a crystal leads to the investigation of*

$$-y'' + q(x)y = \lambda y, \quad \text{where} \quad q(x+1) = q(x).$$

*The parameter* $\lambda \in \mathbb{R}$ *corresponds to the energy of the electron. Only energies for which the equation is stable are allowed and hence the set* $\sigma = \{\lambda \in \mathbb{R} | |\Delta(\lambda)| \leq 1\}$ *is called the spectrum of the crystal. Since* $\Delta(\lambda)$ *is continuous with respect to* $\lambda$, *the spectrum consists of bands with gaps in between.*

*Consider the explicit case*

$$q(x) = q_0, \quad 0 \leq x < \frac{1}{2}, \qquad q(x) = 0, \quad \frac{1}{2} \leq x < 1.$$

*Show that there are no spectral bands below a certain value of* $\lambda$. *Show that there is an infinite number of gaps if* $q_0 \neq 0$. *How many gaps are there for* $q_0 = 0$? *(Hint: Set* $\lambda - q_0 \to (a-\varepsilon)^2$ *and* $\lambda \to (a+\varepsilon)^2$ *in the expression for* $\Delta(\lambda)$. *If* $q_0 \to 0$, *where would you expect gaps to be? Choose these values for* $a$ *and look at the case* $a \to \infty$.*)*

Chapter 4

# Differential equations in the complex domain

## 4.1. The basic existence and uniqueness result

Until now we have only imposed rather weak requirements on the smoothness of our differential equations. However, on the other hand, most examples encountered were in fact (real) analytic. Up to this point we did not use this additional information, but in the present chapter I want to show how to gain a better understanding for these problems by taking the detour over the complex plane.

In this chapter we want to look at differential equations in a complex domain $\Omega \subseteq \mathbb{C}^{n+1}$. We suppose that

$$f : \Omega \to \mathbb{C}^n, \qquad (z, w) \mapsto f(z, w), \tag{4.1}$$

is analytic in $\Omega$ and consider the equation

$$w' = f(z, w), \qquad w(z_0) = w_0. \tag{4.2}$$

Here the prime denotes complex differentiation and hence the equation only makes sense if $w$ is analytic as well. Clearly, the first question to ask is whether solutions exist at all. Fortunately, this can be answered using the same tools as in the real case. It suffices to only point out the differences.

The first step is to rewrite (4.2) as

$$w(z) = w_0 + \int_{z_0}^{z} f(\zeta, w) d\zeta. \tag{4.3}$$

But note that we now have to be more careful since the integral is along a path in the complex plane and independence of the path is not clear. On the other hand, we will only consider values of $z$ in a small disc around $z_0$. Since a disc is simply connected, path independence follows from the Cauchy integral theorem. Next, we need a suitable Banach space. As in the real case we can use the sup norm

$$\sup_{|z-z_0|<\varepsilon} |w(z)| \tag{4.4}$$

since the uniform limit of a sequence of analytic functions is again analytic. Now we can proceed as in the real case to obtain

**Theorem 4.1.** *Suppose $f : \Omega \to \mathbb{C}$ is analytic. Then the initial value problem (4.2) has a unique solution defined in a sufficiently small disc around $z_0$.*

Next, let us look at maximally defined solutions. Unfortunately, this topic is more tricky than in the real case. In fact, let $w_1(z)$ and $w_2(z)$ be two solutions defined on the domains $U_1$ and $U_2$ respectively. Then they coincide in a neighborhood of $z_0$ by our local uniqueness result. Hence they also coincide on the connected component of $U_1 \cap U_2$ containing $z_0$. But this is all we can say in general as the example

$$w' = \frac{1}{z}, \qquad w(1) = 0, \qquad z \in \mathbb{C}\backslash\{0\}, \tag{4.5}$$

shows. Indeed, the solution is given by $w(z) = \ln(z)$ and different choices of the branch cut will give different solutions.

These problems do not arise if $\Omega$ is simply connected.

**Theorem 4.2.** *Suppose $\Omega \subseteq \mathbb{C}$ is simply connected and $z_0 \in \Omega$. Then the initial value problem (4.2) has a unique solution defined on all of $\Omega$.*

**Proof.** Pick $z \in \Omega$ and let $\gamma : [0, 1] \to \Omega$ be a path from $z_0$ to $z$. Around each point $\gamma(t_0)$ we have a solution of the differential equation $w' = f(z, w)$ and by local uniqueness we can choose the solutions in such a way that they coincide for $t$ close to $t_0$. So we can define the value of $w(z)$ by analytic continuation along the path $\gamma$. Since $\Omega$ is simply connected, this value is uniquely defined by the monodromy theorem. □

Finally, let us show how analyticity can be used in the investigation of a simple differential equation,

$$w' + w^2 = z, \qquad w(0) = w_0. \tag{4.6}$$

This is a Riccati equation and we already know that it cannot be solved unless we find a particular solution. However, after you have tried for some time you will agree that it seems not possible to find one and hence we need to try something different. Since we know that the solution is analytic near 0 we can at least write

$$w(z) = \sum_{j=0}^{\infty} w_j z^j \tag{4.7}$$

and plugging this into our equation yields

$$\sum_{j=0}^{\infty} j w_j z^{j-1} + \left( \sum_{j=0}^{\infty} w_j z^j \right)^2 = z. \tag{4.8}$$

Expanding the product and aligning powers of $z$ gives

$$\sum_{j=0}^{\infty} \left( (j+1) w_{j+1} + \sum_{k=0}^{j} w_j w_{k-j} \right) z^j = z. \tag{4.9}$$

Comparing powers of $z$ we obtain

$$w_1 = -w_0^2,\ w_2 = w_0^3 + \frac{1}{2}, \quad w_{j+1} = \frac{-1}{j+1} \sum_{k=0}^{j} w_j w_{k-j}. \tag{4.10}$$

Thus we have at least found a recursive formula for computing the coefficients of the power series of the solution. However, I should point out that this will no longer work if the function $f$ involves $w$ in a too complicated way. Hence we will only investigate the case of linear equations further. In fact, this will eventually allow us to *solve* the above equation using special functions (Problem 4.7). However, we will on the other hand allow for poles in the coefficients, which is often needed in applications.

The following two sections are quite technical and can be skipped if you are not interested in the details of the proof of the generalized power series method alluded to above.

**Problem 4.1.** *Try to find a solution of the initial value problem*

$$w'' = (z^2 - 1) w, \qquad w(0) = 0,$$

*by using the power series method from above.*

## 4.2. Linear equations

For the rest of this chapter we will restrict our attention to linear equations which are the most important ones in applications. That is, we will look at the equation

$$w' = A(z) w, \qquad w(z_0) = w_0, \qquad z, z_0 \in \Omega \subseteq \mathbb{C}, \tag{4.11}$$

where $A(z)$ is a matrix whose coefficients are analytic in $\Omega$. Note that, as in the real case, the superposition principle holds. Hence, we can find a principal matrix solution $\Pi(z, z_0)$ such that the solution of (4.11) is given by

$$w(z) = \Pi(z, z_0) w_0 \tag{4.12}$$

at least for $z$ in a neighborhood of $z_0$. It is also not hard to see that Liouville's formula (3.58) extends to the complex case. Moreover, if $\Omega$ is simply connected, we can extend solutions to the entire domain $\Omega$.

In summary, we now know that the solution is nice whenever the matrix $A(z)$ is analytic. However, in most applications the coefficients will have singularities and one of the main questions is the behavior of the solutions near such a singularity. This will be our next topic. But first let us look at a prototypical example.

The system

$$w' = \frac{1}{z} A w, \qquad z \in \mathbb{C}\backslash\{0\}, \tag{4.13}$$

is called **Euler system**. Obviously it has a first order pole at $z = 0$ and since $\mathbb{C}\backslash\{0\}$ is not simply connected, solutions might not be defined for all $z \in \mathbb{C}\backslash\{0\}$. Hence we introduce a branch cut along the negative real axis and consider the simply connected domain $\Omega = \mathbb{C}\backslash(-\infty, 0]$. To solve (4.13) we will use the transformation

$$\zeta = \ln(z) = \ln|z| + \mathrm{i}\arg(z), \qquad -\pi < \arg(z) < \pi, \tag{4.14}$$

which maps $\Omega$ to the strip $\tilde{\Omega} = \{z \in \mathbb{C} | -\pi < \mathrm{Im}(z) < \pi\}$. The equation in the new coordinates reads

$$\omega' = A\omega, \qquad \omega(\zeta) = w(\mathrm{e}^{\zeta}). \tag{4.15}$$

Hence a fundamental system is given by

$$W(z) = z^A = \exp(\ln(z) A), \tag{4.16}$$

where the last expression is to be understood as the definition of $z^A$. As usual, $z^A$ can be easily computed if $A$ is in Jordan canonical form. In particular, for a Jordan block $J$ we obtain

$$z^J = z^{\alpha} \begin{pmatrix} 1 & \ln(z) & \frac{\ln(z)^2}{2!} & \dots & \frac{\ln(z)^{n-1}}{(n-1)!} \\ & 1 & \ln(z) & \ddots & \vdots \\ & & 1 & \ddots & \frac{\ln(z)^2}{2!} \\ & & & \ddots & \ln(z) \\ & & & & 1 \end{pmatrix}. \tag{4.17}$$

Therefore the solution consists of terms of the form $z^{\alpha} \ln(z)^k$, where $\alpha$ is an eigenvalue of $A$ and $k$ is a nonnegative integer. Note that the logarithmic terms are only present if $A$ is not diagonalizable.

This behavior is in fact typical near any isolated singularity as the following result shows.

**Theorem 4.3.** *Suppose $A(z)$ is analytic in $\Omega = \{z \in \mathbb{C} | 0 < |z - z_0| < \varepsilon\}$. Then a fundamental system of $w' = A(z)w$ is of the form*

$$W(z) = U(z)(z - z_0)^M, \tag{4.18}$$

*where $U(z)$ is analytic in $\Omega$.*

**Proof.** Again we use our change of coordinates $\zeta = \ln(z)$ to obtain

$$\omega' = \mathrm{e}^{\zeta} A(\mathrm{e}^{\zeta})\omega, \qquad \mathrm{Re}(\zeta) < \ln(\varepsilon). \tag{4.19}$$

But this system is periodic with period $2\pi\mathrm{i}$ and hence the result follows as in the proof of Floquet's theorem (Theorem 3.9). □

Observe that any other fundamental system $\tilde{W}(z)$ can be written as

$$\tilde{W}(z) = W(z)C = U(z)C\,(z - z_0)^{C^{-1}MC}, \qquad \det(C) \neq 0, \tag{4.20}$$

and hence has a representation $\tilde{W}(z) = \tilde{U}(z)(z - z_0)^{\tilde{M}}$, where $\tilde{M}$ is linearly equivalent to $M$.

Please note that this theorem does *not* say that all the *bad* terms are sitting in $(z - z_0)^B$. In fact, $U(z)$ might have an essential singularity at $z_0$. However, if this is not the case, the singularity is called **regular** and we can easily absorb the pole of $U(z)$ in the $(z - z_0)^B$ term by using

$$W(z) = U(z)(z - z_0)^n\,(z - z_0)^{B - n\mathbb{I}}. \tag{4.21}$$

But when can this be done? We expect this to be possible if the singularity of $A(z)$ is not too bad. However, the equation $w' = \frac{1}{z^2}w$ has the solution $w(z) = \exp(-\frac{1}{z})$, which has an essential singularity at 0. Hence our only hope left are first order poles. We will say that $z_0$ is a **simple singularity** of our system if $A(z)$ has a pole of (at most) first order at $z_0$.

**Theorem 4.4.** *Suppose $A(z)$ is analytic in $\Omega = \{z \in \mathbb{C} | 0 < |z - z_0| < \varepsilon\}$ and has a simple singularity at $z_0$. Then $U(z)$ in (4.18) can be chosen analytic in $\{z \in \mathbb{C} |\, |z - z_0| < \varepsilon\}$.*

**Proof.** It is no restriction to consider $z_0 = 0$ and it suffices to show that $U(z)$ can have at most a pole. Let $w(z)$ be any solution. Moreover, for given $r_0 > 0$ we can find a number $n$ such that $\|A(z)\| \le \frac{n}{|z|}$. Using polar coordinates $z = r\mathrm{e}^{\mathrm{i}\varphi}$ we have

$$\begin{aligned} |w(r\mathrm{e}^{\mathrm{i}\varphi})| &= |w(r_0\mathrm{e}^{\mathrm{i}\varphi}) + \int_r^{r_0} A(s\mathrm{e}^{\mathrm{i}\varphi})w(s\mathrm{e}^{\mathrm{i}\varphi})\mathrm{e}^{\mathrm{i}\varphi}ds| \\ &\le |w(r_0\mathrm{e}^{\mathrm{i}\varphi})| + \int_r^{r_0} \frac{n}{s}|w(s\mathrm{e}^{\mathrm{i}\varphi})|ds \end{aligned} \tag{4.22}$$

for $0 < r \le r_0$. Applying Gronwall and taking the maximum over all $\varphi$ we obtain

$$|w(z)| \le \sup_{\zeta:|\zeta|=r_0} |w(\zeta)| \left|\frac{r_0}{z}\right|^n, \tag{4.23}$$

which is the desired estimate. □

The converse of this result is in general not true, however, note that

$$A(z) = U'(z)U(z)^{-1} + \frac{1}{z-z_0}U(z)MU(z)^{-1} \tag{4.24}$$

shows that $A(z)$ cannot have an essential singularity if $U(z)$ has none.

**Lemma 4.5.** *If $z_0$ is a regular singularity, then $A(z)$ has at most a pole at $z_0$.*

**Problem 4.2.** *Let $z_0$ be a simple singularity and let $W(z)$ be a fundamental system as in (4.18). Show that*

$$\det(W(z)) = (z-z_0)^{\mathrm{tr}(A_0)} d(z), \qquad d(z_0) \ne 0,$$

*where $d(z)$ is analytic near $z_0$ and $A_0 = \lim_{z\to z_0}(z-z_0)A(z)$. Moreover, conclude that $\mathrm{tr}(A_0 - M) \in \mathbb{Z}$. (Hint: Use Liouville's formula (3.58) for the determinant.)*

## 4.3. The Frobenius method

In this section we pursue our investigation of simple singularities. Without loss of generality we will set $z_0 = 0$. Since we know how a fundamental system looks like from Theorem 4.4, we can make the ansatz

$$W(z) = U(z)z^M, \qquad U(z) = \sum_{j=0}^{\infty} U_j z^j, \quad U_0 \ne 0. \tag{4.25}$$

Using

$$A(z) = \frac{1}{z}\sum_{j=0}^{\infty} A_j z^j \tag{4.26}$$

and plugging everything into our differential equation yields the recurrence relation

$$U_j\,(j+M) = \sum_{k=0}^{j} A_k U_{j-k} \tag{4.27}$$

for the coefficients $U_j$. However, since we don't know $M$ this does not help us much. By (4.16) you could suspect that we just have $M = A_0$ and $U_0 = \mathbb{I}$. Indeed, if we assume $\det(U_0) \ne 0$, we obtain $U_0 M = A_0 U_0$ for $j = 0$ and hence $W(z)U_0^{-1} = U(z)U_0^{-1}z^{A_0}$ is of the anticipated form. Unfortunately, we don't know that $\det(U_0) \ne 0$ and, even worse, this is wrong in general (examples will follow).

So let us be less ambitious and look for a single solution first. If $\mu$ is an eigenvalue with corresponding eigenvector $u_0$ of $M$, then

$$w_0(z) = W(z)u_0 = z^\mu U(z)u_0 \tag{4.28}$$

is a solution of the form

$$w_0(z) = z^\alpha u_0(z), \quad u_0(z) = \sum_{j=0}^\infty u_{0,j} z^j, \quad u_{0,0} \neq 0,\ \alpha = \mu + m. \tag{4.29}$$

Here $m \in \mathbb{N}_0$ is chosen such that $u_0(0) = u_{0,0} \neq 0$. Inserting this ansatz into our differential equation we obtain

$$(A_0 - \alpha - j)u_{0,j} + \sum_{k=1}^{j} A_k u_{0,j-k} = 0. \tag{4.30}$$

In particular, for $j = 0$,

$$(A_0 - \alpha)u_{0,0} = 0, \tag{4.31}$$

we see that $\alpha$ must be an eigenvalue of $A_0$!

Now what about the case where $\mu$ corresponds to a nontrivial Jordan block of size $n > 1$? Then, by (4.17), we have a corresponding set of generalized eigenvectors $u_l$, $1 \leq l \leq n$, such that

$$w_l(z) = W(z)u_l = z^\alpha \left( u_l(z) + \ln(z) u_{l-1}(z) + \cdots + \frac{\ln(z)^l}{l!} u_0(z) \right), \tag{4.32}$$

$1 \leq l \leq n$, are $n$ solutions. Here

$$u_l(z) = z^{\mu-\alpha} U(z) u_l = \sum_{j=m_l}^\infty u_{l,j} z^j, \quad u_{l,m_l} \neq 0, \quad 1 \leq l \leq n, \tag{4.33}$$

As before, $m_\ell \in \mathbb{Z}$ is chosen such that $u_{l,m_l} \neq 0$ (note that $m_l \geq \mu - \alpha = -m$). We set $u_{l,j} = 0$ for $j < m_l$ and $u_{-1,j} = 0$ for notational convenience later on.

Again, inserting this ansatz into our differential equation, we obtain

$$(A_0 - \alpha - j)u_{l,j} + \sum_{k=1}^{j} A_k u_{l,j-k} = u_{l-1,j}. \tag{4.34}$$

Considering $j < m_l$ we see $u_{l-1,j} = 0$ for $j < m_l$ and thus $m_{l-1} \geq m_l$. In particular, $-m_l \in \mathbb{N}_0$ since $m_0 = 0$. Furthermore, for $j = m_l$ we get

$$(A_0 - \alpha - m_l)u_{l,m_l} = u_{l-1,m_l}. \tag{4.35}$$

Hence there are two cases, $m_l = m_{l-1}$ and $(A_0 - \alpha - m_l)u_{l,m_l} = u_{l-1,m_{l-1}}$, that is, $\alpha + m_{l-1}$ corresponds to a nontrivial Jordan block of $A_0$. Or $m_l < m_{l-1}$ and $(A_0 - \alpha - m_l)u_{l,m_l} = 0$, that is, $\alpha + m_l$ is another eigenvalue of $A_0$.

In summary,

**Theorem 4.6.** *If $A(z)$ has a simple pole at $z_0 = 0$ with residue $A_0$, then every solution of $w' = A(z)w$ is of the form*

$$w(z) = z^\alpha \sum_{k=0}^{l} u_{l-k}(z) \frac{\ln(z)^k}{k!}, \qquad u_l(z) = \sum_{j=m_l}^{\infty} u_{l,j} z^j, \quad u_{l,m_l} \neq 0, \quad (4.36)$$

*where $-m_l \in \mathbb{N}_0$ and $m_l \le m_{l-1} \le \cdots \le m_1 \le m_0 = 0$. The vectors $u_{l,m_l}$ are eigenvectors, $(A_0 - \alpha + m_l)u_{l,m_l} = 0$, if $m_l = m_{l-1}$ (set $m_{-1} = 0$) or generalized eigenvectors, $(A_0 - \alpha + m_l)u_{l,m_l} = u_{l,m_{l-1}}$, if $m_l < m_{l-1}$.*

In particular, the Jordan structures of $M$ and $A_0$ are related as follows:

**Theorem 4.7.** *For every eigenvalue $\mu$ of $M$ there must be an eigenvalue $\alpha = \mu + m$, $m \in \mathbb{N}_0$, of $A_0$. For every Jordan block of $\mu$ there is a corresponding Jordan block of $\alpha$, which might be smaller or equal. If it is smaller, there must be eigenvalues $\alpha_j = \alpha + m_j$, $-m_j \in \mathbb{N}$, of $A_0$ with corresponding Jordan blocks, which make up for the missing parts.*

*If no two eigenvalues of $A_0$ differ by an integer, then $A_0$ and $M$ are similar.*

So we have found a quite complete picture of the possible forms of solutions of our differential equation in the neighborhood of the singular point $z = 0$ and we can now try to go the opposite way. Given a solution of the system of linear equations (4.34), where $\alpha$ is an eigenvalue of $A_0$ we get a solution of our differential equation via (4.32) provided we can show that the series converges.

But before turning to the problem of convergence, let us reflect about how to solve the system (4.34). If the numbers $\alpha + j$ are not eigenvalues of $A_0$ for $j > 0$, we can multiply (4.34) by $(\alpha + m_l + j - A_0)^{-1}$ and $u_{l,j}$ is uniquely determined by $u_{l,j-1}$. Whereas this might not always be true, it is at least true for $j > j_0$ with $j_0$ sufficiently large. Hence we are left with a finite system for the coefficients $u_{l,j}$, $0 \le l \le n$, $0 \le j \le j_0$, which we can solve first. All remaining coefficients are then determined uniquely in a recursive manner.

**Theorem 4.8.** *Suppose $u_{l,j}$ solves (4.34), then $u_l(z)$ defined via the power series (4.33) has the same radius of convergence as the power series for $zA(z)$ around $z = 0$. Moreover, $w_l(z)$ defined via (4.32) is a solution of $w' = A(z)w$.*

**Proof.** Suppose $\delta$ is smaller than the radius of convergence of the power series for $zA(z)$ around $z = 0$. We equip the space of expansion coefficients with the norm (Problem 4.3)

$$\|u_j\| = \sum_{j=0}^{\infty} |u_j|\,\delta^j. \tag{4.37}$$

The idea is now to cut off the first $j_0$ terms which cause trouble and view the rest as a fixed point equation in the above Banach space. Let

$$Ku_j = \begin{cases} 0 & j \le j_0 \\ \frac{1}{\gamma+j}\sum_{k=0}^{j} A_k u_{j-k} & j > j_0 \end{cases}, \tag{4.38}$$

then

$$\|Ku_j\| \le \frac{1}{j_0 - |\mathrm{Re}(\gamma)|}\sum_{j=0}^{\infty}\sum_{k=0}^{j} |A_{j-k}|\,|u_k|\delta^j = \frac{\|A_j\|}{j_0 - |\mathrm{Re}(\gamma)|}\|u_j\|. \tag{4.39}$$

Hence for $j_0$ sufficiently large, the equation $u_j = v_j + Ku_j$ has a unique solution by the contraction principle for any fixed $v_j$. Now let $u_{l,j}$ be a solution of (4.34) and choose $\gamma = \alpha + m_l$ and $v_j = u_{l,j}$ for $j \le j_0$ respectively $v_j = -\frac{1}{\alpha+m_l+j}u_{l-1,j}$ for $j > j_0$. Then the solution of our fixed point problem coincides with our solution $u_{l,j}$ of (4.34) by construction. □

In summary, we obtain the following procedure for finding a full set of linearly independent solutions:

For all eigenvalues $\alpha$ of $A_0$ for which $\alpha + j$ is not an eigenvalue for all $j \in \mathbb{N}_0$, take corresponding generalized eigenfunctions $u_{0,l} \neq 0$, $(A_0 - \alpha)u_{0,l} = u_{0,l-1}$. Then $w_l(z)$ as defined in (4.32) with $m_l = 0$ and

$$u_{l,j} = (A_0 - \alpha - j)^{-1}\left(u_{l-1,j} - \sum_{k=1}^{j} a_k u_{l,j-k}\right), \tag{4.40}$$

are linearly independent solutions.

For all other eigenvalues $\tilde{\alpha} = \alpha + m_j$, there are two cases. First try to find solutions for $\tilde{\alpha}$ as in the case before until a sufficient number of solutions has been found or until this is no longer possible (i.e., (4.34) has no nontrivial solution). Next, add further terms in the ansatz (4.32) for $\alpha$ until a sufficient number of solutions has been found. This will produce a full set of linearly independent solutions.

This procedure for finding the general solution near a simple singularity is known as **Frobenius method**. The eigenvalues of $A_0$ are also called **characteristic exponents**. Observe that our requirement of the singularity to be simple is indeed crucial, since it ensures that the algebraic system of equations for the coefficients can be solved recursively.

Finally, let me remark, that we can also try to apply this procedure to get a power series around infinity. To do this, one makes the change of coordinates $\zeta = \frac{1}{z}$, then our system transforms to

$$\omega' = -\frac{1}{\zeta^2} A(\frac{1}{\zeta})\omega, \qquad w(z) = \omega(\frac{1}{z}). \tag{4.41}$$

In particular, $\infty$ is a simple singularity if and only if $A(z)$ has (at least) a first order zero at $\infty$, that is,

$$A(\frac{1}{\zeta}) = \zeta \sum_{j=0}^{\infty} A_j \zeta^j. \tag{4.42}$$

A system is called a **Fuchs system** if it has only finitely many singularities all of which, including infinity, are simple. It then follows from Liouville's theorem (every bounded analytic function is constant) that $A(z)$ must be rational.

**Lemma 4.9.** *Every Fuchs system is of the form*

$$A(z) = \sum_{j=1}^{k} \frac{A_j}{z - z_j}. \tag{4.43}$$

**Problem 4.3.** *Let $w_j > 0$, $J \in \mathbb{N}_0$, be given weights. Show that the set of all complex-valued sequences $\{u_j\}_{j\in\mathbb{N}_0}$ for which the norm*

$$\|u_j\| = \sum_{j=0}^{\infty} |u_j|\, w_j$$

*is finite, form a Banach space.*

## 4.4. Second order equations

In this section we want to apply our theory to second order equations

$$u'' + p(z)u' + q(z)u = 0 \tag{4.44}$$

We will assume that the singular point is $z_0 = 0$ for notational convenience.

Transforming (4.44) to a system as usual shows that $z_0 = 0$ is a simple singularity if both $p(z)$ and $q(z)$ have at most a first order pole. However, we can do even better. Introducing $w(z) = (u(z), z\,u'(z))$ we obtain

$$w' = A(z)w, \qquad A(z) = \begin{pmatrix} 0 & \frac{1}{z} \\ -zq(z) & \frac{1}{z} - p(z) \end{pmatrix} \tag{4.45}$$

and $z_0 = 0$ is a simple singularity if $p(z)$ and $zq(z)$ have at most first order poles. This is even optimal.

**Theorem 4.10** (Fuchs)**.** *The system (4.45) has a regular singularity at $z_0$ if and only if $p(z)$ and $zq(z)$ have at most first order poles.*

**Proof.** If (4.45) has a regular singularity, there is a solution of the form $u(z) = z^\alpha h(z)$, where $h(0) = 1$ and $h(z)$ is analytic near 0. Let $v(z)$ be a linearly independent solution and consider $c(z) = v(z)/u(z)$. Then, since $c(z)$ has no essential singularity,

$$p(z) = -\frac{c''(z)}{c'(z)} - 2\frac{u'(z)}{u(z)} \tag{4.46}$$

has a first order pole. Moreover,

$$q(z) = -\frac{u''(z)}{u(z)} - p(z)\frac{u'(z)}{u(z)} \tag{4.47}$$

has at most a second order pole. □

I remark that using induction on the order of the differential equation, one can show the analogous result for $n$-th order equations.

Hence we will assume that the coefficients are of the form

$$p(z) = \frac{1}{z}\sum_{j=0}^{\infty} p_j z^j, \qquad q(z) = \frac{1}{z^2}\sum_{j=0}^{\infty} q_j z^j, \tag{4.48}$$

such that we can apply the Frobenius method from the previous section.

The characteristic exponents are the eigenvalues of the matrix

$$A_0 = \begin{pmatrix} 0 & 1 \\ -q_0 & 1 - p_0 \end{pmatrix} \tag{4.49}$$

and are given by

$$\alpha_{1,2} = \frac{1}{2}(1 - p_0 \pm \sqrt{(p_0 - 1)^2 - 4q_0}). \tag{4.50}$$

Taking the standard branch of the root, we have $\mathrm{Re}(\alpha_1) \geq \mathrm{Re}(\alpha_2)$ and the analysis in our previous section implies

**Theorem 4.11.** *Suppose the coefficients $p(z)$ and $q(z)$ have poles of order (at most) one and two respectively . Then, using the notation form above, two cases can occur:*

*Case 1. If $\alpha_1 - \alpha_2 \notin \mathbb{N}_0$, a fundamental system of solutions is given by*

$$u_j(z) = z^{\alpha_j} h_j(z), \quad j = 1, 2, \tag{4.51}$$

*where the functions $h_j(z)$ are analytic near $z = 0$ and satisfy $h_j(0) = 1$.*

*Case 2. If $\alpha_1 - \alpha_2 = m \in \mathbb{N}_0$, a fundamental system of solutions is given by*

$$\begin{aligned} u_1(z) &= z^{\alpha_1} h_1(z), \\ u_2(z) &= z^{\alpha_2} h_2(z) + c\ln(z) u_1(z), \end{aligned} \tag{4.52}$$

*where the functions $h_j(z)$ are analytic near $z = 0$ and satisfy $h_j(0) = 1$. The constant $c \in \mathbb{C}$ might be zero unless $m = 0$.*

Now, let us see how this method works by considering an explicit example. This will in addition show that all cases from above can occur. The example is the famous **Bessel equation**

$$z^2 u'' + z u' + (z^2 - \nu^2) u = 0, \qquad \nu \in \mathbb{C}. \tag{4.53}$$

It is no restriction to assume $\mathrm{Re}(\nu) \geq 0$ and hence we will do so. The eigenvalues of $A_0$ are given by $\alpha_{1,2} = \pm\nu$ and hence there is a solution of the form

$$u_1(z) = z^\nu \sum_{j=0}^\infty h_{1,j} z^j, \qquad h_{1,0} = 1. \tag{4.54}$$

Plugging this into our equation yields

$$z^2 \sum_{j=0}^\infty h_{1,j}(j+\nu-1)(j+\nu) z^{j+\nu-2} + z \sum_{j=0}^\infty h_{1,j}(j+\nu) z^{j+\nu-1} + (z^2 - \nu^2) \sum_{j=0}^\infty h_{1,j} z^{j+\nu} = 0 \tag{4.55}$$

and after multiplying by $z^{-\nu}$ and aligning powers of $z$

$$\sum_{j=0}^\infty \left( h_{1,j}(j+\nu-1)(j+\nu) + h_{1,j}(j+\nu) + h_{1,j-2} - h_{1,j}\nu^2 \right) z^j = 0, \tag{4.56}$$

where we set $h_{1,j} = 0$ for $j < 0$. Comparing coefficients we obtain the recurrence relation

$$j(j+2\nu) h_{1,j} + h_{1,j-2} = 0 \tag{4.57}$$

for the unknown expansion coefficients $h_{1,j}$. In particular, this can be viewed as two independent recurrence relations for the even $h_{1,2j}$ and odd $h_{1,2j+1}$ coefficients. The solution is easily seen to be

$$h_{1,2j} = \frac{(-1)^j}{4^j j! (\nu+1)_j}, \qquad h_{2j+1} = 0, \tag{4.58}$$

where we have used the **Pochhammer symbol**

$$(x)_0 = 1, \quad (x)_j = x(x+1)\cdots(x+j-1). \tag{4.59}$$

This solution, with a different normalization, is called **Bessel function**

$$J_\nu(z) = \frac{u_1(z)}{2^\nu \Gamma(\nu+1)} = \sum_{j=0}^\infty \frac{(-1)^j}{j! \Gamma(\nu+j+1)} \left(\frac{z}{2}\right)^{2j+\nu} \tag{4.60}$$

of order $\nu$. Now what about the second solution? So let us investigate the equation for $-\nu$. Replacing $\nu$ by $-\nu$ in the previous calculation, we see that we can find a second (linearly independent) solution $J_{-\nu}(z)$ provided $(-\nu+1)_j \neq 0$ for all $j$, which can only happen if $\nu \in \mathbb{N}_0$. Hence there are no logarithmic terms even for $\nu = \frac{2n+1}{2}$, where $\alpha_1 - \alpha_2 = 2\nu = 2n+1 \in \mathbb{N}$. It remains to look at the case, where $\nu = n \in \mathbb{N}$. All odd coefficients must be zero and the recursion for the even ones gives us a contradiction at $2j = 2n$. Hence the only possibility left is a logarithmic solution

$$u_2(z) = z^{-n} h_2(z) + c\ln(z) u_1(z). \tag{4.61}$$

Inserting this into our equation yields

$$j(j-2n)h_{2,j} + h_{2,j-2} = -2c(j-n)h_{1,j-2\nu}. \tag{4.62}$$

Again all odd coefficients vanish, $h_{2,2j+1} = 0$. The even coefficients $h_{2,2j}$ can be determined recursively for $j < n$ as before

$$h_{2,2j} = \frac{1}{4^j j! (\nu-1)_j}, \qquad j < n. \tag{4.63}$$

The recursion for $j = 2n$ reads $h_{2,2(n-1)} = -2c\,n$ from which

$$c = \frac{-2}{4^n n!(n-1)!} \tag{4.64}$$

follows. The remaining coefficients now follow recursively from

$$4j(j+n)h_{2,2j+2n} + h_{2,2(j-1)+2n} = -2c(2j+n)h_{1,2j} \tag{4.65}$$

once we choose a value for $h_{2,2n}$. This is a first order linear inhomogeneous recurrence relation with solution given by (see Problem 4.4 and note that the solution of the homogeneous equation is $h_{1,2j}$)

$$h_{2,2j+2n} = h_{1,2j}\left(h_{2,2n} - \frac{c}{2}\sum_{k=1}^{j}\frac{2k+n}{k(k+n)}\right). \tag{4.66}$$

Choosing $h_{2,2n} = \frac{c}{2} H_n$, where

$$H_j = \sum_{k=1}^{j}\frac{1}{k} \tag{4.67}$$

are the **harmonic numbers**, we obtain

$$h_{2,2n+2j} = \frac{(-1)^j (H_{j+n} + H_j)}{4^{j+n}(n-1)! j! (j+n)!}. \tag{4.68}$$

Usually, the following linear combination

$$\begin{aligned} Y_n(z) &= -\frac{2^n(n-1)!}{\pi}u_2(z) + \frac{\gamma - \ln(2)}{2^{n-1}\pi n!}u_1(z) \\ &= \frac{2}{\pi}(\gamma + \ln(\frac{z}{2}))J_n(z) - \frac{1}{\pi}\sum_{j=0}^{n-1}\frac{(-1)^j(n-1)!}{j!(1-n)_j}\left(\frac{z}{2}\right)^{2j-n} \\ &\quad - \frac{1}{\pi}\sum_{j=0}^{\infty}\frac{(-1)^j(H_{j+n}+H_j)}{j!(j+n)!}\left(\frac{z}{2}\right)^{2j+n} \end{aligned} \tag{4.69}$$

is taken as second independent solution. Here $\gamma = \lim_{j\to\infty}(H_j - \ln(j))$ is the **Euler constant**.

Finally, let me remark that one usually uses the **Hankel function**

$$Y_\nu(z) = \frac{\cos(\pi\nu)J_\nu(z) - J_{-\nu}(z)}{\sin(\pi\nu)} \tag{4.70}$$

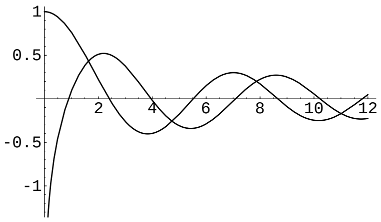
as second solution of the Bessel equation. For fixed $z \neq 0$ the right hand side has a singularity for $\nu \in \mathbb{N}_0$. However, since

$$J_{-\nu}(z) = (-1)^\nu J_\nu(z), \qquad \nu \in \mathbb{N}_0, \tag{4.71}$$

it can be removed and it can be shown that the limit is a second linearly independent solution (Problem 4.5) which coincides with the one from above.

Whereas you might not find Bessel functions on your pocket calculator, they are available in *Mathematica*. For example, here is a plot of the Bessel and Hankel function of order $\nu = 0$.

```
In[1]:= Plot[{BesselJ[0,z],BesselY[0,z]},{z,0,12}];
```




**Problem 4.4.** *Consider the first order liner inhomogeneous difference equation*

$$x(n+1) - f(n)x(n) = g(n), \quad f(n) \neq 0.$$

*Show that the solution of the homogeneous equation ($g = 0$) is given by*

$$x_h(n) = x(0) \begin{cases} \prod_{j=0}^{n-1} f(j) & \text{for } n > 0 \\ 1 & \text{for } n = 0 \\ \prod_{j=n}^{-1} f(j)^{-1} & \text{for } n < 0 \end{cases} .$$

*Use a variation of constants ansatz for the inhomogeneous equation and show that the solution is given by*

$$x(n) = x_h(n) + \begin{cases} x_h(n) \sum_{j=0}^{n-1} \frac{g(j)}{x_h(j+1)} & \text{for } n > 0 \\ 0 & \text{for } n = 0 \\ -x_h(n) \sum_{j=n}^{-1} \frac{g(j)}{x_h(j+1)} & \text{for } n < 0 \end{cases} .$$

**Problem 4.5** (Hankel functions). *Prove that the Hankel function is a second linearly independent solution for all $\nu$ as follows:*

(i) *Prove (4.71) and conclude that the Hankel function is well defined for all $\nu$ and holomorphic in both variables $z$ and $\nu$ (for $z \in \mathbb{C}\backslash(-\infty, 0]$ and $\mathrm{Re}(\nu) > 0$).*

(ii) *Show that the modified Wronskian*

$$W(u(z), v(z)) = z(u(z)v'(z) - u'(z)v(z))$$

*of two solutions of the Bessel equation is constant (Hint: Liouville's formula). Prove*

$$W(J_\nu(z), J_{-\nu}(z)) = \frac{-2}{\Gamma(\nu)\Gamma(1-\nu)} = -\frac{2}{\pi}\sin(\pi\nu).$$

*(Hint: Use constancy of the Wronskian and evaluate it at $z = 0$. You don't need to prove the formula for the gamma functions.)*

(iii) *Now show*

$$W(J_\nu(z), Y_\nu(z)) = \frac{2}{\pi}.$$

*Differentiate this formula with respect to $z$ and show that $Y_\nu(z)$ satisfies the Bessel equation.*

**Problem 4.6.** *Prove the following properties of the Bessel functions.*

(i) $(z^{\pm\nu} J_\nu(z))' = \pm z^{\pm\nu} J_{\nu\mp1}(z)$.

(ii) $J_{\nu+1}(z) + J_{\nu-1}(z) = \frac{2\nu}{z} J_\nu(z)$.

(iii) $J_{\nu+1}(z) - J_{\nu-1}(z) = 2J_\nu(z)'$.

**Problem 4.7.** *Many differential equations occur in practice that are not of the standard form (4.53). Show that the differential equation*

$$w'' + \frac{1-2a}{z}w' + \left((bcz^{c-1})^2 + \frac{a^2 - \nu^2 c^2}{z^2}\right) w = 0.$$

*can be transformed to the Bessel equation via $w(z) = z^a u(bz^c)$.*

*Find the solution of*

- $w' + w^2 = z$,
- $w' = w^2 - z^2$

*in terms of Bessel functions. (Hint: Problem 3.16 (iv).)*

**Problem 4.8** (Legendre polynomials). *The* **Legendre equation** *is given by*

$$(1 - z^2)w'' - 2zw' + n(n+1)w = 0.$$

*Make a power series ansatz at $z = 0$ and show that there is a polynomial solution $p_n(z)$ if $n \in \mathbb{N}_0$. What is the order of $p_n(z)$?*

**Problem 4.9** (Hypergeometric equation). *The* **hypergeometric equation** *is given by*

$$z(1-z)w'' + (c - (1+a+b)z)w' - abw = 0.$$

*Classify all singular points (including $\infty$). Use the Frobenius method to show that*

$$F(a,b,c;z) = \sum_{j=0}^{\infty} \frac{(a)_j (b)_j}{(c)_j j!} z^j, \qquad -c \notin \mathbb{N}_0,$$

*is a solution. This is the* **hypergeometric function**. *Show that $z^{1-c} w(z)$ is again a solution of the hypergeometric equation but with different coefficients. Use this to prove that $F(a-c+1, b-c+1, 2-c; z)$ is a second solution for $c-2 \notin \mathbb{N}_0$. This gives two linearly independent solutions if $c \notin \mathbb{Z}$.*

**Problem 4.10** (Confluent hypergeometric equation). *The* **confluent hypergeometric equation** *is given by*

$$zw'' + (c-z)w' - aw = 0.$$

*Classify all singular points (including $\infty$). Use the Frobenius method to show that*

$$K(a,b;z) = \sum_{j=0}^{\infty} \frac{(a)_j}{(c)_j j!} z^j, \qquad -c \notin \mathbb{N}_0,$$

*is a solution. This is the* **confluent hypergeometric or Kummer function**.

*Show that $z^{1-c} w(z)$ is again a solution of the confluent hypergeometric equation but with different coefficients. Use this prove that $K(a-c+1, 2-c; z)$ is a second solution for $c-2 \notin \mathbb{N}_0$. This gives two linearly independent solutions if $c \notin \mathbb{Z}$.*

**Problem 4.11** (Riemann equation). *A second order equation whose associated system is of Fuchs type is called a* **Riemann equation** *if it has only three singular points (including $\infty$). Solutions of a Riemann equation are denoted by the* **Riemann symbol**

$$P \left\{ \begin{array}{cccc} z_0 & z_1 & z_2 & \\ \alpha_1 & \beta_1 & \gamma_1 & z \\ \alpha_2 & \beta_2 & \gamma_2 & \end{array} \right\},$$

*where the numbers $z_j$ are the singular points and the numbers below $z_j$ are the corresponding characteristic exponents.*

*Recall that given points $z_j$, $j = 0,1,2$, can be mapped to any other given points $\zeta_j = \zeta(z_j)$, $j = 0,1,2$, by a fractional linear transform (Möbius transform)*

$$\zeta(z) = \frac{az+b}{cz+d}, \qquad ad - bc \neq 0.$$

*Pick* $\zeta_0 = 0$, $\zeta_1 = 1$ *and* $\zeta_2 = \infty$ *and show that*

$$P\left\{\begin{array}{cccc} z_0 & z_1 & z_2 & \\ \alpha_1 & \beta_1 & \gamma_1 & z \\ \alpha_2 & \beta_2 & \gamma_2 & \end{array}\right\} = P\left\{\begin{array}{cccc} 0 & 1 & \infty & \\ \alpha_1 & \beta_1 & \gamma_1 & \frac{az+b}{cz+d} \\ \alpha_2 & \beta_2 & \gamma_2 & \end{array}\right\}.$$

*For the case* $z_0 = 0$, $z_1 = 1$ *and* $z_2 = \infty$, *express the coefficients* $p(z)$ *and* $q(z)$ *in terms of the characteristic exponents. Conclude that a Riemann equation is uniquely determined by its symbol.*

*Finally, show*

$$z^\nu (1-z)^\mu P\left\{\begin{array}{cccc} 0 & 1 & \infty & \\ \alpha_1 & \beta_1 & \gamma_1 & z \\ \alpha_2 & \beta_2 & \gamma_2 & \end{array}\right\} = P\left\{\begin{array}{cccc} 0 & 1 & \infty & \\ \alpha_1+\nu & \beta_1+\mu & \gamma_1-\mu-\nu & z \\ \alpha_2+\nu & \beta_2+\mu & \gamma_2-\mu-\nu & \end{array}\right\}$$

*and conclude that any Riemann equation can be transformed into the hypergeometric equation.*

*Show that the Legendre equation is a Riemann equation. Find the transformation which maps it to the hypergeometric equation.*

*Chapter 5*

# Boundary value problems

## 5.1. Introduction

Boundary value problems are of fundamental importance in physics. However, solving such problems usually involves a combination of methods from ordinary differential equations, functional analysis, complex functions, and measure theory. Since the remaining chapters do not depend on the present one, you can also skip it and go directly to Chapter 6.

To motivate the investigation of boundary value problems, let us look at a typical example from physics first. The vibrations of a string can be described by its displacement $u(t,x)$ at the point $x$ and time $t$. The equation of motion for this system is the one dimensional **wave equation**

$$\frac{1}{c^2}\frac{\partial^2}{\partial t^2}u(t,x) = \frac{\partial^2}{\partial x^2}u(t,x), \tag{5.1}$$

where $c$ is the speed of sound in our string. Moreover, we will assume that the string is fixed at both endpoints, that is, $x \in [0,1]$ and $u(t,0) = u(t,1) = 0$, and that the initial displacement $u(0,x) = u(x)$ and the initial velocity $\frac{\partial u}{\partial t}(0,x) = v(x)$ are given.

Unfortunately, this is a partial differential equation and hence none of our methods found thus far apply. In particular, it is unclear how we should solve the posed problem. Hence let us try to find some solutions of the equation (5.1) first. To make it a little easier, let us try to make an ansatz for $u(t,x)$ as a product of two functions, each of which depends on only one variable, that is,

$$u(t,x) = w(t)y(x). \tag{5.2}$$

This ansatz is called **separation of variables**. Plugging everything into the wave equation and bringing all $t$, $x$ dependent terms to the left, right side, respectively, we obtain

$$\frac{1}{c^2}\frac{\ddot{w}(t)}{w(t)} = \frac{y''(x)}{y(x)}. \tag{5.3}$$

Now if this equation should hold for all $t$ and $x$, the quotients must be equal to a constant $-\lambda$. That is, we are lead to the equations

$$-\frac{1}{c^2}\ddot{w}(t) = \lambda w(t) \tag{5.4}$$

and

$$-y''(x) = \lambda y(x), \qquad y(0) = y(1) = 0 \tag{5.5}$$

which can easily be solved. The first one gives

$$w(t) = c_1 \cos(c\sqrt{\lambda}t) + c_2 \sin(c\sqrt{\lambda}t) \tag{5.6}$$

and the second one

$$y(x) = c_3 \cos(\sqrt{\lambda}x) + c_4 \sin(\sqrt{\lambda}x). \tag{5.7}$$

However, $y(x)$ must also satisfy the boundary conditions $y(0) = y(1) = 0$. The first one $y(0) = 0$ is satisfied if $c_3 = 0$ and the second one yields ($c_4$ can be absorbed by $w(t)$)

$$\sin(\sqrt{\lambda}) = 0, \tag{5.8}$$

which holds if $\lambda = (\pi n)^2$, $n \in \mathbb{N}$. In summary, we obtain the solutions

$$u(t,x) = (c_1 \cos(cn\pi t) + c_2 \sin(cn\pi t)) \sin(n\pi x), \qquad n \in \mathbb{N}. \tag{5.9}$$

In particular, the string can only vibrate with certain fixed frequencies!

So we have found a large number of solutions, but we still have not dealt with our initial conditions. This can be done using the superposition principle which holds since our equation is linear. In fact, choosing

$$u(t,x) = \sum_{n=1}^{\infty} \Big( c_{1,n} \cos(cn\pi t) + \frac{c_{2,n}}{cn\pi} \sin(cn\pi t) \Big) \sin(n\pi x), \tag{5.10}$$

where the coefficients $c_{1,n}$ and $c_{2,n}$ decay sufficiently fast, we obtain further solutions of our equation. Moreover, these solutions satisfy

$$u(0,x) = \sum_{n=1}^{\infty} c_{1,n} \sin(n\pi x), \qquad \frac{\partial}{\partial t} u(0,x) = \sum_{n=1}^{\infty} c_{2,n} \sin(n\pi x). \tag{5.11}$$

Hence, expanding the initial conditions into Fourier series

$$u(x) = \sum_{n=1}^{\infty} u_n \sin(n\pi x), \qquad v(x) = \sum_{n=1}^{\infty} v_n \sin(n\pi x), \tag{5.12}$$

we see that the solution of our original problem is given by (5.10) if we choose $c_{1,n} = u_n$ and $c_{2,n} = v_n$.

In general, a vast number of problems in various areas leads to the investigation of the following problem

$$Ly(x) = \lambda y(x), \qquad L = \frac{1}{r(x)} \left( -\frac{d}{dx} p(x) \frac{d}{dx} + q(x) \right), \tag{5.13}$$

subject to the **boundary conditions**

$$\cos(\alpha) y(a) = \sin(\alpha) p(a) y'(a), \quad \cos(\beta) y(b) = \sin(\beta) p(b) y'(b), \tag{5.14}$$

$\alpha, \beta \in \mathbb{R}$. Such a problem is called **Sturm–Liouville boundary value problem**. Our example shows that we should prove the following facts about Sturm–Liouville problems:

(i) The Sturm–Liouville problem has a countable number of eigenvalues $E_n$ with corresponding eigenfunctions $u_n(x)$, that is, $u_n(x)$ satisfies the boundary conditions and $Lu_n(x) = E_n u_n(x)$.

(ii) The eigenfunctions $u_n$ are complete, that is, any *nice* function $u(x)$ can be expanded into a generalized Fourier series

$$u(x) = \sum_{n=1}^{\infty} c_n u_n(x).$$

This problem is very similar to the eigenvalue problem of a matrix. However, our linear operator is now acting on some space of functions which is not finite dimensional. Nevertheless, we can equip such a function space with a scalar product

$$\langle f, g \rangle = \int_a^b f^*(x) g(x) dx, \tag{5.15}$$

where '$*$' denotes complex conjugation. In fact, it turns out that the proper setting for our problem is a Hilbert space and hence we will recall some facts about Hilbert spaces in the next section before proceeding further.

**Problem 5.1.** *Find conditions for the initial values $u(x)$ and $v(x)$ such that (5.10) is indeed a solution (i.e., such that interchanging the order of summation and differentiation is admissible). (Hint: The decay of the Fourier coefficients is related to the smoothness of the function.)*

**Problem 5.2.** *Show that*

$$q_2(x) y'' + q_1(x) y' + q_0(x) y$$

*can be written as*

$$\frac{1}{r(x)} \left( (p(x) y')' + q(x) y \right).$$

*Find $r$, $p$, $q$ in terms of $q_0$, $q_1$, $q_2$.*

*Write the Bessel and Legendre equations (Problem 4.8) in this form.*

**Problem 5.3** (Hanging cable)**.** *Consider the vibrations of a cable suspended at $x = 1$. Denote the displacement by $u(t, x)$. Then the motion is described by the equation*

$$\frac{\partial^2}{\partial t^2} u(t, x) = g \frac{\partial}{\partial x} x \frac{\partial}{\partial x} u(t, x),$$

*with boundary conditions $u(t, 1) = 0$. Find all solutions of the form $u(t, x) = w(t)y(x)$. (Hint: Problem 4.7)*

**Problem 5.4** (Harmonic crystal in one dimension)**.** *Suppose you have a linear chain of identical particles coupled to each other by springs. Then the equation of motion is given by*

$$m \frac{d^2}{dt^2} u(t, n) = k(u(t, n+1) - u(t, n)) + k(u(t, n-1) - u(t, n)),$$

*where $m > 0$ is the mass of the particles and $k > 0$ is the spring constant. (This is an infinite system of differential equations to which our theory does not apply!) Look for a solution in terms of Bessel functions $c(t, n) = J_{an}(bt)$ (Hint: Problem 4.6.). Show that $s(t, n) = \int_0^t c(s, n) ds$ is a second solution. Can you give the solution corresponding to the initial data $u(0, n) = u(n)$, $\frac{du}{dt}(0, n) = v(n)$ provided $u(n)$ and $v(n)$ decay sufficiently fast?*

## 5.2. Symmetric compact operators

Suppose $\mathfrak{H}_0$ is a vector space. A map $\langle ., .. \rangle : \mathfrak{H}_0 \times \mathfrak{H}_0 \to \mathbb{C}$ is called skew linear form if it is conjugate linear in the first and linear in the second argument, that is,

$$\begin{array}{rcl} \langle \lambda_1 f_1 + \lambda_2 f_2, g \rangle &=& \lambda_1^* \langle f_1, g \rangle + \lambda_2^* \langle f_2, g \rangle \\ \langle f, \lambda_1 g_1 + \lambda_2 g_2 \rangle &=& \lambda_1 \langle f, g_1 \rangle + \lambda_2 \langle f, g_2 \rangle \end{array}, \quad \lambda_1, \lambda_2 \in \mathbb{C}. \tag{5.16}$$

A skew linear form satisfying the requirements

(i) $\langle f, f \rangle > 0$ for $f \neq 0$.
(ii) $\langle f, g \rangle = \langle g, f \rangle^*$

is called **inner product** or **scalar product**. Associated with every scalar product is a norm

$$\|f\| = \sqrt{\langle f, f \rangle}. \tag{5.17}$$

(We will prove later that this is indeed a norm.) The pair $(\mathfrak{H}_0, \langle ., .. \rangle)$ is called **inner product space**. If $\mathfrak{H}_0$ is complete with respect to the above norm, it is called a **Hilbert space**. It is usually no restriction to assume that $\mathfrak{H}_0$ is complete since one can easily replace it by its completion $\mathfrak{H}$. However, for our purpose this is not necessary and hence we will not do so here to avoid technical complications later on.

A vector $f \in \mathfrak{H}_0$ is called **normalized** if $\|f\| = 1$. Two vectors $f, g \in \mathfrak{H}_0$ are called **orthogonal** if $\langle f, g\rangle = 0$ and a set of vectors $\{u_j\}$ is called orthonormal set if $\langle u_j, u_k\rangle = 0$ for $j \neq k$ and $\langle u_j, u_j\rangle = 1$. If $f, g \in \mathfrak{H}_0$ are orthogonal we have the **Pythagoras theorem**

$$\|f + g\|^2 = \|f\|^2 + \|g\|^2, \tag{5.18}$$

which is straightforward to check.

**Theorem 5.1.** *Suppose $\{u_j\}_{j=0}^n$ is an orthonormal set. Then every $f \in \mathfrak{H}_0$ can be written as*

$$f = f_n + f^{\perp}, \qquad f_n = \sum_{j=0}^{n} \langle u_j, f\rangle u_j, \tag{5.19}$$

*where $f_n$ and $f^{\perp}$ are orthogonal. In particular,*

$$\|f\|^2 = \sum_{j=0}^{n} |\langle u_j, f\rangle|^2 + \|f^{\perp}\|^2. \tag{5.20}$$

**Proof.** A straightforward calculation shows $\langle u_j, f - f_n\rangle = 0$ and hence $f_n$ and $f^{\perp} = f - f_n$ are orthogonal. The remaining formula follows by applying (5.18) iteratively. □

Out of this result we get three important consequences with almost no effort.

(i) **Bessel inequality**:

$$\|f\|^2 \geq \sum_{j=0}^{n} |\langle u_j, f\rangle|^2. \tag{5.21}$$

(ii) **Schwarz inequality**:

$$|\langle f, g\rangle| \leq \|f\|\|g\|. \tag{5.22}$$

(It suffices to prove the case $\|g\| = 1$. But then $g$ forms an orthonormal set and the result follows from Bessel's inequality.)

(iii) The map $\|.\|$ is indeed a norm. Only the triangle inequality is nontrivial. It follows from the Schwarz inequality since

$$\|f + g\|^2 = \|f\|^2 + \langle f, g\rangle + \langle g, f\rangle + \|g\|^2 \leq (\|f\| + \|g\|)^2. \tag{5.23}$$

In particular, Bessel inequality shows that we can also handle countable orthonormal sets. An orthonormal set is called an **orthonormal basis** if

$$\|f\|^2 = \sum_{j} |\langle u_j, f\rangle|^2 \tag{5.24}$$

for all $f \in \mathfrak{H}_0$. Clearly this is equivalent to $\lim_{n\to\infty} f_n = f$ in (5.19) and hence every $f \in \mathfrak{H}_0$ can be written as

$$f = \sum_j \langle u_j, f \rangle u_j. \tag{5.25}$$

A **linear operator** is a linear mapping

$$A : \mathfrak{D}(A) \to \mathfrak{H}_0, \tag{5.26}$$

where $\mathfrak{D}(A)$ is a linear subspace of $\mathfrak{H}_0$, called the **domain** of $A$. A linear operator $A$ is called **symmetric** if its domain is dense (i.e., its closure is $\mathfrak{H}_0$) and if

$$\langle g, Af \rangle = \langle Ag, f \rangle \qquad f, g \in \mathfrak{D}(A). \tag{5.27}$$

A number $z \in \mathbb{C}$ is called **eigenvalue** of $A$ if there is a nonzero vector $u \in \mathfrak{D}(A)$ such that

$$Au = zu. \tag{5.28}$$

The vector $u$ is called a corresponding **eigenvector** in this case. An eigenvalue is called **simple** if there is only one linearly independent eigenvector.

**Theorem 5.2.** *Let $A$ be symmetric. Then all eigenvalues are real and eigenvectors corresponding to different eigenvalues are orthogonal.*

**Proof.** Suppose $\lambda$ is an eigenvalue with corresponding normalized eigenvector $u$. Then $\lambda = \langle u, Au \rangle = \langle Au, u \rangle = \lambda^*$, which shows that $\lambda$ is real. Furthermore, if $Au_j = \lambda_j u_j$, $j = 1, 2$, we have

$$(\lambda_1 - \lambda_2)\langle u_1, u_2 \rangle = \langle Au_1, u_2 \rangle - \langle u_1, Au_2 \rangle = 0 \tag{5.29}$$

finishing the proof. □

The linear operator $A$ defined on $\mathfrak{D}(A) = \mathfrak{H}_0$ is called **bounded** if

$$\|A\| = \sup_{f: \|f\|=1} \|Af\| \tag{5.30}$$

is finite. It is not hard to see that this is indeed a norm (Problem 5.6) on the space of bounded linear operators. By construction, a bounded operator is Lipschitz continuous

$$\|Af\| \le \|A\| \|f\| \tag{5.31}$$

and hence continuous.

Moreover, a linear operator $A$ defined on $\mathfrak{D}(A) = \mathfrak{H}_0$ is called **compact** if every sequence $Af_n$ has a convergent subsequence whenever $f_n$ is bounded. Every compact linear operator is bounded and the product of a bounded and a compact operator is again compact (Problem 5.7).

**Theorem 5.3.** *A symmetric compact operator has an eigenvalue $\alpha_0$ which satisfies $|\alpha_0| = \|A\|$.*

**Proof.** We set $\alpha = \|A\|$ and assume $\alpha \neq 0$ (i.e, $A \neq 0$) without loss of generality. Since

$$\|A\|^2 = \sup_{f:\|f\|=1} \|Af\|^2 = \sup_{f:\|f\|=1} \langle Af, Af\rangle = \sup_{f:\|f\|=1} \langle f, A^2 f\rangle \tag{5.32}$$

there exists a normalized sequence $u_n$ such that

$$\lim_{n\to\infty} \langle u_n, A^2 u_n\rangle = \alpha^2. \tag{5.33}$$

Since $A$ is compact, it is no restriction to assume that $A^2 u_n$ converges, say $\lim_{n\to\infty} A^2 u_n = \alpha^2 u$. Now

$$\begin{aligned} \|(A^2 - \alpha^2) u_n\|^2 &= \|A^2 u_n\|^2 - 2\alpha^2 \langle u_n, A^2 u_n\rangle + \alpha^4 \\ &\leq 2\alpha^2(\alpha^2 - \langle u_n, A^2 u_n\rangle) \end{aligned} \tag{5.34}$$

(where we have used $\|A^2 u_n\| \leq \|A\|\|Au_n\| \leq \|A\|^2\|u_n\| = \alpha^2$) implies $\lim_{n\to\infty}(A^2 u_n - \alpha^2 u_n) = 0$ and hence $\lim_{n\to\infty} u_n = u$. In addition, $u$ is a normalized eigenvector of $A^2$ since $(A^2 - \alpha^2)u = 0$. Factorizing this last equation according to $(A - \alpha)u = v$ and $(A + \alpha)v = 0$ show that either $v \neq 0$ is an eigenvector corresponding to $-\alpha$ or $v = 0$ and hence $u \neq 0$ is an eigenvector corresponding to $\alpha$. □

Note that for a bounded operator $A$, there cannot be an eigenvalue with absolute value larger than $\|A\|$, that is, the set of eigenvalues is bounded by $\|A\|$ (Problem 5.8).

Now consider a symmetric compact operator $A$ with eigenvalue $\alpha_0$ (as above) and corresponding normalized eigenvector $u_0$. Setting

$$\mathfrak{H}_0^{(1)} = \{f \in \mathfrak{H}_0 | \langle f, u_0\rangle = 0\} \tag{5.35}$$

we can restrict $A$ to $\mathfrak{H}_0^{(1)}$ since $f \in \mathfrak{H}_0^{(1)}$ implies $\langle Af, u_0\rangle = \alpha_0\langle f, u_0\rangle = 0$ and hence $Af \in \mathfrak{H}_0^{(1)}$. Denoting this restriction by $A_1$, it is not hard to see that $A_1$ is again a symmetric compact operator. Hence we can apply Theorem 5.3 iteratively to obtain a sequence of eigenvalues $\alpha_j$ with corresponding normalized eigenvectors $u_j$. Moreover, by construction, $u_n$ is orthogonal to all $u_j$ with $j < n$ and hence the eigenvectors $\{u_j\}$ form an orthonormal set. This procedure will not stop unless $\mathfrak{H}_0$ is finite dimensional. However, note that $\alpha_j = 0$ for $j \geq n$ might happen if $A_n = 0$.

**Theorem 5.4.** *Suppose $\mathfrak{H}_0$ is an inner product space and $A : \mathfrak{H}_0 \to \mathfrak{H}_0$ is a compact symmetric operator. Then there exists a sequence of real eigenvalues $\alpha_j$ converging to 0. The corresponding normalized eigenvectors $u_j$ form an orthonormal set and every $f \in \mathfrak{H}_0$ can be written as*

$$f = \sum_{j=0}^{\infty} \langle u_j, f\rangle u_j + h, \tag{5.36}$$

*where $h$ is in the kernel of $A$, that is, $Ah = 0$.*

*In particular, if 0 is not an eigenvalue, then the eigenvectors form an orthonormal basis.*

**Proof.** Existence of the eigenvalues $\alpha_j$ and the corresponding eigenvectors has already been established. If the eigenvalues should not converge to zero, there is a subsequence such that $v_k = \alpha_{j_k}^{-1} u_{j_k}$ is a bounded sequence for which $Av_k$ has no convergent subsequence since $\|Av_k - Av_l\|^2 = \|u_{j_k} - u_{j_l}\|^2 = 2$.

Next, setting

$$f_n = \sum_{j=0}^{n} \langle u_j, f\rangle u_j, \tag{5.37}$$

we have

$$\|A(f - f_n)\| \le |\alpha_n| \|f - f_n\| \le |\alpha_n| \|f\| \tag{5.38}$$

since $f - f_n \in \mathfrak{H}_0^{(n)}$. Letting $n \to \infty$ shows $A(f_\infty - f) = 0$ proving (5.36). □

Remark: There are two cases where our procedure might fail to construct an orthonormal basis of eigenvectors. One case is where there is an infinite number of nonzero eigenvalues. In this case $\alpha_n$ never reaches 0 and all eigenvectors corresponding to 0 are missed. In the other case, 0 is reached, but there might not be a countable basis and hence again some of the eigenvectors corresponding to 0 are missed. In any case one can show that by adding vectors from the kernel (which are automatically eigenvectors), one can always extend the eigenvectors $u_j$ to an orthonormal basis of eigenvectors.

This is all we need and it remains to apply these results to Sturm-Liouville operators.

**Problem 5.5.** *Prove the* **parallelogram law**

$$\|f + g\|^2 + \|f - g\|^2 = 2\|f\|^2 + 2\|g\|^2.$$

**Problem 5.6.** *Show that (5.30) is indeed a norm. Show that the product of two bounded operators is again bounded.*

**Problem 5.7.** *Show that every compact linear operator is bounded and that the product of a bounded and a compact operator is compact (compact operators form an ideal).*

**Problem 5.8.** *Show that if $A$ is bounded, then every eigenvalue $\alpha$ satisfies $|\alpha| \le \|A\|$.*

## 5.3. Regular Sturm-Liouville problems

Now we want to apply the theory of inner product spaces to the investigation of Sturm-Liouville problem. But first let us look at the corresponding differential equation

$$-(p(x)y')' + (q(x) - z\,r(x))y = 0, \qquad z \in \mathbb{C},\ x \in I = (a,b), \tag{5.39}$$

for $y \in C^2(I, \mathbb{C})$, which is equivalent to the first order system

$$\begin{array}{rcl} y' &=& \frac{1}{p(x)} w \\ w' &=& (q(x) - z\,r(x))y \end{array}, \tag{5.40}$$

where $w(x) = p(x)y'(x)$. Hence we see that there is a unique solution if $p^{-1}(x)$, $q(x)$, and $r(x)$ are continuous in $I$. In fact, as noted earlier, it even suffices to assume that $p^{-1}(x)$, $q(x)$, and $r(x)$ are integrable over each compact subinterval of $I$. I remark that essentially all you have to do is to replace *differentiable* by *absolutely continuous* in the sequel. However, we will assume that

$$r, q \in C^0([a,b], \mathbb{R}),\ p \in C^1([a,b], \mathbb{R}), \quad p(x), r(x) > 0,\ x \in [a,b], \tag{5.41}$$

for the rest of this chapter and call the differential equation (5.39) regular in this case.

Denote by

$$\Pi(z, x, x_0), \qquad z \in \mathbb{C}, \tag{5.42}$$

the principal matrix solution of (5.39). We know that it is continuous with respect to all variables by Theorem 2.7. But with respect to $z$ a much stronger result is true.

**Lemma 5.5.** *The principal matrix solution $\Pi(z, x, x_0)$ is analytic with respect to $z \in \mathbb{C}$.*

**Proof.** It suffices to show that every solution is analytic with respect to $z \in \mathbb{C}$ in a neighborhood of $x_0$ if the initial conditions are analytic. In this case each of the iterations (2.15) is analytic with respect to $z \in \mathbb{C}$. Moreover, for $z$ in a compact set, the Lipschitz constant can be chosen independent of $z$. Hence the series of iterations converges uniformly for $z$ in a compact set, implying that the limit is again analytic by a well-known result from complex analysis. □

Moreover, by Liouville's formula (3.58) the **modified Wronskian**

$$W_x(u, v) = u(x)p(x)v'(x) - p(x)u'(x)v(x) \tag{5.43}$$

is independent of $x$ if $u(x)$ and $v(x)$ both solve (5.39) with the same $z \in \mathbb{C}$.

Now let us look for a suitable scalar product. We consider

$$\langle f, g\rangle = \int_I f(x)^* g(x) r(x) dx, \tag{5.44}$$

and denote $C([a,b],\mathbb{C})$ with this inner product by $\mathfrak{H}_0$.

Next, we want to consider the Sturm-Liouville equation as operator $L$ in $\mathfrak{H}_0$. Since there are function in $\mathfrak{H}_0$ which are not differentiable, we cannot apply it to any function in $\mathfrak{H}_0$. Thus we need a suitable domain

$$\mathfrak{D}(L) = \{f \in C^2([a,b],\mathbb{C}) | BC_a(f) = BC_b(f) = 0\}, \tag{5.45}$$

where

$$\begin{array}{rcl} BC_a(f) & = & \cos(\alpha) f(a) - \sin(\alpha) p(a) f'(a) \\ BC_b(f) & = & \cos(\beta) f(b) - \sin(\beta) p(b) f'(b) \end{array}. \tag{5.46}$$

It is not hard to see that $\mathfrak{D}(L)$ is a dense linear subspace of $\mathfrak{H}_0$. We remark that the case $\alpha = 0$ (i.e., $u(a) = 0$) is called a **Dirichlet boundary condition** at $a$. Similarly, the case $\alpha = \pi/2$ (i.e., $u'(a) = 0$) is called a **Neumann boundary condition** at $a$.

Of course we want $L$ to be symmetric. Using integration by parts it is straightforward to show **Green's formula**

$$\int_I g^* (Lf)\, r dx = W_a(g^*, f) - W_b(g^*, f) + \int_I (Lg)^* f\, r dx \tag{5.47}$$

for $f, g \in C^2([a,b],\mathbb{C})$. Moreover, if $f, g \in \mathfrak{D}(L)$, the above two Wronskians vanish at the boundary and hence

$$\langle g, Lf\rangle = \langle Lg, f\rangle, \qquad f, g \in \mathfrak{D}(L), \tag{5.48}$$

which shows that $L$ is symmetric.

Of course we want to apply Theorem 5.4 and for this we would need to show that $L$ is compact. Unfortunately, it turns out that $L$ is not even bounded (Problem 5.9) and it looks like we are out of luck. However, there is on last chance: the inverse of $L$ might be compact so that we can apply Theorem 5.4 to it.

Since $L$ might not be injective (if 0 is an eigenvalue), we will consider $L - z$ for some fixed $z \in \mathbb{C}$. To compute the inverse of $L - z$ we need to solve the inhomogeneous equation $(L - z)f = g$. This can be easily done by taking two linearly independent solutions $u_+$ and $u_-$ of the homogeneous equation and using the variation of constants formula (3.67). Moreover, in addition to the fact that $f$ is a solution of the differential equation $(L - z)f = g$ it must also be in the domain of $L$, that is, it must satisfy the boundary conditions. Hence we must choose the initial condition in the variation of constants formula such that the boundary conditions are satisfied.

By Problem 3.16 the solutions of the inhomogeneous equation $(L-z)f = g$ can be written as

$$\begin{aligned} f(x) &= \frac{u_+(z,x)}{W(u_+(z),u_-(z))}\Big(c_1 + \int_a^x u_-(z,t)g(t)\,r(t)dt\Big) \\ &\quad + \frac{u_-(z,x)}{W(u_+(z),u_-(z))}\Big(c_2 + \int_x^b u_+(z,t)g(t)\,r(t)dt\Big), \qquad (5.49) \end{aligned}$$

implying

$$\begin{aligned} f'(x) &= \frac{u_+'(z,x)}{W(u_+(z),u_-(z))}\Big(c_1 + \int_a^x u_-(z,t)g(t)\,r(t)dt\Big) \\ &\quad + \frac{u_-'(z,x)}{W(u_+(z),u_-(z))}\Big(c_2 + \int_x^b u_+(z,t)g(t)\,r(t)dt\Big). \qquad (5.50) \end{aligned}$$

Now let us choose $c_1 = 0$, then $f(a) = cu_-(a)$ and $f'(a) = cu_-'(a)$ (where $c = \frac{\langle u_+, g\rangle}{W(u_+,u_-)}$). So choosing $u_-(z,x)$ such that $BC_a(u_-(z)) = 0$, we infer $BC_a(f) = 0$. Similarly, choosing $c_2 = 0$ and $u_+(z,x)$ such that $BC_b(u_+(z)) = 0$, we infer $BC_b(f) = 0$. But can we always do this? Well, setting

$$\begin{array}{ll} u_-(z,a) = \sin(\alpha), & p(a)u_-'(z,a) = \cos(\alpha) \\ u_+(z,b) = \sin(\beta), & p(b)u_+'(z,b) = \cos(\beta) \end{array} \qquad (5.51)$$

we have two solutions of the required type except for the fact that the Wronskian $W(u_+(z),u_-(z))$ might vanish. Now what is so special about the zeros of this Wronskian? Since $W(u_+(z),u_-(z)) = 0$ implies that $u_+(z)$ and $u_-(z)$ are linearly dependent, this implies that $u_+(z,x) = cu_-(z,x)$. Hence $BC_a(u_+(z)) = cBC_a(u_-(z)) = 0$ shows that $z$ is an eigenvalue with corresponding eigenfunction $u_+(z)$. In particular, $z$ must be real, since $L$ is symmetric. Moreover, since $W(u_+(z),u_-(z))$ is analytic in $\mathbb{C}$, the zeros must be discrete.

Let us introduce the operator (the **resolvent** of $L$)

$$R_L(z)g(x) = \int_a^b G(z,x,t)g(t)\,r(t)dt, \qquad (5.52)$$

where

$$G(z,x,t) = \frac{1}{W(u_+(z),u_-(z))}\begin{cases} u_+(z,x)u_-(z,t), & x \ge t \\ u_+(z,t)u_-(z,x), & x \le t \end{cases} \qquad (5.53)$$

is called the **Green function** of $L$. Note that $G(z,x,y)$ is meromorphic with respect to $z \in \mathbb{C}$ with poles precisely at the zeros of $W(u_+(z),u_-(z))$ and satisfies $G(z,x,t)^* = G(z^*,x,t)$ (Problem 5.10) respectively $G(z,x,t) = G(z,t,x)$. Then, by construction we have $R_L(z) : \mathfrak{H}_0 \to \mathfrak{D}(L)$ and

$$(L-z)R_L(z)g = g, \quad R_L(z)(L-z)f = f, \quad g \in \mathfrak{H}_0,\ f \in \mathfrak{D}(L), \qquad (5.54)$$

and hence $R_L(z)$ is the inverse of $L - z$. Our next lemma shows that $R_L(z)$ is compact.

**Lemma 5.6.** *The operator $R_L(z)$ is compact. In addition, for $z \in \mathbb{R}$ it is also symmetric.*

**Proof.** Fix $z$ and note that $G(z,.,..)$ is continuous on $[a,b]\times[a,b]$ and hence uniformly continuous. In particular, for every $\varepsilon > 0$ we can find a $\delta > 0$ such that $|G(z,y,t) - G(z,x,t)| \leq \varepsilon$ whenever $|y - x| \leq \delta$. Let $g(x) = R_L(z)f(x)$, then

$$\begin{aligned} |g(x) - g(y)| &\leq \int_a^b |G(z,y,t) - G(z,x,t)|\,|f(t)|\,r(t)dt \\ &\leq \varepsilon \int_a^b |f(t)|\,r(t)dt \leq \varepsilon \|1\|\,\|f\|, \end{aligned} \tag{5.55}$$

whenever $|y - x| \leq \delta$. Hence, if $f_n(x)$ is a bounded sequence in $\mathfrak{H}_0$, then $g_n(x) = R_L(z)f_n(x)$ is equicontinuous and has a uniformly convergent subsequence by the Arzelà-Ascoli theorem (Theorem 2.13). But a uniformly convergent sequence is also convergent in the norm induced by the scalar product. Therefore $R_L(z)$ is compact.

If $\lambda \in \mathbb{R}$, we have $G(\lambda,t,x)^* = G(\lambda^*,x,t) = G(\lambda,x,t)$ from which symmetry of $R_L(\lambda)$ follows. □

As a consequence we can apply Theorem 5.4 to obtain

**Theorem 5.7.** *The regular Sturm-Liouville problem has a countable number of eigenvalues $E_n$. All eigenvalues are discrete and simple. The corresponding normalized eigenfunctions $u_n$ form an orthonormal basis for $\mathfrak{H}_0$.*

**Proof.** Pick a value $\lambda \in \mathbb{R}$ such that $R_L(\lambda)$ exists. By Theorem 5.4 there are eigenvalues $\alpha_n$ of $R_L(\lambda)$ with corresponding eigenfunctions $u_n$. Moreover, $R_L(\lambda)u_n = \alpha_n u_n$ is equivalent to $Lu_n = (\lambda + \frac{1}{\alpha_n})u_n$, which shows that $E_n = \lambda + \frac{1}{\alpha_n}$ are eigenvalues of $L$ with corresponding eigenfunctions $u_n$. Now everything follows from Theorem 5.4 except that the eigenvalues are simple. To show this, observe that if $u_n$ and $v_n$ are two different eigenfunctions corresponding to $E_n$, then $BC_a(u_n) = BC_a(v_n) = 0$ implies $W_a(u_n,v_n) = 0$ and hence $u_n$ and $v_n$ are linearly dependent. □

It looks like Theorem 5.7 answers all our questions concerning Sturm-Liouville problems. Unfortunately this is not true since the assumptions we have imposed on the coefficients are often too restrictive to be of real practical use! First of all, as noted earlier, it suffices to assume that $r(x)$, $p(x)^{-1}$, $q(x)$ are integrable over $I$. However, this is a minor point. The more important one is, that in most cases at least one of the coefficients will have a (non integrable) singularity at one of the endpoints or the interval might be infinite. For example, the Legendre equation (Problem 4.8) appears on the interval $I = (-1, 1)$, over which $p(x)^{-1} = (1 - x^2)^{-1}$ is not integrable.

In such a situation, the solutions might no longer be extensible to the boundary points and the boundary condition (5.46) makes no sense. However, in this case it is still possible to find two solutions $u_-(z_0, x)$, $u_+(z_0, x)$ (at least for $z_0 \in \mathbb{C}\backslash\mathbb{R}$) which are square integrable near $a$, $b$ and satisfy $\lim_{x\downarrow a} W_x(u_-(z_0)^*, u_-(z_0)) = 0$, $\lim_{x\uparrow b} W_x(u_+(z_0)^*, u_+(z_0)) = 0$, respectively. Introducing the boundary conditions

$$\begin{array}{rcl} BC_a(f) & = & \lim_{x\downarrow a} W_x(u_-(z_0), f) = 0 \\ BC_b(f) & = & \lim_{x\uparrow b} W_x(u_+(z_0), f) = 0 \end{array} \tag{5.56}$$

one obtains again a symmetric operator. The inverse $R_L(z)$ can be computed as before, however, the solutions $u_\pm(z, x)$ might not exist for $z \in \mathbb{R}$ and they might not be holomorphic in the entire complex plane.

It can be shown that Lemma 5.6 (and thus Theorem 5.7) still holds if

$$\int_a^b \int_a^b |G(z, x, y)|^2 r(x) r(y)\, dx\, dy < \infty. \tag{5.57}$$

This can be done for example in the case of Legendre's equation using the explicit behavior of solutions near the singular points $\pm 1$, which follows from the Frobenius method.

However, even for such simple cases as $r(x) = p(x) = 1$, $q(x) = 0$ on $I = \mathbb{R}$, this generalization is still not good enough! In fact, it is not hard to see that there are no eigenfunctions at all in this case. For the investigation of such problems a sound background in measure theory and functional analysis is necessary and hence this is way beyond our scope. I just remark that a similar result holds if the eigenfunction expansion is replaced by an integral transform with respect to a Borel measure. For example, in the case $r(x) = p(x) = 1$, $q(x) = 0$ on $I = \mathbb{R}$ one is lead to the Fourier transform on $\mathbb{R}$.

**Problem 5.9.** *Show directly that $L = -\frac{d^2}{dx^2}$ on $I = (0, \pi)$ with Dirichlet boundary conditions is unbounded. (Hint: Consider $f(x) = \sin(nx)$.)*

**Problem 5.10.** *Show $u_\pm(z, x)^* = u_\pm(z^*, x)$.*

**Problem 5.11** (Periodic boundary conditions)**.** *Show that $L$ defined on*

$$\mathfrak{D}(L) = \{f \in C^2([a, b], \mathbb{C}) | f(a) = f(b), p(a) f'(a) = p(b) f'(b)\} \tag{5.58}$$

*is symmetric.*

**Problem 5.12** (Liouville normal form). *Show that the differential equation (5.39) can be transformed into one with $r = p = 1$ using the transformation*

$$y(x) = \int_a^x \frac{r(t)}{p(t)} dt \qquad v(y) = \sqrt{\tilde{r}(y)\tilde{p}(y)}\, u(x(y)),$$

*where $\tilde{r}(y) = r(x(y))$ and $\tilde{p}(y) = p(x(y))$. Then*

$$-(pu')' + qu = r\lambda u$$

*transforms into*

$$-v'' + Qv = \lambda v,$$

*where*

$$Q = \frac{\tilde{q}}{\tilde{r}^2} + \frac{1}{(\tilde{r}\tilde{p})^2}\left(\frac{1}{2}\tilde{r}\tilde{p}(\tilde{r}\tilde{p})'' - \frac{1}{4}((\tilde{r}\tilde{p})')^2\right).$$

*Moreover,*

$$\int_a^b |u(x)|^2 r(x) dx = \int_0^c |v(y)|^2 dy, \quad c = \int_a^b \frac{r(t)}{p(t)} dt.$$

## 5.4. Oscillation theory

In this section we want to gain further insight by looking at the zeros of the eigenfunctions of a Sturm-Liouville equation.

Let $u$ and $v$ be arbitrary (nonzero) solutions of $Lu = \lambda_0 u$ and $Lv = \lambda v$ for some $\lambda_0, \lambda \in \mathbb{C}$. Then we have

$$W'(u, v) = (\lambda_0 - \lambda) r u v, \tag{5.59}$$

or equivalently for $c, d \in I$

$$W_d(u, v) - W_c(u, v) = (\lambda_0 - \lambda) \int_c^d u(t)v(t)\, r(t) dt. \tag{5.60}$$

This is the key ingredient to the proof of **Sturm's oscillation** theorem.

**Lemma 5.8** (Sturm). *Let $\lambda_0 < \lambda_1$, $(c, d) \subseteq (a, b)$, and $Lu = \lambda_0 u$, $Lv = \lambda_1 v$. Suppose at each end of $(c, d)$ either $W(u, v) = 0$ or $u = 0$. Then $v$ must vanish in $(c, d)$.*

**Proof.** By decreasing $d$ or increasing $c$ to a zero of $u$ (and perhaps flipping signs), we can suppose $u > 0$ on $(c, d)$. If $v$ has no zeros in $(c, d)$, we can suppose $v > 0$ on $(c, d)$ again after perhaps flipping signs. At $c$ either $W(u, v)$ vanishes or else $u(c) = 0$, $v(c) > 0$, and $u'(c) > 0$. Thus, in any case we have $W_c(u, v) \leq 0$. Similarly, $W_d(u, v) \geq 0$. Since the right side of (5.60) is negative, this is inconsistent with (5.60). □

Note that the claim still holds if $\lambda_0 = \lambda_1$ and $W(u,v) \neq 0$ (what happens if $W(u,v) = 0$?).

To gain a better understanding we now introduce **Prüfer variables** defined by

$$u(x) = \rho_u(x)\sin(\theta_u(x)) \qquad p(x)u'(x) = \rho_u(x)\cos(\theta_u(x)). \tag{5.61}$$

If $(u(x), p(x)u'(x))$ is never $(0,0)$ and $u$ is differentiable, then

$$\rho_u(x) = \sqrt{u(x)^2 + (p(x)u'(x))^2} \tag{5.62}$$

is positive and

$$\theta_u(x) = \arctan(\frac{u(x)}{p(x)u'(x)}) = \operatorname{arccot}(\frac{p(x)u'(x)}{u(x)}) \tag{5.63}$$

is uniquely determined once a value of $\theta_u(x_0)$ is chosen by requiring $\theta_u$ to be continuous.

That $u$ satisfies $Lu = \lambda u$ is now equivalent to the system (Problem 5.13)

$$\begin{aligned} \theta_u' &= \frac{\cos^2(\theta_u)}{p} - (q - \lambda r)\sin^2(\theta_u), \\ \rho_u' &= \rho_u\,(\frac{1}{p} + q - \lambda r)\sin(\theta_u)\cos(\theta_u). \end{aligned} \tag{5.64}$$

In addition, notice that

$$W_x(u,v) = \rho_u(x)\rho_v(x)\sin(\theta_u(x) - \theta_v(x)). \tag{5.65}$$

Thus,

**Lemma 5.9.** *Suppose $(u, pu')$ and $(v, pv')$ are never $(0,0)$. Then $u(x_0)$ is zero if and only if $\theta_u(x_0) \equiv 0 \mod \pi$ and $W_{x_0}(u,v)$ is zero if and only if $\theta_u(x_0) \equiv \theta_v(x_0) \mod \pi$.*

In linking Prüfer variables to the number of zeros of $u$, an important role is played by the observation that $\theta_u(x_0) \equiv 0 \mod \pi$ implies

$$\lim_{x\to x_0} \frac{u(x)}{x - x_0} = u'(x_0) \quad \Leftrightarrow \quad \lim_{x\to x_0} \frac{\rho_u(x)\sin(\theta_u(x))}{x - x_0} = \rho_u(x_0)\frac{\cos(\theta_u(x_0))}{p(x_0)} \tag{5.66}$$

and hence we have

$$\lim_{x\to x_0} \frac{\sin(\theta_u(x))}{x - x_0} = \frac{\cos(\theta_u(x_0))}{p(x_0)} \quad \Leftrightarrow \quad \lim_{x\to x_0} \frac{\theta_u(x) - \theta_u(x_0)}{x - x_0} = \frac{1}{p(x_0)}. \tag{5.67}$$

The same result also follows from (5.64), but the present proof does not require that $u$ is a solution of our differential equation.

So we have proven

**Lemma 5.10.** *If $u$ is any $C^1$ function obeying $(u(x), p(x)u'(x)) \neq (0,0)$ on $(a,b)$, then if $\theta_u(x_0) \equiv 0 \mod \pi$,*

$$\lim_{x\to x_0} \frac{\theta_u(x) - \theta_u(x_0)}{x - x_0} = \frac{1}{p(x_0)}. \tag{5.68}$$

In exactly the same way, we have

**Lemma 5.11.** *Let $\lambda_0 < \lambda_1$ and let $u, v$ solve $Lu = \lambda_0 u$, $Lv = \lambda_1 v$. Introduce*

$$\Delta_{u,v}(x) = \theta_v(x) - \theta_u(x). \tag{5.69}$$

*Then, if $\Delta_{u,v}(x_0) \equiv 0 \mod \pi$ but $\theta_u(x_0) \not\equiv 0 \mod \pi$,*

$$\lim_{x\to x_0} \frac{\Delta_{u,v}(x) - \Delta_{u,v}(x_0)}{x - x_0} = (\lambda_1 - \lambda_0) r(x_0) \sin^2 \theta_u(x_0) > 0. \tag{5.70}$$

*And if $\Delta_{u,v}(x_0) \equiv 0 \mod \pi$ but $\theta_u(x_0) \equiv 0 \mod \pi$,*

$$\lim_{x\to x_0} \frac{\Delta_{u,v}(x) - \Delta_{u,v}(x_0)}{(x - x_0)^3} = \frac{(\lambda_1 - \lambda_0) r(x_0)}{3 p(x_0)^2} > 0. \tag{5.71}$$

**Proof.** If $\Delta_{u,v}(x_0) \equiv 0 \mod \pi$ and $\theta_u(x_0) \not\equiv 0 \mod \pi$, then (from (5.65))

$$\lim_{x\to x_0} \frac{\rho_u(x)\rho_v(x) \sin(\Delta_{u,v}(x))}{x - x_0} = -W'_{x_0}(u,v) \tag{5.72}$$

implies the first assertion. If $\Delta_{u,v}(x_0) \equiv 0 \mod \pi$ and $\theta_u(x_0) \equiv \theta_v(x_0) \equiv 0 \mod \pi$, then (using de l'Hospital and again (5.65))

$$\begin{aligned}\lim_{x\to x_0} \frac{\rho_u(x)\rho_v(x) \sin(\Delta_{u,v}(x))}{(x - x_0)^3} &= \lim_{x\to x_0} \frac{-W'_x(u,v)}{3(x-x_0)^2} \\ &= \frac{(\lambda_1 - \lambda_0) r(x_0) \rho_u(x_0) \rho_v(x_0)}{3} \lim_{x\to x_0} \frac{\sin(\theta_u(x)) \sin(\theta_v(x))}{(x - x_0)^2}\end{aligned} \tag{5.73}$$

and the result follows using (5.67). □

Or, put differently, the last two lemmas imply that the integer parts of $\theta_u(x)/\pi$ and $\Delta_{u,v}(x)/\pi$ are increasing.

**Lemma 5.12.** *Let $\lambda_0 < \lambda_1$ and let $u, v$ solve $Lu = \lambda_0 u$, $Lv = \lambda_1 v$. Denote by $\#(u,v)$ the number of zeros of $W(u,v)$ inside the interval $(a,b)$. Then*

$$\#(u,v) = \lim_{x\uparrow b} [\![\Delta_{u,v}(x)/\pi]\!] - \lim_{x\downarrow a} [\![\Delta_{u,v}(x)/\pi]\!], \tag{5.74}$$

*where $[\![x]\!]$ denotes the integer part of a real number $x$, that is, $[\![x]\!] = \sup\{n \in \mathbb{Z} | n \leq x\}$. Moreover, let $\#(u)$ be the number of zeros of $u$ inside $(a,b)$. Then*

$$\#(u) = \lim_{x\uparrow b} [\![\theta_u(x)/\pi]\!] - \lim_{x\downarrow a} [\![\theta_u(x)/\pi]\!]. \tag{5.75}$$

**Proof.** We start with an interval $[x_0, x_1]$ containing no zeros of $W(u, v)$. Hence $[\![\Delta_{u,v}(x_0)/\pi]\!] = [\![\Delta_{u,v}(x_1)/\pi]\!]$. Now let $x_0 \downarrow a$, $x_1 \uparrow b$ and use Lemma 5.9 and Lemma 5.11. The second assertion is proven similar. □

Up to this point $u$ was essentially arbitrary. Now we will take $u(x) = u_\pm(\lambda, x)$, the solutions defined in (5.51), and investigate the dependence of the corresponding Prüfer angle on the parameter $\lambda \in \mathbb{R}$. As a preparation we show

**Lemma 5.13.** *Let $\lambda \in \mathbb{R}$. Then*

$$W_x(u_\pm(\lambda), \dot{u}_\pm(\lambda)) = \left\{ \begin{array}{l} \int_x^b u_+(\lambda, t)^2\, r(t) dt \\ -\int_a^x u_-(\lambda, t)^2\, r(t) dt \end{array} \right. , \tag{5.76}$$

*where the dot denotes a derivative with respect to $\lambda$.*

**Proof.** From (5.60) we know

$$W_x(u_\pm(\lambda), u_\pm(\tilde{\lambda})) = (\tilde{\lambda} - \lambda) \left\{ \begin{array}{l} \int_x^b u_+(\lambda, t) u_+(\tilde{\lambda}, t)\, r(t) dt \\ -\int_a^x u_-(\lambda, t) u_-(\tilde{\lambda}, t)\, r(t) dt \end{array} \right. . \tag{5.77}$$

Now use this to evaluate the limit

$$\lim_{\tilde{\lambda} \to \lambda} W_x\Big(u_\pm(\lambda), \frac{u_\pm(\lambda) - u_\pm(\tilde{\lambda})}{\lambda - \tilde{\lambda}}\Big). \tag{5.78}$$

□

Now, since

$$\dot{\theta}_u(x) = -\frac{W_x(u, \dot{u})}{\rho_u(x)^2}, \tag{5.79}$$

equation (5.76) immediately implies

$$\dot{\theta}_+(\lambda, x) = -\frac{\int_x^b u_+(\lambda, t)^2\, r(t) dt}{\rho_+(\lambda, x)^2} < 0, \quad \dot{\theta}_-(\lambda, x) = \frac{\int_a^x u_-(\lambda, t)^2\, r(t) dt}{\rho_-(\lambda, x)^2} > 0, \tag{5.80}$$

where we have abbreviated $\rho_\pm(\lambda, x) = \rho_{u_\pm(\lambda)}(x)$ and $\theta_\pm(\lambda, x) = \theta_{u_\pm(\lambda)}(x)$. Next let us choose

$$\theta_-(\lambda, a) = \alpha \in [0, \pi), \quad -\theta_+(\lambda, b) = \beta \in [0, \pi) \tag{5.81}$$

and since $\pm\theta_\pm(., x) \geq 0$ is decreasing, the limit

$$\mp\theta_\pm(x) = \mp \lim_{\lambda \downarrow -\infty} \theta_\pm(\lambda, x) \geq 0 \tag{5.82}$$

exists. In fact, the following lemma holds.

**Lemma 5.14.** *We have*

$$\theta_+(x) = 0,\ x \in [a, b), \qquad \theta_-(x) = 0,\ x \in (a, b]. \tag{5.83}$$

**Proof.** We only do the proof for $\theta_-(x)$. Fix $x_0 \in (a,b]$ and consider $w(x) = \pi - (\pi - \varepsilon)\frac{x-a}{x_0-a}$ for $\varepsilon > 0$ small. Then, for sufficiently small $\lambda$, we have

$$\frac{1}{p}\cos^2(w) - (q-\lambda)\sin^2(w) \le \frac{1}{p} - (q-\lambda)\sin^2(\varepsilon) < w' \tag{5.84}$$

for $x \in [a, x_0]$ which shows that $w$ is a super solution (compare page 18). Hence $0 \le \theta_-(x_0) \le \varepsilon$ for any $\varepsilon$. $\square$

Now observe that $u_-(\lambda)$ is an eigenfunction if and only if it satisfies the boundary condition at $b$, that is, if and only if $\theta_-(\lambda, b) = \beta \mod \pi$. This shows that $u_-(\lambda)$ can eventually no longer satisfy the boundary condition at $b$ as $\lambda \to -\infty$. Hence there is a lowest eigenvalue $E_0$ and we note

**Lemma 5.15.** *The eigenvalues of a regular Sturm-Liouville problem can be ordered according to $E_0 < E_1 < \cdots$.*

After these preparations we can now easily establish several beautiful and important results.

**Theorem 5.16.** *Suppose $L$ has a Dirichlet boundary condition at $b$ (i.e., $u(b) = 0$). Then we have*

$$\#_{(-\infty,\lambda)}(L) = \#(u_-(\lambda)), \tag{5.85}$$

*where $\#(u)$ is the number of zeros of $u$ inside $(a,b)$ and $\#_{(\lambda_0,\lambda_1)}(L)$ is the number of eigenvalues of $L$ inside $(\lambda_0, \lambda_1)$. Likewise, suppose $L$ has a Dirichlet boundary condition at $a$. Then we have*

$$\#_{(-\infty,\lambda)}(L) = \#(u_+(\lambda)). \tag{5.86}$$

**Proof.** For $\lambda$ small, $u_-(\lambda)$ has no zeros by Lemma 5.14. Hence the result holds for small $\lambda$. As $\lambda$ increases, $\theta_-(\lambda, b)$ increases and is $0 \mod \pi$ if and only if $\lambda$ is an eigenvalue of $L$ (Lemma 5.9) completing the proof. $\square$

The same proof together with Sturm's result (Lemma 5.8) shows

**Theorem 5.17.** *Suppose the eigenvalues are ordered according to $E_0 < E_1 < \cdots$. Then the eigenfunction $u_n$ corresponding to $E_n$ has precisely $n$ zeros in the interval $(a,b)$ and the zeros of $u_{n+1}$ interlace the zeros of $u_n$. That is, if $x_{n,j}$ are the zeros of $u_n$ inside $(a,b)$, then*

$$a < x_{n+1,1} < x_{n,1} < x_{n+1,2} < \cdots < x_{n+1,n+1} < b. \tag{5.87}$$

In precisely the same way one proves

**Theorem 5.18.** *We have for $\lambda_0 < \lambda_1$*

$$\#_{(\lambda_0,\lambda_1)}(L) = \#(u_-(\lambda_0), u_+(\lambda_1)) = \#(u_+(\lambda_0), u_-(\lambda_1)), \tag{5.88}$$

*where $\#(u,v)$ is the number of zeros of $W(u,v)$ inside $(a,b)$ and $\#_{(\lambda_0,\lambda_1)}(L)$ is the number of eigenvalues of $L$ inside $(\lambda_0, \lambda_1)$.*

**Proof.** We only carry out the proof for the $\#(u_-(\lambda_0), u_+(\lambda_1))$ case. Abbreviate $\Delta(\lambda_1, x) = \Delta_{u_-(\lambda_0), u_+(\lambda_1)}(x)$. Since the Wronskian is constant for $\lambda_1 = \lambda_0$, our claim holds for $\lambda_1$ close to $\lambda_0$. Moreover, since $\Delta(\lambda_1, b) = \beta - \Theta_-(\lambda_0, b)$ is independent of $\lambda_1$, it suffices to look at $\Delta(\lambda_1, a)$ by Lemma 5.12. As $\lambda_1 \geq \lambda_0$ increases, $-\Delta(\lambda_1, a)$ increases by (5.80) and is $0 \mod \pi$ if and only if $\lambda_1$ is an eigenvalue of $L$ (Lemma 5.9) completing the proof. □

**Problem 5.13.** *Prove equation (5.64).*

**Problem 5.14.** *Suppose that $q(x) > 0$ and let $-(pu')' + qu = 0$. Show that at two consecutive zeros $x_k$ and $x_{k+1}$ of $u'(x)$ we have*

$$|u(x_k)| \leq |u(x_{k+1})| \qquad \text{if} \quad (pq)' \geq 0.$$

*Hint: consider*

$$u^2 - \frac{1}{pq}(pu')^2.$$

**Problem 5.15.** *Consider the ordered eigenvalues $E_n(\alpha)$ of our Sturm-Liouville problem as a function of the boundary parameter $\alpha$. Show that the eigenvalues corresponding to different parameters are interlacing. That is, suppose $0 < \alpha_1 < \alpha_2 \leq \pi$ and show $E_n(\alpha_1) < E_n(\alpha_2) < E_{n+1}(\alpha_1)$.*

*Part 2*

# Dynamical systems

Chapter 6

# Dynamical systems

## 6.1. Dynamical systems

You can think of a dynamical system as the time evolution of some physical system, like the motion of a few planets under the influence of their respective gravitational forces. Usually you want to know the fate of system for long times, like, will the planets eventually collide or will the system persist for all times? For some systems (e.g., just two planets) these questions are relatively simple to answer since it turns out that the motion of the system is regular and converges (e.g.) to an equilibrium.

However, many interesting systems are not that regular! In fact, it turns out that for many systems even very close initial conditions might get spread far apart in short times. For example, you probably have heard about the motion of a butterfly which can produce a perturbance of the atmosphere resulting in a thunderstorm a few weeks later.

A **dynamical system** is a semigroup $G$ acting on a space $M$. That is, there is a map

$$\begin{array}{rccc} T: & G \times M & \to & M \\ & (g, x) & \mapsto & T_g(x) \end{array} \tag{6.1}$$

such that

$$T_g \circ T_h = T_{g \circ h}. \tag{6.2}$$

If $G$ is a group, we will speak of an **invertible dynamical system**.

We are mainly interested in **discrete dynamical systems** where

$$G = \mathbb{N}_0 \quad \text{or} \quad G = \mathbb{Z} \tag{6.3}$$

and in **continuous dynamical systems** where

$$G = \mathbb{R}^+ \quad \text{or} \quad G = \mathbb{R}. \tag{6.4}$$

Of course this definition is quite abstract and so let us look at some examples first.

The prototypical example of a discrete dynamical system is an iterated map. Let $f$ map an interval $I$ into itself and consider

$$T^n = f^n = f \circ f^{n-1} = f \circ \cdots \circ f, \qquad G = \mathbb{N}_0. \tag{6.5}$$

Clearly, if $f$ is invertible, so is the dynamical system if we extend this definition for $n = \mathbb{Z}$ in the usual way. You might suspect that such a system is too simple to be of any interest. However, we will see that the contrary is the case and that such simple system bear a rich mathematical structure with lots of unresolved problems.

The prototypical example of a continuous dynamical system is the flow of an autonomous differential equation

$$T_t = \Phi_t, \qquad G = \mathbb{R}, \tag{6.6}$$

which we will consider in the following section.

## 6.2. The flow of an autonomous equation

Now we will have a closer look at the solutions of an autonomous system

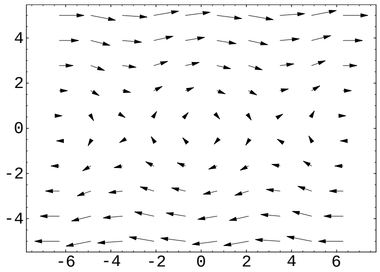
$$\dot{x} = f(x), \qquad x(0) = x_0. \tag{6.7}$$

Throughout this section we will assume $f \in C^k(M, \mathbb{R}^n)$, $k \geq 1$, where $M$ is an open subset of $\mathbb{R}^n$.

Such a system can be regarded as a **vector field** on $\mathbb{R}^n$. Solutions are curves in $M \subseteq \mathbb{R}^n$ which are tangent to this vector field at each point. Hence to get a geometric idea of how the solutions look like, we can simply plot the corresponding vector field.

This can be easily done using *Mathematica*. For example, the vector field of the mathematical pendulum, $f(x, y) = (y, -\sin(x))$, can be plotted as follows.

```
In[1]:= Needs["Graphics`PlotField`"];

In[2]:= PlotVectorField[{y, −Sin[x]}, {x, −2π, 2π}, {y, −5, 5},
          Frame → True, PlotPoints → 10];
```




We will return to this example in Section 6.6.

In particular, solutions of the IVP (6.7) are also called **integral curves** or **trajectories**. We will say that $\phi$ is an integral curve at $x_0$ if it satisfies $\phi(0) = x_0$.

As in the previous chapter, there is a (unique) **maximal integral curve** $\phi_x$ at every point $x$, defined on a maximal interval $I_x = (T_-(x), T_+(x))$.

Introducing the set

$$W = \bigcup_{x \in M} I_x \times \{x\} \subseteq \mathbb{R} \times M \tag{6.8}$$

we define the **flow** of our differential equation to be the map

$$\Phi : W \to M, \qquad (t,x) \mapsto \phi(t,x), \tag{6.9}$$

where $\phi(t,x)$ is the maximal integral curve at $x$. We will sometimes also use $\Phi_x(t) = \Phi(t,x)$ and $\Phi_t(x) = \Phi(t,x)$.

If $\phi(.)$ is an integral curve at $x$, then $\phi(.+s)$ is an integral curve at $y = \phi(s)$. This defines a bijection between integral curves at $x$ and $y$ respectively. Furthermore, it maps maximal integral curves to maximal integral curves and we hence infer $I_x = s + I_y$. As a consequence, we note that for $x \in M$ and $s \in I_x$ we have

$$\Phi(s+t,x) = \Phi(t, \Phi(s,x)) \tag{6.10}$$

for all $t \in I_{\Phi(s,x)} = I_x - s$. In particular, choosing $t = -s$ shows that $\Phi_s(.) = \Phi(s,.)$ is a local diffeomorphism with inverse $\Phi_{-s}(.)$.

Our next goal is to show that $W$ is open and $\Phi \in C^k(W,M)$. Fix a point $(t_0, x_0) \in W$ (implying $t_0 \in I_{x_0}$) and set $\gamma = \Phi_{x_0}([0,t_0])$. By Theorem 2.7 there is an open neighborhood $(-\varepsilon(x), \varepsilon(x)) \times U(x)$ of $(0,x)$ around each point $x \in \gamma$ such that $\Phi$ is defined and $C^k$ on this neighborhood. Since $\gamma$ is compact, finitely many of this neighborhoods cover $\{0\} \times \gamma$ and hence we can find an $\varepsilon > 0$ and an open neighborhood $U$ of $\gamma$ such that $\Phi$ is defined on $(-\varepsilon, \varepsilon) \times U$. Next, pick $m \in \mathbb{N}$ so large that $\frac{t_0}{m} < \varepsilon$ and let $K^j(x) = K(K^{j-1}(x))$, where $K(x) = \Phi_{\frac{t_0}{m}}(x)$ is $C^k$ for $x \in U$ by construction. Since $K^j(x_0) \in \gamma \subset U$ for $1 \le j \le m$, there is an open neighborhood $U_0 \subseteq U$ of

$x_0$ such that $K^m$ is defined on $U_0$. Moreover,

$$\Phi(t,x) = \Phi(t-t_0, \Phi(t_0,x)) = \Phi(t-t_0, K^m(x)) \tag{6.11}$$

is defined and smooth for all $(t,x) \in (t_0+\varepsilon, t_0-\varepsilon) \times U_0$.

In summary, we have proven the following result.

**Theorem 6.1.** *Suppose $f \in C^k$. For all $x \in M$ there exists an interval $I_x \subseteq \mathbb{R}$ containing $0$ and a corresponding unique maximal integral curve $\Phi(.,x) \in C^k(I_x, M)$ at $x$. Moreover, the set $W$ defined in (6.8) is open and $\Phi \in C^k(W, M)$ is a (local) flow on $M$, that is,*

$$\begin{aligned} \Phi(0,x) &= x, \\ \Phi(s+t,x) &= \Phi(t, \Phi(s,x)), \quad x \in M,\ s, t+s \in I_x. \end{aligned} \tag{6.12}$$

Now look at an example illustrating our findings. Let $M = \mathbb{R}$ and $f(x) = x^3$. Then $W = \{(t,x) | 2tx^2 < 1\}$ and $\Phi(t,x) = \frac{x}{\sqrt{1-2x^2t}}$. $T_-(x) = -\infty$ and $T_+(x) = 1/(2x^2)$.

Note that if we replace $f \to -f$ we have to set $\Phi(t,x) \to \Phi(-t,x)$.

Finally, I remark that away from singular points, all vector fields look locally the same.

**Lemma 6.2** (Straightening out of vector fields)**.** *Suppose $f(x_0) \neq 0$. Then there is a local coordinate transform $y = \varphi(x)$ such that $\dot{x} = f(x)$ is transformed to*

$$\dot{y} = (1, 0, \dots, 0). \tag{6.13}$$

**Proof.** It is no restriction to assume $x_0 = 0$. After a linear transformation we see that it is also no restriction to assume $f(0) = (1, 0, \dots, 0)$.

Consider all points starting on the plane $x_1 = 0$. Then the point $\Phi(t, (0, x_2, \dots, x_n))$ should be mapped to the point $(0, x_2, \dots, x_n) + t(1, 0, \dots, 0) = (t, x_2, \dots, x_n)$. Hence the inverse of the map we are looking for should be given by

$$\psi(x) = \Phi(x_1, (0, x_2, \dots, x_n)), \tag{6.14}$$

which is well defined in a neighborhood of 0. The Jacobi determinant at 0 is given by

$$\det \frac{\partial \psi_i}{\partial x_j}\Big|_{x=0} = \det \Big( \frac{\partial \Phi}{\partial t}, \frac{\partial \Phi}{\partial x_2}, \dots, \frac{\partial \Phi}{\partial x_n} \Big)\Big|_{t=0,x=0} = \det \mathbb{I}_n = 1 \tag{6.15}$$

since $\partial \Phi / \partial x|_{t=0,x=0} = \mathbb{I}_n$ and $\partial \Phi / \partial t|_{t=0,x=0} = f(0) = (1, 0, \dots, 0)$ by assumption. So by the inverse function theorem we can assume that $\psi$ is a local diffeomorphism and we can consider new coordinates $y = \psi^{-1}(x)$. Since $\partial \psi_j / \partial x_1 = f_j(\psi(x))$ our system reads in the new coordinates

$$\dot{y}_j = \Big( \frac{\partial \psi_j}{\partial x_i} \Big)^{-1}_{\psi^{-1}(x)} f_i(x) = \delta_{1,j}, \tag{6.16}$$

which is the required form. □

**Problem 6.1.** *Compute the flow for $f(x) = x^2$ defined on $M = \mathbb{R}$.*

**Problem 6.2.** *Find a transformation which straightens out the flow $\dot{x} = x$ defined on $M = \mathbb{R}$.*

**Problem 6.3.** *Show that $\Phi(t, x) = \mathrm{e}^t(1 + x) - 1$ is a flow (i.e., it satisfies (6.12)). Can you find an autonomous system corresponding to this flow?*

**Problem 6.4.** *Suppose $\Phi(t, x)$ is differentiable and satisfies (6.12). Show that $\Phi$ is the flow of the vector field*

$$f(x) = \dot{\Phi}(0, x).$$

## 6.3. Orbits and invariant sets

The **orbit** of $x$ is defined as

$$\gamma(x) = \Phi(I_x, x) \subseteq M. \tag{6.17}$$

Note that $y \in \gamma(x)$ implies $y = \Phi(t, x)$ and hence $\gamma(x) = \gamma(y)$ by (6.12). In particular, different orbits are disjoint (i.e., we have the following equivalence relation on $M$: $x \simeq y$ if $\gamma(x) = \gamma(y)$). If $\gamma(x) = \{x\}$, then $x$ is called a **fixed point** (also **singular**, **stationary**, or **equilibrium point**) of $\Phi$. Otherwise $x$ is called **regular** and $\Phi(., x) : I_x \hookrightarrow M$ is an immersion.

Similarly we introduce the **forward** and **backward orbits**

$$\gamma_\pm(x) = \Phi((0, T_\pm(x)), x). \tag{6.18}$$

Clearly $\gamma(x) = \gamma_-(x) \cup \{x\} \cup \gamma_+(x)$. One says that $x \in M$ is a **periodic point** of $\Phi$ if there is some $T > 0$ such that $\Phi(T, x) = x$. The lower bound of such $T$ is called the **period**, $T(x)$ of $x$, that is, $T(x) = \inf\{T > 0 | \Phi(T, x) = x\}$. By continuity of $\Phi$ we have $\Phi(T(x), x) = x$ and by the flow property $\Phi(t + T(x), x) = \Phi(t, x)$. In particular, an orbit is called **periodic orbit** if one (and hence all) point of the orbit is periodic.

It is not hard to see (Problem 6.7) that $x$ is periodic if and only if $\gamma_+(x) \cap \gamma_-(x) \neq \emptyset$ and hence periodic orbits are also called **closed orbits**.

Hence we may classify the orbits of $f$ as follows:

(i) **fixed orbits** (corresponding to a periodic point with period zero)

(ii) **regular periodic orbits** (corresponding to a periodic point with positive period)

(iii) **non-closed orbits** (not corresponding to a periodic point)

The quantity $T_+(x) = \sup I_x$ (resp. $T_-(x) = \inf I_x$) defined in the previous section is called positive (resp. negative) **lifetime** of $x$. A point

$x \in M$ is called $\sigma$ complete, $\sigma \in \{\pm\}$, if $T_\sigma(x) = \sigma\infty$ and complete if it is both $+$ and $-$ complete (i.e., if $I_x = \mathbb{R}$).

Lemma 2.10 gives us a useful criterion when a point $x \in M$ is $\sigma$ complete.

**Lemma 6.3.** *Let $x \in M$ and suppose that the forward (resp. backward) orbit lies in a compact subset $C$ of $M$. Then $x$ is $+$ (resp. $-$) complete.*

Clearly a periodic point is complete. If all points are complete, the vector field is called **complete**. Thus $f$ complete means that $\Phi$ is globally defined, that is, $W = \mathbb{R} \times M$.

A set $U \subseteq M$ is called $\sigma$ **invariant**, $\sigma \in \{\pm\}$, if

$$\gamma_\sigma(x) \subseteq U, \qquad \forall x \in U, \tag{6.19}$$

and invariant if it is both $\pm$ invariant, that is, if $\gamma(x) \subseteq U$. If $U$ is $\sigma$ invariant, the same is true for $\overline{U}$, the closure of $U$. In fact, $x \in \overline{U}$ implies the existence of a sequence $x_n \in U$ with $x_n \to x$. Fix $t \in I_x$. Then (since $W$ is open) for $N$ sufficiently large we have $t_n \in I_{x_n}$, $n \geq N$ and $\Phi(t,x) = \lim_{n\to\infty} \Phi(t_n, x_n) \in \overline{U}$.

Clearly, arbitrary intersections and unions of $\sigma$ invariant sets are $\sigma$ invariant. Moreover, the closure of a $\sigma$ invariant set is again $\sigma$ invariant.

If $C \subseteq M$ is a compact $\sigma$ invariant subspace, then Lemma 6.3 implies that all points in $C$ are $\sigma$ complete.

A nonempty, compact, $\sigma$ invariant set is called **minimal** if it contains no proper $\sigma$ invariant subset possessing these three properties.

**Lemma 6.4.** *Every nonempty, compact ($\sigma$) invariant set $C \subseteq M$ contains a minimal ($\sigma$) invariant set.*

**Proof.** Consider the family $\mathcal{F}$ of all compact ($\sigma$) invariant subsets of $C$. Every nest in $\mathcal{F}$ has a minimal member by the finite intersection property of compact sets. So by the minimal principle there exists a minimal member of $\mathcal{F}$. □

**Lemma 6.5.** *Every $\sigma$ invariant set $C \subseteq M$ homeomorphic to an $m$-dimensional disc (where $m$ is not necessarily the dimension of $M$) contains a singular point.*

**Proof.** We only prove the case $\sigma = +$. Pick a sequence $T_j \downarrow 0$. By Brouwer's theorem $\Phi(T_j, .) : C \to C$ has a fixed point $x_j$. Since $C$ is compact we can assume $x_j \to x$ after maybe passing to a subsequence. Fix $t > 0$ and pick $n_j \in \mathbb{N}_0$ such that $0 \leq t - n_j T_j < T_j$. Then

$$\Phi(t,x) = \lim_{j\to\infty} \Phi(n_j T_j, x_j) = \lim_{j\to\infty} x_j = x \tag{6.20}$$

and $x$ is fixed. □

The **$\omega_\pm$-limit set** of a point $x \in M$, $\omega_\pm(x)$ is the set of those points $y \in M$ for which there exists a sequence $t_n \to \pm\infty$ with $\Phi(t_n, x) \to y$.

Clearly, $\omega_\pm(x)$ is empty unless $x$ is $\pm$ complete. Observe, that $\omega_\pm(x) = \omega_\pm(y)$ if $y \in \gamma(x)$ (if $y = \Phi(t, x)$ we have $\Phi(t_n, y) = \Phi(t_n, \Phi(t, x)) = \Phi(t_n + t, x)$). Moreover, $\omega_\pm(x)$ is closed. Indeed, if $y \notin \omega_\pm(x)$ there is a neighborhood $U$ of $y$ disjoint from $\Phi(\{t \in I_x | t > T\}, x)$ for some $T > 0$. Hence the complement of $\omega_\pm(x)$ is open.

The set $\omega_\pm(x)$ is invariant since if $\Phi(t_n, x) \to y$ we have

$$\Phi(t_n + t, x) = \Phi(t, \Phi(t_n, x)) \to \Phi(t, y) \tag{6.21}$$

for $|t_n|$ large enough since $x$ is $\pm$ complete.

In summary,

**Lemma 6.6.** *The set $\omega_\pm(x)$ is a closed invariant set.*

In some situations we can say even more.

**Lemma 6.7.** *If $\gamma_\sigma(x)$ is contained in a compact set $C$, then $\omega_\sigma(x)$ is nonempty, compact, and connected.*

**Proof.** We only work out the proof for $\sigma = +$. By Lemma 6.3, $x$ is $\sigma$ complete. Hence $\omega_\sigma(x)$ is nonempty and compact. If $\omega_\sigma(x)$ is disconnected, we can split it up into two closed sets $\omega_{1,2}$ which are also closed in $M$. Since $M$ is normal, we can find two disjoint neighborhoods $U_{1,2}$ of $\omega_{1,2}$, respectively. Now choose a strictly increasing sequence $t_n \to \infty$ such that $\Phi(t_{2m+1}, x) \in U_1$ and $\Phi(t_{2m}, x) \in U_2$. By connectedness of $\Phi((t_{2m}, t_{2m+1}), x)$ we can find $\Phi(\tilde{t}_m, x) \in C \backslash (U_1 \cup U_2)$ with $t_{2m} < \tilde{t}_m < t_{2m+1}$. Since $C \backslash (U_1 \cup U_2)$ is compact, we can assume $\Phi(\tilde{t}_m, x) \to y \in C \backslash (U_1 \cup U_2)$. But $y$ must also be in $\omega_\sigma(x)$, a contradiction. □

Now let us consider an example which shows that the compactness requirement is indeed necessary. Let $M = \mathbb{R}^2$ and consider the vector field

$$f(x) = \begin{pmatrix} \cos^2(x_1)(\sin(x_1) - x_2 \cos(x_1)) \\ \sin(x_1) + x_2 \cos(x_1) \end{pmatrix}, \tag{6.22}$$

Since $f$ is bounded it is complete by Theorem 2.12. The singularities are given by $(\mathbb{Z}\pi, 0)$. One further verifies that for $x \in (-\pi/2, \pi/2) \times \mathbb{R}$ we have

$$\Phi(t, x) = \begin{pmatrix} \arctan(r\mathrm{e}^{\tau(t)} \cos(\tau(t) + \theta)) \\ r\mathrm{e}^{\tau(t)} \sin(\tau(t) + \theta) \end{pmatrix}, \tag{6.23}$$

where $(r, \theta)$ are the polar coordinates of $(\tan(x_1), x_2)$ and

$$\dot{\tau}(t) = \frac{1}{\sqrt{1 + r^2 \mathrm{e}^{2\tau(t)} \cos^2(\tau(t))}}, \quad \tau(0) = 0. \tag{6.24}$$

Clearly, $\tau \in C^\infty(\mathbb{R}, \mathbb{R})$ is a diffeomorphism and hence $\omega_-(x) = (0,0)$ and $\omega_+(x) = \{\pm\pi\} \times \mathbb{R}$ if $x \neq (0,0)$. Moreover,

$$\Phi(t, (\pm\frac{\pi}{2}, x_2)) = \begin{pmatrix} \pm\frac{\pi}{2} \\ x_2 \pm t \end{pmatrix} \tag{6.25}$$

and hence $\omega_-(\pm\frac{\pi}{2}, 0) = \omega_+(\pm\frac{\pi}{2}, 0) = \emptyset$.

Thus far $\Phi$ is only given for $x \in [-\frac{\pi}{2}, \frac{\pi}{2}] \times \mathbb{R}$. The remaining parts of the plane can be investigated using the transformation $(t, x_1, x_2) \to (-t, x_1 \pm \pi, x_2)$.

We end this section with an important lemma. Recall that a set $\Sigma \subset \mathbb{R}^n$ is called a **submanifold** of codimension one (i.e., its dimension is $n-1$), if it can be written as

$$\Sigma = \{x \in U | S(x) = 0\}, \tag{6.26}$$

where $U \subset \mathbb{R}^n$ is open, $S \in C^k(U)$, and $\partial S/\partial x \neq 0$ for all $x \in \Sigma$. The submanifold $\Sigma$ is said to be transversal to the vector field $f$ if $(\partial S/\partial x) f(x) \neq 0$ for all $x \in \Sigma$.

**Lemma 6.8.** *Suppose $x \in M$ and $T \in I_x$. Let $\Sigma$ be submanifold of codimension one transversal to $f$ such that $\Phi(T, x) \in \Sigma$. Then there exists a neighborhood $U$ of $x$ and $\tau \in C^k(U)$ such that $\tau(x) = T$ and*

$$\Phi(\tau(y), y) \in \Sigma \tag{6.27}$$

*for all $y \in U$.*

**Proof.** Consider the equation $S(\Phi(t, y)) = 0$ which holds for $(T, x)$. Since

$$\frac{\partial}{\partial t} S(\Phi(t, y)) = \frac{\partial S}{\partial x}(\Phi(t, y)) f(\Phi(t, y)) \neq 0 \tag{6.28}$$

for $(t, y)$ in a neighborhood $I \times U$ of $(T, x)$ by transversality. So by the implicit function theorem (maybe after restricting $U$), there exists a function $\tau \in C^k(U)$ such that for all $y \in U$ we have $S(\Phi(\tau(y), y)) = 0$, that is, $\Phi(\tau(y), y) \in \Sigma$. □

If $x$ is periodic and $T = T(x)$, then

$$P_\Sigma(y) = \Phi(\tau(y), y) \tag{6.29}$$

is called **Poincaré map**.

**Problem 6.5.** *Consider a first order autonomous system in $\mathbb{R}^1$. Suppose $f(x)$ is differentiable, $f(0) = f(1) = 0$, and $f(x) > 0$ for $x \in (0,1)$. Determine the orbit $\gamma(x)$ and $\omega_\pm(x)$ if $x \in [0,1]$.*

**Problem 6.6.** *Let $\phi(t)$ be the solution of a first order autonomous system. Suppose $\lim_{t\to\infty} \phi(t) = x \in M$. Show that $x$ is a singular point.*

**Problem 6.7** (Periodic points)**.** *Let $\Phi$ be the flow of some differential equation.*

(i) *Show that if $T$ satisfies $\Phi(T, x) = x$, the same is true for any integer multiple of $T$. Moreover, show that we must have $T = nT(x)$ for some $n \in \mathbb{Z}$ if $T(x) \neq 0$.*

(ii) *Show that a point $x$ is stationary if and only if $T(x) = 0$.*

(iii) *Show that $x$ is periodic if and only if $\gamma_+(x) \cap \gamma_-(x) \neq \emptyset$ in which case $\gamma_+(x) = \gamma_-(x)$ and $\Phi(t + T(x), x) = \Phi(t, x)$ for all $t \in \mathbb{R}$. In particular, the period is the same for all points in the same orbit.*

**Problem 6.8.** *A point $x \in M$ is called* **nonwandering** *if for every neighborhood $U$ of $x$ there is a sequence of positive times $t_n \to \infty$ such that $\Phi_{t_n}(U) \cap U \neq \emptyset$ for all $t_n$. The* **set of nonwandering points** *is denoted by $\Omega(f)$.*

(i) *$\Omega(f)$ is a closed invariant set (Hint: show that it is the complement of an open set).*

(ii) *$\Omega(f)$ contains all periodic orbits (including all fixed points).*

(iii) *$\omega_+(x) \subseteq \Omega(f)$ for all $x \in M$.*

*Find the set of nonwandering points $\Omega(f)$ for the system $f(x, y) = (y, -x)$.*

**Problem 6.9.** *Which of the following equations determine a submanifold of codimension one of $\mathbb{R}^2$?*

(i) $x = 0$.

(ii) $x^2 + y^2 = 1$.

(iii) $x^2 - y^2 = 1$.

(iv) $x^2 + y^2 = 0$.

*Which of them is transversal to $f(x, y) = (x, -y)$, $f(x, y) = (1, 0)$, or $f(x, y) = (0, 1)$, respectively.*

## 6.4. Stability of fixed points

As already mentioned earlier, one of the key questions is the long time behavior of the dynamical system (6.7). In particular, one often wants to know whether the solution is *stable* or not. But first we need to define what we mean by stability. Usually one looks at a fixed point and wants to know what happens if one starts close to it. Hence we define the following.

A fixed point $x_0$ of $f(x)$ is called **stable** if for any given neighborhood $U(x_0)$ there exists another neighborhood $V(x_0) \subseteq U(x_0)$ such that any solution starting in $V(x_0)$ remains in $U(x_0)$ for all $t \geq 0$.

Similarly, a fixed point $x_0$ of $f(x)$ is called **asymptotically stable** if it is stable and if there is a neighborhood $U(x_0)$ such that

$$\lim_{t\to\infty} |\phi(t,x) - x_0| = 0 \quad \text{for all } x \in U(x_0). \tag{6.30}$$

For example, consider $\dot{x} = ax$ in $\mathbb{R}^1$. Then $x_0 = 0$ is stable if and only if $a \le 0$ and asymptotically stable if and only if $a < 0$. More generally, suppose the equation $\dot{x} = f(x)$ in $\mathbb{R}^1$ has a fixed point $x_0$. Then it is not hard to see (by looking at the solution found in Section 1.3) that $x_0$ is stable if

$$\frac{f(x) - f(x_0)}{x - x_0} \le 0, \qquad x \in U(x_0)\backslash\{x_0\} \tag{6.31}$$

for some neighborhood $U(x_0)$ and asymptotically stable if strict inequality holds. In particular, if $f'(x_0) \ne 0$ the stability can be read of from the derivative of $f$ at $x_0$ alone. However, if $f'(x_0) = 0$, no information on the stability of the nonlinear system can be read off from the linear one as can be seen from the example

$$f(x) = \mu x^3. \tag{6.32}$$

In $\mathbb{R}^n$, $n > 1$, the equation cannot be solved explicitly in general, and good criteria for stability are needed. This will be the topic of the remainder of this chapter.

But before that, let me point out that it is also interesting to look at the change of a differential equation with respect to a parameter $\mu$. By Theorem 2.8 the flow depends smoothly on the parameter $\mu$ (if $f$ does). Nevertheless very small changes in the parameters can produce large changes in the qualitative behavior of solutions. The systematic study of these phenomena is known as **bifurcation theory**. I do not want to go into further details at this point but I will rather show you some prototypical examples.

The system

$$\dot{x} = \mu x - x^3 \tag{6.33}$$

has one stable fixed point for $\mu \le 0$ which becomes unstable and splits off two stable fixed points at $\mu = 0$. This is known as **pitchfork bifurcation**. The system

$$\dot{x} = \mu x - x^2 \tag{6.34}$$

has two stable fixed point for $\mu \ne 0$ which collide and exchange stability at $\mu = 0$. This is known as **transcritical bifurcation**. The system

$$\dot{x} = \mu + x^2 \tag{6.35}$$

has two stable fixed point for $\mu < 0$ which collide at $\mu = 0$ and vanish. This is known as **saddle-node bifurcation**.

Observe that by the implicit function theorem, the number of fixed points can locally only change at a point $(x_0, \mu_0)$ if $f(x_0, \mu_0) = 0$ and $\frac{\partial f}{\partial x}(x_0, \mu_0) \neq 0$.

**Problem 6.10.** *Draw phase plots as a function of $\mu$ for the three systems from above and prove all statements made above.*

## 6.5. Stability via Liapunov's method

Pick a fixed point $x_0$ of $f$ and an open neighborhood $U(x_0)$ of $x_0$. A **Liapunov function** is a continuous function

$$L : U(x_0) \to \mathbb{R} \tag{6.36}$$

which is zero at $x_0$, positive for $x \neq x_0$, and satisfies

$$L(\phi(t_0)) \geq L(\phi(t_1)), \quad t_0 < t_1, \qquad \phi(t_j) \in U(x_0)\backslash\{x_0\}, \tag{6.37}$$

for any solution $\phi(t)$. It is called a **strict Liapunov function** if equality in (6.37) never occurs. Note that $U(x_0)\backslash\{x_0\}$ can contain no periodic orbits if $L$ is strict (why?).

Since the function $L$ is decreasing along integral curves, we expect the level sets of $L$ to be positively invariant. Let $S_\delta$ be the connected component of $\{x \in U(x_0) | L(x) \leq \delta\}$ containing $x_0$. First of all note that

**Lemma 6.9.** *If $S_\delta$ is compact, then it is positively invariant.*

**Proof.** Suppose $\phi(t)$ leaves $S_\delta$ at $t_0$ and let $x = \phi(t_0)$. Since $S_\delta$ is compact, there is a ball $B_r(x) \subseteq U(x_0)$ such that $\phi(t_0+\varepsilon) \in B_r(x)\backslash S_\delta$ for small $\varepsilon > 0$. But then $L(\phi(t_0 + \varepsilon)) > \delta = L(x)$ contradicting (6.37). □

Moreover, $S_\delta$ is a neighborhood of $x_0$ which shrinks to a point as $\delta \to 0$.

**Lemma 6.10.** *For every $\delta > 0$ there is an $\varepsilon > 0$ such that*

$$S_\varepsilon \subseteq B_\delta(x_0) \quad \textit{and} \quad B_\varepsilon(x_0) \subseteq S_\delta. \tag{6.38}$$

**Proof.** Assume that the first claim in (6.38) is false. Then for every $n \in \mathbb{N}$, there is an $x_n \in S_{1/n}$ such that $|x_n - x_0| > \delta$. Since $S_{1/n}$ is connected, we can even require $|x_n - x_0| = \delta$ and by compactness of the sphere we can pass to a convergent subsequence $x_{n_m} \to y$. By continuity of $L$ we have $L(y) = \lim_{m\to\infty} L(x_{n_m}) = 0$ implying $y = x_0$. This contradicts $|y - x_0| = \delta > 0$.

If the second claim in (6.38) were false, we could find a sequence $x_n$ such that $|x_n - x_0| \leq 1/n$ and $L(x_n) \geq \delta$. But then $\delta \leq \lim_{n\to\infty} L(x_n) = L(x_0) = 0$, again a contradiction. □

Hence, given any neighborhood $V(x_0)$, we can find an $\varepsilon$ such that $S_\varepsilon \subseteq V(x_0)$ is positively invariant. In other words, $x_0$ is stable.

But we can say even more. For every $x$ with $\phi(t,x) \in U(x_0)$, $t \geq 0$, the limit

$$\lim_{t\to\infty} L(\phi(t,x)) = L_0(x) \tag{6.39}$$

exists by monotonicity. Moreover, for every $y \in \omega_+(x)$ we have $L(y) = L_0(x)$. Hence, if $L$ is not constant on any orbit in $U(x_0)\backslash\{x_0\}$ we must have $\omega_+(x) = \{x_0\}$. In particular, this holds for every $x \in S_\varepsilon$ and thus $x_0$ is asymptotically stable.

In summary we have proven **Liapunov's theorem**.

**Theorem 6.11** (Liapunov)**.** *Suppose $x_0$ is a fixed point of $f$. If there is a Liapunov function $L$, then $x_0$ is stable. If, in addition, $L$ is not constant on any orbit lying entirely in $U(x_0)\backslash\{x_0\}$, then $x_0$ is asymptotically stable. This is for example the case if $L$ is a strict Liapunov function.*

Most Liapunov functions will in fact be differentiable. In this case (6.37) holds if and only if

$$\frac{d}{dt}L(\phi(t,x)) = \mathrm{grad}(L)(\phi(t,x))\dot{\phi}(t,x) = \mathrm{grad}(L)(\phi(t,x))f(\phi(t,x)) \leq 0. \tag{6.40}$$

The expression

$$\mathrm{grad}(L)(x)f(x) \tag{6.41}$$

appearing in the previous equation is known as the **Lie derivative** of $L$ along the vector field $f$. A function for which the Lie derivative vanishes is constant on every orbit and is hence called a **constant of motion**.

**Problem 6.11.** *Show that $L(x,y) = x^2 + y^2$ is a Liapunov function for the system*

$$\dot{x} = y, \qquad \dot{y} = -\eta y - x,$$

*where $\eta \geq 0$ and investigate the stability of $(x_0, y_0) = (0,0)$.*

**Problem 6.12** (Gradient systems)**.** *A system of the type*

$$\dot{x} = f(x), \qquad f(x) = -\mathrm{grad}V(x),$$

*is called a* **gradient system***. Investigate the stability of a fixed point. (Hint: Compute Lie derivative of $V$.)*

## 6.6. Newton's equation in one dimension

We have learned in the introduction, that a particle moving in one dimension under the external force field $f(x)$ is described by Newton's equation

$$\ddot{x} = f(x). \tag{6.42}$$

Physicist usually refer to $M = \mathbb{R}^2$ as the **phase space**, to $(x, \dot{x})$ as a **phase point**, and to a solution as a **phase curve**. Theorem 2.3 then says that through every phase point there passes precisely one phase curve.

The **kinetic energy** is the quadratic form

$$T = \frac{\dot{x}^2}{2} \tag{6.43}$$

and the **potential energy** is the function

$$U(x) = -\int_{x_0}^{x} f(\xi) d\xi \tag{6.44}$$

and is only determined up to a constant which can be chosen arbitrarily. The sum of the kinetic and potential energies is called the total **energy** of the system

$$E = T + U(x). \tag{6.45}$$

It is constant along solutions as can be seen from

$$\frac{d}{dt} E = \dot{x}\ddot{x} + U'(x)\dot{x} = \dot{x}(\ddot{x} - f(x)) = 0. \tag{6.46}$$

Hence, the solution can be given implicitly as

$$\int_{x_0}^{x} \frac{d\xi}{\sqrt{2(E - U(\xi))}} = t \tag{6.47}$$

Fixed points of the equation of motion are the solutions of $\dot{x} = 0$, $U'(x) = 0$ and hence correspond to extremal points of the potential. Moreover, if $U'(x_0) = 0$ the energy (more precisely $E - U(x_0)$) can be used as Liapunov function, implying that $x_0$ is stable if $U(x)$ has a local minimum at $U(x_0)$. In summary,

**Theorem 6.12.** *Newton's equations have a fixed point if and only if $\dot{x} = 0$ and $U'(x) = 0$ at this point. Moreover, a fixed point is stable if $U(x)$ has a local minimum there.*

Note that a fixed point cannot be asymptotically stable (why?).

Now let us investigate some examples. We first look at the so called **mathematical pendulum** given by

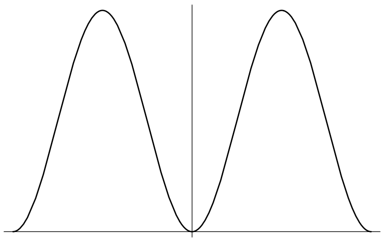
$$\ddot{x} = -\sin(x). \tag{6.48}$$

Here $x$ describes the displacement angle from the position at rest ($x = 0$). In particular, $x$ should be understood modulo $2\pi$. The potential is given by $U(x) = -\cos(x)$. To get a better understanding of this system we will look at some solutions corresponding to various initial conditions. This is usually referred to as phase portrait of the system. We will use *Mathematica* to plot the solutions. The following code will do the computations for us.

```
In[3]:= PhasePlot[f_, ic_, tmax_, opts___] :=
          Block[{i, n = Length[ic], ff, ivp, sol, phaseplot},
            ff = f /. {x → x[t], y → y[t]};
            Off[ParametricPlot :: "ppcom"];
            Do[
              ivp = {x'[t] == ff[[1]], y'[t] == ff[[2]],
                x[0] == ic[[i, 1]], y[0] == ic[[i, 2]]};
              sol = NDSolve[ivp, {x[t], y[t]}, {t, −tmax, tmax}];
              phaseplot[i] =
                ParametricPlot[{x[t], y[t]}/.sol, {t, −tmax, tmax},
                  DisplayFunction → Identity]
              , {i, 1, n}];
            On[ParametricPlot :: "ppcom"];
            Show[Table[phaseplot[i], {i, 1, n}],
              DisplayFunction → $DisplayFunction, opts]
          ];
```

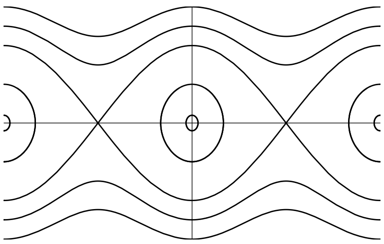
Next, let us define the potential.

```
In[4]:= U[x_] = 1 − Cos[x];
        Plot[U[x], {x, −2π, 2π}, Ticks → False];
```



and plot the phase portrait

```
In[5]:= PhasePlot[{y, −U'[x]}, {{0, 0.2}, {0, 1}, {−2π, 0.2}, {−2π, 1},
          {2π, 0.2}, {2π, 1}, {0, 2}, {2π, −2}, {2π, 2}, {−2π, −2},
          {−2π, 2}, {0, −2}, {0, 2.5}, {0, −2.5}, {0, 3}, {0, −3}},
          2π, PlotRange → {{−2π, 2π}, {−3, 3}}, Ticks → False];
```



Now let us start with a rigorous investigation. We restrict our attention to the interval $x \in (-\pi, \pi]$. The fixed points are $x = 0$ and $x = \pi$. Since the potential has a minimum at $x = 0$, it is stable. Next, the level sets of $E(\dot{x}, x) = const$ are invariant as noted earlier. For $E = 0$ the corresponding level set is the equilibrium position $(\dot{x}, x) = (0, 0)$. For $0 < E < 2$ the level set is homeomorphic to a circle. Since this circle contains no fixed points, it is a regular periodic orbit. Next, for $E = 2$ the level set consists of the fixed point $\pi$ and two non-closed orbits connecting $-\pi$ and $\pi$. It is usually referred to as **separatrix**. For $E > 2$ the level sets are again closed orbits (since we regard everything modulo $2\pi$).

In a neighborhood of the equilibrium position $x = 0$, the system is approximated by its linearization $\sin(x) = x + O(x^2)$ given by

$$\ddot{x} = -x, \tag{6.49}$$

which is called the **harmonic oscillator**. Since the energy is given by $E = \frac{\dot{x}^2}{2} + \frac{x^2}{2}$, the phase portrait consists of circles centered at 0. Moreover, if

$$U'(x_0) = 0, \qquad U''(x_0) = \frac{\omega^2}{2} > 0, \tag{6.50}$$

our system should be approximated by

$$\ddot{y} = -\omega^2 y, \qquad y(t) = x(t) - x_0. \tag{6.51}$$

Finally, let remark that one frequently uses the momentum $p = \dot{x}$ (we have chosen units such that the mass is one) and the location $q = x$ as coordinates. The energy is called the Hamiltonian

$$H(p, q) = \frac{p^2}{2} + U(q) \tag{6.52}$$

and the equations of motion are written as (compare Problem 8.3)

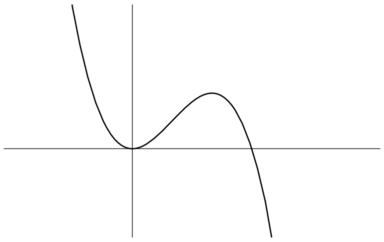
$$\dot{q} = \frac{\partial H(p, q)}{\partial p}, \qquad \dot{p} = -\frac{\partial H(p, q)}{\partial q}. \tag{6.53}$$

This formalism is called **Hamilton mechanics** and it is also useful for systems with more than one degree of freedom. We will return to this point of view in Section 9.3.

**Problem 6.13.** *Consider the mathematical pendulum. If $E = 2$ what is the time it takes for the pendulum to get from $x = 0$ to $x = \pi$?*

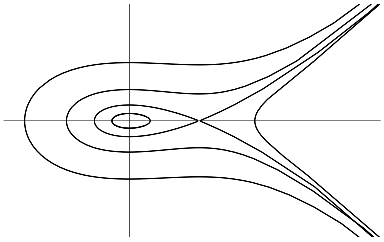
**Problem 6.14.** *Investigate the potential $U(x) = x^2 - 2x^3$.*

```
In[6]:= U[x_] = x^2 - 2x^3;
        Plot[U[x], {x, -0.5, 1}, Ticks → False];
```



*Here are some interesting phase curves to get you started.*

```
In[7]:= PhasePlot[{y, −U'[x]}, {{−0.5, 0}, {−0.3, 0}, {−1/6, 0}, {0.1, 0},
          {0.34, 0}, {0.6, 0}}, 4,
          PlotRange → {{−0.6, 1.2}, {−2, 2}}, Ticks → False];
```



**Problem 6.15.** *The mathematical pendulum with friction is described by*

$$\ddot{x} = -\eta \dot{x} - \sin(x).$$

*Is the energy still conserved in this case? Is the fixed point* $(\dot{x}, x) = (0, 0)$ *(asymptotically) stable? How does the phase portrait change?*

*Discuss also the linearization*

$$\ddot{x} = -\eta \dot{x} - x.$$

**Problem 6.16.** *Consider a more general system with friction*

$$\ddot{x} = -\eta(x)\dot{x} - U'(x), \quad \eta(x) > 0.$$

(i) *Use the energy to show that there are no regular periodic solutions (compare Problem 8.4).*

(ii) *Show that minima of* $U(x)$ *are asymptotically stable.*

Chapter 7

# Local behavior near fixed points

## 7.1. Stability of linear systems

Our aim in this chapter is to show that a lot of information of the stability of a flow near a fixed point can be read off by linearizing the system around the fixed point. But first we need to discuss stability of linear autonomous systems

$$\dot{x} = Ax. \tag{7.1}$$

Clearly, our definition of stability in Section 6.4 is invariant under a linear change of coordinates. Hence it will be no restriction to assume that the matrix $A$ is in Jordan canonical form.

Moreover, from the explicit form (3.39) of $\exp(tJ)$ for a Jordan block $J$ it follows that the long-time behavior of the system is determined by the real part of the eigenvalues. In general it depends on the initial condition and there are two linear manifolds $E^+(\mathrm{e}^A)$ and $E^-(\mathrm{e}^A)$, such that if we start in $E^+(\mathrm{e}^A)$ (resp. $E^-(\mathrm{e}^A)$), then $x(t) \to 0$ as $t \to \infty$ (resp. $t \to -\infty$).

The linear manifold $E^+(\mathrm{e}^A)$ (resp. $E^-(\mathrm{e}^A)$) is called **stable** (resp. **unstable**) **manifold** and is spanned by the generalized eigenvectors corresponding to eigenvalues with negative (resp. positive) real part,

$$E^{\pm}(\mathrm{e}^A) = \bigoplus_{\pm\mathrm{Re}(\alpha_j)<0} \mathrm{Ker}(A - \alpha_j)^{a_j}. \tag{7.2}$$

Similarly one can define the **center manifold** $E^0(\mathrm{e}^A)$ corresponding to the eigenvalues with zero real part. However, these situations are generally of less interest since they are not stable under small perturbations. Hence we will give a system where all eigenvalues have nonzero real part a special name. They are called **hyperbolic** systems.

Observe that (2.31) implies

$$\|\exp(tA)\| \le \mathrm{e}^{|t|\,\|A\|}. \tag{7.3}$$

In the case where all eigenvalues have negative real part we can say much more.

**Theorem 7.1.** *Denote the eigenvalues of $A$ by $\alpha_j$, $1 \le j \le m$, and the corresponding algebraic and geometric multiplicities by $a_j$ and $g_j$, respectively.*

*The system $\dot{x} = Ax$ is globally stable if and only if $\mathrm{Re}(\alpha_j) \le 0$ and $a_j = g_j$ whenever $\mathrm{Re}(\alpha_j) = 0$.*

*The system $\dot{x} = Ax$ is globally asymptotically stable if and only if we have $\mathrm{Re}(\alpha_j) < 0$ for all $j$. Moreover, in this case there is a constant $C$ for every $\alpha < \min\{-\mathrm{Re}(\alpha_j)\}_{j=1}^m$ such that*

$$\|\exp(tA)\| \le C\mathrm{e}^{-t\alpha}. \tag{7.4}$$

**Proof.** As noted earlier, the definition of (asymptotic) stability is of a topological nature and hence invariant under continuous transformations. Moreover, since $\|U\exp(tJ)U^{-1}\| \le \|U\|\|\exp(tJ)\|\|U^{-1}\|$ it is no restriction to assume that $A$ is in Jordan canonical form. Now the first claim is clear from (3.39). For the second claim note that $\|\exp(tA)\| = \mathrm{e}^{-t\alpha}\|\exp(t(A+\alpha))\|$. Since $\mathrm{Re}(\alpha_j + \alpha) < 0$, a look at (3.39) confirms that all entries of the matrix $\exp(t(A+\alpha))$ are bounded. Hence $\exp(t(A+\alpha))$ is bounded and we are done. □

Finally, let us look at the hyperbolic case. In addition, our previous theorem together with the fact that the stable and unstable manifolds are invariant with respect to $A$ (and thus with respect to $\exp(tA)$) immediately give the following result.

**Theorem 7.2.** *The linear stable and unstable manifolds $E^\pm$ are invariant under the flow and every point starting in $E^\pm$ converges exponentially to $0$ as $t \to \pm\infty$. In fact, we have*

$$|\exp(tA)x_\pm| \le C\mathrm{e}^{\mp t\alpha}|x_\pm|, \quad \pm t \ge 0, \quad x_\pm \in E_\pm, \tag{7.5}$$

*for any $\alpha < \min\{|\mathrm{Re}(\alpha)|\,|\alpha \in \sigma(A), \pm\mathrm{Re}(\alpha) > 0\}$ and some $C > 0$ depending on $\alpha$.*

**Problem 7.1.** *For the matrices in Problem 3.6. Determine the stability of the origin and, if the system is hyperbolic, find the corresponding stable and unstable manifolds.*

**Problem 7.2.** *Let $A$ be a two by two matrix and let*

$$\chi_A(z) = z^2 - Tz + D = 0, \qquad T = \mathrm{tr}(A), D = \det(A),$$

*be its characteristic polynomial. Show that $A$ is hyperbolic if $TD \neq 0$. Moreover, $A$ is asymptotically stable if and only if $D > 0$ and $T < 0$. (Hint: $T = \alpha_1 + \alpha_2$, $D = \alpha_1\alpha_2$.)*

*Let $A$ be a three by three matrix and let*

$$\chi_A(z) = z^3 - Tz^2 + Mz - D = 0$$

*be its characteristic polynomial. Show that $A$ is hyperbolic if $(TM - D)D \neq 0$. Moreover, $A$ is asymptotically stable if and only if $D < 0$, $T < 0$ and $TM < D$. (Hint: $T = \alpha_1 + \alpha_2 + \alpha_3$, $M = \alpha_1\alpha_2 + \alpha_2\alpha_3 + \alpha_2\alpha_3$, $D = \alpha_1\alpha_2\alpha_3$, and $TM - D = (\alpha_1 + \alpha_2)(\alpha_1 + \alpha_3)(\alpha_2 + \alpha_3)$.)*

## 7.2. Stable and unstable manifolds

In this section we want to transfer some of our results of the previous section to nonlinear equations. We define the **stable**, **unstable set** of a fixed point $x_0$ as the set of all points converging to $x_0$ for $t \to \infty$, $t \to -\infty$, that is,

$$W^{\pm}(x_0) = \{x \in M | \lim_{t\to\pm\infty} |\Phi(t,x) - x_0| = 0\}. \tag{7.6}$$

Both sets are obviously invariant under the flow. Our goal in this section is to find these sets.

Any function $f \in C^1$ vanishing at $x_0 = 0$ can be decomposed as

$$f(x) = A\,x + g(x), \tag{7.7}$$

where $A$ is the Jacobian of $f$ at 0 and $g(x) = o(|x|)$. Clearly, for small $x$ we expect the solutions to be described by the solutions of the linear equation. This is true for small $t$ by Theorem 2.6, but what about $|t| \to \infty$? In Section 6.4 we saw that for $n = 1$ stability can be read off from $A = f'(0)$ alone as long as $f'(0) \neq 0$. In this section we will generalize this result to higher dimensions.

We will call the fixed point $x_0$ **hyperbolic** if the linearized system is, that is, if none of the eigenvalues of $A$ has zero real part.

We define the **stable** respectively **unstable manifolds** of a fixed point $x_0$ to be the set of all points which converge exponentially to $x_0$ as $t \to \infty$ respectively $t \to -\infty$, that is,

$$M^{\pm}(x_0) = \{x \in M | \sup_{\pm t\geq 0} \mathrm{e}^{\pm\alpha t}|\Phi(t,x) - x_0| < \infty \text{ for some } \alpha > 0\}. \tag{7.8}$$

Both sets are invariant under the flow by construction.

In the linear case we clearly have $M^{\pm}(0) = E^{\pm}(0)$. Our goal is to show, as a generalization of Theorem 7.2, that the sets $M^{\pm}(x_0)$ are indeed manifolds (smooth) and that $E^{\pm}(0)$ is tangent to $M^{\pm}(x_0)$ at 0. Finally, we will show that $M^{\pm}(x_0) = W^{\pm}(x_0)$ in the hyperbolic case.

We will assume that $x_0$ is a hyperbolic fixed point. The key idea is again to formulate our problem as an integral equation which can then be solved by iteration. Since we understand the behavior of the solutions to the linear system we can use the variation of constants formula (3.67) to rewrite our equation as

$$x(t) = \mathrm{e}^{tA}x_0 + \int_0^t \mathrm{e}^{(t-r)A} g(x(r))dr. \tag{7.9}$$

Now denote by $P^{\pm}$ the projectors onto the stable, unstable subspaces $E^{\pm}$ of $\exp(A)$. Moreover, abbreviate $x_{\pm} = P^{\pm}x_0$ and $g_{\pm}(x) = P^{\pm}g(x)$.

What we need is a condition on $x_0 = x_+ + x_-$ such that $x(t)$ remains bounded. Clearly, if $g(x) = 0$, this condition is $x_- = 0$. In the general case, we might still try to express $x_- = h^+(x_+)$. For this we project out the unstable part of our integral equation and solve for $x_-$

$$x_- = \mathrm{e}^{-tA}x_-(t) - \int_0^t \mathrm{e}^{-sA} g_-(x(s))ds. \tag{7.10}$$

Here $x_{\pm}(t) = P^{\pm}x(t)$. If we suppose that $|x(t)|$ is bounded for $t \geq 0$, we can let $t \to \infty$,

$$x_- = -\int_0^\infty \mathrm{e}^{-rA} g_-(x(r))dr, \tag{7.11}$$

where the integral converges absolutely since the integrand decays exponentially. Plugging this back into our equation we see

$$x(t) = \mathrm{e}^{tA}x_+ + \int_0^t \mathrm{e}^{(t-r)A} g_+(x(r))dr - \int_t^\infty \mathrm{e}^{(t-r)A} g_-(x(r))dr. \tag{7.12}$$

Introducing $P(t) = P^+$, $t > 0$, respectively $P(t) = -P^-$, $t \leq 0$, this can be written more compactly as

$$x(t) = K(x)(t), \quad K(x)(t) = \mathrm{e}^{tA}x_+ + \int_0^\infty \mathrm{e}^{(t-r)A} P(t-r) g(x(r))dr. \tag{7.13}$$

To solve this equation by iteration, suppose $|x(t)| \leq \delta$, then, since the Jacobian of $g$ at 0 vanishes, we have

$$\sup_{t\geq 0} |g(x(t)) - g(\tilde{x}(t))| \leq \varepsilon \sup_{t\geq 0} |x(t) - \tilde{x}(t)|, \tag{7.14}$$

where $\varepsilon$ can be made arbitrarily small by choosing $\delta$ sufficiently small. Moreover, for $\alpha < \min\{|\mathrm{Re}(\alpha)|\, |\alpha \in \sigma(A)\}$ we have

$$\|\mathrm{e}^{(t-r)A}P(t-r)\| \leq C\mathrm{e}^{-\alpha|t-r|} \tag{7.15}$$

by (7.5), and we can apply the usual fixed point techniques to conclude existence of a bounded solution $\psi(t, x_+)$ which is $C^k$ with respect to $x_+$ if

$f$ is. The details are deferred to Section 7.4 at the end of this chapter (see Theorem 7.13).

Clearly we have $\psi(t, 0) = 0$. Introducing the function $h^+(a) = P^-\psi(0, a)$ we obtain a good candidate $\{a+h^+(a)|a \in E^+\cap U(0)\}$ for the stable manifold of the nonlinear system in a neighborhood $U(0)$ of 0.

Moreover, I claim that $M^+$ is tangent to $E^+$ at 0. Setting $y(t) = \frac{\partial}{\partial x_+}x(t)$ yields the equation

$$y(t) = \mathrm{e}^{tA}P^+ + \int_0^\infty \mathrm{e}^{(t-r)A}P(t-r)g_x(x(r))y(r)dr \tag{7.16}$$

and in particular, we have

$$y(0)|_{a=0} = P^+ \quad \Rightarrow \quad \frac{\partial}{\partial a}h^+(a)|_{a=0} = 0, \tag{7.17}$$

that is, our candidate is tangent to the linear stable manifold $E^+$ at 0. Details are again deferred to Section 7.4 (see the proof of Theorem 7.13).

Hence we have proven existence of a stable manifold which is tangent to its linear counterpart for a hyperbolic fixed point. The unstable manifold can be obtained by reversing time $t \to -t$.

However, we can do even a little better. I claim that the same proof also shows that

$$M^{\pm,\alpha}(x_0) = \{x \in M| \sup_{\pm t\geq 0} \mathrm{e}^{\pm\alpha t}|\Phi(t,x) - x_0| < \infty\}. \tag{7.18}$$

is a smooth manifold. This is the counterpart of $E^{\pm,\alpha}$, the space spanned by all eigenvectors of $A$ with real part less/bigger than $\mp\alpha$.

**Theorem 7.3.** *Suppose $f \in C^k$ has a fixed point $x_0$ with corresponding Jacobian $A$. Then, if $-\alpha \notin \sigma(A)$, there is a neighborhood $U(x_0)$ and a function $h^{+,\alpha} \in C^k(E^{+,\alpha}, E^{-,\alpha})$ such that*

$$M^{+,\alpha}(x_0) \cap U(x_0) = \{x_0 + a + h^{+,\alpha}(a)|a \in E^{+,\alpha} \cap U\}. \tag{7.19}$$

*Both $h^{+,\alpha}$ and its Jacobian vanish at $x_0$, that is, $M^{+,\alpha}(x_0)$ is tangent to its linear counterpart $E^{+,\alpha}$ at $x_0$.*

**Proof.** To see this, make the change of coordinates $\tilde{x}(t) = \exp(\alpha\, t)x(t)$, transforming $A$ to $\tilde{A} = A + \alpha\mathbb{I}$ and $g(x)$ to $\tilde{g}(t, \tilde{x}) = \exp(\alpha\, t)g(\exp(-\alpha\, t)\tilde{x})$. Since $\tilde{A}$ and $\tilde{g}$ satisfy the same assumptions we conclude, since $\sup_{t\geq 0}|\tilde{x}(t)| \leq \delta$, that $\sup_{t\geq 0}|x(t)| \leq \delta\exp(-\alpha\, t)$. By uniqueness of the solution of our integral equation in a sufficiently small neighborhood of $x_0$ we obtain (7.19). □

As first consequence we obtain existence of stable and unstable manifolds even in the non hyperbolic case, since $M^+(x_0) = M^{+,\varepsilon}(x_0)$ for $\varepsilon > 0$ sufficiently small.

**Theorem 7.4** (Stable manifold)**.** *Suppose $f \in C^k$ has a fixed point $x_0$ with corresponding Jacobian $A$. Then, there is a neighborhood $U(x_0)$ and functions $h^\pm \in C^k(E^\pm, E^\mp)$ such that*

$$M^\pm(x_0) \cap U(x_0) = \{x_0 + a + h^\pm(a) | a \in E^\pm \cap U\}. \tag{7.20}$$

*Both $h^\pm$ and their Jacobians vanish at $x_0$, that is, $M^\pm(x_0)$ are tangent to their respective linear counterpart $E^\pm$ at $x_0$. Moreover,*

$$|\Phi(t,x) - x_0| \le C\mathrm{e}^{\mp t\alpha}, \pm t \ge 0, x \in M^\pm \tag{7.21}$$

*for any $\alpha < \min\{|\mathrm{Re}(\alpha)|\,|\alpha \in \sigma(A), \mathrm{Re}(\alpha) \ne 0\}$ and some $C > 0$ depending on $\alpha$.*

Moreover, we can even get a nonlinear counterpart of the center subspace $E^0$ of the system by considering $M^0(x_0) = M^{+,-\varepsilon}(x_0) \cap M^{-,-\varepsilon}(x_0)$ for $\varepsilon > 0$ sufficiently small.

**Theorem 7.5** (Center manifold)**.** *Suppose $f \in C^k$ has a fixed point $x_0$ with corresponding Jacobian $A$. Then, the set*

$$M^0(x_0) = M^{+,-\alpha}(x_0) \cap M^{-,-\alpha}(x_0) \tag{7.22}$$

*for some $\alpha < \min\{|\mathrm{Re}(\alpha)|\,|\alpha \in \sigma(A),\ \mathrm{Re}(\alpha) \ne 0\}$, is an invariant $C^k$ manifold tangent to $E^0$ at $x_0$.*

For example, consider

$$\dot{x} = -\alpha_0 x, \qquad \dot{y} = y^2 \qquad \alpha_0 > 0. \tag{7.23}$$

Let $\alpha < \alpha_0$, then $M^+(0) = M^{+,\alpha}(0) = \{(x,y)|y = 0\} = E^+$ and $M^-(0) = M^{-,\alpha}(0) = \emptyset = E^-$. Moreover, $M^{+,-\alpha}(0) = \mathbb{R}^2$ and $M^{-,-\alpha}(0) = \{(x,y)|x = 0\} = E^0$ implying $M^0 = \{(x,y)|x = 0\} = E^0$. However, there are infinitely many other smooth invariant manifold tangent to $E^0$ (can you find them?).

In the hyperbolic case we can even say a little more.

**Theorem 7.6.** *Suppose $f \in C^k$ has a hyperbolic fixed point $x_0$. Then there is a neighborhood $U(x_0)$ such that $\gamma_\pm(x) \subset U(x_0)$ if and only if $x \in M^\pm(x_0)$. In particular,*

$$W^\pm(x_0) = M^\pm(x_0). \tag{7.24}$$

**Proof.** This follows since we have shown that any solution staying sufficiently close to $x_0$ solves (7.12). Hence uniqueness of the solution (in a sufficiently small neighborhood of $x_0$) implies that the initial value must lie in $M^+(x_0)$. □

It can happen that an orbit starting in the unstable manifold of one fixed point $x_0$ ends up in the stable manifold of another fixed point $x_1$. Such an orbit is called **heteroclinic orbit** if $x_0 \neq x_1$ and **homoclinic orbit** if $x_0 = x_1$. See the problems for examples.

Moreover, as another consequence we obtain

**Corollary 7.7.** *Suppose $f \in C^k$, $f(x_0) = 0$, and let all eigenvalues of the Jacobian of $f$ at $x_0$ have negative real part. Then the point $x_0$ is asymptotically stable.*

It also follows that, if the fixed point $x_0$ of $f$ is hyperbolic and $A$ has at least one eigenvalue with positive real part, then $x_0$ is unstable (why?).

Finally, it is also possible to include the case where $f$ depends on a parameter $\lambda \in \Lambda$. If $x_0$ is a hyperbolic fixed point for $f(x, 0)$ then, by the implicit function theorem, there is a fixed point $x_0(\lambda)$ (which is again hyperbolic) for $\lambda$ sufficiently small. In particular we have

$$f(x, \lambda) = A(\lambda)(x - x_0(\lambda)) + g(x, \lambda), \tag{7.25}$$

where $A(\lambda)$ is the Jacobian of $f(., \lambda)$ at $x_0(\lambda)$. By Problem 3.4, the projectors $P^\pm(\lambda) = P^\pm(A(\lambda))$ vary smoothly with respect to $\lambda$ and we can proceed as before to obtain (compare Problem 7.10)

**Theorem 7.8.** *Suppose $f \in C^k$ and let $x(\lambda)$ be as above. Then, there is a neighborhood $U(x_0)$ and functions $h^\pm \in C^k(E^\pm \times \Lambda, E^\mp)$ such that*

$$M^\pm(x_0(\lambda)) \cap U(x_0) = \{x(\lambda) + P^\pm(\lambda)a + h^\pm(a, \lambda) | a \in E^\pm \cap U\}. \tag{7.26}$$

**Problem 7.3.** *Find the linearization of*

$$f(x) = (x_2, -\sin(x_1)).$$

*and determine the stability of $x = 0$ if possible.*

**Problem 7.4** (Duffing equation)**.** *Investigate the Duffing equation*

$$\ddot{x} = -\delta\dot{x} + x - x^3, \quad \delta \geq 0.$$

*Determine the stability of the fixed points by linearization. Find the stable and unstable manifolds.*

**Problem 7.5.** *Classify the fixed points of the Lorenz equation*

$$f(x) = (x_2 - x_1, rx_1 - x_2 - x_1x_3, x_1x_2 - x_3), \quad r > 0,$$

*according to stability. At what value of $r$ does the number of fixed points change?*

**Problem 7.6.** *Consider the system*

$$f(x) = (-x_1, x_2 + x_1^2).$$

*Find the flow (Hint: Start with the equation for $x_1$.). Next, find the stable and unstable manifolds. Plot the phase portrait and compare it to the linearization.*

**Problem 7.7** (Heteroclinic orbit)**.** *Determine the stability of the fixed points of the pendulum (6.48) by linearization. Find the stable and unstable manifolds. Find a heteroclinic orbit.*

**Problem 7.8** (Homoclinic orbit)**.** *Determine the stability of the fixed points of the system in Problem 6.14 by linearization. Find the stable and unstable manifolds. Find a homoclinic orbit.*

**Problem 7.9.** *Consider the system*

$$f(x) = (-x_1 - x_2^2, x_2 + x_1^2)$$

*and find an approximation to the stable manifold by computing a few iterations of (7.12). Plot the phase portrait (numerically) and compare it to the linearization.*

**Problem 7.10.** *Suppose $A(\lambda)$ is a matrix which is $C^k$ with respect to $\lambda$ in some compact set. Suppose there is an $0 < \alpha < \min\{|\mathrm{Re}(\alpha)|\,|\alpha \in \sigma(A(\lambda))\}$, then*

$$\| \left(\frac{d}{d\lambda}\right)^n \mathrm{e}^{tA(\lambda)} P(\lambda, t)\| \le C_n(1 + |t|^n)\mathrm{e}^{-\alpha|t|}, \quad n \le k.$$

*(Hint: Start with the case where $A(\lambda)$ is a scalar. In the general case use the power series for the exponential to find the derivative. The problem is that $A(\lambda)$ and its derivatives might not commute. However, once you take the norm ...)*

## 7.3. The Hartman-Grobman theorem

The result of the previous section only tells us something about the orbits in the stable and unstable manifold. In this section we want to prove a stronger result, which shows that the orbits near a hyperbolic fixed point are locally just continuously deformed versions of their linear counterparts.

We begin with a lemma for maps.

**Lemma 7.9.** *Suppose $A$ is an invertible matrix with no eigenvalues on the unit circle and choose a norm such that $\alpha = \max(\|A_-^{-1}\|, \|A_+\|) < 1$. Then for every bounded $g$ satisfying*

$$|g(x) - g(y)| \le \varepsilon|x - y|, \qquad \varepsilon < (1 - \alpha), \tag{7.27}$$

*there is a unique continuous map $\varphi(x) = x + h(x)$ with $h$ bounded such that*

$$\varphi \circ A = f \circ \varphi, \qquad f = A + g. \tag{7.28}$$

*If $f$ is invertible (e.g. if $\varepsilon\|A^{-1}\| < 1$), then $h$ is a homeomorphism and if $g(0) = 0$ then $\varphi(0) = 0$.*

**Proof.** The condition (7.28) is equivalent to

$$h(Ax) - Ah(x) = g(x + h(x)). \tag{7.29}$$

We will investigate this equation in the Banach space of continuous functions with the sup norm. Introduce the linear operators $L : (Lh)(x) = h(Ax) - Ah(x)$ and $U : (Uh)(x) = h(Ax)$ The operator $U$ is clearly invertible (since $A$) is and we have $\|U\| = \|U^{-1}\| = 1$ (it even preserves the norm). Moreover, I claim that $L$ is invertible as well. To show this we use the decomposition $A = A_- \oplus A_+$ which induces the decompositions $L = L_- \oplus L_+$, where $L_\pm h_\pm(x) = h_\pm(Ax) - A_\pm h_\pm(x)$, and $U = U_- \oplus U_+$, where $U_\pm h_\pm(x) = h_\pm(Ax)$. Then we have

$$\begin{aligned}
\|(U_- - A_-)^{-1}\| &= \| - A_-^{-1} \sum_{n=0}^{\infty} A_-^{-n} U^n \| \le \frac{\alpha}{1-\alpha} \le \frac{1}{1-\alpha}, \\
\|(U_+ - A_+)^{-1}\| &= \|U^{-1} \sum_{n=0}^{\infty} A_+^n U^{-n} \| \le \frac{1}{1-\alpha},
\end{aligned} \tag{7.30}$$

which shows that $L^{-1} = (U_- - A_-)^{-1} \oplus (U_+ - A_+)^{-1}$ exists. Hence it remains to solve the fixed point equation

$$h(x) = L^{-1} g(x + h(x)). \tag{7.31}$$

Since the operator on the right is a contraction,

$$\begin{aligned}
&\|L^{-1} g(x + h_1(x)) - L^{-1} g(x + h_2(x))\| \\
&\le \frac{1}{1-\alpha} \|g(x + h_1(x)) - g(x + h_2(x))\| \\
&\le \frac{\varepsilon}{1-\alpha} \|h_1 - h_2\|,
\end{aligned} \tag{7.32}$$

it follows that there is a unique solution by the contraction principle.

Now suppose $f$ is invertible, then there is a map $\vartheta(x) = x + k(x)$ such that $\vartheta \circ A^{-1} = f^{-1} \circ \vartheta$, that is, $A \circ \vartheta = \vartheta \circ f$. Hence $A \circ \vartheta \circ \varphi = \vartheta \circ f \circ \varphi = \vartheta \circ \varphi \circ A$ and thus $\vartheta \circ \varphi = \mathbb{I}$ by the uniqueness part of our result (in the case $g \equiv 0$). Similarly, $A^{-1} \circ \varphi \circ \vartheta = \varphi \circ \vartheta \circ A^{-1}$ implies $\varphi \circ \vartheta = \mathbb{I}$ and thus $\varphi$ is a homeomorphism.

To show $\varphi(0) = 0$ evaluate $A\varphi^{-1}(x) = \varphi^{-1}(f(x))$ at $x = 0$ which shows $A\varphi^{-1}(0) = \varphi^{-1}(0)$. But this equation has only the solution $\varphi^{-1}(0) = 0$. □

**Corollary 7.10.** *Suppose there is a homeomorphism $\varphi(x) = x + h(x)$ with $h$ bounded such that*

$$\varphi \circ A = f \circ \varphi, \tag{7.33}$$

*then $\varphi$ is unique.*

**Proof.** Suppose there are two such maps $\varphi_1$ and $\varphi_2$. Then $(\varphi_1 \varphi_2^{-1})A = A(\varphi_1 \varphi_2^{-1})$ shows that $\varphi_1 \varphi_2^{-1} = \mathbb{I}$ by our above lemma (in the case $g \equiv 0$). □

Now we are able to prove the anticipated result.

**Theorem 7.11** (Hartman-Grobman)**.** *Suppose $f$ is a differentiable vector field with $0$ as a hyperbolic fixed point. Denote by $\Phi(t,x)$ the corresponding flow and by $A = df_0$ the Jacobian of $f$ at $0$. Then there is a homeomorphism $\varphi(x) = x + h(x)$ with $h$ bounded such that*

$$\varphi \circ \mathrm{e}^{tA} = \Phi_t \circ \varphi \tag{7.34}$$

*in a sufficiently small neighborhood of* $0$.

**Proof.** Set $y(t,x) = \frac{\partial\Phi}{\partial x}(t,x)$, then

$$y(t,x) = \mathbb{I} + \int_0^t \frac{\partial f}{\partial x}(\Phi(s,x))\, y(s,x)\, ds. \tag{7.35}$$

Setting $x = 0$ the solution is given by

$$y(t,0) = \mathrm{e}^{tA}. \tag{7.36}$$

So let us try to apply Lemma 7.9 to show (7.34) for fixed $t$, say $t = 1$.

Let $\phi_\delta$ be a smooth bump function such that $\phi_\delta(x) = 0$ for $|x| \le \delta$ and $\phi_\delta(x) = 1$ for $|x| \ge 2\delta$. Replacing $f$ by the function $f + \phi_\delta(A - f)$, it is no restriction to consider the global problem with $f = A$ for $|x| \ge 2\delta$. To be able to apply Lemma 7.9 we need to show that $z(1,x)$, defined by

$$y(t,x) = \mathrm{e}^{tA} + z(t,x), \tag{7.37}$$

can be made arbitrarily small by choosing $\delta$ small. This follows by applying Gronwall's inequality (Problem 2.8) to

$$z(t,x) = \int_0^t g(\Phi(s,x))\mathrm{e}^{sA} ds + \int_0^t f(\Phi(s,x)) z(s,x) ds \tag{7.38}$$

and using that $g(x) = df_x - A$ can be made arbitrarily small by choosing $\delta$ small.

Hence, there is a $\varphi$ such that (7.34) holds at least for $t = 1$. Furthermore, the map $\varphi_s = \Phi_s \circ \varphi \circ \mathrm{e}^{-sA}$ also satisfies (7.34) for $t = 1$. Hence, if we can show that $\varphi_t(x) = x + h_t(x)$ with $h_t$ bounded, then Corollary 7.10 will tell us $\varphi = \varphi_t$ which is precisely (7.34). Now observe

$$h_t = \Phi_t \circ \varphi \circ \mathrm{e}^{-tA} - \mathbb{I} = (\Phi_t - \mathrm{e}^{tA}) \circ \mathrm{e}^{-tA} + \Phi_t \circ h \circ \mathrm{e}^{-tA}, \tag{7.39}$$

where the first term is bounded since $\Phi_t(x) = \mathrm{e}^{tA}x$ for $|x| \ge 2\delta$ and the second is since $h$ is. □

Two systems with vector fields $f$, $g$ and respective flows $\Phi_f$, $\Phi_g$ are said to be **topologically conjugate** if there is a homeomorphism $\varphi$ such that

$$\varphi \circ \Phi_{f,t} = \Phi_{g,t} \circ \varphi. \tag{7.40}$$

Note that topological conjugacy of flows is an equivalence relation.

The Hartman-Grobman theorem hence states that $f$ is locally conjugate to its linearization $A$ at a hyperbolic fixed point. In fact, there is even a stronger results which says that two vector fields are locally conjugate near hyperbolic fixed points if and only if the dimensions of the stable and unstable subspaces coincide.

To show this, it suffices to show this result for linear systems. The rest then follows from transitivity of the equivalence relations and the Hartman-Grobman theorem.

**Theorem 7.12.** *Suppose $A$ and $B$ are two matrices with no eigenvalues on the imaginary axis. If the dimensions of their respective stable and unstable subspaces for their flows are equal, then their flows are topologically conjugate.*

**Proof.** First of all, it is no restriction to assume that $\mathbb{R}^n = \mathbb{R}^s \oplus \mathbb{R}^u$, where $\mathbb{R}^s$ and $\mathbb{R}^u$ are the stable and unstable subspaces for both flows (in fact, we could even assume that both matrices are in Jordan canonical form using a linear conjugacy). Treating both parts separately, it suffices to prove the two cases $s = n$ and $u = n$. Moreover, it even suffices to prove the case $s = n$, since the other one follows by considering $A^{-1}$, $B^{-1}$.

So let us assume $s = n$, that is, all eigenvalues have negative real part. Hence there is a norm such that $|\exp(tA)x|_A \le \exp(-t\alpha)|x|_A$ for all $t \ge 0$ (Problem 3.3). From this and $|\exp(tA)x|_A \ge \exp(-t\alpha)|x|_A$ for all $t \le 0$ it follows that any nonzero solution $x(t) = \exp(tA)x$ satisfies $\frac{d}{dt}|x(t)|_A < 0$ and hence there is a unique time $\tau_A(x)$ such that $|\exp(\tau(x)A)x|_A = 1$. Since this unit sphere is transversal, $\tau_A$ is even a smooth function by Lemma 6.8. Note $\tau_A(\exp(tA)x) = \tau_A(x) - t$. Similar considerations can be made for $B$.

Then the function $h_{AB}(x) = x/|x|_B$ maps the unit sphere for $A$ continuously to the one for $B$. Moreover, since the inverse is given by $h_{BA}(x) = x/|x|_A$ it is a homeomorphism. Now consider the map

$$h(x) = \exp(-\tau_A(x)B)h_{AB}(\exp(\tau_A(x)A)x), \qquad x \neq 0, \tag{7.41}$$

which is a homeomorphism from $\mathbb{R}^n\backslash\{0\}$ to itself. In fact its inverse is given by

$$h^{-1}(x) = \exp(-\tau_B(x)A)h_{BA}(\exp(\tau_B(x)B)x), \qquad x \neq 0, \tag{7.42}$$

which follows easily since $\tau(x) = \tau(y)$ if $y = h(x)$. Furthermore, since $\tau(x) \to -\infty$ as $x \to 0$ we have $|h(x)| \le c\|\exp(-\tau(x)B)\| \to 0$ as $x \to 0$. Thus we can extend $h$ to a homeomorphism from $\mathbb{R}^n$ to itself by setting $h(0)$.

Finally, $h$ is a topological conjugation since

$$\begin{aligned} h(\exp(tA)x) &= \exp((t-\tau_A(x))B)h_{AB}(\exp((\tau_A(x)-t)A)\exp(tA)x) \\ &= \exp(tB)h(x), \end{aligned} \tag{7.43}$$

where we have used $\tau_A(\exp(tA)x) = \tau_A(x) - t$. □

**Problem 7.11.** *Let*

$$A = \begin{pmatrix} \alpha & \beta \\ -\beta & \alpha \end{pmatrix}, \quad B = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

*Explicitly compute the conjugacy found in the proof of Theorem 7.12 in polar coordinates.*

## 7.4. Appendix: Hammerstein integral equations

During Section 7.2 we encountered the following **Hammerstein integral equation**

$$K_\lambda(x)(t) = k(t,\lambda) + \int_0^\infty \kappa(s-t,\lambda)K(s,x(s),\lambda)ds, \tag{7.44}$$

where

$$k, \kappa \in C([0,\infty)\times\Lambda, \mathbb{R}^n), \qquad K \in C([0,\infty)\times U\times\Lambda, \mathbb{R}^n), \tag{7.45}$$

with $\Lambda \subset \mathbb{R}^n$ compact. Now we are going to show the analog of Theorem 2.20 for this equation, which we used in Section 7.2. Again this result is rather technical and you can skip this section.

We assume that for every compact set $C \subseteq U$, $k$ and $K$ are uniformly continuous and bounded

$$|k(t,\lambda)| \le m, \quad |K(t,x,\lambda)| \le M, \quad (t,x,\lambda) \in [0,\infty)\times C\times\Lambda, \tag{7.46}$$

and that there is a dominating function $\alpha(s)$ such that

$$|\kappa(s+t,\lambda)| \le \alpha(s) \quad \text{for} \quad |t| \le \varepsilon. \tag{7.47}$$

In addition, suppose

$$|K(s,x,\lambda) - K(s,y,\lambda)| \le L|x-y|, \qquad x,y \in U, \tag{7.48}$$

where $L$ is independent of $\lambda$, and that

$$L\int_{-\infty}^{\infty} |\kappa(s,\lambda)|ds \le \theta < 1. \tag{7.49}$$

**Theorem 7.13.** *Let $K_\lambda$ satisfy the requirements from above. Then the fixed point equation $K_\lambda(x) = x$ has a unique solution $\overline{x}(t,\lambda) \in C([0,\infty)\times\Lambda, U)$.*

*Assume in addition that all partial derivatives of order up to $r$ with respect to $\lambda$ and $x$ of $k(t,\lambda)$, $\kappa(s,\lambda)$, and $K(s,x,\lambda)$ are continuous. Furthermore, for all partial derivatives of order up to $r$ with respect to $\lambda$ of $\kappa(s,\lambda)$*

*there are dominating functions as in (7.47) and all partial derivatives of order up to $r$ with respect to $\lambda$ and $x$ of $K(s,x,\lambda)$ are uniformly continuous and bounded when $x$ is restricted to compacts as in (7.46). Then all partial derivatives of order up to $r$ with respect to $\lambda$ of $\overline{x}(t,\lambda)$ are continuous.*

**Proof.** As in Theorem 2.20 it is no restriction to assume $k(t,\lambda)\equiv 0$. Choose

$$\delta = (1-\theta)^{-1}\|K_\lambda(0)\|, \tag{7.50}$$

then $\|x\|\le\delta$ implies

$$\begin{aligned}\|K_\lambda(x)\| &\le \int_0^\infty |\kappa(s-t,\lambda)|(|K(s,0,\lambda)|+|K(s,x(s),\lambda)-K(s,0,\lambda)|)ds\\ &\le \|K_\lambda(0)\|+\theta\|x\|\le\delta \end{aligned}\tag{7.51}$$

and hence $K_\lambda$ maps $C([0,\infty),B_\delta(0))$ into itself. Moreover, by assumption $K_\lambda$ is a contraction with contraction constant $\theta$ implying that there is a unique solution $\overline{x}(\lambda,t)$.

Next, we want to show that $K_\lambda(x)$ is continuous with respect to $\lambda$,

$$\begin{aligned}&|K_\lambda(x)(t)-K_\eta(x)(t)|\le\\ &\qquad\int_0^\infty |\kappa(s-t,\lambda)|\,|K(s,x(s),\lambda)-K(s,x(s),\eta)|ds\\ &\qquad\int_0^\infty |\kappa(s-t,\lambda)-\kappa(s-t,\eta)|\,|K(s,x(s),\eta)|ds.\end{aligned}\tag{7.52}$$

By uniform continuity of $K$, for every $\varepsilon>0$ we have $|K(s,x,\lambda)-K(s,x,\eta)|\le\varepsilon$ provided $|\lambda-\eta|$ is sufficiently small and hence

$$\|K_\lambda(x)(t)-K_\eta(x)(t)\|\le\frac{\varepsilon\theta}{L}+M\int_{-\infty}^\infty|\kappa(s-t,\lambda)-\kappa(s-t,\eta)|ds. \tag{7.53}$$

Since the right hand side can be made arbitrarily small by choosing $|\lambda-\eta|$ small, the claim follows.

Now we can show that $\overline{x}$ is continuous. By our previous consideration, the first term in

$$|\overline{x}(t,\lambda)-\overline{x}(s,\eta)|\le|\overline{x}(t,\lambda)-\overline{x}(t,\eta)|+|\overline{x}(t,\eta)-\overline{x}(s,\eta)| \tag{7.54}$$

converges to zero as $(t,\lambda)\to(s,\eta)$ and so does the second since

$$\begin{aligned}&|\overline{x}(t,\eta)-\overline{x}(s,\eta)|\\ &\qquad\le\int_0^\infty|\kappa(r-t,\eta)-\kappa(r-s,\eta)|\,|K(r,\overline{x}(r,\eta),\eta)|dr\\ &\qquad\le M\int_0^\infty|\kappa(r-t,\eta)-\kappa(r-s,\eta)|dr.\end{aligned}\tag{7.55}$$

Hence the case $r=0$ is finished.

Now let us turn to the second claim. Suppose that $\overline{x}(t,\lambda) \in C^1$, then $\overline{y}(t,\lambda) = \frac{\partial}{\partial\lambda}\overline{x}(t,\lambda)$ is a solution of the fixed point equation $\tilde{K}_\lambda(\overline{x}(\lambda), y) = y$. Here

$$\begin{aligned}
\tilde{K}_\lambda(x,y)(t) &= \int_0^\infty \kappa_\lambda(s-t,\lambda)K(s,x(s),\lambda)ds \\
&\quad + \int_0^\infty \kappa(s-t,\lambda)K_\lambda(s,x(s),\lambda)ds \\
&\quad + \int_0^\infty \kappa(s-t,\lambda)K_x(s,x(s),\lambda)y(s)ds, \qquad (7.56)
\end{aligned}$$

where the subscripts denote partial derivatives. The rest follows as in the proof of the Theorem 2.20. To show that $\tilde{K}_\lambda(x,y)$ depends continuously on $x$ you need to use uniform continuity of $K$ and its derivatives. $\square$

Chapter 8

# Planar dynamical systems

## 8.1. The Poincaré–Bendixson theorem

This section is devoted to the case where $M$ is an open subset of $\mathbb{R}^2$. Flows in $\mathbb{R}^2$ are particularly simple because of the **Jordan Curve Theorem**: Every **Jordan curve** $J$ (i.e., a homeomorphic image of the circle $S^1$) dissects $\mathbb{R}^2$ into two connected regions. In particular, $\mathbb{R}^2 \backslash J$ has two components.

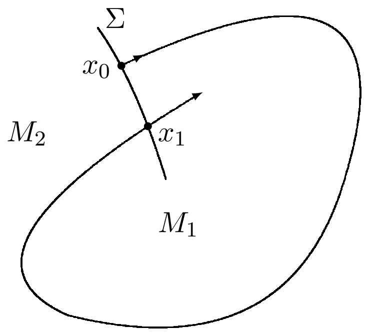
By an **arc** $\Sigma \subset \mathbb{R}^2$ we mean a submanifold of dimension one given by a smooth map $t \to s(t)$. Using this map the points of $\Sigma$ can be ordered. Moreover, for each regular $x \in M$ (i.e., $f(x) \neq 0$), we can find an arc $\Sigma$ containing $x$ which is transversal to $f$ (i.e., $\dot{s}_1(t) f_2(s(t)) - \dot{s}_2(t) f_1(s(t)) \neq 0$).

**Lemma 8.1.** *Let $x_0 \in M$ be a regular point and $\Sigma$ a transversal arc containing $x_0$. Denote by $x_n = x(t_n)$, $n \geq 1$, the (maybe finite) ordered (according to $t_n$) sequence of intersections of $\gamma_\sigma(x_0)$ with $\Sigma$. Then $x_n$ is monotone (with respect to the order of $\Sigma$).*

**Proof.** We only consider $\sigma = +$. If $x_0 = x_1$ we are done. Otherwise consider the curve $J$ from $x_0$ to $x_1$ along $\gamma_+(x_0)$ and back from $x_1$ to $x_0$ along $\Sigma$. This curve $J$ is the image of a continuous bijection from $S^1$ to $J$. Since $S^1$ is compact, it is a homeomorphism. Hence $J$ is a Jordan curve and $M \backslash J = M_1 \cup M_2$.

Now let $\tilde{\Sigma}$ be the arc from $x_0$ to $x_1$ along $\Sigma$. Then $f$ always points either in the direction of $M_1$ or $M_2$ since it cannot change direction by transversality of $\Sigma$. Hence either $\gamma_+(x_1) \subset M_1$ or $\gamma_+(x_1) \subset M_2$. Moreover, if $x_0 < x_1$, then $\gamma_+(x_1)$ must remain in the component containing all points

$x \in \Sigma$, $x_1 < x$, and if $x_0 > x_1$, then $\gamma_+(x_1)$ must remain in the component containing all points $x \in \Sigma$, $x_1 > x$.




Iterating this procedure proves the claim. □

Let $y \in \Sigma \cap \omega_\sigma(x)$ and $t_n \to \sigma\infty$ such that $x_n = \Phi(t_n, x) \to y$. Then, by Lemma 6.8 (with $x = y$ and $T = 0$), we can use $\tilde{t}_n = t_n + \tau(x_n)$ to obtain a sequence $\tilde{t}_n \to \sigma\infty$, $\tilde{x}_n = \Phi(\tilde{t}_n, x) \to y$ such that $\tilde{x}_n \in \Sigma \cap \gamma_\sigma(x)$.

**Corollary 8.2.** *Let $\Sigma$ be a transversal arc, then $\omega_\sigma(x)$ intersects $\Sigma$ in at most one point.*

**Proof.** Suppose there are two points of intersections $y_{1,2}$. Then there exist sequences $x_{1,n}$, $x_{2,n} \in \Sigma \cap \gamma_\sigma(x)$ converging to $y_1$, $y_2$, respectively. But this is not possible by monotonicity found in Lemma 8.1. □

**Corollary 8.3.** *Suppose $\omega_\sigma(x) \cap \gamma_\sigma(x) \neq \emptyset$. Then $x$ is periodic and hence $\omega_+(x) = \omega_-(x) = \gamma(x)$.*

**Proof.** First of all note that our assumption implies $\gamma_\sigma(x) \subseteq \omega_\sigma(x)$ by invariance of $\omega_\sigma(x)$. Assume $y \in \omega_\sigma(x) \cap \gamma_\sigma(x)$ is not fixed. Pick a transversal arc $\Sigma$ containing $y$ and a sequence $x_n \in \Sigma \cap \gamma_\sigma(x) \subseteq \Sigma \cap \omega_\sigma(x)$. By the previous corollary we must have $x_n = y$ and hence $y$ is periodic. □

**Corollary 8.4.** *A minimal compact $\sigma$ invariant set $C$ is a periodic orbit.*

**Proof.** Pick $x \in C$. Then $\omega_\sigma(x) = C$ and hence $\omega_\sigma(x) \cap \gamma_\sigma(x) \neq \emptyset$. Therefore $x$ is periodic by the previous corollary. □

After this sequence of corollaries we proceed with our investigation of $\omega_\pm$ limit sets.

**Lemma 8.5.** *If $\omega_\sigma(x) \neq \emptyset$ is compact and contains no fixed points, then $\omega_\sigma(x)$ is a regular periodic orbit.*

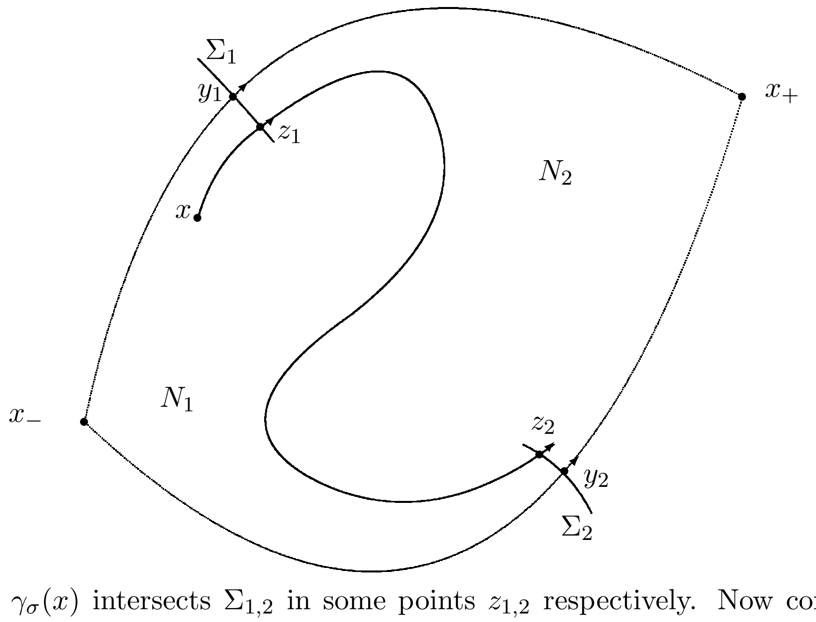
**Proof.** Let $y \in \omega_\sigma(x)$. Take $z \in \omega_\sigma(y) \subseteq \omega_\sigma(x)$ which is not fixed by assumption. Pick a transversal arc $\Sigma$ containing $z$ and a sequence $y_n \to z$ with $y_n \in \Sigma \cap \gamma_\sigma(y)$. Since $\Sigma \cap \gamma_\sigma(y) \subseteq \Sigma \cap \omega_\sigma(x) = \{z\}$ by Corollary 8.2 we conclude $y_n = z$ and hence $\omega_\sigma(x)$ is a regular periodic orbit. □

**Lemma 8.6.** *Suppose $\omega_\sigma(x)$ is connected and contains a regular periodic orbit $\gamma(y)$. Then $\omega_\sigma(x) = \gamma(y)$.*

**Proof.** If $\omega_\sigma(x)\backslash\gamma(y)$ is nonempty, then, by connectedness, there is a point $\tilde{y} \in \gamma(y)$ such that we can find a point $z \in \omega_\sigma(x)\backslash\gamma(y)$ arbitrarily close to $\tilde{y}$. Pick a transversal arc $\Sigma$ containing $\tilde{y}$. By Lemma 6.8 we can find $\tau(z)$ such that $\Phi(\tau(z), z) \in \Sigma$. But then we even have $\Phi(\tau(z), z) \in \Sigma \cap \omega_\sigma(x) = \{\tilde{y}\}$ (by Corollary 8.2) and hence $z \in \gamma(y)$ contradicting our assumption. □

**Lemma 8.7.** *Let $x \in M$, $\sigma \in \{\pm\}$, and suppose $\omega_\sigma(x)$ is compact. Let $x_\pm \in \omega_\sigma(x)$ be distinct fixed points. Then there exists at most one orbit $\gamma(y) \subset \omega_\sigma(x)$ with $\omega_\pm(y) = x_\pm$.*

**Proof.** Suppose there are two orbits $\gamma(y_{1,2})$. Since $\lim_{t\to\pm\infty} \Phi(t, y_{1,2}) = x_\pm$, we can extend $\Phi(t, y_{1,2})$ to continuous functions on $\mathbb{R} \cup \{\pm\infty\}$ by $\Phi(\pm\infty, y_{1,2}) = x_\pm$. Hence the curve $J$ from $x_-$ to $x_+$ along $\gamma(y_1)$ and back from $x_+$ to $x_-$ along $\gamma(y_2)$ is a Jordan curve. Writing $M\backslash J = M_1 \cup M_2$ we can assume $x \in M_1$ (since $x \in J$ is prohibited by Corollary 8.3). Pick two transversal arcs $\Sigma_{1,2}$ containing $y_{1,2}$ respectively.




Then $\gamma_\sigma(x)$ intersects $\Sigma_{1,2}$ in some points $z_{1,2}$ respectively. Now consider the Jordan curve from $y_1$ to $z_1$ to $z_2$ to $y_2$ to $x_+$ and back to $y_1$ (along

$\Sigma_1$, $\gamma_\sigma(x)$, $\Sigma_2$, $\gamma(y_2)$, $\gamma(y_1)$). It dissects $M$ into two parts $N_1, N_2$ such that $\gamma_\sigma(z_1)$ or $\gamma_\sigma(z_2)$ must remain in one of them, say $N_2$ (as in the proof of Lemma 8.1).But now $\gamma_\sigma(x)$ cannot return close to points of $\gamma(y_{1,2}) \cap N_1$ contradicting our assumption. □

These preparations now yield the following theorem.

**Theorem 8.8** (Poincaré–Bendixson)**.** *Let $M$ be an open subset of $\mathbb{R}^2$ and $f \in C^1(M, \mathbb{R}^2)$. Fix $x \in M$, $\sigma \in \{\pm\}$, and suppose $\omega_\sigma(x) \neq \emptyset$ is compact, connected, and contains only finitely many fixed points. Then one of the following cases holds:*

(i) *$\omega_\sigma(x)$ is a fixed orbit.*

(ii) *$\omega_\sigma(x)$ is a regular periodic orbit.*

(iii) *$\omega_\sigma(x)$ consists of (finitely many) fixed points $\{x_j\}$ and unique non-closed orbits $\gamma(y)$ such that $\omega_\pm(y) \in \{x_j\}$.*

**Proof.** If $\omega_\sigma(x)$ contains no fixed points it is a regular periodic orbit by Lemma 8.5. If $\omega_\sigma(x)$ contains at least one fixed point $x_1$ but no regular points, we have $\omega_\sigma(x) = \{x_1\}$ since fixed points are isolated and $\omega_\sigma(x)$ is connected.

Suppose that $\omega_\sigma(x)$ contains both fixed and regular points. Let $y \in \omega_\sigma(x)$ be regular. We need to show that $\omega_\pm(y)$ consists of one fixed point. Therefore it suffices to show that it cannot contain regular points. Let $z \in \omega_\pm(y)$ be regular. Take a transversal arc $\Sigma$ containing $z$ and a sequence $y_n \to z$, $y_n \in \gamma(y) \cap \Sigma$. By Corollary 8.2 $\gamma(y) \subseteq \omega_\sigma(x)$ can intersect $\Sigma$ only in $y$. Hence $y_n = z$ and $\gamma(y)$ is regular periodic. Now Lemma 8.6 implies $\gamma(y) = \omega_\sigma(x)$ which is impossible since $\omega_\sigma(x)$ contains fixed points. □

Finally let me remark, that since the domain surrounded by a periodic orbit is invariant, Lemma 6.5 implies

**Lemma 8.9.** *The interior of every periodic orbit must contain a fixed point.*

**Problem 8.1.** *Can*

$$\phi(t) = \begin{pmatrix} \cos(2t) \\ \sin(t) \end{pmatrix}$$

*be the solution of an autonomous system $\dot{x} = f(x)$? (Hint: Plot the orbit.) Can it be the solution of $\dot{x} = f(t, x)$?*

**Problem 8.2.** *Find and prove a "Poincaré–Bendixson theorem" in $\mathbb{R}^1$?*

**Problem 8.3.** *Suppose* $\mathrm{div} f = 0$. *Show that there is a function $F(x)$ such that $f_1(x) = \frac{\partial F(x)}{\partial x_2}$ and $f_2(x) = -\frac{\partial F(x)}{\partial x_1}$. Show that every orbit $\gamma(x)$ satisfies $F(\gamma(x)) = const.$ Apply this to Newton's equation $\ddot{x} = f(x)$ in $\mathbb{R}$.*

**Problem 8.4** (Bendixson's criterion)**.** *Suppose* div$f$ *does not change sign and does not vanish identically in a simply connected region $U \subseteq M$. Show that there are no regular periodic orbits contained (entirely) inside $U$. (Hint: Suppose there is one and consider the line integral of $f$ along this curve. Recall the Gauss theorem in $\mathbb{R}^2$.)*

*Use this to show that*

$$\ddot{x} + p(x)\dot{x} + q(x) = 0$$

*has no regular periodic solutions if $p(x) > 0$.*

**Problem 8.5** (Dulac's criterion)**.** *Show the following generalization of Bendixson's criterion. Suppose there is a scalar function $\alpha(x)$ such that* div$(\alpha f)$ *does not change sign and does not vanish identically in a simply connected region $U \subseteq M$, then there are no regular periodic orbits contained (entirely) inside $U$.*

**Problem 8.6.** *If the intersection $\omega_+(x) \cap \omega_-(x) \neq \emptyset$ contains a non fixed point, then $x$ is periodic.*

## 8.2. Examples from ecology

In this section we want to consider a model from ecology. It describes two populations, one predator species $y$ and one prey species $x$. Suppose the growth rate of the prey without predators is $A$ (compare Problem 1.12). If predators are present, we assume that the growth rate is reduced proportional to the number of predators, that is,

$$\dot{x} = (A - By)x, \qquad A, B > 0. \tag{8.1}$$

Similarly, if there is no prey, the numbers of predators will decay at a rate $-D$. If prey is present, we assume that this rate increases proportional to the amount of prey, that is

$$\dot{y} = (Cx - D)y, \qquad C, D > 0. \tag{8.2}$$

Scaling $x$, $y$, and $t$ we arrive at the system

$$\begin{array}{l} \dot{x} = (1-y)x \\ \dot{y} = \alpha(x-1)y \end{array}, \qquad \alpha > 0, \tag{8.3}$$

which are the predator-prey equations of **Volterra and Lotka**.

There are two fixed points. First of all, $(0,0)$ is a hyperbolic saddle whose stable manifold is $x = 0$ and whose unstable manifold is $y = 0$. In particular, the first quadrant $Q = \{(x,y) | x > 0, y > 0\}$ is invariant. This is the region we are interested in. The second fixed point $(1,1)$ is not hyperbolic and hence the stability cannot be obtained by linearization.

Hence let us try to eliminate $t$ from our differential equations to get a single first order equation for the orbits. Writing $y = y(x)$, we infer from the chain rule

$$\frac{dy}{dx} = \frac{dy}{dt}\left(\frac{dx}{dt}\right)^{-1} = \alpha\frac{(x-1)y}{(1-y)x}. \tag{8.4}$$

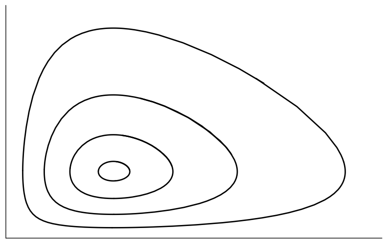
This equation is separable and solving it shows that the orbits are given implicitly by

$$L(x,y) = f(y) + \alpha f(x) = const, \qquad f(x) = \ln(x) - x + 1. \tag{8.5}$$

The function $f$ cannot be inverted in terms of elementary functions (its inverse is $-W(-\exp(const - 1 + \alpha(x-1))x^{-\alpha})$, where $W$ is a branch of the product log function). However, it is not hard to see that the level sets are compact. Hence each orbit is periodic and surrounds the fixed point $(1,1)$.

**Theorem 8.10.** *All orbits of the Volterra–Lotka equations (8.3) in $Q$ are closed and encircle the only fixed point* $(1,1)$.

The phase portrait is depicted below.



Next, let us refine this model by assuming limited grow for both species (compare again Problem 1.12). The corresponding system is given by

$$\begin{aligned} \dot{x} &= (1 - y - \lambda x)x \\ \dot{y} &= \alpha(x - 1 - \mu y)y \end{aligned}, \qquad \alpha, \lambda, \mu > 0. \tag{8.6}$$

Again the fixed point $(0,0)$ is a hyperbolic saddle whose stable manifold is $x = 0$ and whose unstable manifold is $y = 0$.

We first look at the case where $\lambda \ge 1$ and hence where there is only one additional fixed point in $\overline{Q}$, namely $(\lambda^{-1}, 0)$. It is a hyperbolic sink if $\lambda > 1$ and if $\lambda = 1$, one eigenvalue is zero. Unfortunately, the equation for the orbits is no longer separable and hence a more thorough investigation is necessary to get a complete picture of the orbits.

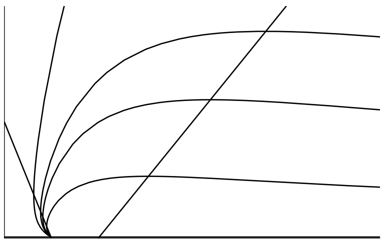
The key idea now is to split $Q$ into regions where $\dot{x}$ and $\dot{y}$ have definite signs and then use the following elementary observation (Problem 8.7).

**Lemma 8.11.** *Let $\phi(t) = (x(t), y(t))$ be the solution of a planar system. Suppose $U$ is open and $\overline{U}$ is compact. If $x(t)$ and $y(t)$ are strictly monotone in $U$, then either $\phi(t)$ hits the boundary at some finite time $t = t_0$ or $\phi(t)$ converges to a fixed point $(x_0, y_0) \in \overline{Q}$.*

Now let us see how this applies to our case. These regions where $\dot{x}$ and $\dot{y}$ have definite signs are separated by the two lines

$$L_1 = \{(x, y)|y = 1 - \lambda x\}, \qquad L_2 = \{(x, y)|\mu y = x - 1\}. \tag{8.7}$$

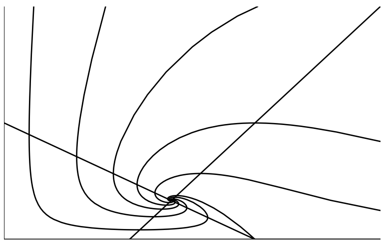
A typical situation for $\alpha = \mu = 1$, $\lambda = 2$ is depicted below.



This picture seems to indicate that all trajectories converge to the fixed point $(\lambda^{-1}, 0)$. Now let us try to prove this. Denote the regions in $Q$ enclosed by these lines by (from left to right) by $Q_1$, $Q_2$, and $Q_3$. Suppose we start at a point $(x_0, y_0) \in Q_3$. Then, adding to $Q_3$ the constraint $x \le x_0$, we can apply Lemma 8.11 to conclude that the trajectory enters $Q_2$ through $L_2$ or converges to a fixed point in $\overline{Q_2}$. The last case is only possible if $(\lambda^{-1}, 0) \in \overline{Q_2}$, that is, if $\lambda = 1$. Similarly, starting in $Q_2$ the trajectory will enter $Q_1$ via $L_1$ or converge to $(\lambda^{-1}, 0)$. Finally, if we start in $Q_1$, the only possibility for the trajectory is to converge to $(\lambda^{-1}, 0)$.

In summary, we have proven that for $\lambda \ge 1$ every trajectory in $Q$ converges to $(\lambda^{-1}, 0)$.

Now consider the remaining case $0 < \lambda < 1$. Then $(\lambda^{-1}, 0)$ is a hyperbolic saddle and there is a second fixed point $(\frac{1+\mu}{1+\mu\lambda}, \frac{1-\lambda}{1+\mu\lambda})$, which is a sink. A phase portrait for $\alpha = \mu = 1$, $\lambda = \frac{1}{2}$ is shown below.



Again it looks like all trajectories converge to the sink in the middle. We will use the same strategy as before. Now the lines $L_1$ and $L_2$ split $Q$ into four regions $Q_1$, $Q_2$, $Q_3$, and $Q_4$ (where $Q_4$ is the new one). As before we can show that trajectories pass through these sets according to $Q_4 \to Q_3 \to Q_2 \to Q_1 \to Q_4$ unless they get absorbed by the sink in the middle. Note that since the stable manifold of $(\lambda^{-1}, 0)$ is still $y = 0$, no trajectory in $Q$ can converge to it. However, there is now a big difference to the previous case: A trajectory starting in $Q_4$ can return to $Q_4$ and hence there could be periodic orbits.

To exclude periodic orbits we will try to find a Liapunov function. Inspired by (8.5) we introduce

$$L(x,y) = y_0 f(\frac{y}{y_0}) + \alpha\, x_0 f(\frac{x}{x_0}), \tag{8.8}$$

where we have abbreviated $(x_0, y_0) = (\frac{1+\mu}{1+\mu\lambda}, \frac{1-\lambda}{1+\mu\lambda})$ for our fixed point. In fact, using

$$\dot{x} = (y_0 - y - \lambda(x - x_0))x, \qquad \dot{y} = \alpha(x - x_0 - \mu(y - y_0))y \tag{8.9}$$

we compute

$$\dot{L} = \frac{\partial V}{\partial x}\dot{x} + \frac{\partial V}{\partial y}\dot{y} = -\alpha\lambda(x - x_0)^2 - \alpha\mu(y - y_0)^2 < 0. \tag{8.10}$$

Hence we again see that all orbits starting in $Q$ converge to the fixed point $(x_0, y_0)$.

**Theorem 8.12.** *Suppose $\lambda \geq 1$, then there is no fixed point of the equations (8.6) in $Q$ and all trajectories in $Q$ converge to the point $(0, \lambda^{-1})$.*

*If $0 < \lambda < 1$ there is only one fixed point $(\frac{1+\mu}{1+\mu\lambda}, \frac{1-\lambda}{1+\mu\lambda})$ in $Q$. It is asymptotically stable and all trajectories converge to this point.*

For our original model this means that the predators can only survive if their growth rate is positive at the limiting population $\lambda^{-1}$ of the prey species.

**Problem 8.7.** *Prove Lemma 8.11.*

**Problem 8.8** (Volterra principle)**.** *Show that for any orbit of the Volterra–Lotka system (8.3), the time average over one period*

$$\frac{1}{T}\int_0^T x(t)dt = 1, \qquad \frac{1}{T}\int_0^T y(t)dt = 1$$

*is independent of the orbit. (Hint: Integrate $\frac{d}{dt}\ln(x(t))$ over one period.)*

**Problem 8.9.** *Show that the change of coordinates $x = \exp(q)$, $y = \exp(p)$ transforms the Volterra–Lotka system (8.3) into a Hamiltonian system with Hamiltonian $H(p, q) = L(\exp(q), \exp(p))$.*

*Moreover, use the same change of coordinates to transform (8.6). Then use the Bendixson's criterion (Problem 8.4) to show that there are no periodic orbits.*

**Problem 8.10.** *Show that (8.6) has no periodic orbits in the case $\lambda < 1$ if $\mu\lambda \geq 1$ as follows:*

*If there is a periodic orbit it must contain a point* $(x_0, y_0)$ *on* $L_1$ *which satisfies*

$$\frac{1+\mu}{1+\mu\lambda} < x_0 < \frac{1}{\lambda}, \qquad y_0 = 1 - \lambda x_0. \tag{8.11}$$

*The trajectory enters* $Q_1$ *and satisfies* $x(t) < x_0$ *in* $Q_1$ *since* $x(t)$ *decreases there. Hence we must have* $y(t) < y_1 = \frac{x_0-1}{\mu}$ *when it hit* $L_2$*. Now we enter* $Q_2$*, where* $y(t)$ *decreases implying* $x(t) < x_1 = \frac{1-y_1}{\lambda}$ *when we hit* $L_1$*. Proceeding like this we finally see* $y(t) > y_2 = \frac{x_1-1}{\mu}$ *when we return to* $L_1$*. If* $y_2 \geq y_0$*, that is if*

$$(1+\mu)(1-\mu\lambda) \geq (1-(\mu\lambda)^2)x_0, \tag{8.12}$$

*the trajectory is spiraling inwards and we get a contradiction to our assumption that it is periodic. This is the case when* $\mu\lambda \geq 1$.

**Problem 8.11** (Competing species)**.** *Suppose you have two species* $x$ *and* $y$ *such that one inhibits the growth of the other. A simple model describing such a situation would be*

$$\begin{array}{l} \dot{x} = (A - By)x \\ \dot{y} = (C - Dx)y \end{array}, \qquad A, B, C, D > 0.$$

*Find out as much as possible about this system.*

**Problem 8.12** (Competing species with limited growth)**.** *Consider the same setting as in the previous problem but now with limited growth. The equations read*

$$\begin{array}{l} \dot{x} = (1 - y - \lambda x)x \\ \dot{y} = \alpha(1 - x - \mu y)y \end{array}, \qquad \alpha, \lambda, \mu > 0.$$

*Again, find out as much as possible about this system.*

## 8.3. Examples from electrical engineering

An electrical circuit consists of elements each of which has two connectors (*in* and *out*), where every connector of one element is connected to one or more connectors of the other elements. Mathematically speaking we have an ordered graph.

At each time $t$, there will be a certain current $I(t)$ flowing through each element and a certain voltage difference $V(t)$ between its connectors. It is of no importance which connector is called *in* and which one *out*. However, the current is counted positively if it flows from *in* to *out* and similarly for the voltage differences. The state space of the system is given by the pairs $(I, V)$ of all elements in the circuit. These pairs must satisfy two requirements. By Kirchhoff's first law, the sum over all currents in a vertex must vanish (conservation of charge) and by Kirchhoff's second law, the sum over all voltage differences in a closed loop must vanish (the voltage corresponds to a potential).

Usually one has three types of different elements, inductors, capacitors, and resistors. For an inductor we have

$$L\dot{I}_L = V_L, \tag{8.13}$$

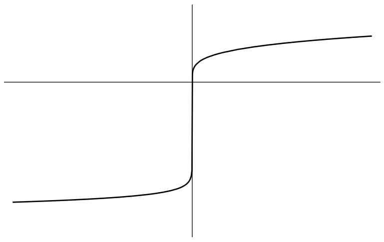
where $L > 0$ is the inductance, $I_L(t)$ is the current through the inductor and $V_L(t)$ is the voltage difference between the connectors. For a capacitor we have

$$C\dot{V}_C = I_C, \tag{8.14}$$

where $C > 0$ is the capacity, $I_C(t)$ is the current through the capacitor and $V_C(t)$ is the voltage difference. For a resistor we have

$$V_R = R(I_R), \tag{8.15}$$

where the function $R(.)$ is called the characteristic of the resistor. Since there is no potential difference if there is no current we must have $R(0) = 0$. One often can assume $R(I) = R\,I$, where the resistance $R$ is a constant (Ohm's law), but for sophisticated elements like semiconductors this is not possible. For example, the characteristic of a diode looks as follows.



In the positive direction you need only a very small voltage to get a large current whereas in the other direction you will get almost no current even for fairly large voltages. Hence one says that a diode lets the current only pass in one direction.

We will look at the case of one inductor, one capacitor, and one resistor arranged in a loop. Kirchhoff's laws yield $I_R = I_L = I_C$ and $V_R + V_L + V_C = 0$. Using the properties of our three elements and eliminating, say, $I_C$, $I_R$, $V_L$, $V_R$ we obtain the system

$$\begin{array}{l} L\dot{I}_L = -V_C - R(I_L) \\ C\dot{V}_C = I_L \end{array}, \qquad f(0) = 0, \quad L, C > 0. \tag{8.16}$$

In addition, note that the change of energy in each element is given by $IV$. By Kirchhoff's laws we have

$$I_L V_L + I_C V_C + I_R V_R = 0, \tag{8.17}$$

which can be rewritten as

$$\frac{d}{dt}\left(\frac{L}{2}I_L^2 + \frac{C}{2}V_C^2\right) = -I_R R(I_R). \tag{8.18}$$

That is, the energy dissipated in the resistor has to come from the inductor and the capacitor.

Finally, scaling $V_C$ and $t$ we end up with **Liénard's equation** (compare Problem 8.13)

$$\begin{array}{l}\dot{x} = y - f(x)\\ \dot{y} = -x\end{array}, \qquad f(0) = 0. \tag{8.19}$$

Equation (8.18) now reads

$$\frac{d}{dt}W(x,y) = -xf(x), \qquad W(x,y) = \frac{x^2+y^2}{2}. \tag{8.20}$$

This equation will be our topic for the rest of this section. First of all, the only fixed point is $(0,0)$. If $xf(x) > 0$ in a neighborhood of $x = 0$, then $W$ is a Liapunov function and hence $(0,0)$ is stable. Moreover, we even have

**Theorem 8.13.** *Suppose $xf(x) \geq 0$ for all $x \in \mathbb{R}$ and $xf(x) > 0$ for $0 < |x| < \varepsilon$. Then every trajectory of Liénard's equation (8.19) converges to $(0,0)$.*

**Proof.** If $W(x,y)$ is constant on an orbit, say $W(x,y) = R^2/2$, then the orbit must be a circle of radius $R$. Hence we must have $f(x) = 0$ for $0 \leq |x| \leq R$ and the result follows from Liapunov's theorem (Theorem 6.11). □

Conversely, note that $(0,0)$ is unstable if $xf(x) < 0$ for $0 < |x| < \varepsilon$.

We will now show that Liénard's equation has periodic orbits if $f$ is odd and if $xf(x)$ is negative for $x$ small and positive for $x$ large. More precisely, we will need the following assumptions.

(i) $f$ is odd, that is, $f(-x) = -f(x)$.

(ii) $f(x) < 0$ for $0 < x < \alpha$ ($f(\alpha) = 0$ without restriction).

(iii) $\liminf_{x\to\infty} f(x) > 0$ and in particular $f(x) > 0$ for $x > \beta$ ($f(\beta) = 0$ without restriction).

(iv) $f(x)$ is monotone increasing for $x > \alpha$.

Furthermore, let us abbreviate $Q_\pm = \{(x,y) | \pm x > 0\}$ and $L_\pm = \{(x,y) | x = 0, \pm y > 0\}$. Our symmetry requirement (i) will allow us to restrict our attention to $Q_+$ since the corresponding results for $Q_-$ will follow via the transformation $(x,y) \to (-x,-y)$ which maps $Q_+$ to $Q_-$ and leave the differential equation (8.19) invariant if $f$ is odd.

As a first observation we note that

**Lemma 8.14.** *Every trajectory of Liénard's equation (8.19) in $Q_+$ can cross the graph of $f(x)$ at most once.*

**Proof.** Suppose a trajectory starts below the graph of $f$, that is $y_0 < f(x_0)$. We need to show that it cannot get above again. Suppose at some time $t_1$ we cross the graph of $f$. Then $y(t_1-\delta) < f(x(t_1-\delta))$ and $y(t_1+\varepsilon) > f(x(t_1+\varepsilon))$ for $\varepsilon, \delta > 0$ sufficiently small. Moreover, we must also have $x(t_1-\delta) > x(t_1)$ and $x(t_1+\varepsilon) > x(t_1)$ by our differential equation. In particular, we can find $\varepsilon$ and $\delta$ such that $x(t_1-\delta) = x(t_1+\varepsilon)$ implying

$$y(t_1+\varepsilon) > f(x(t_1+\varepsilon)) = f(x(t_1-\delta)) > y(t_1-\delta). \tag{8.21}$$

This contradicts that $y(t)$ is decreasing (since $x(t) > 0$). □

Next we show

**Lemma 8.15.** *Suppose $f$ satisfies the requirements (ii) and (iii). Then, every trajectory starting at $L_+$ will hit $L_-$ at a finite positive time.*

**Proof.** Suppose we start at $(0, y_0)$, $y_0 > 0$. First of all note that the trajectory must satisfy $W(x(t), y(t)) \geq \varepsilon^2/2$, where $\varepsilon = \min\{\alpha, y_0\}$. Next, our trajectory must hit the line $\{(x,y)|x = \alpha, y > 0\}$ by Lemma 8.11. Moving on we must hit $\{(x,y)|x > 0, y = 0\}$. Otherwise we would have $x(t) \to \infty$ in finite time (since $\dot{y}(t) \geq \alpha$) which is impossible since $\dot{x}(t) \leq y_0$. But from this point on we must stay within the region $x(t) \leq R$ and $x^2+(y-C)^2 \leq R^2$, where $R > \beta$ is sufficiently large and $C < \inf f(x)$. This follows since the vector field always points to the interior of this region. Applying again Lemma 8.11 finishes the proof. □

Now suppose $f$ satisfies (i)–(iii). Denote the first intersection point of the trajectory starting at $(x(0), y(0)) = (0, y_0) \in L_+$ with $L_-$ by $(x(T), y(T)) = (0, P(y_0))$. Then, every periodic orbit orbit must encircle $(0,0)$ and satisfy $P(y_0) = -y_0$. Hence every periodic orbit corresponds to a zero of the function

$$\Delta(y_0) = W(0, P(y_0)) - W(0, y_0) = -\int_0^T x(t) f(x(t)) dt. \tag{8.22}$$

Now what can we say about this function? Clearly, for $y_0 < \alpha$ we have $\Delta(y_0) > 0$. Moreover, there is a number $r > 0$ such that the trajectory starting at $(0, r)$ intersects the graph of $f$ at $(\beta, 0)$ (show this). So for $y_0 > r$ our trajectory intersects the line $x = \beta$ at $t_1$ and $t_2$. Furthermore, since the intersection with $f$ can only be for $t \in (t_1, t_2)$, we have $y(t) > f(x(t))$ for $0 \leq t \leq t_1$ and $y(t) < f(x(t))$ for $t_2 \leq t \leq T$. Now let us split $\Delta$ into three parts by splitting the integral at $t_1$ and $t_2$. For the first part we obtain

$$\Delta_1(y_0) = -\int_0^{t_1} x(t) f(x(t)) dt = \int_0^{\beta} \frac{-x f(x)}{y(x) - f(x)} dx. \tag{8.23}$$

Since $y(x)$ is increasing as $y_0$ increases (orbits cannot intersect), the absolute value of the integrand in $\Delta_1(y_0)$ decreases. In addition, since $y(t_1) \uparrow \infty$ we have $\Delta_1(y_0) \to 0$. The second part is

$$\Delta_2(y_0) = -\int_{t_1}^{t_2} x(t) f(x(t)) dt = \int_{y(t_1)}^{y(t_2)} f(x(y)) dy < 0. \tag{8.24}$$

By (iii) this part cannot tend to 0. Finally, the absolute value of the integrand in the last part

$$\Delta_3(y_0) = -\int_{t_2}^{T} x(t) f(x(t)) dt = \int_{\beta}^{0} \frac{-x f(x)}{y(x) - f(x)} dx \tag{8.25}$$

also decreases, with a similar argument as for $\Delta_1$.

Moreover, I claim that $\Delta(y_0)$ eventually becomes negative. If $y(t_2) \downarrow -\infty$ then $\Delta_3(y_0) \to 0$ as in the case of $\Delta_1$ and the claim holds. Otherwise, there is a solution which stays below $f$ for all time and we must have $f(x) \to \infty$ (since it must stay above this solution). But then $\Delta_2(y_0) \to -\infty$ (show this) and the claim again holds.

If in addition (iv) holds, it is no restriction to assume $\alpha = \beta$ and we have that $\Delta(y_0)$ is monotone decreasing for $y_0 > r$. Since we must also have $\alpha > r$, there is precisely one zero in this case. This proves

**Theorem 8.16.** *Suppose $f$ satisfies the requirements (i)–(iii). Then Liénard's equation (8.19) has at least one periodic orbit encircling* $(0, 0)$.

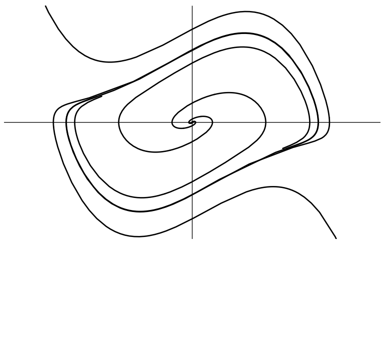
*If in addition (iv) holds, this periodic orbit is unique and every trajectory converges to this orbit as $t \to \infty$.*

The classical application is **van der Pol's equation**

$$\ddot{x} - \mu(1 - x^2)\dot{x} + x = 0, \qquad \mu > 0, \tag{8.26}$$

which models a triode circuit. By Problem 8.13 it is equivalent to Liénard's equation with $f(x) = \mu(\frac{x^3}{3} - x)$. All requirements of Theorem 8.16 are satisfied and hence van der Pol's equation has a unique periodic orbit and all trajectories converge to this orbit as $t \to \infty$.

The phase portrait for $\mu = 1$ is shown below.



**Problem 8.13.** *The equation*

$$\ddot{x} + g(x)\dot{x} + x = 0$$

*is also often called Liénard's equation. Show that it is equivalent to (8.19) if we set $y = \dot{x} + f(x)$, where $f(x) = \int_0^x g(t)dt$.*

Chapter 9

# Higher dimensional dynamical systems

## 9.1. Attracting sets

In most applications, the main interest is to understand the long time behavior of the flow of a differential equation (which we assume $\sigma$ complete from now on for simplicity). In this respect it is important to understand the fate of all points starting in some set $X$. Hence we will extend some of our previous definitions to sets first.

Given a set $X \subseteq M$ we can always obtain a $\sigma$ invariant set by considering

$$\gamma_\pm(X) = \bigcup_{\pm t \ge 0} \Phi(t, X) = \bigcup_{x \in X} \gamma_\pm(x). \tag{9.1}$$

Taking the closure $\overline{\gamma_\sigma(X)}$ we even obtain a closed $\sigma$ invariant set. Moreover, the **$\omega_\pm$-limit set** of $X$ is the set $\omega_\pm(X)$ of all points $y \in M$ for which there exists sequences $t_n \to \pm\infty$ and $x_n \in X$ with $\Phi(t_n, x_n) \to y$.

We will only consider the case $\sigma = +$ from now on for notational simplicity. The set $\omega_+(X)$ can equivalently be characterized as,

$$\omega_+(X) = \bigcap_{t \ge 0} \Phi(t, \overline{\gamma_+(X)}) = \bigcap_{t \ge 0} \overline{\bigcup_{s \ge t} \Phi(s, X)}. \tag{9.2}$$

Clearly, $\omega_+(X)$ is closed as the intersection of closed sets and it is also not hard to see that is invariant (Problem 9.1).

**Lemma 9.1.** *The set $\omega_\pm(X)$ is a closed invariant set.*

In addition, by $\Phi(t, \overline{\gamma_+(X)}) \subseteq \overline{\gamma_+(X)}$ we have $\Phi(s, \overline{\gamma_+(X)}) \subseteq \Phi(t, \overline{\gamma_+(X)})$ for $s > t$ and hence it is immediate that

$$\omega_+(X) = \bigcap_{t \geq t_0} \Phi(t, \overline{\gamma_+(X)}) = \bigcap_{n \in \mathbb{N}} \Phi(n, \overline{\gamma_+(X)}). \tag{9.3}$$

So if $\overline{\gamma_+(X)} \neq \emptyset$ is compact, $\omega_+(X)$ is the intersection of countably many nonempty compact nesting sets and thus it is also a nonempty compact set by the finite intersection property of compact sets.

**Lemma 9.2.** *Suppose $X$ is nonempty. If the set $\overline{\gamma_\sigma(X)}$ is compact, then $\omega_\sigma(X)$ is nonempty and compact. If $\overline{\gamma_\sigma(X)}$ is in addition connected (e.g., if $X$ is connected), then so is $\omega_\sigma(X)$.*

**Proof.** It remains to show that $\Lambda = \omega_+(X)$ is connected. Suppose it is not and can be split into two disjoint closed sets, $\Lambda = \Lambda_0 \cup \Lambda_1$, none of which is empty. Since $\mathbb{R}^n$ is normal, there are disjoint open sets $U_0$ and $U_1$ such that $\Lambda_0 \subset U_0$ and $\Lambda_1 \subset U_1$. Moreover, the set $V_n = \Phi(n, \overline{\gamma_+(X)}) \backslash (U_0 \cup U_1)$ is compact. Hence $V = \bigcap_n V_n$ is either nonempty or $V_n$ is eventually empty. In the first case we must have $V \subset \Lambda$ which is impossible since $V \cap (U_0 \cup U_1) = \emptyset$. Otherwise, if $V_n$ is eventually empty, then $\phi(n, \overline{\gamma_+(X)})$ must be eventually in $U_0$ or in $U_1$ (since $\phi(n, \overline{\gamma_+(X)})$ is connected) implying $\Lambda \subset U_0$ respectively $\Lambda \subset U_1$. Again a contradiction. □

Note that we have

$$\bigcup_{x \in X} \omega_+(x) \subseteq \omega_+(X) \tag{9.4}$$

but equality will not hold in general as the example

$$\dot{x} = x(1 - x^2), \qquad \dot{y} = -y \tag{9.5}$$

shows. In this case it is not hard to see that

$$\omega_+(B_r(0)) = [-1, 1] \times \{0\}, \qquad r > 0, \tag{9.6}$$

but

$$\bigcup_{x \in B_r(0)} \omega_+(x) = \{(-1, 0), (0, 0), (1, 0)\}. \tag{9.7}$$

In particular $\omega_+(B_r(0))$ contains the three fixed points plus their unstable manifolds. That is, all orbits which lie entirely in $B_r(0)$. This is also true in general.

**Theorem 9.3.** *The set $\omega_+(X)$ is the union over all complete orbits lying entirely in $\overline{\gamma_+(X)}$.*

**Proof.** Let $\gamma(y)$ be such a orbit, then $\gamma(y) \subseteq \overline{\gamma_+(X)}$ and invariance of $\gamma(y)$ implies $\gamma(y) \subseteq \overline{\Phi(t, \gamma_+(X))}$ for all $t$ and hence $\gamma(y) \subseteq \omega_+(X)$. The converse follows since $\omega_+(X) \subseteq \overline{\gamma_+(X)}$. □

An invariant set $\Lambda$ is called **attracting** if there exists some neighborhood $U$ of $\Lambda$ such that $U$ is positively invariant and $\Phi_t(x) \to \Lambda$ as $t \to \infty$ for all $x \in U$. The sets

$$W^{\pm}(\Lambda) = \{x \in M | \lim_{t\to\pm\infty} d(\Phi_t(x), \Lambda) = 0\} \tag{9.8}$$

are the **stable** respectively **unstable sets** of $\Lambda$. Here $d(A, B) = \inf\{|x - y| \, | x \in A, y \in B\}$ denotes the distance between two sets $A, B \subseteq \mathbb{R}^n$. The set $W^+(\Lambda)$ is also called the **domain** or **basin of attraction** for $\Lambda$. It is not hard to see that we have

$$W^+(\Lambda) = \bigcup_{t<0} \Phi_t(U) = \{x \in M | \omega_+(x) \subseteq \Lambda\}. \tag{9.9}$$

But how can we find such a set? Fortunately, using our considerations from above, there is an easy way of doing so. An open connected set $E$ whose closure is compact is called a **trapping region** for the flow if $\Phi_t(\overline{E}) \subset E$, $t > 0$. In this case

$$\Lambda = \omega_+(E) = \bigcap_{t\ge 0} \Phi(t, E) \tag{9.10}$$

is an attracting set by construction.

Unfortunately the definition of an attracting set is not always good enough. In our example (9.5) any ball $B_r(0)$ with radius $r > 1$ is a trapping region. However, whereas only the two fixed points $(\pm 1, 0)$ are *really* attracting, the corresponding attracting set $\Lambda$ also contains the repelling fixed point $(0, 0)$ plus its unstable manifold. In particular, the domain of attraction of the two attracting fixed points $W^+(\{(-1, 0), (1, 0)\}) = \{(x, y) \in \mathbb{R}^2 | x = 0\}$ is up to a set of measure zero the same as $W^+(\Lambda) = \mathbb{R}^2$.

In fact, an attracting set will always contain the unstable manifolds of all its points.

**Lemma 9.4.** *Let $E$ be a trapping region, then*

$$W^-(x) \subseteq \omega_+(E), \qquad \forall x \in \omega_+(E). \tag{9.11}$$

**Proof.** From $y \in W^-(x)$ we infer $\Phi(t, y) \in \overline{\gamma_+(E)}$ for $t \to -\infty$. Hence $\gamma(y) \subseteq \overline{\gamma_+(E)}$ and the claim follows from Theorem 9.3. □

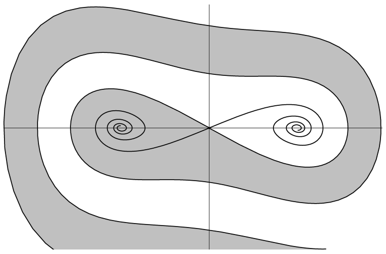
To exclude such situations, we can define an **attractor** to be an attracting set which is topologically transitive. Here a closed invariant set $\Lambda$ is called **topologically transitive** if for any two open sets $U, V \subseteq \Lambda$ there is some $t \in \mathbb{R}$ such that $\Phi(t, U) \cap V \neq \emptyset$. In particular, an attractor cannot be split into smaller attracting sets. Note that $\Lambda$ is topologically transitive if it contains a dense orbit (Problem 9.2).

This implies that only the sets $\{(-1,0)\}$ or $\{(1,0)\}$ are attractors for the above example. The domains of attraction are $W^+(\{(\pm 1,0)\}) = \{(x,y) \in \mathbb{R}^2 | \pm x > 0\}$.

As another example let us look at the **Duffing equation**

$$\ddot{x} = -\delta \dot{x} + x - x^3, \quad \delta \geq 0, \tag{9.12}$$

from Problem 7.4. It has a sink at $(-1,0)$, a hyperbolic saddle at $(0,0)$, and a sink at $(1,0)$. The basin of attraction of the sink $(-1,0)$ is bounded by the stable and unstable manifolds of the hyperbolic saddle $(0,0)$. The situation for $\delta = 0.3$ is depicted below.



Finally, let us consider the van der Pol equation (8.26). The unique periodic orbit is an attractor and its basin of attraction is $\mathbb{R}^2 \backslash \{0\}$. However, not all attractors are fixed points or periodic orbits, as the example in our next section will show.

**Problem 9.1.** *Show that $\omega_\pm(X)$ is invariant under the flow.*

**Problem 9.2.** *Show that a closed invariant set which has a dense orbit is topologically transitive.*

## 9.2. The Lorenz equation

One of the most famous dynamical systems which exhibits chaotic behavior is the **Lorenz equation**

$$\begin{aligned}\dot{x} &= -\sigma(x-y), \\ \dot{y} &= rx - y - xz, \\ \dot{z} &= xy - bz,\end{aligned} \tag{9.13}$$

where $\sigma, r, b > 0$. Lorenz arrived at these equations when modelling a two-dimensional fluid cell between two parallel plates which are at different temperatures. The corresponding situation is described by a complicated system of nonlinear partial differential equations. To simplify the problem, he expanded the unknown functions into Fourier series with respect to the spacial coordinates and set all coefficients except for three equal to zero. The resulting equation for the three time dependent coefficients is (9.13). The variable $x$ is proportional to the intensity of convective motion, $y$ is proportional to the temperature difference between ascending and descending currents, and $z$ is proportional to the distortion from linearity of the vertical temperature profile.

So let us start with an investigation of this system. First of all observe that the system is invariant under the transformation

$$(x, y, z) \to (-x, -y, z). \tag{9.14}$$

Moreover, the $z$ axis is an invariant manifold since

$$x(t) = 0, \quad y(t) = 0, \quad z(t) = z_0 \mathrm{e}^{-bt} \tag{9.15}$$

is a solution of our system.

But now let us come to some deeper results. We first show that the dynamic is quite simple if $r \le 1$. If $r \le 1$ there is only one fixed point of the vector field, namely the origin. The linearization is given by

$$\begin{pmatrix} -\sigma & \sigma & 0 \\ r & -1 & 0 \\ 0 & 0 & -b \end{pmatrix} \tag{9.16}$$

and the corresponding eigenvalues are

$$-b, \qquad -\frac{1}{2}(1 + \sigma \pm \sqrt{(1+\sigma)^2 + 4(r-1)\sigma}). \tag{9.17}$$

Hence the origin is asymptotically stable for $r < 1$. Moreover, it is not hard to see that

$$L(x, y, z) = rx^2 + \sigma y^2 + \sigma z^2 \tag{9.18}$$

is a Liapunov function in this case since one readily verifies

$$\dot{L}(x, y, z) = -\sigma(r(x+y)^2 + (1-r)y^2 + bz^2). \tag{9.19}$$

In particular, the following lemma follows easily from Theorem 6.11 (Problem 9.3).

**Lemma 9.5.** *Suppose $r \le 1$, then the Lorenz equation has only the origin as fixed point and all solutions converge to the origin as $t \to \infty$.*

If $r$ grows above 1, there are two new fixed points

$$(x, y, z) = (\pm\sqrt{b(r-1)}, \pm\sqrt{b(r-1)}, r-1), \tag{9.20}$$

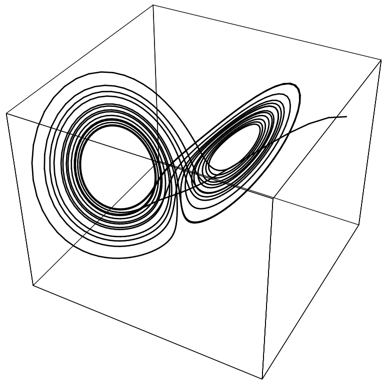
and the linearization is given by

$$\begin{pmatrix} -\sigma & \sigma & 0 \\ 1 & -1 & \mp\sqrt{b(r-1)} \\ \pm\sqrt{b(r-1)} & \pm\sqrt{b(r-1)} & -b \end{pmatrix}. \tag{9.21}$$

One can again compute the eigenvalues but the result would almost fill one page. Note however that by (9.14) the eigenvalues are the same for both points. From (9.17) we can read off that one eigenvalue is now positive and hence the origin is no longer stable. It can be shown that the two new fixed points are asymptotically stable for $1 < r < 470/19 = 2.74$.

Next, let us try to plot some solutions using *Mathematica*.

```
In[1]:= σ = 10; r = 28; b = 8/3;
        sol = NDSolve[{x'[t] == −σ(x[t] − y[t]),
                       y'[t] == −x[t] z[t] + r x[t] − y[t],
                       z'[t] == x[t] y[t] − b z[t],
                       x[0] == 30, y[0] == 10, z[0] == 40},
                       {x, y, z}, {t, 0, 20}, MaxSteps → 5000];
        ParametricPlot3D[Evaluate[{x[t], y[t], z[t]}/.sol], {t, 0, 20},
             PlotPoints → 2000, Axes → False, PlotRange → All];
```



We observe that all trajectories first move inwards and then encircle the two fixed points in a pretty irregular way.

To get a better understanding, let us show that there exists an ellipsoid $E_\varepsilon$ which all trajectories eventually enter and never leave again. To do this, let us consider a small modification of our Liapunov function from above,

$$L(x, y, z) = rx^2 + \sigma y^2 + \sigma(z - 2r)^2. \tag{9.22}$$

A quick computation shows

$$\dot{L}(x, y, z) = -2\sigma(rx^2 + y^2 + b(z - r)^2 - br^2). \tag{9.23}$$

Now let $E$ be the ellipsoid defined by $E = \{(x, y, z)|\dot{L}(x, y, z) \geq 0\}$ and let $M = \max_{(x,y,z)\in E} L(x, y, z)$. Define $E_\varepsilon = \{(x, y, z)|L(x, y, z) \leq M + \varepsilon\}$ for positive $\varepsilon$. Any point outside $E_\varepsilon$ also lies outside $E$ and hence $\dot{L} \leq -\delta < 0$ for such points. That is, for $x \in \mathbb{R}^3\backslash E_\varepsilon$ the value of $L$ is strictly decreasing along its trajectory and hence it must enter $E_\varepsilon$ after some finite time.

Moreover, $E_\varepsilon$ is a trapping region for the Lorenz equation and there is a corresponding attracting set

$$\Lambda = \bigcap_{n\in\mathbb{N}} \Phi(n, E_0), \tag{9.24}$$

which is called the *attractor* of the Lorenz equation. In particular, we see that solutions exist for all positive times. Note also that $W^+(\Lambda) = \mathbb{R}^3$. All fixed points plus their unstable manifolds (if any) must also be contained in $\Lambda$. Moreover, I even claim that $\Lambda$ is of Lebesgue measure zero. To see this we need a generalized version of Liouville's formula (3.58).

**Lemma 9.6.** *Let $\dot{x} = f(x)$ be a dynamical system on $\mathbb{R}^n$ with corresponding flow $\Phi(t, x)$. Let $M$ be a bounded measurable subset of $\mathbb{R}^n$ and let $V = \int_M dx$ be its volume. Abbreviate $M(t) = \Phi(t, M)$ respectively $V(t) = \int_{M(t)} dx$, then*

$$\dot{V}(t) = \int_{M(t)} \operatorname{div}(f(x))\, dx. \tag{9.25}$$

**Proof.** By the change of variable formula for multiple integrals we have

$$V(t) = \int_{M(t)} dx = \int_M \det(d\Phi_t(x))\, dx. \tag{9.26}$$

Since $d\Phi_t = \mathbb{I} + df\, t + o(t)$ we infer $V(t) = \int_M (1 + \operatorname{tr}(df)t + o(t))\, dx$ and hence

$$\dot{V}(0) = \lim_{t\to 0} \frac{V(t) - V(0)}{t} = \lim_{t\to 0} \int_M (\operatorname{tr}(df) + o(1))\, dx = \int_M \operatorname{tr}(df)\, dx \tag{9.27}$$

by the dominated convergence theorem. Replacing $M$ with $M(t)$ shows that the above result holds for all $t$ and not only for $t = 0$. □

Applying this lemma to the Lorenz equation we obtain

$$V(t) = V \mathrm{e}^{-(1+\sigma+b)t} \tag{9.28}$$

since

$$\operatorname{div}(f) = -(1 + \sigma + b). \tag{9.29}$$

In particular, we see that the measure of $\Phi(t, E_0)$ decreases exponentially, and the measure of $\Lambda$ must be zero. Note that this result also implies that none of the three fixed points can be a source.

Our numerical experiments from above show that $\Lambda$ seems to be a quite complicated set. This is why it was called the ***strange attractor*** of the Lorenz equation.

However, this is clearly no satisfying mathematical definition of a strange attractor. One possibility is to call an attractor strange if the dynamical system generated by the **time-one map**

$$\Phi_1 : \Lambda \to \Lambda \tag{9.30}$$

is *chaotic* and if $\Lambda$ is *fractal*. It is still unknown whether the Lorenz attractor is strange in the sense of this definition. See the book by Sparrow [**27**] for a survey of results.

I will not go into any further details at this point. We will see how these terms are defined in Section 12.3 and Section 12.6, respectively. However, I hope that this example shows that even simple systems in $\mathbb{R}^3$ can exhibit very complicated dynamics. I also hope that you can now better appreciate the Poincaré–Bendixson which excludes such strange behavior in $\mathbb{R}^2$.

**Problem 9.3.** *Prove Lemma 9.5.*

**Problem 9.4.** *Solve the Lorenz equation for the case $\sigma = 0$.*

**Problem 9.5.** *Investigate the Lorenz equation for the case $r = \infty$ as follows. First introduce $\varepsilon = r^{-1}$. Then use the change of coordinates $(t, x, y, x) \mapsto (\tau, \xi, \eta, \zeta)$, where $\tau = \varepsilon^{-1} t$, $\xi = \varepsilon x$, $\eta = \sigma\varepsilon^2 y$, and $\zeta = \sigma(\varepsilon^2 z - 1)$.*

*Show that the resulting system for $\varepsilon = 0$ corresponds to a single third order equation $\xi''' = -\xi^2 \xi'$. Integrate this equation once and observe that the result is of Newton type (see Section 6.6). Now what can you say about the solutions?*

## 9.3. Hamiltonian mechanics

In the previous sections we have seen that even simple looking dynamical systems in three dimension can be extremely complicated. In the rest of this chapter we want to show that it is still possible to get some further insight if the system has a special structure. Hence we will look again at systems arising in classical mechanics.

The point of departure in classical mechanics is usually the **Hamilton principle**. Suppose a mechanical system has $n$ degrees of freedom described by coordinates $q \in U \subseteq \mathbb{R}^n$. Associated with such a system is a **Lagrange function**

$$L(v, q), \qquad v = \dot{q}, \tag{9.31}$$

and an integral curve $q(t)$ for which the **action integral**

$$\mathcal{I}(q) = \int_{t_0}^{t_1} L(\dot{q}(t), q(t)) dt \tag{9.32}$$

subject to the boundary conditions $q(t_0) = q_0$, $q(t_1) = q_1$ is extremal.

If $L$ is differentiable, extremal curves can be found by setting the Gateaux derivative of $I$ equal to zero. That is, setting

$$q_\varepsilon(t) = q(t) + \varepsilon r(t), \tag{9.33}$$

we see that a necessary condition for $q$ to be extremal is that

$$\frac{d}{d\varepsilon} \mathcal{I}(q_\varepsilon)\Big|_{\varepsilon=0} = 0. \tag{9.34}$$

Using integration by parts this immediately yields (Problem 9.6) the corresponding **Euler-Lagrange equation**

$$\frac{\partial L}{\partial q} - \frac{d}{dt}\frac{\partial L}{\partial v} = 0. \tag{9.35}$$

In the situation of particles under the influence of some forces we have

$$L(v,q) = \frac{1}{2} vMv - U(q), \tag{9.36}$$

where $M$ is a positive diagonal matrix with the masses of the particles as entries and $U$ is the potential corresponding to the forces. The associated Euler-Lagrange equations are just Newton's equations

$$M\ddot{q} = -\operatorname{grad} U(q). \tag{9.37}$$

If the **momentum**

$$p(v,q) = \frac{\partial L}{\partial v}(v,q) \tag{9.38}$$

is a diffeomorphism for fixed $q$, and hence

$$\det \frac{\partial^2 L}{\partial v^2} \neq 0, \tag{9.39}$$

then we can consider the **Legendre transform** of $L$,

$$H(p,q) = pv - L(v,q), \quad v = v(p,q), \tag{9.40}$$

which is known as the **Hamilton function** of the system. The associated variational principle is that the integral

$$\mathcal{I}(p,q) = \int_{t_0}^{t_1} \Big( p(t)\dot{q}(t) - H(p(t), q(t)) \Big) dt \tag{9.41}$$

subject to the boundary conditions $q(t_0) = q_0$, $q(t_1) = q_1$ is extremal. The corresponding Euler-Lagrange equations are Hamilton's equations

$$\dot{q} = \frac{\partial H(p,q)}{\partial p}, \qquad \dot{p} = -\frac{\partial H(p,q)}{\partial q}. \tag{9.42}$$

This formalism is called **Hamilton mechanics**.

In the special case of some particles we have

$$p = Mv, \qquad H(p,q) = \frac{1}{2} pM^{-1}p + U(q) \tag{9.43}$$

and the Hamiltonian corresponds to the total energy of the system.

Introducing the **symplectic matrix**

$$J = \begin{pmatrix} 0 & \mathbb{I} \\ -\mathbb{I} & 0 \end{pmatrix}, \quad J^{-1} = J^T = -J, \tag{9.44}$$

Hamilton's equation can also be written as

$$\frac{d}{dt}\begin{pmatrix} p \\ q \end{pmatrix} = -\mathrm{grad}_s H(p,q), \tag{9.45}$$

where $\mathrm{grad}_s = -J\,\mathrm{grad}$ is called the **symplectic gradient**.

A straightforward calculation shows that $H$ is a **constant of motion**, that is,

$$\frac{d}{dt}H(p(t),q(t)) = \frac{\partial H}{\partial p}\dot{p} + \frac{\partial H}{\partial q}\dot{q} = -\frac{\partial H}{\partial p}\frac{\partial H}{\partial q} + \frac{\partial H}{\partial q}\frac{\partial H}{\partial p} = 0. \tag{9.46}$$

More generally, for a function $I(p,q)$ its change along a trajectory is given by its Lie derivative (compare (6.41))

$$\frac{d}{dt}I(p(t),q(t)) = \{I(p(t),q(t)), H(p(t),q(t))\}, \tag{9.47}$$

where

$$\{I,H\} = \frac{\partial I}{\partial p}\frac{\partial H}{\partial q} - \frac{\partial I}{\partial q}\frac{\partial H}{\partial p} \tag{9.48}$$

is called **Poisson bracket**. This should be compared with the Heisenberg equation of Problem 3.18.

A function $I(p,q)$ is called a **first integral** if it is constant along trajectories, that is, if

$$\{I,H\} = 0. \tag{9.49}$$

But how can we find first integrals? One source are symmetries.

**Theorem 9.7** (Noether)**.** *Let $\Phi(t,q)$ be the flow generated by $f(q)$. If $\Phi$ leaves the Lagrangian invariant, then*

$$I(v,q) = \frac{\partial L(v,q)}{\partial v} f(q) \tag{9.50}$$

*is a constant of motion.*

**Proof.** Abbreviate $q^s(t) = \Phi(s,q(t))$. The invariance of $L(v,q)$ implies

$$\begin{aligned} 0 &= \frac{d}{ds}L(\dot{q}^s(t),q^s(t))\Big|_{s=0} \\ &= \frac{\partial L}{\partial v}(\dot{q}(t),q(t))\frac{\partial f}{\partial q}(q(t))\dot{q}(t) + \frac{\partial L}{\partial q}(\dot{q}(t),q(t))f(q(t)) \end{aligned} \tag{9.51}$$

and hence

$$\begin{aligned} \frac{d}{dt}I(\dot{q}(t),q(t)) &= \left(\frac{d}{dt}\frac{\partial L}{\partial v}(\dot{q},q)\right) f(q) + \frac{\partial L}{\partial v}(\dot{q},q)\frac{\partial f}{\partial q}(q)\dot{q} \\ &= \left(\frac{d}{dt}\frac{\partial L}{\partial v}(\dot{q},q) - \frac{\partial L}{\partial q}(\dot{q},q)\right) f(q) = 0 \end{aligned} \tag{9.52}$$

by the Euler-Lagrange equation. □

Another important property of Hamiltonian systems is that they are volume preserving. This follows immediately form Lemma 9.6 since the divergence of a Hamiltonian vector field is zero.

**Theorem 9.8** (Liouville)**.** *The volume in phase space is preserved under a Hamiltonian flow.*

This property can often give important information concerning the motion via **Poincaré's recurrence theorem**.

**Theorem 9.9** (Poincaré)**.** *Suppose $\Phi$ is a volume preserving bijection of a bounded region $D \subseteq \mathbb{R}^n$. Then in any neighborhood $U \subseteq D$ there is a point $x$ returning to $U$, that is, $\Phi^n(x) \in U$ for some $n \in \mathbb{N}$.*

**Proof.** Consider the sequence $\Phi^n(U) \subseteq D$. There are two numbers $l$, $k$ such that $\Phi^l(U) \cap \Phi^k(U) \neq \emptyset$ since otherwise their volume would be infinite. Hence $U \cap \Phi^{k-l}(U) \neq \emptyset$. If $y$ is a point in the intersection we have $y = \Phi^{k-l}(x)$, which proves the claim. □

**Problem 9.6.** *Derive the Euler-Lagrange equation (9.35).*

**Problem 9.7** (Legendre transform)**.** *Let $F(v)$ be such that*

$$\det \frac{\partial^2 F}{\partial v^2}(v_0) \neq 0.$$

*Show that the function $p(v) = \frac{\partial F}{\partial v}(v)$ is a local diffeomorphism near $v_0$ and that the Legendre transform*

$$G(p) = pv(p) - F(v(p))$$

*is well defined. Show that*

$$p = \frac{\partial F}{\partial v}(v) \quad \Leftrightarrow \quad v = \frac{\partial G}{\partial p}(p)$$

*and conclude that the Legendre transformation is involutive.*

**Problem 9.8.** *Suppose that $D$ is bounded and positively invariant under a volume preserving flow. Then $D$ belongs to the set of nonwandering points. (Hint: The Poincaré's recurrence theorem and Problem 6.8.)*

**Problem 9.9** (Relativistic mechanics)**. Einstein's equation** *says that the kinetic energy of a relativistic particle is given by*

$$T(v) = m(v)c^2, \qquad m(v) = m_0\sqrt{1 + \frac{v^2}{c^2}},$$

*where $c$ is the speed of light and $m_0$ is the (rest) mass of the particle. Derive the equation of motions from Hamilton's principle using the Lagrangian $L(v,q) = T(v) - U(q)$. Derive the corresponding Hamilton equations.*

## 9.4. Completely integrable Hamiltonian systems

Finally we want to show that there is also a canonical form for a Hamilton system under certain circumstances. To do this we need to transform our system in such a way that the Hamilton structure is preserved. More precisely, if our transformation is given by

$$(P, Q) = \varphi(p, q), \qquad (p, q) = \psi(P, Q), \tag{9.53}$$

we have

$$\begin{pmatrix} \dot{P} \\ \dot{Q} \end{pmatrix} = d\varphi \begin{pmatrix} \dot{p} \\ \dot{q} \end{pmatrix} = -d\varphi J \operatorname{grad} H(p, q) = -(d\varphi J d\varphi^T) \operatorname{grad} K(P, Q), \tag{9.54}$$

where $K = H \circ \varphi$ is the transformed Hamiltonian. Hence, we need to require that the Jacobian of $\varphi$ is a symplectic matrix, that is,

$$d\varphi \in \mathrm{Sp}(2n) = \{M \in \mathrm{Gl}(2n) | M J M^T = J\}, \tag{9.55}$$

where $\mathrm{Sp}(2n)$ is the **symplectic group**. Such a map is called a **symplectic map**. In this case $\varphi$ is also called a **canonical transform**. Alternatively they can be characterized as those transformations which leave the **symplectic two form**

$$\omega((p_1, q_1), (p_2, q_2)) = (p_1, q_1) J (p_2, q_2) = p_1 q_2 - p_2 q_1 \tag{9.56}$$

invariant.

To find canonical transformations, recall that we have derived Hamilton's equations from the variational principle (9.41). Hence, our transform will be canonical if the integrands of (9.41) and

$$\tilde{I}(P, Q) = \int_{t_0}^{t_1} P(t)\dot{Q}(t) - K(P(t), Q(t)) dt \tag{9.57}$$

only differ by a total differential. By $H(p, q) = K(P, Q)$ we are lead to

$$p dq - P dQ = dS, \tag{9.58}$$

where $dq$ has to be understood as $dq(t) = \dot{q}(t)dt$ for a given curve $q(t)$. The function $S$ is called a generating function and could depend on all four variables $p$, $q$, $P$, and $Q$. However, since only two of them are independent in general, it is more natural to express two of them by the others.

For example, we could use

$$S = S_1(q, Q) \tag{9.59}$$

and

$$p dq - P dQ = \frac{\partial S_1}{\partial q} dq + \frac{\partial S_1}{\partial Q} dQ \tag{9.60}$$

shows we have

$$p = \frac{\partial S_1}{\partial q}, \qquad P = -\frac{\partial S_1}{\partial Q}, \tag{9.61}$$

since the previous equation must hold for all curves $q(t)$ and $Q(t)$. Moreover, if we require

$$\det \frac{\partial S_1}{\partial q \partial Q} \neq 0, \tag{9.62}$$

we can solve $p = \frac{\partial S_1(q,Q)}{\partial q}$ locally for $Q = Q(p,q)$ and hence our canonical transformation is given by

$$(P, Q) = (\frac{\partial S_1}{\partial Q}(q, Q(p,q)), Q(p,q)). \tag{9.63}$$

Similarly we could choose

$$S = -PQ + S_2(P, q), \tag{9.64}$$

where

$$pdq - PdQ = -QdP - PdQ + \frac{\partial S_2}{\partial P}dP + \frac{\partial S_2}{\partial Q}dQ \tag{9.65}$$

implies

$$Q = \frac{\partial S_2}{\partial P}, \qquad p = \frac{\partial S_2}{\partial q}. \tag{9.66}$$

Again, if we require

$$\det \frac{\partial S_2}{\partial P \partial q} \neq 0, \tag{9.67}$$

we obtain a canonical transformation

$$(P, Q) = (P(p,q), \frac{\partial S_2}{\partial P}(P(p,q), q)). \tag{9.68}$$

The remaining two cases

$$S = qp + S_3(Q, p) \quad \text{and} \quad S = qp - PQ + S_4(P, p) \tag{9.69}$$

are left as an exercise.

Now let us return to our canonical form. We will start with one dimension, that is, $n = 1$ with $H(p, q)$ as in (6.52). Let $q_0$ be a local minimum of $U(q)$ surrounded by periodic orbits $\gamma_E$ which are uniquely determined by the energy $E$ of a point on the orbit. The two intersection points of $\gamma_E$ with the $q$ axis to the left and right of $q_0$ will be denoted by $q_-(E)$ and $q_+(E)$, respectively. In particular, note $U(q_\pm(E)) = E$.

The integral over the momentum along such a periodic orbit

$$I(E) = \frac{1}{2\pi}\int_{\gamma_E} p\, dq = \frac{1}{\pi}\int_{q_-(E)}^{q_+(E)} \sqrt{2(E - U(q))}dq \tag{9.70}$$

is called the **action variable**. Next, by (6.47)

$$I'(E) = \frac{1}{\pi}\int_{q_-(E)}^{q_+(E)} \frac{dq}{\sqrt{2(E-U(q))}} = \frac{T(E)}{2\pi} > 0, \tag{9.71}$$

where $T(E)$ is the period of $\gamma_E$ and thus we can express $E$ as a function of $I$, say $E = K(I)$. Hence if we take $I$ as one of our new variables, the new Hamiltonian $K$ will depend on $I$ only. To find a suitable second variable we will look for a generating function $S_2(I,q)$. Since we want $p = \frac{\partial S_2}{\partial q}$ we set

$$S_2(I,q) = \int_{q_-(K(I))}^{q} p dq = \int_{q_-(K(I))}^{q} \sqrt{2(K(I)-U(q))} dq \tag{9.72}$$

and the second variable is

$$\theta = \frac{\partial S_2}{\partial I} = \int_{q_-(E)}^{q} \frac{I'(E)^{-1} dq}{\sqrt{2(E-U(q))}} = \frac{2\pi}{T(E)} t, \tag{9.73}$$

where $t$ is the time it takes from $q_-(E)$ to $q$ (compare again (6.47) and note $K'(I) = I'(E)^{-1}$). The variable $\theta$ is called the **angle variable** and is only defined modulo $2\pi$. The equation of motion read

$$\begin{aligned} \dot{I} &= -\frac{\partial K}{\partial \theta} = 0, \\ \dot{\theta} &= \frac{\partial K}{\partial I} = \Omega(I), \end{aligned} \tag{9.74}$$

where $\Omega(I) = 2\pi/T(K(I))$.

The main reason why we could find such a canonical transform to action-angle variables is the existence of a first integral, namely the Hamiltonian. In one dimension this single first integral suffices to decompose the surfaces of constant energy into periodic orbits. In higher dimensions this is no longer true unless one can find $n$ first integrals $L_j$ which are functionally independent and in involution, $\{L_j, L_k\} = 0$. Such systems are called **completely integrable**. If the system is integrable, the $n$ first integrals can be used to define the $n$-dimensional manifolds $\Gamma_c = \{(p,q) | L_j(p,q) = c_j,\ 1 \le j \le n\}$ which can be shown to be diffeomorphic to an $n$-dimensional torus (if they are compact). Taking a basis of cycles $\{\gamma_j(c)\}_{j=1}^n$ on the torus $\Gamma_c$ one can define the action variables as before via

$$I_j(c) = \frac{1}{2\pi}\int_{\gamma_j(c)} p\, dq \tag{9.75}$$

and the angle variables via a generating function $S_2(I,q) = \int^q p\, dq$. I do not want to go into further details here but I refer to the excellent book by Arnold [**2**]. However, I will at least illustrate the situation for the prototypical example. Approximating the potential $U(q)$ near a local minimum we obtain

$$U(q) = U(q_0) + \frac{1}{2} qWq + o(|q|^2), \tag{9.76}$$

where $W$ is a positive matrix and $U(q_0)$ can be chosen zero. Neglecting the higher order terms, the resulting model

$$H(p,q) = \frac{1}{2}(pMp + qWq) \tag{9.77}$$

is known as **harmonic oscillator**. Let $V$ be the (real) orthogonal matrix which transforms the symmetric matrix $M^{-1/2}WM^{-1/2}$ to diagonal form and let $\omega_j^2$ be the eigenvalues. Then the symplectic transform $(P,Q) = (VM^{1/2}p, VM^{-1/2}q)$ (Problem 9.11) gives the decoupled system

$$\dot{Q}_j = P_j, \qquad \dot{P}_j = -\omega_j^2 Q_j, \qquad j = 1, \dots, n. \tag{9.78}$$

In particular,

$$K(P,Q) = \sum_{j=1}^{n} K_j, \qquad K_j = \frac{1}{2}(P_j^2 + Q_j^2), \tag{9.79}$$

where the $K_j$'s are $n$ first integrals in involution (check this). The corresponding action-angle variables are given by (Problem 9.13)

$$I_j = \frac{1}{2}(\frac{P_j^2}{\omega_j} + \omega_j Q_j^2), \qquad \theta_j = \operatorname{arccot} \frac{P_j}{\omega_j Q_j}. \tag{9.80}$$

For example, consider the following Hamiltonian

$$H(p,q) = \sum_{j=1}^{n} \frac{p_j}{2m} + U_0\left(q_{j+1} - q_j\right), \quad q_0 = q_{n+1} = 0 \tag{9.81}$$

which describes a lattice of $n$ equal particles (with mass $m$) with nearest neighbor interaction described by the potential $U_0(x)$. The zeroth and $n$-th particle are considered fixed and $q_j$ is the displacement of the $j$-th particle from its equilibrium position. If we assume that the particles are coupled by springs, the potential would be $U_0(x) = \frac{k}{2}x^2$, where $k > 0$ is the so called spring constant, and we have a harmonic oscillator. The motion is decomposed into $n$ modes corresponding to the eigenvectors of the Jacobian of the potential. Physicists believed for a long time that a nonlinear perturbation of the force will lead to thermalization. That is, if the system starts in a certain mode of the linearized system, the energy will eventually be distributed equally over all modes. However, Fermi, Pasta, and Ulam showed with computer experiments that this is not true (Problem 9.14). This is related to the existence of solitons, see for example [**20**].

**Problem 9.10** (Symplectic group)**.** *Show that* $\mathrm{Sp}(2n)$ *is indeed a group. Suppose* $M \in \mathrm{Sp}(2n)$*, show that* $\det(M)^2 = 1$ *and* $\chi_M(z) = z^{2n}\chi_M(z^{-1})$.

**Problem 9.11.** *Show that the linear transformation* $(P, Q) = (Up, (U^{-1})^T q)$, *where* $U$ *is an arbitrary matrix, is canonical.*

**Problem 9.12.** *Show that the transformation generated by a function* $S$ *is canonical by directly proving that* $d\varphi$ *is symplectic. (Hint: Prove* $-Jd\varphi = Jd\psi^T$ *using*

$$\frac{\partial p}{\partial Q} = \frac{\partial^2 S_1}{\partial Q \partial q} = -\left(\frac{\partial P}{\partial q}\right)^T$$

*and similar for the others.)*

**Problem 9.13.** *Consider the harmonic oscillator in one dimension*

$$H(p, q) = \frac{1}{2}p^2 + \frac{\omega^2}{2}q^2$$

*and show that* $S_1(q, \theta) = \frac{\omega}{2}q^2 \cot(\theta)$ *generates a canonical transformation to action-angle variables.*

**Problem 9.14** (Fermi-Pasta-Ulam experiment)**.** *Consider the Hamiltonian (9.81) with the interaction potential* $U_0(x) = \frac{k}{2}(x^2 + \alpha x^3)$. *Note that it is no restriction to use* $m = k = 1$ *(why?).*

*Compute the eigenvalues and the eigenvectors of the linearized system* $\alpha = 0$. *Choose an initial condition in an eigenspace and (numerically) compute the time evolution. Investigate how the state is distributed with respect to the eigenvectors as a function of* $t$. *(Choose* $N = 32$, $\alpha = 1/6$.*)*

**Problem 9.15.** *Show that the Poisson bracket is a skew-symmetric bilinear form satisfying the* **Jacobi identity**

$$\{I, \{J, K\}\} + \{J, \{K, I\}\} + \{K, \{I, J\}\} = 0$$

*and* **Leibniz' rule**

$$\{I, J\,K\} = J\{I, K\} + K\{I, J\}.$$

**Problem 9.16** (Lax pair)**.** *Let* $L(p, q)$ *and* $P(p, q)$ *be* $n$ *by* $n$ *matrices. They are said to form a* **Lax pair** *for a Hamiltonian system if the equations of motion (9.42) are equivalent to the* **Lax equation**

$$\dot{L} = [P, L].$$

*Show that the quantities*

$$\mathrm{tr}(L^j), \qquad 1 \le j \le n,$$

*are first integrals (Hint: Compare Problem 3.18).*

## 9.5. The Kepler problem

Finally, as an application of our results we will show how to solve equation (1.11) from Section 1.1. In fact, we will even consider a slightly more general case, the **two body problem**. Suppose we have two masses placed at $x_1 \in \mathbb{R}^3$ and $x_2 \in \mathbb{R}^3$. They interact with a force $F$ depending only on the distance of the masses and lies on the line connecting both particles. The kinetic energy is given by

$$T(\dot{x}) = \frac{m_1}{2}\dot{x}_1^2 + \frac{m_2}{2}\dot{x}_2^2 \tag{9.82}$$

and the potential energy is

$$U(x) = U(|x_1 - x_2|). \tag{9.83}$$

The Lagrangian is the difference of both

$$L(\dot{x}, x) = T(\dot{x}) - U(x). \tag{9.84}$$

Clearly it is invariant under translations $(x_1, x_2) \mapsto (x_1+sa, x_2+sa)$, $a \in \mathbb{R}^3$, and so Theorem 9.7 tells us that all three components of the total momentum

$$m_1\dot{x}_1 + m_2\dot{x}_2 \tag{9.85}$$

are first integrals. Hence we will choose new coordinates

$$q_1 = \frac{m_1 x_1 + m_2 x_2}{m_1 + m_2}, \qquad q_2 = x_1 - x_2 \tag{9.86}$$

in which our Lagrangian reads

$$L(\dot{q}, q) = \frac{M}{2}\dot{q}_1^2 + \frac{\mu}{2}\dot{q}_2^2 - U(q_2), \qquad M = m_1 + m_2,\ \mu = \frac{m_1 m_2}{M}. \tag{9.87}$$

In particular, the system decouples and the solution of the first part is given by $q_1(t) = q_1(0) + \dot{q}_1(0)t$. To solve the second, observe that it is invariant under rotations and, invoking again Theorem 9.7, we infer that the **angular momentum**

$$l = \mu q_2 \wedge \dot{q}_2 \tag{9.88}$$

is another first integral. Hence we have found three first integrals and we suspect that our system is integrable. However, since

$$\{l_1, l_2\} = l_3, \quad \{l_1, l_3\} = -l_2, \quad \{l_2, l_3\} = l_1 \tag{9.89}$$

they are not in involution. But using $\{l, |l|^2\} = 0$ it is not hard to see

**Theorem 9.10.** *The two body problem is completely integrable. A full set of first integrals which are functionally independent and in involution is given by*

$$p_{11}, \quad p_{12}, \quad p_{13}, \quad \frac{\mu}{2}p_2^2 + U(q_2), \quad |l|^2, \quad l_3, \tag{9.90}$$

*where* $p_1 = M\dot{q}_1$ *and* $p_2 = \mu\dot{q}_2$.

Our next step would be to compute the action angle variables. But since this is quite cumbersome, we will use a more direct approach to solve the equation of motions. Since the motion is confined to the plane perpendicular to $l$ (once the initial condition has been chosen), it suggests itself to choose polar coordinates $(r,\varphi)$ in this plane. The angular momentum now reads

$$l_0 = |l| = \mu r^2 \dot{\varphi} \tag{9.91}$$

and conservation of energy implies

$$\frac{\mu}{2}\left(\dot{r}^2 + \frac{l_0^2}{\mu^2 r^2}\right) + U(r) = E. \tag{9.92}$$

Hence, $r(t)$ follows (implicitly) from

$$\dot{r} = \sqrt{\frac{2(E-U(r))}{\mu} - \frac{l_0^2}{\mu^2 r^2}} \tag{9.93}$$

via separation of variables. In case of the Kepler problem (gravitational force)

$$U(r) = -\frac{\gamma}{r} \tag{9.94}$$

it is possible to compute the integral, but not to solve for $r$ as a function of $t$. However, if one is only interested in the shape of the orbit one can look at $r = r(\varphi)$ which satisfies

$$\frac{1}{r^2}\frac{dr}{d\varphi} = \sqrt{\frac{2\mu(E-U(r))}{l_0^2} - \frac{1}{r^2}}. \tag{9.95}$$

The solution is given by (Problem 9.17)

$$r(\varphi) = \frac{p}{1-\varepsilon\cos(\varphi-\varphi_0)}, \qquad p = \frac{l_0^2}{\gamma\mu},\ \varepsilon = \sqrt{1 + \frac{2El_0^2}{\mu\gamma^2}} \tag{9.96}$$

Thus the orbit is an ellipsis if $\varepsilon < 1$, a parabola if $\varepsilon = 1$, and a hyperbola if $\varepsilon > 1$.

**Problem 9.17.** *Solve (9.95). (Hint: Use the transformation $\rho = r^{-1}$.)*

## 9.6. The KAM theorem

In the last section we were quite successful solving the two body problem. However, if we want to investigate the motion of planets around the sun under the influence of the gravitational force we need to consider the general **$N$-body problem** where the kinetic energy is given by

$$T(\dot{x}) = \sum_{j=1}^{N} \frac{m_j}{2}\dot{x}_j^2 \tag{9.97}$$

and the potential energy is

$$U(x) = \sum_{1 \le j < k \le N} U_{jk}(|x_j - x_k|). \tag{9.98}$$

In case of the gravitational force one has

$$U_{jk}(|x_j - x_k|) = \frac{m_j m_k}{|x_j - x_k|}. \tag{9.99}$$

However, whereas we could easily solve this problem for $N = 2$, this is no longer possible for $N \ge 3$. In fact, despite of the efforts of many astronomers and mathematicians, very little is known for this latter case.

The reason is of course that the $N$-body problem is no longer integrable for $N \ge 3$. In fact, it can be even shown that a *generic* Hamiltonian system (with more than one degree of freedom) is not integrable. So integrable systems are the exception from the rule. However, many interesting physical systems are nearly integrable systems. That is, they are small perturbations of integrable systems. For example, if we neglect the forces between the planets and only consider the attraction by the sun, the resulting system is integrable. Moreover, since the mass of the sun is much larger than those of the planets, the neglected term can be considered as a small perturbation.

This leads to the study of systems

$$H(p, q) = H_0(p, q) + \varepsilon\, H_1(p, q), \tag{9.100}$$

where $H_0$ is completely integrable and $\varepsilon$ is *small*. Since $H_0$ is integrable, we can choose corresponding action angle variables $(I, \theta)$ and it hence suffices to consider systems of the type

$$H(I, \theta) = H_0(I) + \varepsilon\, H_1(I, \theta), \tag{9.101}$$

where $I \in \mathbb{R}^n$ and all components of $\theta$ have to be taken modulo $2\pi$, that is, $\theta$ lives on the torus $\mathbb{T}^n$.

By (9.74) the unperturbed motion for $\varepsilon = 0$ is given by

$$I(t) = I_0, \qquad \theta(t) = \theta_0 + \Omega(I_0)t. \tag{9.102}$$

Hence the solution curve is a line winding around the invariant torus $\Gamma_{I_0} = \{I_0\} \times \mathbb{T}^n$. Such tori with a linear flow are called **Kronecker tori**. Two cases can occur.

If the frequencies $\Omega(I_0)$ are **nonresonant** or **rationally independent**,

$$k\Omega(I_0) \ne 0 \quad \text{for all } k \in \mathbb{Z}^n \backslash \{0\}, \tag{9.103}$$

then each orbit is dense. On the other hand, if the frequencies $\Omega(I_0)$ are **resonant**,

$$k\Omega(I_0) = 0 \quad \text{for some } k \in \mathbb{Z}^n \backslash \{0\}, \tag{9.104}$$

the torus can be decomposed into smaller ones with the same property as before.

The corresponding solutions are called **quasi-periodic**. They will be periodic if and only if all frequencies in $\Omega(I_0)$ are rationally dependent, that is,

$$\Omega(I_0) = k\omega \quad \text{for some } k \in \mathbb{Z}^n,\ \omega \in \mathbb{R}. \tag{9.105}$$

In case of the solar system such quasi-periodic solutions correspond to a stable motion (planets neither collide nor escape to infinity) and the question is whether they persist for small perturbations or not. Hence this problem is also known as "stability problem" for the solar system.

As noted by Kolmogorov *most* tori whose frequencies are nonresonant survive under small perturbations. More precisely, let $I \in D \subseteq \mathbb{R}^n$ and denote by $\Omega(D)$ the set of all possible frequencies for our system. Let $\Omega_\alpha(D)$ be the set of frequencies $\Omega$ satisfying the following **diophantine condition**

$$|k\Omega| \geq \frac{\alpha}{|k|^n} \quad \text{for all } k \in \mathbb{Z}^n\backslash\{0\}. \tag{9.106}$$

Then the following famous result by Kolmogorov, Arnold, and Moser holds

**Theorem 9.11** (KAM)**.** *Suppose $H_0$, $H_1$ are analytic on $D \times \mathbb{T}^n$ and $H_0$ is nondegenerate, that is,*

$$\det\left(\frac{\partial H_0}{\partial I}\right) \neq 0. \tag{9.107}$$

*Then there exists a constant $\delta > 0$ such that for*

$$|\varepsilon| < \delta\alpha^2 \tag{9.108}$$

*all Kronecker tori $\Gamma_I$ of the unperturbed system with $I \in \Omega_\alpha(D)$ persist as slightly deformed tori. They depend continuously on $I$ and form a subset of measure $O(\alpha)$ of the phase space $D \times \mathbb{T}^n$.*

The proof of this result involves what is know as "**small divisor**" problem and is beyond the scope of this manuscript. However, we will at least consider a simpler toy problem which illustrates some of the ideas and, in particular, explains where the diophantine condition (9.106) comes from. See the books by Arnold [**2**] or Moser [**19**] for further details and references.

But now we come to our toy problem. We begin with the system

$$\dot{x} = Ax, \qquad A = \begin{pmatrix} \mathrm{i}\omega_1 & & \\ & \ddots & \\ & & \mathrm{i}\omega_n \end{pmatrix}, \quad \omega_j \in \mathbb{R}, \tag{9.109}$$

where the solution is quasi-periodic and given by

$$x_j(t) = (\mathrm{e}^{At}c)_j = c_j \mathrm{e}^{\mathrm{i}\omega_j t}. \tag{9.110}$$

Next we perturb this system according to

$$\dot{x} = Ax + g(x), \tag{9.111}$$

where $g(x)$ has a convergent power series

$$g(x) = \sum_{|k|\geq 2} g_k x^k, \qquad k \in \mathbb{N}_0^n, \tag{9.112}$$

where $k = (k_1, \dots, k_n)$, $|k| = k_1 + \cdots + k_n$, and $x^k = x_1^{k_1} \cdots x_n^{k_n}$. For the solution of the perturbed system we can make the ansatz

$$x(t) = \sum_{|k|\geq 1} c_k \mathrm{e}^{\mathrm{i}\omega k\, t} \tag{9.113}$$

or equivalently

$$x(t) = u(\mathrm{e}^{At} c), \tag{9.114}$$

where

$$u(x) = x + \sum_{|k|\geq 2} u_k x^k. \tag{9.115}$$

Inserting this ansatz into (9.111) gives

$$\frac{\partial u}{\partial x}(x) Ax = Au(x) + g(u(x)), \tag{9.116}$$

that is,

$$\sum_{|k|\geq 2} (\omega k - A) u_k x^k = g(x + \sum_{|k|\geq 2} u_k x^k). \tag{9.117}$$

Comparing coefficients of $x^k$ shows that

$$(\mathrm{i}\omega k - A) u_k = \text{terms involving } u_\ell \text{ for } |\ell| < |k|. \tag{9.118}$$

Hence the coefficients $u_k$ can be determined recursively provided

$$\omega k - \omega_j \neq 0 \text{ for all } |k| \geq 2, 1 \leq j \leq n. \tag{9.119}$$

Next one needs to show that the corresponding series converges and it is clear that this will only be the case if the divisors $\omega k - \omega_j$ do not tend to zero too fast. In fact, it can be shown that this is the case if there are positive constants $\delta$, $\tau$ such that

$$|\omega k - \omega_j| \geq \frac{\delta}{|k|^\tau} \tag{9.120}$$

holds. Moreover, it can be shown that the set of frequencies $\omega$ satisfying (9.120) for some constants is dense and of full Lebesgue measure in $\mathbb{R}^n$.

An example which shows that the system is unstable if the frequencies are resonant is given in Problem 9.18.

**Problem 9.18.** *Consider*
$$g(x) = \begin{pmatrix} x_1^{k_1+1} x_2^{k_2} \\ 0 \end{pmatrix}, \qquad \omega_1 k_1 + \omega_2 k_2 = 0,$$
*and show that the associated system is unstable. (Hint: Bernoulli equation.)*

*Part 3*

# Chaos

Chapter 10

# Discrete dynamical systems

## 10.1. The logistic equation

This chapter gives a brief introduction to discrete dynamical systems. Most of the results are similar to the ones obtained for continuous dynamical systems. Moreover, they won't be needed until Chapter 11. We begin with a simple example.

Let $N(t)$ be the size of a certain species at time $t$ whose growth rate is proportional to the present amount, that is,

$$\dot{N}(t) = \kappa N(t). \tag{10.1}$$

The solution of this equation is clearly given by $N(t) = N_0 \exp(\kappa\, t)$. Hence the population grows exponentially if $\kappa > 0$ and decreases exponentially if $\kappa < 0$. Similarly, we could model this situation by a difference equation

$$N(n+1) - N(n) = kN(n) \tag{10.2}$$

or equivalently

$$N(n+1) = (1+k)N(n), \tag{10.3}$$

where $N(n)$ is now the population after $n$ time intervals (say years). The solution is given by $N(n) = N_0(1+k)^n$ and we have again exponential growth respectively decay according to the sign of $k > -1$. In particular, there is no big difference between the continuous and the discrete case and we even get the same results at $t = n$ if we set $\kappa = \ln(1+k)$.

However, this result can be quite misleading as the following example shows. A refined version of the above growth model is given by

$$\dot{N}(t) = \kappa N(t)(L - N(t)), \tag{10.4}$$

where the population is limited by a maximum $L$. It is not hard to see (e.g., by computing the solution explicitly), that for any positive initial population $N_0$, the species will eventually tend to the limiting population $L$. The discrete version reads

$$N(n+1) - N(n) = kN(n)(L - N(n)) \tag{10.5}$$

or equivalently

$$N(n+1) = kN(n)(\tilde{L} - N(n)), \qquad \tilde{L} = L + \frac{1}{k}. \tag{10.6}$$

Introducing $x_n = N(n)/\tilde{L}$, $\mu = k\tilde{L}$ we see that it suffices to consider

$$x_{n+1} = \mu x_n(1 - x_n), \tag{10.7}$$

which is known as the **logistic equation**. Introducing the quadratic function

$$L_\mu(x) = \mu x(1 - x) \tag{10.8}$$

you can formally write the solution as $n$-th iterate of this map, $x_n = L_\mu^n(x_0)$. But if you try to work out a closed expression for these iterates, you will soon find out that this is not as easy as in the continuous case. Moreover, the above difference equation leads to very complicated dynamics and is still not completely understood.

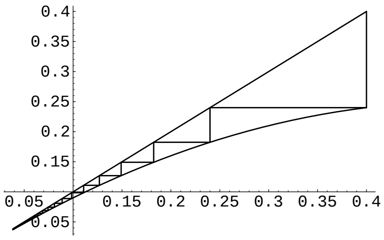
To get a first impression of this structure let us do some numerical experiments. We will consider $0 \le \mu \le 4$ in which case the interval $[0, 1]$ is mapped into itself under $f$.

First of all, we will use the following *Mathematica* code

```
In[1]:= ShowWeb[f_, xstart_, nmax_] :=
          Block[{x, xmin, xmax, graph, web},
            x[0] := xstart;
            x[n_] := x[n] = f[x[n - 1]];
            web = Flatten[Table[{{x[n], x[n]}, {x[n], x[n + 1]}},
              {n, 0, nmax}], 1];
            xmax = Max[web]; xmin = Min[web];
            graph = Plot[{f[x], x}, {x, xmin, xmax},
              DisplayFunction → Identity];
            Show[graph, Graphics[Line[web]],
                DisplayFunction → $DisplayFunction]
          ];
```

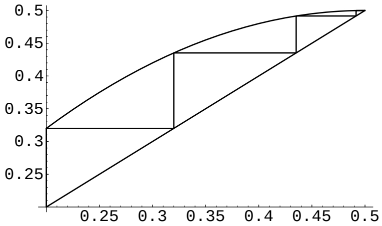
to visualize `nmax` iterations of a function $f(x)$ starting at `xstart`. If $\mu$ is small, say $\mu = 1$,

*In[2]:=* `ShowWeb[1#(1 − #)&, 0.4, 20];`



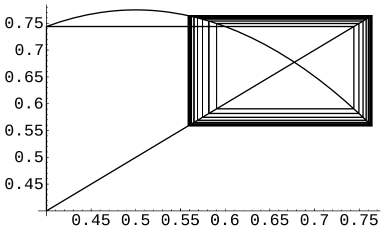

we see that all initial conditions in $(0, 1)$ eventually converge to 0 which is one solution of the fixed point equation $x = L_\mu(x)$. If $\mu$ increases beyond 1, it turns out that all initial conditions converge to the second solution $1 - \frac{1}{\mu}$ of the fixed point equation.

*In[3]:=* `ShowWeb[2#(1 − #)&, 0.2, 20];`




At $\mu = 3$ the behavior changes again and all initial conditions eventually jump back and forth between the two solutions of the equation $L_\mu^2(x) = x$ which are not solutions of $L_\mu(x) = x$.

*In[4]:=* `ShowWeb[3.1#(1 − #)&, 0.4, 20];`




Clearly this method of investigating the system gets quite cumbersome. We will return to this problem in Section 12.1.

**Problem 10.1.** *If the iteration converges, will the limit always be a fixed point?*

**Problem 10.2.** *Consider an m-th order* **difference equation**

$$x_{n+m} = F(n, x_n, \ldots, x_{n+m-1}). \tag{10.9}$$

*Show that it can be reduced to the iteration of a single map.*

## 10.2. Fixed and periodic points

Now let us introduce some notation for later use. To set the stage let $M$ be a metric space and let $f : M \to M$ be continuous. We are interested in investigating the dynamical system corresponding to the iterates

$$f^n(x) = f^{n-1}(f(x)), \quad f^0(x) = x. \tag{10.10}$$

In most cases $M$ will just be a subset of $\mathbb{R}^n$, however, the more abstract setting chosen here will turn out useful later on.

A point $p \in M$ satisfying

$$f(p) = p \tag{10.11}$$

is called a **fixed point** of $f$. Similarly, a fixed point of $f^n$,

$$f^n(p) = p, \tag{10.12}$$

is called a **periodic point** of period $n$. We will usually assume that $n$ is the **prime period** of $p$, that is, we have $f^m(p) \neq p$ for all $1 \le m < n$.

The forward **orbit** of $x$ is defined as

$$\gamma_+(x) = \{f^n(x) | n \in \mathbb{N}_0\}. \tag{10.13}$$

It is clearly positively invariant, that is, $f(\gamma_+(x)) \subseteq \gamma_+(x)$. An orbit for $x$ is a set of points

$$\gamma(x) = \{x_n | n \in \mathbb{Z} \text{ such that } x_0 = x,\ x_{n+1} = f(x_n)\}. \tag{10.14}$$

It is important to observe that the points $x_{-n}$, $n \in \mathbb{N}$, are not uniquely defined unless $f$ is one to one. Moreover, there might be no such points at all (if $f^{-1}(x) = \emptyset$ for some $x_n$). An orbit is invariant, that is, $f(\gamma(x)) = \gamma(x)$. The points $x_n \in \gamma(x)$ are also called a past history of $x$.

If $p$ is periodic with period $n$, then $\gamma_+(p)$ is finite and consists of precisely $n$ points

$$\gamma_+(p) = \{p, f(p), \dots, f^{n-1}(x)\}. \tag{10.15}$$

The converse is not true since a point might be **eventually periodic** (fixed), that is, it might be that $f^k(x)$ is periodic (fixed) for some $k$.

For example, if $M = \mathbb{R}$ and $f = 0$, then $p = 0$ is the only fixed point and every other point is eventually fixed.

A point $x \in M$ is called **forward asymptotic** to a periodic point $p$ of period $n$ if

$$\lim_{k\to\infty} f^{nk}(x) = p. \tag{10.16}$$

The **stable set** $W^+(p)$ is the set of all $x \in M$ for which (10.16) holds. Clearly, if $p_1$, $p_2$ are distinct periodic points, their stable sets are disjoint. In fact, if $x \in W^+(p_1) \cap W^+(p_2)$ we would have $\lim_{k\to\infty} f^{n_1 n_2 k}(x) = p_1 = p_2$, a contradiction. We call $p$ **attracting** if there is an open neighborhood $U$

of $p$ such that $U \subseteq W^+(p)$. The set $W^+(p)$ is clearly positively invariant (it is even invariant $f(W^+(p)) = W^+(p)$ if $f$ is invertible).

Similarly, a point $x \in M$ is called **backward asymptotic** to a periodic point $p$ of period $n$ if there is a past history $x_n$ of $x$ such that $\lim_{k\to\infty} x_{-nk}(x) = p$. The **unstable set** $W^-(p)$ is the set of all $x \in M$ for which this condition holds. Again unstable sets of distinct periodic points are disjoint. We call $p$ **repelling** if there is an open neighborhood $U$ of $p$ such that $U \subseteq W^-(p)$.

Note that if $p$ is repelling, every $x \in U$ will eventually leave $U$ under iterations. Nevertheless $x$ can still return to $U$ (Problem 10.5).

Note that if one point in the orbit $\gamma_+(p)$ of a periodic point $p$ is attracting (repelling) so are all the others (show this).

Now let us look at the logistic map $L_\mu(x) = \mu x(1-x)$ with $M = [0,1]$. We have already seen that if $\mu = 0$, then the only fixed point is $0$ with $W^+(0) = [0,1]$ and all points in $(0,1]$ are eventually periodic.

So let us next turn to the case $0 < \mu < 1$. Then we have $L_\mu(x) \le \mu x$ and hence $L_\mu^n(x) \le \mu^n x$ shows that every point converges exponentially to 0. In particular, we have $W^+(0) = [0,1]$.

Note that locally this follows since $L_\mu'(0) = \mu < 1$. Hence $L_\mu$ is contracting in a neighborhood of the fixed point and so all points in this neighborhood converge to the fixed point.

This result can be easily generalized to differentiable maps $f :\in C^1(U,U)$, where $U \subset \mathbb{R}^n$.

**Theorem 10.1.** *Suppose $f \in C^1(U,U)$, $U \subset \mathbb{R}^n$, then a periodic point $p$ with period $n$ is attracting if all eigenvalues of $d(f^n)_p$ are inside the unit circle and repelling if all eigenvalues are outside.*

**Proof.** In the first case there is a suitable norm such that $\|d(f^n)_p\| < \theta < 1$ for any fixed $\theta$ which is larger than all eigenvalues (Problem 3.3). Moreover, since the norm is continuous, there is an open ball $B$ around $p$ such that we have $\|d(f^n)_x\| \le \theta$ for all $x \in B$. Hence we have $|f^n(x) - p| = |f^n(x) - f^n(p)| \le \theta|x-p|$ and the claim is obvious.

The second case can now be reduced to the first by considering the local inverse of $f$ near $p$. □

If none of the eigenvalues of $d(f^n)$ at a periodic point $p$ lies on the unit circle, then $p$ is called **hyperbolic**. Note that by the chain rule the derivative is given by

$$d(f^n)(p) = \prod_{x\in\gamma_+(p)} df_x = df_{f^{n-1}(p)} \cdots df_{f(p)} df_p. \tag{10.17}$$

Finally, stability of a periodic point can be defined as in the case of differential equations. A periodic orbit $\gamma_+(p)$ of $f(x)$ is called **stable** if for any given neighborhood $U(\gamma_+(p))$ there exists another neighborhood $V(\gamma_+(p)) \subseteq U(\gamma_+(p))$ such that any point in $V(\gamma_+(p))$ remains in $U(\gamma_+(p))$ under all iterations. Note that this is equivalent to the fact that for any given neighborhood $U(p)$ there exists another neighborhood $V(p) \subseteq U(p)$ such that any point in $x \in V(p)$ satisfies $f^{nm}(x) \in U(p)$ for all $m \in \mathbb{N}_0$.

Similarly, a periodic orbit $\gamma_+(p)$ of $f(x)$ is called **asymptotically stable** if it is stable and attracting.

Pick a periodic point $p$ of $f$, $f^n(p) = p$, and an open neighborhood $U(p)$ of $p$. A **Liapunov function** is a continuous function

$$L : U(p) \to \mathbb{R} \tag{10.18}$$

which is zero at $p$, positive for $x \neq p$, and satisfies

$$L(x) \geq L(f^n(x)), \quad x, f^n(x) \in U(p) \backslash \{p\}. \tag{10.19}$$

It is called a **strict Liapunov function** if equality in (10.19) never occurs.

As in the case of differential equations we have the following analog of Liapunov's theorem (Problem 10.6).

**Theorem 10.2.** *Suppose $p$ is a periodic point of $f$. If there is a Liapunov function $L$, then $p$ is stable. If, in addition, $L$ is strict, then $p$ is asymptotically stable.*

**Problem 10.3.** *Consider the logistic map $L_\mu$ for $\mu = 1$. Show that $W^+(0) = [0, 1]$.*

**Problem 10.4.** *Determine the stability of all fixed points of the logistic map $L_\mu$, $0 \leq \mu \leq 4$.*

**Problem 10.5.** *Consider the logistic map $L_\mu$ for $\mu = 4$. show that 0 is a repelling fixed point. Find an orbit which is both forward and backward asymptotic to 0.*

**Problem 10.6.** *Prove Theorem 10.2.*

## 10.3. Linear difference equations

As in the case of differential equations, the behavior of nonlinear maps near fixed (periodic) points can be investigated by looking at the linearization. We begin with the study of the homogeneous linear first order difference equations

$$x(m + 1) = A(m)x(m), \qquad x(m_0) = x_0, \tag{10.20}$$

where $A(m) \in \mathbb{R}^n \times \mathbb{R}^n$. Clearly, the solution corresponding to $x(m_0) = x_0$ is given by

$$x(m, m_0, x_0) = \Pi(m, m_0)x_0, \tag{10.21}$$

where $\Pi(m, m_0)$ is the **principal matrix solution** given by

$$\Pi(m, m_0) = \prod_{j=m_0}^{m-1} A(j), \quad m \geq m_0. \tag{10.22}$$

In particular, linear combinations of solutions are again solutions and the set of all solutions forms an $n$-dimensional vector space.

The principal matrix solution solves the matrix valued initial value problem

$$\Pi(m+1, m_0) = A(m)\Pi(m, m_0), \qquad \Pi(m_0, m_0) = \mathbb{I} \tag{10.23}$$

and satisfies

$$\Pi(m, m_1)\Pi(m_1, m_0) = \Pi(m, m_0). \tag{10.24}$$

Moreover, if $A(m)$ is invertible for all $m$, we can set

$$\Pi(m, m_0) = \prod_{j=m}^{m_0-1} A(j)^{-1}, \quad m < m_0 \tag{10.25}$$

In this case, $\Pi(m, m_0)$ is an isomorphism with inverse given by $\Pi(m, m_0)^{-1} = \Pi(m_0, m)$ and all formulas from above hold for all $m$.

The analog of Liouville's formula is just the usual product rule for determinants

$$\det(\Pi(m, m_0)) = \prod_{j=m_0}^{m-1} \det(A(j)). \tag{10.26}$$

Finally, let us turn to the inhomogeneous system

$$x(m+1) = A(m)x(m) + g(m), \qquad x(m_0) = x_0, \tag{10.27}$$

where $A(m) \in \mathbb{R}^n \times \mathbb{R}^n$ and $g(m) \in \mathbb{R}^n$. Since the difference of two solutions of the inhomogeneous system (10.27) satisfies the corresponding homogeneous system (10.20), it suffices to find one particular solution. In fact, it is straight forward to verify that the solution is given by the following formula.

**Theorem 10.3.** *The solution of the inhomogeneous initial value problem is given by*

$$x(m) = \Pi(m, m_0)x_0 + \sum_{j=m_0}^{m-1} \Pi(m, j)g(j), \tag{10.28}$$

*where $\Pi(m, m_0)$ is the principal matrix solution of the corresponding homogeneous system.*

If $A(m)$ is invertible, the above formula also holds for $m < m_0$ if we set

$$x(m) = \Pi(m, m_0)x_0 - \sum_{j=m-1}^{m_0} \Pi(m, j)g(j), \quad m < m_0. \qquad (10.29)$$

**Problem 10.7.** *Find an explicit formula for the* **Fibonacci numbers** *defined via*

$$x(m) = x(m-1) + x(m-2), \qquad x(1) = x(2) = 1.$$

## 10.4. Local behavior near fixed points

In this section we want to investigate the local behavior of a differentiable map $f : \mathbb{R}^n \to \mathbb{R}^n$ near a fixed point $p$. We will assume $p = 0$ without restriction and write

$$f(x) = Ax + g(x), \qquad (10.30)$$

where $A = df_0$. The analogous results for periodic points are easily obtained by replacing $f$ with $f^n$.

First we show the Hartman-Grobman theorem for maps (compare Theorem 7.11).

**Theorem 10.4** (Hartman-Grobman)**.** *Suppose $f$ is a local diffeomorphism with hyperbolic fixed point 0. Then there is a homeomorphism $\varphi(x) = x + h(x)$, with bounded $h$, such that*

$$\varphi \circ A = f \circ \varphi, \qquad A = df_0, \qquad (10.31)$$

*in a sufficiently small neighborhood of* 0*.*

**Proof.** Let $\phi_\delta$ be a smooth bump function such that $\phi_\delta(x) = 0$ for $|x| \le \delta$ and $\phi_\delta(x) = 1$ for $|x| \ge 2\delta$. Then the function $g_\delta = (1 - \varphi_\delta)(f - A)$ satisfies the assumptions of Lemma 7.9 (show this) for $\delta$ sufficiently small. Since $f$ and $f_\delta$ coincide for $|x| \le \delta$ the homeomorphism for $f_\delta$ is also the right one for $f$ for $x$ in the neighborhood $\varphi^{-1}(\{x| \, |x| \le \delta\})$. □

Let me emphasize that the homeomorphism $\varphi$ is in general not differentiable! In particular, this shows that the stable and unstable sets $W^+(0)$ and $W^-(0)$ (defined in Section 10.2) are given (locally) by homeomorphic images of the corresponding linear ones $E^+(A)$ and $E^-(A)$, respectively. In fact, it can even be shown that (in contradistinction to $\varphi$) they are differentiable manifolds as we will see in a moment.

We will assume that $f$ is a local diffeomorphism for the rest of this section.

We define the **stable** respectively **unstable manifolds** of a fixed point $p$ to be the set of all points which converge exponentially to $p$ under iterations of $f$ respectively $f^{-1}$, that is,

$$M^{\pm}(p) = \{x \in M | \sup_{\pm m \in \mathbb{N}_0} \alpha^{\pm m} |f^m(x) - p| < \infty \text{ for some } \alpha \in (0,1)\}. \tag{10.32}$$

Both sets are obviously invariant under the flow In particular and are called the **stable** and **unstable manifold** of $p$.

It is no restriction to assume that $p = 0$. In the linear case we clearly have $M^{\pm}(0) = E^{\pm}(A)$.

Our goal is to show, the sets $M^{\pm}(x_0)$ are indeed manifolds (smooth) tangent to $E^{\pm}(A)$. As in the continuous case, the key idea is to formulate our problem as a fixed point equation which can then be solved by iteration.

Now writing

$$f(x) = Ax + g(x) \tag{10.33}$$

our difference equation can be rephrased as

$$x(m) = A^m x_0 + \sum_{j=0}^{m-1} A^{m-j} g(x(j)) \tag{10.34}$$

by Theorem 10.3.

Next denote by $P^{\pm}$ the projectors onto the stable, unstable subspaces $E^{\pm}(A)$. Moreover, abbreviate $x_{\pm} = P^{\pm} x_0$ and $g_{\pm}(x) = P^{\pm} g(x)$.

What we need is a condition on $x_0 = x_+ + x_-$ such that $x(m)$ remains bounded. If we project out the unstable part of our summation equation

$$x_- = A^{-m} x_-(m) - \sum_{j=0}^{m-1} A^j g_-(x(j)). \tag{10.35}$$

and suppose $|x(m)|$ bounded for $m \geq 0$, we can let $m \to \infty$,

$$x_- = -\sum_{j=0}^{\infty} A^{-j} g_-(x(j)), \tag{10.36}$$

where the sum converges since the summand decays exponentially. Plugging this back into our equation and introducing $P(m) = P^+$, $m > 0$, respectively $P(m) = -P^-$, $m \leq 0$, we arrive at

$$x(m) = K(x)(m), \quad K(x)(m) = A^m x_+ + \sum_{j=0}^{\infty} A^{m-j} P(m-j) g(x(j)). \tag{10.37}$$

To solve this equation by iteration, suppose $|x(m)| \leq \delta$, then since the Jacobian of $g$ at 0 vanishes, we have

$$\sup_{m\geq 0} |g(x(m)) - g(\tilde{x}(m))| \leq \varepsilon \sup_{m\geq 0} |x(m) - \tilde{x}(m)|, \tag{10.38}$$

where $\varepsilon$ can be made arbitrarily small by choosing $\delta$ sufficiently small. Since we have

$$\|A^{m-j} P(m-j)\| \leq C\alpha^{|m-j|}, \quad \alpha < 1. \tag{10.39}$$

existence of a solution follows by Theorem 2.1. Proceeding as in the case of differential equations we obtain

**Theorem 10.5** (Stable manifold)**.** *Suppose $f \in C^k$ has a fixed point $p$ with corresponding invertible Jacobian $A$. Then, there is a neighborhood $U(p)$ and functions $h^{\pm} \in C^k(E^{\pm}(A), E^{\mp}(A))$ such that*

$$M^{\pm}(x_0) \cap U(p) = \{p + a + h^{\pm}(a) | a \in E^{\pm} \cap U\}. \tag{10.40}$$

*Both $h^{\pm}$ and their Jacobians vanish at $p$, that is, $M^{\pm}(p)$ are tangent to their respective linear counterpart $E^{\pm}(A)$ at $p$. Moreover,*

$$|f^{\pm m}(x) - p| \leq C\alpha^{\pm m}, m \in \mathbb{N}_0, x \in M^{\pm}(p) \tag{10.41}$$

*for any $\alpha < \min\{|\alpha|\, |\alpha \in \sigma(A_+) \cup \sigma(A_-)^{-1}\}$ and some $C > 0$ depending on $\alpha$.*

**Proof.** The proof is similar to the case of differential equations. The details are left to the reader. □

In the hyperbolic case we can even say a little more.

**Theorem 10.6.** *Suppose $f \in C^k$ has a hyperbolic fixed point $p$ with invertible Jacobian. Then there is a neighborhood $U(p)$ such that $\gamma_{\pm}(x) \subset U(p)$ if and only if $x \in M^{\pm}(p)$. In particular,*

$$W^{\pm}(p) = M^{\pm}(p). \tag{10.42}$$

**Proof.** The proof again follows as in the case of differential equations. □

It happens that an orbit starting in the unstable manifold of one fixed point $p_0$ ends up in the stable manifold of another fixed point $p_1$. Such an orbit is called **heteroclinic orbit** if $p_0 \neq p_1$ and **homoclinic orbit** if $p_0 = p_1$.

Note that the same considerations apply to fixed points if we replace $f$ by $f^n$.

Chapter 11

# Periodic solutions

## 11.1. Stability of periodic solutions

In Section 6.4 we have defined stability for a fixed point. In this section we want to extend this notation to periodic solutions.

An orbit $\gamma(x_0)$ is called **stable** if for any given neighborhood $U(\gamma(x_0))$ there exists another neighborhood $V(\gamma(x_0)) \subseteq U(\gamma(x_0))$ such that any solution starting in $V(\gamma(x_0))$ remains in $U(\gamma(x_0))$ for all $t \geq 0$.

Similarly, an orbit $\gamma(x_0)$ is called **asymptotically stable** if it is stable and if there is a neighborhood $U(\gamma(x_0))$ such that

$$\lim_{t\to\infty} d(\Phi(t,x), \gamma(x_0)) = 0 \quad \text{for all } x \in U(x_0). \tag{11.1}$$

Here $d(x,U) = \sup_{y\in U} |x-y|$.

Note that this definition ignores the time parametrization of the orbit. In particular, if $x$ is close to $x_1 \in \gamma(x_0)$, we do *not* require that $\Phi(t,x)$ stays close to $\Phi(t,x_1)$ (we only require that it stays close to $\gamma(x_0)$). To see that this definition is the right one, consider the mathematical pendulum (6.48). There all orbits are periodic, but the period is not the same. Hence, if we fix a point $x_0$, any point $x \neq x_0$ starting close will have a slightly larger respectively smaller period and thus $\Phi(t,x)$ does not stay close to $\Phi(t,x_0)$. Nevertheless, it will still stay close to the orbit of $x_0$.

But now let us turn to the investigation of the stability of periodic solutions. Suppose the differential equation

$$\dot{x} = f(x) \tag{11.2}$$

has a periodic solution $\Phi(t,x_0)$ of period $T = T(x_0)$.

Since linearizing the problem was so successful for fixed points, we will try to use a similar approach for periodic points. Abbreviating the linearization of $f$ along the periodic orbit by

$$A(t) = df_{\Phi(t,x_0)}, \qquad A(t+T) = A(t), \tag{11.3}$$

or problem suggests to investigate the **first variational equation**

$$\dot{y} = A(t)y, \tag{11.4}$$

which we already encountered in (2.39). Note that choosing a different point of the periodic orbit $x_0 \to \Phi(s, x_0)$ amounts to $A(t) \to A(t+s)$.

Our goal is to show that stability of the periodic orbit $\gamma(x_0)$ is related to stability of the first variational equation. As a first useful observation we note that the corresponding principal matrix solution $\Pi(t, t_0)$ can be obtained by linearizing the flow along the periodic orbit.

**Lemma 11.1.** *The principal matrix solution of the first variational equation is given by*

$$\Pi_{x_0}(t, t_0) = \frac{\partial \Phi_{t-t_0}}{\partial x}(\Phi(t_0, x_0)). \tag{11.5}$$

*Moreover,* $f(\Phi(t, x_0))$ *is a solution of the first variational equation*

$$f(\Phi(t, x_0)) = \Pi_{x_0}(t, t_0) f(\Phi(t_0, x_0)). \tag{11.6}$$

**Proof.** Abbreviate $J(t,x) = \frac{\partial \Phi_t}{\partial x}(x)$, then $J(0,x) = \mathbb{I}$ and by interchanging $t$ and $x$ derivatives it follows that $\dot{J}(t,x) = df_{\Phi(t,x)} J(t,x)$. Hence $J(t - t_0, \Phi(t_0, x_0))$ is the principal matrix solution of the first variational equation. It remains to show that (11.6) satisfies the first variational equation which is a straightforward calculation. $\square$

Since $A(t)$ is periodic, all considerations of Section 3.4 apply. In particular, the principal matrix solution is of the form

$$\Pi_{x_0}(t, t_0) = P_{x_0}(t, t_0) \exp((t - t_0) Q_{x_0}(t_0)) \tag{11.7}$$

and the monodromy matrix $M_{x_0}(t_0) = \exp(T Q_{x_0}(t_0)) = \frac{\partial \Phi_{T-t_0}}{\partial x}(\Phi(t_0, x_0))$ has eigenvalues independent of the point in the orbit chosen. Note that one of the eigenvalues is one, since

$$M_{x_0}(t_0) f(\Phi(t_0, x_0)) = f(\Phi(t_0, x_0)). \tag{11.8}$$

## 11.2. The Poincaré map

Recall the Poincaré map

$$P_\Sigma(y) = \Phi(\tau(y), y) \tag{11.9}$$

introduced in Section 6.3. It is one of the major tools for investigating periodic orbits. Stability of the periodic orbit $\gamma(x_0)$ is directly related to stability of $x_0$ as a fixed point of $P_\Sigma$.

**Lemma 11.2.** *The periodic orbit $\gamma(x_0)$ is an (asymptotically) stable orbit of $f$ if and only if $x_0$ is an (asymptotically) stable fixed point of $P_\Sigma$.*

**Proof.** Suppose $x_0$ is a stable fixed point of $P_\Sigma$. Let $U$ be a neighborhood of $\gamma(x_0)$. Choose a neighborhood $\tilde{U} \subseteq U \cap \Sigma$ of $x_0$ such that $\Phi([0,T],\tilde{U}) \subseteq U$. If $x_0$ is a stable fixed point of $P_\Sigma$ there is another neighborhood $\tilde{V} \subseteq \Sigma$ of $x_0$ such that $P^n(\tilde{V}) \subseteq \tilde{U}$ for all $n$. Now let $V$ be a neighborhood of $\gamma(x_0)$ such that $V \subseteq \Phi([0,T],\tilde{V})$. Then if $y \in V$ there is a smallest $t_0 \geq 0$ such that $y_0 = \Phi(t_0, y) \in \tilde{V}$. Hence $y_n = P_\Sigma^n(y_0) \in \tilde{U}$ and thus $\phi(t,V) \subseteq U$ for all $t \geq 0$.

Moreover, if $y_n \to x_0$ then $\Phi(t,y) \to \gamma(x_0)$ by continuity of $\Phi$ and compactness of $[0,T]$. Hence $\gamma(x_0)$ is asymptotically stable if $x_0$ is. The converse is trivial. □

As an immediate consequence of this result and Theorem 10.1 we obtain

**Corollary 11.3.** *Suppose $f \in C^k$ has a periodic orbit $\gamma(x_0)$. If all eigenvalues of the Poincaré map lie inside the unit circle then the periodic orbit is asymptotically stable.*

We next show how this approach is related to the first variational equation.

**Theorem 11.4.** *The eigenvalues of the derivative of the Poincaré map $dP_\Sigma$ at $x_0$ plus the single value 1 coincide with the eigenvalues of the monodromy matrix $M_{x_0}(t_0)$.*

*In particular, the eigenvalues of the Poincaré map are independent of the base point $x_0$ and the transversal arc $\Sigma$.*

**Proof.** After a linear transform it is no restriction to assume $f(x_0) = (0,\dots,0,1)$. Write $x = (y,z) \in \mathbb{R}^{n-1} \times \mathbb{R}$. Then $\Sigma$ is locally the graph of a function $s : \mathbb{R}^{n-1} \to \mathbb{R}$ and we can take $y$ as local coordinates for the Poincaré map. Since

$$\frac{\partial}{\partial x}\Phi(\tau(x),x)\Big|_{x=x_0} = f(x_0)d\tau_{x_0} + \frac{\partial \Phi_T}{\partial x}(x_0) \tag{11.10}$$

we infer $dP_\Sigma(x_0)_{j,k} = M_{x_0}(t_0)_{j,k}$ for $1 \leq j,k \leq n-1$ by Lemma 11.1. Moreover, $M_{x_0}(0)f(x_0) = f(x_0)$ and thus

$$M_{x_0}(0) = \begin{pmatrix} dP_\Sigma(x_0) & 0 \\ m & 1 \end{pmatrix} \tag{11.11}$$

from which the claim is obvious. □

As a consequence we obtain

**Corollary 11.5.** *The determinants of the derivative of the Poincaré map at $x_0$ and of the monodromy matrix are equal*

$$\det(dP_\Sigma(x_0)) = \det(M_{x_0}(t_0)). \tag{11.12}$$

*In particular, since the determinant of the monodromy matrix does not vanish, $P_\Sigma(y)$ is a local diffeomorphism at $x_0$.*

By Liouville's formula (3.58) we have

$$\det(M_{x_0}(t_0)) = \exp\left(\int_0^T \mathrm{tr}(A(t))\,dt\right) = \exp\left(\int_0^T \mathrm{div}(f(\Phi(t,x_0))\,dt\right). \tag{11.13}$$

In two dimensions there is only one eigenvalue which is equal to the determinant and hence we obtain

**Lemma 11.6.** *Suppose $f$ is a planar vector field. Then a periodic point $x_0$ is asymptotically stable if*

$$\int_0^T \mathrm{div}(f(\Phi(t,x_0))\,dt < 0 \tag{11.14}$$

*and unstable if the integral is positive.*

As another application of the use of the Poincaré map we will show that hyperbolic periodic orbits persist under small perturbations.

**Lemma 11.7.** *Let $f(x,\lambda)$ be $C^k$ and suppose $f(x,0)$ has a hyperbolic periodic orbit $\gamma(x_0)$. Then, in a sufficiently small neighborhood of 0 there is a $C^k$ map $\lambda \mapsto x_0(\lambda)$ such that $x_0(0) = x_0$ and $\gamma(x_0(\lambda))$ is a periodic orbit of $f(x,\lambda)$.*

**Proof.** Fix a transversal arc $\Sigma$ for $f(x,0)$ at $x_0$. That arc is also transversal for $f(x,\lambda)$ with $\lambda$ sufficiently small. Hence there is a corresponding Poincaré map $P_\Sigma(x,\varepsilon)$ (which is $C^k$). Since $P_\Sigma(x_0,0) = x_0$ and no eigenvalue of $P_\Sigma(x,0)$ lies on the unit circle the result follows from the implicit function theorem. □

## 11.3. Stable and unstable manifolds

To show that the stability of a periodic point $x_0$ can be read off from the first variational equation, we will first simplify the problem by applying some transformations.

Using $y(t) = x(t) - \Phi(t,x_0)$ we can reduce it to the problem

$$\dot{y} = \tilde{f}(t,y), \qquad \tilde{f}(t,y) = f(y + \Phi(t,x_0)) - f(\Phi(t,x_0)), \tag{11.15}$$

where $\tilde{f}(t, 0) = 0$ and $\tilde{f}(t + T, x) = \tilde{f}(t, x)$. This equation can be rewritten as

$$\dot{y} = A(t)y + \tilde{g}(t, y) \tag{11.16}$$

with $\tilde{g}$ $T$-periodic, $\tilde{g}(t, 0) = 0$, and $(\partial g/\partial y)(t, 0) = 0$.

We will see that hyperbolic periodic orbits are quite similar to hyperbolic fixed points. (You are invited to show that this definition coincides with our previous one for fixed points in the special case $T = 0$.)

Moreover, by Corollary 3.10 the transformation $z(t) = P(t)^{-1}y(t)$ will transform the system to

$$\dot{z} = Qz + g(t, z). \tag{11.17}$$

Hence we can proceed as in Section 7.2 to show the existence of stable and unstable manifolds at $x_0$ defined as

$$M^{\pm}(x_0) = \{x \in M| \sup_{\pm t \geq 0} \mathrm{e}^{\pm\gamma t}|\Phi(t, x) - \Phi(t, x_0)| < \infty \text{ for some } \gamma > 0\}. \tag{11.18}$$

Making this for different points $\Phi(t_0, x_0)$ in our periodic orbit we set

$$M^{\pm}_{t_0}(x_0) = M^{\pm}(\Phi(t_0, x_0)). \tag{11.19}$$

Note that the linear counterparts are the linear subspaces

$$E^{\pm}(t_0) = \Pi_{x_0}(t_1, 0)E^{\pm}(0) \tag{11.20}$$

corresponding to the stable and unstable subspace of $M_{x_0}(t_0)$ (compare (3.76)).

**Theorem 11.8** (Stable manifold for periodic orbits). *Suppose $f \in C^k$ has a hyperbolic periodic orbit $\gamma(x_0)$ with corresponding monodromy matrix $M(t_0)$.*

*Then, there is a neighborhood $U(\gamma(x_0))$ and functions $h^{\pm} \in C^k([0, T] \times E^{\pm}, E^{\mp})$ such that*

$$M^{\pm}_{t_0}(x_0) \cap U(\gamma(x_0)) = \{\Phi(t_0, x_0) + a + h^{\pm}(t_0, a)|a \in E^{\pm}(t_0) \cap U\}. \tag{11.21}$$

*Both $h^{\pm}(t_0, .)$ and their Jacobians vanish at $x_0$, that is, $M^{\pm}_{t_0}(x_0)$ are tangent to their respective linear counterpart $E^{\pm}(t_0)$ at $\Phi(t_0, x_0)$. Moreover,*

$$|\Phi(t, x) - \Phi(x_0, t + t_0)| \leq C\mathrm{e}^{\mp t\gamma}, \pm t \geq 0, x \in M^{\pm}_{t_0}(x_0) \tag{11.22}$$

*for any $\gamma < \min\{|\mathrm{Re}(\gamma_j)|\}_{j=1}^{m}$ and some $C > 0$ depending on $\gamma$. Here $\gamma_j$ are the eigenvalues of $Q(t_0)$.*

**Proof.** As already pointed out before, the same proof as in Section 7.2 applies. The only difference is that $g$ now depends on $t$. However, since $g$ is periodic we can restrict $t$ to the compact interval $[0, T]$ for all estimates and no problems arise. Hence we get $M^{\pm}_{t_0}$ for each point in the orbit.

Parametrizing each point by $t_0 \in [0,T]$ it is not hard to see that $g$ is $C^k$ as a function of this parameter. Moreover, by (11.20), so are the stable and unstable subspaces of the monodromy matrix $M(t_0)$. □

Now we can take the union over all $t_0$ and define

$$\begin{aligned} M^{\pm}(\gamma(x_0)) &= \\ &= \{x \mid \sup_{\pm t \geq 0} \mathrm{e}^{\pm\gamma t}|\Phi(t,x) - \Phi(t+t_0,x_0)| < \infty \text{ for some } t_0, \gamma > 0\} \\ &= \bigcup_{t_0 \in [0,T]} M^{\pm}_{t_0}(x_0). \end{aligned} \tag{11.23}$$

as the **stable and unstable manifold**, respectively. They are clearly invariant under the flow and are locally given by

$$\begin{aligned} &M^{\pm}(\gamma(x_0)) \cap U(\gamma(x_0)) = \\ &\quad \{\Phi(t_0,x_0) + \Pi_{x_0}(t_0,0)a + h^{\pm}(t_0, \Pi_{x_0}(t_0,0)a)| \\ &\qquad a \in E^{\pm}(0) \cap U,\ t_0 \in [0,T]\}. \end{aligned} \tag{11.24}$$

The points in $M^{\pm}(\gamma(x_0))$ are said to have an **asymptotic phase**, that is, there is a $t_0$ such that

$$\Phi(t,x) \to \Phi(t+t_0,x_0) \quad \text{as} \quad t \to \infty \text{ or } t \to -\infty. \tag{11.25}$$

As in the case of a fixed point, the (un)stable manifold coincides with the (un)stable set

$$W^{\pm}(\gamma(x_0)) = \{x \mid \lim_{t \to \pm\infty} d(\Phi(t,x), \gamma(x_0)) = 0\} \tag{11.26}$$

of $\gamma(x_0)$ if the orbit is hyperbolic.

**Theorem 11.9.** *Suppose $f \in C^k$ has a hyperbolic periodic orbit $\gamma(x_0)$. Then there is a neighborhood $U(x_0)$ such that $\gamma_{\pm}(x) \subset U(\gamma(x_0))$ if and only if $x \in M^{\pm}(\gamma(x_0))$. In particular,*

$$W^{\pm}(\gamma(x_0)) = M^{\pm}(\gamma(x_0)). \tag{11.27}$$

**Proof.** Suppose $d(\Phi(t,x), \gamma(x_0)) \to 0$ as $t \to \infty$. Note that it is no restriction to assume that $x$ is sufficiently close to $\gamma(x_0)$. Choose a transversal arc $\Sigma$ containing $x$ and consider the corresponding Poincaré map $P_\Sigma$. Then $M^{\pm}(\gamma(x_0)) \cap \Sigma$ must be the stable and unstable manifolds of the Poincaré map. By the Hartman-Grobman theorem for flows, $x$ must lie on the stable manifold of the Poincaré map and hence it lies in $M^{\pm}(\gamma(x_0))$. □

Moreover, if $f$ depends on a parameter $\lambda$, then we already know that a hyperbolic periodic orbit persists under small perturbations and depends smoothly on the parameter by Lemma 11.7. Moreover, the same is true for the stable and unstable manifolds (which can be proven as in Theorem 7.8).

**Theorem 11.10.** *Let $f(x,\lambda)$ be $C^k$ and suppose $f(x,0)$ has a hyperbolic periodic orbit $\gamma(x_0)$. Then, in a sufficiently small neighborhood of* 0 *there is a $C^k$ map $\lambda \mapsto x_0(\lambda)$ such that $x_0(0) = x_0$ and $\gamma(x_0(\lambda))$ is a periodic orbit of $f(x,\lambda)$. Moreover, the corresponding stable and unstable manifolds are locally given by*

$$M^{\pm}(\gamma(x_0(\lambda))) \cap U(\gamma(x_0(\lambda))) = \{\Phi(t_0, x_0(\lambda), \lambda) + a(\lambda) + h^{\pm}(t_0, a(\lambda)) | \\ a \in E^{\pm}(0) \cap U,\ t_0 \in [0,T]\}, \tag{11.28}$$

*where $a(\lambda) = \Pi_{x_0(\lambda)}(t_0, 0, \lambda) P^{\pm}(\lambda) a$, $h^{\pm} \in C^k$.*

**Problem 11.1** (Hopf bifurcation)**.** *Investigate the system*

$$\dot{x} = -y + (\mu + \sigma(x^2 + y^2)x, \qquad \dot{y} = x + (\mu + \alpha(x^2 + y^2)y$$

*as a function of the parameter $\mu$ for $\sigma = 1$ and $\sigma = -1$. Compute the stable and unstable manifolds in each case. (Hint: Use polar coordinates.)*

## 11.4. Melnikov's method for autonomous perturbations

In Lemma 11.7 we have seen that hyperbolic periodic orbits are stable under small perturbations. However, there is a quite frequent situations in applications where this result is not good enough! In Section 6.6 we have learned that many physical models are given as Hamiltonian systems. Clearly such systems are idealized and a more realistic model can be obtained by perturbing the original one a little. This will usually render the equation unsolvable. The typical situation for a Hamiltonian system in two dimensions is that there is a fixed point surrounded by periodic orbits. As we have seen in Problem 6.16, adding an (arbitrarily small) friction term will render the fixed point asymptotically stable and all periodic orbits disappear. In particular, the periodic orbits are unstable under small perturbations and hence cannot be hyperbolic. On the other hand, van der Pol's equation (8.26) is also Hamiltonian for $\mu = 0$ and in Theorem 8.16 we have shown that one of the periodic orbits persists for $\mu > 0$.

So let us consider a Hamiltonian system

$$H(p,q) = \frac{p^2}{2} + U(q), \tag{11.29}$$

with corresponding equation of motions

$$\dot{p} = -U'(q), \quad \dot{q} = p. \tag{11.30}$$

Moreover, let $q_0$ be an equilibrium point surrounded by periodic orbits. Without restriction we will choose $q_0 = 0$. We are interested in the fate of these periodic orbits under a small perturbation

$$\dot{p} = -U'(q) + \varepsilon f(p,q), \quad \dot{q} = p + \varepsilon g(p,q), \tag{11.31}$$

which is not necessarily Hamiltonian. Choosing the section $\Sigma = \{(0,q)|q > 0\}$, the corresponding Poincaré map is given by

$$P_\Sigma((0,q),\varepsilon) = \Phi(\tau(q,\varepsilon),(0,q),\varepsilon), \tag{11.32}$$

where $\tau(q,\varepsilon)$ is the first return time. The orbit starting at $(0,q)$ will be periodic if and only if $q$ is a zero of the displacement function

$$\Delta(q,\varepsilon) = \Phi_1(\tau(q,\varepsilon),(0,q),\varepsilon) - q. \tag{11.33}$$

Since $\Delta(q,0)$ vanishes identically, so does the derivative with respect to $q$ and hence we cannot apply the implicit function theorem. Of course this just reflects the fact that the periodic orbits are not hyperbolic and hence was to be expected from the outset.

The way out of this dilemma is to consider the reduced displacement function $\tilde{\Delta}(q,\varepsilon) = \varepsilon^{-1}\Delta(q,\varepsilon)$ (which is as good as the original one for our purpose). Now $\tilde{\Delta}(q,0) = \Delta_\varepsilon(q,0)$ and $\tilde{\Delta}_q(q,0) = \Delta_{\varepsilon,q}(q,0)$. Thus, if we find a simple zero of $\Delta_\varepsilon(q,0)$, then the implicit function theorem applied to $\tilde{\Delta}(q,\varepsilon)$ tells us that the corresponding periodic orbit persists under small perturbations.

Well, whereas this might be a nice result, it is still of no use unless we can compute $\Delta_\varepsilon(q,0)$ somehow. Abbreviate

$$(p(t,\varepsilon), q(t,\varepsilon)) = \Phi(t,(0,q),\varepsilon), \tag{11.34}$$

then

$$\begin{aligned} \frac{\partial}{\partial\varepsilon}\Delta(q,\varepsilon)\Big|_{\varepsilon=0} &= \frac{\partial}{\partial\varepsilon}q(\tau(q,\varepsilon),\varepsilon)\Big|_{\varepsilon=0} = \dot{q}(T(q),0)\tau_\varepsilon(q,0) + q_\varepsilon(T(q),0) \\ &= p(T(q),0)\tau_\varepsilon(q,0) + q_\varepsilon(T(q),0) = q_\varepsilon(T(q),0), \end{aligned} \tag{11.35}$$

where $T(q) = \tau(q,0)$ is the period of the unperturbed orbit. Next, observe that $(p_\varepsilon(t), q_\varepsilon(t)) = \frac{\partial}{\partial\varepsilon}(p(t,\varepsilon), q(t,\varepsilon))|_{\varepsilon=0}$ is the solution of the first variational equation

$$\dot{p}_\varepsilon(t) = -U''(q_\varepsilon(t))q_\varepsilon(t) + f(p(t),q(t)), \quad \dot{q}_\varepsilon(t) = p_\varepsilon(t) + g(p(t),q(t)) \tag{11.36}$$

corresponding to the initial conditions $(p_\varepsilon(t), q_\varepsilon(t)) = (0,0)$. Here we have abbreviated $(p(t),q(t)) = (p(t,0),q(t,0))$. By the variation of constants formula the solution is given by

$$\begin{pmatrix} p_\varepsilon(t) \\ q_\varepsilon(t) \end{pmatrix} = \int_0^t \Pi_q(t,s) \begin{pmatrix} f(p(s),q(s)) \\ g(p(s),q(s)) \end{pmatrix} ds. \tag{11.37}$$

We are only interested in the value at $t = T(q)$, where

$$\Pi_q(T(q),s) = \Pi_q(T(q),0)\Pi_q(0,s) = \Pi_q(T(q),0)\Pi_q(s,0)^{-1}. \tag{11.38}$$

Furthermore, using Lemma 11.1,

$$\Pi_q(t,0)\begin{pmatrix} -U'(q) \\ 0 \end{pmatrix} = \begin{pmatrix} -U'(q(t)) \\ p(t) \end{pmatrix} \tag{11.39}$$

and we infer

$$\Pi_q(t,0) = \frac{1}{U'(q)} \begin{pmatrix} U'(q(t)) & -\alpha(t)U'(q(t)) + \beta(t)p(t) \\ -p(t) & \alpha(t)p(t) + \beta(t)U'(q(t)) \end{pmatrix}, \tag{11.40}$$

where $\alpha(t)$ and $\beta(t)$ are given by

$$\Pi_q(t,0) \begin{pmatrix} 0 \\ U'(q) \end{pmatrix} = \alpha(t) \begin{pmatrix} -U'(q(t)) \\ p(t) \end{pmatrix} + \beta(t) \begin{pmatrix} p(t) \\ U'(q(t)) \end{pmatrix}. \tag{11.41}$$

Moreover, by Liouville's formula we have $\det \Pi_q(t,s) = 1$ and hence

$$\beta(t) = \frac{U'(q)^2}{U'(q(t))^2 + p(t)^2} \det \Pi_q(t,0) = \frac{U'(q)^2}{U'(q(t))^2 + p(t)^2}. \tag{11.42}$$

Now putting everything together we obtain

$$\Delta_\varepsilon(q,0) = \frac{1}{U'(q)} \int_0^{T(q)} \big(p(s)f(p(s),q(s)) + U'(q(s))g(p(s),q(s))\big)\, ds. \tag{11.43}$$

The integral on the right hand side is known as the **Melnikov integral** for periodic orbits.

For example, let me show how this applies to the van der Pol equation (8.26). Here we have ($q = x$ and $p = y$) the harmonic oscillator $U(q) = q^2/2$ as unperturbed system and the unperturbed orbit is given by $(p(t), q(t)) = (q\sin(t), q\cos(t))$. Hence, using $f(p,q) = 0$, $g(p,q) = q - q^3/3$ we have

$$\Delta_\varepsilon(q,0) = q \int_0^{2\pi} \cos(s)^2 \big(\frac{\cos(s)^2}{3q^2} - 1\big) ds = \frac{\pi q}{4}(q^2 - 4) \tag{11.44}$$

and $q = 2$ is a simple zero of $\Delta_\varepsilon(q,0)$.

This result is not specific to the Hamiltonian form of the vector field as we will show next. In fact, consider the system

$$\dot{x} = f(x) + \varepsilon\, g(x,\varepsilon). \tag{11.45}$$

Suppose that the unperturbed system $\varepsilon = 0$ has a **period annulus**,, that is, an annulus of periodic orbits. Denote the period of a point $x$ in this annulus by $T(x)$.

Fix a periodic point $x_0$ in this annulus and let us derive some facts about the unperturbed system first. Let $\Phi(t,x,\varepsilon)$ be the flow of (11.45) and abbreviate $\Phi(t,x) = \Phi(t,x,0)$. Using the orthogonal vector field

$$f^\perp(x) = J f(x), \qquad J = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}. \tag{11.46}$$

we can make the following ansatz for the principal matrix solution of the first variational equation of the unperturbed system

$$\begin{aligned} \Pi_{x_0}(t,0) f(x_0) &= f(x(t)), \\ \Pi_{x_0}(t,0) f^\perp(x_0) &= \alpha_{x_0}(t) f(x(t)) + \beta_{x_0}(t) f^\perp(x(t)), \end{aligned} \tag{11.47}$$

where $x(t) = \Phi(t, x_0)$.

**Lemma 11.11.** *The coefficients $\alpha_{x_0}(t)$ and $\beta_{x_0}(t)$ are given by*

$$\begin{aligned}
\beta_{x_0}(t) &= \frac{|f(x_0)|^2}{|f(x(t))|^2} \mathrm{e}^{\int_0^t \mathrm{div}(f(x(s)))ds} \\
\alpha_{x_0}(t) &= \int_0^t \frac{\beta_{x_0}(s)}{|f(x(s))|^2} f(x(s))[J, A(s)] f(x(s)) ds,
\end{aligned} \tag{11.48}$$

*where $x(t) = \Phi(t, x_0)$ and $A(t) = df_{x(t)}$.*

**Proof.** Since $\beta(t) = \frac{|f(x_0)|^2}{|f(x(t))|^2} \det(\Pi_{x_0})$ the first equation follows from Liouville's formula. Next, differentiating (11.47) with respect to $t$ shows

$$\dot\alpha(t) f(x(t)) + \dot\beta(t) f^\perp(x(t)) = \beta(t)(A(t) f^\perp(x(t)) - (A(t) f(x(t)))^\perp) \tag{11.49}$$

since $\dot f(x(t)) = A(t) f(x(t))$. Multiplying both sides with $f(x(t))$ and integrating with respect to $t$ proves the claim since $\alpha(0) = 0$. □

Now denote by $\Psi(t, x)$ the flow of the orthogonal vector field $f^\perp(x)$ and let us introduce the more suitable coordinates

$$x(u, v) = \Phi(u, \Psi(v, x_0)). \tag{11.50}$$

Abbreviate $T(v) = T(x(u, v))$ and differentiate $\Phi(T(v), x(u, v)) - x(u, v) = 0$ with respect to $v$ producing

$$\dot\Phi(T(v), x(u, v)) \frac{\partial T}{\partial v}(v) + \frac{\partial \Phi}{\partial x}(T(v), x(u, v)) \frac{\partial x}{\partial v}(u, v) = \frac{\partial x}{\partial v}(u, v). \tag{11.51}$$

Evaluating at $(u, v) = (0, 0)$ gives

$$\Pi_{x_0}(T(x_0), 0) f^\perp(x_0) + \frac{\partial T}{\partial v}(0) f(x_0) = f^\perp(x_0). \tag{11.52}$$

Using (11.47) we obtain

$$(\alpha_{x_0}(T(x_0)) - \frac{\partial T}{\partial v}(0)) f(x_0) = (1 - \beta_{x_0}(T(x_0))) f^\perp(x_0) \tag{11.53}$$

or equivalently

$$\alpha_{x_0}(T(x_0)) = \frac{\partial T}{\partial v}(0) = \frac{\partial T}{\partial x}(x_0) f^\perp(x_0), \qquad \beta_{x_0}(T(x_0)) = 1. \tag{11.54}$$

After these preparations, let us consider the Poincaré map

$$P_\Sigma(x, \varepsilon) = \Phi(\tau(x, \varepsilon), x, \varepsilon), \qquad x \in \Sigma, \tag{11.55}$$

corresponding to some section $\Sigma$ (to be specified later). Since we expect the $\varepsilon$ derivative to be of importance, we fix $x_0 \in \Sigma$ and compute

$$\begin{aligned}
&\frac{\partial}{\partial \varepsilon}\Phi(\tau(x_0,\varepsilon),x_0,\varepsilon) - x_0\Big|_{\varepsilon=0} \\
&\quad = \dot{\Phi}(T(x_0),x_0)\frac{\partial \tau}{\partial \varepsilon}(x_0,0) + \frac{\partial}{\partial \varepsilon}\Phi(T(x_0),x_0,\varepsilon)\Big|_{\varepsilon=0} \\
&\quad = \frac{\partial \tau}{\partial \varepsilon}(x_0,0) f(x_0) + x_\varepsilon(T(x_0)),
\end{aligned} \tag{11.56}$$

where $x_\varepsilon(t)$ is the solution of the variational equation

$$\dot{x}_\varepsilon(t) = A(t)x_\varepsilon(t) + g(x(t),0) \tag{11.57}$$

corresponding to the initial condition $x_\varepsilon(0) = 0$. Splitting $g$ according to

$$g(x(s),0) = \frac{f(x(s))g(x(s),0)}{|f(x(s))|^2} f(x(s)) + \frac{f(x(s)) \wedge g(x(s),0)}{|f(x(s))|^2} f^\perp(x(s)) \tag{11.58}$$

and invoking (11.47) we obtain after a little calculation

$$\begin{aligned}
x_\varepsilon(T(x_0)) &= \int_0^{T(x_0)} \Pi_{x_0}(T(x_0),s) g(x(s),0) ds \\
&= (N(x_0) + \alpha_{x_0}(T(x_0))M(x_0)) f(x_0) + M(x_0) f^\perp(x_0),
\end{aligned} \tag{11.59}$$

where

$$M(x_0) = \int_0^{T(x_0)} \frac{f(x(s)) \wedge g(x(s),0)}{\beta_{x_0}(s)|f(x(s))|^2} ds \tag{11.60}$$

and

$$\begin{aligned}
N(x_0) \;\; = \;\; & \int_0^{T(x_0)} \frac{f(x(s))g(x(s),0)}{|f(x(s))|^2} ds \\
& - \int_0^{T(x_0)} \alpha_{x_0}(s) \frac{f(x(s)) \wedge g(x(s),0)}{\beta_{x_0}(s)|f(x(s))|^2} ds.
\end{aligned} \tag{11.61}$$

Putting everything together we have

$$\begin{aligned}
&\frac{\partial}{\partial \varepsilon}\Phi(\tau(x,\varepsilon),x,\varepsilon) - x\Big|_{\varepsilon=0} \\
&\quad = (\frac{\partial \tau}{\partial \varepsilon}(x,0) + N(x) + \alpha_x(T(x))M(x)) f(x) + M(x) f^\perp(x)
\end{aligned} \tag{11.62}$$

at any point $x \in \Sigma$.

Now let us fix $x_0$ and choose $\Sigma = \{x_0 + f(x_0)^\perp v | v \in \mathbb{R}\}$. Then the displacement function is

$$\Delta(v,\varepsilon) = (\Phi(\tau(x,\varepsilon),x,\varepsilon) - x) f^\perp(x_0), \quad x = x_0 + f(x_0)^\perp v, \tag{11.63}$$

and

$$\frac{\partial \Delta}{\partial \varepsilon}(0,0) = |f^\perp(x_0)|^2 M(x_0). \tag{11.64}$$

Moreover, since $\Phi(\tau(x_0,\varepsilon),x_0,\varepsilon)\in\Sigma$ we have

$$\frac{\partial\tau}{\partial\varepsilon}(x_0,0)+N(x_0)+\alpha_{x_0}(T(x_0))=0 \tag{11.65}$$

and, if $M(x_0)=0$,

$$\frac{\partial^2\Delta}{\partial\varepsilon\partial v}(0,0)=|f^{\perp}(x_0)|^2\frac{\partial M}{\partial x}(x_0)f^{\perp}(x_0). \tag{11.66}$$

**Theorem 11.12.** *Suppose (11.45) for $\varepsilon=0$ has a period annulus. If the Melnikov integral $M(x)$ has a zero $x_0$ at which the derivative of $M(x)$ in the direction of $f^{\perp}(x_0)$ does not vanish, then the periodic orbit at $x_0$ persists for small $\varepsilon$.*

Note that we have

$$M(x(t))=\beta_{x_0}(t)M(x_0). \tag{11.67}$$

**Problem 11.2.** *Show*

$$\begin{aligned}\beta_{x(s)}(t)&=\frac{\beta_{x_0}(t+s)}{\beta_{x_0}(s)},\\ \alpha_{x(s)}(t)&=\frac{1}{\beta_{x_0}(s)}\left(\alpha_{x_0}(t+s)-\alpha_{x_0}(s)\right)\end{aligned}$$

*and*

$$\beta_{x(s)}(T(x_0))=1,\quad \alpha_{x(s)}(T(x_0))=\frac{\alpha_{x_0}(T(x_0))}{\beta_{x_0}(s)}.$$

## 11.5. Melnikov's method for nonautonomous perturbations

Now let us consider the more general case of nonautonomous perturbations. We consider the nonautonomous system

$$\dot{x}(t)=f(x(t))+\varepsilon\, g(t,x(t),\varepsilon) \tag{11.68}$$

ore equivalently the extended autonomous one

$$\dot{x}=f(x)+\varepsilon\, g(\tau,x,\varepsilon),\qquad \dot{\tau}=1. \tag{11.69}$$

We will assume that $g(t,x,\varepsilon)$ is periodic with period $T$ and that the unperturbed system $\varepsilon=0$ has a period annulus.

To find a periodic orbit which persists we need of course require that the extended unperturbed system has a periodic orbit. Hence we need to suppose that the resonance condition

$$mT=nT(x_0),\qquad n,m\in\mathbb{N}, \tag{11.70}$$

where $T(x)$ denotes the period of $x$, holds for some periodic point $x_0$ in this annulus. It is no restriction to assume that $m$ and $n$ are relatively prime. Note that we have $\beta_{x_0}(nT(x_0))=1$ and $\alpha_{x_0}(nT(x_0))=n\,\alpha_{x_0}(T(x_0))$.

The Poincaré map corresponding to $\Sigma = \{\tau = t_0 \mod mT\}$ is given by

$$P_\Sigma(x,\varepsilon) = \Phi(mT,(x,t_0),\varepsilon) \tag{11.71}$$

and the displacement function is

$$\Delta(x,\varepsilon) = x(mT,\varepsilon) - x, \tag{11.72}$$

where $x(t,\varepsilon)$ is the solution corresponding to the initial condition $x(t_0,\varepsilon) = x$. Note that it is no restriction to assume $t_0 = 0$ and replace $g(s,x,\varepsilon)$ by $g(s+t_0,x,\varepsilon)$.

Again it is not possible to apply the implicit function theorem directly to $\Delta(x,\varepsilon)$ since the derivative in the direction of $f(x_0)$ vanishes. We will handle this problem as in the previous section by a regularization process. However, since $\Delta(x,\varepsilon)$ is now two dimensional, two cases can occur.

One is if the derivative of $\Delta(x,\varepsilon)$ in the direction of $f^\perp(x_0)$ also vanishes. This is the case if, for example, the period in the annulus is constant and hence $\Delta(x,0) = 0$. Here we can divide by $\varepsilon$ and proceed as before.

The second case is if the derivative of $\Delta(x,\varepsilon)$ in the direction of $f^\perp(x_0)$ does not vanish. Here we have to use a Liapunov-Schmidt type reduction and split $\mathbb{R}^2$ according to $f(x_0)$ and $f^\perp(x_0)$. One direction can be handled by the implicit function theorem directly and the remaining one can be treated as in the first case.

We will express $\Delta$ in more suitable coordinates $x(u,v)$ from (11.50). Using the results from the previous section we have

$$\frac{\partial\Delta}{\partial u}(x_0,0) = 0, \qquad \frac{\partial\Delta}{\partial v}(x_0,0) = n\,\alpha_{x_0}(T(x_0))f(x_0) \tag{11.73}$$

and

$$\begin{aligned}\frac{\partial\Delta}{\partial\varepsilon}(x_0,0) = x_\varepsilon(mT) &= (N(t_0,x_0) + n\,\alpha_{x_0}(T(x_0))M(t_0,x_0))f(x_0)\\ &\quad + M(t_0,x_0)f^\perp(x_0),\end{aligned} \tag{11.74}$$

where

$$M(t_0,x_0) = \int_0^{nT(x_0)} \frac{f(x(s)) \wedge g(s+t_0,x(s),0)}{\beta_{x_0}(s)|f(x(s))|^2} ds \tag{11.75}$$

and

$$\begin{aligned}N(t_0,x_0) &= \int_0^{nT(x_0)} \frac{f(x(s))g(s+t_0,x(s),0)}{|f(x(s))|^2} ds\\ &\quad - \int_0^{nT(x_0)} \alpha_{x_0}(s)\frac{f(x(s)) \wedge g(s+t_0,x(s),0)}{\beta_{x_0}(s)|f(x(s))|^2} ds.\end{aligned} \tag{11.76}$$

Note that $M(t_0+T,x_0) = M(t_0,x_0)$ and $N(t_0+T,x_0) = N(t_0,x_0)$.

With this notation we can now easily treat the case of an **isochronous period annulus**, where $T(x) = T(x_0)$ is constant, respectively $\alpha_x(T(x)) = 0$. Since $\Delta(x, 0) = 0$ we can proceed as before to obtain

**Theorem 11.13.** *Suppose (11.68) for $\varepsilon = 0$ has an isochronous period annulus. If the function $x \mapsto (M(t_0, x), N(t_0, x))$ has a simple zero at $(t_0, x_0)$, then the periodic orbit at $(t_0, x_0)$ persists for small $\varepsilon$.*

The case $\alpha_x(T(x)) \neq 0$ will be considered next. We will call the period annulus a **regular period annulus** in this case.

We split the displacement function according to (compare (11.50))

$$\Delta(x(u, v), \varepsilon) = \Delta_1(u, v, \varepsilon) f(x_0) + \Delta_2(u, v, \varepsilon) f^{\perp}(x_0). \tag{11.77}$$

Then

$$\frac{\partial \Delta_1}{\partial v}(0, 0, 0) = n\, \alpha_{x_0}(T(x_0)) \neq 0 \tag{11.78}$$

and hence there is a function $v(u, \varepsilon)$ such that $\Delta_1(u, v(u, \varepsilon), \varepsilon) = 0$ by the implicit function theorem. Moreover, by $\Delta(x(u, 0), 0) = 0$ we even have $v(u, 0) = 0$. Hence it remains to find a zero of

$$\tilde{\Delta}_2(u, \varepsilon) = \Delta_2(u, v(u, \varepsilon), \varepsilon). \tag{11.79}$$

Since $\tilde{\Delta}_2(u, 0) = \Delta_2(u, 0, 0) = 0$, we can divide by $\varepsilon$ and apply the implicit function theorem as before.

Now using

$$\frac{\partial \tilde{\Delta}_2}{\partial \varepsilon}(0, 0) = M(t_0, x_0). \tag{11.80}$$

and, if $M(t_0, x_0) = 0$,

$$\frac{\partial^2 \tilde{\Delta}_2}{\partial \varepsilon \partial u}(0, 0) = \frac{\partial M}{\partial x}(t_0, x_0) f(x_0) \tag{11.81}$$

we obtain the following result.

**Theorem 11.14.** *Suppose (11.68) for $\varepsilon = 0$ has a regular period annulus. If the function $x \mapsto M(t_0, x)$ has a zero at $(t_0, x_0)$ at which the derivative of $M(t_0, x)$ in the direction of $f(x_0)$ does not vanish, then the periodic orbit at $(t_0, x_0)$ persists for small $\varepsilon$.*

Chapter 12

# Discrete dynamical systems in one dimension

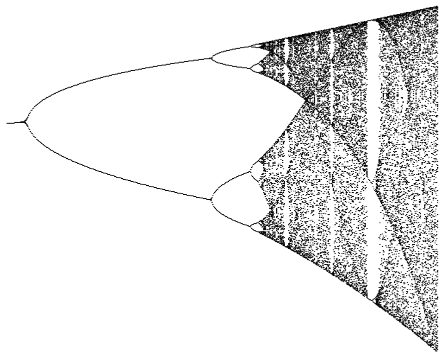
## 12.1. Period doubling

We now return to the logistic equation and the numerical investigation started in Section 10.1. Let us try to get a more complete picture by iterating one given initial condition for different values of $\mu$. Since we are only interested in the asymptotic behavior we first iterate 200 times and then plot the next 100 iterations.

```
In[1]:= BifurcationList[f_, x0_, {μ_, μ0_, μ1_}, opts___] :=
        Block[{Nmin, Nmax, Steps},
         Nmin, Nmax, Steps = {Nmin, Nmax, Steps} /. {opts} /.
          {Nmin → 200, Nmax → 300, Steps → 300};
         Flatten[
          Table[Module[{x},
           x = Nest[f, x0, Nmin];
           Map[{μ, #}&, NestList[f, x, Nmax − Nmin]]],
          {μ, μ0, μ1, (μ1 − μ0)/Steps}],
         1]];
```

The result is shown below.

```
In[2]:= ListPlot[
          BifurcationList[μ#(1 − #)&, 0.4, {μ, 2.95, 4}],
          PlotStyle → {PointSize[0.002]}, PlotRange → All,
          Axes → False];
```



So we see that at certain point the *attracting set* just doubles its size and gets more and more complicated. I do not want to say more about this picture right now, however, I hope that you are convinced that the dynamics of this simple system is indeed quite complicated. Feel free to experiment with the above code and try to plot some parts of the above diagram in more detail.

In particular we see that there are certain points $\mu$ where there is a qualitative change in the dynamics of a dynamical system. Such a point is called a **bifurcation point** of the system.

The first point was $\mu = 1$, where a second fixed point entered our interval $[0, 1]$. Now when can such a situation happen? First of all, fixed points are zeros of the function

$$g(x) = f(x) - x. \tag{12.1}$$

If $f$ is differentiable, so is $g$ and by the implicit function theorem the number of zeros can only change locally if $g'(x) = 0$ at a zero of $g$. In our case of the logistic equation this yields the following system

$$\begin{aligned} L_\mu(x) &= x = \mu x(1-x), \\ L'_\mu(x) &= 1 = \mu(1-2x), \end{aligned} \tag{12.2}$$

which has the only solution $x = 0$ and $\mu = 1$. So what precisely happens at the value $\mu = 1$? Obviously a second fixed point $p = 1 - 1/\mu$ enters our interval. The fixed point 0 is no longer attracting since $L'_\mu(0) = \mu > 1$ but $p$ is for $1 < \mu < 3$ since $L'_\mu(p) = 2 - \mu$. Moreover, I claim $W^s(0) = \{0, 1\}$ and $W^s(p) = (0, 1)$ for $1 < \mu \le 3$. To show this first observe that we have

$$\frac{L_\mu(x) - p}{x - p} = 1 - \mu x. \tag{12.3}$$

If $1 < \mu \le 2$ the right hand side is in $(-1, 1)$ for $x \in (0, 1)$. Hence $x \in (0, 1)$ converges to $p$. If $2 < \mu \le 3$ the right hand side is in $(-1, 1)$ only for $x \in (0, \frac{2}{\mu})$. If $x$ stays in this region for all iterations, it will converge to $p$. Otherwise, we have $x \in [\frac{2}{\mu}, 1]$ after some iterations. After the next iteration we are in $[0, 2 - \frac{4}{\mu}]$ and in particular below $p$. Next, we stay below $p$ until we reach $[\frac{1}{\mu}, p]$. For this case consider the second iterate which satisfies

$$\frac{L_\mu^2(x) - p}{x - p} = (1 - \mu x)(1 - \mu L_\mu(x)). \tag{12.4}$$

For $x \in (\frac{1}{\mu}, p)$ the right hand side is in $(-1, 1)$ implying $L_\mu^{2n}(x) \to p$. Thus we also have $L_\mu^{2n+1}(x) \to L_\mu(p) = p$ and hence $L_\mu^n(x) \to p$ for all $x \in (0, 1)$.

Now what happens for $\mu > 3$? Since we have $L_\mu'(p) = 2 - \mu < -1$ for $\mu > 3$ the fixed point $p$ is no longer attracting. Moreover, a look at our numeric investigation shows that there should be a periodic orbit of period two. And indeed, solving the equation

$$L_\mu^2(x) = x \tag{12.5}$$

shows that, in addition to the fixed points, there is a periodic orbit

$$p_\pm = \frac{1 + \mu \pm \sqrt{(\mu + 1)(\mu - 3)}}{2\mu} \tag{12.6}$$

for $\mu > 3$. Moreover, we have $(L_\mu^2)'(p_\pm) = L_\mu'(p_+) L_\mu'(p_-) = 4 + 2\mu - \mu^2$ which is in $(-1, 1)$ for $3 < \mu < 1 + \sqrt{6}$. Hence, the attracting fixed point $p$ is replaced by the attracting periodic orbit $p_+$, $p_-$. This phenomenon is known as **period doubling**. Our numerical bifurcation diagram shows that this process continues. The attracting period two orbit is replaced by an attracting period four orbit at $\mu = 1 + \sqrt{6}$ (period doubling bifurcation in $f^2$) and so forth. Clearly it is no longer possible to analytically compute all these points since the degrees of the arising polynomial equations get too high.

So let us try to better understand the period doubling bifurcation. Suppose we have a map $f : I \to I$ depending on a parameter $\mu$. Suppose that at $\mu_0$ the number of zeros of $f^2(x) - x$ changes locally at $p$, that is, suppose there are two new zeros $p_\pm(\mu)$ such that $p_\pm(\mu_0) = p$ and $f(p_\pm(\mu)) = p_\mp(\mu)$. By continuity of $f$ we must have $f([p_-(\mu), p_+(\mu)]) \subseteq [p_-(\mu), p_+(\mu)]$ and hence there must be a fixed point $p(\mu) \in [p_-(\mu), p_+(\mu)]$. So the fixed point $p$ persists. That should only happen if $f'(p) \neq 1$. But since we must have $(f^2)'(p) = f'(p)^2 = 1$ this implies $f'(p) = -1$.

In summary, orbits of period two will appear in general only at fixed points where $f'(p) = -1$.

Note that in the above argument we have shown that existence of an orbit of period two implies existence of an orbit of period one. In fact, a much stronger result is true which will be presented in the next section.

## 12.2. Sarkovskii's theorem

In this section we want to show that certain periods imply others for continuous maps $f : I \to I$, where $I \subseteq \mathbb{R}$ is some compact interval. As our first result we will show that period three implies all others.

**Lemma 12.1.** *Suppose $f : I \to I$ is continuous and has an orbit of period three. Then it also has orbits with (prime) period $n$ for all $n \in \mathbb{N}$.*

**Proof.** The proof is based on the following two elementary facts. If $I, J$ are two closed intervals satisfying $f(J) \supseteq I$, then there is a subinterval $J_0$ of $J$ such that $f(J_0) = I$. If $f(J) \supseteq J$, there is a fixed point in $J$.

Let $a < b < c$ be the period three orbit. And suppose $f(a) = b$, $f(b) = c$ (the case $f(a) = c$, $f(b) = a$ is similar). Abbreviate $I_0 = [a, b]$ and $I_1 = [b, c]$.

Set $J_0 = I_1$ and observe that $f(I_1) \supseteq I_1$ by continuity of $f$. Hence we can find a subinterval $J_1 \subseteq J_0$ (prove this) such that $f(J_1) = J_0$. Moreover, since $f(J_1) = J_0 \supseteq J_1$ we can iterate this procedure to obtain a sequence of nesting sets $J_k$ such that $f(J_k) = J_{k-1}$. In particular, we have $f^k(J_k) = J_0 \supseteq J_k$ and thus $f^n$ has a fixed point in $J_n$. The only problem is, is the prime period of this point $n$ if $n > 1$? Unfortunately, since all iterations stay in $I_1$, we might always get the same fixed point of $f$. To ensure that this does not happen we need to refine our analysis by going to $I_0$ in the $(n-1)$-th step and then back to $I_1$.

So let $n > 1$ and define $J_0 \supseteq \cdots \supseteq J_{n-2}$ as before. Now observe $f^{n-1}(J_{n-2}) = f(I_1) \supseteq I_0$. Hence we can choose a subinterval $J_{n-1} \subseteq J_{n-2}$ such that $f^{n-1}(J_{n-1}) = I_0$ and thus $f^n(J_{n-1}) = f(I_0) \supseteq I_1$. Again there is a subinterval $J_n \subseteq J_{n-1}$ such that $f^n(J_n) = I_1$. Hence there is a fixed point $x \in J_n$ of $f^n$ such that $f^j(x) \in I_1$ for $j \neq n-1$ and $f^{n-1}(x) \in I_0$. Moreover, if $f^j(x) \in I_1$ for all $j$, then $f^{n-1}(x) = b$ contradicting $a = f^{n-2}(x) \in I_1$. The prime period of $x$ cannot be $n-1$ since $f^{n-1}(x) \in [a, b)$ and if it were smaller than $n-1$, all iterates would stay in the interior of $I_1$, a contradiction. So the prime period is $n$ and we are done. □

So when does the first period three orbit appear for the logistic map $L_\mu$? For $\mu = 4$ the equation $L_\mu^3(x) = x$ can be solved using *Mathematica* showing that there are two period three orbits. One of them is given by

$$\{\frac{1}{2}(1+c), 1-c^2, 4c^2(1-c^2)\}, \qquad c = \cos(\frac{\pi}{9}), \tag{12.7}$$

the other one is slightly more complicated. Since there are no period three orbits for $0 \le \mu \le 3$, there must be a local change in the zero set of $L_\mu^3(x) - x$. Hence we need to search for a solution of the system of equations $L_\mu^3(x) = x$, $(L_\mu^3)'(x) = 1$. Plugging this equation into *Mathematica* gives a rather complicated solution for the orbit, but a simple one for $\mu = 1 + 2\sqrt{2} = 3.828$. Since this is the only solution for $\mu \in \mathbb{R}$ other than $x = 0, \mu = 1$ we know that the logistic equation has orbits of all periods for $\mu \ge 1 + 2\sqrt{2}$.

In fact, this result is only a special case of a much more general theorem due to Sarkovskii. We first introduce a quite unusual ordering of the natural numbers as follows. First note that all integers can be written as $2^m(2n+1)$ with $m, n \in \mathbb{N}_0$. Now for all $m \in \mathbb{N}_0$ and $n \in \mathbb{N}$ we first arrange them by $m$ and then, for equal $m$, by $n$ in increasing order. Finally we add all powers of two ($m = 0$) in decreasing order. That is, denoting the **Sarkovskii ordering** by $\succ$ we have

$$3 \succ 5 \succ \cdots \succ 2 \cdot 3 \succ 2 \cdot 5 \succ \cdots \succ 2^m(2n+1) \succ \cdots \succ 2^2 \succ 2 \succ 1 \quad (12.8)$$

With this notation the following claim holds.

**Theorem 12.2** (Sarkovskii). *Suppose $f : I \to I$ is continuous and has an orbit of period $m$. Then it also has orbits with prime period $n$ for all $m \succ n$.*

The proof is in spirit similar to that of Lemma 12.1 but quite tedious. Hence we omit it here. It can be found (e.g.) in [**23**].

## 12.3. On the definition of chaos

In this section we want to define when we consider a discrete dynamical system to be chaotic. We return to our abstract setting and consider a continuous map $f : M \to M$ on a metric space $M$.

It is quite clear from the outset, that defining chaos is a difficult task. Hence it will not surprise you that different authors use different definitions. But before giving you a definition, let us reflect on the problem for a moment.

First of all, you will certainly agree that a chaotic system should exhibit **sensitive dependence on initial conditions**. That is, there should be a $\delta > 0$ such that for any $x \in M$ and any $\varepsilon > 0$ there is a $y \in M$ and an $n \in \mathbb{N}$ such that $d(x, y) < \varepsilon$ and $d(f^n(x), f^n(y)) > \delta$.

However, the example

$$M = (0, \infty), \quad f(x) = (1 + \mu)x, \quad \mu > 0, \qquad (12.9)$$

exhibits sensitive dependence on initial conditions but should definitely not be considered chaotic since all iterates in the above example converge to infinity. To rule out such a situation we introduce another condition.

A map $f$ as above is called **topologically transitive** if for any given open sets $U, V \subseteq M$ there is an $n \in \mathbb{N}$ such that $f^n(U) \cap V \neq \emptyset$. Observe that a system is transitive if it contains a dense orbit (Problem 12.1).

A system having both properties is called chaotic in the book by Robinson [**23**]. However, we will still consider another definition since this one has one draw back. It involves the metric structure of $M$ and hence is not preserved under topological equivalence. Two dynamical systems $(M_j, f_j)$, $j = 1, 2$, are called **topological equivalent** if there is a homeomorphism $\varphi : M_1 \to M_2$ such that the following diagram commutes.

$$\begin{array}{ccc} M_1 & \xrightarrow{f_1} & M_1 \\ \varphi \updownarrow & & \updownarrow \varphi \\ M_2 & \xrightarrow{f_2} & M_2 \end{array} \tag{12.10}$$

Clearly $p_2 = \varphi(p_1)$ is a periodic point of period $n$ for $f_2$ if and only if $p_1$ is for $f_1$. Moreover, we have $W^s(p_2) = \varphi(W^s(p_1))$ and all topological properties (e.g., transitivity) hold for one system if and only if they hold for the other.

On the other hand, properties involving the metric structure might not be preserved. For example, take $\varphi = x^{-1}$, then the above example is mapped to the system

$$M = (0, \infty), \quad f(x) = (1 + \mu)^{-1} x, \quad \mu > 0, \tag{12.11}$$

which no longer exhibits sensitive dependence on initial conditions. (Note that the problem here is that $M$ is not compact. If $M$ is compact, $f$ is uniformly continuous and sensitive dependence on initial conditions is preserved.)

Hence we will use the following definition for chaos due to Devaney [**7**]. A discrete dynamical system $(M, f)$ with continuous $f$ and infinite $M$ as above is called **chaotic** if it is transitive and if the periodic orbits are dense. If $M$ is finite and transitive it is not hard to see that it consists of one single periodic orbit.

The following lemma shows that chaotic dynamical systems exhibit sensitive dependence on initial conditions.

**Lemma 12.3.** *Suppose $f : M \to M$ is chaotic, then it exhibits sensitive dependence on initial conditions.*

**Proof.** First observe that there is a number $8\delta$ such that for all $x \in M$ there exists a periodic point $q \in M$ whose orbit is of distance at least $4\delta$ from $x$. In fact, since $M$ is not finite we can pick two periodic points $q_1$ and $q_2$ with disjoint orbits. Let $8\delta$ be the distance between the two orbits. Then, by the triangle inequality the distance from at least one orbit to $x$ must be larger than $4\delta$.

Fix $x \in M$ and $\varepsilon > 0$ and let $q$ be a periodic orbit with distance at least $4\delta$. Without restriction we assume $\varepsilon < \delta$. Since periodic orbits are dense, there is a periodic point $p \in B_\varepsilon(x)$ of period $n$.

Now the idea is as follows. By transitivity there is a $y$ close to $x$ which gets close to $q$ after $k$ iterations. Now iterate another $j$ times such that $k+j$ is a multiple of $n$. Since $0 \le j < n$ is small, $f^{k+j}(y)$ is still close to the orbit of $q$. Hence $f^{k+j}(y)$ is far away from $x$ and $f^{k+j}(p) = p$ is close to $x$. Since $f^{k+j}(x)$ cannot be close to both, we have sensitive dependence on initial conditions.

Now to the boring details. Let $V = \bigcap_{i=0}^{n-1} f^{-i}(B_\delta(f^i(q)))$ (i.e., $z \in V$ implies that $f^i(z) \in B_\delta(f^i(q))$ for $0 \le i < n$). By transitivity there is a $y \in B_\varepsilon(x)$ such that $f^k(y) \in V$ and hence $f^{k+j}(y) \in B_\delta(f^j(q))$. Now by the triangle inequality and $f^{k+j}(p) = p$ we have

$$\begin{aligned} d(f^{k+j}(p), f^{k+j}(y)) &\ge d(x, f^j(q)) - d(f^j(q), f^{k+j}(y)) - d(p, x) \\ &> 4\delta - \delta - \delta = 2\delta. \end{aligned} \tag{12.12}$$

Thus either $d(f^{k+j}(x), f^{k+j}(y)) > \delta$ or $d(f^{k+j}(p), f^{k+j}(x)) > \delta$ and we are done. □

Now we have defined what a chaotic dynamical system is, but we haven't seen one yet! Well, in fact we have, I claim that the logistic map is chaotic for $\mu = 4$.

To show this we will take a detour via the **tent map**

$$M = [0, 1], \quad T_\mu(x) = \frac{\mu}{2}(1 - |2x - 1|) \tag{12.13}$$

using topological equivalence. The tent map $T_2$ is equivalent to the logistic map $L_4$ by virtue of the homeomorphism $\varphi(x) = \sin^2(\frac{\pi x}{2})$ (Problem 12.2). Hence it follows that $L_4$ is chaotic once we have shown that $T_2$ is.

The main advantage of $T_2$ is that the iterates are easy to compute. Using

$$T_2(x) = \left\{ \begin{array}{cc} 2x, & 0 \le x \le \frac{1}{2} \\ 2 - 2x, & \frac{1}{2} \le x \le 1 \end{array} \right\} \tag{12.14}$$

it is not hard to verify that

$$T_2^n(x) = \left\{ \begin{array}{cc} 2^n x - 2j, & \frac{2j}{2^n} \le x \le \frac{2j+1}{2^n} \\ 2(j+1) - 2^n x, & \frac{2j+1}{2^n} \le x \le \frac{2j+2}{2^n} \end{array} \right\}_{0 \le j \le 2^{n-1}-1}. \tag{12.15}$$

Moreover, each of the intervals $I_{n,j} = [\frac{j}{2^n}, \frac{j+1}{2^n}]$ is mapped to $[0, 1]$ under $T_2^n$. Hence each of the intervals $I_{n,j}$ contains (precisely) one solution of $T_2^n(x) = x$ implying that periodic points are dense. For given $x \in [0, 1]$ and $\varepsilon > 0$ we can find $n, j$ such that $I_{n,j} \subset B_\varepsilon(x)$. Hence $T_2^n(B_\varepsilon(x)) = [0, 1]$, which shows that $T_2$ is transitive. Hence the system is chaotic. It is also not hard to show directly that $T_2$ has sensitive dependence on initial conditions (exercise).

Suppose $f(0) = f(1) = 0$, $f(\frac{1}{2}) = 1$, and suppose $f$ is monotone increasing, decreasing on $[0, \frac{1}{2}]$, $[\frac{1}{2}, 1]$. Does any such map have similar properties? Is such a map always chaotic?

**Problem 12.1.** *Show that a closed invariant set which has a dense orbit is topologically transitive.*

**Problem 12.2.** *Show that $T_2$ and $L_4$ are topologically equivalent via the map $\varphi(x) = \sin^2(\frac{\pi x}{2})$. (i.e., show that $\varphi$ is a homeomorphism and that $\varphi \circ T_2 = L_4 \circ \varphi$).*

## 12.4. Cantor sets and the tent map

Now let us further investigate the tent map $T_\mu$ for $\mu > 2$. Unfortunately, in this case $T_\mu$ does no longer map $[0,1]$ into itself. Hence we must consider it as a map on $\mathbb{R}$,

$$M = \mathbb{R}, \quad T_\mu(x) = \frac{\mu}{2}(1 - |2x - 1|). \tag{12.16}$$

It is not hard to show that $T_\mu^n(x) \to -\infty$ if $x \in \mathbb{R}\backslash[0,1]$. Hence most points will escape to $-\infty$. However, there are still some points in $[0,1]$ which stay in $[0,1]$ for all iterations (e.g., 0 and 1). But how can we find these points?

Let $\Lambda_0 = [0,1]$. Then the points which are mapped to $\Lambda_0$ under one iteration are given by $(\frac{1}{\mu}\Lambda_0) \cup (1 - \frac{1}{\mu}\Lambda_0)$. Denote this set by

$$\Lambda_1 = [0, \frac{1}{\mu}] \cup [1 - \frac{1}{\mu}, 1]. \tag{12.17}$$

All points in $\mathbb{R}\backslash\Lambda_1$ escape to $-\infty$ since the points in $(\frac{1}{\mu}, 1 - \frac{1}{\mu})$ are mapped to $\mathbb{R}\backslash[0,1]$ after one iteration.

Similarly, the points which are mapped to $\Lambda_1$ under one iteration are given by $(\frac{1}{\mu}\Lambda_1) \cup (1 - \frac{1}{\mu}\Lambda_1)$. Hence the corresponding set

$$\Lambda_2 = [0, \frac{1}{\mu^2}] \cup [\frac{1}{\mu} - \frac{1}{\mu^2}, \frac{1}{\mu}] \cap [1 - \frac{1}{\mu}, 1 - \frac{1}{\mu} + \frac{1}{\mu^2}] \cup [1 - \frac{1}{\mu^2}, 1] \tag{12.18}$$

has the property that points starting in this set stay in $[0,1]$ during two iterations. Proceeding inductively we obtain sets $\Lambda_n = (\frac{1}{\mu}\Lambda_{n-1}) \cup (1 - \frac{1}{\mu}\Lambda_{n-1})$ having the property that points starting in $\Lambda_n$ stay in $[0,1]$ for at least $n$ iterations. Moreover, each set $\Lambda_n$ consists of $2^n$ closed subintervals of length $\mu^{-n}$.

Now if we want to stay in $[0,1]$ we have to take the intersection of all these sets, that is, we define

$$\Lambda = \bigcap_{n\in\mathbb{N}} \Lambda_n \subset [0,1]. \tag{12.19}$$

Since the sets $\Lambda_n$ form a nesting sequence of compact sets, the set $\Lambda$ is also compact and nonempty. By construction the set $\Lambda$ is invariant since we have

$$T_\mu(\Lambda) = \Lambda \tag{12.20}$$

and all points in the open set $\mathbb{R}\backslash\Lambda$ converge to $-\infty$.

Moreover, since the endpoints of the subintervals of $\Lambda_n$ are just given by $f^{-n}(\{0,1\})$, we see that these points are in $\Lambda$. Now the set $\Lambda$ has two more interesting properties. First of all it is totally disconnected, that is, it contains no open subintervals. In fact, this easily follows since its Lebesgue measure $|\Lambda| \le \lim_{n\to\infty}|\Lambda_n| = \lim_{n\to\infty}(2/\mu)^n = 0$ vanishes. Second, it is perfect, that is, every point is an accumulation point. This is also not hard to see, since $x\in\Lambda$ implies that $x$ must lie in some subinterval of $\Lambda_n$ for every $n$. Since the endpoints of these subintervals are in $\Lambda$ (as noted earlier) and converge to $x$, the point $x$ is an accumulation point.

Compact sets which are totally disconnected and perfect are called **Cantor sets**. Hence we have proven,

**Lemma 12.4.** *The set $\Lambda$ is a Cantor set.*

This result is also not surprising since the construction very much reassembles the construction of the Cantor middle-thirds set you know from your calculus course. Moreover, we obtain precisely the Cantor middle-thirds set if we choose $\mu = 3$. Maybe you also recall, that this case can be conveniently described if one writes $x$ in the base three number system. Hence fix $\mu = 3$ and let us write

$$x = \sum_{n\in\mathbb{N}} \frac{x_n}{3^n}, \qquad x_n \in \{0,1,2\}. \tag{12.21}$$

Then we have $\Lambda_n = \{x | x_j \neq 1,\ 1 \le j \le n\}$ and hence

$$\Lambda = \{x | x_j \neq 1, j \in \mathbb{N}\}. \tag{12.22}$$

Moreover, the action of $T_3$ can also be transparently described using this notation

$$\left\{ \begin{array}{lcl} x_1 = 0 & \Rightarrow & T_3(x) = \sum_{n\in\mathbb{N}} \frac{x_{n+1}}{3^n} \\ x_1 = 1 & \Rightarrow & T_3(x) \notin [0,1] \\ x_1 = 2 & \Rightarrow & T_3(x) = \sum_{n\in\mathbb{N}} \frac{x'_{n+1}}{3^n} \end{array} \right. , \tag{12.23}$$

where $x'_n = 2 - x_j$ (i.e., $0' = 2$, $1' = 1$, $2' = 0$). Unfortunately this description still has a few draw backs. First of all, the map $x \mapsto \{x_n\}$ is not well defined, since for some points there is more than one possible expansion ($\frac{1}{3} = \sum_{n=2}^{\infty} \frac{2}{3^n}$). Next, it is not easy to tell when two point $x$, $y$ are close by looking at $x_n$, $y_n$ and the fact that $T_3$ does not simply shift the sequence $x_n$ is a little annoying. Finally, it only works for $\mu = 3$.

So let us return to arbitrary $\mu > 2$ and let us see whether we can do better. Let $\Sigma_2 = \{0,1\}^{\mathbb{N}_0}$ be the set of sequences taking only the values 0 and 1.

Set $I_0 = [0, \frac{1}{\mu}]$, $I_1 = [1 - \frac{1}{\mu}, 1]$ and define the **itinerary map**

$$\begin{array}{rccl} \varphi: & \Lambda & \to & \Sigma_2 \\ & x & \mapsto & x_n = j \text{ if } T_\mu^n(x) \in I_j \end{array} . \tag{12.24}$$

Then $\varphi$ is well defined and $T_\mu$ acts on $x_n$ just by a simple shift. That is, if we introduce the shift map $\sigma: \Sigma_2 \to \Sigma_2$, $(x_0, x_1, \dots) \mapsto (x_1, x_2, \dots)$, we have $\sigma \circ \varphi = \varphi \circ T_\mu$ and it looks like we have a topological equivalence between $(\Lambda, T_\mu)$ and $(\Sigma_2, \sigma)$. But before we can show this, we need some further definitions first.

First of all we need to make sure that $(\Sigma_2, \sigma)$ is a dynamical system. Hence we need a metric on $\Sigma_2$. We will take the following one

$$d(x, y) = \sum_{n \in \mathbb{N}_0} \frac{|x_n - y_n|}{2^n} \tag{12.25}$$

(prove that this is indeed a metric). Moreover, we need to make sure that $\sigma$ is continuous. But since

$$d(\sigma(x), \sigma(y)) \le 2\, d(x, y) \tag{12.26}$$

it is immediate that $\sigma$ is even uniformly continuous.

So it remains to show that $\varphi$ is a homeomorphism.

We start by returning to the construction of $\Lambda_n$. If we set $I = [0,1]$ we have seen that $\Lambda_1$ consists of two subintervals $I_0 = \frac{1}{\mu} I$ and $I_1 = 1 - \frac{1}{\mu} I$. Proceeding inductively we see that the set $\Lambda_n$ consist of $2^n$ subintervals $I_{s_0,\cdots,s_{n-1}}$, $s_j \in \{0,1\}$, defined recursively via $I_{0,s_0,\cdots,s_n} = \frac{1}{\mu} I_{s_0,\cdots,s_n}$ and $I_{1,s_0,\cdots,s_n} = 1 - \frac{1}{\mu} I_{s_0,\cdots,s_n}$. Note that $T_\mu(I_{s_0,\cdots,s_n}) = I_{s_1,\cdots,s_n}$.

By construction we have $x \in I_{s_0,\cdots,s_n}$ if and only if $\varphi(x)_i = s_i$ for $0 \le i \le n$. Now pick a sequence $s \in \Sigma_2$ and consider the intersection of nesting intervals

$$I_s = \bigcap_{n \in \mathbb{N}_0} I_{s_0,\cdots,s_n}. \tag{12.27}$$

By the finite intersection property of compact sets it is a nonempty interval, hence $\varphi$ is onto. By $|I_{s_0,\cdots,s_n}| = \mu^{-n-1}$ its length is zero and thus it can contain only one point, that is, $\varphi$ is injective.

If $x$ and $y$ are close so are $T_\mu(x)^n$ and $T_\mu(y)^n$ by continuity of $T_\mu$. Hence, for $y$ sufficiently close to $x$ the first $n$ iterates will stay sufficiently close such that $x_j = y_j$ for $0 \le j \le n$. But this implies that $\varphi(x)$ and $\varphi(y)$ are close and hence $\varphi$ is continuous. Similarly, $\varphi(x)$ and $\varphi(y)$ close implies that the first $n$ terms are equal. Hence $x, y \in I_{x_0,\cdots,x_n} = I_{y_0,\cdots,y_n}$ are close, implying that $\varphi^{-1}$ is continuous.

In summary,

**Theorem 12.5.** *The two dynamical systems $(\Lambda, T_\mu)$, $\mu > 2$, and $(\Sigma_2, \sigma)$ are topologically equivalent via the homeomorphism $\varphi : \Lambda \to \Sigma_2$.*

Hence in order to understand the tent map for $\mu > 2$, all we have to do is to study the shift map $\sigma$ on $\Sigma_2$. In fact, we will show that $(\Sigma_2, \sigma)$, and hence $(\Lambda, T_\mu)$, $\mu > 2$, is chaotic in the next section.

## 12.5. Symbolic dynamics

The considerations of the previous section have shown that the shift map on a sequence space of finitely many symbols is hidden in the tent map. This turns out to be true for other systems as well. Hence it deserves a thorough investigation which will be done now.

Let $N \in \mathbb{N}\backslash\{1\}$ and define the **space on $N$ symbols**

$$\Sigma_N = \{0, 1, \dots, N-1\}^{\mathbb{N}_0} \tag{12.28}$$

to be the set of sequences taking only the values $0, \dots, N-1$. Note that $\Sigma_N$ is not countable (why?).

Defining

$$d(x, y) = \sum_{n\in\mathbb{N}_0} \frac{|x_n - y_n|}{N^n}, \tag{12.29}$$

$\Sigma_N$ becomes a metric space. Observe that two points $x$ and $y$ are close if and only if their first $n$ values coincide. More precisely,

**Lemma 12.6.** *We have $d(x, y) \le N^{-n}$ if $x_j = y_j$ for all $j \le n$ and we have $d(x, y) \ge N^{-n}$ if $x_j \ne y_j$ for at least one $j \le n$.*

**Proof.** Suppose $x_j = y_j$ for all $j \le n$, then

$$d(x, y) = \sum_{j>n} \frac{|x_j - y_j|}{N^j} \le \frac{1}{N^{n+1}} \sum_{j\ge 0} \frac{N-1}{N^j} = \frac{1}{N^n}. \tag{12.30}$$

Conversely, if $x_j \ne y_j$ for at least one $j \le n$, we have

$$d(x, y) = \sum_{k\in\mathbb{N}} \frac{|x_k - y_k|}{N^k} \ge \frac{1}{N^j} \ge \frac{1}{N^n}. \tag{12.31}$$

□

We first show that $\Sigma_N$ is a Cantor set, that is, it is compact, perfect and totally disconnected. Here a topological space $M$ is called totally disconnected if for any two points $x$ and $y$ there are disjoint respective open neighborhoods $U$ and $V$ such that $U \cup V = M$. I leave it as an exercise to prove that this is equivalent to our previous definition for subsets of the real line (Hint: If $x, y \in M \subset \mathbb{R}$ and $M$ contains no open interval, then there is a $z \notin M$ between $x$ and $y$).

**Lemma 12.7.** *The set $\Sigma_N$ is a Cantor set.*

**Proof.** We first prove that $\Sigma_N$ is compact. We need to show that every sequence $x^n$ contains a convergent subsequence. Given $x^n$, we can find a subsequence $x^{0,n}$ such that $x_0^{0,n}$ is the same for all $n$. Proceeding inductively, we obtain subsequences $x^{m,n}$ such that $x_k^{j,n} = x_k^{m,n}$ is the same for all $n$ if $0 \le k \le j \le m$. Now observe that $x^{n,n}$ is a subsequence which converges since $x_j^{n,n} = x_j^{m,m}$ for all $j \le \min(m,n)$.

To see that $\Sigma_N$ is perfect, fix $x$ and define $x^n$ such that $x_j^n = x_j$ for $0 \le j \le n$ and $x_{n+1}^n \ne x_{n+1}$. Then $x \ne x^n$ and $x^n$ converges to $x$.

To see that $\Sigma_N$ is totally disconnected, observe that the map $\delta_{j_0} : \Sigma_N \to \{0, \dots, N-1\}$, $x \mapsto x_{j_0}$ is continuous. Hence the set $U = \{x | x_{j_0} = c\} = \delta_{j_0}^{-1}(c)$ for fixed $j_0$ and $c$ is open and so is $V = \{x | x_{j_0} \ne c\}$. Now let $x, y \in \Sigma_N$, if $x \ne y$ there is a $j_0$ such that $x_{j_0} \ne y_{j_0}$. Now take $c = x_{j_0}$ then $U$ and $V$ from above are disjoint open sets whose union is $\Sigma_N$ and which contain $x$ and $y$ respectively. □

On $\Sigma_N$ we have the **shift map**

$$\begin{array}{rccc} \sigma: & \Sigma_N & \to & \Sigma_N \\ & x_n & \mapsto & x_{n+1} \end{array}, \tag{12.32}$$

which is uniformly continuous since we have

$$d(\sigma(x), \sigma(y)) \le N d(x, y). \tag{12.33}$$

Furthermore, it is chaotic as we will prove now. Observe that a point $x$ is periodic for $\sigma$ if and only if it is a periodic sequence.

**Lemma 12.8.** *The shift map has a countable number of periodic points which are dense.*

**Proof.** Since a sequence satisfying $\sigma^n(x) = x$ is uniquely determined by its first $n$ coefficients, there are precisely $N^n$ solutions to this equation. Hence there are countably many periodic orbits. Moreover, if $x$ is given we can define $x^n$ by taking the first $n$ coefficients of $x$ and then repeating them periodically. Then $x^n$ is a sequence of periodic points converging to $x$. Hence the periodic points are dense. □

**Lemma 12.9.** *The shift map has a dense orbit.*

**Proof.** Construct an orbit as follows. Start with the values $0, \dots, N-1$ as first coefficients. Now add all $N^2$ two digit combinations of $0, \dots, N-1$. Next add all $N^3$ three digit combinations. Proceeding inductively we obtain a sequence $x$. For example for $N = 2$ we have to take $0, 1; 00, 01, 10, 11; \dots$, that is, $x = (0, 1, 0, 0, 0, 1, 1, 0, 1, 1, \dots)$. I claim that the orbit of $x$ is dense. In fact, let $y$ be given. The first $n$ coefficients of $y$ appear as a block somewhere in $x$ by construction. Hence shifting $x$ $k$ times until this block reaches the start, we have $d(y, \sigma^k(x)) \le N^{-n}$. Hence the orbit is dense. □

Combining the two lemmas we see that $(\Sigma_N, \sigma)$ is chaotic. I leave it as an exercise to show that $\sigma$ has sensitive dependence on initial conditions directly.

It turns out that, as we have already seen in the previous section, many dynamical systems (or at least some subsystem) can be shown to be topologically equivalent to the shift map. Hence it is the prototypical example of a chaotic map.

However sometimes it is also necessary to consider only certain subsets of $\Sigma_N$ since it might turn out that only certain transitions are admissible in a given problem. For example, consider the situation in the previous section. There we had $\Sigma_2$ and, for $x \in \Sigma_2$, $x_n$ told us whether the $n$-th iterate is in $I_0$ or $I_1$. Now for a different system it could be that a point starting in $I_1$ could never return to $I_1$ once it enters $I_0$. In other words, a zero can never be followed by a one. Such a situation can be conveniently described by introducing a transition matrix.

A **transition matrix** $A$ is an $N \times N$ matrix all whose entries are zero or one. Suppose the ordered pair $j, k$ may only appear as adjacent entries in the sequence $x$ if $A_{j,k} = 1$. Then the corresponding subset is denoted by

$$\Sigma_N^A = \{x \in \Sigma_N | A_{x_n, x_{n+1}} = 1 \text{ for all } n \in \mathbb{N}_0\}. \tag{12.34}$$

Clearly $\sigma$ maps $\Sigma_N^A$ into itself and the dynamical system $(\Sigma_N^A, \sigma)$ is called a **subshift of finite type**. It is not hard to see that $\Sigma_N^A$ is a closed subset of $\Sigma_N$ and thus compact. Moreover, $\sigma$ is continuous on $\Sigma_N^A$ as the restriction of a continuous map.

Now let us return to our example. Here we have

$$A = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}. \tag{12.35}$$

A quick reflection shows that the only sequences which are admissible are those which contain finitely many ones first (maybe none) and then only zeroes. In particular, all points are eventually fixed and converge to the only fixed point $x = (0, 0, 0, \dots)$. So the system is definitely not chaotic. The same is true for all other possibilities except

$$A = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \tag{12.36}$$

in which case we have $\Sigma_2^A = \Sigma_2$. Hence we need an additional condition to ensure that the subshift is chaotic.

A transition matrix is called **irreducible** if there is an integer $l \in \mathbb{N}$ such that $A^l_{j,k} \neq 0$ for all $0 \leq j, k \leq N-1$. The following lemma is the key ingredient for our proof that irreducible subshifts are chaotic.

**Lemma 12.10.** *Let $A$ be a transition matrix and let $(x_1, \dots, x_k)$ be an admissible block of length $k$, that is $A_{x_j, x_{j+1}} = 1$ for $1 \leq j \leq k-1$. Then, if $A$ is irreducible and $l$ is as above, there is an admissible block $(x_1, \dots, x_{l-1})$ such that $(j, x_1, \dots, x_{l-1}, k)$ is admissible for all $0 \leq j, k \leq N-1$.*

**Proof.** Fix $j, k$ and note that

$$A^l_{j,k} = \sum_{x_1, \dots, x_{l-1}} A_{j,x_1} A_{x_1,x_2} \cdots A_{x_{l-2},x_{l-1}} A_{x_{l-1},k} \neq 0. \tag{12.37}$$

Hence at least one product in the sum must be one. Consequently all terms in this product must be one and we have found a block with the required property. □

This lemma ensures that, if $A$ is irreducible, there is an admissible block of length $l-1$ such that we can glue admissible blocks to both ends in such a way that the resulting block is again admissible!

As first application we prove

**Lemma 12.11.** *Suppose $A$ is irreducible, then $\Sigma_N^A$ is a Cantor set.*

**Proof.** As noted earlier, $\Sigma_N^A$ is compact. Moreover, as the subset of a totally disconnected set it is totally disconnected. Now let $x \in \Sigma_N^A$ be given. To show that there are points arbitrarily close to $x$ start by taking the first $n$ coefficients and add our admissible block of length $l-1$ from Lemma 12.10 to the end. Next add a single coefficient to the end such that the resulting block is different from the corresponding one of $x$. Finally, add our admissible block of length $l-1$ recursively to fill up the sequence. The constructed point can be made arbitrarily close to $x$ by choosing $n$ large and so we are done. □

As second application we show that $(\Sigma_N^A, \sigma)$ is chaotic.

**Lemma 12.12.** *Suppose $A$ is irreducible, then the shift map on $\Sigma_N^A$ has a countable number of periodic points which are dense.*

**Proof.** The proof is similar to the last part of the previous proof. We first show that the periodic points are dense. Let $x$ be given and take the first $n$ coefficients and add our admissible block of length $l-1$ from Lemma 12.10 to the end. Now take this entire block and repeat it periodically. The rest is straightforward. □

**Lemma 12.13.** *Suppose $A$ is irreducible, then the shift map on $\Sigma_N^A$ has a dense orbit.*

**Proof.** The proof is as in the case of the full shift. Take all admissible blocks of length $1, 2, 3, \dots$ and glue them together using our admissible block of length $l-1$ from Lemma 12.10. □

Finally, let me remark that similar results hold if we replace $\mathbb{N}_0$ by $\mathbb{Z}$. Let $N \in \mathbb{N}\backslash\{1\}$ and define the

$$\Sigma_N = \{0, 1, \dots, N-1\}^{\mathbb{Z}} \tag{12.38}$$

to be the set of doubly infinite sequences taking only the values $0, \dots, N-1$. Defining

$$d(x, y) = \frac{1}{2} \sum_{n \in \mathbb{N}_0} \frac{|x_n - y_n| + |x_{-n} - y_{-n}|}{N^n}, \tag{12.39}$$

$\Sigma_N$ becomes a metric space. Again we have

**Lemma 12.14.** *We have $d(x, y) \le N^{-n}$ if $x_j = y_j$ for all $|j| \le n$ and we have $d(x, y) \ge N^{-n}$ if $x_j \ne y_j$ for at least one $|j| \le n$.*

The shift map $\sigma$ is defined as before. However, note that $\sigma$ is invertible in this case. All other results hold with no further modifications. The details are left to the reader.

## 12.6. Strange attractors/repellors and fractal sets

A compact invariant set $\Lambda$, $f(\Lambda) = \Lambda$, is called **attracting** if there is a neighborhood $U$ of $\Lambda$ such that $d(f^n(x), \Lambda) \to 0$ as $n \to \infty$ for all $x \in U$. A compact invariant set $\Lambda$, $f(\Lambda) = \Lambda$, is called **repelling** if there is a neighborhood $U$ of $\Lambda$ such that for all $x \in U\backslash\Lambda$ there is an $n$ such that $f^n(x) \notin U$.

For example, let $f(x) = x^3$, then $\{0\}$ is an attracting set and $[-1, 1]$ is an repelling set. To exclude sets like $[-1, 1]$ in the above example we will introduce another condition. An attracting respectively repelling set is called an **attractor** respectively **repellor** if it is topologically transitive.

If $f$ is differentiable, there is a simple criterion when an invariant set is attracting respectively repelling.

**Theorem 12.15.** *Suppose $f : I \to I$ is continuously differentiable and $\Lambda$ is a compact invariant set. If there is an $n_0 \in \mathbb{N}$ such that $|f^{n_0}(x)| < 1$ for all $x \in \Lambda$, then $\Lambda$ is attracting. Similarly, if there is an $n_0 \in \mathbb{N}$ such that $|f^{n_0}(x)| > 1$ for all $x \in \Lambda$, then $\Lambda$ is repelling.*

**Proof.** We only prove the first claim, the second is similar. Choose $\alpha$ such that $\max_{x\in\Lambda} |f'(x)| < \alpha < 1$. For every $y$ in $\Lambda$ there is a (nonempty) open interval $I_y$ containing $y$ such that $|f'(x)| \le \alpha$ for all $x \in I_y$. Now let $U$ be the union of all those intervals. Fix $x \in U$ and let $y \in \Lambda$ be such that $d(x, \Lambda) = |x - y|$. Then, by the mean value theorem, $d(f^{n_0}(x), \Lambda) \le |f^{n_0}(x) - f^{n_0}(y)| \le \alpha|x - y| = \alpha d(x, \Lambda)$. Hence $d(f^{n_0 n}(x), \Lambda) \to 0$ and by continuity of $f$ and invariance of $\Lambda$ we also have $d(f^{n_0 n+j}(x), \Lambda) \to 0$ for $0 \le j \le n_0$. Thus the claim is proven. □

Repelling, attracting sets as above are called **hyperbolic repelling, attracting sets**, respectively.

An attractor, repellor $\Lambda$ is called **strange** if the dynamical system $(\Lambda, f)$ is chaotic and if $\Lambda$ is **fractal**.

We have already learned what the first condition means, but you might not know what fractal means. The short answer is that a set is called fractal if its Hausdorff dimension is not an integer. However, since you might also not know what the Hausdorff dimension is, let me give you the long answer as well.

I will first explain what the Hausdorff measure is, omitting all technical details (which can be found e.g. in [**24**]).

Recall that the **diameter** of a (nonempty) subset $U$ of $\mathbb{R}^n$ is defined by $d(U) = \sup_{x,y\in U} |x - y|$. A **cover** $\{V_j\}$ of $U$ is called a $\delta$**-cover** if it is countable and if $d(V_j) \le \delta$ for all $j$.

For $U$ a subset of $\mathbb{R}^n$ and $\alpha \ge 0$, $\delta > 0$ we define

$$h^\alpha_\delta(U) = \inf\left\{ \sum_j d(V_j)^\alpha \middle| \{V_i\} \text{ is a } \delta\text{-cover of } U \right\} \in [0, \infty]. \tag{12.40}$$

As $\delta$ decreases the number of admissible covers decreases and hence the limit

$$h^\alpha(U) = \lim_{\delta\downarrow 0} h^\alpha_\delta(U) \tag{12.41}$$

exists. Moreover it is not hard to show that $h^\alpha(U) \le h^\alpha(V)$ if $U \subseteq V$ and that for countable unions we have

$$h^\alpha(\bigcup_j U_j) \le \sum_j h^\alpha(U_j). \tag{12.42}$$

Hence $h^\alpha$ is an **outer measure** and the resulting **measure** on the Borel $\sigma$-algebra is called $\alpha$ dimensional **Hausdorff measure**. As any measure it satisfies

$$\begin{aligned} h^\alpha(\emptyset) &= 0 \\ h^\alpha(\bigcup_j U_j) &= \sum_j h^\alpha(U_j) \end{aligned} \tag{12.43}$$

for any countable union of disjoint sets $U_j$. It follows that $h^0$ is the counting measure and it can be shown that $h^n(U) = c_n|U|$, where $|U|$ denotes the Lebesgue measure of $U$ and $c_n = \pi^{n/2}/2^n\Gamma(n/2-1)$ is the volume of a ball with diameter one in $\mathbb{R}^n$.

Using the fact that for $\lambda > 0$ the map $\lambda : x \mapsto \lambda x$ gives rise to a bijection between $\delta$-covers and $(\delta/\lambda)$-covers, we easily obtain the following scaling property of Hausdorff measures.

**Lemma 12.16.** *Let $\lambda > 0$ and $U$ be a Borel set of $\mathbb{R}^n$, then*

$$h^\alpha(\lambda U) = \lambda^\alpha h^\alpha(U). \tag{12.44}$$

Moreover, Hausdorff measures also behave nicely under uniformly Hölder continuous maps.

**Lemma 12.17.** *Suppose $f : U \to \mathbb{R}^n$ is uniformly Hölder continuous with exponent $\gamma > 0$, that is,*

$$|f(x) - f(y)| \le c|x-y|^\gamma \quad \textit{for all } x, y \in U, \tag{12.45}$$

*then*

$$h^\alpha(f(U)) \le c^\alpha h^{\alpha\gamma}(U). \tag{12.46}$$

**Proof.** A simple consequence of the fact that for every $\delta$-cover $\{V_j\}$ of a Borel set $U$, the set $\{f(U \cap V_j)\}$ is a $(c\delta^\gamma)$-cover for the Borel set $f(U)$. □

Now we are ready to define the Hausdorff dimension. First of all note that $h^\alpha_\delta$ is non increasing with respect to $\alpha$ for $\delta < 1$ and hence the same is true for $h^\alpha$. Moreover, for $\alpha \le \beta$ we have $\sum_j d(V_j)^\beta \le \delta^{\beta-\alpha}\sum_j d(V_j)^\alpha$ and hence

$$h^\beta_\delta(U) \le \delta^{\beta-\alpha}\, h^\alpha_\delta(U). \tag{12.47}$$

Thus if $h^\alpha(U)$ is finite, then $h^\beta(U) = 0$ for every $\beta > \alpha$. Hence there must be one value of $\alpha$ where the Hausdorff measure of a set jumps from $\infty$ to 0. This value is called the **Hausdorff dimension**

$$\dim_H(U) = \inf\{\alpha | h^\alpha(U) = 0\} = \sup\{\alpha | h^\alpha(U) = \infty\}. \tag{12.48}$$

It can be shown that the Hausdorff dimension of an $m$ dimensional submanifold of $\mathbb{R}^n$ is again $m$. Moreover, it is also not hard to see that we have $\dim_H(U) \le n$ (prove this! Hint: It suffices to take for $U$ the unit cube. Now split $U$ into $k^n$ cubes of length $1/k$.).

The following observations are useful when computing Hausdorff dimensions. First of all the Hausdorff dimension is monotone, that is, for $U \subseteq V$ we have $\dim_H(U) \le \dim_H(V)$. Furthermore, if $U_j$ is a (countable) sequence of Borel sets we have $\dim_H(\bigcup_j U_j) = \sup_j \dim_H(U_j)$ (prove this).

Using Lemma 12.17 it is also straightforward to show

**Lemma 12.18.** *Suppose $f : U \to \mathbb{R}^n$ is uniformly Hölder continuous with exponent $\gamma > 0$, that is,*

$$|f(x) - f(y)| \le c|x - y|^\gamma \quad \text{for all } x, y \in U, \tag{12.49}$$

*then*

$$\dim_H(f(U)) \le \frac{1}{\gamma} \dim_H(U). \tag{12.50}$$

*Similarly, if $f$ is bi-Lipschitz, that is,*

$$a|x - y| \le |f(x) - f(y)| \le b|x - y| \quad \text{for all } x, y \in U, \tag{12.51}$$

*then*

$$\dim_H(f(U)) = \dim_H(U). \tag{12.52}$$

We end this section by computing the Hausdorff dimension of the repellor $\Lambda$ of the tent map.

**Theorem 12.19.** *The Hausdorff dimension of the repellor $\Lambda$ of the tent map is*

$$\dim_H(\Lambda) = \frac{\ln(2)}{\ln(\mu)}, \qquad \mu \ge 2. \tag{12.53}$$

*In particular, it is a strange repellor.*

**Proof.** Let $\delta = \mu^{-n}$. Using the $\delta$-cover $I_{s_0,\dots,s_{n-1}}$ we see $h_\delta^\alpha(\Lambda) \le (\frac{2}{\mu^\alpha})^n$. Hence for $\alpha = d = \ln(2)/\ln(\mu)$ we have $h_\delta^d(\Lambda) \le 1$ implying $\dim_H(\Lambda) \le d$.

The reverse inequality is a little harder. Let $V_j$ be a cover. It is clearly no restriction to assume that all $V_j$ are intervals. Moreover, finitely many of these sets cover $\Lambda$ by compactness. Drop all others and fix $j$. For $V_j$ there is a $k$ such that

$$\frac{1 - 2\mu^{-1}}{\mu^k} \le |V_j| < \frac{1 - 2\mu^{-1}}{\mu^{k-1}}. \tag{12.54}$$

Since the distance of two intervals in $\Lambda_k$ is at least $\frac{1-2\mu^{-1}}{\mu^{k-1}}$ we can intersect at most one such interval. For $n \ge k$ we see that $V_j$ intersects at most $2^{n-k} = 2^n \mu^{dk} \le 2^n (1 - 2\mu^{-1})^{-d} |V_j|^d$ intervals of $\Lambda_n$.

Choosing $n$ larger than all $k$ (for all $V_j$) and using that we must intersect all $2^n$ intervals in $\Lambda^n$, we end up with

$$2^n \le \sum_j \frac{2^n}{(1 - 2\mu^{-1})^d} |V_j|^d \tag{12.55}$$

which together with our first estimate yields

$$(1 - \frac{2}{\mu})^d \leq h^d(\Lambda) \leq 1. \tag{12.56}$$

□

Observe that this result can also formally be derived from the scaling property of the Hausdorff measure by solving the identity

$$\begin{aligned} h^\alpha(\Lambda) &= h^\alpha(\Lambda \cap I_0) + h^\alpha(\Lambda \cap I_1) \\ &= \frac{1}{\mu^\alpha} h^\alpha(T_\mu(\Lambda \cap I_0)) + \frac{1}{\mu^\alpha} h^\alpha(T_\mu(\Lambda \cap I_1)) \\ &= \frac{2}{\mu^\alpha} h^\alpha(\Lambda) \end{aligned} \tag{12.57}$$

for $\alpha$. However, this is only possible if we already know that $0 < h^\alpha(\Lambda) < \infty$ for some $\alpha$.

## 12.7. Homoclinic orbits as source for chaos

In this section we want to show that similar considerations as for the tent map can be made for other maps as well. We start with the logistic map for $\mu > 4$. As for the tent map, it is not hard to show that that $L_\mu^n(x) \to -\infty$ if $x \in \mathbb{R}\backslash[0,1]$. Hence most points will escape to $-\infty$ and we want to find the points which stay in $[0,1]$ for all iterations.

Set $\Lambda_0 = [0,1]$, then $\Lambda_1 = L_\mu^{-1}(\Lambda_0)$ is given by

$$\Lambda_1 = I_0 \cup I_1 = [0, G_\mu(1)] \cup [1 - G_\mu(1), 1], \tag{12.58}$$

where

$$G_\mu(x) = \frac{1}{2} - \sqrt{\frac{1}{4} - \frac{x}{\mu}}, \quad L_\mu(G_\mu(x)) = x, \quad 0 \leq x \leq 1. \tag{12.59}$$

To make our life a little easier we will make the additional assumption that

$$L'_\mu(x) \geq \alpha > 1 \quad \text{for} \quad x \in I_0, \tag{12.60}$$

which implies $\mu > 2 + \sqrt{5} = 4.236$. The general case $\mu > 4$ can be found in the book by Robinson [**23**].

Now proceeding as in the case of the tent map, we see that there is a sequence of nesting sets $\Lambda_n$ consisting of $2^n$ subintervals $I_{s_0,\cdots,s_{n-1}}$, $s_j \in \{0,1\}$, defined recursively via $I_{0,s_0,\cdots,s_n} = G_\mu(I_{s_0,\cdots,s_n})$ and $I_{1,s_0,\cdots,s_n} = 1 - G_\mu(I_{s_0,\cdots,s_n})$. The only difference is that, since $L_\mu$ is not (piecewise) linear, we do not know the length of the interval $I_{s_0,\cdots,s_n}$. However, by our assumption (12.60), we know $G'_\mu(x) \leq \alpha^{-1}$ and thus $|I_{s_0,\cdots,s_n}| \leq \alpha^{-n-1}$. But this is all we have used for the tent map and hence the same proof shows

**Theorem 12.20.** *Suppose $\mu > 2+\sqrt{5}$. Then the logistic map $L_\mu$ leaves the set*

$$\Lambda = \bigcap_{n\in\mathbb{N}} \Lambda_n \subset [0,1] \tag{12.61}$$

*invariant. All points $x \in \mathbb{R}\backslash\Lambda$ satisfy $\lim_{n\to\infty} L_\mu^n(x) = -\infty$. The set $\Lambda$ is a Cantor set and the dynamical system $(\Lambda, L_\mu)$ is topologically equivalent to the shift on two symbols $(\Sigma_2, \sigma)$ by virtue of the itinerary map*

$$\begin{array}{rccl} \varphi: & \Lambda & \to & \Sigma_2 \\ & x & \mapsto & x_n = j \text{ if } L_\mu^n(x) \in I_j \end{array}. \tag{12.62}$$

*In particular, $(\Lambda, L_\mu)$ is chaotic.*

Clearly we also want to know whether the repellor $\Lambda$ of the logistic map is strange.

**Theorem 12.21.** *The Hausdorff dimension of the repellor $\Lambda$ of the logistic map satisfies*

$$d(\mu) \le \dim_H(\Lambda) \le d(\mu(1-2G_\mu(1))), \qquad d(x) = \frac{\ln(2)}{\ln(x)}. \tag{12.63}$$

*In particular, it is strange if $\mu > 2+\sqrt{8} = 4.828$.*

**Proof.** The proof is analogous to the one of Theorem 12.19. The only difference is that we have to use different estimates for $L_\mu'$ from above and below,

$$\mu(1-2G_\mu(1)) = \alpha \le |L_\mu'(x)| \le \beta = \mu, \quad x \in I_0 \cup I_1. \tag{12.64}$$

Using the $\delta$-cover $I_{s_0,\dots,s_{n-1}}$ we see $h^{d(\alpha)}(\Lambda) \le (a/\alpha)^{d(\alpha)}$ where $a = |I_0| = |I_1| = G_\mu(1)$.

Similarly, using that the distance of two intervals in $\Lambda_k$ is at least $\frac{b}{\beta^{k-1}}$, where $b = d(I_0, I_1) = 1 - 2G_\mu(1)$ we obtain

$$b^{d(\beta)} \le h^{d(\beta)}(\Lambda) \tag{12.65}$$

which finishes the proof. □

Well, if you look at the proof for a moment, you will see that only a few properties of the logistic map have been used in the proof. And it is easy to see that the same proof applies to the following more general situation.

**Theorem 12.22.** *Let $f: M \to M$ be a continuously differentiable interval map. Suppose there are two disjoint compact intervals $I_0$, $I_1$ such that $I_0 \cup I_1 \subseteq f(I_0)$, $I_0 \cup I_1 \subseteq f(I_1)$, and $1 < \alpha \le |f'(x)| \le \beta$ for all $x \in I_0 \cup I_1$. Set*

$$\Lambda = \{x \in I_0 \cup I_1 | f^n(x) \in I_0 \cup I_1 \text{ for all } n \in \mathbb{N}\} \tag{12.66}$$

*and define the itinerary map as*

$$\begin{array}{rccl} \varphi : & \Lambda & \to & \Sigma_2 \\ & x & \mapsto & x_n = j \text{ if } f^n(x) \in I_j \end{array} . \tag{12.67}$$

*Then the set $\Lambda$ is a Cantor set and the dynamical system $(\Lambda, f)$ is topologically equivalent to the shift on two symbols $(\Sigma_2, \sigma)$. The Hausdorff dimension of $\Lambda$ satisfies*

$$d(\beta) \le \dim_H(\Lambda) \le d(\alpha), \qquad d(x) = \frac{\ln(2)}{\ln(x)}, \tag{12.68}$$

*and it is strange if $\alpha > 2$.*

**Proof.** By assumption, the restricted maps $f : I_0 \to f(I_0)$ and $f : I_1 \to f(I_1)$ are invertible. Denote by $g_0 : f(I_0) \to I_0$ and $g_1 : f(I_1) \to I_1$ the respective inverses. Now proceeding as usual, we see that there is a sequence of nesting sets $\Lambda_n$ consisting of $2^n$ subintervals $I_{s_0,\cdots,s_{n-1}}$, $s_j \in \{0,1\}$, defined recursively via $I_{0,s_0,\cdots,s_n} = g_0(I_{s_0,\cdots,s_n})$ and $I_{1,s_0,\cdots,s_n} = g_1(I_{s_0,\cdots,s_n})$. By assumption we also know at least $|I_{s_0,\cdots,s_n}| \le \alpha^{-n}|I_{s_0}|$ and hence the proof follows as before. □

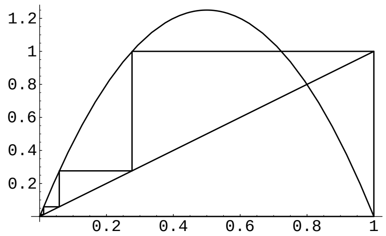
You should try to draw a picture for $f$ as in the above theorem. Moreover, it clearly suffices to assume that $f$ is absolutely continuous on $I_0 \cup I_1$.

Next, let $f$ be as in Theorem 12.22 and note that $I_0 \subseteq f(I_0)$ implies that there is a (unique) fixed point $p \in I_0$. Since $I_0 \subseteq f(I_1)$ there is a point $q \in I_1$ such that $f(q) = p$. Moreover, denoting by $g_0 : f(I_0) \to I_0$ the inverse of $f : I_0 \to f(I_0)$, we see that there is a whole sequence $g_0^n(q)$ which converges to $p$ as $n \to \infty$. In the case of the logistic map we can take $q = G_\mu(1)$.

In[3]:= $\mu = 5;$

$$\mathtt{x_0} = \mathtt{Nest}\left[\left(\frac{1}{2} - \sqrt{\frac{1}{4} - \frac{\#}{\mu}}\right)\&, 1., 5\right];$$

$\mathtt{ShowWeb}[\mu\#(1 - \#)\&, \mathtt{x_0}, 6];$




The fact that $x_0$ reaches the fixed point 0 after finitely many iterations (and not only asymptotically) is related to dimension one. Since the fixed point 0 is repelling ($T_\mu'(0) = \mu > 1$) it cannot converge to 0 unless it reaches it after finitely many steps.

In general, let $f : I \to I$ be continuously differentiable. A fixed point $p$ is called a **hyperbolic repellor** if $|f'(p)| > 1$. Hence there is a closed interval $W$ containing $p$ such that $|f'(x)| \geq \alpha > 1$ for all $x \in W$. Moreover, by the inverse function theorem there is a local inverse $g : f(W) \to W$ such that $g(f(x)) = x$, $x \in W$. Note that $g$ is a contraction. A point $q \in W$ is called a **homoclinic point** if there exists an $l \in \mathbb{N}_0$ such that $f^l(q) = p$. The set $\gamma(q) = \{f^j(q)|j \in \mathbb{N}_0\} \cup \{g^j(q)|j \in \mathbb{N}\}$ is called the corresponding **homoclinic orbit**. It is called nondegenerate if $(f^l)'(q) \neq 0$ (which implies $f'(x) \neq 0$ for all $x \in \gamma(q)$. A hyperbolic repellor with a homoclinic orbit is also called a **snap back repellor**.

**Theorem 12.23.** *Suppose $f \in C^1(I, I)$ has a repelling hyperbolic fixed point $p$ and a corresponding nondegenerate homoclinic point $q$.*

*In every sufficiently small neighborhood $U$ of $p$ there is an $n \in \mathbb{N}$ and an $f^n$ invariant Cantor set $\Lambda$ (i.e., $f^n(\Lambda) = \Lambda$) such that $(f^n, \Lambda)$ is topologically equivalent to the shift on two symbols $(\Sigma_2, \sigma)$.*

**Proof.** We will need to construct two disjoint intervals $I_j \subset U \cap W$, $j = 0, 1$, as in Theorem 12.22 for the map $F = f^n$ with $n$ suitable. By shrinking $W$ it is no restriction to assume $W \subseteq U$.

The idea is to take compact intervals $I_0$ containing $p$ and $I_1$ containing $q$. Since $f^l(q) = p$, the interval $f^l(I_1)$ contains again $p$. Taking sufficiently many iterations we can blow up both intervals such that the iterated images contain both original ones. The only tricky part is to ensure that the derivative of the iterated map is larger than one.

So we start with an interval $I_1 \subset W$ containing $q \in W$. Since $q$ is nondegenerate we can choose $I_1$ such that $|(f^l)'(x)| \geq \varepsilon > 0$ for all $x \in I_1$. Moreover, by shrinking $I_1$ if necessary we can also assume $f^l(I_1) \cap I_1 = \emptyset$. Next pick $m$ so large that $g^m(I_1) \subseteq f^l(I_1)$ ($g$ being the local inverse of $f$ as above) and $\alpha^m \varepsilon > 1$. Set $n = m + l$. Since $g^m(W)$ contains $p$ and $g^m(I_1)$ we can further shrink $I_1$ such that $f^l(I_1) \subseteq g^m(W)$, that is, $f^n(I_1) \subseteq W$. By construction $|(f^n)'(x)| \geq \varepsilon\alpha^m > 1$ for $x \in I_1$.

Next we will choose $I_0 = g^l(f^l(I_1))$. Then we have $I_0 \cap I_1 = \emptyset$ and $I_0 \subseteq f^n(I_1)$ since $I_0 \subseteq f^l(I_1)$. Furthermore, by $p \in I_0$ we have $I_0 \subseteq f^n(I_0)$ and by $g^m(I_1) \subseteq f^l(I_1) = f^l(I_0)$ we have $I_1 \subseteq f^n(I_0)$. Finally, since $I_0 \subseteq g^n(W)$ we have $|(f^n)'(x)| \geq \alpha^n > 1$ for $x \in I_0$ and we are done. $\square$

Why is the degeneracy condition necessary? Can you give a counter example?

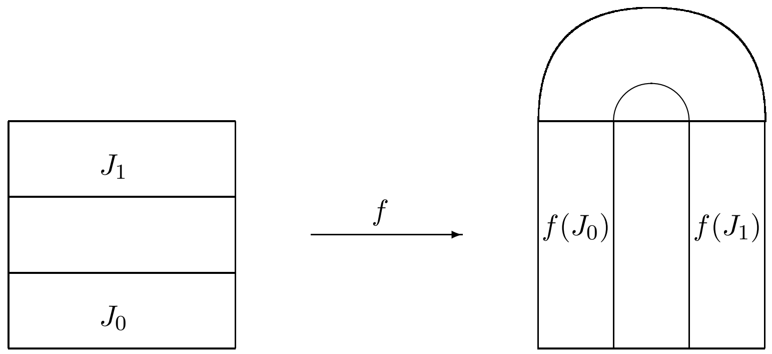
*Chapter 13*

# Chaos in higher dimensional systems

## 13.1. The Smale horseshoe

In this section we will consider a two dimensional analog of the tent map and show that it has an invariant Cantor set on which the dynamics is chaotic. We will see in the following section that it is a simple model for the behavior of a map in the neighborhood of a hyperbolic fixed point with a homoclinic orbit.

The **Smale horseshoe map** $f : D \to \mathbb{R}^2$, $D = [0,1]^2$, is defined by contracting the $x$ direction, expanding the $y$ direction, and then twist the result around as follows.




Since we are only interested in the dynamics on $D$, we only describe this part of the map analytically. We fix $\lambda \in (0, \frac{1}{2}]$, $\mu \in [2, \infty)$, set

$$J_0 = [0,1] \times [0, \frac{1}{\mu}], \qquad J_1 = [0,1] \times [1 - \frac{1}{\mu}, 1], \tag{13.1}$$

and define

$$f : J_0 \to f(J_0), \quad (x, y) \mapsto (\lambda x, \mu y), \tag{13.2}$$

respectively

$$f : J_1 \to f(J_1), \quad (x, y) \mapsto (1 - \lambda x, \mu(1 - y)). \tag{13.3}$$

A look at the two coordinates shows that $f_1(x, y) \in [0, 1]$ whenever $x \in [0, 1]$ and that $f_2(x, y) = T_\mu(y)$. Hence if we want to stay in $D$ during the first $n$ iterations we need to start in $\Lambda_{+,n} = [0, 1] \times \Lambda_n(T_\mu)$, where $\Lambda_n(T_\mu) = \Lambda_n$ is the same as for $T_\mu$. In particular, if we want to stay in $D$ for all positive iterations we have to start in

$$\Lambda_+ = [0,1] \times \Lambda(T_\mu) = \bigcap_{n \in \mathbb{N}_0} f^n(D). \tag{13.4}$$

But note that $f$ is invertible, with inverse given by

$$g = f^{-1} : K_0 = f(J_0) \to J_0, \quad (x, y) \mapsto (\lambda^{-1} x, \mu^{-1} y), \tag{13.5}$$

respectively

$$g = f^{-1} : K_1 = f(J_1) \to J_1, \quad (x, y) \mapsto (\lambda^{-1}(1 - x), 1 - \mu^{-1} y). \tag{13.6}$$

Hence, by the same consideration, if we want to stay in $D$ for all negative iterations, we have to start in

$$\Lambda_- = \Lambda(T_{1/\lambda}) \times [0,1] = \bigcap_{n \in \mathbb{N}_0} f^{-n}(D). \tag{13.7}$$

Finally, if we want to stay in $D$ for all (positive and negative) iterations we have to start in

$$\Lambda = \Lambda_- \cap \Lambda_+ = \Lambda(T_{1/\lambda}) \times \Lambda(T_\mu). \tag{13.8}$$

The set $\Lambda$ is a Cantor set since any product of two Cantor sets is again a Cantor set (prove this).

Now by our considerations for the tent map, the $y$ coordinate of every point in $\Lambda$ can uniquely defined by a sequence $y_n$, $n \in \mathbb{N}_0$. Similarly, the $x$ coordinate of every point in $\Lambda$ can be uniquely defined by a sequence $x_n$, $n \in \mathbb{N}_0$. Hence defining $s_n = y_n$ and $s_{-n} = x_{n-1}$ for $n \in \mathbb{N}_0$ we see that there is a one to one correspondence between points in $\Lambda$ and doubly infinite sequences on two symbols. Hence we have found again an itinerary map

$$\begin{array}{rccl} \varphi : & \Lambda & \to & \Sigma_2 \\ & (x, y) & \mapsto & s_n = \begin{cases} y_n & n \geq 0 \\ x_{-n-1} & n < 0 \end{cases}, \end{array} \tag{13.9}$$

where $y_n$ is defined by $f^n(x, y) \in J_{y_n}$ and $x_n$ is defined by $g^n(x, y) \in K_{x_n}$. As in the case of the tent map it is easy to see $\varphi$ is continuous (exercise). Now what about the action of $\sigma = \varphi \circ f \circ \varphi^{-1}$? By construction, $\sigma$ shifts $y_n$ to the left, $\sigma(s)_n = y_{n+1}$, $n \geq 0$, and $\sigma^{-1}$ shifts $x_n$ to the left, $\sigma^{-1}(s)_n = x_{-n-1}$, $n < 0$. Hence $\sigma$ shifts $x_n$ to the right, $\sigma(s)_n = x_{-n-2}$, $n < -1$, and we need to figure out what the new first element $\sigma(s)_{-1}$ is. Well, since $(x, y) \in J_{y_0}$ is equivalent to $f(x, y) \in K_{y_0}$, we see that this element is $\sigma(s)_{-1} = y_0$ and hence $\sigma$ just shifts $s_n$ to the left, $\sigma(s)_n = s_{n+1}$. In summary, we have shown

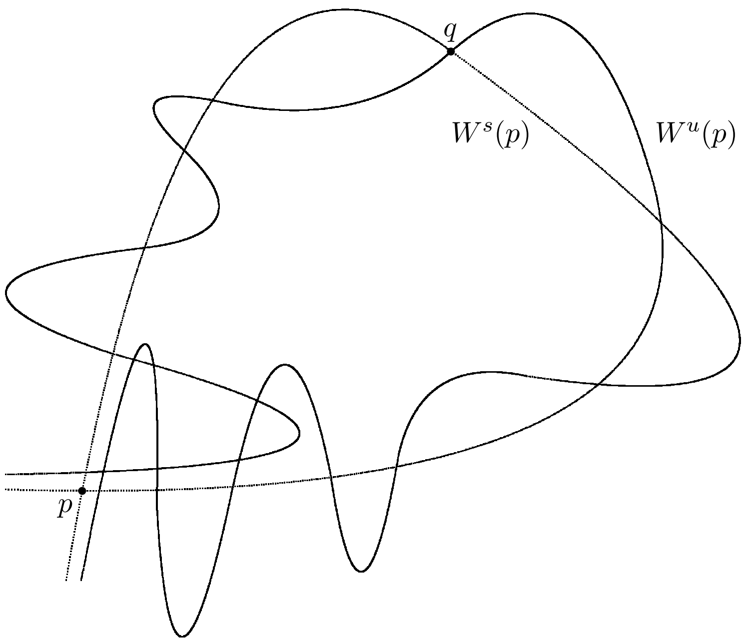
**Theorem 13.1.** *The Smale horseshoe map has an invariant Cantor set $\Lambda$ on which the dynamics is equivalent to the double sided shift on two symbols. In particular it is chaotic.*

## 13.2. The Smale-Birkhoff homoclinic theorem

In this section I will present the higher dimensional analog of Theorem 12.23.

Let $f$ be a diffeomorphism ($C^1$) and suppose $p$ is a hyperbolic fixed point. A **homoclinic point** is a point $q \neq p$ which is in the stable and unstable manifold. If the stable and unstable manifold intersect transversally at $q$, then $q$ is called **transverse**. This implies that there is a homoclinic orbit $\gamma(q) = \{q_n\}$ such that $\lim_{n\to\infty} q_n = \lim_{n\to-\infty} q_n = p$. Since the stable and unstable manifolds are invariant, we have $q_n \in W^s(p) \cap W^u(p)$ for all $n \in \mathbb{Z}$. Moreover, if $q$ is transversal, so are all $q_n$ since $f$ is a diffeomorphism.

The typical situation is depicted below.



This picture is known as **homoclinic tangle**.

**Theorem 13.2** (Smale–Birkhoff)**.** *Suppose $f$ is a diffeomorphism with a hyperbolic fixed point $p$ and a corresponding transversal homoclinic point $q$. Then some iterate $f^n$ has a hyperbolic invariant set $\Lambda$ on which it is topologically equivalent to the bi-infinite shift on two symbols.*

The idea of proof is to find a horseshoe map in some iterate of $f$. Intuitively, the above picture shows that this can be done by taking an open set containing one peak of the unstable manifold between two successive homoclinic points. Taking iterations of this set you will eventually end up with a horseshoe like set around the stable manifold lying over our original set. For details see [**23**].

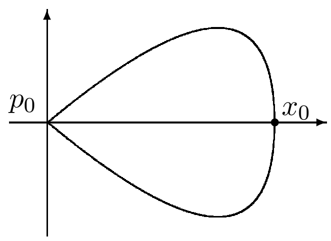
## 13.3. Melnikov's method for homoclinic orbits

Finally we want to combine the Smale–Birkhoff theorem from the previous section with Melnikov's method from Section 11.5 to obtain a criterion for chaos in ordinary differential equations.

Again we will start with a planar system

$$\dot{x} = f(x) \tag{13.10}$$

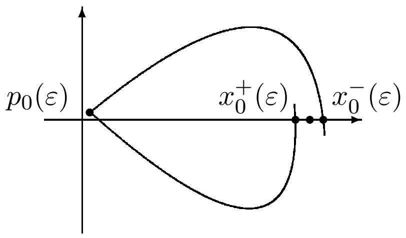
which has a homoclinic orbit $\gamma(x_0)$ at a fixed point $p_0$. For example, we could take Duffing's equation from Problem 7.4 (with $\delta = 0$). The typical situation for the unperturbed system is depicted below.




Now we will perturb this system a little and consider

$$\dot{x} = f(x) + \varepsilon\, g(x). \tag{13.11}$$

Since the original fixed point $p_0$ is hyperbolic it will persist for $\varepsilon$ small, lets call it $p_0(\varepsilon)$. On the other hand, it is clear that in general the stable and unstable manifold of $p_0(\varepsilon)$ will no longer coincide for $\varepsilon \neq 0$ and hence there is no homoclinic orbit at $p_0(\varepsilon)$ for $\varepsilon \neq 0$. Again the typical situation is displayed in the picture below




However, it is clear that we will not be able to produce chaos with such a perturbation since the Poincaré–Bendixson theorem implies that the motion of a planar system must be quite regular. Hence we need at least another dimension and hence we will take a nonautonomous perturbation and consider

$$\dot{x} = f(x) + \varepsilon\, g(\tau, x, \varepsilon), \qquad \dot{\tau} = 1, \tag{13.12}$$

where $g(\tau, x, \varepsilon)$ is periodic with respect to $\tau$, say $g(\tau + 2\pi, x, \varepsilon) = g(\tau, x, \varepsilon)$. We will abbreviate $z = (x, \tau)$.

Of course our pictures from above do no longer show the entire system but they can be viewed as a slice for some fixed $\tau = t_0$. Note that the first picture will not change when $\tau$ varies but the second will. In particular, $p_0(\tau, \varepsilon)$ will now correspond to a hyperbolic periodic orbit and the manifolds in our pictures are the intersection of the stable and unstable manifolds of $p_0(\tau, \varepsilon)$ with the plane $\Sigma = \{(x, \tau) | \tau = t_0\}$. Moreover, taking $\Sigma$ as the section of a corresponding Poincaré map $P_\Sigma$, these intersections are just the stable and unstable manifold of the fixed point $p_0(\varepsilon) = p_0(t_0, \varepsilon)$ of $P_\Sigma$. Hence if we can find a transverse intersection point, the Smale–Birkhoff theorem will tell us that there is an invariant Cantor set close to this point, where the Poincaré map is chaotic.

Now it remains to find a good criterion for the existence of such a transversal intersection. Replacing $g(\tau, x, \varepsilon)$ with $g(\tau - t_0, x, \varepsilon)$ it is no restriction to assume $t_0 = 0$. Denote the (un)stable manifold of the periodic orbit $(p_0, \tau)$ by $W(p_0) = \{(\Phi(x_0, s), \tau) | (s, \tau) \in \mathbb{R} \times S^1\}$. Then for any given point $z_0 = (x_0, t_0) \in W(p_0)$ a good measure of the splitting of the perturbed stable and unstable manifolds is the distance of the respective intersections points with the line through $z_0$ and orthogonal to the vector field. That is, denote by $z_0^+(\varepsilon)$, $z_0^-(\varepsilon)$ the intersection of the stable, unstable manifold with the line $\{(x_0 + u f(x_0)^\perp, 0) | u \in \mathbb{R}\}$, respectively. Then the separation of the manifolds is measured by

$$\Delta(z_0, \varepsilon) = f(x_0)^\perp (x_0^-(\varepsilon) - x_0^+(\varepsilon)) = f(x_0) \wedge (x_0^-(\varepsilon) - x_0^+(\varepsilon)). \tag{13.13}$$

Since $\Delta(z_0, 0) = 0$ we can apply the same analysis as in Section 11.4 to conclude that $\Delta(z_0, \varepsilon)$ has a zero for small $\varepsilon$ if $\frac{\partial \Delta}{\partial \varepsilon}(z_0, 0)$ has a simple zero. Moreover, if the zero of $\frac{\partial \Delta}{\partial \varepsilon}(z_0, 0)$ is simple, this is also equivalent to the fact that the intersection of the stable and unstable manifolds is transversal.

It remains to compute $\frac{\partial \Delta}{\partial \varepsilon}(z_0, 0)$ which can be done using the same ideas as in Section 11.4. Let $z^\pm(t, \varepsilon) = (x^\pm(t, \varepsilon), t)$ be the orbit in $W^\pm(\gamma(p_0(\varepsilon)))$ which satisfies $z^\pm(0, \varepsilon) = z_0^\pm(\varepsilon)$. Then we have

$$\frac{\partial \Delta}{\partial \varepsilon}(z_0, 0) = f(x_0) \wedge (x_\varepsilon^-(0) - x_\varepsilon^+(0)), \tag{13.14}$$

where $x_\varepsilon^\pm(t) = \frac{\partial}{\partial \varepsilon} x^\pm(t, \varepsilon)|_{\varepsilon=0}$ are solutions of the corresponding variational equation. However, since we do not know the initial conditions (we know only the asymptotic behavior), it is better to consider

$$y^\pm(t) = f(x_0(t)) \wedge x_\varepsilon^\pm(t), \quad x_0(t) = \Phi(t, x_0). \tag{13.15}$$

Using the variational equation

$$\dot{x}_\varepsilon^\pm(z_0, t) = A(t) x_\varepsilon^\pm(t) + g(t - t_0, x_0(t), 0), \quad A(t) = df_{x_0(t)}, \tag{13.16}$$

we obtain after a little calculation (Problem 13.1)

$$\dot{y}^\pm(t) = \operatorname{tr}(A(t)) y^\pm(t) + f(x_0(t)) \wedge g(t - t_0, x_0(t), 0) \tag{13.17}$$

and hence

$$\dot{y}^\pm(t) = \dot{y}^\pm(T_\pm) + \int_{T_\pm}^t \mathrm{e}^{\int_s^t \operatorname{tr}(A(r)) dr} f(x_0(s)) \wedge g(s - t_0, x_0(s), 0)\, ds. \tag{13.18}$$

Next, we want to get rid of the boundary terms at $T_\pm$ by taking the limit $T_\pm \to \pm\infty$. They will vanish provided $x_\varepsilon^\pm(T_\pm)$ remains bounded since $\lim_{t\to\pm\infty} f(x_0(t)) = f(p_0) = 0$. In fact, this is shown in the next lemma.

**Lemma 13.3.** *The stable and unstable manifolds of the perturbed periodic orbit $p_0(\varepsilon)$ are locally given by*

$$W^\pm(\gamma(p_0(\varepsilon))) = \{(\Phi(s, x_0) + h^\pm(\tau, s)\varepsilon + o(\varepsilon), \tau) | (s, \tau) \in S^1 \times \mathbb{R}\}, \tag{13.19}$$

*where $x_0 \in W(p_0)$ is fixed and $h^\pm(\tau, s)$ is bounded as $s \to \pm\infty$.*

**Proof.** By Theorem 11.10 a point in $W^\pm(\gamma(p_0(\varepsilon)))$ can locally be written as

$$(p_0 + h_0^\pm(\tau, a) + h_1^\pm(\tau, a)\varepsilon + o(\varepsilon), \tau). \tag{13.20}$$

Moreover, fixing $x_0 \in W(p_0)$ there is a unique $s = s(\tau, a)$ such that

$$p_0 + h_0^\pm(\tau, a, 0) = \Phi(s, x_0) \tag{13.21}$$

and hence we can choose $h^\pm(\tau, s) = h_1^\pm(\tau, a(\tau, s))$. □

Hence we even have

$$y^{\pm}(t) = \int_{\pm\infty}^{t} \mathrm{e}^{\int_s^t \mathrm{tr}(A(r))dr} f(x_0(s)) \wedge g(s - t_0, x_0(s), 0)\, ds \tag{13.22}$$

and thus finally

$$\frac{\partial \Delta}{\partial \varepsilon}(z_0, 0) = M_{x_0}(t_0), \tag{13.23}$$

where $M_{x_0}(t_0)$ is the **homoclinic Melnikov integral**

$$M_{x_0}(t) = \int_{-\infty}^{\infty} \mathrm{e}^{-\int_0^s \mathrm{div}(f(\Phi(r,x_0)))dr} f(\Phi(s, x_0)) \wedge g(s - t, \Phi(s, x_0), 0)\, ds. \tag{13.24}$$

Note that the base point $x_0$ on the homoclinic orbit is not essential since we have (Problem 13.2)

$$M_{\Phi(t,x_0)}(t_0) = \mathrm{e}^{\int_0^t \mathrm{div}(f(\Phi(r,x_0)))dr} M_{x_0}(t + t_0). \tag{13.25}$$

In summary we have proven

**Theorem 13.4** (Melnikov). *Suppose the homoclinic Melnikov integral $M_{x_0}(t)$ has a simple zero for some $t \in \mathbb{R}$, then the Poincaré map $P_\Sigma$ has a transversal homoclinic orbit for sufficiently small $\varepsilon \neq 0$.*

For example, consider the forced Duffing equation (compare Problem 7.4)

$$\dot{q} = p, \quad \dot{p} = q - q^3 - \varepsilon(\delta p + \gamma \cos(\omega \tau)), \quad \dot{\tau} = 1. \tag{13.26}$$

The homoclinic orbit is given by

$$q_0(t) = \sqrt{2}\,\mathrm{sech}(t), \quad p_0(t) = -\sqrt{2}\tanh(t)\mathrm{sech}(t) \tag{13.27}$$

and hence

$$\begin{aligned} M(t) &= \int_{-\infty}^{\infty} q_0(s)\,(\delta p_0(s) + \gamma \cos(\omega(s - t)))\, ds \\ &= \frac{4\delta}{3} - \sqrt{2}\pi\gamma\omega\mathrm{sech}(\frac{\pi\omega}{2}) \sin(\omega t) \end{aligned} \tag{13.28}$$

Thus the Duffing equation is chaotic for $\delta$, $\gamma$ sufficiently small provided

$$\left|\frac{\delta}{\gamma}\right| < \frac{3\sqrt{2}\pi|\omega|}{4}\mathrm{sech}(\frac{\pi\omega}{2}). \tag{13.29}$$

**Problem 13.1.** *Prove the following formula for $x, y \in \mathbb{R}^2$ and $A \in \mathbb{R}^2 \otimes \mathbb{R}^2$,*

$$Ax \wedge y + x \wedge Ay = \mathrm{tr}(A) x \wedge y.$$

**Problem 13.2.** *Show (13.25).*

**Problem 13.3.** *Apply the Melnikov method to the forced mathematical pendulum (compare Section 6.6)*

$$\dot{q} = p, \qquad \dot{q} = -\sin(q) + \varepsilon \sin(t).$$

The End

# Bibliography


[1] R. Abraham, J. E. Marsden, and T. Ratiu, *Manifolds, Tensor Analysis, and Applications*, $2^{\text{nd}}$ edition, Springer, New York, 1983.

[2] V.I. Arnold, *Mathematical methods of classical mechanics*, $2^{\text{nd}}$ ed., Springer, New York, 1989.

[3] V.I. Arnold, *Gewöhnliche Differentialgleichungen*, Springer, Berlin, 1980.

[4] F. Brauer and J.A. Nohel, *Ordinary Differential Equations: A First Course*, $2^{\text{nd}}$ edition, W.A. Benjamin, New York, 1973.

[5] C. Chicone, *Ordinary Differential Equations with Applications*, Springer, New York, 1999.

[6] E.A. Coddington and N. Levinson, *Theory of Ordinary Differential Equations*, McGraw-Hill, New York, 1955.

[7] R.L. Devaney, *An introduction to Chaotic Dynamical Systems*, $2^{\text{nd}}$ ed., Addison-Wesley, Redwood City, 1989.

[8] K. Falconer, *Fractal Geometry*, Benjamin/Clummings Publishing, Menlo Park, 1986.

[9] A. Gray, M. Mezzino, and M. A. Pinsky, *Introduction to Ordinary Differential Equations with Mathematica*, Springer, New York, 1997.

[10] J. Guckenheimer and P. Holmes, *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*, Springer, New York, 1983.

[11] P. Hartman, *Ordinary Differential Equations*, Wiley, New York, 1964.

[12] M. W. Hirsch and S. Smale, *Differential Equations, Dynamical Systems, and Linear Algebra*, Academic Press, San Diego, 1989.

[13] J. Hofbauer and K. Sigmund, *Evolutionary Games and Replicator Dynamics*, Cambridge University Press, Cambridge, 1998.

[14] R.A. Holmgren, *A First Course in Discrete Dynamical Systems*, 2nd ed.,Springer, New York, 1996.

[15] K. Jänich, *Analysis*, $2^{\text{nd}}$ ed., Springer, Berlin, 1990.

[16] E.L. Ince, *Ordinary Differential Equations*, Dover Publ., New York, 1956.

[17] E. Kamke, *Differentialgleichungen, I. Gewöhnliche Differentialgleichungen*, Springer, New York, 1997.

[18] B.M. Levitan and I.S. Sargsjan, *Introduction to Spectral Theory*, Amer. Math. Soc., Providence, 1975.

[19] J. Moser, *Stable and Random Motions in Dynamical Systems: With Special Emphasis on Celestial Mechanics*, Princeton University Press, Princeton, 2001.

[20] R.S. Palais, *The symmetries of solitons*, Bull. Amer. Math. Soc., **34**, 339–403 (1997).

[21] J. Palis and W. de Melo, *Geometric Theory of Dynamical Systems*, Springer, New York, 1982.

[22] L. Perko, *Differential Equations and Dynamical Systems*, $2^{\text{nd}}$ ed., Springer, New York, 1996.

[23] C. Robinson, *Dynamical Systems: Stability, Symbolic Dynamics, and Chaos*, CRC Press, Boca Raton, 1995.

[24] C.A. Rogers, *Hausdorff Measures*, Cambridge University Press, Cambridge, 1970.

[25] D. Ruelle, *Elements of Differentiable Dynamics and Bifurcation Theory*, Academic Press, San Diego, 1988.

[26] D. Schwalbe and S. Wagon, *VisualDSolve. Visualizing Differential Equations with Mathematica*, Springer, New York, 1997.

[27] C. Sparrow, *The Lorenz Equation, Bifurcations, Chaos and Strange Attractors*, Springer, New York, 1982.

[28] F. Verhulst, *Nonlinear Differential Equations and Dynamical Systems*, Springer, Berlin, 1990.

[29] W. Walter, *Gewöhnliche Differentialgleichungen*, Akademische Verlagsgesellschaft, Leipzig, 1962.

[30] J. Weidmann, *Linear Operators in Hilbert Spaces*, Springer, New York, 1980.

[31] S. Wiggins, *Global Bifurcations and Chaos*, $2^{\text{nd}}$ ed., Springer, New York, 1988.

[32] S. Wiggins, *Introduction to Applied Nonlinear Dynamical Systems and Chaos*, Springer, New York, 1990.

[33] S. Wolfram, *The Mathematica Book*, 4th ed., Wolfram Media/Cambridge University Press, Champaign/Cambridge, 1999.

[34] D. Zwillinger, *Handbook of Differential Equations*, Academic Press, San Diego, 1989.

# Glossary of notations

| | |
|---|---|
| $A_\pm$ | ... matrix $A$ restricted to $E^\pm(A)$. |
| $B_\varepsilon(x)$ | ... ball of radius $\varepsilon$ centered at $x$. |
| $C(U,V)$ | ... set of continuous functions from $U$ to $V$. |
| $C(U)$ | $= C(U,\mathbb{R})$ |
| $C^k(U,V)$ | ... set of $k$ times continuously differentiable functions. |
| $\mathbb{C}$ | ... the set of complex numbers |
| $\chi_A$ | ... Characteristic polynomial of $A$, 44 |
| $d(U)$ | ... diameter of $U$, 212 |
| $d(x,y)$ | ... distance in a metric space |
| $df_x$ | ... Jacobian of a differentiable mapping $f$ at $x$ |
| $E^0(A)$ | ... center subspace of a matrix, 46 |
| $E^\pm(A)$ | ... (un)stable subspace of a matrix, 46 |
| $\gamma(x)$ | ... orbit of $x$, 109 |
| $\gamma_\pm(x)$ | ... forward, backward orbit of $x$, 109 |
| $\mathfrak{H}_0$ | ... inner product space, 86 |
| $I_x$ | $= (T_-(x), T_+(x))$ |
| $L_\mu$ | ... logistic map, 174 |
| $\Lambda$ | ... a compact invariant set |
| $M^\pm$ | ... (un)stable manifold, 123, 188 |
| $\mathbb{N}$ | ... the set of positive integers |
| $\mathbb{N}_0$ | $= \mathbb{N} \cup \{0\}$ |
| $o(.)$ | ... Landau symbol |
| $O(.)$ | ... Landau symbol |
| $\Omega(f)$ | ... set of nonwandering points, 113 |

| | |
|---|---|
| $P_\Sigma(y)$ | ... Poincaré map, 112 |
| $\Phi(t, x_0)$ | ... flow of a dynamical system, 107 |
| $\Pi(t, t_0)$ | ... principal matrix of a linear system, 55 |
| $\mathbb{R}$ | ... the set of reals |
| $\sigma$ | ... shift map on $\Sigma_N$, 208 |
| $\sigma(A)$ | ... spectrum (set of eigenvalues) of a matrix |
| $\Sigma_N$ | ... sequence space over $N$ symbols, 207 |
| $T_\pm(x)$ | ... positive, negative lifetime of $x$, 109 |
| $T(x)$ | ... period of $x$ (if $x$ is periodic), 109 |
| $T_\mu$ | ... tent map, 203 |
| $\omega_\pm(x)$ | ... positive, negative $\omega$-limit set of $x$, 111 |
| $W^\pm$ | ... (un)stable set, 123, 151 , 176 |
| $\mathbb{Z}$ | ... the set of integers |
| $z$ | ... a complex number |
| $\sqrt{z}$ | ... square root of $z$ with branch cut along $(-\infty, 0)$ |
| $z^*$ | ... complex conjugation |
| $\Vert . \Vert$ | ... norm |
| $\langle ., .. \rangle$ | ... scalar product in $\mathfrak{H}_0$, 86 |
| $(\lambda_1, \lambda_2)$ | $= \{\lambda \in \mathbb{R} \,\vert\, \lambda_1 < \lambda < \lambda_2\}$, open interval |
| $[\lambda_1, \lambda_2]$ | $= \{\lambda \in \mathbb{R} \,\vert\, \lambda_1 \le \lambda \le \lambda_2\}$, closed interval |

# Index