Ordinary Differential Equations and Dynamical Systems

Gerald Teschl

To Susanne, Simon, and Jakob

Contents

Part 3. Chaos

Preface

About

When you publish a textbook on such a classical subject the first question you will be faced with is: Why the heck another book? Well, everything started when I was supposed to give the basic course on *Ordinary Differential Equations* in Summer 2000 (which at that time met 5 hours per week). While there were many good books on the subject available, none of them quite fitted my needs. I wanted a concise but rigorous introduction with full proofs also covering classical topics such as Sturm–Liouville boundary value problems, differential equations in the complex domain as well as modern aspects of the qualitative theory of differential equations. The course was continued with a second part on *Dynamical Systems and Chaos* in Winter 2000/01 and the notes were extended accordingly. Since then the manuscript has been rewritten and improved several times according to the feedback I got from students over the years when I redid the course. Moreover, since I had the notes on my homepage from the very beginning, this triggered a significant amount of feedback as well. Beginning from students who reported typos, incorrectly phrased exercises, etc. over colleagues who reported errors in proofs and made suggestions for improvements, to editors who approached me about publishing the notes. Last but not least, this also resulted in a chinese translation. Moreover, if you google for the manuscript, you can see that it is used at several places worldwide, linked as a reference at various sites including Wikipedia. Finally, Google Scholar will tell you that it is even cited in several publications. Hence I decided that it is time to turn it into a *real* book.

Content

Its main aim is to give a self contained introduction to the field of ordinary differential equations with emphasis on the dynamical systems point of view while still keeping an eye on classical tools as pointed out before.

The first part is what I typically cover in the introductory course for bachelor students. Of course it is typically not possible to cover everything and one has to skip some of the more advanced sections. Moreover, it might also be necessary to add some material from the first chapter of the second part to meet curricular requirements.

The second part is a natural continuation beginning with planar examples (culminating in the generalized Poincaré–Bendixon theorem), continuing with the fact that things get much more complicated in three and more dimensions, and ending with the stable manifold and the Hartman–Grobman theorem.

The third and last part gives a brief introduction to chaos focusing on two selected topics: Interval maps with the logistic map as the prime example plus the identification of homoclinic orbits as a source for chaos and the Melnikov method for perturbations of periodic orbits and for finding homoclinic orbits.

Prerequisites

It only requires some basic knowledge from calculus, complex functions, and linear algebra which should be covered in the usual courses. In addition, I have tried to show how a computer system, *Mathematica*, can help with the investigation of differential equations. However, the course is not tied to *Mathematica* and any similar program can be used as well.

Updates

The AMS is hosting a web page for this book at

`http://www.ams.org/bookpages/gsm-XXX/`

where updates, corrections, and other material may be found, including a link to material on my own web site:

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

There you can also find an accompanying *Mathematica* notebook with the code from the text plus some additional material. **Please do not put a**

copy of this file on your personal webpage but link to the page above.

Acknowledgments

I wish to thank my students, Ada Akerman, Kerstin Ammann, Jörg Arnberger, Alexander Beigl, Paolo Capka, Jonathan Eckhardt, Michael Fischer, Anna Geyer, Ahmed Ghneim, Hannes Grimm-Strele, Tony Johansson, Klaus Kröncke, Alice Lakits, Simone Lederer, Oliver Leingang, Johanna Michor, Thomas Moser, Markus Müller, Andreas Németh, Andreas Pichler, Tobias Preinerstorfer, Jin Qian, Dominik Rasipanov, Martin Ringbauer, Simon Rößler, Robert Stadler, Shelby Stanhope, Raphael Stuhlmeier, Gerhard Tulzer, Paul Wedrich, Florian Wisser, and colleagues, Edward Dunne, Klemens Fellner, Giuseppe Ferrero, Ilse Fischer, Delbert Franz, Heinz Hanßmann, Daniel Lenz, Jim Sochacki, and Eric Wahlén, who have pointed out several typos and made useful suggestions for improvements. Finally, I also like to thank the anonymous referees for valuable suggestions improving the presentation of the material.

If you also find an error or if you have comments or suggestions (no matter how small), please let me know.

I have been supported by the Austrian Science Fund (FWF) during much of this writing, most recently under grant Y330.

Gerald Teschl

Vienna, Austria
April 2012

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

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

Part 1

Classical theory

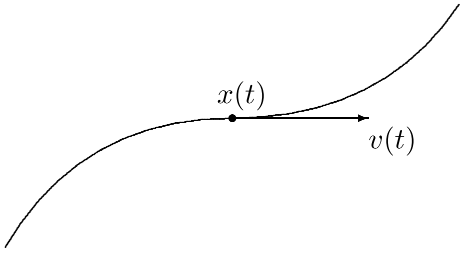
Chapter 1

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 of the particle

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

and the derivative of the velocity is the 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

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

which exerts a force $F(x)$ on the particle at x. Then **Newton's second law of motion** states that, at each point x in space, the force acting on the particle must be equal to the acceleration times the mass m (a positive constant) 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 of second order since the highest derivative is of second degree. More precisely, we 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 adding v to the dependent variables and considering $(x, v) \in \mathbb{R}^6$. 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 $x(t)$ that satisfy (1.5) (respectively (1.6)). To be 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 8.5.)

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 also has 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^∞ (with respect to both t and ε). Show that

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

(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 ε. Then solve the corresponding differential equations for $r_0(t)$, $r_1(t)$, ... and note that the initial conditions follow from $r(0,\varepsilon) = h$ respectively $\dot{r}(0,\varepsilon) = 0$. A rigorous justification for this procedure will be given in Section 2.5.)

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 $C(U,V) = C^0(U,V)$, $C^\infty(U,V) = \bigcap_{k\in\mathbb{N}} C^k(U,V)$, and $C^k(U) = C^k(U,\mathbb{R})$.

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

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

for the unknown function $x \in C^k(J)$, $J \subseteq \mathbb{R}$, and its derivatives

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

Here $F \in C(U)$ with U an open subset of $\mathbb{R}^{k+2}$. 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 general 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, $\frac{\partial F}{\partial y_k}(t, y) \neq 0$. This is the type of differential equations we will consider from now on.

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 $g_i(t) \equiv 0$.

Moreover, any system can always be reduced to a first-order system by changing to the new set of dependent 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 dependent 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, we will not pursue this case here.

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. Is the equation linear, autonomous? What is its order?*

(i) $y'(x) + y(x) = 0$.

(ii) $\frac{d^2}{dt^2}u(t) = t\,\sin(u(t))$.

(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 for $y(x)$ are linear?*

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

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

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

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?

Problem 1.8. *Let $x^{(k)} = f(x, x^{(1)}, \dots, x^{(k-1)})$ be an autonomous equation (or system). Show that if $\phi(t)$ is a solution, then so is $\phi(t - t_0)$.*

1.3. First order autonomous equations

Let us look at the simplest (nontrivial) case of a first-order autonomous equation and let us try to find the solution starting at a certain point x_0 at time $t = 0$:

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

We could of course also ask for the solution starting at x_0 at time t_0. However, once we have a solution $\phi(t)$ with $\phi(0) = x_0$, the solution $\psi(t)$ with $\psi(t_0) = x_0$ is given by a simple shift $\psi(t) = \phi(t - t_0)$ (this holds in fact for any autonomous equation – compare Problem 1.8).

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 $F(x) = \int_{x_0}^x \frac{dy}{f(y)}$ 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. Here $F^{-1}(t)$ is the inverse map of $F(t)$.

Now let us look at the maximal interval where ϕ is defined by this procedure. If $f(x_0) > 0$ (the case $f(x_0) < 0$ follows analogously), then f remains positive in some interval (x_1, x_2) around x_0 by continuity. Define

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

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, ϕ is defined for all $t > 0$ if and only if

$$T_+ = \int_{x_0}^{x_2} \frac{dy}{f(y)} = +\infty, \tag{1.25}$$

that is, if $1/f(x)$ is *not* integrable near x_2. Similarly, ϕ is defined for all $t < 0$ if and only if $1/f(x)$ is *not* integrable near x_1.

If $T_+ < \infty$ there are two possible cases: Either $x_2 = \infty$ or $x_2 < \infty$. In the first case the solution ϕ diverges to $+\infty$ and there is no way to extend it beyond T_+ in a continuous way. In the second case the solution ϕ reaches the point x_2 at the finite time T_+ and we could extend it as follows: If $f(x_2) > 0$ then x_2 was not chosen maximal and we can increase it which provides the required extension. Otherwise, if $f(x_2) = 0$, we can extend ϕ by setting $\phi(t) = x_2$ for $t \geq T_+$. However, in the latter case this might not be the only possible extension as we will see in the examples below. Clearly, similar arguments apply for $t < 0$.

Now let us look at some examples.

Example. If $f(x) = x$, $x_0 > 0$, we have $(x_1, x_2) = (0, \infty)$ and

$$F(x) = \log(\frac{x}{x_0}). \tag{1.26}$$

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

$$\phi(t) = x_0 \mathrm{e}^t. \tag{1.27}$$

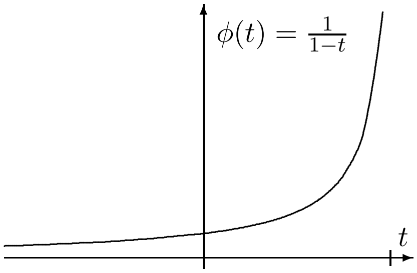
Thus the solution is globally defined for all $t \in \mathbb{R}$. Note that this is in fact a solution for all $x_0 \in \mathbb{R}$. ◇

Example. Let $f(x) = x^2$, $x_0 > 0$. We have $(x_1, x_2) = (0, \infty)$ and

$$F(x) = \frac{1}{x_0} - \frac{1}{x}. \tag{1.28}$$

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

$$\phi(t) = \frac{x_0}{1 - x_0 t}. \tag{1.29}$$



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 $f(x_0) = 0$, there is a trivial solution

$$\phi(t) = x_0 \tag{1.30}$$

to the initial condition $x(0) = x_0$. But is this the only one? If we have

$$\left| \int_{x_0}^{x_0+\varepsilon} \frac{dy}{f(y)} \right| < \infty, \tag{1.31}$$

then there is another solution

$$\varphi(t) = F^{-1}(t), \qquad F(x) = \int_{x_0}^{x} \frac{dy}{f(y)}, \tag{1.32}$$

with $\varphi(0) = x_0$ which is different from $\phi(t)$!

Example. Consider $f(x) = \sqrt{|x|}$, $x_0 > 0$. Then $(x_1, x_2) = (0, \infty)$,

$$F(x) = 2(\sqrt{x} - \sqrt{x_0}). \tag{1.33}$$

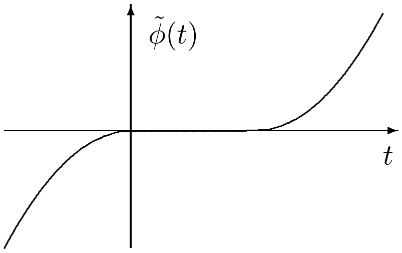
and

$$\varphi(t) = (\sqrt{x_0} + \frac{t}{2})^2, \quad -2\sqrt{x_0} < t < \infty. \tag{1.34}$$

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

$$\tilde{\phi}(t) = \begin{cases} -\frac{(t-t_0)^2}{4}, & t \le t_0, \\ 0, & t_0 \le t \le t_1, \\ \frac{(t-t_1)^2}{4}, & t_1 \le t. \end{cases} \tag{1.35}$$

The solution $\tilde{\phi}$ for $t_0 = 0$ and $t_1 = 1$ is depicted below:



◇

As a conclusion of the previous examples we have:

- Solutions might only exist locally in t, even for perfectly nice f.
- Solutions might not be unique. Note however, that $f(x) = \sqrt{|x|}$ is not differentiable at the point $x_0 = 0$ which causes the problems.

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

$$\dot{x} = f(x)g(t) \tag{1.36}$$

(see Problem 1.11).

Problem 1.9. *Solve the following differential equations:*

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

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

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

Problem 1.10. *Show that the solution of* (1.20) *is unique if* $f \in C^1(\mathbb{R})$.

Problem 1.11 (Separable equations). *Show that the equation* ($f, g \in C^1$)

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

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

Problem 1.12. *Solve the following differential equations:*

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

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

(iii) $\dot{x} = \sin(t)\mathrm{e}^x$.

Sketch the solutions. For which initial conditions (if any) are the solutions bounded?

Problem 1.13. *Investigate uniqueness of the differential equation*

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

Show that the initial value problem $x(0) = x_0$ has a unique global solution for every $x_0 \in \mathbb{R}$. However, show that the global solutions still intersect! (Hint: Note that if $x(t)$ is a solution so is $-x(t)$ and $x(-t)$, so it suffices to consider $x_0 \geq 0$ and $t \geq 0$.)

Problem 1.14. *Charging a capacitor is described by the differential equation*

$$R\dot{Q}(t) + \frac{1}{C}Q(t) = V_0,$$

where $Q(t)$ is the charge at the capacitor, C is its capacitance, V_0 is the voltage of the battery, and R is the resistance of the wire.

Compute $Q(t)$ assuming the capacitor is uncharged at $t = 0$. What charge do you get as $t \to \infty$?

Problem 1.15 (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, $\kappa > 0$ is the growth rate, 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 1.16 (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*

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

Rescale by

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

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

$$\ddot{x} = \eta \dot{x}^2 - g, \qquad \eta > 0.$$

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 $t > t_0$ and match the solutions at 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. In many cases the solvable equation will be the

Linear equation:

The solution of the linear homogeneous equation

$$\dot{x} = a(t)x \tag{1.37}$$

is given by

$$\phi(t) = x_0 A(t, t_0), \qquad A(t, s) = \mathrm{e}^{\int_s^t a(s)ds}, \tag{1.38}$$

and the solution of the corresponding inhomogeneous equation

$$\dot{x} = a(t)x + g(t), \tag{1.39}$$

is given by

$$\phi(t) = x_0 A(t, t_0) + \int_{t_0}^{t} A(t, s) g(s) ds. \tag{1.40}$$

This can be verified by a straightforward computation.

Next we turn to the problem of transforming differential equations. Given the point with coordinates (t, x), we may change to new coordinates (s, y) given by

$$s = \sigma(t, x), \qquad y = \eta(t, x). \tag{1.41}$$

Since we do not want to lose information, we require this transformation to be a diffeomorphism (i.e., invertible with differentiable inverse).

A given function $\phi(t)$ will be transformed into a function $\psi(s)$ which has to be obtained by eliminating t from

$$s = \sigma(t, \phi(t)), \qquad \psi = \eta(t, \phi(t)). \tag{1.42}$$

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**

$$s = \sigma(t), \qquad y = \eta(t, x) \tag{1.43}$$

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

$$t = \tau(s), \qquad x = \xi(s, y), \tag{1.44}$$

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

$$\dot{x} = f(t, x) \tag{1.45}$$

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

$$\dot{y} = \dot{\tau}\left(\frac{\partial \eta}{\partial t}(\tau, \xi) + \frac{\partial \eta}{\partial x}(\tau, \xi)\, f(\tau, \xi)\right), \tag{1.46}$$

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

$$\frac{dy}{ds} = \frac{dy(t(s), x(t(s)))}{ds} = \frac{\partial y}{\partial t}\frac{dt}{ds} + \frac{\partial y}{\partial x}\frac{dx}{dt}\frac{dt}{ds}. \tag{1.47}$$

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

Homogeneous equation:

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

$$\dot{x} = f(\frac{x}{t}). \tag{1.48}$$

This special form suggests the change of variables $y = \frac{x}{t}$ $(t \neq 0)$, which (by (1.47)) transforms our equation into

$$\dot{y} = \frac{\partial y}{\partial t} + \frac{\partial y}{\partial x}\dot{x} = -\frac{x}{t^2} + \frac{1}{t}\dot{x} = \frac{f(y) - y}{t}. \tag{1.49}$$

This equation is separable.

More generally, consider the differential equation

$$\dot{x} = f(\frac{ax + bt + c}{\alpha x + \beta t + \gamma}). \tag{1.50}$$

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

$$\dot{x} = \tilde{f}(ax + bt), \tag{1.51}$$

which transforms into

$$\dot{y} = a\tilde{f}(y) + b \tag{1.52}$$

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 (1.50) to the homogeneous equation

$$\dot{y} = \hat{f}(\frac{ay + bs}{\alpha y + \beta s}) \tag{1.53}$$

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

Bernoulli equation:

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

$$\dot{x} = f(t)x + g(t)x^n, \qquad n \neq 0, 1. \tag{1.54}$$

The transformation

$$y = x^{1-n} \tag{1.55}$$

gives the linear equation

$$\dot{y} = (1 - n)f(t)y + (1 - n)g(t). \tag{1.56}$$

(Note: If $n = 0$ or $n = 1$ the equation is already linear and there is nothing to do.)

Riccati equation:

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

$$\dot{x} = f(t)x + g(t)x^2 + h(t). \tag{1.57}$$

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

$$y = \frac{1}{x - x_p(t)} \tag{1.58}$$

yields the linear equation

$$\dot{y} = -(f(t) + 2x_p(t)g(t))y - g(t). \tag{1.59}$$

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 [**24**] or Zwillinger [**48**], where you can look up a given equation and find out if it is known to be solvable explicitly. 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

$$\dot{x} = \sin(t)x \tag{1.60}$$

you would use the command

```
In[1]:= DSolve[x'[t] == x[t]Sin[t], x[t], t]
Out[1]=  {{x[t] → e^(-Cos[t]) C[1]}}
```

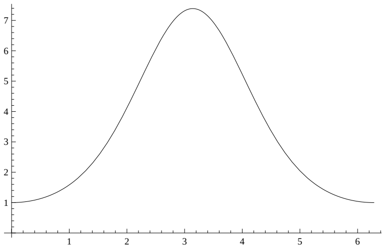
Here the constant C[1] introduced by *Mathematica* can be chosen arbitrarily (e.g. to satisfy an initial condition). We can also solve the corresponding initial value problem using

```
In[2]:= DSolve[{x'[t] == Sin[t]x[t], x[0] == 1}, x[t], t]
Out[2]=  {{x[t] → e^(1-Cos[t])}}
```

and plot it using

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

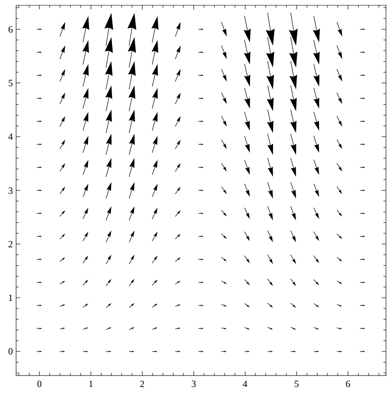
Out[3]=



In some situations it is also useful to visualize the corresponding **directional field**. That is, to every point (t, x) we attach the vector $(1, f(t, x))$. Then the solution curves will be tangent to this vector field in every point:

```
In[4]:= VectorPlot[{1, Sin[t] x}, {t, 0, 2π}, {x, 0, 6}]
```

Out[4]=



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 be 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

$$\textit{In[5]:=}\ \mathtt{DSolve}[\{\mathtt{x}'[\mathtt{t}] == \sqrt{\mathtt{x}[\mathtt{t}]}, \mathtt{x}[0] == 0\}, \mathtt{x}[\mathtt{t}], \mathtt{t}]$$

$$\textit{Out[5]=}\ \{\{\mathtt{x}[\mathtt{t}] \to \frac{\mathtt{t}^2}{4}\}\}$$

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.

Moreover, if you try to solve the general initial value problem it gets even worse:

$$\textit{In[6]:=}\ \mathtt{DSolve}[\{\mathtt{x}'[\mathtt{t}] == \sqrt{\mathtt{x}[\mathtt{t}]}, \mathtt{x}[0] == \mathtt{x0}\}, \mathtt{x}[\mathtt{t}], \mathtt{t}]\ //\ \mathtt{Simplify}$$

$$\textit{Out[6]=}\ \{\{\mathtt{x}[\mathtt{t}] \to \frac{1}{4}(\mathtt{t} - 2\sqrt{\mathtt{x}_0})^2\}, \{\mathtt{x}[\mathtt{t}] \to \frac{1}{4}(\mathtt{t} + 2\sqrt{\mathtt{x}_0})^2\}\}$$

The first "solution" is no solution of our initial value problem at all! It satisfies $\dot{x} = -\sqrt{x}$.

Problem 1.18. *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 1.19 (Euler equation)**.** *Transform the differential equation*

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

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

Problem 1.20. *Pick some differential equations from the previous problems and solve them using your favorite computer algebra system. Plot the solutions.*

Problem 1.21 (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 1.22 (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 μ works.

Consider

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

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

Problem 1.23. *Show that*

$$\dot{x} = t^{n-1} f(\frac{x}{t^n})$$

can be solved using the new variable $y = \frac{x}{t^n}$*.*

Problem 1.24 (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 normal of the mirror and the incoming respectively reflected ray must be equal. Considering the scalar products of the vectors with the normal vector this yields

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

which is the differential equation for $y = y(x)$ *you have to solve. I recommend the substitution* $u = \frac{y}{x}$*.)*

Problem 1.25 (Catenary)**.** *Solve the differential equation describing the shape* $y(x)$ *of a hanging chain suspended at two points:*

$$y'' = a\sqrt{1 + (y')^2}, \qquad a > 0.$$

Problem 1.26 (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. Assume that the initial value problem $y(x_0) = y_0$*,* $y'(x_0) = y_1$ *has a unique solution.*

- *Show that a nontrivial solution of the boundary value problem must satisfy* $y'(0) = p_0 > 0$*.*
- *If a solution satisfies* $y'(x_0) = 0$*, then the solution is symmetric with respect to this point:* $y(x) = y(x_0 - x)$*. (Hint: Uniqueness.)*

- *Solve the initial value problem $y(0) = 0$, $y'(0) = p_0 > 0$ as follows: 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 $y'' = f(y)$.)*
- *Does the solution found in the previous item attain $y'(x_0) = 0$ at some x_0? What value should x_0 have for $y(x)$ to solve our boundary value problem?*
- *Can you find a value for p_0 in terms of special functions?*

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.

Moreover, even in situations where an exact solution can be obtained, a qualitative analysis can give a better overview of the behavior than the formula for the solution. To get more specific, let us look at the first-order autonomous initial value problem

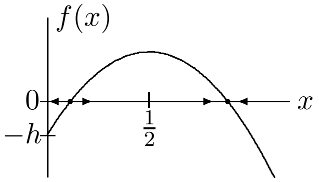
$$\dot{x} = f(x), \qquad x(0) = x_0, \tag{1.61}$$

where $f \in C(\mathbb{R})$ is such that solutions are unique (e.g. $f \in C^1(\mathbb{R})$). We already saw how to solve this equation in Section 1.3. However, for a given f we might well shipwreck when computing the integral $F(x) = \int_{x_0}^{x} \frac{dy}{f(y)}$ or when trying to solve $F(x(t)) = t$ for $x(t)$. On the other hand, to get a qualitative understanding of the solution an explicit formula turns out to be unessential.

Example. For example, consider the logistic growth model (Problem 1.16)

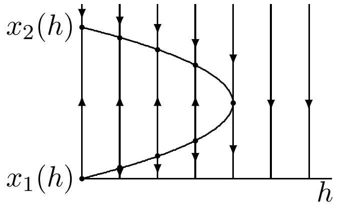
$$\dot{x}(t) = (1 - x(t))x(t) - h, \tag{1.62}$$

which can be solved by separation of variables. To get an overview we plot the corresponding right-hand side $f(x) = (1 - x)x - h$:



Since the sign of $f(x)$ tells us in what direction the solution will move, all we have to do is to discuss the sign of $f(x)$!

For $0 < h < \frac{1}{4}$ there are two zeros $x_{1,2} = \frac{1}{2}(1 \pm \sqrt{1-4h})$. If we start at one of these zeros, the solution will stay there for all t. If we start below x_1 the solution will decrease and converge to $-\infty$. If we start above x_1 the solution will increase and converge to x_2. If we start above x_2 the solution will decrease and again converge to x_2.



At $h = \frac{1}{4}$ a bifurcation occurs: The two zeros coincide $x_1 = x_2$ but otherwise the analysis from above still applies. For $h > \frac{1}{4}$ there are no zeros and all solutions decrease and converge to $-\infty$. ◇

So we get a complete picture just by discussing the sign of $f(x)$! More generally, we have the following result (Problem 1.28).

Lemma 1.1. *Consider the first-order autonomous initial value problem* (1.61)*, where $f \in C(\mathbb{R})$ is such that solutions are unique.*

(i) *If $f(x_0) = 0$, then $x(t) = x_0$ for all t.*

(ii) *If $f(x_0) \neq 0$, then $x(t)$ converges to the first zero left ($f(x_0) < 0$) respectively right ($f(x_0) > 0$) of x_0. If there is no such zero the solution converges to $-\infty$, respectively ∞.*

If our differential equation is not autonomous, the situation becomes a bit more involved. As a prototypical example let us investigate the differential equation

$$\dot{x} = x^2 - t^2. \tag{1.63}$$

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 4.13, 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

$$\dot{x} = f(t,x), \qquad x(t_0) = x_0 \tag{1.64}$$

defined in a neighborhood of t_0 (Theorem 2.2). 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 2.16).

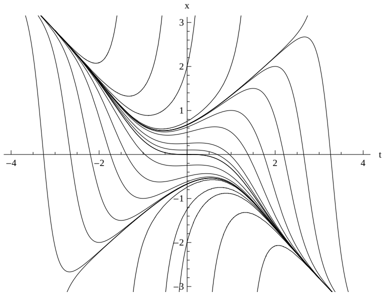
In order to get some feeling of what we should expect, a good starting point is a numerical investigation. Using the command

```
In[7]:= 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[7]=  {{x[t] → InterpolatingFunction[{{-2., 1.03747}}, <>][t]}}
```

we can compute a numerical solution on the interval $(-2, 2)$. Numerically solving an ordinary differential equation 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 2.7). 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 though we have requested the solution on $(-2, 2)$. This indicates that the solution only exists for finite time.

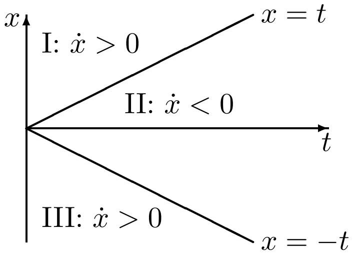
Combining the solutions for different initial conditions into one plot we get the following picture:



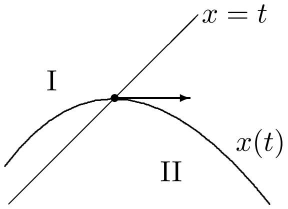
First of all we note the symmetry with respect to the transformation $(t, x) \to (-t, -x)$. Hence it suffices to consider $t \geq 0$. Moreover, observe that different solutions never cross, which is a consequence of uniqueness.

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.

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 $-\infty$ (why can't it remain bounded?). 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 these questions we need to generalize the idea from above that a solution can only cross the line $x = t$ from above and the line $x = -t$ from below.

A differentiable function $x_+(t)$ satisfying

$$\dot{x}_+(t) > f(t, x_+(t)), \qquad t \in [t_0, T), \tag{1.65}$$

is called a **super solution** (or **upper solution**) of our equation. Similarly, a differentiable function $x_-(t)$ satisfying

$$\dot{x}_-(t) < f(t, x_-(t)), \qquad t \in [t_0, T), \tag{1.66}$$

is called a **sub solution** (or **lower solution**).

Example. For example, $x_+(t) = t$ is a super solution and $x_-(t) = -t$ is a sub solution of our equation for $t \geq 0$. ⋄

Lemma 1.2. *Let $x_+(t)$, $x_-(t)$ be super, sub solutions of the differential equation $\dot{x} = f(t,x)$ on $[t_0, T)$, respectively. Then for every solution $x(t)$ on $[t_0, T)$ we have*

$$x(t) < x_+(t), \quad t \in (t_0, T), \quad \textit{whenever} \quad x(t_0) \leq x_+(t_0), \tag{1.67}$$

respectively

$$x_-(t) < x(t), \quad t \in (t_0, T), \quad \textit{whenever} \quad x(t_0) \geq x_-(t_0). \tag{1.68}$$

Proof. In fact, consider $\Delta(t) = x_+(t) - x(t)$. Then we have $\Delta(t_0) \geq 0$ and $\dot{\Delta}(t) > 0$ whenever $\Delta(t) = 0$. Hence $\Delta(t)$ can cross 0 only from below. Since we start with $\Delta(t_0) \geq 0$, we have $\Delta(t) > 0$ for $t > t_0$ sufficiently close to t_0. In fact, if $\Delta(t_0) > 0$ this follows from continuity and otherwise, if $\Delta(t_0) = 0$, this follows from $\dot{\Delta}(t_0) > 0$. Now let $t_1 > t_0$ be the first value with $\Delta(t_1) = 0$. Then $\Delta(t) > 0$ for $t \in (t_0, t_1)$, which contradicts $\dot{\Delta}(t_1) > 0$. □

Similar results hold for $t < t_0$. The details are left to the reader (Problem 1.29).

Now we are able to answer our remaining questions. Since we were already successful by considering the curves given by $f(t,x) = 0$, let us look at the **isoclines** $f(t,x) = const$.

Considering $x^2 - t^2 = -2$ the corresponding curve is

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

which is easily seen to be a super solution

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

for $t > 2\sqrt{2/3}$. 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.

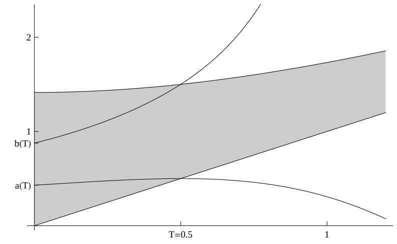
But will every solution in region II eventually end up between $y_+(t)$ and $x_-(t)$? The answer is yes: Since $x(t)$ is decreasing in region II, every solution will eventually be below $-y_+(t)$. Furthermore, every solution $x(t)$ starting at a point (t_0, x_0) below $-y_+(t)$ and above $y_+(t)$ satisfies $\dot{x}(t) < -2$ as long as it remains between $-y_+(t)$ and $y_+(t)$. Hence, by integrating this inequality, $x(t) - x_0 < -2(t - t_0)$, we see that $x(t)$ stays below the line $x_0 - 2(t - t_0)$ as long as it remains between $-y_+(t)$ and $y_+(t)$. 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

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

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. See the picture below which shows two solutions entering the shaded region between $x_+(t)$ and $y_-(t)$ at $T = 0.5$:



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 $[a(T), b(T)]$! Moreover, this also implies that $a(T)$ is strictly increasing and $b(T)$ is strictly decreasing. 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 $x \mapsto f(t, x) = x^2 - t^2$ is increasing in region I, we see that the distance between two solutions

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

must increase as well. If there were two such solutions, their distance would consequently increase. But this is impossible, since the distance of $x_+(t)$

and $y_-(t)$ tends to zero. 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 $x_0(t)$ will eventually be above $y_-(t)$ and converge to $+\infty$. It remains to show that this happens in finite time.

This is not surprising, since the $x(t)^2$ term should dominate over the $-t^2$ term and we already know that the solutions of $\dot{x} = x^2$ diverge. So let us try to make this precise: First of all

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

for every solution above $y_-(t)$ implies $x(t) > x_0 + 2(t - t_0)$. Thus there is an $\varepsilon > 0$ such that

$$x(t) > \frac{t}{\sqrt{1-\varepsilon}}.$$

This implies

$$\dot{x}(t) = x(t)^2 - t^2 > x(t)^2 - (1-\varepsilon)x(t)^2 = \varepsilon x(t)^2$$

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

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

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 that 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 to 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!

We end this section with a generalization of Lemma 1.2 which is often useful. Indeed, you might wonder what happens if we allow equality in the definition of a super solution (1.65). At first sight you might expect that this should not do much harm and the conclusion of Lemma 1.2 should still hold if we allow for equality there as well. However, if you apply this conclusion to two solutions of the same equation it will automatically give you uniqueness of solutions. Hence this generalization cannot be true without further assumptions on f. One assumption which will do the trick (and which will hence also guarantee uniqueness of solutions) is the following condition: We will say that f is locally **Lipschitz continuous** in the second argument, uniformly with respect to the first argument, if

$$L = \sup_{(t,x)\neq(t,y)\in V} \frac{|f(t,x)-f(t,y)|}{|x-y|} \tag{1.69}$$

is finite for every compact set V contained in the domain of f. We will meet this condition again in Section 2.2 where we will also further discuss it. For now notice that it will hold if f has a continuous partial derivative with respect to x by the mean value theorem.

Theorem 1.3. *Suppose f is locally Lipschitz continuous with respect to x uniformly in t. Let $x(t)$ and $y(t)$ be two differentiable functions such that*

$$x(t_0) \le y(t_0), \qquad \dot{x}(t) - f(t,x(t)) \le \dot{y}(t) - f(t,y(t)),\ t \in [t_0,T) \tag{1.70}$$

Then we have $x(t) \le y(t)$ for every $t \in [t_0,T)$. Moreover, if $x(t) < y(t)$ for some t this remains true for all later times.

Proof. We argue by contradiction. Suppose the first claim were not true. Then we could find some time t_1 such that $x(t_1) = y(t_1)$ and $x(t) > y(t)$ for $t \in (t_1, t_1+\varepsilon)$. Introduce $\Delta(t) = x(t) - y(t)$ and observe

$$\dot{\Delta}(t) = \dot{x}(t) - \dot{y}(t) \le f(t,x(t)) - f(t,y(t)) \le L\Delta(t), \qquad t \in [t_1, t_1+\varepsilon),$$

where the first inequality follows from assumption and the second from (1.69). But this implies that the function $\tilde{\Delta}(t) = \Delta(t)\mathrm{e}^{-Lt}$ satisfies $\dot{\tilde{\Delta}}(t) \le 0$ and thus $\tilde{\Delta}(t) \le \tilde{\Delta}(t_1) = 0$, that is, $x(t) \le y(t)$ for $t \in [t_0,T)$ contradicting our assumption.

So the first part is true. To show the second part set $\Delta(t) = y(t) - x(t)$ which is now nonnegative by the first part. Then, as in the previous case one shows $\dot{\tilde{\Delta}}(t) \ge 0$ where $\tilde{\Delta}(t) = \Delta(t)\mathrm{e}^{Lt}$ and the claim follows. □

A few consequences are worth while noting:

First of all, if $x(t)$ and $y(t)$ are two solutions with $x(t_0) \le y(t_0)$, then $x(t) \le y(t)$ for all $t \ge t_0$ (for which both solutions are defined). In particular, in the case $x(t_0) = y(t_0)$ this shows uniqueness of solutions: $x(t) = y(t)$.

Second, we can extend the notion of a super solution by requiring only $x_+(t) \geq f(t, x_+(t))$. Then $x_+(t_0) \geq x(t_0)$ implies $x_+(t) \geq x(t)$ for all $t \geq t_0$ and if strict inequality becomes true at some time it remains true for all later times.

Problem 1.27. *Let x be a solution of* (1.61) *which satisfies* $\lim_{t\to\infty} x(t) = x_1$. *Show that* $\lim_{t\to\infty} \dot{x}(t) = 0$ *and* $f(x_1) = 0$. *(Hint: If you prove* $\lim_{t\to\infty} \dot{x}(t) = 0$ *without using* (1.61) *your proof is wrong! Can you give a counter example?)*

Problem 1.28. *Prove Lemma 1.1. (Hint: This can be done either by using the analysis from Section 1.3 or by using the previous problem.)*

Problem 1.29. *Generalize the concept of sub and super solutions to the interval* (T, t_0), *where* $T < t_0$.

Problem 1.30. *Discuss the equation* $\dot{x} = x^2 - \frac{t^2}{1+t^2}$.

- *Make a numerical analysis.*
- *Show that there is a unique solution which asymptotically approaches the line* $x = 1$.
- *Show that all solutions below this solution approach the line* $x = -1$.
- *Show that all solutions above go to* ∞ *in finite time.*

Problem 1.31. *Discuss the equation* $\dot{x} = x^2 - t$.

Problem 1.32. *Generalize Theorem 1.3 to the interval* (T, t_0), *where* $T < t_0$.

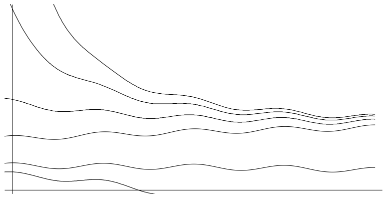
1.6. Qualitative analysis of first-order periodic equations

Some of the most interesting examples are periodic ones, where $f(t+1, x) = f(t, x)$ (without loss we have assumed the period to be one). So let us consider the logistic growth model with a time dependent harvesting term

$$\dot{x}(t) = (1 - x(t))x(t) - h \cdot (1 - \sin(2\pi t)), \tag{1.71}$$

where $h \geq 0$ is some positive constant. In fact, we could replace $1 - \sin(2\pi t)$ by any nonnegative periodic function $g(t)$ and the analysis below will still hold.

The solutions corresponding to some initial conditions for $h = 0.2$ are depicted below.



It looks like all solutions starting above some value x_1 converge to a periodic solution starting at some other value $x_2 > x_1$, while solutions starting below x_1 diverge to $-\infty$.

The key idea is to look at the fate of an arbitrary initial value x after one period. More precisely, let us denote the solution which starts at the point x at time $t = 0$ by $\phi(t, x)$. Then we can introduce the **Poincaré map** via

$$P(x) = \phi(1, x). \tag{1.72}$$

By construction, an initial condition x_0 will correspond to a periodic solution if and only if x_0 is a fixed point of the Poincaré map, $P(x_0) = x_0$. In fact, this follows from uniqueness of solutions of the initial value problem, since $\phi(t + 1, x)$ again satisfies $\dot{x} = f(t, x)$ if $f(t + 1, x) = f(t, x)$. So $\phi(t + 1, x_0) = \phi(t, x_0)$ if and only if equality holds at the initial time $t = 0$, that is, $\phi(1, x_0) = \phi(0, x_0) = x_0$.

We begin by trying to compute the derivative of $P(x)$ as follows. Set

$$\theta(t, x) = \frac{\partial}{\partial x}\phi(t, x) \tag{1.73}$$

and differentiate

$$\dot{\phi}(t, x) = \big(1 - \phi(t, x)\big)\phi(t, x) - h \cdot \big(1 - \sin(2\pi t)\big), \tag{1.74}$$

with respect to x (we will justify this step in Theorem 2.10). Then we obtain

$$\dot{\theta}(t, x) = \big(1 - 2\phi(t, x)\big)\theta(t, x) \tag{1.75}$$

and assuming $\phi(t, x)$ is known we can use (1.38) to write down the solution

$$\theta(t, x) = \exp\left(\int_0^t \big(1 - 2\phi(s, x)\big) ds\right). \tag{1.76}$$

Setting $t = 1$ we obtain

$$P'(x) = \theta(1, x) = \exp\left(1 - 2\int_0^1 \phi(s, x) ds\right). \tag{1.77}$$

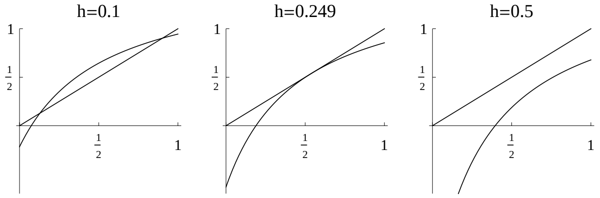
While it might look as if this formula is of little help since we do not know $\phi(t, x)$, it at least tells us that that $P'(x) > 0$, that is, $P(x)$ is strictly increasing. Note that this latter fact also follows since different solutions cannot cross in the (t, x) plane by uniqueness (show this!).

Moreover, differentiating this last expression once more we obtain

$$P''(x) = -2\left(\int_0^1 \theta(s,x)ds\right) P'(x) < 0 \tag{1.78}$$

since $\theta(t,x) > 0$ by (1.76). Thus $P(x)$ is concave and there are at most two intersections with the line x (why?). In other words, there are at most two periodic solutions. Note that so far we did not need any information on the harvesting term.

To see that all cases can occur, we will now consider the dependence with respect to the parameter h. A numerically computed picture of the Poincaré map for different values of h is shown below.



It seems to indicate that $P(x)$ is decreasing as a function of h. To prove this we proceed as before. Set

$$\psi(t,x) = \frac{\partial}{\partial h}\phi(t,x) \tag{1.79}$$

and differentiate the differential equation with respect to h (again this step will be justified by Theorem 2.10) to obtain

$$\dot{\psi}(t,x) = \big(1 - 2\phi(t,x)\big)\psi(t,x) + \big(1 - \sin(2\pi t)\big). \tag{1.80}$$

Hence, since $\psi(0,x) = \frac{\partial}{\partial h}\phi(0,x) = \frac{\partial}{\partial h}x = 0$, equation (1.40) implies

$$\psi(t,x) = -\int_0^t \exp\left(\int_s^t \big(1 - 2\phi(r,x)\big)dr\right)\big(1 - \sin(2\pi s)\big)ds < 0 \tag{1.81}$$

and setting $t = 1$ we infer

$$\frac{\partial}{\partial h}P_h(x) < 0, \tag{1.82}$$

where we have added h as a subscript to emphasize the dependence on the parameter h. Moreover, for $h = 0$ we have

$$P_0(x) = \frac{\mathrm{e}\,x}{1 + (\mathrm{e} - 1)x} \tag{1.83}$$

and there are two fixed points $x_1 = 0$ and $x_2 = 1$. As h increases these points will approach each other and collide at some critical value h_c. Above this value there are no periodic solutions and all orbits converge to $-\infty$ since $P(x) < x$ for all $x \in \mathbb{R}$ (show this).

To complete our analysis suppose $h < h_c$ and denote by $x_1 < x_2$ the two fixed points of $P(x)$. Define the iterates of $P(x)$ by $P^0(x) = x$ and $P^n(x) = P(P^{n-1}(x))$. We claim

$$\lim_{n\to\infty} P^n(x) = \begin{cases} x_2, & x > x_1, \\ x_1, & x = x_1, \\ -\infty, & x < x_1. \end{cases} \tag{1.84}$$

For example, let $x \in (x_1, x_2)$. Then, since $P(x)$ is strictly increasing we have $x_1 = P(x_1) < P(x) < P(x_2) = x_2$. Moreover, since $P(x)$ is concave we have $x < P(x)$, which shows that $P^n(x)$ is a strictly increasing sequence. Let $x_0 \in (x, x_2]$ be its limit. Then $P(x_0) = P(\lim_{n\to\infty} P^n(x)) = \lim_{n\to\infty} P^{n+1}(x) = x_0$ shows that x_0 is a fixed point, that is, $x_0 = x_2$. The other cases can be shown similar (Problem 1.33).

So for $x < x_1$ the solution diverges to $-\infty$ and for $x > x_1$ we have

$$\lim_{n\to\infty} |\phi(n,x) - x_2| = 0, \tag{1.85}$$

which implies (show this)

$$\lim_{t\to\infty} |\phi(t,x) - \phi(t,x_2)| = 0. \tag{1.86}$$

Similar considerations can be made for the case $h = h_c$ and $h > h_c$.

Problem 1.33. *Suppose $P(x)$ is a continuous, monotone, and concave function with two fixed points $x_1 < x_2$. Show the remaining cases in* (1.84).

Problem 1.34. *Find* $\lim_{n\to\infty} P^n(x)$ *in the case $h = h_c$ and $h > h_c$.*

Problem 1.35. *Suppose $f \in C^2(\mathbb{R})$ and $g \in C(\mathbb{R})$ is a nonnegative periodic function $g(t+1) = g(t)$. Show that the above discussion still holds for the equation*

$$\dot{x} = f(x) + h \cdot g(t)$$

if $f''(x) < 0$ and $g(t) \ge 0$.

Problem 1.36. *Suppose $a \in \mathbb{R}$ and $g \in C(\mathbb{R})$ is a nonnegative periodic function $g(t+1) = g(t)$. Find conditions on a, g such that the linear inhomogeneous equation*

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

has a periodic solution. When is this solution unique? (Hint: (1.40)*.)*

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

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

From the triangle inequality we also get the **inverse triangle inequality** (Problem 2.1)

$$|\|f\| - \|g\|| \le \|f - g\|. \tag{2.1}$$

The pair $(X, \|.\|)$ is called a **normed vector space**. Given a normed vector space X, we say that a sequence of vectors f_n **converges** to a vector f if $\lim_{n\to\infty} \|f_n - f\| = 0$. We will write $f_n \to f$ or $\lim_{n\to\infty} f_n = f$, as usual, in this case. Moreover, a mapping $F : X \to Y$ between two normed spaces is called **continuous** if $f_n \to f$ implies $F(f_n) \to F(f)$. In fact, it is not hard to see that the norm, vector addition, and multiplication by scalars are continuous (Problem 2.2).

In addition to the concept of convergence we also have the concept of a **Cauchy sequence** and hence the concept of completeness: A normed space is called **complete** if every Cauchy sequence has a limit. A complete normed space is called a **Banach space**.

Example. Clearly $\mathbb{R}^n$ (or $\mathbb{C}^n$) is a Banach space with the usual Euclidean norm

$$|x| = \sqrt{\sum_{j=1}^{n} |x_j|^2}. \tag{2.2}$$

◇

We will be mainly interested in the following example: Let I be a compact interval and consider the continuous functions $C(I)$ on this interval. They form a vector space if all operations are defined pointwise. Moreover, $C(I)$ becomes a normed space if we define

$$\|x\| = \sup_{t\in I} |x(t)|. \tag{2.3}$$

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 $x(t)$ if and only if

$$\lim_{n\to\infty} \|x_n - x\| = \lim_{n\to\infty} \sup_{t\in I} |x_n(t) - x(t)| = 0. \tag{2.4}$$

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

$$|x_n(t) - x_m(t)| \le \varepsilon \qquad \forall n, m > N_\varepsilon,\ t \in I \tag{2.5}$$

we see

$$|x_n(t) - x(t)| \le \varepsilon \qquad \forall n > N_\varepsilon,\ t \in I, \tag{2.6}$$

that is, $x_n(t)$ converges uniformly to $x(t)$. However, up to this point we do not 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: Fix $t \in I$ and $\varepsilon > 0$. To show that x is continuous we need to find a δ such that $|t - s| < \delta$ implies $|x(t) - x(s)| < \varepsilon$. Pick n so that $\|x_n - x\| < \varepsilon/3$ and δ so that $|t-s| < \delta$ implies $|x_n(t) - x_n(s)| < \varepsilon/3$. Then $|t - s| < \delta$ implies

$$|x(t)-x(s)| \le |x(t)-x_n(t)|+|x_n(t)-x_n(s)|+|x_n(s)-x(s)| < \frac{\varepsilon}{3}+\frac{\varepsilon}{3}+\frac{\varepsilon}{3} = \varepsilon$$

as required. 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 result.

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

$$\|K(x) - K(y)\| \le \theta \|x - y\|, \qquad x, y \in C. \tag{2.7}$$

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*

$$\|K^n(x) - \overline{x}\| \le \frac{\theta^n}{1-\theta} \|K(x) - x\|, \qquad x \in C. \tag{2.8}$$

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

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

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

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

$$\begin{aligned} \|x_n - x_m\| &\le \sum_{j=m+1}^{n} \|x_j - x_{j-1}\| \le \theta^m \sum_{j=0}^{n-m-1} \theta^j \|x_1 - x_0\| \\ &= \theta^m \frac{1-\theta^{n-m}}{1-\theta} \|x_1 - x_0\| \le \frac{\theta^m}{1-\theta} \|x_1 - x_0\|. \end{aligned} \tag{2.9}$$

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

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

shows that $\overline{x}$ is a fixed point and the estimate (2.8) follows after taking the limit $n \to \infty$ in (2.9). □

Question: Why is closedness of C important?

Problem 2.1. *Show that $|\|f\| - \|g\|| \le \|f - g\|$.*

Problem 2.2. *Let X be a Banach space. Show that the norm, vector addition, and multiplication by scalars are continuous. That is, if $f_n \to f$, $g_n \to g$, and $\alpha_n \to \alpha$, then $\|f_n\| \to \|f\|$, $f_n + g_n \to f + g$, and $\alpha_n f_n \to \alpha f$.*

Problem 2.3. *Show that the space $C(I,\mathbb{R}^n)$ together with the sup norm (2.3) is a Banach space if I is a compact interval. Show that the same is true for $I=[0,\infty)$ and $I=\mathbb{R}$ if one considers the vector space of bounded continuous functions $C_b(I,\mathbb{R}^n)$.*

Problem 2.4. *Derive Newton's method for finding the zeros of a twice continuously differentiable function $f(x)$,*

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

from the contraction principle by showing that if $\overline{x}$ is a zero with $f'(\overline{x})\neq 0$, then there is a corresponding closed interval C around $\overline{x}$ such that the assumptions of Theorem 2.1 are satisfied.

2.2. The basic existence and uniqueness result

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

$$\dot{x}=f(t,x), \quad x(t_0)=x_0. \tag{2.10}$$

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 (2.10) is equivalent to the following **integral equation**

$$x(t)=x_0+\int_{t_0}^{t} f(s,x(s))\,ds. \tag{2.11}$$

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

$$x_1(t)=x_0+\int_{t_0}^{t} f(s,x_0(s))\,ds. \tag{2.12}$$

Iterating this procedure we get a sequence of approximating solutions

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

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 (2.11).

We will set $t_0=0$ for notational simplicity and consider only the case $t\geq 0$ to avoid excessive numbers of absolute values in the following estimates.

First of all we will need a Banach space. The obvious choice is $X=C([0,T],\mathbb{R}^n)$for some suitable $T>0$. Furthermore, we need a closed subset

$C \subseteq X$ such that $K : C \to C$. We will try a closed ball of radius δ around the constant function x_0.

Since U is open and $(0, x_0) \in U$ we can choose $V = [0,T] \times \overline{B_\delta(x_0)} \subset U$, where $B_\delta(x_0) = \{x \in \mathbb{R}^n \,|\, |x - x_0| < \delta\}$, and abbreviate

$$M = \max_{(t,x)\in V} |f(t,x)|, \tag{2.14}$$

where the maximum exists by continuity of f and compactness of V. Then

$$|K(x)(t) - x_0| \le \int_0^t |f(s,x(s))|ds \le t\,M \tag{2.15}$$

whenever the graph of $x(t)$ lies within V, that is, $\{(t,x(t))|t \in [0,T]\} \subset V$. Hence, for $t \le T_0$, where

$$T_0 = \min\{T, \frac{\delta}{M}\}, \tag{2.16}$$

we have $T_0\,M \le \delta$ and the graph of $K(x)$ restricted to $[0,T_0]$ is again in V. In the special case $M = 0$ one has to understand this as $\frac{\delta}{M} = \infty$ such that $T_0 = \min\{T, \infty\} = T$. Moreover, note that since $[0,T_0] \subseteq [0,T]$ the same constant M will also bound $|f|$ on $V_0 = [0,T_0] \times \overline{B_\delta(x_0)} \subseteq V$.

So if we choose $X = C([0,T_0], \mathbb{R}^n)$ as our Banach space, with norm $\|x\| = \max_{0\le t\le T_0} |x(t)|$, and $C = \{x \in X \,|\, \|x - x_0\| \le \delta\}$ as our closed subset, then $K : C \to C$ and it remains to show that K is a contraction.

To show this, we need to estimate

$$|K(x)(t) - K(y)(t)| \le \int_0^t |f(s,x(s)) - f(s,y(s))|ds. \tag{2.17}$$

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, uniformly with respect to the first argument, that is, for every compact set $V_0 \subset U$ the following number

$$L = \sup_{(t,x)\ne(t,y)\in V_0} \frac{|f(t,x) - f(t,y)|}{|x - y|} \tag{2.18}$$

(which depends on V_0) is finite. Then,

$$\begin{aligned}\int_0^t |f(s,x(s)) - f(s,y(s))|ds &\le L \int_0^t |x(s) - y(s)|ds \\ &\le L\,t \sup_{0\le s\le t} |x(s) - y(s)| \end{aligned} \tag{2.19}$$

provided the graphs of both $x(t)$ and $y(t)$ lie in V_0. In other words,

$$\|K(x) - K(y)\| \le L\,T_0 \|x - y\|, \quad x, y \in C. \tag{2.20}$$

Moreover, choosing $T_0 < L^{-1}$ we see that K is a contraction and existence of a unique solution follows from the contraction principle:

Theorem 2.2 (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, uniformly with respect to the first, then there exists a unique local solution $\overline{x}(t) \in C^1(I)$ of the IVP (2.10), where I is some interval around t_0.*

More specific, if $V = [t_0, t_0 + T] \times \overline{B_\delta(x_0)} \subset U$ and M denotes the maximum of $|f|$ on V. Then the solution exists at least for $t \in [t_0, t_0 + T_0]$ and remains in $\overline{B_\delta(x_0)}$, where $T_0 = \min\{T, \frac{\delta}{M}\}$. The analogous result holds for the interval $[t_0 - T, t_0]$.

Proof. We have already shown everything except for the fact that our proof requires $T_0 < L^{-1}$ in addition to $T_0 \leq T$ and $T_0 \leq \frac{\delta}{M}$. That this condition is indeed superfluous will be shown in the next section. □

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 evaluating the integrals is often too time consuming. However, if $f(t, x)$ is analytic, the m'th Picard iterate $x_m(t)$ matches the Taylor expansion of the solution $\overline{x}(t)$ around t_0 up to order m and this can be used for numerical computations (cf. Problem 4.4). In any event, the important fact for us is that there exists a unique solution to the initial value problem.

In many cases, f will be even differentiable. In particular, recall that $f \in C^1(U, \mathbb{R}^n)$ implies that f is locally Lipschitz continuous in the second argument, uniformly with respect to the first, as required in Theorem 2.2 (see Problem 2.5 below).

Lemma 2.3. *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 (2.10) is $C^{k+1}(I)$.*

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. □

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

$$|f(y) - f(x)| \leq \sup_{\varepsilon \in [0,1]} \left\| \frac{\partial f(x + \varepsilon(y - x))}{\partial x} \right\| |x - y|,$$

where $\frac{\partial f(x_0)}{\partial x}$ denotes the Jacobian matrix at x_0 and $\|.\|$ denotes the matrix norm (cf. (3.8)*). Conclude that a function $f \in C^1(U, \mathbb{R}^n)$, $U \subseteq \mathbb{R}^{n+1}$, is locally Lipschitz continuous in the second argument, uniformly with respect to the first, and thus satisfies the hypothesis of Theorem 2.2. (Hint: Start with the case $m = n = 1$.)*

Problem 2.6. *Are the following functions Lipschitz continuous near 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 2.7. *Apply the Picard iteration to the first-order linear equation*

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

Problem 2.8. *Apply the Picard iteration to the first-order equation*

$$\dot{x} = 2t - 2\sqrt{\max(0, x)}, \qquad x(0) = 0.$$

Does it converge?

2.3. Some extensions

In this section we want to derive some further extensions of the Picard–Lindelöf theorem. They are of a more technical nature and can be skipped on first reading.

As a preparation we need a slight generalization of the contraction principle. In fact, looking at its proof, observe that we can replace θ^n by any other summable sequence θ_n (Problem 2.10).

Theorem 2.4 (Weissinger)**.** *Let C be a (nonempty) closed subset of a Banach space X. Suppose $K : C \to C$ satisfies*

$$\|K^n(x) - K^n(y)\| \le \theta_n \|x - y\|, \qquad x, y \in C, \tag{2.21}$$

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

$$\|K^n(x) - \overline{x}\| \le \left(\sum_{j=n}^{\infty} \theta_j \right) \|K(x) - x\|, \qquad x \in C. \tag{2.22}$$

Our first objective is to give some concrete values for the existence time T_0. Using Weissinger's theorem instead of the contraction principle, we can avoid the restriction $T_0 < L^{-1}$:

Theorem 2.5 (improved Picard–Lindelöf)**.** *Suppose $f \in C(U, \mathbb{R}^n)$, where U is an open subset of $\mathbb{R}^{n+1}$, and f is locally Lipschitz continuous in the second argument. Choose $(t_0, x_0) \in U$ and $\delta, T > 0$ such that $[t_0, t_0 + T] \times \overline{B_\delta(x_0)} \subset U$. Set*

$$M(t) = \int_{t_0}^{t} \sup_{x \in B_\delta(x_0)} |f(s,x)| ds, \tag{2.23}$$

$$L(t) = \sup_{x \neq y \in B_\delta(x_0)} \frac{|f(t,x) - f(t,y)|}{|x-y|}. \tag{2.24}$$

Note that $M(t)$ is nondecreasing and define T_0 via

$$T_0 = \sup\{0 < t \leq T |\, M(t_0 + t) \leq \delta\}. \tag{2.25}$$

Suppose

$$L_1(T_0) = \int_{t_0}^{t_0+T_0} L(t) dt < \infty. \tag{2.26}$$

Then the unique local solution $\overline{x}(t)$ of the IVP (2.10) *is given by*

$$\overline{x} = \lim_{m \to \infty} K^m(x_0) \in C^1([t_0, t_0 + T_0], \overline{B_\delta(x_0)}), \tag{2.27}$$

where $K^m(x_0)$ is defined in (2.13)*, and satisfies the estimate*

$$\sup_{t_0 \leq t \leq T_0} |\overline{x}(t) - K^m(x_0)(t)| \leq \frac{L_1(T_0)^m}{m!} \mathrm{e}^{L_1(T_0)} \int_{t_0}^{t_0+T_0} |f(s, x_0)| ds. \tag{2.28}$$

An analogous result holds for $t < t_0$.

Proof. Again we choose $t_0 = 0$ for notational simplicity. Our aim is to verify the assumptions of Theorem 2.4 choosing $X = C([0, T_0], \mathbb{R}^n)$ with norm $\|x\| = \max_{0 \leq t \leq T_0} |x(t)|$ and $C = \{x \in X \,|\, \|x - x_0\| \leq \delta\}$.

First of all, if $x \in C$ we have

$$|K(x)(t) - x_0| \leq \int_0^t |f(s, x(s))| ds \leq M(t) \leq \delta, \qquad t \in [0, T_0],$$

that is, $K(x) \in C$ as well. In particular, this explains our choice for T_0.

Next we claim

$$|K^m(x)(t) - K^m(y)(t)| \leq \frac{L_1(t)^m}{m!} \sup_{0 \leq s \leq t} |x(s) - y(s)|, \tag{2.29}$$

where $L_1(t) = \int_0^t L(s)ds$. This follows by induction:

$$\begin{aligned} |K^{m+1}(x)(t) - K^{m+1}(y)(t)| &\le \int_0^t |f(s, K^m(x)(s)) - f(s, K^m(y)(s))|ds \\ &\le \int_0^t L(s)|K^m(x)(s) - K^m(y)(s)|ds \\ &\le \int_0^t L(s)\frac{L_1(s)^m}{m!} \sup_{r\le s} |x(r) - y(r)|ds \\ &\le \sup_{r\le t} |x(r) - y(r)| \int_0^t L_1'(s)\frac{L_1(s)^m}{m!}ds \\ &= \frac{L_1(t)^{m+1}}{(m+1)!} \sup_{r\le t} |x(r) - y(r)|. \end{aligned}$$

Hence K satisfies the assumptions of Theorem 2.4 which finally yields

$$\sup_{0\le t\le T_0} |\overline{x}(t) - K^m(x_0)(t)| \le \sum_{j=m}^{\infty} \left(\frac{L_1(T_0)^j}{j!}\right) \int_0^{T_0} |f(s, x_0)|ds.$$

□

Note that if we set

$$M = \sup_{(t,x)\in[t_0,T]\times B_\delta(x_0)} |f(t,x)| \tag{2.30}$$

then we can chose

$$T_0 = \min(T, \frac{M}{\delta}). \tag{2.31}$$

If $f(t,x)$ is defined for all $x \in \mathbb{R}^n$ and we can find a *global* Lipschitz constant, then we can say more about the interval where the solution exists:

Corollary 2.6. *Suppose* $[t_0, T] \times \mathbb{R}^n \subset U$ *and*

$$\int_{t_0}^T L(t)dt < \infty, \qquad L(t) = \sup_{x\ne y\in\mathbb{R}^n} \frac{|f(t,x) - f(t,y)|}{|x-y|}, \tag{2.32}$$

then $\overline{x}$ *is defined for all* $t \in [t_0, T]$.

In particular, if $U = \mathbb{R}^{n+1}$ *and* $\int_{-T}^T L(t)dt < \infty$ *for all* $T > 0$, *then* $\overline{x}$ *is defined for all* $t \in \mathbb{R}$.

Proof. In this case we can simply choose our closed set C to be the entire Banach space $X = C([0,T], \mathbb{R}^n)$ (i.e., $\delta = \infty$) and proceed as in the proof of the previous theorem with $T_0 = T$. □

Note that this corollary applies for example if the differential equation is linear, that is, $f(t,x) = A(t)x + b(t)$, where $A(t)$ is a matrix and $b(t)$ is a vector which both have continuous entries.

Finally, let me remark that the requirement that f is continuous in Theorem 2.2 is already more than we actually needed in its proof. In fact, all one needs to require is that f is measurable with $M(t)$ finite and $L(t)$ locally integrable (i.e., $\int_I L(t)dt < \infty$ for any compact interval I).

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

$$\dot{x} = \operatorname{sgn}(t)x, \qquad x(0) = 1. \tag{2.33}$$

Then $x(t) = \exp(|t|)$ might be considered a solution even though it is not differentiable at $t = 0$. This generalization is known as **differential equations in the sense of Carathéodory**.

Problem 2.9. *Consider the initial value problem $\dot{x} = x^2$, $x(0) = x_0 > 0$. What is the maximal value for T_0 (as a function of x_0) according to Theorem 2.2 respectively Theorem 2.5? What maximal value do you get from the explicit solution? (Hint: Compute T_0 as a function of δ and find the optimal δ.)*

Problem 2.10. *Prove Theorem 2.4. 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 2.4 (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.*

2.4. 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 2.7 (Generalized Gronwall's inequality). *Suppose $\psi(t)$ satisfies*

$$\psi(t) \le \alpha(t) + \int_0^t \beta(s)\psi(s)ds, \qquad t \in [0,T], \tag{2.34}$$

with $\alpha(t) \in \mathbb{R}$ and $\beta(t) \ge 0$. Then

$$\psi(t) \le \alpha(t) + \int_0^t \alpha(s)\beta(s)\exp\left(\int_s^t \beta(r)dr\right)ds, \qquad t \in [0,T]. \tag{2.35}$$

Moreover, if in addition $\alpha(s) \le \alpha(t)$ for $s \le t$, then

$$\psi(t) \le \alpha(t)\exp\left(\int_0^t \beta(s)ds\right), \qquad t \in [0,T]. \tag{2.36}$$

Proof. Abbreviate $\phi(t) = \exp\big(-\int_0^t \beta(s)ds\big)$. Then one computes

$$\frac{d}{dt}\phi(t)\int_0^t \beta(s)\psi(s)ds = \beta(t)\phi(t)\left(\psi(t) - \int_0^t \beta(s)\psi(s)ds\right) \leq \alpha(t)\beta(t)\phi(t)$$

by our assumption (2.34). Integrating this inequality with respect to t and dividing the resulting equation by $\phi(t)$ shows

$$\int_0^t \beta(s)\psi(s)ds \leq \int_0^t \alpha(s)\beta(s)\frac{\phi(s)}{\phi(t)}ds.$$

Adding $\alpha(t)$ on both sides and using again (2.34) finishes the proof of the first claim. The second claim is left as an exercise (Problem 2.11). □

We will also frequently use the following simple consequence (Problem 2.12): If

$$\psi(t) \leq \alpha + \int_0^t (\beta\,\psi(s) + \gamma)ds, \qquad t \in [0,T], \tag{2.37}$$

for given constants $\alpha \in \mathbb{R}$, $\beta \geq 0$, and $\gamma \in \mathbb{R}$, then

$$\psi(t) \leq \alpha\,\exp(\beta t) + \frac{\gamma}{\beta}(\exp(\beta t) - 1), \qquad t \in [0,T]. \tag{2.38}$$

In the case $\beta = 0$ the right-hand side has to be replaced by its limit $\psi(t) \leq \alpha + \gamma t$. Of course this last inequality does not provide any new insights.

Now we can show that our IVP is well-posed.

Theorem 2.8. *Suppose $f, g \in C(U, \mathbb{R}^n)$ and let f be locally Lipschitz continuous in the second argument, uniformly with respect to the first. If $x(t)$ and $y(t)$ are respective solutions of the IVPs*

$$\begin{array}{c} \dot{x} = f(t,x) \\ x(t_0) = x_0 \end{array} \quad \textit{and} \quad \begin{array}{c} \dot{y} = g(t,y) \\ y(t_0) = y_0 \end{array}, \tag{2.39}$$

then

$$|x(t) - y(t)| \leq |x_0 - y_0|\,\mathrm{e}^{L|t-t_0|} + \frac{M}{L}(\mathrm{e}^{L|t-t_0|} - 1), \tag{2.40}$$

where

$$L = \sup_{(t,x)\neq(t,y)\in V} \frac{|f(t,x) - f(t,y)|}{|x-y|}, \quad M = \sup_{(t,x)\in V} |f(t,x) - g(t,x)|, \tag{2.41}$$

with $V \subset U$ some set containing the graphs of $x(t)$ and $y(t)$.

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

$$|x(t) - y(t)| \leq |x_0 - y_0| + \int_0^t |f(s,x(s)) - g(s,y(s))|ds.$$

Estimating the integrand shows

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

Hence the claim follows from (2.38). □

In particular, denote the solution of the IVP (2.10) by

$$\phi(t, t_0, x_0) \tag{2.42}$$

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

$$|\phi(t, t_0, x_0) - \phi(t, t_0, y_0)| \le |x_0 - y_0| \, \mathrm{e}^{L|t-t_0|}, \tag{2.43}$$

shows that ϕ depends continuously on the initial value. Of course this bound blows up exponentially as t increases, but the linear equation $\dot{x} = x$ in one dimension shows that we cannot do better in general.

Moreover, we even have

Theorem 2.9. *Suppose $f \in C(U, \mathbb{R}^n)$ is locally Lipschitz continuous in the second argument, uniformly with respect to the first. Around each point $(t_0, x_0) \in U$ we can find a compact set $I \times B \subset U$ such that $\phi(t, s, x) \in C(I \times I \times B, \mathbb{R}^n)$. Moreover, $\phi(t, t_0, x_0)$ is Lipschitz continuous,*

$$|\phi(t, t_0, x_0) - \phi(s, s_0, y_0)| \le |x_0 - y_0| \, \mathrm{e}^{L|t-t_0|} + (|t - s| + |t_0 - s_0| \mathrm{e}^{L|t-s_0|}) M, \tag{2.44}$$

where

$$L = \sup_{(t,x) \ne (t,y) \in V} \frac{|f(t,x) - f(t,y)|}{|x - y|}, \qquad M = \max_{(t,x) \in V} |f(t,x)|, \tag{2.45}$$

with $V \subset U$ some compact set containing $I \times \phi(I \times I \times B)$.

Proof. Using the same notation as in the proof of Theorem 2.2 we can find a compact set $V = [t_0 - \varepsilon, t_0 + \varepsilon] \times \overline{B_\delta(x_0)}$ such that $\phi(t, t_0, x_0)$ exists for $|t - t_0| \le \varepsilon$. But then it is straightforward to check that $V_1 = [t_1 - \varepsilon/2, t_1 + \varepsilon/2] \times \overline{B_{\delta/2}(x_1)}$ works to show that $\phi(t, t_1, x_1)$ exists for $|t - t_1| \le \varepsilon/2$ whenever $|t_1 - t_0| \le \varepsilon/2$ and $|x_1 - x_0| \le \delta/2$. Hence we can choose $I = [t_0 - \varepsilon/2, t_0 + \varepsilon/2]$ and $B = \overline{B_{\delta/2}(x_0)}$.

To obtain the estimate observe

$$\begin{aligned}|\phi(t,t_0,x_0)-\phi(s,s_0,y_0)| \le{}& |\phi(t,t_0,x_0)-\phi(t,t_0,y_0)|\\ &+|\phi(t,t_0,y_0)-\phi(t,s_0,y_0)|\\ &+|\phi(t,s_0,y_0)-\phi(s,s_0,y_0)|\\ \le{}& |x_0-y_0|\,\mathrm{e}^{L|t-t_0|}\\ &+|\int_{t_0}^t f(r,\phi(r,t_0,y_0))dr-\int_{s_0}^t f(r,\phi(r,s_0,y_0))dr|\\ &+|\int_s^t f(r,\phi(r,s_0,y_0))dr|,\end{aligned}$$

where we have used (2.43) for the first term. Moreover, the third term can clearly be estimated by $M|t-s|$. To estimate the second term, abbreviate $\Delta(t)=\phi(t,t_0,y_0)-\phi(t,s_0,y_0)$ and use (assume $t_0\le s_0\le t$ without loss of generality)

$$\begin{aligned}\Delta(t)\le{}&\left|\int_{t_0}^{s_0} f(r,\phi(r,t_0,y_0))dr\right|+\int_{s_0}^t|f(r,\phi(r,t_0,y_0))-f(r,\phi(r,s_0,y_0))|dr\\ \le{}& |t_0-s_0|M+L\int_{s_0}^t\Delta(r)dr.\end{aligned}$$

Hence an application of Gronwall's inequality finishes the proof. □

Note that in the case of an autonomous system we have $\phi(t,t_0,x_0)=\phi(t-t_0,0,x_0)$ by Problem 1.8 and it suffices to consider $\phi(t,x_0)=\phi(t,0,x_0)$ in such a situation.

However, in many cases the previous result is not good enough and we need to be able to differentiate with respect to the initial condition. Hence we will assume $f\in C^k(U,\mathbb{R}^n)$ for some $k\ge 1$.

We first suppose that $\phi(t,t_0,x)$ is differentiable with respect to x. Then the same is true for $\dot\phi(t,t_0,x)$ by (2.10) combined with the chain rule and differentiating (2.10) yields

$$\frac{\partial^2\phi}{\partial x\partial t}(t,t_0,x)=\frac{\partial f}{\partial x}(t,\phi(t,t_0,x))\frac{\partial\phi}{\partial x}(t,t_0,x). \tag{2.46}$$

Hence, if we further assume that we can interchange the partial derivatives on the left-hand side,

$$\frac{\partial^2\phi}{\partial x\partial t}(t,t_0,x)=\frac{\partial^2\phi}{\partial t\partial x}(t,t_0,x), \tag{2.47}$$

we see that

$$\frac{\partial\phi}{\partial x}(t,t_0,x) \tag{2.48}$$

necessarily satisfies the **first variational equation**

$$\dot{y} = A(t,x)y, \qquad A(t,x) = \frac{\partial f}{\partial x}(t, \phi(t,t_0,x)). \tag{2.49}$$

Note that this equation is linear and the corresponding integral equation reads

$$y(t) = \mathbb{I} + \int_{t_0}^{t} A(s,x)y(s)ds, \tag{2.50}$$

where we have used $\phi(t_0,t_0,x) = x$ and hence $\frac{\partial\phi}{\partial x}(t_0,t_0,x) = \mathbb{I}$. Applying similar fixed point techniques as before, one can show that the first variational equation has a solution which is indeed the derivative of $\phi(t,t_0,x)$ with respect to x.

Theorem 2.10. *Suppose $f \in C^k(U,\mathbb{R}^n)$, $k \geq 1$. Around each point $(t_0,x_0) \in U$ we can find an open set $I \times B \subseteq U$ such that $\phi(t,s,x) \in C^k(I \times I \times B, \mathbb{R}^n)$. Moreover, $\frac{\partial}{\partial t}\phi(t,s,x) \in C^k(I \times I \times B, \mathbb{R}^n)$ and if D_k is a partial derivative of order k, then $D_k\phi$ satisfies the higher order variational equation obtained from*

$$\frac{\partial}{\partial t} D_k \phi(t,s,x) = D_k \frac{\partial}{\partial t}\phi(t,s,x) = D_k f(t, \phi(t,s,x)) \tag{2.51}$$

by applying the chain rule. In particular, this equation is linear in $D_k\phi$ and it also follows that the corresponding higher order derivatives commute.

Proof. By adding t to the dependent variables it is no restriction to assume that our equation is autonomous and consider $\phi(t,x) = \phi(t,0,x)$. Existence of a set $I \times B \subseteq U$ such that $\phi(t,x_0)$ is continuous has been established in the previous theorem and it remains to investigate differentiability.

We start by showing the case $k = 1$. We have to prove that $\phi(t,x)$ is differentiable at every given point $x_1 \in B$. Without loss of generality we will assume $x_1 = 0$ for notational convenience. We will take $I = (-T,T)$ and B some open ball around x_0 such that the closure of $I \times B$ still lies in U.

Abbreviate $\phi(t) = \phi(t,x_1)$, $A(t) = A(t,x_1)$ and denote by $\psi(t)$ the solution of the first variational equation $\dot{\psi}(t) = A(t)\psi(t)$ corresponding to the initial condition $\psi(t_0) = \mathbb{I}$. Set

$$\theta(t,x) = \frac{\phi(t,x) - \phi(t) - \psi(t)x}{|x|},$$

then $\frac{\partial\phi}{\partial x}$ at $x_1 = 0$ will exist (and be equal to ψ) if we can show $\lim_{x\to 0}\theta(t,x) = 0$.

Our assumption $f \in C^1$ implies

$$f(y) = f(x) + \frac{\partial f}{\partial x}(x)\,(y-x) + \left(\int_0^1 \Big(\frac{\partial f}{\partial x}(x + t(y-x)) - \frac{\partial f}{\partial x}(x)\Big)dt\right)(y-x),$$

and thus

$$f(y) - f(x) = \frac{\partial f}{\partial x}(x)\,(y-x) + |y-x| R(y,x), \tag{2.52}$$

where

$$|R(y,x)| \le \max_{t\in[0,1]} \Big\| \frac{\partial f}{\partial x}(x + t(y-x)) - \frac{\partial f}{\partial x}(x) \Big\|.$$

Here $\|.\|$ denotes the matrix norm (cf. Section 3.1). By uniform continuity of the partial derivatives $\frac{\partial f}{\partial x}$ in a neighborhood of $x_1 = 0$ we infer $\lim_{y\to x} |R(y,x)| = 0$ again uniformly in x in some neighborhood of 0.

Using (2.52) we see

$$\begin{aligned}\dot\theta(t,x) &= \frac{1}{|x|}(f(\phi(t,x)) - f(\phi(t)) - A(t)\psi(t)x)\\ &= A(t)\theta(t,x) + \frac{|\phi(t,x) - \phi(t)|}{|x|} R(\phi(t,x), \phi(t)).\end{aligned}$$

Now integrate and take absolute values (note $\theta(0,x) = 0$ and recall (2.43)) to obtain

$$|\theta(t,x)| \le \tilde R(x) + \int_0^t \|A(s)\| |\theta(s,x)| ds,$$

where

$$\tilde R(x) = \mathrm{e}^{LT} \int_0^T |R(\phi(s,x), \phi(s))| ds.$$

Then Gronwall's inequality implies $|\theta(t,x)| \le \tilde R(x) \exp(\int_0^T \|A(s)\| ds)$. Since $\lim_{y\to x} |R(y,x)| = 0$ uniformly in x in some neighborhood of 0, we have $\lim_{x\to 0} \tilde R(x) = 0$ and hence $\lim_{x\to 0} \theta(t,x) = 0$. Moreover, $\frac{\partial\phi}{\partial x}(t,x)$ is C^0 as the solution of the first variational equation. This settles the case $k = 1$ since all partial derivatives (including the one with respect to t) are continuous.

For the general case $k \ge 1$ we use induction: Suppose the claim holds for k and let $f \in C^{k+1}$. Then $\phi(t,x) \in C^1$ and the partial derivative $\frac{\partial\phi}{\partial x}(t,x)$ solves the first variational equation. But $A(t,x) \in C^k$ and hence $\frac{\partial\phi}{\partial x}(t,x) \in C^k$, which, together with Lemma 2.3, shows $\phi(t,x) \in C^{k+1}$. □

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

$$\dot x(t) = f(t,x,\lambda), \qquad x(t_0) = x_0, \tag{2.53}$$

with corresponding solution

$$\phi(t, t_0, x_0, \lambda). \tag{2.54}$$

Theorem 2.11. *Suppose $f \in C^k(U \times \Lambda, \mathbb{R}^n)$, $k \ge 1$. Around each point $(t_0, x_0, \lambda_0) \in U \times \Lambda$ we can find an open set $I \times B \times \Lambda_0 \subseteq U \times \Lambda$ such that $\phi(t,s,x,\lambda) \in C^k(I \times I \times B \times \Lambda_0, \mathbb{R}^n)$.*

Proof. This follows from the previous result by adding the parameters λ to the dependent variables and requiring $\dot{\lambda} = 0$. Details are left to the reader. □

Problem 2.11. *Show* (2.36).

Problem 2.12. *Show* (2.38). *(Hint: Introduce $\tilde{\psi}(t) = \psi(t) + \frac{\gamma}{\beta}$.)*

Problem 2.13. *Find different functions $f(t,x) = f(x)$ and $g(t,x) = g(x)$ such that the inequality in* (2.40) *becomes an equality.*

Problem 2.14. *Suppose $f \in C(U, \mathbb{R}^n)$ satisfies $|f(t,x) - f(t,y)| \le L(t)|x - y|$. Show that the solution $\phi(t,x_0)$ of* (2.10) *satisfies*

$$|\phi(t,x_0) - \phi(t,y_0)| \le |x_0 - y_0|\, \mathrm{e}^{|\int_{t_0}^t L(s)ds|}.$$

Problem 2.15. *Show that in the one dimensional case, we have*

$$\frac{\partial \phi}{\partial x}(t,x) = \exp\left(\int_{t_0}^t \frac{\partial f}{\partial x}(s, \phi(s,x))ds\right).$$

2.5. Regular perturbation theory

Using Theorem 2.11 we can now also justify the perturbation method proposed in Problem 1.2 for initial value problems depending on a small parameter ε. In general, such a problem is of the form

$$\dot{x} = f(t,x,\varepsilon), \qquad x(t_0) = x_0, \tag{2.55}$$

and known as a **regular perturbation problem**.

If we suppose $f \in C^1$ then Theorem 2.11 ensures that the same is true for the solution $\phi(t,\varepsilon)$, where we do not display the dependence on the initial conditions (t_0, x_0) for notational simplicity. In particular, we have the following Taylor expansions

$$\phi(t,\varepsilon) = \phi_0(t) + \phi_1(t)\varepsilon + o(\varepsilon) \tag{2.56}$$

with respect to ε in a neighborhood of $\varepsilon = 0$.

Clearly the unperturbed term $\phi_0(t) = \phi(t,0)$ is given as the solution of the unperturbed equation

$$\dot{\phi}_0 = f_0(t,\phi_0), \qquad \phi_0(t_0) = x_0, \tag{2.57}$$

where $f_0(t,x) = f(t,x,0)$. Moreover the derivative $\phi_1(t) = \frac{\partial}{\partial\varepsilon}\phi(t,\varepsilon)|_{\varepsilon=0}$ solves the corresponding first variational equation

$$\dot{\phi}_1 = f_{10}(t,\phi_0(t))\phi_1 + f_{11}(t,\phi_0(t)), \qquad \phi_1(t_0) = 0, \tag{2.58}$$

where $f_{10}(t,x) = \frac{\partial}{\partial x} f(t,x,0)$ and $f_{11}(t,x) = \frac{\partial}{\partial\varepsilon} f(t,x,\varepsilon)|_{\varepsilon=0}$. The initial condition $\phi_1(t_0) = 0$ follows from the fact that the initial condition x_0 does not depend on ε, implying $\phi_1(t_0) = \frac{\partial}{\partial\varepsilon}\phi(t_0,\varepsilon)|_{\varepsilon=0} = \frac{\partial}{\partial\varepsilon}x_0|_{\varepsilon=0} = 0$.

Hence once we have the solution of the unperturbed problem $\phi_0(t)$, we can then compute the correction term $\phi_1(t)$ by solving another linear equation.

Note that the procedure can be equivalently described as follows: Plug the Taylor expansion for $\phi(t,\varepsilon)$ into the differential equation, expand the right-hand side with respect to ε, and compare coefficients with respect to powers of ε.

Example. Let us look at a simple example. Consider the equation

$$\dot{v} = -\varepsilon v - g, \quad v(0) = 0, \qquad \varepsilon \geq 0,$$

which models the velocity of a falling object with air resistance (cf. Problem 1.17). The solution can be easily found

$$\phi(t,\varepsilon) = g\frac{\mathrm{e}^{-\varepsilon t} - 1}{\varepsilon}$$

and there is no need for any perturbation techniques. However, we will still apply it to illustrate the method. The unperturbed problem is

$$\dot{\phi}_0 = -g, \quad \phi_0(0) = 0,$$

and the solution is given by $\phi_0(t) = -gt$. Similarly, since $f(t,v,\varepsilon) = -\varepsilon v - g$ it follows that $f_{10}(t,v) = 0$, $f_{11}(t,v) = -v$ and the equation for the first correction term is

$$\dot{\phi}_1 = -\phi_0(t), \quad \phi_1(0) = 0,$$

with solution given by $\phi_1(t) = \frac{g}{2}t^2$. Hence our approximation is

$$v(t) = -g\left(t - \varepsilon\frac{t^2}{2} + o(\varepsilon)\right)$$

which of course coincides with the Taylor expansion of the exact solution. However note, the approximation is only valid for fixed time and will in general get worse as t increases. In fact, observe that for $\varepsilon > 0$ the approximation diverges to $+\infty$ while the exact solution converges to $\frac{g}{\varepsilon}$. ◇

Clearly we can extend this procedure to get further approximations:

Theorem 2.12. *Let Λ be some open interval. Suppose $f \in C^k(U \times \Lambda, \mathbb{R}^n)$, $k \geq 1$ and fix some values $(t_0, x_0, \varepsilon_0) \in U \times \Lambda$. Let $\phi(t,\varepsilon) \in C^k(I \times \Lambda_0, \mathbb{R}^n)$ be the solution of the initial value problem*

$$\dot{x} = f(t,x,\varepsilon), \qquad x(t_0) = x_0, \tag{2.59}$$

guaranteed to exist by Theorem 2.11.

Then

$$\phi(t,\varepsilon) = \sum_{j=0}^{k} \frac{\phi_j(t)}{j!}(\varepsilon - \varepsilon_0)^j + o((\varepsilon - \varepsilon_0)^k), \tag{2.60}$$

where the coefficients can be obtained by recursively solving

$$\dot{\phi}_j = f_j(t, \phi_0, \dots, \phi_j, \varepsilon_0), \qquad \phi_j(t_0) = \begin{cases} x_0, & j = 0, \\ 0, & j \geq 1, \end{cases} \tag{2.61}$$

where the function f_j is recursively defined via

$$\begin{aligned} f_{j+1}(t, x_0, \dots, x_{j+1}, \varepsilon) =& \frac{\partial f_j}{\partial \varepsilon}(t, x_0, \dots, x_j, \varepsilon) \\ & + \sum_{k=0}^{j} \frac{\partial f_j}{\partial x_k}(t, x_0, \dots, x_j, \varepsilon) x_{k+1}, \\ f_0(t, x_0, \varepsilon) =& f(t, x_0, \varepsilon). \end{aligned} \tag{2.62}$$

If we assume $f \in C^{k+1}$ the error term will be $O((\varepsilon - \varepsilon_0)^{k+1})$ uniformly for $t \in I$.

Proof. The result follows by plugging (2.60) into the differential equation and comparing powers of ε. If $f \in C^{k+1}$ we know that $\frac{\partial^{k+1}}{\partial \varepsilon^{k+1}} \phi$ is continuous and hence bounded on $I \times \Lambda_0$, which gives the desired estimate on the remainder in the Taylor expansion. □

A few remarks are in order: Of course we could admit more than one parameter if we are willing to deal with Taylor series in more than one variable. Moreover, we could include the case where the initial condition depends on ε by simply replacing the initial conditions for $\phi_j(t_0)$ by the corresponding expansion coefficients of $x_0(\varepsilon)$.

Finally, we remark that the Taylor expansion will converge if f is analytic with respect to all variables. This will be shown in Theorem 4.2.

Problem 2.16. *Compute the next term ϕ_2 in the above example.*

Problem 2.17. *Approximate the solutions of $\ddot{x} + x + \varepsilon x^3 = 0$, $x(0) = 1$, $\dot{x}(0) = 0$ up to order one. (Hint: It is not necessary to convert this second order equation to a first order system. In order to solve the second order equations you need to use the computer or preview Section 3.3.)*

2.6. 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 of the IVP (2.10) can be defined.

Suppose that solutions of the IVP (2.10) exist locally and are unique (e.g., f is Lipschitz). Let ϕ_1, ϕ_2 be two solutions of the IVP (2.10) 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 with initial condition $x(t_+) = \phi_1(t_+) = \phi_2(t_+)$ shows that both solutions coincide in a neighborhood of t_+ by local uniqueness. This contradicts maximality of t_+ and hence $t_+ = T_+$. Similarly, $t_- = T_-$.

Moreover, we get a solution

$$\phi(t) = \begin{cases} \phi_1(t), & t \in I_1, \\ \phi_2(t), & t \in I_2, \end{cases} \tag{2.63}$$

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.

Theorem 2.13. *Suppose the IVP* (2.10) *has a unique local solution (e.g. the conditions of Theorem 2.5 are satisfied). Then there exists a unique maximal solution defined on some maximal interval* $I_{(t_0,x_0)} = (T_-(t_0, x_0), T_+(t_0, x_0))$.

Proof. Let $\mathcal{S}$ be the set of all solutions ϕ of (2.10) which are defined on an open interval I_ϕ. Let $\mathcal{I} = \bigcup_{\phi\in\mathcal{S}} I_\phi$, which is again open. Moreover, if $t_1 > t_0 \in \mathcal{I}$, then $t_1 \in I_\phi$ for some ϕ and thus $[t_0, t_1) \subseteq I_\phi \subseteq \mathcal{I}$. Similarly for $t_1 < t_0$ and thus $\mathcal{I}$ is an open interval containing t_0. In particular, it is of the form $\mathcal{I} = (T_-, T_+)$. Now define $\phi_{max}(t)$ on $\mathcal{I}$ by $\phi_{max}(t) = \phi(t)$ for some $\phi \in \mathcal{S}$ with $t \in I_\phi$. By our above considerations any two ϕ will give the same value, and thus $\phi_{max}(t)$ is well-defined. Moreover, for every $t_1 > t_0$ there is some $\phi \in \mathcal{S}$ such that $t_1 \in I_\phi$ and $\phi_{max}(t) = \phi(t)$ for $t \in (t_0 - \varepsilon, t_1 + \varepsilon)$ which shows that ϕ_{max} is a solution. By construction there cannot be a solution defined on a larger interval. □

The solution found in the previous theorem is called the **maximal solution**. A solution defined for all $t \in \mathbb{R}$ is called a **global solution**. Clearly every global solution is maximal.

Remark: If we drop the requirement that f is Lipschitz, we still have existence of solutions (see Theorem 2.19 below), but we already know that we might lose uniqueness. Even without uniqueness, two given solutions of the IVP (2.10) 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 can be used to ensure existence of maximal solutions in this case. For example, consider the differential equation $\dot{x} = \sqrt{|x|}$ where we have found global (and thus maximal) solutions which are however not unique.

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

Lemma 2.14. *Let $\phi(t)$ be a solution of* (2.10) *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 there exists a sequence $t_m \in (t_-, t_+)$ such that*

$$\lim_{m\to\infty} (t_m, \phi(t_m)) = (t_+, y) \in U. \tag{2.64}$$

Similarly for t_-.

Proof. Clearly, if there is an extension, then (2.64) holds for any sequence $t_m \uparrow t_+$. Conversely, suppose there is a sequence satisfying (2.64). We first show that in this case

$$\lim_{t\uparrow t_+} \phi(t) = y. \tag{2.65}$$

Intuitively this follows, since otherwise the solution would need to oscillate faster and faster as t approaches t_+. Consequently its derivative would need to grow, which is impossible since $f(t,x)$ is bounded near y. More precisely, since U is open there is some $\delta > 0$ such that $V = [t_+ - \delta, t_+] \times \overline{B_\delta(y)} \subset U$ and $M = \max_{(t,x)\in V} |f(t,x)| < \infty$. Moreover, after maybe passing to a subsequence, we can assume that $t_m \in (t_+ - \delta, t_+)$, $\phi(t_m) \in B_\delta(y)$, and $t_m < t_{m+1}$. If (2.65) were wrong, we could find a sequence $\tau_m \uparrow t_+$ such that $|\phi(\tau_m) - y| \ge \gamma > 0$. Without loss we can choose $\gamma < \delta$ and $\tau_m \ge t_m$. Moreover, by the intermediate value theorem we can even require $|\phi(\tau_m) - y| = \gamma$ and $|\phi(t) - y| < \delta$ for $t \in [t_m, \tau_m]$. But then

$$\begin{aligned} 0 < \gamma = |\phi(\tau_m) - y| &\le |\phi(\tau_m) - \phi(t_m)| + |\phi(t_m) - y| \\ &\le \int_{t_m}^{\tau_m} |f(s,\phi(s))| ds + |\phi(t_m) - y| \le M|\tau_m - t_m| + |\phi(t_m) - y|, \end{aligned}$$

where the right-hand side converges to 0 as $m \to \infty$. A contradiction and thus (2.65) holds.

Now take 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 function on $(t_-, t_+ + \varepsilon)$. This function is continuous by construction and the limits of its left and right derivative are both equal to $f(t_+, y)$. Hence it is differentiable at $t = t_+$ and thus a solution defined on $(t_-, t_+ + \varepsilon)$. □

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.

Corollary 2.15. *Let $\phi(t)$ be a solution of* (2.10) *defined on the interval (t_-, t_+). Suppose there is a compact set $[t_0, t_+] \times C \subset U$ such that $\phi(t_m) \in C$ for some sequence $t_m \in [t_0, t_+)$ converging to 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 $t_m \to t_+$. By compactness $\phi(t_m)$ has a convergent subsequence and the claim follows from the previous lemma. □

The logical negation of this result is also of interest.

Corollary 2.16. *Let $I_{(t_0,x_0)} = (T_-(t_0,x_0), T_+(t_0,x_0))$ be the maximal interval of existence of a solution starting at $x(t_0) = x_0$. If $T_+ = T_+(t_0,x_0) < \infty$, then the solution must eventually 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 2.17. *Suppose $U = \mathbb{R} \times \mathbb{R}^n$ and for every $T > 0$ there are constants $M(T)$, $L(T)$ such that*

$$|f(t,x)| \le M(T) + L(T)|x|, \qquad (t,x) \in [-T,T] \times \mathbb{R}^n. \tag{2.66}$$

Then all solutions of the IVP (2.10) are defined for all $t \in \mathbb{R}$.

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

$$|\phi(t)| \le |x_0| + \int_0^t (M + L|\phi(s)|)ds, \quad t \in [0,T] \cap I,$$

and the variant (2.38) of Gronwall's inequality shows

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

Thus ϕ lies in a compact ball and the result follows by the previous lemma. □

Again, let me remark that it suffices to assume

$$|f(t,x)| \le M(t) + L(t)|x|, \qquad x \in \mathbb{R}^n, \tag{2.67}$$

where $M(t)$, $L(t)$ are locally integrable. A slight extension of the above result is outlined in Problem 2.22.

Problem 2.18. *Show that Theorem 2.17 is false (in general) if the estimate is replaced by*

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

with $\alpha > 1$.

Problem 2.19. *Consider a first-order autonomous system in $\mathbb{R}^n$ 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 $\gamma_j = \{x_j(t) | t \in I_j\} \subset \mathbb{R}^n$ either coincide or are disjoint.*

Problem 2.20. *Consider a first-order autonomous equation 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 (T_-, T_+) for solutions starting in $[0,1]$? Does such a solution have a limit as $t \to T_\pm$?*

Problem 2.21. *Consider a first-order equation in $\mathbb{R}^1$ with $f(t,x)$ defined on $\mathbb{R}\times\mathbb{R}$. Suppose $x\,f(t,x) < 0$ for $|x| > R$. Show that all solutions exist for all $t > 0$.*

Problem 2.22. *Suppose $U = \mathbb{R}\times\mathbb{R}^n$ and that*

$$|f(t,x)| \le g(|x|)$$

for some positive continuous function $g \in C([0,\infty))$ which satisfies

$$\int_0^\infty \frac{dr}{g(r)} = \infty.$$

Then all solutions of the IVP (2.10) are defined for all $t \ge 0$.

Show that the same conclusion still holds if there is such a function $g_T(r)$ for every $t \in [0,T]$.

(Hint: Look at the differential equation for $r(t)^2 = |x(t)|^2$. Estimate the right-hand side and recall the analysis from Sections 1.3 and 1.5.)

2.7. 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 (2.10).

If $\phi(t)$ is a solution, then by Taylor's theorem we have

$$\phi(t_0+h) = x_0 + \dot\phi(t_0)h + o(h) = x_0 + f(t_0,x_0)h + o(h). \tag{2.68}$$

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

$$x_h(t_{m+1}) = x_h(t_m) + f(t_m, x_h(t_m))h, \qquad t_m = t_0 + mh, \tag{2.69}$$

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 2.18 (Arzelà–Ascoli)**.** *Suppose the sequence of functions $x_m(t) \in C(I,\mathbb{R}^n)$, $m \in \mathbb{N}$, on a compact interval I is (uniformly)* **equicontinuous***, that is, for every $\varepsilon > 0$ there is a $\delta > 0$ (independent of m) such that*

$$|x_m(t) - x_m(s)| \le \varepsilon \quad \text{if} \quad |t-s| < \delta,\ m \in \mathbb{N}. \tag{2.70}$$

If the sequence x_m is bounded, then there is a uniformly convergent subsequence.

Proof. Let $\{t_j\}_{j=1}^\infty \subset I$ be a dense subset of our interval (e.g., all rational numbers in I). Since $x_m(t_1)$ is bounded, we can choose a subsequence $x_m^{(1)}(t)$ such that $x_m^{(1)}(t_1)$ converges (Bolzano–Weierstraß). Similarly we can extract a subsequence $x_m^{(2)}(t)$ from $x_m^{(1)}(t)$ which converges at t_2 (and hence also at t_1 since it is a subsequence of $x_m^{(1)}(t)$). By induction we get a sequence $x_m^{(j)}(t)$ converging at $t_1, \dots, t_j$. The diagonal sequence $\tilde{x}_m(t) = x_m^{(m)}(t)$ will hence converge for all $t = t_j$ (why?). We will show that it converges uniformly for all t:

Fix $\varepsilon > 0$ and choose δ such that $|x_m(t) - x_m(s)| \le \frac{\varepsilon}{3}$ for $|t-s| < \delta$. The balls $B_\delta(t_j)$ cover I and by compactness even finitely many, say $1 \le j \le p$, suffice. Furthermore, choose N_ε such that $|\tilde{x}_m(t_j) - \tilde{x}_n(t_j)| \le \frac{\varepsilon}{3}$ for $n, m \ge N_\varepsilon$ and $1 \le j \le p$.

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

$$\begin{aligned} |\tilde{x}_m(t) - \tilde{x}_n(t)| \le & |\tilde{x}_m(t) - \tilde{x}_m(t_j)| + |\tilde{x}_m(t_j) - \tilde{x}_n(t_j)| \\ & + |\tilde{x}_n(t_j) - \tilde{x}_n(t)| \le \varepsilon \end{aligned}$$

for $n, m \ge N_\varepsilon$, which shows that $\tilde{x}_m$ is Cauchy with respect to the maximum norm. By completeness of $C(I,\mathbb{R}^n)$ it has a limit. □

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

$$M = \max_{(t,x)\in V} |f(t,x)|. \tag{2.71}$$

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

$$|x_h(t) - x_h(s)| \le M|t-s|. \tag{2.72}$$

Hence any subsequence of the family $x_h(t)$ is equicontinuous and there is a uniformly convergent subsequence $\phi_m(t) \to \phi(t)$. It remains to show that the limit $\phi(t)$ solves our initial value problem (2.10). We will show this by verifying that the corresponding integral equation (2.11) holds. Since f is uniformly continuous on V, we can find a sequence $\Delta(h) \to 0$ as $h \to 0$, such that

$$|f(s,y) - f(t,x)| \le \Delta(h) \quad \text{for} \quad |y-x| \le Mh,\ |s-t| \le h. \tag{2.73}$$

To be able to estimate the difference between left and right-hand side of (2.11) for $x_h(t)$ we choose an m with $t \le t_m$ and write

$$x_h(t) = x_0 + \sum_{j=0}^{m-1} \int_{t_j}^{t_{j+1}} \chi(s) f(t_j, x_h(t_j)) ds, \tag{2.74}$$

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

$$\begin{aligned} &\left| x_h(t) - x_0 - \int_{t_0}^t f(s, x_h(s)) ds \right| \\ &\quad \le \sum_{j=0}^{m-1} \int_{t_j}^{t_{j+1}} \chi(s) |f(t_j, x_h(t_j)) - f(s, x_h(s))| ds \\ &\quad \le \Delta(h) \sum_{j=0}^{m-1} \int_{t_j}^{t_{j+1}} \chi(s) ds = |t - t_0| \Delta(h), \end{aligned} \tag{2.75}$$

from which it follows that ϕ is indeed a solution

$$\phi(t) = \lim_{m\to\infty} \phi_m(t) = x_0 + \lim_{m\to\infty} \int_{t_0}^t f(s, \phi_m(s)) ds = x_0 + \int_{t_0}^t f(s, \phi(s)) ds \tag{2.76}$$

since we can interchange limit and integral by uniform convergence.

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

Theorem 2.19 (Peano). *Suppose f is continuous on $V = [t_0, t_0 + T] \times \overline{B_\delta(x_0)} \subset U$ and denote the maximum of $|f|$ by M. Then there exists at least one solution of the initial value problem (2.10) for $t \in [t_0, t_0 + T_0]$ which remains in $\overline{B_\delta(x_0)}$, where $T_0 = \min\{T, \frac{\delta}{M}\}$. The analogous result holds for the interval $[t_0 - T_0, t_0]$.*

Of course this theorem raises the question if there are also conditions on f which are weaker than the Lipschitz condition but still guarantee uniqueness. One such condition is presented in Problem 2.25.

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. By (2.29) with $x(t) = x_h(t)$ and $y(t) = K(x_h)(t)$

this yields

$$\begin{aligned}\|x_h - K^m(x_h)\| &\le \sum_{j=0}^{m-1} \|K^j(x_h) - K^{j+1}(x_h)\| \\ &\le \|x_h - K(x_h)\| \sum_{j=0}^{m-1} \frac{(LT_0)^j}{j!}, \end{aligned} \tag{2.77}$$

using the same notation as in the proof of Theorem 2.2. Taking $n \to \infty$ we finally obtain

$$\|x_h - \phi\| \le T_0 \mathrm{e}^{LT_0} \Delta(h), \qquad t \in [t_0, t_0 + T_0], \tag{2.78}$$

since our above estimate (2.75) for $t = t_0 + T_0$ reads

$$\|x_h - K(x_h)\| \le T_0 \Delta(h). \tag{2.79}$$

Note that if we can find some Lipschitz constant L_0 such that $|f(t,x) - f(s,x)| \le L_0|t-s|$, then we can choose $\Delta(h) = (L_0 + LM)h$.

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. Setting $t_m = t_0 + hm$ and $x_m = x_h(t_m)$ the resulting algorithm

$$x_{m+1} = x_m + \frac{h}{6}(k_{1,m} + 2k_{2,m} + 2k_{3,m} + k_{4,m}), \tag{2.80}$$

where

$$\begin{array}{ll} k_{1,m} = f(t_m, x_m), & k_{2,m} = f(t_m + \frac{h}{2}, x_m + \frac{h}{2}k_{1,m}), \\ k_{3,m} = f(t_m + \frac{h}{2}, x_m + \frac{h}{2}k_{2,m}), & k_{4,m} = f(t_{m+1}, x_m + hk_{3,m}), \end{array} \tag{2.81}$$

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

Problem 2.23. Heun's method *(or* **improved Euler***) is given by*

$$x_{m+1} = x_m + \frac{h}{2}\big(f(t_m, x_m) + f(t_{m+1}, y_m)\big), \quad y_m = x_m + f(t_m, x_m)h.$$

Show that using this method the error during one step is of $O(h^3)$ (provided $f \in C^2$):

$$\phi(t_0 + h) = x_0 + \frac{h}{2}\big(f(t_0, x_0) + f(t_1, y_0)\big) + O(h^3).$$

Note that this is not the only possible scheme with this error order since

$$\phi(t_0 + h) = x_0 + \frac{h}{2}\big(f(t_1, x_0) + f(t_0, y_0)\big) + O(h^3)$$

as well.

Problem 2.24. *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.*

Problem 2.25 (Osgood uniqueness criterion)**.** *We call a continuous nondecreasing function $\rho : [0, \infty) \to [0, \infty)$ with $\rho(0) = 0$ a* **module of continuity***. It is said to satisfy the* **Osgood condition** *if*

$$\int_0^1 \frac{dr}{\rho(r)} = \infty.$$

We will say that a function $f : \mathbb{R} \to \mathbb{R}$ is ρ-continuous if $|f(x) - f(y)| \leq C\rho(|x - y|)$ for some constant C. For example in the case $\rho(r) = r^\alpha$, $\alpha \in (0, 1)$, we obtain the Hölder continuous functions and in the case $\rho(r) = r$ the Lipschitz continuous functions. Note that only in the Lipschitz case the Osgood condition holds. Another module satisfying the Osgood condition is $\rho(r) = r(1 + |\log(r)|)$, the corresponding functions are known as almost Lipschitz functions.

Let $f(t, x)$ be as in the Peano theorem and suppose

$$|(x - y) \cdot (f(t, x) - f(t, y))| \leq C|x - y|\rho(|x - y|),$$

$t \in [t_0, t_0 + T]$, $x, y \in B_\delta(x_0)$, for some modulus of continuity which satisfies the Osgood condition (here the $\cdot$ indicates the scalar product). Then the solution is unique.

(Hint: Consider the difference of two solutions $R(t) = |x(t) - y(t)|^2$ and suppose $R(t_1) = 0$ but $R(t) > 0$ for $t \in (t_1, t_2)$. Estimate $\dot{R}$ using the assumptions and proceed as for a separable equation to obtain a contradiction.)

Chapter 3

Linear equations

3.1. The matrix exponential

We begin with the study of the autonomous linear first-order system

$$\dot{x}(t) = Ax(t), \qquad x(0) = x_0, \tag{3.1}$$

where A is an n by n matrix. Here, as usual, we write Ax for the matrix product whose components are given by

$$(Ax)_i = \sum_{j=1}^{n} A_{i,j} x_j, \tag{3.2}$$

where $(A_{i,j})_{1 \le i,j \le n}$ are the entries of A and $(x_j)_{1 \le j \le n}$ are the components of x. We also recall the definition of the scalar product and norm

$$x \cdot y = \sum_{j=1}^{n} x_j^* y_j, \qquad |x| = \sqrt{x \cdot x} = \left(\sum_{j=1}^{n} |x_j|^2 \right)^{1/2}. \tag{3.3}$$

Here $*$ denotes complex conjugation which can of course be omitted in the real case. We will also use A^j for the powers of A defined inductively via $A^j = A^{j-1} A$ and $A^0 = \mathbb{I}$.

If we perform the Picard iteration we obtain

$$\begin{aligned}
x_0(t) &= x_0\\
x_1(t) &= x_0 + \int_0^t A x_0(s) ds = x_0 + A x_0 \int_0^t ds = x_0 + t A x_0\\
x_2(t) &= x_0 + \int_0^t A x_1(s) ds = x_0 + A x_0 \int_0^t ds + A^2 x_0 \int_0^t s\, ds\\
&= x_0 + t A x_0 + \frac{t^2}{2} A^2 x_0
\end{aligned}$$

and hence by induction

$$x_m(t) = \sum_{j=0}^{m} \frac{t^j}{j!} A^j x_0. \tag{3.4}$$

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

$$x(t) = \lim_{m\to\infty} x_m(t) = \sum_{j=0}^{\infty} \frac{t^j}{j!} A^j x_0. \tag{3.5}$$

In the one dimensional case ($n = 1$) this series is just the usual exponential and hence we will write

$$x(t) = \exp(tA) x_0, \tag{3.6}$$

where we define the **matrix exponential** by

$$\exp(A) = \sum_{j=0}^{\infty} \frac{1}{j!} A^j. \tag{3.7}$$

Hence, in order to understand our original problem, we have to understand the matrix exponential! The Picard iteration ensures convergence of $\exp(A)x_0$ for every vector x_0 and choosing the canonical basis vectors of $\mathbb{R}^n$ we see that all matrix elements converge. However, for later use we want to introduce a suitable norm for matrices and give a direct proof for convergence of the above series in this norm.

We will use $\mathbb{C}^n$ rather than $\mathbb{R}^n$ as underlying vector space since $\mathbb{C}$ is algebraically closed (which will be important later on, when we compute the matrix exponential with the help of the Jordan canonical form). So let A be a complex matrix acting on $\mathbb{C}^n$ and introduce the **matrix norm**

$$\|A\| = \sup_{x\colon |x|=1} |Ax|. \tag{3.8}$$

It is not hard to see that the vector space of n by n matrices $\mathbb{C}^{n\times n}$ becomes a Banach space with this norm (Problem 3.1). In fact, we have

$$\max_{j,k} |A_{jk}| \le \|A\| \le n \max_{j,k} |A_{jk}| \tag{3.9}$$

and thus a sequence of matrices converges in the matrix norm if and only if all matrix entries converge. Moreover, using (Problem 3.2)

$$\|A^j\| \le \|A\|^j \tag{3.10}$$

convergence of the series (3.7) follows from convergence of $\sum_{j=0}^{\infty} \frac{\|A\|^j}{j!} = \exp(\|A\|)$ (Problem 3.4).

However, note that in general $\exp(A+B) \neq \exp(A)\exp(B)$ unless A and B commute, that is, unless the **commutator**

$$[A, B] = AB - BA \tag{3.11}$$

vanishes. In this case you can mimic the proof of the one dimensional case to obtain

Lemma 3.1. *Suppose A and B commute. Then*

$$\exp(A+B) = \exp(A)\exp(B), \qquad [A, B] = 0. \tag{3.12}$$

If we perform a linear change of coordinates,

$$y = U^{-1} x, \tag{3.13}$$

then the matrix exponential in the new coordinates is given by

$$U^{-1}\exp(A)U = \exp(U^{-1}AU). \tag{3.14}$$

This follows from (3.7) by using $U^{-1}A^jU = (U^{-1}AU)^j$ together with continuity of the matrix product (Problem 3.3). Hence in order to compute $\exp(A)$ we need a coordinate transform which renders A as simple as possible:

Theorem 3.2 (Jordan canonical form). *Let A be a complex n by n matrix. Then there exists a linear change of coordinates U such that A transforms into a block matrix,*

$$U^{-1}AU = \begin{pmatrix} J_1 & & \\ & \ddots & \\ & & J_m \end{pmatrix}, \tag{3.15}$$

with each block of the form

$$J = \alpha\mathbb{I} + N = \begin{pmatrix} \alpha & 1 & & & \\ & \alpha & 1 & & \\ & & \alpha & \ddots & \\ & & & \ddots & 1 \\ & & & & \alpha \end{pmatrix}. \tag{3.16}$$

Here N is a matrix with ones in the first diagonal above the main diagonal and zeros elsewhere.

The numbers α are the eigenvalues of A and the new basis vectors u_j (the columns of U) consist of generalized eigenvectors of A. The general procedure of finding the Jordan canonical form is quite cumbersome and hence further details will be deferred to Section 3.8. In particular, since most computer algebra systems can easily do this job for us!

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
```

◇

To compute the exponential we observe

$$\exp(U^{-1}AU) = \begin{pmatrix} \exp(J_1) & & \\ & \ddots & \\ & & \exp(J_m) \end{pmatrix}, \tag{3.17}$$

and hence it remains to compute the exponential of a single Jordan block $J = \alpha\mathbb{I}+N$ as in (3.16). Since $\alpha\mathbb{I}$ commutes with N, we infer from Lemma 3.1 that

$$\exp(J) = \exp(\alpha\mathbb{I})\exp(N) = \mathrm{e}^{\alpha}\sum_{j=0}^{k-1}\frac{1}{j!}N^j. \tag{3.18}$$

Here we have used the fact that the series for $\exp(N)$ terminates after k terms, where k is the size of N. In fact, it is not hard to see that N^j is a matrix with ones in the j'th diagonal above the main diagonal and vanishes once j reaches the size of J:

$$N = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \quad N^2 = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \quad N^3 = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix},$$

and $N^4 = 0$. In summary, $\exp(J)$ explicitly reads

$$\exp(J) = \mathrm{e}^{\alpha} \begin{pmatrix} 1 & 1 & \frac{1}{2!} & \cdots & \frac{1}{(k-1)!} \\ & 1 & 1 & \ddots & \vdots \\ & & 1 & \ddots & \frac{1}{2!} \\ & & & \ddots & 1 \\ & & & & 1 \end{pmatrix}. \tag{3.19}$$

Example. In two dimensions the exponential matrix of

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \tag{3.20}$$

is given by

$$\exp(A) = \mathrm{e}^{\delta} \left(\cosh(\Delta)\mathbb{I} + \frac{\sinh(\Delta)}{\Delta} \begin{pmatrix} \gamma & b \\ c & -\gamma \end{pmatrix} \right), \tag{3.21}$$

where

$$\delta = \frac{a+d}{2}, \quad \gamma = \frac{a-d}{2}, \quad \Delta = \sqrt{\gamma^2 + bc}. \tag{3.22}$$

Here one has to set $\frac{\sinh(\Delta)}{\Delta} = 1$ for $\Delta = 0$. Moreover, note $\cosh(\mathrm{i}\Delta) = \cos(\Delta)$ and $\frac{\sinh(\mathrm{i}\Delta)}{\mathrm{i}\Delta} = \frac{\sin(\Delta)}{\Delta}$.

To see this set $A = \delta\mathbb{I} + B$ and use $\exp(A) = \mathrm{e}^{\delta}\exp(B)$ plus

$$B^m = \begin{cases} \Delta^{2k} B, & m = 2k+1, \\ \Delta^{2k}\mathbb{I}, & m = 2k, \end{cases} \qquad B = \begin{pmatrix} \gamma & b \\ c & -\gamma \end{pmatrix}.$$

Hence

$$\exp(A) = \mathrm{e}^{\delta} \left(\sum_{k=0}^{\infty} \frac{\Delta^{2k}}{(2k)!}\mathbb{I} + \sum_{k=0}^{\infty} \frac{\Delta^{2k}}{(2k+1)!} B \right)$$

establishing the claim. ◇

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 for arbitrary matrices. In fact, we even have:

Lemma 3.3. *A vector u is an eigenvector of A corresponding to the eigenvalue α if and only if u is an eigenvector of* $\exp(A)$ *corresponding to the eigenvalue* e^{α}.

Moreover, the Jordan structure of A and $\exp(A)$ *are the same except for the fact that eigenvalues of A which differ by a multiple of* $2\pi\mathrm{i}$ *(as well as the corresponding Jordan blocks) are mapped to the same eigenvalue of* $\exp(A)$. *In particular, the geometric and algebraic multiplicity of* e^{α} *is the sum of the geometric and algebraic multiplicities of the eigenvalues which differ from α by a multiple of* $2\pi\mathrm{i}$.

Proof. The first part is straightforward. To see the second it suffices to consider one Jordan block with $\alpha = 0$. We are looking for generalized eigenvectors u_k such that $\exp(N)u_k = u_{k-1}$, that is,

$$\sum_{l=j+1}^{n} \frac{1}{(j-l)!} u_{k,l} = u_{k-1,l}, \qquad 2 \le k \le n,\ 1 \le j \le n.$$

Setting $u_{k,l} = \frac{(l-1)!}{(k-1)!} s(k-1, l-1)$ with $s(k,k) = 1$ and $s(k,l) = 0$ for $l > k$ this requirement transforms into

$$\sum_{l=j+1}^{k} \binom{l}{j} s(k,l) = k\, s(k-1,j), \qquad 0 \le j \le k-1,$$

which is satisfied if we choose $s(k,l)$ to be the Stirling numbers of the first kind (Problem 3.6).

Hence the transformation matrix U we are looking for is $U = (\frac{(j-1)!}{(k-1)!} s(k-1, j-1))_{1\le j,k\le n}$ and its inverse is given by $U^{-1} = (\frac{(j-1)!}{(k-1)!} S(k-1, j-1))_{1\le j,k\le n}$ where $S(j,k)$ are the Stirling numbers of the second kind defined via

$$\sum_{k=j}^{n} S(l,k) s(k,j) = \delta_{j,k}, \qquad 1 \le j, l \le n.$$

Then, by construction, $U^{-1}\exp(N)U = N$ and the claim follows. □

Clearly *Mathematica* can also compute the exponential for us:

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

Out[5]//MatrixForm=
```

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

To end this section let me emphasize, that both the eigenvalues and generalized eigenvectors can be complex even if the matrix A has only real entries. However, in many applications only real solutions are of interest. For such a case there is also a **real Jordan canonical form** which I want to mention briefly.

So suppose the matrix A has only real entries. If an eigenvalue α is real, both real and imaginary parts of a generalized eigenvector are again generalized eigenvectors. In particular, they can be chosen real and there is nothing else to do for such an eigenvalue.

If α is nonreal, there must be a corresponding complex conjugate block $J^* = \alpha^*\mathbb{I}+N$ 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

$$\begin{pmatrix} J & 0 \\ 0 & J^* \end{pmatrix} \tag{3.24}$$

is replaced by

$$\begin{pmatrix} R & \mathbb{I} & & & \\ & R & \mathbb{I} & & \\ & & R & \ddots & \\ & & & \ddots & \mathbb{I} \\ & & & & R \end{pmatrix}, \tag{3.25}$$

where

$$R = \begin{pmatrix} \mathrm{Re}(\alpha) & \mathrm{Im}(\alpha) \\ -\mathrm{Im}(\alpha) & \mathrm{Re}(\alpha) \end{pmatrix} \quad \text{and} \quad \mathbb{I} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}. \tag{3.26}$$

Since the matrices

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \tag{3.27}$$

commute, the exponential is given by

$$\begin{pmatrix} \exp(R) & \exp(R) & \exp(R)\frac{1}{2!} & \dots & \exp(R)\frac{1}{(n-1)!} \\ & \exp(R) & \exp(R) & \ddots & \vdots \\ & & \exp(R) & \ddots & \exp(R)\frac{1}{2!} \\ & & & \ddots & \exp(R) \\ & & & & \exp(R) \end{pmatrix}, \tag{3.28}$$

where

$$\exp(R) = \mathrm{e}^{\mathrm{Re}(\alpha)} \begin{pmatrix} \cos(\mathrm{Im}(\alpha)) & \sin(\mathrm{Im}(\alpha)) \\ -\sin(\mathrm{Im}(\alpha)) & \cos(\mathrm{Im}(\alpha)) \end{pmatrix}. \tag{3.29}$$

Problem 3.1. *Show that the space of n by n matrices $\mathbb{C}^{n\times n}$ together with the matrix norm is a Banach space. Show* (3.9)*.*

Problem 3.2. *Show that the matrix norm satisfies*

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

(This shows that $\mathbb{C}^{n\times n}$ *is even a* **Banach algebra***.) Conclude* $\|A^j\| \le \|A\|^j$*.*

Problem 3.3. *Show that the matrix product is continuous with respect to the matrix norm. That is, if* $A_j \to A$ *and* $B_j \to B$ *we have* $A_j B_j \to AB$*. (Hint: Problem 3.2).*

Problem 3.4. *Let* A_j *be a sequence in* $\mathbb{C}^{n\times n}$*. Show that*

$$\sum_{j=0}^{\infty} A_j$$

converges if $\sum_{j=0}^{\infty} \|A_j\|$ *does.*

Problem 3.5. *Is there a real matrix* A *such that*

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

(Hint: (3.21)*.)*

Problem 3.6. *The Stirling numbers of the first kind are define as the coefficients of the polynomials*

$$S_n(x) = x(x-1)\cdots(x-n+1) = \sum_{k=0}^{n} s(n,k) x^k.$$

and satisfy the basic recursion $s(n,k) = s(n-1,k-1) - (n-1)s(n-1,k)$*.*

Show the Stirling numbers satisfy the recursion from the proof of Lemma 3.3. (Hint: Insert the definition into $S_n(1+x) = (1+x)S_{n-1}(x)$*, apply the binomial theorem and compare coefficients. Finally use the basic recursion.)*

3.2. Linear autonomous first-order systems

In the previous section we have seen that the solution of the autonomous linear first-order system (3.1) is given by

$$x(t) = \exp(tA)x_0. \qquad (3.30)$$

In particular, the map $\exp(tA)$ provides an isomorphism between all initial conditions x_0 and all solutions. Hence the set of all solutions is a vector space isomorphic to $\mathbb{R}^n$ (respectively $\mathbb{C}^n$ if we allow complex initial values).

In order to understand the dynamics of the system (3.1), we need to understand the properties of the function $\exp(tA)$. We will start with the case of two dimensions which covers all prototypical cases. Furthermore, we will assume A as well as x_0 to be real-valued.

In this situation there are two eigenvalues, α_1 and α_2, which are either both real or otherwise complex conjugates of each other. We begin with the generic case where A is diagonalizable and hence there are two linearly independent eigenvectors, u_1 and u_2, which form the columns of U. In particular,

$$U^{-1}AU = \begin{pmatrix} \alpha_1 & 0 \\ 0 & \alpha_2 \end{pmatrix}. \tag{3.31}$$

Now using the change of coordinates

$$y(t) = U^{-1}x(t), \qquad y_0 = U^{-1}x_0, \tag{3.32}$$

the solution of the transformed equation

$$\dot{y} = (U^{-1}AU)y, \qquad y(0) = y_0, \tag{3.33}$$

is given by

$$y(t) = \exp(tU^{-1}AU)y_0 = \begin{pmatrix} \mathrm{e}^{\alpha_1 t} & 0 \\ 0 & \mathrm{e}^{\alpha_2 t} \end{pmatrix} y_0 \tag{3.34}$$

and the solution of our original equation (3.30) is given by

$$x(t) = U\exp(tU^{-1}AU)U^{-1}x_0 = U \begin{pmatrix} \mathrm{e}^{\alpha_1 t} & 0 \\ 0 & \mathrm{e}^{\alpha_2 t} \end{pmatrix} U^{-1}x_0. \tag{3.35}$$

Using $y_0 = U^{-1}x_0 = (y_{0,1}, y_{0,2})$ we obtain

$$x(t) = y_{0,1}\mathrm{e}^{\alpha_1 t}u_1 + y_{0,2}\mathrm{e}^{\alpha_2 t}u_2. \tag{3.36}$$

In the case where both eigenvalues are real, all quantities in (3.36) are real. Otherwise we have $\alpha_2 = \alpha_1^*$ and we can assume $u_2 = u_1^*$ without loss of generality. Let us write $\alpha_1 \equiv \alpha = \lambda + \mathrm{i}\omega$ and $\alpha_2 \equiv \alpha^* = \lambda - \mathrm{i}\omega$. Then **Euler's formula**

$$\mathrm{e}^{\mathrm{i}\omega} = \cos(\omega) + \mathrm{i}\sin(\omega) \tag{3.37}$$

implies

$$\mathrm{e}^{\alpha t} = \mathrm{e}^{\lambda t}\left(\cos(\omega t) + \mathrm{i}\sin(\omega t)\right), \qquad \alpha = \lambda + \mathrm{i}\omega. \tag{3.38}$$

Moreover, $x_0^* = x_0$ implies $y_{0,1}u_1 + y_{0,2}u_2 = y_{0,1}^* u_2 + y_{0,2}^* u_1$ which shows $y_{0,1}^* = y_{0,2}$. Hence, both terms in (3.36) are complex conjugates of each other implying

$$\begin{aligned} x(t) &= 2\mathrm{Re}(y_{0,1}\mathrm{e}^{\alpha_1 t}u_1) \\ &= 2\cos(\omega t)\mathrm{e}^{\lambda t}\mathrm{Re}(y_{0,1}u_1) - 2\sin(\omega t)\mathrm{e}^{\lambda t}\mathrm{Im}(y_{0,1}u_1). \end{aligned} \tag{3.39}$$

This finishes the case where A is diagonalizable.

If A is not diagonalizable, both eigenvalues must be equal $\alpha_1 = \alpha_2 \equiv \alpha$. The columns u_1 and u_2 of the matrix U are the eigenvector and generalized eigenvector of A, respectively. Hence

$$U^{-1}AU = \begin{pmatrix} \alpha & 1 \\ 0 & \alpha \end{pmatrix} \tag{3.40}$$

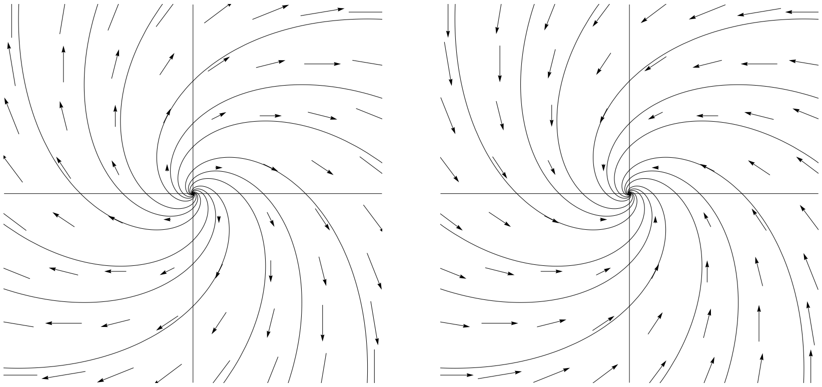


Figure 3.1. Phase portrait for a planar system where both eigenvalues have positive respectively negative real part.

and with a similar computation as before the solution is given by

$$x(t) = (y_{0,1} + y_{0,2}t)\mathrm{e}^{\alpha t}u_1 + y_{0,2}\mathrm{e}^{\alpha t}u_2. \tag{3.41}$$

This finishes the case where A is not diagonalizable.

Next, let us try to understand the qualitative behavior for large t. For this we need to understand the function $\exp(\alpha t)$. From (3.38) we can read off that $\exp(\alpha t)$ will converge to 0 as $t \to \infty$ if $\lambda = \mathrm{Re}(\alpha) < 0$ and grow exponentially if $\lambda = \mathrm{Re}(\alpha) > 0$. It remains to discuss the possible cases according to the respective signs of $\mathrm{Re}(\alpha_1)$ and $\mathrm{Re}(\alpha_2)$.

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. 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 solution curves are traversed in the opposite direction. The origin is called a **sink** in this case. The typical phase portrait is depicted in Figure 3.1 for the case of complex and in Figure 3.2 for the case of real eigenvalues. Note that in the case of real eigenvalues the two lines (plotted thick in the figures) correspond to the two eigenvectors of the coefficient matrix (why are there no eigenvectors visible in the case of complex eigenvalues?). In the complex case, the imaginary part ω causes a rotational component of the solutions and the origin is also called a spiral source respectively spiral sink.

If one eigenvalue is positive and one eigenvalue is negative, the phase portrait is shown in Figure 3.3 and the origin is called a **saddle**. Again the two lines correspond to the two eigenvectors of the coefficient matrix. The long-time behavior now depends on the initial condition x_0. If x_0 lies in the eigenspace corresponding to the negative eigenvalue, the solution will decay exponentially as $t \to \infty$ and grow exponentially as $t \to -\infty$. If x_0 lies in

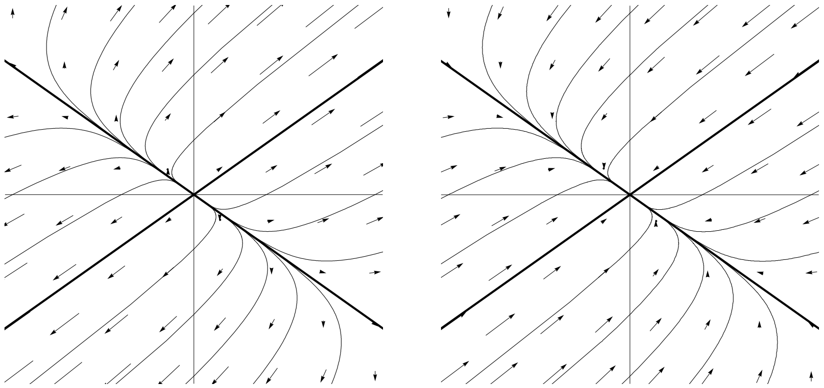


Figure 3.2. Phase portrait for a planar system where both eigenvalues are positive respectively negative.

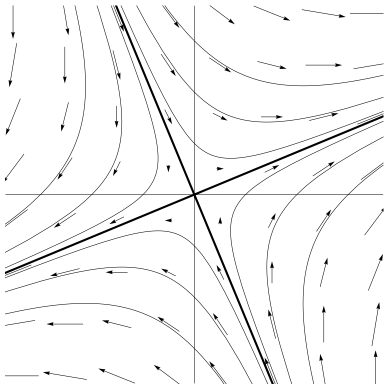


Figure 3.3. Phase portrait for a planar system with real eigenvalues of opposite sign.

the eigenspace corresponding to the positive eigenvalue, it is the other way round. If x_0 has components in both eigenspaces, it will grow exponentially as $t \to \pm\infty$.

If both eigenvalues are purely imaginary, the solutions will be periodic and encircle the origin. The phase portrait looks as in Figure 3.4 and the origin is called a **center**. All solutions are clearly bounded in this case.

In the case where the matrix is not diagonalizable, the phase portrait looks as in Figure 3.5. As before, the line corresponds to the eigenvector. If α is negative, all solutions will converge to 0, whereas if α is positive, all solutions will grow exponentially as $t \to \infty$. The polynomial term t does not play a role since it is dominated by the exponential term $\exp(\alpha t)$ unless $\alpha = 0$ (cf. Problem 3.7). If $\alpha = 0$ the solution is constant if we start in the subspace spanned by the eigenvector (i.e., $y_{0,2} = 0$ in (3.41)) and grows like t otherwise (i.e., $y_{0,2} \neq 0$).

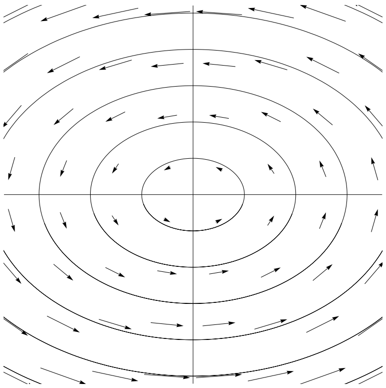


Figure 3.4. Phase portrait for a planar system with purely imaginary eigenvalues.

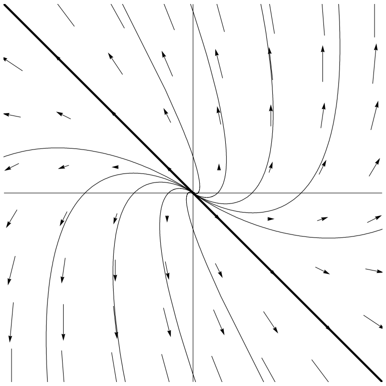


Figure 3.5. Phase portrait for a planar system with equal real eigenvalues (not diagonalizable).

Finally, we turn to the general case. As before, the considerations of the previous section show that it suffices to consider the case of one Jordan block

$$\exp(tJ) = \mathrm{e}^{\alpha t} \begin{pmatrix} 1 & t & \frac{t^2}{2!} & \dots & \frac{t^{n-1}}{(n-1)!} \\ & 1 & t & \ddots & \vdots \\ & & 1 & \ddots & \frac{t^2}{2!} \\ & & & \ddots & t \\ & & & & 1 \end{pmatrix}. \tag{3.42}$$

In particular, every solution is a linear combination of terms of the type $t^j \exp(\alpha t)$. Since $\exp(\alpha t)$ decays faster than any polynomial, our entire Jordan block converges to zero if $\lambda = \mathrm{Re}(\alpha) < 0$ (cf. Problem 3.7). If $\lambda = 0$, $\exp(\alpha t) = \exp(\mathrm{i}\omega t)$ will remain at least bounded, but the polynomial terms will diverge. However, if we start in the direction of the eigenvector $(1, 0, \dots, 0)$, we won't see the polynomial terms. In summary,

Theorem 3.4. *A solution of the linear system* (3.1) *converges to* 0 *as* $t \to \infty$ *if and only 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 as $t \to \infty$ *if and only 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 *positive*.

A linear system (not necessarily autonomous) is called **stable** if all solutions remain bounded as $t \to \infty$ and **asymptotically stable** if all solutions converge to 0 as $t \to \infty$.

Corollary 3.5. *The linear system* (3.1) *is stable if and only if all eigenvalues* α_j *of* A *satisfy* $\mathrm{Re}(\alpha_j) \le 0$ *and for all eigenvalues with* $\mathrm{Re}(\alpha_j) = 0$ *the corresponding algebraic and geometric multiplicities are equal. Moreover, in this case there is a constant* C *such that*

$$\|\exp(tA)\| \le C, \qquad t \ge 0. \tag{3.43}$$

In the case of an asymptotically stable matrix we can even specify the decay rate.

Corollary 3.6. *The linear system* (3.1) *is asymptotically stable if and only if all eigenvalues* α_j *of* A *satisfy* $\mathrm{Re}(\alpha_j) < 0$. *Moreover, in this case there is a constant* $C = C(\alpha)$ *for every* $\alpha < \min\{-\mathrm{Re}(\alpha_j)\}_{j=1}^m$ *such that*

$$\|\exp(tA)\| \le C\mathrm{e}^{-t\alpha}, \qquad t \ge 0. \tag{3.44}$$

Proof. It remains to prove the second claim. 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 note that $\|\exp(tA)\| = \mathrm{e}^{-t\alpha}\|\exp(t(A + \alpha\mathbb{I}))\|$. Since $\mathrm{Re}(\alpha_j + \alpha) < 0$ all entries of the matrix $\exp(t(A + \alpha\mathbb{I}))$ are bounded and consequently $\|\exp(t(A + \alpha\mathbb{I}))\| \le C$ is bounded (cf. Problem 3.7) as required. □

Note that one can choose $\alpha = \min\{-\mathrm{Re}(\alpha_j)\}_{j=1}^m$ if and only if for all eigenvalues α_j with $-\mathrm{Re}(\alpha_j) = \alpha$ the corresponding algebraic and geometric multiplicities are equal.

A matrix all whose eigenvalues satisfy $\mathrm{Re}(\alpha_j) < 0$ is also known as a **Hurwitz matrix**. The **Routh-Hurwitz criterion** (cf. [**9**, Sect. V.6]) states that a real matrix is Hurwitz if and only if the following determinants are strictly positive,

$$\det\begin{pmatrix} a_1 & 1 & 0 & 0 & 0 & 0 & \cdots & 0 \\ a_3 & a_2 & a_1 & 1 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{2k-1} & a_{2k-2} & a_{2k-3} & a_{2k-4} & a_{2k-5} & a_{2k-6} & \cdots & a_k \end{pmatrix} > 0, \qquad (3.45)$$

for $1 \le k \le n$. Here the numbers a_j are the coefficients of the characteristic polynomial of A,

$$\det(z\mathbb{I} - A) = z^n + a_1 z^{n-1} + \cdots + a_{n-1} z + a_n, \qquad (3.46)$$

and $a_j = 0$ for $j \ge n$.

Finally, observe that the solution of the inhomogeneous equation

$$\dot{x}(t) = Ax(t) + g(t), \qquad x(0) = x_0, \qquad (3.47)$$

is given by

$$x(t) = \exp(tA)x_0 + \int_0^t \exp((t-s)A)g(s)ds, \qquad (3.48)$$

which can be verified by a straightforward computation (however, we will in fact prove a more general result in Theorem 3.12 below). This formula is sometimes called **Duhamel's formula**. As always for linear equations, note that the solution consists of the general solution of the homogeneous equation plus a particular solution of the inhomogeneous equation. However, if the inhomogeneous term is of a special form, an ansatz might be faster than evaluating the integral in (3.48) — see Problem 3.13.

Problem 3.7. *Show*

$$\lim_{t\to\infty} t^m e^{\alpha t} = 0, \qquad m \in \mathbb{N}_0,\ \mathrm{Re}(\alpha) < 0,$$

and

$$\max_{0\le t<\infty} |t^m e^{\alpha t}| = \left(\frac{m}{-\mathrm{Re}(\alpha)}\right)^m e^{-m}, \qquad m \in \mathbb{N}_0,\ \mathrm{Re}(\alpha) < 0.$$

(Hint: l'Hôpital's rule.)

Problem 3.8. *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 3.9. *Solve the systems corresponding to the following matrices:*

$(i).\ A = \begin{pmatrix} 2 & 1 \\ 0 & 2 \end{pmatrix}, \quad x_0 = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \qquad (ii).\ A = \begin{pmatrix} -1 & 1 \\ 0 & 1 \end{pmatrix}, \quad x_0 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}.$

Problem 3.10. *Solve*

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

Problem 3.11. *Find a two by two matrix such that* $x(t) = (\sinh(t), \mathrm{e}^t)$ *is a solution.*

Problem 3.12. *Which of the following functions*

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

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

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

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

(v) $x(t) = (\mathrm{e}^t + 2\mathrm{e}^{-t}, \mathrm{e}^t + 2\mathrm{e}^{-t})$

can be solutions of a first-order autonomous homogeneous system? (Hint: Compare with the necessary structure of the solution found in this section.)

Problem 3.13. *Let A be an n by n matrix and β a constant. Consider the special inhomogeneous equation*

$$\dot{x} = Ax + p(t)\mathrm{e}^{\beta t},$$

where $p(t)$ is a vector all whose entries are polynomials. Set $\deg(p(t)) = \max_{1\le j\le n} \deg(p_j(t))$. *Show that this equation has a particular solution of the form*

$$q(t)\mathrm{e}^{\beta t},$$

where $q(t)$ is a polynomial vector with $\deg(q(t)) = \deg(p(t))$ *if β is not an eigenvalue of A and* $\deg(q(t)) = \deg(p(t)) + a$ *if β is an eigenvalue of algebraic multiplicity a.*

(Hint: Investigate (3.48) *using the following fact:* $\int p(t)\mathrm{e}^{\beta t} dt = q(t)\mathrm{e}^{\beta t}$, *where $q(t)$ is a polynomial of degree* $\deg(q) = \deg(p)$ *if $\beta \neq 0$ and* $\deg(q) = \deg(p) + 1$ *if $\beta = 0$.)*

Problem 3.14. *Let A be a real 2 by 2 matrix. Then the eigenvalues can be expressed in terms of the determinant $D = \det(A)$ and the trace $T = \mathrm{tr}(A)$. In particular, (T, D) can take all possible values in $\mathbb{R}^2$ if A ranges over all possible matrices in $\mathbb{R}^{2\times 2}$. Split the (T, D) plane into regions in which the various cases discussed in this section occur (source, spiral source, sink, spiral sink, saddle, center).*

Problem 3.15 (Laplace transform). *Let $x : [0, \infty) \to \mathbb{C}^n$ such that $|x(t)| \le M\mathrm{e}^{at}$ for some constants $M \ge 0$ and $a \in \mathbb{R}$. Then the* **Laplace transform**

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

exists and is analytic for $\mathrm{Re}(s) > a$. *Show that for* $x \in C^1([0,\infty))$ *satisfying* $|x(t)| + |\dot{x}(t)| \le M\mathrm{e}^{at}$ *we have*

$$\mathcal{L}(\dot{x})(s) = s\mathcal{L}(x)(s) - x(0)$$

for $\mathrm{Re}(s) > a$. *Moreover, 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.

Problem 3.16. *Suppose all eigenvalues of* A *satisfy* $\mathrm{Re}(\alpha_j) < 0$. *Show that every solution of* (3.47) *satisfies*

$$\lim_{t\to\infty} x(t) = 0.$$

if $\lim_{t\to\infty} |g(t)| = 0$ *(Hint:* (3.48).*) What if* $\lim_{t\to\infty} g(t) = g_0$*?*

3.3. Linear autonomous equations of order n

In this section, we want to have a brief look at the case of the n'th order equations

$$x^{(n)} + c_{n-1}x^{(n-1)} + \cdots + c_1\dot{x} + c_0 x = 0, \tag{3.49}$$

which appear frequently in applications. Here $c_0, \dots, c_{n-1}$ are some real (or complex) constants. Again the solutions form an n dimensional vector space since a solution is uniquely determined by the initial conditions

$$x(0) = x_0, \quad \dots, \quad x^{(n-1)}(0) = x_{n-1}. \tag{3.50}$$

The corresponding system is given by

$$A = \begin{pmatrix} 0 & 1 & & & \\ & 0 & 1 & & \\ & & \ddots & \ddots & \\ & & & 0 & 1 \\ -c_0 & -c_1 & \cdots & \cdots & -c_{n-1} \end{pmatrix} \tag{3.51}$$

and hence all our considerations apply: The characteristic polynomial can be computed by performing the Laplace expansion with respect to the last row and is given by

$$\chi_A(z) = \det(z\mathbb{I} - A) = z^n + c_{n-1}z^{n-1} + \cdots + c_1 z + c_0. \tag{3.52}$$

One can show that the geometric multiplicity of every eigenvalue is one (Problem 3.24).

Theorem 3.7. *Let* α_j, $1 \le j \le m$, *be the zeros of the characteristic polynomial*

$$z^n + c_{n-1}z^{n-1} + \cdots + c_1 z + c_0 = \prod_{j=1}^m (z - \alpha_j)^{a_j} \tag{3.53}$$

associated with (3.49) *and let* a_j *be the corresponding multiplicities. Then the functions*

$$x_{j,k}(t) = t^k \exp(\alpha_j t), \qquad 0 \le k < a_j, \quad 1 \le j \le m, \tag{3.54}$$

are n *linearly independent solutions of* (3.49).

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

Proof. Let us look at a solution of the corresponding first-order system. By construction, the first component of every solution of the system will solve our n'th order equation. By collecting functions from each Jordan block (3.42), this first component must be a linear combination of the functions $x_{j,k}(t)$. So the solution space of (3.49) is spanned by these functions. Since this space is n dimensional, all functions must be present. In particular, these functions must be linearly independent. □

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

$$t^k \mathrm{e}^{\lambda_j t} \cos(\omega_j t), \qquad t^k \mathrm{e}^{\lambda_j t} \sin(\omega_j t). \tag{3.55}$$

Example. Consider the differential equation

$$\ddot{x} + \omega_0^2 x = 0, \quad \omega_0 \ge 0.$$

The characteristic polynomial is $\alpha^2 + \omega_0^2 = 0$ and the zeros are $\alpha_1 = \mathrm{i}\omega_0$, $\alpha_2 = -\mathrm{i}\omega_0$. Hence for $\omega_0 > 0$ a basis of solutions is

$$x_1(t) = \mathrm{e}^{\mathrm{i}\omega_0 t}, \qquad x_2(t) = \mathrm{e}^{-\mathrm{i}\omega_0 t}$$

or, if we want real solutions,

$$x_1(t) = \cos(\omega_0 t), \qquad x_2(t) = \sin(\omega_0 t).$$

For $\omega_0 = 0$ we have only one zero $\alpha_1 = 0$ of multiplicity $a_1 = 2$ and a basis of solutions is given by

$$x_{1,0}(t) = 1, \qquad x_{1,1}(t) = t.$$

◇

By (3.48) the solution of the inhomogeneous equation

$$x^{(n)} + c_{n-1} x^{(n-1)} + \cdots + c_1 \dot{x} + c_0 x = g(t) \tag{3.56}$$

is given by

$$x(t) = x_h(t) + \int_0^t u(t-s) g(s) ds, \tag{3.57}$$

where $x_h(t)$ is an arbitrary solution of the homogeneous equation and $u(t)$ is the solution of the homogeneous equation corresponding to the initial condition $u(0) = \dot{u}(0) = \cdots = u^{(n-2)}(0) = 0$ and $u^{(n-1)}(0) = 1$ (Problem 3.21).

Hence the algorithm for solving a linear n'th order equation with constant coefficients is as follows: Start with the homogeneous equation, compute the zeros of the characteristic polynomial and write down the general solution as a linear combination of the fundamental solutions (3.54). Find a particular solution of the inhomogeneous equation and determine the unknown constants of the homogeneous equation from the initial conditions. The particular solution of the inhomogeneous equation can be found by evaluating the integral in (3.57). However, in many situations it is more efficient to make a suitable ansatz for the solution (Problem 3.22):

Lemma 3.8. *Consider the inhomogeneous equation* (3.56) *with right-hand side of the form* $g(t) = p(t)\mathrm{e}^{\beta t}$*, where* $p(t)$ *is a polynomial. Then there is a particular solution of the same form* $x_p(t) = q(t)\mathrm{e}^{\beta t}$*, where* $q(t)$ *is a polynomial which satisfies* $\deg(q) = \deg(p)$ *if* $\beta \notin \{\alpha_j\}_{j=1}^m$ *is not equal to any of the characteristic eigenvalues and* $\deg(q) = \deg(p) + a_j$ *if* $\beta = \alpha_j$ *is equal to one of the characteristic eigenvalues whose algebraic multiplicity is* a_j*.*

Note that in the case $\beta = \alpha_j$ you can assume the first a_j coefficients of q to be zero, since they correspond to a homogenous solution. Moreover, if you allow complex values for $\beta = \lambda + \mathrm{i}\omega$, this also includes the case where $g(t) = p(t)\mathrm{e}^{\lambda t}\cos(\omega t)$ or $g(t) = p(t)\mathrm{e}^{\lambda t}\sin(\omega t)$ after taking real and imaginary parts. Finally, the case of linear combinations of such terms comes for free by linearity.

Of special importance is the case of second order, which appears in a vast number of applications. For example when modeling electrical circuits:

Example. 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 a directed 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).

In a simple circuit one has three types of different elements, inductors, capacitors, and resistors. For an inductor we have

$$L\dot{I}_L = V_L, \tag{3.58}$$

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

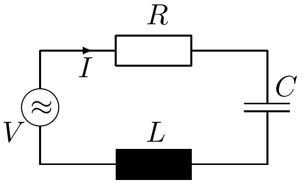
$$C\dot{V}_C = I_C, \tag{3.59}$$

where $C > 0$ is the capacitance, $I_C(t)$ is the current through the capacitor and $V_C(t)$ is the voltage difference. For a resistor we have (Ohm's law)

$$V_R = R\, I_R, \tag{3.60}$$

where $R > 0$ is the resistance, $I_R(t)$ is the current through the resistor and $V_R(t)$ is the voltage difference.

We will look at the case of one inductor L, one capacitor C, and one resistor R arranged in a loop together with an external power source V (the classical **RLC circuit**).



Kirchhoff's laws yield $I_R = I_L = I_C$ and $V_R + V_L + V_C = V$. Using the properties of our three elements we arrive at the second-order linear differential equation

$$L\ddot{I}(t) + R\dot{I}(t) + \frac{1}{C}I(t) = \dot{V}(t) \tag{3.61}$$

for the current I. Let us try to solve this equation for an external sinusoidal voltage

$$V(t) = V_0\cos(\omega t). \tag{3.62}$$

It turns out convenient to use the complex voltage $V(t) = V_0\mathrm{e}^{\mathrm{i}\omega t}$:

$$\ddot{I} + \frac{R}{L}\dot{I} + \frac{1}{LC}I = \mathrm{i}\frac{\omega V_0}{L}\mathrm{e}^{\mathrm{i}\omega t}. \tag{3.63}$$

We get the solutions for $V(t) = V_0\cos(\omega t)$ and $V(t) = V_0\sin(\omega t)$ by taking real and imaginary part of the complex solution, respectively.

The eigenvalues are

$$\alpha_{1,2} = -\eta \pm \sqrt{\eta^2 - \omega_0^2}, \tag{3.64}$$

where we have introduced the convenient abbreviations

$$\eta = \frac{R}{2L} \quad \text{and} \quad \omega_0 = \frac{1}{\sqrt{LC}}. \tag{3.65}$$

If $\eta > \omega_0$ (**over damping**), both eigenvalues are negative and the solution of the homogeneous equation is given by

$$I_h(t) = k_1 \mathrm{e}^{\alpha_1 t} + k_2 \mathrm{e}^{\alpha_2 t}. \tag{3.66}$$

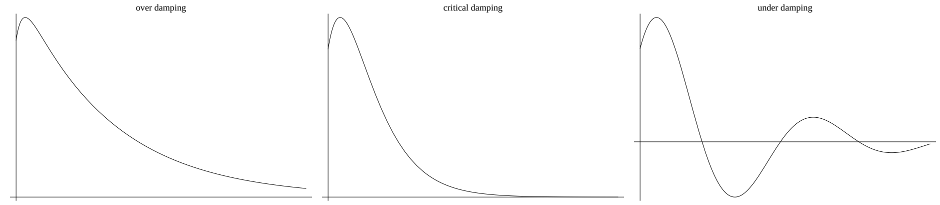
If $\eta = \omega_0$ (**critical damping**), both eigenvalues are equal and the solution of the homogeneous equation is given by

$$I_h(t) = (k_1 + k_2 t)\mathrm{e}^{-\eta t}. \tag{3.67}$$

Finally, for $\eta < \omega_0$ (**under damping**) we have complex conjugate eigenvalues and the solution of the homogeneous equation is given by

$$I_h(t) = k_1 \mathrm{e}^{-\eta t}\cos(\beta t) + k_2 \mathrm{e}^{-\eta t}\sin(\beta t), \qquad \beta = \sqrt{\omega_0^2 - \eta^2} > 0. \tag{3.68}$$

In every case the real part of both eigenvalues is negative and the homogeneous solution decays exponentially as $t \to \infty$:



Observe that for fixed $\eta > 0$, the choice $\omega_0 = \eta$ gives that fastest decay without an oscillatory component.

For the inhomogeneous solution we make the ansatz

$$I_i(t) = k\,\mathrm{e}^{\mathrm{i}\omega t} \tag{3.69}$$

with an unknown constant k. This produces

$$k = \frac{V_0}{R + \mathrm{i}(L\omega - \frac{1}{\omega C})}. \tag{3.70}$$

Since the homogeneous solution decays exponentially, we have after a short time

$$I(t) = I_h(t) + I_i(t) \approx I_i(t) = \frac{V_0}{Z}\mathrm{e}^{\mathrm{i}\omega t} = \frac{1}{Z}V(t), \tag{3.71}$$

where

$$Z = R + Z_L + Z_C, \qquad Z_L = \mathrm{i}L\omega, \quad Z_C = -\mathrm{i}\frac{1}{\omega C} \tag{3.72}$$

is known as the complex impedance. The current $I(t) = \frac{1}{Z}V(t)$ attains its maximum when

$$|Z|^2 = R^2 + (L\omega - \frac{1}{\omega C})^2 \tag{3.73}$$

gets minimal, that is, if $L\omega - \frac{1}{\omega C} = 0$ and hence

$$\omega = \omega_0 = \frac{1}{\sqrt{LC}}. \tag{3.74}$$

The frequency $\frac{\omega_0}{2\pi}$ is called the resonance frequency of the circuit.

By changing one of the parameters, say C, you can tune the circuit to a specific resonance frequency. This idea is for example used to filter your favorite radio station out of many other available ones. In this case the external power source corresponds to the signal picked up by your antenna and the RLC circuit starts only oscillating if the carrying frequency of your radio station matches its resonance frequency. ◇

Furthermore, our example is not only limited to electrical circuits. Many other systems can be described by the differential equation

$$\ddot{x} + 2\eta\,\dot{x} + \omega_0^2 x = 0, \qquad \eta, \omega_0 > 0, \tag{3.75}$$

at least for small amplitudes $x(t)$. Here $\frac{\omega_0}{2\pi}$ is the **resonance frequency** of the system and η is the damping factor. If you add a periodic **forcing** term,

$$\ddot{x} + 2\eta\,\dot{x} + \omega_0^2 x = \cos(\omega t), \tag{3.76}$$

you will get a maximal effect if the forcing is resonant, that is, ω coincides with ω_0. If $\eta = 0$, the solution corresponds to a free (undamped) oscillation $x(t) = k_1 \cos(\omega_0 t) + k_2 \sin(\omega_0 t)$ and a resonant forcing will result in a solution whose amplitude tends to ∞ (cf. Problem 3.18).

Problem 3.17. *Solve the following differential equations:*

(i) $\ddot{x} + 3\dot{x} + 2x = \sinh(t)$.

(ii) $\ddot{x} + 2\dot{x} + 2x = \exp(t)$.

(iii) $\ddot{x} + 2\dot{x} + x = t^2$.

Problem 3.18 (Resonance catastrophe)**.** *Solve the equation*

$$\ddot{x} + \omega_0^2 x = \cos(\omega t), \quad \omega_0, \omega > 0.$$

Discuss the behavior of solutions as $t \to \infty$. The inhomogeneous term is also known as a forcing term. It is **resonant** *if $\omega = \omega_0$. What happens in this case?*

Problem 3.19 (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 $\tau = \log(t)$. Discuss the possible solutions for $c_0, c_1 \in \mathbb{R}$.

Problem 3.20. *Find a formula for the Wronskian $W(x, y) = x\dot{y} - \dot{x}y$ of two solutions of the second-order autonomous equation*

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

Problem 3.21. *Prove* (3.57) *(either by reducing it to* (3.48) *or by a direct verification – I recommend doing both;-)*

Problem 3.22. *Look at the second-order autonomous equation*

$$\ddot{x} + c_1\dot{x} + c_0 x = g(t)$$

and let α_1, α_2 be the corresponding eigenvalues (not necessarily distinct). Show that the equation can be factorized as

$$\ddot{x} + c_1\dot{x} + c_0 x = \left(\frac{d}{dt} - \alpha_2\right)\left(\frac{d}{dt} - \alpha_1\right)x.$$

Hence the equation can be reduced to solving two first order equations

$$\left(\frac{d}{dt} - \alpha_2\right)y = g(t), \qquad \left(\frac{d}{dt} - \alpha_1\right)x = y.$$

Use this to prove Theorem 3.7 as well as Lemma 3.8 in the case $n = 2$. Extend this to the general case $n \in \mathbb{N}$. (Hint: The solution for the first order case is given in (3.48)*. Moreover, $\int p(t)\mathrm{e}^{\beta t}dt = q(t)\mathrm{e}^{\beta t}$, where $q(t)$ is a polynomial of degree $\deg(q) = \deg(p)$ if $\beta \neq 0$ and $\deg(q) = \deg(p) + 1$ if $\beta = 0$. For the general case use induction.)*

Problem 3.23. *Derive Taylor's formula with remainder*

$$x(t) = \sum_{j=0}^{n} \frac{x^{(j)}(t_0)}{j!}(t - t_0)^j + \frac{1}{n!}\int_{t_0}^{t} x^{(n+1)}(s)(t - s)^n ds$$

for $x \in C^{n+1}$ from (3.57)*.*

Problem 3.24. *Show that the geometric multiplicity of every eigenvalue of the matrix A from* (3.51) *is one. (Hint: Can you find a cyclic vector? Why does this help you?)*

3.4. General linear first-order systems

In this section we want to consider the case of linear systems, where the coefficient matrix can depend on t. As a preparation 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

$$\frac{d}{dt}A(t)B(t) = \dot{A}(t)B(t) + A(t)\dot{B}(t) \qquad (3.77)$$

and, if $\det(A(t)) \neq 0$,

$$\frac{d}{dt} A(t)^{-1} = -A(t)^{-1} \dot{A}(t) A(t)^{-1} \tag{3.78}$$

(Problem 3.25). Note that the order is important!

Given vectors $a_1, \dots, a_n$ we will write $A = (a_1, \dots, a_n)$ for the matrix which has these vectors as rows. Observe that BA is the matrix whose rows are $Ba_1, \dots, Ba_n$, that is, $BA = (Ba_1, \dots, Ba_n)$. Again note that the order is important here.

We now turn to the general linear first-order system

$$\dot{x}(t) = A(t)x(t), \tag{3.79}$$

where $A \in C(I, \mathbb{R}^{n\times n})$. Clearly our theory from Section 2.2 applies:

Theorem 3.9. *The linear first-order system* (3.79) *has a unique solution satisfying the initial condition $x(t_0) = x_0$. Moreover, this solution is defined for all $t \in I$.*

Proof. This follows directly from Theorem 2.17 (or alternatively from Corollary 2.6) since we can choose $L(T) = \max_{[0,T]} \|A(t)\|$ for every $T \in I$. □

It seems tempting to suspect that the solution is given by the formula $x(t) = \exp(\int_{t_0}^t A(s)ds)x_0$. However, as soon as you try to verify this guess, noncommutativity of matrices will get into your way. In fact, this formula only solves our initial value problem if $[A(t), A(s)] = 0$ for all $t, s \in \mathbb{R}$. Hence it is of little use. So we still need to find the right generalization of $\exp((t - t_0)A)$.

We start by observing 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

$$\phi(t, t_0, x_0) = \sum_{j=1}^{n} \phi(t, t_0, \delta_j) x_{0,j}, \tag{3.80}$$

where δ_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

$$\Pi(t, t_0) = (\phi(t, t_0, \delta_1), \dots, \phi(t, t_0, \delta_n)), \tag{3.81}$$

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

$$\phi(t, t_0, x_0) = \Pi(t, t_0) x_0. \tag{3.82}$$

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

$$\dot{\Pi}(t,t_0) = A(t)\Pi(t,t_0), \qquad \Pi(t_0,t_0) = \mathbb{I}. \tag{3.83}$$

Again observe that our basic existence and uniqueness result applies. In fact, it is easy to check, that a matrix $X(t)$ satisfies $\dot{X} = A(t)X$ if and only if every column satisfies (3.79). In particular, $X(t)c$ solves (3.79) for every constant vector c in this case. In summary,

Theorem 3.10. *The solutions of the system* (3.79) *form an n dimensional vector space. Moreover, there exists a matrix-valued solution $\Pi(t,t_0)$ such that the solution satisfying the initial condition $x(t_0) = x_0$ is given by $\Pi(t,t_0)x_0$.*

Example. In the simplest case, where $A(t) \equiv A$ is constant, we of course have $\Pi(t,t_0) = \mathrm{e}^{(t-t_0)A}$. ⋄

Example. Consider the system

$$\dot{x} = \begin{pmatrix} 1 & t \\ 0 & 2 \end{pmatrix} x, \tag{3.84}$$

which explicitly reads

$$\dot{x}_1 = x_1 + t\, x_2, \qquad \dot{x}_2 = 2x_2. \tag{3.85}$$

We need to find the solution corresponding to the initial conditions $x(t_0) = \delta_1 = (1,0)$ respectively $x(t_0) = \delta_2 = (0,1)$. In the first case $x(t_0) = \delta_1$, the second equation gives $x_2(t) = 0$ and plugging this into the first equation shows $x_1(t) = \mathrm{e}^{t-t_0}$, that is, $\phi(t,t_0,\delta_1) = (\mathrm{e}^{t-t_0}, 0)$. Similarly, in the second case $x(t_0) = (0,1)$, the second equation gives $x_2(t) = \mathrm{e}^{2(t-t_0)}$ and plugging this into the first equation shows $x_1(t) = \mathrm{e}^{2(t-t_0)}(t-1) - \mathrm{e}^{t-t_0}(t_0-1)$, that is, $\phi(t,t_0,\delta_2) = (\mathrm{e}^{2(t-t_0)}(t-1) - \mathrm{e}^{t-t_0}(t_0-1), \mathrm{e}^{2(t-t_0)})$. Putting everything together we obtain

$$\Pi(t,t_0) = \begin{pmatrix} \mathrm{e}^{t-t_0} & \mathrm{e}^{2(t-t_0)}(t-1) - \mathrm{e}^{t-t_0}(t_0-1) \\ 0 & \mathrm{e}^{2(t-t_0)} \end{pmatrix}. \tag{3.86}$$

⋄

Note that using Gronwall's inequality (cf. Problem 2.14) one can get a rough estimate on the norm of the principal matrix solution

$$\|\Pi(t,t_0)\| \le \mathrm{e}^{|\int_{t_0}^t \|A(s)\| ds|}. \tag{3.87}$$

A better estimate is derived in Problem 3.31.

Furthermore, $\Pi(t,t_0)$ satisfies

$$\Pi(t,t_1)\Pi(t_1,t_0) = \Pi(t,t_0) \tag{3.88}$$

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

$$\Pi(t, t_0)^{-1} = \Pi(t_0, t). \tag{3.89}$$

More generally, taking n solutions $\phi_1, \dots, \phi_n$ we obtain a matrix solution $U(t) = (\phi_1(t), \dots, \phi_n(t))$. Note that the differential equation is uniquely determined by n linearly independent solutions by virtue of $A(t) = \dot{U}(t)U(t)^{-1}$.

The determinant of $U(t)$ is called **Wronski determinant**

$$W(t) = \det(\phi_1(t), \dots, \phi_n(t)). \tag{3.90}$$

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 3.11. *The Wronski determinant of n solutions satisfies*

$$W(t) = W(t_0) \exp\left(\int_{t_0}^{t} \operatorname{tr}(A(s))\, ds \right). \tag{3.91}$$

This is known as **Abel's identity** *or* **Liouville's formula**.

Proof. By (3.83) we have

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

and using $U(t + \varepsilon) = \Pi(t + \varepsilon, t)U(t)$ we obtain (Problem 3.26)

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

implying

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

This equation is separable and the solution is given by (3.91). □

Now let us turn to the inhomogeneous system

$$\dot{x} = A(t)x + g(t), \tag{3.92}$$

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 (3.92) satisfies the corresponding homogeneous system (3.79), it suffices to find one particular solution. This can be done using the following ansatz

$$x(t) = \Pi(t, t_0)c(t), \qquad c(t_0) = x(t_0) = x_0, \tag{3.93}$$

which is known as **variation of constants** (also **variation of parameters**). Differentiating this ansatz we see

$$\dot{x}(t) = A(t)x(t) + \Pi(t, t_0)\dot{c}(t) \tag{3.94}$$

and comparison with (3.92) yields

$$\dot{c}(t) = \Pi(t_0, t)g(t). \tag{3.95}$$

Integrating this equation shows

$$c(t) = x_0 + \int_{t_0}^{t} \Pi(t_0, s)g(s)ds \tag{3.96}$$

and we obtain (using (3.88)):

Theorem 3.12. *The solution of the inhomogeneous system* (3.92) *corresponding to the initial condition* $x(t_0) = x_0$ *is given by*

$$x(t) = \Pi(t, t_0)x_0 + \int_{t_0}^{t} \Pi(t, s)g(s)ds, \tag{3.97}$$

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

To end this section, let me emphasize that there is no general way of solving linear systems except for the trivial case $n = 1$ (recall (1.40)). However, if one solution $\phi_1(t)$ is known, one can use the following method known as **reduction of order** (d'Alembert): At least one component of $\phi_1(t)$ is nonzero, say the first one for notational simplicity. Let $X(t)$ be the identity matrix with the first row replaced by $\phi_1(t)$,

$$X(t) = (\phi_1(t), \delta_2, \dots, \delta_n) \tag{3.98}$$

and consider the transformation

$$x(t) = X(t)y(t). \tag{3.99}$$

Then the differential equation for $y(t) = X(t)^{-1}x(t)$ reads

$$\dot{y} = X^{-1}\dot{x} - X^{-1}\dot{X}X^{-1}x = X^{-1}(AX - \dot{X})y \tag{3.100}$$

with

$$AX - \dot{X} = AX - (\dot{\phi}_1, 0, \dots, 0) = A(X - (\phi_1, 0, \dots, 0)) = A(0, \delta_2, \dots, \delta_n). \tag{3.101}$$

In particular, the right-hand side of the resulting system does not contain y_1. Hence we can first solve the $n - 1$ by $n - 1$ system for $(y_2, \dots, y_n)$ and finally determine y_1 by one additional integration.

Example. Consider the system

$$A(t) = \begin{pmatrix} t^2 & -1 \\ 2t & 0 \end{pmatrix}$$

and note that $\phi_1(t) = (1, t^2)$ is a solution. Hence we can make the change of coordinates

$$x(t) = X(t)y(t), \qquad \text{where} \qquad X(t) = \begin{pmatrix} 1 & 0 \\ t^2 & 1 \end{pmatrix}$$

in which the differential equation reads

$$\dot{y} = X(t)^{-1}A(t)\begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} y = \begin{pmatrix} 0 & -1 \\ 0 & t^2 \end{pmatrix} y.$$

In particular, the right-hand side does not involve y_1. Hence this system can be solved by first solving the second component $\dot{y}_2 = t^2 y_2$ which gives

$$y_2(t) = \mathrm{e}^{t^3/3}.$$

Now integrating the first component $\dot{y}_1 = -y_2$ gives

$$y_1(t) = -\int \mathrm{e}^{t^3/3} dt$$

and thus a second solution is given by

$$\phi_2(t) = \begin{pmatrix} 1 & 0 \\ t^2 & 1 \end{pmatrix} \begin{pmatrix} -\int \mathrm{e}^{t^3/3} dt \\ \mathrm{e}^{t^3/3} \end{pmatrix} = \begin{pmatrix} -\int \mathrm{e}^{t^3/3} dt \\ \mathrm{e}^{t^3/3} - t^2 \int \mathrm{e}^{t^3/3} dt \end{pmatrix}.$$

$\diamond$

Problem 3.25 (Differential calculus for matrices.). *Suppose $A(t)$ and $B(t)$ are differentiable. Prove* (3.77) *and* (3.78). *(Hint: $AA^{-1} = \mathbb{I}$.)*

Problem 3.26. *Show that for any n by n matrix A we have*

$$\det(\mathbb{I} + \varepsilon\, A + o(\varepsilon)) = 1 + \varepsilon\, \mathrm{tr}(A) + o(\varepsilon),$$

where $o(\varepsilon)$ (Landau symbol) collects terms which vanish faster than ε as $\varepsilon \to 0$. (Hint: E.g. Jordan canonical form.)

Problem 3.27. *Compute $\Pi(t, t_0)$ for the system*

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

Problem 3.28. *Compute $\Pi(t, t_0)$ for the system*

$$A(t) = \begin{pmatrix} 2 + 2t & 3 + 2t \\ -1 - 2t & -2 - 2t \end{pmatrix}.$$

(Hint: $\phi_1(t) = \mathrm{e}^{-t}(1, -1)$ is a solution.)

Problem 3.29 (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 $|\psi_j(t)|^2$ is 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, $|\psi(t)| = 1$. 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 A^* is the adjoint (complex conjugate of the transposed) matrix.) The matrix $H(t)$ is called the Hamiltonian and describes the interaction. Show that the solution is given by

$$\psi(t) = U(t, t_0)\psi_0, \qquad U(t_0, t_0) = \mathbb{I},$$

where $U(t, t_0)$ is unitary, that is, $U(t, t_0)^{-1} = U(t, t_0)^*$ (Hint: Problem 3.25). Conclude that $\psi(t)$ remains normalized for all t if ψ_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 $\langle\varphi, \psi\rangle = \varphi^* \cdot \psi$ is the scalar product in $\mathbb{C}^n$. Show that

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

where $[H, L] = HL - LH$ is the commutator.

Problem 3.30. *Show that if* $\liminf_{t\to\infty} \int_{t_0}^t \mathrm{tr}(A(s))ds = \infty$*, then* (3.79) *has an unbounded solution. (Hint:* (3.91)*.)*

Problem 3.31. *For any matrix A, the matrix* $\mathrm{Re}(A) = \frac{1}{2}(A + A^*)$ *is symmetric and hence has only real eigenvalues (cf. Theorem 3.29). Let α_0 be its largest eigenvalue.*

Let $A(t)$ be given and define $\alpha_0(t)$ as above. Show that

$$\|\Pi(t, t_0)\| \le \exp\left(\int_{t_0}^t \alpha_0(s)ds\right), \qquad t \ge t_0.$$

A similar formula holds for $t \le t_0$ if we take the lowest eigenvalue. (Hint: Compute $\frac{d}{dt}|x(t)|^2$ for $x(t) = \Pi(t, t_0)x_0$ and note that $\langle x, \mathrm{Re}(A)x\rangle \le \alpha_0|x|^2$ for every $x \in \mathbb{R}^n$.)

Remark: If $A(t) \equiv A$ is constant, we know that one can do much better and replace α_0 by the real part of the largest eigenvalue of A plus an arbitrarily small ε (the ε is necessary to cover possible polynomial terms) – cf. also Corollary 3.6. Hence one might conjecture that the same is true in the general case. However, this is not the case as Problem 3.40 below shows.

3.5. Linear equations of order n

In this section, we want to have a brief look at the case of the n'th order equations

$$x^{(n)} + q_{n-1}(t)x^{(n-1)} + \cdots + q_1(t)\dot{x} + q_0(t)x = 0, \tag{3.102}$$

where $q_0(t), \dots, q_{n-1}(t)$ are some continuous functions. Again the solutions form an n dimensional vector space since a solution is uniquely determined by the initial conditions

$$x(t_0) = x_0, \quad \dots, \quad x^{(n-1)}(t_0) = x_{n-1} \tag{3.103}$$

and, as in the case of constant coefficients (cf. Section 3.3), the corresponding system is given by

$$A(t) = \begin{pmatrix} 0 & 1 & & & \\ & 0 & 1 & & \\ & & \ddots & \ddots & \\ & & & 0 & 1 \\ -q_0(t) & -q_1(t) & \cdots & \cdots & -q_{n-1}(t) \end{pmatrix}. \tag{3.104}$$

If we denote by $\phi_j(t, t_0)$ the solution corresponding to the initial condition $(x(t_0), \dots, x^{(n-1)}(t_0)) = \delta_j$, the principal matrix solution is given by

$$\Pi(t, t_0) = \begin{pmatrix} \phi_1(t, t_0) & \cdots & \phi_n(t, t_0) \\ \dot{\phi}_1(t, t_0) & \cdots & \dot{\phi}_n(t, t_0) \\ \vdots & \vdots & \vdots \\ \phi_1^{(n-1)}(t, t_0) & \cdots & \phi_n^{(n-1)}(t, t_0) \end{pmatrix}. \tag{3.105}$$

As a consequence of Theorem 3.12 we obtain:

Theorem 3.13. *The solution of the inhomogeneous n-th order linear equation*

$$x^{(n)} + q_{n-1}(t)x^{(n-1)} + \cdots + q_1(t)\dot{x} + q_0(t)x = g(t) \tag{3.106}$$

corresponding to the initial condition

$$x(t_0) = x_0, \quad \dots \quad x^{(n-1)}(t_0) = x_{n-1}, \tag{3.107}$$

is given by

$$x(t) = x_0\phi_1(t, t_0) + \cdots + x_{n-1}\phi_n(t, t_0) + \int_{t_0}^{t} \phi_n(t, s)g(s)ds, \tag{3.108}$$

where $\phi_j(t, t_0)$, $1 \le j \le n$, are the solutions corresponding to the initial conditions $(\phi_j(t_0, t_0), \dots, \phi_j^{(n-1)}(t_0, t_0)) = \delta_j$.

Next, given sufficiently smooth functions $f_1, \dots, f_m$ we define their **Wronski determinant** (or simply their **Wronskian**) as

$$W(f_1, \dots, f_m) = \det \begin{pmatrix} f_1 & \cdots & f_m \\ f_1' & \cdots & f_m' \\ \vdots & \vdots & \vdots \\ f_1^{(m-1)} & \cdots & f_m^{(m-1)} \end{pmatrix}. \tag{3.109}$$

Note that the Wronskian will vanish identically if the functions are linearly dependent, but the converse is in general not true (cf. Problem 3.33).

By Lemma 3.11 the Wronskian of n solutions satisfies

$$W(\phi_1, \dots, \phi_n)(t) = W(\phi_1, \dots, \phi_n)(t_0) \exp\left(-\int_{t_0}^{t} q_{n-1}(s) ds \right) \tag{3.110}$$

and it will vanish if and only if the solutions are linearly dependent.

Finally, note that the differential equation (3.102) is uniquely determined by n linearly independent solutions $\phi_1, \dots, \phi_n$ since this is true for the corresponding system. Explicitly we have

$$\frac{W(\phi_1, \dots, \phi_n, x)(t)}{W(\phi_1, \dots, \phi_n)(t)} = x^{(n)}(t) + q_{n-1}(t) x^{(n-1)}(t) + \cdots + q_0(t) x(t). \tag{3.111}$$

In fact, by expanding the Wronski determinant with respect to the last column we see that the left-hand side is of the same form as the right-hand side with possibly different coefficients $\tilde{q}_j$. However, since the Wronskian on the left-hand side vanishes whenever we choose $x = \phi_j$, the corresponding differential equation has the same solutions and thus $\tilde{q}_j = q_j$.

Example. For example, in the case of second order equations we obtain using Laplace expansion along the last column

$$W(\phi_1, \phi_2, x) = W(\phi_1, \phi_2)\ddot{x} - \dot{W}(\phi_1, \phi_2)\dot{x} + W(\dot{\phi}_1, \dot{\phi}_2) x \tag{3.112}$$

and thus

$$q_1 = -\frac{\dot{W}(\phi_1, \phi_2)}{W(\phi_1, \phi_2)}, \quad q_0 = \frac{W(\dot{\phi}_1, \dot{\phi}_2)}{W(\phi_1, \phi_2)}. \tag{3.113}$$

Note that the formula for q_1 is consistent with (3.110). ◇

As for the case of systems, there is no general way of solving linear n'th order equations except for the trivial case $n = 1$ (recall (1.40)). However, if one solution $\phi_1(t)$ is known, one can again use the following method known as **reduction of order** (d'Alembert):

Given one solution $\phi_1(t)$ of (3.102), the variation of constants ansatz

$$x(t) = c(t)\phi_1(t) \tag{3.114}$$

gives a $(n-1)$'th order equation for $\dot{c}$: Setting $q_n = 1$ and using Leibniz rule we obtain

$$\sum_{j=0}^{n} q_j x^{(j)} = \sum_{j=0}^{n} q_j \sum_{k=0}^{j} \binom{j}{k} c^{(k)} \phi_1^{(j-k)} = \sum_{j=0}^{n} q_j \sum_{k=1}^{j} \binom{j}{k} c^{(k)} \phi_1^{(j-k)}, \quad (3.115)$$

where we have used $\sum_{j=0}^{n} q_j c \phi_1^{(j)} = 0$ for $k = 0$. Thus x solves (3.102) if and only if $d = \dot{c}$ solves

$$\sum_{k=0}^{n-1} d^{(k)} \sum_{j=k+1}^{n} \binom{j}{k+1} q_j \phi_1^{(j-k-1)} = 0. \quad (3.116)$$

Hence it remains to solve this $(n-1)$'th order equation for d and perform one additional integration to obtain c.

Example. Consider the differential equation

$$\ddot{x} - 2t\dot{x} - 2x = 0$$

and observe that $\phi_1(t) = \mathrm{e}^{t^2}$ is a solution. Hence we can set $x(t) = \mathrm{e}^{t^2} c(t)$ to obtain

$$\begin{aligned}&\big(\mathrm{e}^{t^2}\ddot{c}(t) + 4t\mathrm{e}^{t^2}\dot{c}(t) + (2+4t^2)\mathrm{e}^{t^2}c(t)\big) - 2\big(\mathrm{e}^{t^2}\dot{c}(t) + 2t\mathrm{e}^{t^2}c(t)\big) - 2\mathrm{e}^{t^2}c(t)\\ &\quad = \mathrm{e}^{t^2}(\ddot{c}(t) + 2t\dot{c}(t)) = 0.\end{aligned}$$

The solution of this equation is given by

$$\dot{c}(t) = \mathrm{e}^{-t^2}$$

implying

$$c(t) = \int_0^t \mathrm{e}^{-s^2} ds = \frac{\sqrt{\pi}}{2}\operatorname{erf}(t),$$

where $\operatorname{erf}(z) = \frac{2}{\sqrt{\pi}}\int_0^z \mathrm{e}^{-x^2} dx$ is the Gauss error function. Hence a second solution is given by $\phi_2(t) = \mathrm{e}^{t^2}\operatorname{erf}(t)$. ⋄

There is also an alternative method based on factorizing the differential equation outlined in Problems 3.35 and 3.36. Moreover, one can choose $q_{n-1}(t) = 0$ without loss of generality by Problem 3.37.

Problem 3.32. *Use reduction of order to find the general solution of the following equations:*

(i) $t\,\ddot{x} - 2(t+1)\dot{x} + (t+2)x = 0$, $\phi_1(t) = \mathrm{e}^t$.

(ii) $t^2\ddot{x} - 3t\,\dot{x} + 4x = 0$, $\phi_1(t) = t^2$.

Problem 3.33. *Show that the Wronskian of the two functions $f_1(t) = t^2$ and $f_2(t) = t|t|$ vanishes identically even though the two solutions are not linearly dependent.*

Problem 3.34. *Consider the equation $\ddot{x} + q_0(t)x$. Assume one solution is ϕ_1 and use reduction of order to show that a second solution is given by*

$$\phi_2(t) = \phi_1(t) \int^t \frac{1}{\phi_1(s)^2} ds.$$

Problem 3.35. *Verify that the second-order equation*

$$\ddot{x} + (1 - t^2)x = 0$$

can be factorized as

$$\left(\frac{d}{dt} - t\right)\left(\frac{d}{dt} + t\right) x = 0$$

(note that the order is important). Use this to find the solution. (Hint: The solution can be found by solving two first order problems.)

Problem 3.36. *Show that any linear n-th order equation can be factorized into first order equations:*

Let $\phi_1, \dots, \phi_n$ be linearly independent solutions of the n-th order equation $L_n(f) = 0$. Set

$$L_1(f) = \frac{W(\phi_1, f)}{\phi_1} = f' - \frac{\phi_1'}{\phi_1} f$$

and define $\psi_j = L_1(\phi_j)$. Show that $\psi_2, \dots, \psi_n$ are linearly independent and

$$L_n(f) = L_{n-1}(L_1(f)), \qquad L_{n-1}(f) = \frac{W(\psi_2, \dots, \psi_n, f)}{W(\psi_2, \dots, \psi_n)}.$$

Problem 3.37. *Consider the change of variables*

$$y(t) = Q(t)x(t), \qquad Q(t) = \mathrm{e}^{\frac{1}{n}\int^t q_{n-1}(s)ds}.$$

Show that if $x(t)$ satisfies (3.102)*, then $y(t)$ satisfies*

$$y^{(n)} + \sum_{k=0}^{n-2} \sum_{j=k}^{n} \binom{j}{k} q_j(t) Q^{(j-k)}(t) y^{(k)},$$

where $q_n(t) = 1$. In particular, the new equation does not contain $y^{(n-1)}$.

Problem 3.38. *Show that x solves*

$$\ddot{x} + q_1(t)\dot{x} + q_0(t)x = 0$$

if and only if

$$y(t) = \mathrm{e}^{Q(t)} \frac{\dot{x}(t)}{x(t)}, \qquad Q(t) = \int^t q_1(s) ds,$$

solves the Riccati equation

$$\dot{y} + \mathrm{e}^{-Q(t)} y^2 + \mathrm{e}^{Q(t)} q_0(t).$$

3.6. Periodic linear systems

In this section we want to consider (3.79) in the special case where $A(t)$ is periodic,

$$A(t+T) = A(t), \qquad T > 0. \tag{3.117}$$

This periodicity condition implies that $x(t+T)$ is again a solution if $x(t)$ is. Moreover, we even have

Lemma 3.14. *Suppose $A(t)$ is periodic with period T. Then the principal matrix solution satisfies*

$$\Pi(t+T, t_0+T) = \Pi(t, t_0). \tag{3.118}$$

Proof. By $\frac{d}{dt}\Pi(t+T,t_0+T) = A(t+T)\Pi(t+T,t_0+T) = A(t)\Pi(t+T,t_0+T)$ and $\Pi(t_0+T, t_0+T) = \mathbb{I}$ we see that $\Pi(t+T, t_0+T)$ solves (3.83). Thus it is equal to $\Pi(t, t_0)$ by uniqueness. □

Hence it suggests itself to investigate what happens if we move on by one period, that is, to look at the **monodromy matrix**

$$M(t_0) = \Pi(t_0+T, t_0). \tag{3.119}$$

Note that $M(t_0)$ is periodic by our previous lemma, that is, $M(t_0+T) = M(t_0)$.

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.

However, we have

$$\begin{aligned}\Pi(t_0+\ell T, t_0) &= \Pi(t_0+\ell T, t_0+(\ell-1)T)\Pi(t_0+(\ell-1)T, t_0)\\ &= M(t_0+(\ell-1)T)\Pi(t_0+(\ell-1)T, t_0)\\ &= M(t_0)\Pi(t_0+(\ell-1)T, t_0)\\ &= M(t_0)^\ell\Pi(t_0, t_0) = M(t_0)^\ell. \end{aligned} \tag{3.120}$$

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.

To factor out the exponential term we need to give a meaning to $M(t_0)^\ell$ for the case where $\frac{t}{T} = \ell$ is not an integer. If $M(t_0)$ is a number, the usual way of doing this is to set $M(t_0)^{t/T} = \exp(\frac{t}{T}\log(M(t_0)))$. To mimic this trick we need to find a matrix $Q(t_0)$ such that

$$M(t_0) = \exp(TQ(t_0)), \qquad Q(t_0+T) = Q(t_0). \tag{3.121}$$

This is possible if and only if $\det(M(t_0)) \neq 0$ (see Section 3.8). Note however, that $Q(t_0)$ is not unique.

That $\det(M(t_0)) \neq 0$ follows from Liouville's formula (3.91) which implies that the determinant of the monodromy matrix

$$\det(M(t_0)) = \exp\left(\int_{t_0}^{t_0+T} \mathrm{tr}(A(s))ds\right) = \exp\left(\int_0^T \mathrm{tr}(A(s))ds\right) \tag{3.122}$$

is independent of t_0 and positive.

Now writing

$$\Pi(t,t_0) = P(t,t_0)\exp((t-t_0)Q(t_0)) \tag{3.123}$$

a straightforward computation shows that

$$\begin{aligned} P(t+T,t_0) &= \Pi(t+T,t_0)M(t_0)^{-1}\mathrm{e}^{-(t-t_0)Q(t_0)} \\ &= \Pi(t+T,t_0+T)\mathrm{e}^{-(t-t_0)Q(t_0)} \\ &= \Pi(t,t_0)\mathrm{e}^{-(t-t_0)Q(t_0)} = P(t,t_0) \end{aligned} \tag{3.124}$$

as anticipated. In summary we have proven **Floquet's theorem**.

Theorem 3.15 (Floquet). *Suppose $A(t)$ is periodic. Then the principal matrix solution of the corresponding linear system has the form*

$$\Pi(t,t_0) = P(t,t_0)\exp((t-t_0)Q(t_0)), \tag{3.125}$$

where $P(.,t_0)$ has the same period as $A(.)$ and $P(t_0,t_0) = \mathbb{I}$.

Example. Consider the one-dimensional case

$$\dot{x} = a(t)x, \qquad a(t+T) = a(t).$$

Then the principal matrix solution is

$$\Pi(t,t_0) = \mathrm{e}^{\int_{t_0}^t a(s)ds}$$

and the monodromy matrix is

$$M(t_0) = \mathrm{e}^{\int_{t_0}^{t_0+T} a(s)ds} = \mathrm{e}^{T\bar{a}}, \qquad \bar{a} = \frac{1}{T}\int_0^T a(s)ds.$$

Moreover,

$$P(t,t_0) = \mathrm{e}^{\int_{t_0}^t (a(s)-\bar{a})ds}, \qquad Q(t_0) = \bar{a}.$$

⋄

Note that any fundamental matrix solution can be written in this form (Problem 3.41). Moreover, note that $Q(t_0)$ will be complex even if $A(t)$ is real unless all real eigenvalues of $M(t_0)$ are positive. However, since $A(t)$ also has the period $2T$ and $\Pi(t_0+2T,t_0) = M(t_0)^2$, we infer from Lemma 3.34:

Corollary 3.16. *Suppose $A(t)$ is real and periodic. Then the principal matrix solution of the corresponding linear system has the form*

$$\Pi(t,t_0) = \tilde{P}(t,t_0)\exp((t-t_0)\tilde{Q}(t_0)), \tag{3.126}$$

where both $\tilde{P}(t,t_0)$, $\tilde{Q}(t_0)$ are real and $\tilde{P}(.,t_0)$ has twice the period of $A(.)$.

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

$$\begin{aligned} M(t_1) &= \Pi(t_1+T,t_0+T)M(t_0)\Pi(t_0,t_1) \\ &= \Pi(t_1,t_0)M(t_0)\Pi(t_1,t_0)^{-1}. \end{aligned} \tag{3.127}$$

Thus the eigenvalues and the Jordan structure are independent of t_0 (hence the same also follows for $Q(t_0)$).

The eigenvalues ρ_j of $M(t_0)$ are known as **Floquet multipliers** (also **characteristic multipliers**) and the eigenvalues γ_j of $Q(t_0)$ are known as **Floquet exponents** (**characteristic exponents**). By Lemma 3.3 they are related via $\rho_j = \mathrm{e}^{T\gamma_j}$. Since the periodic part $P(t,t_0)$ is bounded we obtain as in Corollary 3.5

Corollary 3.17. *A periodic linear system is stable if all Floquet multipliers satisfy $|\rho_j| \le 1$ (respectively all Floquet exponents satisfy $\mathrm{Re}(\gamma_j) \le 0$) and for all Floquet multipliers with $|\rho_j| = 1$ (respectively all Floquet exponents with $\mathrm{Re}(\gamma_j) = 0$) the algebraic and geometric multiplicities are equal.*

A periodic linear system is asymptotically stable if all Floquet multipliers satisfy $|\rho_j| < 1$ (respectively all Floquet exponents satisfy $\mathrm{Re}(\gamma_j) < 0$).

Before I show how this result is used in a concrete example, let me note another consequence of Theorem 3.15. The proof is left as an exercise (Problem 3.42).

Corollary 3.18. *The transformation $y(t) = P(t,t_0)^{-1}x(t)$ renders the system into one with constant coefficients,*

$$\dot{y}(t) = Q(t_0)y(t). \tag{3.128}$$

Note also that we have $P(t,t_0)^{-1} = \exp((t-t_0)Q(t_0))P(t_0,t)\exp(-(t-t_0)Q(t))$ by virtue of $\Pi(t,t_0)^{-1} = \Pi(t_0,t)$.

In the remainder of this section we will look at one of the most prominent examples, **Hill's equation**

$$\ddot{x} + q(t)x = 0, \qquad q(t+T) = q(t). \tag{3.129}$$

In this case the associated system is

$$\dot{x} = y, \qquad \dot{y} = -qx \tag{3.130}$$

and the principal matrix solution is given by

$$\Pi(t,t_0) = \begin{pmatrix} c(t,t_0) & s(t,t_0) \\ \dot{c}(t,t_0) & \dot{s}(t,t_0) \end{pmatrix}, \tag{3.131}$$

where $c(t,t_0)$ is the solution of (3.129) corresponding to the initial condition $c(t_0,t_0)=1$, $\dot{c}(t_0,t_0)=0$ and similarly for $s(t,t_0)$ but corresponding to the initial condition $s(t_0,t_0)=0$, $\dot{s}(t_0,t_0)=1$. Liouville's formula (3.91) shows

$$\det \Pi(t,t_0) = 1 \tag{3.132}$$

and hence the characteristic equation for the monodromy matrix

$$M(t_0) = \begin{pmatrix} c(t_0+T,t_0) & s(t_0+T,t_0) \\ \dot{c}(t_0+T,t_0) & \dot{s}(t_0+T,t_0) \end{pmatrix}, \tag{3.133}$$

is given by

$$\rho^2 - 2\Delta\rho + 1 = 0, \tag{3.134}$$

where

$$\Delta = \frac{\operatorname{tr}(M(t_0))}{2} = \frac{c(t_0+T,t_0) + \dot{s}(t_0+T,t_0)}{2}. \tag{3.135}$$

If $\Delta^2 > 1$ we have two different real eigenvalues

$$\rho_\pm = \Delta \pm \sqrt{\Delta^2 - 1}, \tag{3.136}$$

with corresponding eigenvectors

$$u_\pm(t_0) = \begin{pmatrix} 1 \\ m_\pm(t_0) \end{pmatrix}, \tag{3.137}$$

where

$$m_\pm(t_0) = \frac{\rho_\pm - c(t_0+T,t_0)}{s(t_0+T,t_0)} = \frac{\dot{c}(t_0+T,t_0)}{\rho_\pm - \dot{s}(t_0+T,t_0)}. \tag{3.138}$$

Note that $u_\pm(t_0)$ are also eigenvectors of $Q(t_0)$ corresponding to the eigenvalues $\gamma_\pm = \frac{1}{T}\log(\rho_\pm)$ (Lemma 3.3). From $\rho_+\rho_- = 1$ we obtain $\gamma_+ + \gamma_- = 0$ mod $2\pi\mathrm{i}$ and it is no restriction to assume $|\rho_+| \geq 1$ respectively $\mathrm{Re}(\gamma_+) \geq 0$. If we set $\gamma = \mathrm{Re}(\gamma_+)$, we have $\gamma_\pm = \pm\gamma$ if $\rho_\pm > 0$ (i.e. $\Delta = (\rho_+ + \rho_-)/2 > 0$) and $\gamma_\pm = \pm\gamma + \mathrm{i}\pi$ if $\rho_\pm < 0$ (i.e. $\Delta < 0$). In summary, the characteristic multipliers are of the form

$$\rho_\pm = \sigma\, \mathrm{e}^{\pm T\gamma}, \qquad \sigma = \operatorname{sgn}(\Delta),\ \gamma = \frac{1}{T}\log|\rho_+| > 0. \tag{3.139}$$

Considering

$$\begin{aligned} \Pi(t,t_0)u_\pm(t_0) &= P(t,t_0)\exp((t-t_0)Q(t_0))u_\pm(t_0) \\ &= \mathrm{e}^{\gamma_\pm(t-t_0)}P(t,t_0)u_\pm(t_0), \end{aligned} \tag{3.140}$$

we see that there are two solutions of the form

$$\mathrm{e}^{\pm\gamma t}p_\pm(t), \qquad p_\pm(t+T) = \sigma\, p_\pm(t). \tag{3.141}$$

If $\Delta^2 < 1$ we have two different complex conjugate eigenvalues and hence two solutions

$$\mathrm{e}^{\pm\mathrm{i}\gamma t} p_\pm(t), \qquad p_\pm(t+T) = p_\pm(t), \quad \gamma > 0, \tag{3.142}$$

where $\gamma = \mathrm{Im}(\gamma_+)$.

If $\Delta^2 = 1$ we have $\rho_\pm = \Delta$ and either two solutions

$$p_\pm(t), \qquad p_\pm(t+T) = \sigma\, p_\pm(t), \tag{3.143}$$

or two solutions

$$p_+(t), \quad p_-(t) + t\, p_+(t), \qquad p_\pm(t+T) = \sigma\, p_\pm(t), \tag{3.144}$$

where $\sigma = \mathrm{sgn}(\Delta) = \Delta$.

Since a periodic equation is called **stable** if all solutions are bounded, we have shown:

Theorem 3.19. *Hill's equation is stable if $|\Delta| < 1$ and unstable if $|\Delta| > 1$.*

This result is of practical importance in applications. For example, the potential of a charged particle moving in the electric field of a quadrupole is given by

$$U(x) = e\frac{V}{a^2}(x_1^2 - x_2^2). \tag{3.145}$$

The corresponding equations of motion are Newton's equation (1.5), where the force is given by

$$F(x) = -\frac{\partial}{\partial x} U(x). \tag{3.146}$$

If the voltage V varies with respect to time according to $V(t) = V_0 + V_1 \cos(t)$, one gets the following equations of motion (neglecting the induced magnetic field)

$$\begin{aligned} \ddot{x}_1 &= -\frac{2e}{ma^2}(V_0 + V_1\cos(t))x_1, \\ \ddot{x}_2 &= +\frac{2e}{ma^2}(V_0 + V_1\cos(t))x_2, \\ \ddot{x}_3 &= 0. \end{aligned} \tag{3.147}$$

The equation for the x_1 and x_2 coordinates is the **Mathieu equation**

$$\ddot{x} = -\omega^2(1 + \varepsilon\cos(t))x. \tag{3.148}$$

A numerically computed stability diagram is depicted in Figure 3.6. 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_1 or x_2 direction) if the mass of the particle lies within

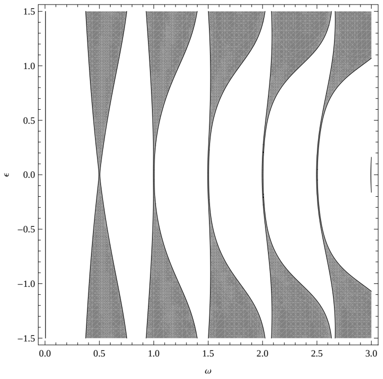


Figure 3.6. Numerically computed stability diagram for the Mathieu equation with $0 \leq \omega \leq 3$ and $-1.5 \leq \varepsilon \leq 1.5$.

a certain region. This can be used to filter charged particles according to their mass (**quadrupole mass spectrometry**).

Hill's equation also can be used as a simple one-dimensional model in quantum mechanics to describe a single electron moving in a periodic field (cf. Problem 5.36). We will further investigate this problem in Section 5.6.

Problem 3.39. *Consider*

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

where $a : \mathbb{R} \to \mathbb{R}$ is periodic with period T and A is a constant two by two matrix. Compute the Floquet exponent, and find $P(t, t_0)$ and $Q(t_0)$ in this case.

Problem 3.40. *Compute the monodromy matrix where $A(t)$ is of period 1 and given by*

$$A(t) = \begin{cases} \begin{pmatrix} \alpha & 1 \\ 0 & \alpha \end{pmatrix}, & 0 \leq t < \frac{1}{2}, \\ \begin{pmatrix} \alpha & 0 \\ 1 & \alpha \end{pmatrix}, & \frac{1}{2} \leq t < 1, \end{cases} \qquad \alpha \in \mathbb{C}.$$

Note that since $A(t)$ is not continuous you have to match solutions at every discontinuity such that the solutions are continuous (cf. Section 2.3).

For which values of α remain all solutions bounded? Show that the bound found in Problem 3.31 is optimal by considering $A(t/T)$ as $T \to 0$.

(Note that we could approximate $A(t)$ by continuous matrices and obtain the same qualitative result with an arbitrary small error.)

Problem 3.41. *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 3.42. *Prove Corollary 3.18.*

Problem 3.43. *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 unique periodic solution of period T if and only 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 (3.97)*.)*

3.7. Perturbed linear first order systems

In this section we want to consider stability of perturbed linear systems of the form

$$\dot{x} = (A(t) + B(t))x, \tag{3.149}$$

where the asymptotic behavior as $t \to \infty$ of the system associated with $A(t)$ is well understood and $B(t)$ is supposed to be some *small* perturbation. We begin by looking at the one-dimensional case.

Example. The solution of the equation

$$\dot{x} = (-a + b(t))x, \qquad x(0) = x_0,$$

is given by

$$x(t) = x_0 \exp\left(-at + \int_0^t b(s)ds\right).$$

If we assume $a > 0$, the unperturbed system is asymptotically stable and all solutions tend to 0 exponentially fast, $|x(t)| \le |x_0|\mathrm{e}^{-at}$, as $t \to \infty$. The same is true for the perturbed system if we, for example, assume that eventually $b(t) \le b_0 < a$. However, note that even if $b(t) \to 0$, the asymptotic form of the solution will in general differ from the unperturbed one. For example, in the case $b(t) = (1+t)^{-1}$ we obtain $x(t) = x_0(1+t)\mathrm{e}^{-at}$. In particular, in the case $a = 0$ the unperturbed system is stable and for the above choice of $b(t) = (1+t)^{-1}$ the perturbed system is unstable. If we make the stronger requirement $\int_0^\infty |b(t)|dt < \infty$, then the perturbed system is again stable even if $a = 0$. ◇

Our aim is to transfer the above observations for the one-dimensional case to general first order systems.

Theorem 3.20. *Consider the system* (3.149) *and suppose that the principal matrix solution of the unperturbed system corresponding to $B(t) \equiv 0$ satisfies*

$$\|\Pi_A(t,s)\| \le C\mathrm{e}^{-\alpha(t-s)}, \qquad t \ge s \ge t_0, \tag{3.150}$$

for some constants $C, \alpha > 0$ and a time $t_0 \ge 0$. Suppose that

$$\|B(t)\| \le b_0, \qquad t \ge t_0. \tag{3.151}$$

Then, if $b_0 C < \alpha$*, we have*

$$\|\Pi_{A+B}(t,s)\| \le D\mathrm{e}^{-(\alpha - b_0 C)(t-s)}, \qquad t \ge s \ge 0, \tag{3.152}$$

for some constant $D > 0$.

Proof. The key ingredient is the variation of constants formula (3.97), where we rewrite (3.149) as

$$\dot{x} - A(t)x = B(t)x$$

and regard the right-hand side as an inhomogeneous term $g(t) = B(t)x(t)$. Then

$$x(t) = \Pi_A(t,s)x(s) + \int_s^t \Pi_A(t,r)B(r)x(r)dr.$$

By our assumptions we obtain

$$|x(t)| \le C\mathrm{e}^{-\alpha(t-s)}|x(s)| + \int_s^t C\mathrm{e}^{-\alpha(t-r)} b_0 |x(r)| dr$$

for $t \ge s \ge t_0$. Introducing $y(t) = |x(t)|\mathrm{e}^{\alpha(t-s)}$ we get

$$y(t) \le C|x(s)| + \int_s^t C b_0 y(r) dr$$

and Gronwall's inequality implies $y(t) \le C|x(s)|\mathrm{e}^{Cb_0(t-s)}$ and hence

$$|x(t)| \le C|x(s)|\mathrm{e}^{-(\alpha - Cb_0)(t-s)}, \qquad t \ge s \ge t_0.$$

Finally, to obtain the general case use (3.88) to reduce it to the case where all times are either $\ge t_0$ or $\le t_0$ and (3.87) to estimate the latter case. This shows that the claim holds with $D = \mathrm{e}^{(\beta + (\alpha - Cb_0))t_0} C$, where $\beta = \max_{0 \le t \le t_0} \|A(t) + B(t)\|$. □

In order to apply this result note that estimates for $\Pi_A(t,s)$ of the required type are provided in Problem 3.31.

As a first consequence we conclude that asymptotic stability is preserved for perturbed linear systems of the form

$$\dot{x} = (A + B(t))x, \tag{3.153}$$

where $B(t)$ is continuous and satisfies $\|B(t)\| \to 0$ as $t \to \infty$. To this end recall that by Corollary 3.6 the unperturbed system corresponding to $B(t) = 0$ is asymptotically stable if and only if all eigenvalues of A have negative real part. Moreover, in this case (3.44) shows that the assumptions of our theorem are satisfied.

Corollary 3.21. *Suppose all eigenvalues* α_j *of* A *have negative real part and* $B(t)$ *satisfies*

$$\lim_{t\to\infty} \|B(t)\| = 0. \tag{3.154}$$

Then the linear system (3.153) *is asymptotically stable. More precisely, for every* $\alpha < \min\{-\mathrm{Re}(\alpha_j)\}_{j=1}^m$ *there is a constant* C *such that*

$$|x(t)| \le C\mathrm{e}^{-t\alpha}|x_0|, \qquad t \ge 0, \tag{3.155}$$

where $x(t)$ *is the solution corresponding to the initial condition* $x(0) = x_0$.

Example. Consider the two dimensional system with

$$A = \begin{pmatrix} -a & 0 \\ 0 & -a \end{pmatrix}, \qquad B(t) = \begin{pmatrix} 0 & \sin(t) \\ \cos(t) & 0 \end{pmatrix}.$$

Since

$$\|B(t)\| = \max(|\sin(t)|, |\cos(t)|)$$

does not tend to 0 (use Problem 3.49 to compute the norm), our corollary does not apply. However, A satisfies (3.150) with $C = 1$, $\alpha = a$ and hence we can conclude that this system is asymptotically stable if $a > 1$. ⋄

Since, by Floquet's theorem (Theorem 3.15), the principal matrix solution of a periodic linear system looks like the one of a constant system up to periodic factors, the above result applies even in this more general case.

Corollary 3.22. *Let* $A(t)$ *be periodic. Suppose all Floquet exponents* γ_j *of* $A(t)$ *have negative real part and* $B(t)$ *satisfies*

$$\lim_{t\to\infty} \|B(t)\| = 0. \tag{3.156}$$

Then the linear system (3.149) *is asymptotically stable. More precisely, for every* $\gamma < \min\{-\mathrm{Re}(\gamma_j)\}_{j=1}^m$ *there is a constant* C *such that*

$$|x(t)| \le C\mathrm{e}^{-t\gamma}|x_0|, \qquad t \ge 0, \tag{3.157}$$

where $x(t)$ *is the solution corresponding to the initial condition* $x(0) = x_0$.

As our second result we will show that stability is preserved under such perturbations if the norm of the perturbation is integrable.

Theorem 3.23. *Consider the system* (3.149) *and suppose that the principal matrix solution of the unperturbed system corresponding to* $B(t) \equiv 0$ *satisfies*

$$\|\Pi_A(t,s)\| \le C, \qquad t \ge s \ge t_0, \tag{3.158}$$

for some constant $C > 0$ *and a time* $t_0 \ge 0$. *Suppose that*

$$\int_0^\infty \|B(t)\| dt < \infty. \tag{3.159}$$

Then we have

$$\|\Pi_{A+B}(t,0)\| \le D, \qquad t \ge 0, \tag{3.160}$$

for some constant $D > 0$.

Proof. As in the previous proof our point of departure is

$$x(t) = \Pi_A(t,t_0)x_0 + \int_0^t \Pi_A(t,s)B(s)x(s)ds$$

and using our estimate for Π_A we obtain

$$|x(t)| \le C|x_0| + \int_0^t C\|B(s)\||x(s)|ds.$$

Hence an application of Gronwall's inequality

$$|x(t)| \le C|x_0| \exp\left(C \int_0^\infty \|B(s)\|ds\right)$$

finishes the proof. □

Again we can apply this to the case where A is constant. To this end recall that by Corollary 3.5 the system corresponding to $B(t) = 0$ is stable if and only if all eigenvalues of A have nonpositive real part. Moreover, (3.43) provides the necessary estimate.

Corollary 3.24. *Suppose all eigenvalues α_j of A satisfy $\mathrm{Re}(\alpha_j) \le 0$ and for all eigenvalues with $\mathrm{Re}(\alpha_j) = 0$ the corresponding algebraic and geometric multiplicities are equal, and $B(t)$ satisfies*

$$\int_0^\infty \|B(t)\|dt < \infty. \tag{3.161}$$

Then the linear system (3.153) *is stable. More precisely, there is a constant C such that*

$$\|x(t)\| \le C|x_0|, \qquad t \ge 0, \tag{3.162}$$

where $x(t)$ is the solution corresponding to the initial condition $x(0) = x_0$.

Again the result applies to perturbed periodic systems as well.

Corollary 3.25. *Let $A(t)$ be periodic. Suppose all Floquet exponents γ_j of $A(t)$ satisfy $\mathrm{Re}(\gamma_j) \le 0$ and for all Floquet exponents with $\mathrm{Re}(\gamma_j) = 0$ the corresponding algebraic and geometric multiplicities are equal, and $B(t)$ satisfies*

$$\int_0^\infty \|B(t)\|dt < \infty. \tag{3.163}$$

Then the linear system (3.149) *is stable. More precisely, there is a constant C such that*

$$|x(t)| \le C|x_0|, \qquad t \ge 0, \tag{3.164}$$

where $x(t)$ is the solution corresponding to the initial condition $x(0) = x_0$.

Finally, note that we also could admit nonlinear perturbations,

$$\dot{x} = A(t)x + g(t,x), \tag{3.165}$$

as long as the nonlinear term satisfies a linear estimate. For example, the same proof as for Theorem 3.20 shows:

Theorem 3.26. *Consider the system* (3.165) *and suppose that the principal matrix solution of the unperturbed system corresponding to* $g(t,x) \equiv 0$ *satisfies*

$$\|\Pi_A(t,s)\| \le C\mathrm{e}^{-\alpha(t-s)}, \qquad t \ge s \ge 0, \tag{3.166}$$

for some constants $C, \alpha > 0$. *Suppose that*

$$|g(t,x)| \le b_0|x|, \qquad |x| < \delta,\ t \ge 0, \tag{3.167}$$

for some constant $0 < \delta \le \infty$. *Then, if* $b_0 C < \alpha$, *the solution* $x(t)$ *starting at* $x(0) = x_0$ *satisfies*

$$\|x(t)\| \le D\mathrm{e}^{-(\alpha - b_0 C)t}|x_0|, \qquad |x_0| < \frac{\delta}{C},\ t \ge 0, \tag{3.168}$$

for some constant $D > 0$.

As an important consequence we single out a useful criterion for asymptotic stability of a fixed point of an autonomous system.

Corollary 3.27. *Suppose* $f \in C^1$ *satisfies* $f(0) = 0$ *and suppose that all eigenvalues of the Jacobian matrix at* 0 *have negative real part. Then there is a* $\delta > 0$ *and an* $\alpha > 0$ *such that solutions of*

$$\dot{x} = f(x), \qquad x(0) = x_0, \tag{3.169}$$

satisfy

$$|x(t)| \le C\mathrm{e}^{-\alpha t}|x_0|, \qquad |x_0| \le \delta. \tag{3.170}$$

Proof. We first write our system in the form (3.165), where $A(t) = A$ is the Jacobian matrix of $f(x)$ at 0 and $g(t,x) = g(x)$ is the remainder. Then, by assumption, A satisfies our requirements and the same is true for $g(x)$ where b_0 can be made arbitrarily small by making δ small (since the Jacobian matrix of g vanishes at 0). □

Example. (Perron) Consider the nonlinear system

$$A(t) = \begin{pmatrix} -\alpha & 0 \\ 0 & -2\alpha + \sin(\log(t)) + \cos(\log(t)) \end{pmatrix}, \qquad g(t,x) = \begin{pmatrix} 0 \\ x_1^2 \end{pmatrix}.$$

The solution of the corresponding unperturbed system is given by

$$\Pi_A(t,t_0) = \begin{pmatrix} \mathrm{e}^{-\alpha(t-t_0)} & 0 \\ 0 & \mathrm{e}^{-2\alpha(t-t_0)+t\sin(\log(t))-t_0\sin(\log(t_0)))} \end{pmatrix}, \qquad t, t_0 > 0.$$

However, while solutions decay exponentially for $\alpha > \frac{1}{2}$ it is not clear for what α the stronger estimates (3.166) holds. Since the derivative of $t\sin(\log(t))$ does not exceed $\sqrt{2}$, we have $|t\sin(\log(t)) - s\sin(\log(s))| \le \sqrt{2}|t-s|$, and we get asymptotic stability from Theorem 3.26 at least for $\alpha > \frac{1}{\sqrt{2}}$.

The general solution of the nonlinear system is given by

$$x(t) = \begin{pmatrix} c_1 \mathrm{e}^{-\alpha t} \\ \mathrm{e}^{-2\alpha t + t\sin(\log(t))}\left(c_2 + c_1^2 \int_0^t \mathrm{e}^{-s\sin(\log(s))} ds\right) \end{pmatrix}.$$

Now for the sequence $t_n = \mathrm{e}^{(2n+11/6)\pi}$, $n \in \mathbb{N}$, we see that by

$$\begin{aligned} \int_0^{t_n} \mathrm{e}^{-s\sin(\log(s))} ds &> \int_{t_n \exp(-2\pi/3)}^{t_n} \mathrm{e}^{-s\sin(\log(s))} ds \\ &> t_n(1-\exp(-2\pi/3))\mathrm{e}^{t_n \exp(-2\pi/3)/2} \end{aligned}$$

the solutions with $c_1 \ne 0$ are unbounded as $t \to \infty$ for $\frac{1}{2} < \alpha < \frac{1}{2} + \frac{1}{4}\mathrm{e}^{-\pi}$. This shows that the condition (3.166) cannot be replaced by exponential decay of solutions. ⋄

Of course we can also obtain a nonlinear version of Theorem 3.23 by making the obvious changes in its proof.

Theorem 3.28. *Consider the system* (3.165) *and suppose that the principal matrix solution of the unperturbed system corresponding to* $g(t,x) \equiv 0$ *satisfies*

$$\|\Pi_A(t,s)\| \le C, \qquad t \ge s \ge 0, \tag{3.171}$$

for some constant $C > 0$. *Suppose that*

$$|g(t,x)| \le b(t)|x|, \qquad |x| < \delta,\ t \ge 0, \tag{3.172}$$

for some constant $0 < \delta \le \infty$ *and some function* $b(t)$ *with* $B = \int_0^\infty b(t) < 0$. *Then the solution* $x(t)$ *starting at* $x(0) = x_0$ *satisfies*

$$|x(t)| \le C\exp(CB)|x_0|, \qquad |x_0| \le \frac{\delta}{C\exp(CB)},\ t \ge 0. \tag{3.173}$$

Problem 3.44 (Long-time asymptotics)**.** *Suppose*

$$\int_0^\infty \|A(t)\| dt < \infty.$$

Show that every solution $x(t)$ *of* (3.79) *converges to some limit:*

$$\lim_{t\to\infty} x(t) = x_\infty.$$

(Hint: First show that all solutions are bounded and then use the corresponding integral equation.)

3.8. Appendix: Jordan canonical form

In this section we want to review some further facts on the Jordan canonical form. In addition, we want to draw some further consequences to be used later on.

Consider a decomposition of $\mathbb{C}^n$ into a direct sum of two linear subspaces, $\mathbb{C}^n = K_1 \oplus K_2$. Such a decomposition is said to **reduce** A if both subspaces K_1 and K_2 are **invariant** under A, that is, $AK_j \subseteq K_j$, $j = 1, 2$. Changing to a new basis $u_1, \dots, u_n$ such that $u_1, \dots, u_m$ is a basis for K_1 and $u_{m+1}, \dots, u_n$ is a basis for K_2, implies that A is transformed to the block form

$$U^{-1}AU = \begin{pmatrix} A_1 & 0 \\ 0 & A_2 \end{pmatrix}, \qquad U = (u_1, \dots, u_n), \tag{3.174}$$

in these new coordinates. Here $A_j = A|_{K_j}$. Moreover, we even have

$$U^{-1}\exp(A)U = \exp(U^{-1}AU) = \begin{pmatrix} \exp(A_1) & 0 \\ 0 & \exp(A_2) \end{pmatrix}. \tag{3.175}$$

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

$$Ax = \alpha x, \qquad x \neq 0, \tag{3.176}$$

for some $\alpha \in \mathbb{C}$. If (3.176) holds, α is called an **eigenvalue** of A and x is called **eigenvector**. In particular, α is an eigenvalue if and only if $\mathrm{Ker}(A - \alpha\mathbb{I}) \neq \{0\}$ and hence $\mathrm{Ker}(A - \alpha)$ is called the **eigenspace** of α in this case. Here we have used the shorthand notation $A - \alpha$ for $A - \alpha\mathbb{I}$. Since $\mathrm{Ker}(A - \alpha\mathbb{I}) \neq \{0\}$ implies that $A - \alpha\mathbb{I}$ is *not* invertible, the eigenvalues are the zeros of the **characteristic polynomial** of A,

$$\chi_A(z) = \prod_{j=1}^{m} (z - \alpha_j)^{a_j} = \det(z - A), \tag{3.177}$$

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

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

$$\sigma(A) = \{\alpha \in \mathbb{C} |\ \mathrm{Ker}(A - \alpha) \neq \{0\}\} = \{\alpha \in \mathbb{C} | \, \chi_A(\alpha) = 0\}. \tag{3.178}$$

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.

There is an important class of matrices where this will indeed work. To this end recall the definition of the **adjoint** matrix A^* which is defined as the complex conjugate of the transposed matrix such that

$$x \cdot (Ay) = (A^* x) \cdot y, \tag{3.179}$$

where $x \cdot y = \sum_{j=1}^{n} x_j^* y_j$ is the scalar product in $\mathbb{C}^n$. A matrix is called **symmetric** if $A^* = A$. A matrix U is called **orthogonal** (or **unitary**) if $U^* = U^{-1}$. Note that a matrix is orthogonal if and only if it preserves the scalar product,

$$(Ux) \cdot (Uy) = x \cdot (U^* U y) = x \cdot y. \tag{3.180}$$

In particular, the equation $U^* U = \mathbb{I}$ is equivalent to the fact that the row vectors of U form an **orthonormal basis**, that is, they are mutually orthogonal and normalized such that they have norm one. The same is true for the column vectors.

Theorem 3.29. *The eigenvalues of a symmetric matrix are real and there is an orthonormal basis of eigenvectors. In particular, there is an orthogonal matrix U which transforms A to diagonal form.*

Proof. Start with one normalized eigenvector u_1. Extend this vector to an orthogonal basis $u_1, \dots u_n$ (e.g. using the Gram–Schmidt procedure). Now observe that, by symmetry, we obtain $u_1 \cdot (A u_j) = (A u_1) \cdot u_j = \alpha_1 (u_1 \cdot u_j) = 0$ for $j = 2, \dots, n$. Hence in this new basis A is of the form

$$U^{-1} A U = \begin{pmatrix} \alpha_1 & 0 \\ 0 & A_2 \end{pmatrix}, \qquad U = (u_1, \cdots, u_n).$$

Since the transformation U is unitary, it preserves the scalar product and the $(n-1)$ by $(n-1)$ matrix A_2 is again symmetric and we can repeat this procedure until we have found an orthonormal basis of eigenvalues. □

However, life is not that simple and we only have $g_j \le a_j$ in general. It turns out that the right objects to look at are kernels of powers of $A - \alpha_j$:

$$K_{j,k} = \operatorname{Ker}(A - \alpha_j)^k, \qquad j = 1, \dots, m, \quad k = 1, \dots \tag{3.181}$$

First of all observe that

$$K_{j,1} \subseteq K_{j,2} \subseteq \cdots \tag{3.182}$$

and since our vector space is n dimensional there must be a smallest index $d_j \le n$ such that equality holds. In fact, it turns out that these spaces will remain the same from this index on:

$$K_{j,1} \subset K_{j,2} \subset \cdots \subset K_{j,d_j} = K_{j,d_j+1} = \dots \tag{3.183}$$

To see this note that $(A - \alpha_j)^{d_j + l} u = 0$ for some $l \ge 1$ implies $(A - \alpha_j)^{l-1} u \in K_{j,d_j+1} = K_{j,d_j}$ and thus $(A - \alpha_j)^{d_j + l - 1} u = (A - \alpha_j)^{d_j} (A - \alpha_j)^{l-1} u = 0$. We call

$$K_j = \operatorname{Ker}(A - \alpha_j)^{d_j}. \tag{3.184}$$

the **generalized eigenspace** corresponding to α_j and its elements are called **generalized eigenvectors**. For a generalized eigenvector u the smallest k with $(A-\alpha_j)^k u = 0$ is called its **order**.

Lemma 3.30. *Suppose $e_j \in \mathbb{N}_0$ are given numbers. Then*

$$\prod_{j=1}^{m} (A-\alpha_j)^{e_j} v = 0 \tag{3.185}$$

if and only if

$$v \in K_{1,e_1} \oplus \cdots \oplus K_{m,e_m}. \tag{3.186}$$

Proof. We show that (3.185) implies (3.186) via induction on $e = \sum_j e_j$ (the other direction being obvious). The case $e = 1$ is trivial. Now suppose $e \geq 2$ and assume there are two indices j, k such that $e_j \geq 1$ and $e_k \geq 1$ (otherwise the claim is again trivial). Then by induction hypothesis

$$v_j = (A-\alpha_j)v = \sum_l u_{j,l} \quad \text{and} \quad v_k = (A-\alpha_k)v = \sum_l u_{k,l},$$

where $u_{j,l} \in K_{l,e_l}$ for $l \neq j$ and $u_{j,j} \in K_{j,e_j-1}$ as well as $u_{k,l} \in K_{l,e_l}$ for $l \neq k$ and $u_{k,k} \in K_{k,e_k-1}$. But then the claim also holds for e since

$$v = \frac{1}{\alpha_j - \alpha_k}(v_k - v_j) = \frac{1}{\alpha_j - \alpha_k} \sum_l (u_{k,l} - u_{j,l}).$$

To show that we have a direct sum let $\sum_j x_j = 0$, $x_j \in K_{j,e_j}$, and set $p_k(z) = p(z)/(z-\alpha_k)^{e_k - l}$ with $p(z) = \prod_j (z-\alpha_j)^{e_j}$ and $l < e_k$ chosen such that $y_k = (A-\alpha_k)^l x_k \neq 0$ but $(A-\alpha_k)y_k = (A-\alpha_k)^{l+1}x_k = 0$. Then $0 = p_k(A)\sum_j x_j = \prod_{j\neq k}(\alpha_k - \alpha_j)^{d_j} y_k$ which contradicts $y_k \neq 0$. □

Lemma 3.31. *There is a unique monic polynomial $\mu_A(z)$ of minimal degree which annihilates A in the sense that $\mu_A(A) = 0$. It is of the form*

$$\mu_A(z) = \prod_{j=1}^{m} (z-\alpha_j)^{d_j}, \qquad d_j \geq 1, \tag{3.187}$$

and called the **minimal polynomial** *of A. Moreover, we can decompose our vector space as the following direct sum of invariant subspaces:*

$$\mathbb{C}^n = K_1 \oplus \cdots \oplus K_m. \tag{3.188}$$

Proof. There are clearly polynomials which annihilate A since the matrices A^j, $j = 0, \dots, n^2$ cannot be linearly independent. If there were more than one monic of minimal degree, their difference would also annihilate A and be of smaller degree.

Now let α_j be an eigenvalue with corresponding eigenvector u_j. Then $0 = \mu_A(A)u_j = \mu_A(\alpha_j)u_j$ shows that α_j is a zero of $\mu_A(z)$. Conversely, let

α be a zero and write $\mu_A(z) = (z-\alpha)\tilde{\mu}(z)$. Since $\tilde{\mu}(z)$ does not annihilate A, there is some u with $v = \tilde{\mu}(z)u \neq 0$. But then $(A-\alpha)v = \mu_A(A)u = 0$ shows that α is an eigenvalue.

Hence $\mu_A(z) = \prod_j (z-\alpha_j)^{e_j}$ for some numbers $e_j \geq 1$. By the previous lemma we have $\oplus_j K_{j,e_j} = V$ which shows $e_j \geq d_j$. The converse direction of the lemma shows $d_j \leq e_j$. □

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

$$U^{-1}AU = \begin{pmatrix} A_1 & & \\ & \ddots & \\ & & A_m \end{pmatrix}, \tag{3.189}$$

where each matrix $A_j = A|_{K_j}$ has only the eigenvalue α_j (why?). 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 $A^k u$, $0 \leq k \leq n-1$ span $\mathbb{C}^n$, that is,

$$\mathbb{C}^n = \{\sum_{k=0}^{n-1} a_k A^k u | a_k \in \mathbb{C}\}. \tag{3.190}$$

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

$$U = (u, (A-\alpha)u, \dots, (A-\alpha)^{n-1}u), \tag{3.191}$$

then U transforms A to

$$J = U^{-1}AU = \begin{pmatrix} \alpha & 1 & & & \\ & \alpha & 1 & & \\ & & \alpha & \ddots & \\ & & & \ddots & 1 \\ & & & & \alpha \end{pmatrix}, \tag{3.192}$$

since $(A-\alpha)^n u = 0$ by $K = \mathrm{Ker}(A-\alpha)^n = \mathbb{C}^n$. The matrix (3.192) 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 K_j into a direct sum of spaces K_{jk}, each of which has a cyclic vector u_{jk}.

We again restrict our attention to the case where A has only one eigenvalue α and consider again the spaces

$$K_k = \mathrm{Ker}(A-\alpha)^k. \tag{3.193}$$

To begin with we define M_n such that

$$K_n = K_{n-1} \oplus M_n. \tag{3.194}$$

Since $(A-\alpha)M_n \subseteq (A-\alpha)K_n \subseteq K_{n-1}$ we can proceed to define M_{n-1} such that

$$K_{n-1} = M_{n-1} \oplus (A-\alpha)M_n \oplus K_{n-2}. \tag{3.195}$$

This can be done since by construction of M_n, the space $(A-\alpha)M_n$ cannot contain any nontrivial vector from K_{n-2}. Proceeding like this we can find M_l such that

$$\mathbb{C}^n = \bigoplus_{l=1}^{n} \bigoplus_{k=0}^{l-1} (A-\alpha)^k M_l. \tag{3.196}$$

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 $(A-\alpha)^k u_j$, $k = 0, \dots, l-1$. Then $\mathbb{C}^n = V_1 \oplus \cdots \oplus V_m$ by construction of the sets M_k and each V_j has a cyclic vector u_j. In summary, we get

Theorem 3.32 (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* (3.192).

This also leads to the following algorithm for computing the Jordan canonical form:

(i) For every eigenvalue α_j compute a basis of generalized eigenvectors by solving $(A-\alpha_j)u = v$ recursively.

(ii) Pick a generalized eigenvector u of highest order k and choose $(A-\alpha_j)^l u$, $l = 0, \dots, k-1$, as new basis elements. Remove all generalized eigenvectors which are in the linear span of the already chosen ones and repeat the last step until no generalized eigenvectors are left.

Furthermore, from the Jordan canonical form we can read off that

$$\dim(\mathrm{Ker}(A-\alpha_j)^{d_j}) = a_j \tag{3.197}$$

and since $(A-\alpha_j)^{d_j}$ annihilates the Jordan block corresponding to α_j we see

$$\prod_{j=1}^{m} (A-\alpha_j)^{d_j} = 0. \tag{3.198}$$

In particular, since $1 \le d_j \le a_j$ we obtain

Theorem 3.33 (Cayley–Hamilton)**.** *Every matrix satisfies its own characteristic equation*

$$\chi_A(A) = 0. \tag{3.199}$$

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

$$A = \exp(B). \tag{3.200}$$

In this case we will call $B = \log(A)$ a **matrix logarithm** of A. 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,

$$\log(1+x) = \sum_{j=1}^{\infty} \frac{(-1)^{j+1}}{j} x^j, \qquad |x| < 1, \tag{3.201}$$

we set

$$\begin{aligned} B &= \log(\alpha)\mathbb{I} + \sum_{j=1}^{n-1} \frac{(-1)^{j+1}}{j\alpha^j} N^j \\ &= \begin{pmatrix} \log(\alpha) & \frac{1}{\alpha} & \frac{-1}{2\alpha^2} & \cdots & \frac{(-1)^n}{(n-1)\alpha^{n-1}} \\ & \log(\alpha) & \frac{1}{\alpha} & \ddots & \vdots \\ & & \log(\alpha) & \ddots & \frac{-1}{2\alpha^2} \\ & & & \ddots & \frac{1}{\alpha} \\ & & & & \log(\alpha) \end{pmatrix}. \end{aligned} \tag{3.202}$$

By construction we have $\exp(B) = A$. Note that B is not unique since different branches of $\log(\alpha)$ will give different matrices. Moreover, it might be complex even if A is real. In fact, if A has a negative eigenvalue, then $\log(\alpha) = \log(|\alpha|) + \mathrm{i}\pi$ implies that $\log(A)$ will be complex. We can avoid this situation by taking the logarithm of A^2.

Lemma 3.34. *A matrix A has a logarithm if and only if* $\det(A) \neq 0$. *Moreover, if A is real and all real eigenvalues are positive, then there is a real logarithm. In particular, if A is real we can find a real logarithm for A^2.*

Proof. Since the eigenvalues of A^2 are the squares of the eigenvalues of A (show this), it remains to show that B is real if all real eigenvalues are positive.

In this case only the Jordan block corresponding to complex eigenvalues could cause problems. We consider the real Jordan canonical form (3.25) and note that for

$$R = \begin{pmatrix} \mathrm{Re}(\alpha) & \mathrm{Im}(\alpha) \\ -\mathrm{Im}(\alpha) & \mathrm{Re}(\alpha) \end{pmatrix} = r \begin{pmatrix} \cos(\varphi) & -\sin(\varphi) \\ \sin(\varphi) & \cos(\varphi) \end{pmatrix}, \quad \alpha = r\mathrm{e}^{\mathrm{i}\varphi},$$

the logarithm is given by

$$\log(R) = \log(r)\mathbb{I} + \begin{pmatrix} 0 & -\varphi \\ \varphi & 0 \end{pmatrix}.$$

Now write the real Jordan block $R\mathbb{I} + N$ as $R(\mathbb{I} + R^{-1}N)$. Then one can check that

$$\log(R\mathbb{I} + N) = \log(R)\mathbb{I} + \sum_{j=1}^{n-1} \frac{(-1)^{j+1}}{j} R^{-j} N^j$$

is the required logarithm. □

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

$$(J - z)^{-1} = \frac{1}{\alpha - z} \sum_{j=0}^{n-1} \frac{1}{(z - \alpha)^j} N^j. \qquad (3.203)$$

In particular, note that the resolvent has a pole at each eigenvalue with the residue being the linear projector (cf. Problem 3.46) onto the corresponding generalized eigenspace.

For later use we also introduce the subspaces

$$\begin{aligned} E^{\pm}(A) &= \bigoplus_{|\alpha_j|^{\pm 1} < 1} \operatorname{Ker}(A - \alpha_j)^{a_j}, \\ E^{0}(A) &= \bigoplus_{|\alpha_j| = 1} \operatorname{Ker}(A - \alpha_j)^{a_j}, \end{aligned} \qquad (3.204)$$

where α_j are the eigenvalues of A and a_j are the corresponding algebraic multiplicities. The subspaces $E^+(A)$, $E^-(A)$, $E^0(A)$ are called **contracting**, **expanding**, **unitary subspace** of A, respectively. For each of these subspaces we can define the corresponding **projections** $P^+(A)$, $P^0(A)$, $P^-(A)$ as the linear projections whose image is the corresponding subspace and whose kernel is the direct sum of the other two subspaces. The restriction of A to these subspaces is denoted by A_+, A_-, A_0, respectively.

Problem 3.45. *Let $f(z)$ be a function analytic in a disc around 0 of radius R:*

$$f(z) = \sum_{k=0}^{\infty} \frac{f^{(k)}(0)}{k!} z^k, \qquad |z| < R. \qquad (3.205)$$

Show that the corresponding power series for $f(A)$ converges if $r(A) = \max_j\{|\alpha_j|\} < R$. In particular, show that for one Jordan block $J = \alpha\mathbb{I} + N$

the result is

$$f(J) = \begin{pmatrix} f(\alpha) & f'(\alpha) & \frac{f''(\alpha)}{2!} & \dots & \frac{f^{(k-1)}(\alpha)}{(k-1)!} \\ & f(\alpha) & f'(\alpha) & \ddots & \vdots \\ & & f(\alpha) & \ddots & \frac{f''(\alpha)}{2!} \\ & & & \ddots & f'(\alpha) \\ & & & & f(\alpha) \end{pmatrix}. \tag{3.206}$$

Problem 3.46. *A* **linear projection** $P : \mathbb{C}^n \to \mathbb{C}^n$ *is a linear map satisfying* $P^2 = P$*. Show that the kernel and range of* P *are complementary subspaces:* $\mathrm{Ker}(P) \oplus \mathrm{Ran}(P) = \mathbb{C}^n$*. Show that* $\mathbb{I} - P$ *is also a projection with* $\mathrm{Ker}(\mathbb{I} - P) = \mathrm{Ran}(P)$ *and* $\mathrm{Ran}(\mathbb{I} - P) = \mathrm{Ker}(P)$*. Show that* P *has the only possible eigenvalues* 0 *and* 1*. Show that given two complementary subspace* $U \oplus V = \mathbb{C}^n$ *there is a unique projection* P *with* $\mathrm{Ker}(P) = U$ *and* $\mathrm{Ran}(P) = V$*.*

Problem 3.47. *Suppose* $A(\lambda)$ *is* C^k *and has no unitary subspace. Then the projectors* $P^{\pm}(A(\lambda))$ *onto the contracting, expanding subspace are given by*

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

In particular, conclude that they are C^k*. (Hint: Jordan canonical form and* (3.203)*.)*

Problem 3.48. *Denote by* $r(A) = \max_j\{|\alpha_j|\}$ *the* **spectral radius** *of* A*. Show that for every* $\varepsilon > 0$ *there is a norm* $\|.\|_\varepsilon$ *on* $\mathbb{C}^n$ *such that*

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

(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 3.49. *Show that for a symmetric matrix* A *the norm is equal to the spectral radius* $r(A) = \max_j\{|\alpha_j|\}$*. Show that for an arbitrary matrix* $\|A\|^2 = \|A^*A\| = r(A^*A)$*. (Hint: Observe that since* A^*A *is symmetric we obtain* $\|A^*A\| = \max_{|x|=1} x \cdot A^*Ax = \max_{|x|=1} |Ax|^2$*, where* $x \cdot y = \sum_{j=1}^n x_j^* y_j$ *denotes the scalar product in* $\mathbb{C}^n$*.)*

Chapter 4

Differential equations in the complex domain

This chapter requires some basic knowledge from complex analysis. Readers only interested in dynamical systems can skip this and the next chapter and go directly to Chapter 6.

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.

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** (complex differentiable) in Ω and consider the equation

$$w' = f(z, w), \qquad w(z_0) = w_0. \tag{4.2}$$

Here the prime denotes complex differentiation,

$$w'(z_0) = \frac{dw(z_0)}{dz} = \lim_{z \to z_0} \frac{w(z) - w(z_0)}{z - z_0}, \tag{4.3}$$

and hence the equation only makes sense if $w(z)$ is analytic as well.

We recall that the existence of the complex derivative is a much stronger condition than existence of the real derivative. In fact, it implies that $w(z)$ can be expanded in a convergent **power series** near z_0:

$$w(z) = \sum_{j=0}^{\infty} \frac{w^{(j)}(z_0)}{j!}(z-z_0)^j, \quad w^{(j)}(z_0) = \frac{d^j w(z_0)}{dz^j}. \tag{4.4}$$

By the Cauchy–Hadamard theorem the **radius of convergence** of this series is given by

$$R^{-1} = \limsup_{j\to\infty} |w_j|^{1/j}, \qquad w_j = \frac{w^{(j)}(z_0)}{j!}. \tag{4.5}$$

If $f(w) = f(w_1, \dots, w_n)$ depends on more than one variable, it is called analytic if the partial complex derivatives

$$\frac{\partial}{\partial w_j} f(w), \qquad 1 \le j \le n, \tag{4.6}$$

exist (in the complex sense as defined in (4.3)). Again it can be shown that $f(w)$ can be expanded in a convergent power series. However, we will not need this result here. Just observe that the definition implies that if $f(z,w)$ is analytic in the $n+1$ variables (z,w) and $w(z)$ is analytic in the single variable z, then $f(z,w(z))$ is analytic in the single variable z by the chain rule.

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

$$w(z) = w_0 + \int_{z_0}^{z} f(\zeta, w(\zeta)) d\zeta. \tag{4.7}$$

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

$$\sup_{|z-z_0|<\varepsilon} |w(z)| \tag{4.8}$$

since the (locally) uniform limit of a sequence of analytic functions is again analytic by the Weierstraß convergence theorem. Now we can proceed as in the real case to obtain

Theorem 4.1. *Let $\Omega = \{(z,w) | |z-z_0| < \varepsilon, |w-w_0| < \delta\}$ be an open rectangle and suppose $f : \Omega \to \mathbb{C}$ is analytic and bounded. Then the initial value problem* (4.2) *has a unique analytic solution defined in the disc $\{z| |z-z_0| < \varepsilon_0\}$, where*

$$\varepsilon_0 = \min(\varepsilon, \frac{\delta}{M}), \qquad M = \sup_{(z,w)\in\Omega} |f(z,w)|. \tag{4.9}$$

Example. The following example shows that the estimates for the convergence radius ε_0 of the solution cannot be improved in general (of course it cannot be larger than ε in general). Consider

$$w' = M\left(\frac{1}{2}\left(1+\frac{w}{\delta}\right)\right)^{1/m}, \qquad M, \delta > 0,\, m > 1.$$

Observe that the right-hand side satisfies the assumptions of our theorem with $\varepsilon = \infty$ and the constants M, δ are equal to the ones used in the differential equation above.

The solution corresponding to the initial condition $w(0) = 0$ can be obtained by separation of variables and is given by

$$w(z) = \delta\left(\left(1+\frac{z}{a_m}\right)^{m/(m-1)} - 1\right), \quad a_m = \left(\frac{m2^{1/m}}{m-1}\right)\frac{\delta}{M} > 0.$$

The solution has a branch point at $z = -a_m$ and hence the convergence radius around zero is a_m. Finally observe that $a_m \to \varepsilon_0 = \frac{\delta}{M}$ as $m \to \infty$. ⋄

Note that we even get analytic dependence on the initial condition and on parameters.

Theorem 4.2. *Suppose $f : \Omega \times \Lambda \to \mathbb{C}$ is analytic. Then the initial value problem*

$$w' = f(z,w,\lambda), \qquad w(z_0) = w_0, \tag{4.10}$$

has a unique solution $w(z, w_0, \lambda)$ defined in a sufficiently small neighborhood around $(z_0, w_0, \lambda_0) \in \Omega \times \Lambda$ which is analytic with respect to all variables.

Proof. This follows again from the Weierstraß convergence theorem since the Picard iterates are analytic together with the fact that constants in the convergence estimates can be chosen uniform in some sufficiently small compact neighborhood around (z_0, w_0, λ_0) (cf. the proof of Theorem 2.9). □

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. If they coincide at a point $z_1 \in U_1 \cap U_2$, they also coincide in a neighborhood of z_1

by our local uniqueness result. Hence the set where they coincide is open. By continuity of $w_j(z)$ it is also closed (in the relative topology) and hence both solutions coincide on the connected component of $U_1 \cap U_2$ containing z_1. But this is all we can say in general as the following example shows.

Example. Consider

$$w' = \frac{1}{z}, \qquad w(1) = 0, \qquad z \in \mathbb{C}\backslash\{0\}. \tag{4.11}$$

The solution is given by

$$w(z) = \log(z) = \log|z| + \mathrm{i}\arg(z) \tag{4.12}$$

and different choices of the branch cut (i.e., the half-ray along which $\arg(z)$ will jump by 2π) will give different solutions. In particular, note that there is no unique maximal domain of definition. ◇

Finally, let us show how analyticity can be used in the investigation of a simple differential equation.

Example. Consider

$$w' + w^2 = z, \qquad w(0) = w_0. \tag{4.13}$$

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

$$w(z) = \sum_{j=0}^{\infty} w_j z^j, \qquad w'(z) = \sum_{j=0}^{\infty} j w_j z^{j-1}, \tag{4.14}$$

and plugging this into our equation yields

$$\sum_{j=0}^{\infty} j w_j z^{j-1} + \left(\sum_{j=0}^{\infty} w_j z^j \right)^2 = z. \tag{4.15}$$

Expanding the product (using the Cauchy product formula) and aligning powers of z gives

$$\sum_{j=0}^{\infty} \left((j+1) w_{j+1} + \sum_{k=0}^{j} w_k w_{j-k} \right) z^j = z. \tag{4.16}$$

Comparing powers of z we obtain

$$w_1 = -w_0^2, \; w_2 = w_0^3 + \frac{1}{2}, \quad w_{j+1} = \frac{-1}{j+1} \sum_{k=0}^{j} w_k w_{j-k}. \tag{4.17}$$

Hence we have at least found a recursive formula for computing the coefficients of the power series of the solution. ◇

In general one obtains the following result by differentiating the differential equation:

Theorem 4.3. *Suppose $f : \Omega \to \mathbb{C}$ is analytic. Then the expansion coefficients in the power series* (4.4) *of the solution $w(z)$ for the initial value problem* (4.2) *can be found recursively via*

$$w^{(j)}(z_0) = f^j(z_0, w(z_0), \dots, w^{(j-1)}(z_0)), \tag{4.18}$$

where the function f^j is recursively defined via

$$\begin{aligned} f^{j+1}(z, w^{(0)}, \dots, w^{(j)}) &= \frac{\partial f^j}{\partial z}(z, w^{(0)}, \dots, w^{(j-1)}) \\ &\quad + \sum_{k=0}^{j-1} \frac{\partial f^j}{\partial w^{(k)}}(z, w^{(0)}, \dots, w^{(j-1)}) w^{(k+1)}, \\ f^1(z, w^{(0)}) &= f(z, w^{(0)}). \end{aligned} \tag{4.19}$$

However, this procedure gets too cumbersome if the function f involves w in a too complicated way. Hence we will only investigate the case of linear equations further. But, to make things a bit more exciting, we will allow for poles in the coefficients, which is often needed in applications. In fact, this will eventually allow us to *solve* the Riccati equation from the last example using special functions (Problem 4.13).

Problem 4.1. *Make a power series ansatz for the following equations:*

(i) $w' + w = z, \qquad w(0) = w_0.$

(ii) $w' + w^2 = z^2, \qquad w(0) = w_0.$

(iii) $w' + w = \frac{1}{1-z}, \qquad w(0) = w_0.$

Problem 4.2. *Try to find a solution of the initial value problem*

$$w'' = (z^2 - 1)w, \qquad w(0) = 1,\ w'(0) = 0,$$

by using the power series method from above. Can you find a closed form for the solution? What is a second solution? (Hint: Problem 3.34)

Problem 4.3. *Make a power series ansatz for the differential equation*

$$z^2 w' = w - z.$$

What is the radius of convergence of the resulting series?

Problem 4.4. *Consider* (4.2) *at $z_0 = 0$. Show that the power series for the n'th Picard iteration and the solution coincide up to order n. This* *can be used to derive an effective numerical scheme known as the Parker–Sochacki algorithm. (Hint: Let $w_n(z)$ be the Picard iterates and suppose $w(z) = w_n(z) + O(z^{n+1})$. What does the Lipschitz estimate tell you about the relation between $f(z, w(z))$ and $f(z, w_n(z))$?)*

4.2. The Frobenius method for second-order equations

To begin with, we will restrict our attention to second-order linear equations

$$u'' + p(z)u' + q(z)u = 0, \tag{4.20}$$

which are among the most important ones in applications. Clearly, everything we know from the real case (superposition principle, etc.) carries over to the complex case and we know that the solutions are analytic whenever the coefficients $p(z)$ and $q(z)$ are. However, in many 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. We will assume that the singular point is $z_0 = 0$ for notational convenience.

Recall that a function $u(z)$, which is analytic in the domain $\Omega = \{z \in \mathbb{C} \,|\, 0 < |z| < r\}$, can be expanded into a (convergent) **Laurent series**

$$u(z) = \sum_{j\in\mathbb{Z}} u_j z^j, \qquad z \in \Omega. \tag{4.21}$$

It is analytic at $z = 0$ if all negative coefficients u_j, $j < 0$, vanish. If all but finitely many vanish, $u(z)$ is said to have a **pole**. The smallest n with $u_{-m} = 0$ for $m > n$ is called the **order** of the pole. Otherwise, if infinitely many negative coefficients are nonzero, $z = 0$ is called an **essential singularity**.

Now let us begin by considering the prototypical example.

Example. The **Euler equation** is given by

$$u'' + \frac{p_0}{z}u' + \frac{q_0}{z^2}u = 0, \qquad z \in \mathbb{C}\backslash\{0\}. \tag{4.22}$$

Obviously the coefficients have poles 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 (4.22) we will use the transformation

$$\zeta = \log(z) = \log|z| + \mathrm{i}\arg(z), \qquad -\pi < \arg(z) < \pi, \tag{4.23}$$

which maps Ω to the strip $\tilde{\Omega} = \{z \in \mathbb{C} | -\pi < \mathrm{Im}(z) < \pi\}$. The equation in the new coordinates reads

$$\omega'' + (p_0 - 1)\omega' + q_0\omega = 0, \qquad \omega(\zeta) = u(\mathrm{e}^\zeta). \tag{4.24}$$

Since it has constant coefficients, a basis of solutions can be given in terms of the characteristic eigenvalues

$$\alpha_{1,2} = \frac{1}{2}(1 - p_0 \pm \sqrt{(p_0-1)^2 - 4q_0}) \tag{4.25}$$

according to Theorem 3.7. If they are different, $\alpha_1 \neq \alpha_2$, we have two linearly independent solutions

$$u_1(z) = z^{\alpha_1}, \qquad u_2(z) = z^{\alpha_2} \tag{4.26}$$

and if they are equal, $\alpha_1 = \alpha_2$, we have two linearly independent solutions

$$u_1(z) = z^{\alpha_1}, \qquad u_2(z) = \log(z) z^{\alpha_1}. \tag{4.27}$$

⋄

Now let us turn to the general case. As a warm up, we will look at first-order equations.

Lemma 4.4. *The first-order equation*

$$u' + p(z)u = 0 \tag{4.28}$$

has a solution of the form

$$u(z) = z^\alpha h(z), \qquad h(z) = \sum_{j=0}^\infty h_j z^j, \quad h_0 = 1, \tag{4.29}$$

if and only if $p(z)$ has at most a first-order pole. In this case we have $\alpha = -\lim_{z\to 0} z\,p(z)$ and the radius of convergence for the power series of $h(z)$ and the Laurent series of $p(z)$ are the same.

Proof. If $p(z) = \frac{p_0}{z} + p_1 + p_2 z + \dots$ has a first-order pole, the solution of the above equation is explicitly given by (cf. (1.38))

$$\begin{aligned} u(z) &= \exp\left(-\int^z p(t)dt\right) = \exp\left(-p_0 \log(z) + c - p_1 z + \dots\right) \\ &= z^{-p_0} \exp\left(c - p_1 z + \dots\right). \end{aligned}$$

Conversely we have

$$p(z) = -\frac{u'(z)}{u(z)} = -\frac{\alpha}{z} - \frac{h'(z)}{h(z)}.$$

□

Now we are ready for our second-order equation (4.20). Motivated by our example, we will assume that the coefficients are of the form

$$p(z) = \frac{1}{z}\sum_{j=0}^\infty p_j z^j, \qquad q(z) = \frac{1}{z^2}\sum_{j=0}^\infty q_j z^j, \tag{4.30}$$

and we will search for a solution of the form

$$u(z) = z^{\alpha} h(z), \tag{4.31}$$

where $\alpha \in \mathbb{C}$ and $h(z)$ is analytic near $z = 0$ with $h(0) = 1$. This is the generalized power series method, or Frobenius method.

Using our ansatz we obtain

$$q(z)u(z) = \frac{1}{z^2}\sum_{k=0}^{\infty} q_k z^k \sum_{j=0}^{\infty} h_j z^{\alpha+j} = z^{\alpha-2}\sum_{j=0}^{\infty}\sum_{k=0}^{j} q_k h_{j-k} z^j, \tag{4.32}$$

$$\begin{aligned} p(z)u'(z) &= \frac{1}{z}\sum_{k=0}^{\infty} p_k z^k \sum_{j=0}^{\infty}(\alpha+j)h_j z^{\alpha+j-1} \\ &= z^{\alpha-2}\sum_{j=0}^{\infty}\sum_{k=0}^{j}(\alpha+j-k)p_k h_{j-k} z^j, \end{aligned} \tag{4.33}$$

$$u''(z) = z^{\alpha-2}\sum_{j=0}^{\infty}(\alpha+j)(\alpha+j-1)h_j z^j. \tag{4.34}$$

Plugging this into (4.20) and comparing coefficients we obtain

$$\Big((\alpha+j)^2+(p_0-1)(\alpha+j)+q_0\Big)h_j+\sum_{k=1}^{j}\big((\alpha+j-k)p_k+q_k\big)h_{j-k}=0. \tag{4.35}$$

Since $h_0 = 1$, this gives for $j = 0$ the **indicial equation**

$$\alpha^2 + (p_0 - 1)\alpha + q_0 = 0. \tag{4.36}$$

Hence the possible choices for α are the **characteristic exponents**

$$\alpha_{1,2} = \frac{1}{2}(1 - p_0 \pm \sqrt{(p_0-1)^2 - 4q_0}). \tag{4.37}$$

Here we will take the standard branch of the root (with branch cut along the negative real axis), such that $\mathrm{Re}(\alpha_1) \geq \mathrm{Re}(\alpha_2)$. Using

$$\alpha^2 + (p_0 - 1)\alpha + q_0 = (\alpha - \alpha_1)(\alpha - \alpha_2) \tag{4.38}$$

we obtain in the case $\alpha = \alpha_1$

$$h_j = \frac{-1}{(\alpha_1 - \alpha_2 + j)j}\sum_{k=1}^{j}\big((\alpha_1+j-k)p_k+q_k\big)h_{j-k}, \quad j > 0, \tag{4.39}$$

which is always solvable since $\mathrm{Re}(\alpha_1 - \alpha_2) \geq 0$ by assumption. In the case $\alpha = \alpha_2$ we obtain

$$h_j = \frac{-1}{(\alpha_2 - \alpha_1 + j)j}\sum_{k=1}^{j}\big((\alpha_2+j-k)p_k+q_k\big)h_{j-k}, \tag{4.40}$$

which might have a problem at $j = m$ if $\alpha_1 = \alpha_2 + m$ for some $m \in \mathbb{N}_0$.

In this case, h_j, $1 \le j \le m-1$, are uniquely determined by our choice $h_0 = 1$, whereas for $j = m$ we obtain

$$0 = \sum_{k=1}^{m} \big((\alpha_1 - k)p_k + q_k\big) h_{m-k}. \tag{4.41}$$

If this equation is fulfilled, we can choose h_m as we like (this freedom reflects the fact that we can add an arbitrary multiple of u_1 to u_2) and the remaining h_j, $j > m$, are again determined recursively. Otherwise there is no solution of the form $z^{\alpha_2} h(z)$.

Hence we need a different ansatz in this last case. To find the form of the second solution we use the variation of constants ansatz (compare Section 3.5)

$$u_2(z) = c(z) u_1(z) = c(z) z^{\alpha_1} h_1(z). \tag{4.42}$$

Then

$$c''(z) + \Big(2\frac{\alpha_1}{z} + 2\frac{h_1'(z)}{h_1(z)} + p(z)\Big) c'(z) = 0, \tag{4.43}$$

where

$$\Big(2\frac{\alpha_1}{z} + 2\frac{h_1'(z)}{h_1(z)} + p(z)\Big) = \frac{1 - \alpha_2 + \alpha_1}{z} + 2h_1'(0) + p_1 + \dots \tag{4.44}$$

Hence, by Lemma 4.4,

$$c'(z) = z^{\alpha_2 - \alpha_1 - 1} \sum_{j=0}^{\infty} c_j z^j, \quad c_0 \ne 0. \tag{4.45}$$

Integrating once we obtain (neglecting the integration constant)

$$c(z) = z^{\alpha_2 - \alpha_1} \sum_{j=0}^{\infty} \frac{c_j}{\alpha_2 - \alpha_1 + j} z^j, \tag{4.46}$$

if $\alpha_1 - \alpha_2 \not\in \mathbb{N}_0$ and

$$c(z) = z^{\alpha_2 - \alpha_1} \sum_{j=0, j\ne m}^{\infty} \frac{c_j}{\alpha_2 - \alpha_1 + j} z^j + c_m \log(z), \tag{4.47}$$

if $\alpha_1 - \alpha_2 = m \in \mathbb{N}_0$. In the latter case c_m could be zero unless $m = 0$.

In summary we have:

Theorem 4.5 (Fuchs). *Suppose the coefficients $p(z)$ and $q(z)$ of the second order equation* (4.20) *have poles of order (at most) one and two respectively. Then, if α_1, α_2 are the characteristic exponents defined in* (4.37) *and ordered according to* $\mathrm{Re}(\alpha_1) \ge \mathrm{Re}(\alpha_2)$, *two cases can occur:*

Case 1. If $\alpha_1 - \alpha_2 \not\in \mathbb{N}_0$, a fundamental system of solutions is given by

$$u_j(z) = z^{\alpha_j} h_j(z), \quad j = 1, 2, \tag{4.48}$$

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

$$\begin{aligned} u_1(z) &= z^{\alpha_1} h_1(z), \\ u_2(z) &= z^{\alpha_2} h_2(z) + c \log(z) u_1(z), \end{aligned} \tag{4.49}$$

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 $m = 0$.

Moreover, in both cases the radius of convergence of the power series for $h_1(z)$ and $h_2(z)$ is at least equal to the minimum of the radius of convergence for $p(z)$ and $q(z)$.

Proof. Since u_1 and u_2 are clearly linearly independent, the only item remaining is to show that the power series for $h_1(z)$ has a nonzero radius of convergence. Let h_j be the coefficients defined via (4.39) and let $R > 0$ be smaller than the radius of convergence of the series for $p(z)$ and $q(z)$. We will show that $|h_j| R^j \le C$ for some $C > 0$.

Abbreviate

$$P = \sum_{j=1}^{\infty} |p_j| R^j, \qquad Q = \sum_{j=1}^{\infty} |q_j| R^j.$$

Then there is a $j_0 > 0$ such that

$$\frac{(|\alpha_1| + j)P + Q}{(\mathrm{Re}(\alpha_1 - \alpha_2) + j)j} \le 1$$

for $j > j_0$. Choose $C = \max_{0 \le j \le j_0} |h_j| R^j$. Then the claim holds for $j \le j_0$ and we can proceed by induction: Suppose it holds up to $j - 1$. Then we obtain from (4.39)

$$\begin{aligned} |h_j| R^j &\le \frac{1}{(\mathrm{Re}(\alpha_1 - \alpha_2) + j)j} \sum_{k=1}^{j} \big((|\alpha_1| + j)|p_k| + |q_k|\big) C\, R^k \\ &\le \frac{(|\alpha_1| + j)P + Q}{(\mathrm{Re}(\alpha_1 - \alpha_2) + j)j} C \le C, \end{aligned}$$

which proves the claim. □

For the practical application of this result it remains to discuss the case $\alpha_1 - \alpha_2 = m \in \mathbb{N}_0$. One option is to use the variation of constants ansatz (4.42). However, unless one is able to find a closed form for the power series of the quotient $\frac{h_1'(z)}{h(z)}$ it might be better to work directly with the ansatz

$$u_2(z) = \hat{u}_2(z) + c \log(z) u_1(z), \qquad \hat{u}_2(z) = z^{\alpha_2} h_2(z), \tag{4.50}$$

from our theorem. Inserting this ansatz into our differential equation we obtain

$$\hat{u}_2''(z)+p(z)\hat{u}_2'(z)+q(z)\hat{u}_2(z)=-c\left(\frac{2}{z}u_1'(z)+\big(\frac{p(z)}{z}-\frac{1}{z^2}\big)u_1(z)\right), \quad (4.51)$$

where the logarithmic terms cancel since u_1 solves our equation. For the generalized power series of the expression on the right-hand side we obtain

$$-\,c\,z^{\alpha_2-2}\sum_{j=m}^{\infty}\left((2j-m)h_{1,j-m}+\sum_{k=1}^{j-m}p_k h_{1,j-m-k}\right)z^j. \quad (4.52)$$

Now comparing powers between both sides (for the left-hand sides the coefficients are given by (4.35) with $\alpha=\alpha_2$) we obtain the following cases: For $j<m$ the right-hand side does not contribute and thus $h_{2,j}$, $1\le j<m$, are uniquely determined by $h_{2,0}=1$ and

$$h_{2,j}=\frac{-1}{(j-m)j}\sum_{k=1}^{j}\big((\alpha_2+j-k)p_k+q_k\big)h_{2,j-k}. \quad (4.53)$$

At $j=m$ we obtain

$$\sum_{k=1}^{m}\big((\alpha_1-k)p_k+q_k\big)h_{2,m-k}=-c\,m. \quad (4.54)$$

If $m=0$ this equation is trivially satisfied and we can choose any (nonzero) c. Otherwise we obtain the unique value

$$c=-\frac{1}{m}\sum_{k=1}^{m}\big((\alpha_1-k)p_k+q_k\big)h_{2,m-k}, \qquad m\in\mathbb{N}. \quad (4.55)$$

Finally, for $j>m$ we obtain

$$\begin{aligned}h_{2,j}=&\frac{-1}{(j-m)j}\sum_{k=1}^{j}\big((\alpha_2+j-k)p_k+q_k\big)h_{2,j-k}\\&-c\left((2j-m)h_{1,j-m}+\sum_{k=1}^{j-m}p_k h_{1,j-m-k}\right)\end{aligned} \quad (4.56)$$

which determines the remaining coefficients uniquely once a value for $h_{2,m}$ is chosen.

Furthermore, the conditions on p and q are optimal:

Theorem 4.6 (Fuchs). *The equation* (4.20) *has two solutions* $u_1(z)$, $u_2(z)$ *as in the previous theorem if and only if* $p(z)$ *and* $zq(z)$ *have at most first-order poles.*

Proof. Consider $v(z) = (u_2(z)/u_1(z))'$ and observe that it is of the form $v(z) = z^\beta k(z)$, where $k(z)$ is analytic near $z = 0$.

Now a straightforward calculation shows

$$p(z) = -\frac{v'(z)}{v(z)} - 2\frac{u_1'(z)}{u_1(z)}$$

and since the right-hand side has at most a first-order pole, so does p. Similarly,

$$q(z) = -\frac{u_1''(z)}{u_1(z)} - p(z)\frac{u_1'(z)}{u_1(z)}$$

has at most a second-order pole. □

Note that (3.113) implies that $p(z)$ and $q(z)$ will be holomorphic near $z = 0$ if and only if there are two linearly independent holomorphic solutions $u_1(z)$ and $u_2(z)$.

Finally, let me remark that this characterization can also be applied to classify singularities at $z_0 = \infty$. To this end one makes the change of variables $\zeta = \frac{1}{z}$ which transforms our equation to

$$\omega'' + \left(2\zeta^{-1} - \zeta^{-2}p(\zeta^{-1})\right)\omega' + \zeta^{-4}q(\zeta)^{-1}\omega = 0, \quad \omega(\zeta) = u(\zeta^{-1}). \tag{4.57}$$

In particular, the equation will satisfy (4.30) in the new variable ζ if and only if the following limits

$$2 - \lim_{z\to\infty} z\,p(z) = p_0, \qquad \lim_{z\to\infty} z^2\,q(z) = q_0 \tag{4.58}$$

exist in $\mathbb{C}$. 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.

Example. Consider the famous **Bessel equation**

$$z^2u'' + zu' + (z^2 - \nu^2)u = 0, \qquad \nu \in \mathbb{C}. \tag{4.59}$$

After dividing by z^2 we see that it is of the form (4.20) with $p(z) = \frac{1}{z}$ and $q(z) = 1 - \frac{\nu^2}{z^2}$. In particular, $p_0 = 1$ and $q_0 = -\nu^2$. Moreover, it is no restriction to assume $\mathrm{Re}(\nu) \geq 0$ and hence we will do so.

The characteristic exponents are given by $\alpha_{1,2} = \pm\nu$ and hence there is a solution of the form

$$u_1(z) = z^\nu \sum_{j=0}^{\infty} h_{1,j} z^j, \qquad h_{1,0} = 1. \tag{4.60}$$

Plugging this into our equation yields

$$z^2 \sum_{j=0}^{\infty} h_{1,j}(j+\nu-1)(j+\nu) z^{j+\nu-2} + z \sum_{j=0}^{\infty} h_{1,j}(j+\nu) z^{j+\nu-1} + (z^2-\nu^2) \sum_{j=0}^{\infty} h_{1,j} z^{j+\nu} = 0 \qquad (4.61)$$

and after multiplying by $z^{-\nu}$ and aligning powers of z

$$\sum_{j=0}^{\infty} \left(h_{1,j}(j+\nu-1)(j+\nu) + h_{1,j}(j+\nu) + h_{1,j-2} - h_{1,j}\nu^2 \right) z^j = 0, \quad (4.62)$$

where we set $h_{1,j} = 0$ for $j < 0$. Comparing coefficients we obtain the recurrence relation

$$j(j+2\nu) h_{1,j} + h_{1,j-2} = 0 \qquad (4.63)$$

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

$$h_{1,2j} = \frac{(-1)^j}{4^j j! (\nu+1)_j}, \qquad h_{2j+1} = 0, \qquad (4.64)$$

where we have used the **Pochhammer symbol**

$$(x)_0 = 1, \quad (x)_j = x(x+1)\cdots(x+j-1) = \frac{\Gamma(x+j)}{\Gamma(x)}. \qquad (4.65)$$

Here $\Gamma(x)$ is the usual **Gamma function** (cf. Problem 4.5). This solution, with a different normalization, is called **Bessel function**

$$J_\nu(z) = \frac{u_1(z)}{2^\nu \Gamma(\nu+1)} = \sum_{j=0}^{\infty} \frac{(-1)^j}{j! \Gamma(\nu+j+1)} \left(\frac{z}{2}\right)^{2j+\nu} \qquad (4.66)$$

of order ν. Now what about the second solution? So let us investigate the equation for $-\nu$. Replacing ν by $-\nu$ in the previous calculation, we see that we can find a second (linearly independent) solution $J_{-\nu}(z)$ provided $\nu \neq 0$ and $(-\nu+1)_j \neq 0$ for all j, that is, provided $\nu \notin \mathbb{N}_0$. 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 $\nu = n \in \mathbb{N}_0$. We begin with the case $n \in \mathbb{N}$. All odd coefficients must be zero and the recursion for the even ones gives us a contradiction at $j = 2n$. Hence the only possibility left is a logarithmic solution

$$u_2(z) = z^{-n} h_2(z) + c \log(z) u_1(z). \qquad (4.67)$$

Inserting this into our equation yields

$$j(j-2n) h_{2,j} + h_{2,j-2} = -2c(j-n) h_{1,j-2n}, \qquad (4.68)$$

where we again set $h_{2,j} = 0$ for $j < 0$. 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

$$h_{2,2j} = \frac{1}{4^j j!(n-1)_j}, \qquad j < n. \tag{4.69}$$

The recursion (4.68) for $j = 2n$ reads $h_{2,2(n-1)} = -2cn$ from which

$$c = \frac{-2}{4^n n!(n-1)!} \tag{4.70}$$

follows. The remaining coefficients now follow recursively from

$$4j(j+n)h_{2,2j+2n} + h_{2,2(j-1)+2n} = -2c(2j+n)h_{1,2j} \tag{4.71}$$

once we choose a value for $h_{2,2n}$ (this freedom just reflects the fact that we can always add a multiple of $u_1(z)$ to $u_2(z)$). This is a first-order linear inhomogeneous recurrence relation with solution given by (see Problem 4.9 and note that the solution of the homogeneous equation is $h_{1,2j}$)

$$h_{2,2j+2n} = h_{1,2j}\left(h_{2,2n} - \frac{c}{2}\sum_{k=1}^{j}\frac{2k+n}{k(k+n)}\right). \tag{4.72}$$

Choosing $h_{2,2n} = -\frac{c}{2}H_n$, where

$$H_j = \sum_{k=1}^{j}\frac{1}{k} \tag{4.73}$$

are the **harmonic numbers**, we obtain

$$h_{2,2n+2j} = \frac{(-1)^j(H_{j+n}+H_j)}{4^{j+n}(n-1)!j!(j+n)!}. \tag{4.74}$$

Usually, the following linear combination

$$\begin{aligned}Y_n(z) =& -\frac{2^n(n-1)!}{\pi}u_2(z) + \frac{\gamma - \log(2)}{2^{n-1}\pi n!}u_1(z)\\ =&\frac{2}{\pi}(\gamma + \log(\frac{z}{2}))J_n(z) - \frac{1}{\pi}\sum_{j=0}^{n-1}\frac{(-1)^j(n-1)!}{j!(1-n)_j}\left(\frac{z}{2}\right)^{2j-n}\\ &-\frac{1}{\pi}\sum_{j=0}^{\infty}\frac{(-1)^j(H_{j+n}+H_j)}{j!(j+n)!}\left(\frac{z}{2}\right)^{2j+n}\end{aligned} \tag{4.75}$$

is taken as second independent solution. Here $\gamma = \lim_{j\to\infty}(H_j - \log(j)) \approx 0.577216$ is the **Euler–Mascheroni constant**.

So the only remaining case is $n = 0$. In this case the recursion does not give a contradiction at $j = 2n$ but we still need to take the logarithmic term in order to get a *different* solution. In particular, we still make the ansatz (4.67) and the recursion (4.68) remains valid for $n = 0$. However, in this case c will not follow from the recursion for $j = 2n$ (which just reads $0 = 0$) but can be chosen arbitrary. The recursion can be solved as before and (4.72) is still valid, that is,

$$h_{2,2j} = h_{1,2j}\,(1 - cH_j) = \frac{(-1)^j}{(j!)^2}(1 - cH_j). \tag{4.76}$$

Choosing $c = \frac{2}{\pi}$ the linear combination

$$\begin{aligned} Y_0(z) &= u_2(z) + \big(-1 + \frac{2}{\pi}(\gamma - \log(2))\big) u_1(z) \\ &= \frac{2}{\pi}(\gamma + \log(\frac{z}{2}))J_0(z) - \frac{2}{\pi}\sum_{j=0}^{\infty} \frac{(-1)^j H_j}{(j!)^2}\left(\frac{z}{2}\right)^{2j} \end{aligned} \tag{4.77}$$

will agree with (4.75) in the special case $n = 0$.

Finally, let me remark that one usually uses the **Hankel function**

$$Y_\nu(z) = \frac{\cos(\pi\nu)J_\nu(z) - J_{-\nu}(z)}{\sin(\pi\nu)} \tag{4.78}$$

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

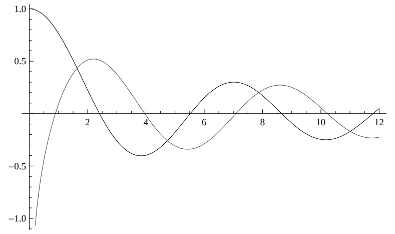
$$\begin{aligned} J_{-n}(z) &= \sum_{j=0}^{\infty} \frac{(-1)^j}{j!\Gamma(-n+j+1)}\left(\frac{z}{2}\right)^{2j-n} \\ &= \sum_{j=n}^{\infty} \frac{(-1)^j}{j!\Gamma(-n+j+1)}\left(\frac{z}{2}\right)^{2j-n} = (-1)^n J_n(z), \quad n \in \mathbb{N}_0, \end{aligned} \tag{4.79}$$

(here we used $\Gamma(-n+j+1)^{-1} = 0$ for $j = 0, 1, \dots, n-1$) it can be removed and it can be shown that the limit is a second linearly independent solution (Problem 4.10) which coincides with (4.75) 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}]
```

Out[1]=



◇

Problem 4.5. *The* **Gamma function** *is defined via*

$$\Gamma(z) = \int_0^\infty x^{z-1} \mathrm{e}^{-x} dx, \qquad \mathrm{Re}(z) > 0.$$

Verify that the integral converges and defines an analytic function in the indicated half plane. Use integration by parts to show

$$\Gamma(z+1) = z\Gamma(z), \qquad \Gamma(1) = 1.$$

Conclude $\Gamma(n) = (n-1)!$ *for* $n \in \mathbb{N}$. *Show that the relation* $\Gamma(z) = \Gamma(z+1)/z$ *can be used to define* $\Gamma(z)$ *for all* $z \in \mathbb{C}\backslash\{0, -1, -2, \dots\}$. *Show that near* $z = -n$, $n \in \mathbb{N}_0$, *the Gamma functions behaves like* $\Gamma(z) = \frac{(-1)^n}{n! z} + O(1)$.

Problem 4.6. *Show that the change of variables*

$$v(z) = \mathrm{e}^{\frac{1}{2}\int^z p(\zeta)d\zeta} u(z)$$

transforms (4.20) *into*

$$v'' + \left(q(z) - \frac{1}{2}p'(z) - \frac{1}{4}p(z)^2\right) v = 0.$$

Problem 4.7. *Solve the following differential equations by the Frobenius method:*

(i) $z\,u' + (1+z)u = 0$.

(ii) $u'' - 2u' + (1 + \frac{1}{4z^2})u = 0$.

(iii) $u'' + \frac{1-z}{z(1+z)}u' - \frac{1-z}{z(1+z)^2}u = 0$.

(iv) $z\,u'' + u' + u = 0$.

Problem 4.8. *Show that the coefficients of* $h(x)$ *from Lemma 4.4 are recursively given by*

$$h_j = \frac{1}{j}\sum_{k=0}^{j-1} p_{j-k} h_k,$$

if $p(z) = z^{-1}\sum_{j=0}^\infty p_j z^j$.

Problem 4.9. *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*

$$x_h(n) = x(0)\begin{cases} \prod\limits_{j=0}^{n-1} f(j), & n > 0, \\ 1, & n = 0, \\ \prod\limits_{j=n}^{-1} f(j)^{-1}, & n < 0, \end{cases}$$

Use a variation of constants ansatz for the inhomogeneous equation and show that the solution is given by

$$x(n) = x_h(n) + \begin{cases} x_h(n) \sum_{j=0}^{n-1} \frac{g(j)}{x_h(j+1)}, & n > 0, \\ 0, & n = 0, \\ -x_h(n) \sum_{j=n}^{-1} \frac{g(j)}{x_h(j+1)}, & n < 0. \end{cases}$$

Problem 4.10 (Hankel functions)**.** *Prove that the Hankel function is a second linearly independent solution for all ν as follows:*

(i) *Use* (4.79) *to prove that the Hankel function is well defined for all ν and analytic in both variables z and ν (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 second equality which is just **Euler's reflection formula** *for the Gamma function.)*

(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 4.11. *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 4.12. *Show*

$$\int_0^a J_\nu(z)^2 z\, dz = \frac{a^2}{2} J_\nu'(a)^2 + \frac{a^2 - \nu^2}{2} J_\nu(a)^2, \quad \nu \geq 0.$$

(Hint: Multiply Bessel's equation by $u'(z)$ and show that the result is a complete differential up to one term.)

Problem 4.13. *Many differential equations occur in practice that are not of the standard form* (4.59)*. 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 3.38.)

Problem 4.14 (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 4.15 (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 ∞*). Use the Frobenius method to show that*

$$F(a,b,c;z) = \sum_{j=0}^{\infty} \frac{(a)_j (b)_j}{(c)_j j!} z^j, \quad -c \notin \mathbb{N}_0,$$

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* $z^{1-c}F(a-c+1, b-c+1, 2-c; z)$ *is a second solution for* $c - 2 \notin \mathbb{N}_0$*. This gives two linearly independent solutions if* $c \notin \mathbb{Z}$*.*

Problem 4.16 (Confluent hypergeometric equation)**.** *The* **confluent hypergeometric equation** *is given by*

$$zw'' + (c-z)w' - aw = 0.$$

Classify all singular points (including ∞*). Use the Frobenius method to show that*

$$K(a,c;z) = \sum_{j=0}^{\infty} \frac{(a)_j}{(c)_j j!} z^j, \quad -c \notin \mathbb{N}_0,$$

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* $z^{1-c}K(a-c+$

$1, 2-c; z)$ *is a second solution for* $c-2 \notin \mathbb{N}_0$. *This gives two linearly independent solutions if* $c \notin \mathbb{Z}$.

Problem 4.17. *Show that any second-order equation* (4.20) *with finitely many singular points* $z_0, \dots, z_n, \infty$ *of Fuchs type is of the form*

$$p(z) = \sum_{j=0}^{n} \frac{p_j}{z-z_j}, \qquad q(z) = \sum_{j=0}^{n} \left(\frac{q_j}{(z-z_j)^2} + \frac{r_j}{z-z_j} \right),$$

where $p_j, q_j, r_j \in \mathbb{C}$ *and necessarily*

$$\sum_{j=0}^{n} r_j = 0.$$

Show that there is no singularity at ∞ *if in addition* $p_\infty = q_\infty = r_\infty = 0$, *where*

$$p_\infty = 2 - \sum_{j=0}^{n} p_j, \quad q_\infty = \sum_{j=0}^{n} (q_j + r_j z_j), \quad r_\infty = \sum_{j=0}^{n} z_j(2q_j + r_j z_j).$$

Problem 4.18 (Riemann equation)**.** *A second-order equation is called a* **Riemann equation** *if it has only three singular points (including* ∞*) of Fuchs type. Solutions of a Riemann equation are denoted by the* **Riemann symbol**

$$P \begin{Bmatrix} z_0 & z_1 & z_2 & \\ \alpha_1 & \beta_1 & \gamma_1 & z \\ \alpha_2 & \beta_2 & \gamma_2 & \end{Bmatrix},$$

where the numbers z_j *are the singular points and the numbers below* z_j *are the corresponding characteristic exponents.*

Recall that given distinct 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*

$$P \begin{Bmatrix} z_0 & z_1 & z_2 & \\ \alpha_1 & \beta_1 & \gamma_1 & z \\ \alpha_2 & \beta_2 & \gamma_2 & \end{Bmatrix} = P \begin{Bmatrix} 0 & 1 & \infty & \\ \alpha_1 & \beta_1 & \gamma_1 & \frac{(z-z_0)(z_1-z_2)}{(z-z_2)(z_1-z_0)} \\ \alpha_2 & \beta_2 & \gamma_2 & \end{Bmatrix}.$$

For the case $z_0 = 0$, $z_1 = 1$, $z_2 = \infty$ *show that*

$$p(z) = \frac{p_0}{z} + \frac{p_1}{z-1}, \qquad q(z) = \frac{q_0}{z^2} + \frac{r_0}{z} + \frac{q_1}{(z-1)^2} - \frac{r_0}{z-1}.$$

Express the coefficients $p(z)$ *and* $q(z)$ *in terms of the characteristic exponents and show that*

$$\alpha_1 + \alpha_2 + \beta_1 + \beta_2 + \gamma_1 + \gamma_2 = 1.$$

Conclude that a Riemann equation is uniquely determined by its symbol.

Finally, show

$$z^{\nu}(1-z)^{\mu} P\begin{Bmatrix} 0 & 1 & \infty & \\ \alpha_1 & \beta_1 & \gamma_1 & z \\ \alpha_2 & \beta_2 & \gamma_2 & \end{Bmatrix} = P\begin{Bmatrix} 0 & 1 & \infty & \\ \alpha_1+\nu & \beta_1+\mu & \gamma_1-\mu-\nu & z \\ \alpha_2+\nu & \beta_2+\mu & \gamma_2-\mu-\nu & \end{Bmatrix}$$

and conclude that any Riemann equation can be transformed into the hypergeometric equation

$$P\begin{Bmatrix} 0 & 1 & \infty & \\ 0 & 0 & a & z \\ 1-c & c-a-b & b & \end{Bmatrix}.$$

Show that the Legendre equation is a Riemann equation. Find the transformation which maps it to the hypergeometric equation.

4.3. Linear systems with singularities

Now we want to extend the results from the previous section to linear systems

$$w' = A(z)w, \qquad w(z_0) = w_0, \qquad z, z_0 \in \Omega \subseteq \mathbb{C}, \tag{4.80}$$

where $A(z)$ is a matrix whose coefficients are analytic in Ω.

As in the real case one can show that one can always extend solutions. However, extensions along different paths might give different solutions in general, as we have seen in example (4.11). These problems do not arise if Ω is simply connected.

Theorem 4.7. *Suppose $w' = A(z)w + b(z)$ is linear, where $A : \Omega \to \mathbb{C}^{n\times n}$ and $b : \Omega \to \mathbb{C}^n$ are analytic in a simply connected domain $\Omega \subseteq \mathbb{C}$. Then for every $z_0 \in \Omega$ the corresponding initial value problem has a unique solution defined on all of Ω.*

In particular, the power series for every solution will converge in the largest disc centered at z_0 and contained in Ω.

Proof. If Ω is a disc centered at z_0 the result follows as in Corollary 2.6. For general Ω, pick $z \in \Omega$ and let $\gamma : [0,1] \to \Omega$ be a path from z_0 to z. Around each point $\gamma(t)$ we have a solution in a ball with radius independent of the initial condition and of $t \in [0,1]$. So we can define the value of $w(z)$ by analytic continuation along the path γ. Since Ω is simply connected, this value is uniquely defined by the monodromy theorem. □

This result has the important consequence that a solution of a linear equation can have singularities (poles, essential singularities, or branch points) only at the points where the coefficients have isolated singularities.

That is, the singularities are fixed and do not depend on the initial condition. On the other hand, nonlinear equations will in general have **movable singularities**, as the simple example

$$w' = -w^2, \tag{4.81}$$

whose general solution is

$$w(z) = \frac{1}{z - z_0}, \tag{4.82}$$

shows. Equations whose only movable singularities are poles play an important role in applications. It can be shown that a first order equation

$$w' = f(z, w) \tag{4.83}$$

which is rational in w and meromorphic in z has this property if it is of Riccati type, that is, $f(z,w) = f_0(z) + f_1(z)w + f_2(z)w^2$, and can hence be transformed to a second order linear equation (cf. Problem 3.38). In the case of a second order equation

$$w'' = f(z, w, w') \tag{4.84}$$

which is rational in w, w' and meromorphic in z, Painlevé and his coworkers showed that there are six equations which cannot be linearized or solved by well-known special functions. These are nowadays known as the **Painlevé transcendents**. For example, the first two are given by

$$\begin{aligned} P_I: &\quad w'' = 6w^2 + z, \\ P_{II}: &\quad w'' = z\,w + 2w^3 + \alpha, \qquad \alpha \in \mathbb{C}. \end{aligned} \tag{4.85}$$

They play an important role in nonlinear physics just as special functions (like Bessel functions) play in linear physics. However, this is beyond this introduction, see for example the book by Ince [**23**], and we return to linear equations.

Again, 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 (4.80) is given by

$$w(z) = \Pi(z, z_0) w_0. \tag{4.86}$$

It is also not hard to see that Liouville's formula (3.91) extends to the complex case.

Again we will allow singularities at $z_0 = 0$. So let us start with the prototypical example. The system

$$w' = \frac{1}{z} A w, \qquad z \in \mathbb{C}\backslash\{0\}, \tag{4.87}$$

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 (4.87) we will use the transformation

$$\zeta = \log(z) = \log|z| + \mathrm{i}\arg(z), \qquad -\pi < \arg(z) < \pi, \tag{4.88}$$

which maps Ω to the strip $\tilde{\Omega} = \{z \in \mathbb{C}| -\pi < \mathrm{Im}(z) < \pi\}$. The equation in the new coordinates reads

$$\omega' = A\omega, \qquad \omega(\zeta) = w(\mathrm{e}^{\zeta}). \tag{4.89}$$

Hence a fundamental system is given by

$$W(z) = z^A = \exp(\log(z)A), \tag{4.90}$$

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

$$z^J = z^{\alpha}\begin{pmatrix} 1 & \log(z) & \frac{\log(z)^2}{2!} & \dots & \frac{\log(z)^{n-1}}{(n-1)!} \\ & 1 & \log(z) & \ddots & \vdots \\ & & 1 & \ddots & \frac{\log(z)^2}{2!} \\ & & & \ddots & \log(z) \\ & & & & 1 \end{pmatrix}. \tag{4.91}$$

Therefore the solution consists of terms of the form $z^{\alpha}\log(z)^k$, where α 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 4.8. *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*

$$W(z) = U(z)(z - z_0)^M, \tag{4.92}$$

where $U(z)$ is analytic in Ω.

Proof. Again we use our change of coordinates $\zeta = \log(z)$ to obtain

$$\omega' = \mathrm{e}^{\zeta}A(\mathrm{e}^{\zeta})\omega, \qquad \mathrm{Re}(\zeta) < \log(\varepsilon).$$

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 3.15). □

Observe that any other fundamental system $\tilde{W}(z)$ can be written as

$$\tilde{W}(z) = W(z)C = U(z)C\,(z - z_0)^{C^{-1}MC}, \qquad \det(C) \neq 0, \tag{4.93}$$

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 $(z-z_0)^M$. 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 $(z-z_0)^M$ term by using

$$W(z) = U(z)(z-z_0)^m\,(z-z_0)^{M-m\mathbb{I}}. \tag{4.94}$$

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** (or **weak singularity**) of our system if $A(z)$ has a pole of (at most) first order at z_0.

Theorem 4.9. *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 $W(z)$ is of the form* (4.92) *and $U(z)$ 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 m such that $\|A(z)\| \le \frac{m}{|z|}$ for $|z| \le r_0$. Using polar coordinates $z = r\mathrm{e}^{\mathrm{i}\varphi}$ we have

$$\begin{aligned} |w(r\mathrm{e}^{\mathrm{i}\varphi})| &= |w(r_0\mathrm{e}^{\mathrm{i}\varphi}) + \int_r^{r_0} A(s\mathrm{e}^{\mathrm{i}\varphi})w(s\mathrm{e}^{\mathrm{i}\varphi})\mathrm{e}^{\mathrm{i}\varphi}ds| \\ &\le |w(r_0\mathrm{e}^{\mathrm{i}\varphi})| + \int_r^{r_0} \frac{m}{s}|w(s\mathrm{e}^{\mathrm{i}\varphi})|ds \end{aligned}$$

for $0 < r \le r_0$. Applying Gronwall and taking the maximum over all φ we obtain

$$|w(z)| \le \sup_{\zeta:|\zeta|=r_0} |w(\zeta)| \left|\frac{r_0}{z}\right|^m,$$

which is the desired estimate. □

The converse of this result is in general not true (except in one dimension; cf. Lemma 4.4). However,

Lemma 4.10. *If z_0 is a regular singularity, then $A(z)$ has at most a pole at z_0.*

Proof. This follows from

$$A(z) = U'(z)U(z)^{-1} + \frac{1}{z-z_0}U(z)MU(z)^{-1},$$

since $\det(U(z))$ can have at most a finite order zero, and hence the entries of $U(z)^{-1}$ can have at most poles of the same order. □

There are no restrictions on the order of the pole as can be seen from the following

Example.

$$A(z) = \frac{1}{z}\begin{pmatrix} 0 & z^{-m} \\ z^m & m \end{pmatrix}, \quad U(z) = \begin{pmatrix} 1 & 0 \\ 0 & z^m \end{pmatrix}, \quad M = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}. \tag{4.95}$$

◇

Problem 4.19. *Let z_0 be a simple singularity and let $W(z)$ be a fundamental system as in* (4.92)*. Show that*

$$\det(W(z)) = (z - z_0)^{\operatorname{tr}(A_0)} d(z), \qquad d(z_0) \neq 0,$$

where $d(z)$ is analytic near z_0 and $A_0 = \lim_{z\to z_0}(z - z_0)A(z)$. Moreover, conclude that $\operatorname{tr}(A_0 - M) \in \mathbb{Z}$*. (Hint: Use Abel's identity* (3.91) *for the determinant.)*

4.4. 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 4.9, we can make the ansatz

$$W(z) = U(z) z^M, \qquad U(z) = \sum_{j=0}^{\infty} U_j z^j, \quad U_0 \neq 0. \tag{4.96}$$

Using

$$A(z) = \frac{1}{z}\sum_{j=0}^{\infty} A_j z^j \tag{4.97}$$

and plugging everything into our differential equation yields the recurrence relation

$$U_j\,(j + M) = \sum_{k=0}^{j} A_k U_{j-k} \tag{4.98}$$

for the coefficient matrices U_j. However, since we don't know M, this does not help us much. By (4.90) you could suspect that we just have $M = A_0$ and $U_0 = \mathbb{I}$. Indeed, if we assume $\det(U_0) \neq 0$, 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 $\det(U_0) \neq 0$ 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 μ is an eigenvalue with corresponding eigenvector u_0 of M, then

$$w_0(z) = W(z)u_0 = z^{\mu} U(z) u_0 \tag{4.99}$$

is a solution of the form

$$w_0(z) = z^\alpha u_0(z), \quad u_0(z) = \sum_{j=0}^\infty u_{0,j} z^j, \quad u_{0,0} \neq 0, \ \alpha = \mu + m. \tag{4.100}$$

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

$$(\alpha + j) u_{0,j} = \sum_{k=0}^{j} A_k u_{0,j-k} \tag{4.101}$$

respectively

$$(A_0 - \alpha - j) u_{0,j} + \sum_{k=1}^{j} A_k u_{0,j-k} = 0. \tag{4.102}$$

In particular, for $j = 0$,

$$(A_0 - \alpha) u_{0,0} = 0, \tag{4.103}$$

we see that α must be an eigenvalue of A_0!

Now what about the case where μ corresponds to a nontrivial Jordan block of size $n > 1$? Then, by (4.91), we have a corresponding set of generalized eigenvectors u_l, $1 \leq l \leq n$, such that

$$w_l(z) = W(z) u_l = z^\alpha \left(u_l(z) + \log(z) u_{l-1}(z) + \cdots + \frac{\log(z)^l}{l!} u_0(z) \right), \tag{4.104}$$

$1 \leq l \leq n$, are n solutions. Here

$$u_l(z) = z^{\mu - \alpha} U(z) u_l = \sum_{j=m_l}^\infty u_{l,j} z^j, \quad u_{l,m_l} \neq 0, \quad 1 \leq l \leq n, \tag{4.105}$$

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

$$u_{l-1,j} = 0, \qquad j < m_l, \tag{4.106}$$

and

$$(\alpha + j) u_{l,j} + u_{l-1,j} = \sum_{k=1}^{j-m_l} A_k u_{l,j-k}, \quad j \geq m_l. \tag{4.107}$$

The first part implies $m_{l-1} \geq m_l$ and in particular $m_l \leq m_0 = 0$. The second implies

$$(A_0 - \alpha - j) u_{l,j} + \sum_{k=1}^{j} A_k u_{l,j-k} = u_{l-1,j}, \quad j \geq m_l. \tag{4.108}$$

Furthermore, for $j = m_l$ we get

$$(A_0 - \alpha - m_l) u_{l,m_l} = u_{l-1,m_l}. \tag{4.109}$$

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 4.11. *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*

$$w(z) = z^\alpha \sum_{k=0}^{l} u_{l-k}(z) \frac{\log(z)^k}{k!}, \qquad u_l(z) = \sum_{j=m_l}^{\infty} u_{l,j} z^j, \quad u_{l,m_l} \neq 0, \tag{4.110}$$

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 4.12. *For every eigenvalue μ of M there must be an eigenvalue $\alpha = \mu + m$, $m \in \mathbb{N}_0$, of A_0. For every Jordan block of μ there is a corresponding Jordan block of α, 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 (4.108), where α is an eigenvalue of A_0 we get a solution of our differential equation via (4.104) 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 (4.108). If the numbers $\alpha + j$ are not eigenvalues of A_0 for $j > 0$, we can multiply (4.108) by $(A_0 - \alpha - j)^{-1}$ 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 4.13. *Suppose $u_{l,j}$ solves* (4.108). *Then $u_l(z)$ defined via the power series* (4.105) *has the same radius of convergence as the power series for $zA(z)$ around $z = 0$. Moreover, $w_l(z)$ defined via* (4.104) *is a solution of $w' = A(z)w$.*

Proof. Suppose δ is smaller than the radius of convergence of the power series for $zA(z)$ around $z = 0$ and abbreviate

$$M = \sum_{j=0}^{\infty} \|A_j\| \, \delta^j < \infty.$$

We equip the space of vector valued $\underline{u} = (u_j)_{j\in\mathbb{N}_0}$ expansion coefficients with the norm (Problem 4.20)

$$\|\underline{u}\| = \sum_{j=0}^{\infty} |u_j| \, \delta^j.$$

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

$$K u_j = \begin{cases} ,0 & j \le j_0, \\ \frac{1}{\gamma+j} \sum_{k=0}^{j} A_k u_{j-k}, & j > j_0, \end{cases}$$

then

$$\begin{aligned} \|K\underline{u}\| &\le \frac{1}{j_0 - |\mathrm{Re}(\gamma)|} \sum_{j=0}^{\infty} \sum_{k=0}^{j} \|A_k\| \, |u_{j-k}| \delta^j \\ &= \sum_{j=0}^{\infty} \sum_{k=0}^{\infty} \|A_k\| \, |u_j| \delta^{j+k} = \frac{M}{j_0 - |\mathrm{Re}(\gamma)|} \|\underline{u}\|. \end{aligned}$$

Hence for j_0 sufficiently large, the equation $u_j = v_j + K u_j$ has a unique solution by the contraction principle for any fixed v_j. Now let $u_{l,j}$ be a solution of (4.107)

$$u_{l,m_l+j} = \frac{1}{\alpha + m_l + j} \sum_{k=1}^{j} A_k u_{l,m_l+j-k} - \frac{1}{\alpha + m_l + j} u_{l-1,m_l+j}$$

and choose $\gamma = \alpha + m_l$ and $v_j = u_{l,m_l+j}$ for $j \le j_0$ respectively $v_j = -\frac{1}{\alpha+m_l+j} u_{l-1,m_l+j}$ for $j > j_0$. Then the solution of our fixed point problem u_j coincides with our solution u_{l,m_l+j} of (4.108) by construction. □

In summary, we obtain the following procedure for finding a full set of linearly independent solutions:

For all eigenvalues α of A_0 for which $\alpha + j$ is not an eigenvalue for all $j \in \mathbb{N}_0$, take corresponding generalized eigenvectors $u_{0,l} \ne 0$, $(A_0 - \alpha)u_{0,l} =$

$u_{0,l-1}$. Then $w_l(z)$ as defined in (4.104) with $m_l = 0$ and

$$u_{l,j} = (A_0 - \alpha - j)^{-1} \left(u_{l-1,j} - \sum_{k=1}^{j} a_k u_{l,j-k} \right), \tag{4.111}$$

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., (4.108) has no nontrivial solution). Next, add further terms in the ansatz (4.104) for α 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.

Clearly we can also try to apply this procedure to get a power series around infinity. To this end, one makes the change of coordinates $\zeta = \frac{1}{z}$. Then our system transforms to

$$\omega' = -\frac{1}{\zeta^2} A(\frac{1}{\zeta})\omega, \qquad w(z) = \omega(\frac{1}{z}). \tag{4.112}$$

In particular, ∞ is a simple singularity if and only if $A(z)$ has (at least) a first-order zero at ∞, that is,

$$A(\frac{1}{\zeta}) = \zeta \sum_{j=0}^{\infty} A_j \zeta^j. \tag{4.113}$$

A system is called a **Fuchs system** if it has only finitely many singularities all of which, including infinity, are simple.

Lemma 4.14. *Every Fuchs system is of the form*

$$A(z) = \sum_{j=1}^{k} \frac{A_j}{z - z_j}. \tag{4.114}$$

Proof. Consider,

$$B(z) = A(z) - \sum_{j=1}^{k} \frac{A_j}{z - z_j},$$

where $A_j = \lim_{z \to z_j} (z - z_j) A(z)$. Then $B(z)$ is analytic on all of $\mathbb{C}$ by construction. Moreover, since $A(z)$ vanishes at ∞, so does $B(z)$ und thus $B(z)$ vanishes by Liouville's theorem (every bounded analytic function is constant). □

Note that a Fuchs system is regular at ∞ if and only if $\sum_{j=1}^k A_j = 0$. Hence every nontrivial ($A(z) \neq 0$) Fuchs system has at least two singularities.

Finally, let me remark, that all results for systems apply to the n'th order linear equation

$$u^{(n)}(z) + q_{n-1}(z) u^{(n-1)}(z) + \cdots + q_1(z) u'(z) + q_0(z) u(z) = 0. \tag{4.115}$$

Transforming this equation to a system as usual, shows that $z_0 = 0$ is a simple singularity if the coefficients $q_j(z)$, $0 \le j \le n-1$ have at most first-order poles. However, we can do even better. Introducing

$$w(z) = (u(z), z\, u'(z), \dots, z^{n-1} u^{(n-1)}(z)). \tag{4.116}$$

shows that

$$A(z) = \frac{1}{z} \begin{pmatrix} 0 & 1 & & & & \\ & 1 & 1 & & & \\ & & 2 & 1 & & \\ & & & \ddots & \ddots & \\ & & & & \ddots & 1 \\ -z^n q_0 & -z^{n-1} q_1 & \cdots & \cdots & -z^2 q_{n-2} & n-1-z\, q_{n-1} \end{pmatrix} \tag{4.117}$$

has a simple singularity at $z = 0$ if $q_j(z)$, $0 \le j \le n-1$, has a pole of order at most $n - j$ at $z = 0$.

For example, transforming (4.20) we obtain the system

$$w' = A(z) w, \qquad A(z) = \begin{pmatrix} 0 & \frac{1}{z} \\ -z q(z) & \frac{1}{z} - p(z) \end{pmatrix}. \tag{4.118}$$

Problem 4.20. *Let $w_j > 0$, $j \in \mathbb{N}_0$, be given weights. Show that the set of all sequences $\underline{u} = (u_j)_{j \in \mathbb{N}_0}$ with $u_j \in \mathbb{C}^n$ for which the norm*

$$\|\underline{u}\| = \sum_{j=0}^{\infty} |u_j|\, w_j$$

is finite, form a Banach space.

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. The present chapter tries to give a brief introduction under minimal requirements. 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 propagation speed of waves 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:

$$u(t,x) = w(t)y(x). \tag{5.2}$$

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

$$\frac{1}{c^2}\frac{\ddot{w}(t)}{w(t)} = \frac{y''(x)}{y(x)}. \tag{5.3}$$

Here we have used dots to indicate derivatives with respect to t and primes to indicate derivatives with respect to x. Now, if this equation should hold for all t and x, the quotients must be equal to a constant $-\lambda$ (the extra minus is chosen for convenience later on). That is, we are led 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!

Note that if λ is negative, then the trigonometric functions have to be replaced by their hyperbolic counterparts. However, since $\sinh(x)$ only vanishes at $x = 0$ this does not produce any further solutions (check this).

So we have found a large number of solutions satisfying the boundary conditions, 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. Moreover, since we have infinitely many solutions we can consider infinite linear combinations under appropriate assumptions on the coefficients.

Lemma 5.1. *Suppose $c_{1,n}$ and $c_{2,n}$ are sequences satisfying*

$$\sum_{n=1}^{\infty} n^2 |c_{1,n}| < \infty, \qquad \sum_{n=1}^{\infty} n^2 |c_{2,n}| < \infty. \tag{5.10}$$

Then

$$u(t,x) = \sum_{n=1}^{\infty} (c_{1,n}\cos(cn\pi t) + c_{2,n}\sin(cn\pi t))\sin(n\pi x), \tag{5.11}$$

is in $C^2(\mathbb{R}\times[0,1])$ *and satisfies the wave equation* (5.1) *as well as the boundary conditions* $u(t,0) = u(t,1) = 0$.

Proof. Consider

$$u_N(x,t) = \sum_{n=1}^{N} (c_{1,n}\cos(cn\pi t) + c_{2,n}\sin(cn\pi t))\sin(n\pi x),$$
$$w_N(x,t) = \sum_{n=1}^{N} (c_{1,n}\cos(cn\pi t) + c_{2,n}\sin(cn\pi t))\, n\pi\cos(n\pi x).$$

By our assumption (5.10) the Weierstraß M-test implies that both series converge uniformly to continuous functions $u(x,t)$, $w(x,t)$, respectively. Furthermore, since $w_N(t,x) = \frac{\partial}{\partial x}u_N(x,t)$ this also shows that $u(x,t)$ has a continuous partial derivative with respect to x given by $\frac{\partial}{\partial x}u(x,t) = w(t,x)$. Similarly one shows existence of the remaining derivatives. In particular, the fact that u_N solves the wave equation remains valid in the limit. □

Next, under the assumptions (5.10), the proof of the previous lemma also shows

$$u(0,x) = \sum_{n=1}^{\infty} c_{1,n}\sin(n\pi x), \qquad \frac{\partial}{\partial t}u(0,x) = \sum_{n=1}^{\infty} cn\pi c_{2,n}\sin(n\pi x). \tag{5.12}$$

Now observe that the sums on the right-hand side are nothing else but Fourier sine series. Moreover, recall that the trigonometric functions form a complete orthonormal system and, under mild assumptions, arbitrary functions can be expanded in such a series (do not worry if you are not familiar with this result, it will follow as a special case of our analysis in this chapter).

Hence, expanding the initial conditions into Fourier sine series

$$u(x) = \sum_{n=1}^{\infty} u_n\sin(n\pi x), \qquad v(x) = \sum_{n=1}^{\infty} v_n\sin(n\pi x), \tag{5.13}$$

where

$$u_n = 2\int_0^1 \sin(n\pi x)u(x)dx, \quad v_n = 2\int_0^1 \sin(n\pi x)v(x)dx, \tag{5.14}$$

we see that the solution of our original problem is given by (5.11) with $c_{1,n} = u_n$ and $c_{2,n} = \frac{v_n}{cn\pi}$, provided the Fourier coefficients satisfy

$$\sum_{n=1}^{\infty} n^2|u_n| < \infty, \qquad \sum_{n=1}^{\infty} n|v_n| < \infty. \tag{5.15}$$

It can be shown that this last condition holds if $u \in C^3[0,1]$ with $u(0) = u''(0) = u(1) = u''(1) = 0$ and $v \in C^2[0,1]$ with $v(0) = v(1) = 0$. We will consider this issue in the example after Theorem 5.11 and in Problem 5.22. For a different method of solving the one-dimensional wave equation see Problem 5.1.

In general, a vast number of problems in various areas lead to the investigation of the following problem

$$Ly(x) = \lambda y(x), \qquad L = \frac{1}{r(x)}\left(-\frac{d}{dx}p(x)\frac{d}{dx} + q(x)\right), \tag{5.16}$$

subject to the **boundary conditions**

$$\cos(\alpha)y(a) = \sin(\alpha)p(a)y'(a), \quad \cos(\beta)y(b) = \sin(\beta)p(b)y'(b), \tag{5.17}$$

$\alpha, \beta \in \mathbb{R}$. 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

$$\langle f, g \rangle = \int_a^b f^*(x) g(x) dx, \tag{5.18}$$

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. *Note that the wave equation* (5.1) *can be factorized according to*

$$\left(\frac{\partial}{\partial x} - \frac{1}{c}\frac{\partial}{\partial t}\right)\left(\frac{\partial}{\partial x} + \frac{1}{c}\frac{\partial}{\partial t}\right) u = \left(\frac{\partial}{\partial x} + \frac{1}{c}\frac{\partial}{\partial t}\right)\left(\frac{\partial}{\partial x} - \frac{1}{c}\frac{\partial}{\partial t}\right) u = 0.$$

Hence $f(x+ct)$ and $g(x-ct)$ as well as $f(x+ct) + g(x-ct)$ are solutions of the wave equation for arbitrary $f, g \in C^2(\mathbb{R})$.

Express f and g in terms of the initial conditions $u(0,x) = u(x) \in C^2(\mathbb{R})$ and $\frac{\partial}{\partial t}u(0,x) = v(x) \in C^1(\mathbb{R})$ to obtain **d'Alembert's formula**

$$u(t,x) = \frac{u(x+ct) + u(x-ct)}{2} + \frac{1}{2c}\int_{x-ct}^{x+ct} v(y)dy.$$

In order to obtain a solution on $x \in [0,1]$ satisfying the boundary conditions $u(t,0) = u(t,1) = 0$, use the following reflection technique: Extend the initial condition $u(x) \in C^2[0,1]$ to $[-1,1]$ using reflection $u(-x) = -u(x)$ and then to $\mathbb{R}$ using periodicity $u(x+2) = u(x)$. Show that the resulting function u will be $C^2(\mathbb{R})$ provided $u(0) = u''(0) = u(1) = u''(1) = 0$. Similarly we can extend $v \in C^1[0,1]$ to a function $v \in C^1(\mathbb{R})$ provided $v(0) = v(1) = 0$.

Problem 5.2. *Show that*

$$q_2(x)y'' + q_1(x)y' + q_0(x)y, \qquad q_2(x) > 0,$$

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 4.14) 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 $u(t,1) = u'(t,0) = 0$. Find all solutions of the form $u(t,x) = w(t)y(x)$. (Hint: Problem 4.13.)

Problem 5.4 (Heat equation)**.** *Use the method described in this section to solve the* **heat equation**

$$\frac{\partial}{\partial t}u(t,x) = \frac{\partial^2}{\partial x^2}u(t,x)$$

with boundary conditions $u(t,0) = u_0$, $u(t,1) = u_1$ and initial condition $u(0,x) = u(x)$. It models the temperature distribution of a thin wire whose edges are kept at a fixed temperature u_0 and u_1. What can you say about $\lim_{t\to\infty} u(t,x)$. (Hint: If $u(x,t)$ solves the heat equation so does $u(x,t) + a + bx$. Use this to reduce the boundary conditions to the case $u_0 = u_1 = 0$.)

Problem 5.5 (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*

$$m\frac{d^2}{dt^2}u(t,n) = k(u(t,n+1) - u(t,n)) + k(u(t,n-1) - u(t,n)),\ (t,n) \in \mathbb{R}\times\mathbb{Z},$$

where $m > 0$ is the mass of the particles and $k > 0$ is the spring constant. Here $u(t,n)$ is the displacement of the n'th particle from its equilibrium position at time t. (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 4.11.) 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. Compact symmetric operators

Suppose $\mathfrak{H}$ is a vector space. A map $\langle .,..\rangle : \mathfrak{H}\times\mathfrak{H} \to \mathbb{C}$ is called a **sesquilinear form** if it is conjugate linear in the first argument and linear in the second; that is,

$$\begin{array}{rcl} \langle \alpha_1 f_1 + \alpha_2 f_2, g\rangle &=& \alpha_1^*\langle f_1, g\rangle + \alpha_2^*\langle f_2, g\rangle,\\ \langle f, \alpha_1 g_1 + \alpha_2 g_2\rangle &=& \alpha_1\langle f, g_1\rangle + \alpha_2\langle f, g_2\rangle, \end{array} \quad \alpha_1,\alpha_2\in\mathbb{C}, \tag{5.19}$$

where '$*$' denotes complex conjugation. A sesquilinear form satisfying the requirements

(i) $\langle f, f\rangle > 0$ for $f \neq 0$ (positive definiteness),

(ii) $\langle f, g\rangle = \langle g, f\rangle^*$ (symmetry)

is called an **inner product** or **scalar product**. Associated with every scalar product is a norm

$$\|f\| = \sqrt{\langle f, f\rangle}. \tag{5.20}$$

Only the triangle inequality is nontrivial (cf. Section 2.1). It will follow from the Cauchy–Schwarz inequality below. Until then, just regard (5.20) as a convenient short hand notation.

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.

Example. Clearly $\mathbb{C}^n$ with the usual scalar product

$$\langle a, b\rangle = \sum_{j=1}^n a_j^* b_j \tag{5.21}$$

is a (finite dimensional) Hilbert space. ◇

A vector $f \in \mathfrak{H}_0$ is called **normalized** or **unit vector** if $\|f\| = 1$. Two vectors $f, g \in \mathfrak{H}_0$ are called **orthogonal** or **perpendicular** ($f \perp g$) if $\langle f, g\rangle = 0$ and **parallel** if one is a multiple of the other.

If f and g are orthogonal we have the **Pythagorean theorem**:

$$\|f+g\|^2 = \|f\|^2 + \|g\|^2, \qquad f \perp g, \tag{5.22}$$

which is one line of computation.

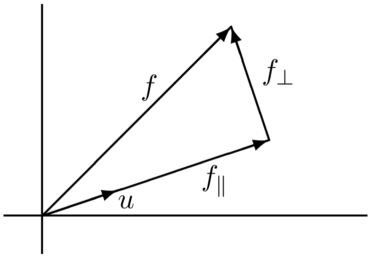
Suppose u is a unit vector. Then the projection of f in the direction of u is given by

$$f_\| = \langle u, f\rangle u \tag{5.23}$$

and $f_\perp$ defined via

$$f_\perp = f - \langle u, f\rangle u \tag{5.24}$$

is perpendicular to u since $\langle u, f_\perp\rangle = \langle u, f - \langle u, f\rangle u\rangle = \langle u, f\rangle - \langle u, f\rangle\langle u, u\rangle = 0$.



Taking any other vector parallel to u it is easy to see

$$\|f - \alpha u\|^2 = \|f_\perp + (f_\| - \alpha u)\|^2 = \|f_\perp\|^2 + |\langle u, f\rangle - \alpha|^2 \tag{5.25}$$

and this expression attains its minimum precisely if $\alpha = \langle u, f\rangle$. Hence $f_\| = \langle u, f\rangle u$ is the unique vector parallel to u which is closest to f.

As a first consequence we obtain the **Cauchy–Schwarz–Bunjakowski** inequality:

Theorem 5.2 (Cauchy–Schwarz–Bunjakowski). *Let $\mathfrak{H}_0$ be an inner product space. Then for every $f, g \in \mathfrak{H}_0$ we have*

$$|\langle f, g\rangle| \le \|f\|\,\|g\| \tag{5.26}$$

with equality if and only if f and g are parallel.

Proof. It suffices to prove the case $\|g\| = 1$ and use $f = \langle g, f\rangle g + f_\perp$. But then the claim follows from $\|f\|^2 = |\langle g, f\rangle|^2 + \|f_\perp\|^2 \ge |\langle g, f\rangle|^2$ with equality if and only if $f_\perp = 0$. □

Note that the Cauchy–Schwarz inequality entails that the scalar product is continuous in both variables, that is, if $f_n \to f$ and $g_n \to g$ we have $\langle f_n, g_n\rangle \to \langle f, g\rangle$.

As another consequence we infer that the map $\|.\|$ is indeed a norm since it satisfies the triangle inequality:

$$\|f+g\|^2 = \|f\|^2 + \langle f, g\rangle + \langle g, f\rangle + \|g\|^2 \le (\|f\| + \|g\|)^2. \tag{5.27}$$

The remaining two requirements are easy.

Next, let us generalize the projection to more than one vector. A set of vectors $\{u_j\}$ is called **orthonormal set** if $\langle u_j, u_k\rangle = 0$ for $j \neq k$ and $\langle u_j, u_j\rangle = 1$.

Lemma 5.3. *Suppose $\{u_j\}_{j=0}^n$ is an orthonormal set. Then every $f \in \mathfrak{H}_0$ can be written as*

$$f = f_\parallel + f_\perp, \qquad f_\parallel = \sum_{j=0}^n \langle u_j, f\rangle u_j, \tag{5.28}$$

where $f_\parallel$ and $f_\perp$ are orthogonal. Moreover, $\langle u_j, f_\perp\rangle = 0$ for all $0 \le j \le n$. In particular,

$$\|f\|^2 = \sum_{j=0}^n |\langle u_j, f\rangle|^2 + \|f_\perp\|^2. \tag{5.29}$$

Moreover, every $\hat{f}$ in the span of $\{u_j\}_{j=0}^n$ satisfies

$$\|f - \hat{f}\| \ge \|f_\perp\| \tag{5.30}$$

with equality holding if and only if $\hat{f} = f_\parallel$. In other words, $f_\parallel$ is uniquely characterized as the vector in the span of $\{u_j\}_{j=0}^n$ closest to f.

Proof. A straightforward calculation shows $\langle u_j, f - f_\parallel\rangle = 0$ and hence $f_\parallel$ and $f_\perp = f - f_\parallel$ are orthogonal. The formula for the norm follows by applying (5.22) iteratively.

Now, fix a vector

$$\hat{f} = \sum_{j=0}^n \alpha_j u_j.$$

in the span of $\{u_j\}_{j=0}^n$. Then one computes

$$\begin{aligned} \|f - \hat{f}\|^2 &= \|f_\parallel + f_\perp - \hat{f}\|^2 = \|f_\perp\|^2 + \|f_\parallel - \hat{f}\|^2 \\ &= \|f_\perp\|^2 + \sum_{j=0}^n |\alpha_j - \langle u_j, f\rangle|^2 \end{aligned}$$

from which the last claim follows. □

From (5.29) we obtain **Bessel's inequality**

$$\sum_{j=0}^n |\langle u_j, f\rangle|^2 \le \|f\|^2 \tag{5.31}$$

with equality holding if and only if f lies in the span of $\{u_j\}_{j=0}^n$.

In particular, the Bessel inequality shows that we can also handle countable orthonormal sets (cf. Problem 5.7). An orthonormal set $\{u_j\}_{j=0}^N$, $N \in \mathbb{N}_0 \cup \{\infty\}$ is called an **orthonormal basis** if

$$\|f\|^2 = \sum_{j=0}^N |\langle u_j, f\rangle|^2 \tag{5.32}$$

for all $f \in \mathfrak{H}_0$. Abbreviating

$$f_n = \sum_{j=0}^n \langle u_j, f\rangle u_j, \tag{5.33}$$

equation (5.29) implies $f - f_n \to 0$ as $n \to N$ and hence (5.32) is equivalent to

$$f = \sum_{j=0}^N \langle u_j, f\rangle u_j \tag{5.34}$$

for every $f \in \mathfrak{H}_0$.

A **linear operator** is a linear mapping

$$A : \mathfrak{D}(A) \to \mathfrak{H}_0, \tag{5.35}$$

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

$$\langle g, Af\rangle = \langle Ag, f\rangle, \qquad f, g \in \mathfrak{D}(A). \tag{5.36}$$

A number $z \in \mathbb{C}$ is called **eigenvalue** of A if there is a nonzero vector $u \in \mathfrak{D}(A)$ such that

$$Au = zu. \tag{5.37}$$

The vector u is called a corresponding **eigenvector** in this case. The set of all eigenvectors corresponding to z augmented by the zero vector is called the **eigenspace**

$$\operatorname{Ker}(A - z) = \{u \in \mathfrak{D}(A) | (A - z)u = 0\} \tag{5.38}$$

corresponding to z. Here we have used the shorthand notation $A - z$ for $A - z\mathbb{I}$. An eigenvalue is called **simple** if there is only one linearly independent eigenvector.

Theorem 5.4. *Let A be symmetric. Then all eigenvalues are real and eigenvectors corresponding to different eigenvalues are orthogonal.*

Proof. Suppose λ is an eigenvalue with corresponding normalized eigenvector u. Then $\lambda = \langle u, Au\rangle = \langle Au, u\rangle = \lambda^*$, which shows that λ is real. Furthermore, if $Au_j = \lambda_j u_j$, $j = 1, 2$, we have

$$(\lambda_1 - \lambda_2)\langle u_1, u_2\rangle = \langle Au_1, u_2\rangle - \langle u_1, Au_2\rangle = 0$$

finishing the proof. □

Unfortunately this theorem does not tell us anything about the existence of eigenvalues. In fact, a general symmetric operators might have no eigenvalues at all. Hence we need to impose some further requirements.

The linear operator A defined on $\mathfrak{D}(A) = \mathfrak{H}_0$ is called **bounded** if

$$\|A\| = \sup_{f:\|f\|=1} \|Af\| \tag{5.39}$$

is finite. It is not hard to see that this is indeed a norm (Problem 5.8) on the space of bounded linear operators. By construction, a bounded operator is Lipschitz continuous

$$\|Af\| \le \|A\|\|f\| \tag{5.40}$$

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 5.9).

In combination with symmetry compactness will turn out to guarantee the existence of an orthonormal basis of eigenfunctions. The crucial step is to prove existence of *one* eigenvalue.

Theorem 5.5. *A compact symmetric operator has an eigenvalue α_0 which satisfies $|\alpha_0| = \|A\|$.*

Proof. We set $\alpha = \|A\|$ and assume $\alpha \ne 0$ (i.e., $A \ne 0$) without loss of generality. Since

$$\|A\|^2 = \sup_{f:\|f\|=1} \|Af\|^2 = \sup_{f:\|f\|=1} \langle Af, Af\rangle = \sup_{f:\|f\|=1} \langle f, A^2 f\rangle$$

there exists a normalized sequence u_n such that

$$\lim_{n\to\infty} \langle u_n, A^2 u_n\rangle = \alpha^2.$$

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

$$\begin{aligned}\|(A^2-\alpha^2)u_n\|^2 &= \|A^2 u_n\|^2 - 2\alpha^2\langle u_n, A^2 u_n\rangle + \alpha^4\\ &\le 2\alpha^2(\alpha^2 - \langle u_n, A^2 u_n\rangle)\end{aligned}$$

(where we have used $\|A^2 u_n\| \le \|A\|\|Au_n\| \le \|A\|^2\|u_n\| = \alpha^2$) 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 $(A+\alpha)v = (A+\alpha)(A-\alpha)u = (A^2-\alpha^2)u = 0$ shows that either $v \ne 0$ is an eigenvector corresponding to $-\alpha$ or $v = 0$ and hence $u \ne 0$ is an eigenvector corresponding to α. □

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 5.10).

Now consider a compact symmetric operator A with eigenvalue α_0 (as above) and corresponding normalized eigenvector u_0. Then we can establish existence of an orthonormal basis of eigenfunctions by mimicking the proof of the finite dimensional case from Theorem 3.29: Set

$$\mathfrak{H}_0^{(1)} = \{f \in \mathfrak{H}_0 | \langle u_0, f\rangle = 0\} \tag{5.41}$$

and observe that $\mathfrak{H}_0^{(1)}$ is a closed linear subspace and hence an inner product space of its own. Moreover, 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 clearly inherits both the symmetry and compactness from A (check this!). Hence we can apply Theorem 5.5 iteratively to obtain a sequence of eigenvalues α_j with corresponding normalized eigenvectors u_j. Moreover, by construction, u_j is orthogonal to all u_k with $k < j$ 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 5.6 (Spectral theorem for compact symmetric operators)**.** *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 α_j converging to* 0*. The corresponding normalized eigenvectors u_j form an orthonormal set and every $f \in \mathrm{Ran}(A) = \{Ag|g \in \mathfrak{H}_0\}$ can be written as*

$$f = \sum_{j=0}^{N} \langle u_j, f\rangle u_j. \tag{5.42}$$

If $\mathrm{Ran}(A)$ *is dense, then the eigenvectors form an orthonormal basis.*

Proof. We assume that $\mathfrak{H}_0$ is infinite dimensional without loss of generality. Existence of the eigenvalues α_j and the corresponding eigenvectors u_j 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, let $f = Ag \in \mathrm{Ran}(A)$. Set

$$f_n = \sum_{j=0}^{n} \langle u_j, f\rangle u_j, \qquad g_n = \sum_{j=0}^{n} \langle u_j, g\rangle u_j$$

and observe

$$f_n = \sum_{j=0}^{n} \langle u_j, Ag\rangle u_j = \sum_{j=0}^{n} \langle Au_j, g\rangle u_j = \sum_{j=0}^{n} \alpha_j \langle u_j, g\rangle u_j = Ag_n.$$

Thus

$$\|f - f_n\| = \|A(g - g_n)\| = \|A_{n+1}(g - g_n)\| \le |\alpha_{n+1}|\|g - g_n\| \le |\alpha_{n+1}|\|g\|$$

since $g - g_n \in \mathfrak{H}_0^{(n+1)}$. Letting $n \to \infty$ shows $f_n \to f$ proving (5.42) in the case $f \in \operatorname{Ran}(A)$.

Next, let $f \in \mathfrak{H}_0$ be arbitrary and suppose $\operatorname{Ran}(A)$ is dense. For fixed $\varepsilon > 0$, there is an $\tilde{f}_\varepsilon \in \operatorname{Ran}(A)$ such that $\|f - \tilde{f}_\varepsilon\| < \frac{\varepsilon}{2}$. Moreover, by the previous part, there is an $\hat{f}_\varepsilon$ in the span of $\{u_j\}_{j=0}^n$ for some sufficiently large n, such that $\|\tilde{f}_\varepsilon - \hat{f}_\varepsilon\| < \frac{\varepsilon}{2}$. That is, $\|f - \hat{f}_\varepsilon\| < \varepsilon$ and since, by Lemma 5.3, f_n is the best approximation within the span of $\{u_j\}_{j=0}^n$ we even have $\|f - f_n\| \le \|f - \hat{f}_\varepsilon\| < \varepsilon$ for n sufficiently large. □

This is all we need and it remains to apply these results to Sturm–Liouville operators.

Problem 5.6. *Prove the* **parallelogram law**

$$\|f + g\|^2 + \|f - g\|^2 = 2\|f\|^2 + 2\|g\|^2$$

for $f, g \in \mathfrak{H}_0$.

Problem 5.7. *Let $\{u_j\}_{j=0}^\infty \subset \mathfrak{H}_0$ be a countable orthonormal set and $f \in \mathfrak{H}_0$. Show that*

$$f_n = \sum_{j=0}^{n} \langle u_j, f\rangle u_j$$

is a Cauchy sequence.

Problem 5.8. *Show that* (5.39) *is indeed a norm. Show that the product of two bounded operators is again bounded with $\|AB\| \le \|A\|\|B\|$.*

Problem 5.9. *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 5.10. *Show that if A is bounded, then every eigenvalue α satisfies $|\alpha| \le \|A\|$.*

5.3. Sturm–Liouville equations

Before we will apply the theory of inner product spaces to the investigation of Sturm–Liouville problems we have a look at the underlying differential equation

$$-(p(x)y')' + (q(x) - z\,r(x))y = 0, \qquad z \in \mathbb{C},\ x \in I = (a, b), \tag{5.43}$$

for $y \in C^2(I, \mathbb{C})$, which is equivalent to the first-order system

$$\begin{array}{rcl} y' &=& \frac{1}{p(x)} w \\ w' &=& (q(x) - z\,r(x))y \end{array}, \tag{5.44}$$

where $w(x) = p(x)y'(x)$. Hence we see that there is a unique solution if $p(x)^{-1}$, $q(x)$, and $r(x)$ are continuous in I. In fact, as noted earlier, it even suffices to assume that $p(x)^{-1}$, $q(x)$, 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* (respectively differentiable in the weak sense) in the sequel. However, we will assume that

$$r, q \in C^0([a, b], \mathbb{R}),\ p \in C^1([a, b], \mathbb{R}), \quad p(x), r(x) > 0,\ x \in [a, b], \tag{5.45}$$

for the rest of this chapter and call the differential equation (5.43) regular in this case. Note that if we only assume $p \in C^0([a, b], \mathbb{R})$, we will still be within the framework of the theory developed so far, but then y might no longer be C^2 since we only know $w = py' \in C^1$.

By (3.105) the principal matrix solution of (5.44) is given by

$$\Pi(z, x, x_0) = \begin{pmatrix} c(z, x, x_0) & s(z, x, x_0) \\ p(x)c'(z, x, x_0) & p(x)s'(z, x, x_0) \end{pmatrix}, \qquad z \in \mathbb{C}, \tag{5.46}$$

where $c(z, x, x_0)$ is the solution of (5.43) corresponding to the initial condition $c(z, x_0, x_0) = 1$, $p(x_0)c'(z, x_0, x_0) = 0$ and similarly for $s(z, x, x_0)$ but corresponding to the initial condition $s(z, x_0, x_0) = 0$, $p(x_0)s'(z, x_0, x_0) = 1$.

We know that $\Pi(z, x, x_0)$ is continuous with respect to x and x_0 by Theorem 2.9. But with respect to z a much stronger result is true. Recall that a function is said to be **entire** if it is analytic on all of $\mathbb{C}$.

Lemma 5.7. *The principal matrix solution $\Pi(z, x, x_0)$ is entire with respect to z for every fixed $(x, x_0) \in I \times I$.*

Proof. It suffices to show that every solution is entire with respect to z in a neighborhood of x_0 if the initial conditions are constant. In this case each of the iterations (2.13) is entire (in fact even polynomial) with respect to z. Moreover, for z in a compact set, the Lipschitz constant can be chosen independently of z. Hence the series of iterations converges uniformly for z in any compact set, implying that the limit is again analytic by the Weierstraß convergence theorem. □

Moreover, by Liouville's formula (3.91) the **modified Wronskian**

$$W_x(u,v) = u(x)p(x)v'(x) - p(x)u'(x)v(x) \tag{5.47}$$

is independent of x if $u(x)$ and $v(x)$ both solve (5.43) with the same $z \in \mathbb{C}$. In particular,

$$\det \Pi(z,x,x_0) = W(c(z,.,x_0), s(z,.,x_0)) = 1. \tag{5.48}$$

Moreover, by (3.97) the solution of the inhomogeneous equation

$$-(p(x)y')' + (q(x) - z\,r(x))y = g(x)r(x) \tag{5.49}$$

is given by

$$y(x) = y(x_0)c(z,x,x_0) + y'(x_0)s(z,x,x_0) + \int_{x_0}^{x} s(z,x,t)g(t)r(t)dt. \tag{5.50}$$

Moreover, note that given two linearly independent solutions u, v of (5.43) we have

$$\begin{aligned} c(z,x,x_0) &= \frac{u(x)p(x_0)v'(x_0) - p(x_0)u'(x_0)v(x)}{W(u,v)}, \\ s(z,x,x_0) &= \frac{u(x)v(x_0) - u(x_0)v(x)}{W(u,v)}. \end{aligned} \tag{5.51}$$

(Since both functions are solutions it suffice to check the initial conditions.)

Problem 5.11. *Given one solution $u(x)$ of (5.43), make a variation of constants ansatz $v(x) = c(x)u(x)$ and show that a second solution is given by*

$$v(x) = u(x)\int^{x} \frac{1}{p(t)u(t)^2}dt.$$

While this formula breaks down at points where u vanishes, Rofe-Beketov's formula works even at such points:

$$\begin{aligned} v(x) =& u(x)\int^{x} \frac{(q(t) + p(t)^{-1} - z\,r(t))(u(t)^2 - (p(t)u'(t))^2)}{(u(t)^2 + (p(t)u'(t))^2)^2}dt \\ & - \frac{p(x)u'(x)}{u(x)^2 + (p(x)u'(x))^2}. \end{aligned}$$

Problem 5.12. *Show that if u is a solution of (5.43), then $w = pu'/u$ satisfies the Riccati equation*

$$w' + p(x)^{-1}w^2 = q(x) - z\,r(x).$$

Problem 5.13 (Liouville normal form)**.** *Show that if $p, r \in C^2[a,b]$, the differential equation (5.43) can be transformed into one with $r = p = 1$ using the diffeomorphism*

$$y(x) = \int_a^x \sqrt{\frac{r(t)}{p(t)}}dt,$$

which maps the interval (a,b) *to the interval* $(0,c)$, $c = \int_a^b \sqrt{\frac{r(t)}{p(t)}}dt$. *By a slight abuse of notation we will denote the inverse of this diffeomorphism by* $x(y)$. *Then, setting*

$$v(y) = \sqrt[4]{r(x(y))p(x(y))}\, u(x(y))$$

the Sturm–Liouville equation

$$-(p(x)u'(x))' + q(x)u(x) = r(x)zu(x), \qquad x \in (a,b),$$

transforms into

$$-v''(y) + Q(y)v(y) = zv(y), \qquad y \in (0,c),$$

where

$$Q(y) = q(x(y)) - \frac{(p(x(y))r(x(y)))^{1/4}}{r(x(y))}\big(p(x(y))((p(x((y))r(x(y)))^{-1/4})')\big)'.$$

Moreover,

$$\int_a^b |u(x)|^2 r(x)dx = \int_0^c |v(y)|^2 dy.$$

Problem 5.14. *Suppose* $u(x)$ *satisfies*

$$u''(x) + g(x)u'(x) + f(x)u(x) = h(x).$$

Show that

$$v(x) = \mathrm{e}^{\frac{1}{2}\int^x g(y)dy} u(x)$$

satisfies

$$v''(x) + \left(f(x) - \frac{1}{2}g'(x) - \frac{1}{4}g(x)^2\right) v(x) = \mathrm{e}^{\frac{1}{2}\int^x g(y)dy} h(x).$$

5.4. Regular Sturm–Liouville problems

Now we want to apply the theory of inner product spaces to the investigation of Sturm–Liouville problems. As in the previous section we continue to assume (5.45).

We first need a suitable scalar product. We will consider

$$\langle f,g\rangle = \int_I f(x)^* g(x) r(x) dx, \tag{5.52}$$

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 an operator

$$L = \frac{1}{r(x)}\left(-\frac{d}{dx}p(x)\frac{d}{dx} + q(x)\right) \tag{5.53}$$

in $\mathfrak{H}_0$. Since there are functions in $\mathfrak{H}_0$ which are not differentiable, we cannot apply it to arbitrary function in $\mathfrak{H}_0$. Thus we need a suitable domain

$$\mathfrak{D}(L) = \{f \in C^2([a,b],\mathbb{C}) | BC_a(f) = BC_b(f) = 0\}, \tag{5.54}$$

where

$$\begin{aligned} BC_a(f) &= \cos(\alpha)f(a) - \sin(\alpha)p(a)f'(a),\\ BC_b(f) &= \cos(\beta)f(b) - \sin(\beta)p(b)f'(b). \end{aligned} \tag{5.55}$$

In other words, we allow linear combinations of the boundary values $f(a)$ and $f'(a)$ (resp. $f(b)$ and $f'(b)$) as boundary conditions. This choice ensures that $\mathfrak{D}(L)$ is a linear subspace of $\mathfrak{H}_0$ and one can even show that it is dense:

Lemma 5.8. *The set of twice differentiable functions with compact support $C_c^2((a,b),\mathbb{C})$ is dense in $\mathfrak{H}_0$.*

Proof. Let $P(x) = 30\int_0^x y^2(y-1)^2 dy = x^3(6x^2 - 15x + 10)$. Note that by construction $P(x)$ is monotone increasing from 0 to 1 (in particular $0 \le P(x) \le 1$ for $0 \le x \le 1$) and both $P'(x)$ as well as $P''(x)$ vanish at $x = 0, 1$. We set $P(x) = 0$ for $x \le 0$ and $P(x) = 1$ for $x \ge 1$ such that $P(x) \in C^2(\mathbb{R})$.

Next pick $f \in C([a,b],\mathbb{C})$. Since f is uniformly continuous we can find a $\delta > 0$ for every $\varepsilon > 0$ such that $|f(x) - f(y)| \le \varepsilon$ whenever $|x - y| \le \delta$. By decreasing δ we can assume $b - a = n\delta$ for some integer n and $\delta \le \varepsilon$. Now set $x_j = a + j\delta$, $0 \le j \le n$, and define

$$\begin{aligned} f_\varepsilon(x) =& f(x_1)P(\frac{x - a - \delta/2}{\delta/2}) + \sum_{j=1}^{n-1}(f(x_{j+1}) - f(x_j))P(\frac{x - x_j}{\delta})\\ &- f(x_{n-1})P(\frac{x - b + \delta}{\delta/2}). \end{aligned}$$

Then $f_\varepsilon \in C_c^2((a,b),\mathbb{C})$ and $\max_{x\in[x_1,x_{n-1}]}|f(x) - f_\varepsilon(x)| \le \varepsilon$. Hence

$$\|f - f_\varepsilon\|^2 \le 8M^2R^2\delta + \varepsilon^2R^2(b-a) \le \varepsilon(8M^2 + \varepsilon(b-a))R^2,$$

where $M = \max_{x\in[a,b]}|f(x)|$, $R = \max_{x\in[a,b]}|r(x)|$, and the claim follows. □

It is not hard to show that the same is true for $C_c^\infty((a,b),\mathbb{C})$ (Problem 5.18).

The two most important cases are $\alpha = 0$ (i.e., $u(a) = 0$) and $\alpha = \pi/2$ (i.e., $u'(a) = 0$). The condition $u(a) = 0$ is called a **Dirichlet boundary condition** at a and the condition $u'(a) = 0$ is called a **Neumann boundary condition** at a. The general case is also known as **Robin boundary condition**. Note that without loss of generality one can assume $\alpha \in [0, \pi)$.

Clearly we want L to be symmetric. In order to get L from one side in the scalar product to the other we use integration by parts (twice) to obtain the **Lagrange identity**

$$\int_c^d g(Lf)\, r dx = W_c(g,f) - W_d(g,f) + \int_c^d (Lg) f\, r dx \tag{5.56}$$

for $f, g \in C^2([a,b],\mathbb{C})$ and $a \le c < d \le b$. Specializing to the case $(c,d) = (a,b)$ and replacing g by g^*,

$$\langle g, Lf\rangle = W_a(g^*, f) - W_b(g^*, f) + \langle Lg, f\rangle, \tag{5.57}$$

this is almost what we want except for the extra boundary terms and here is where the boundary conditions come into play: If f and g satisfy the same boundary conditions the above two Wronskians vanish (Problem 5.19) and hence

$$\langle g, Lf\rangle = \langle Lg, f\rangle, \qquad f, g \in \mathfrak{D}(L), \tag{5.58}$$

which shows that L is symmetric.

Of course we want to apply Theorem 5.6 next and for this we would need to show that L is compact. Unfortunately, it turns out that L is not even bounded (Problem 5.16) and it looks like we are out of luck. However, there is one last chance: the inverse of L might be compact so that we can apply Theorem 5.6 to it.

Since L might not be injective (0 might be 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 $(L-z)f = g$ which can be done by virtue of (5.50). 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 unknown constants in (5.50) such that the boundary conditions are satisfied. To this end we will choose two solutions u_b and u_a of the homogeneous equation, which will be adapted to our boundary conditions, and use (5.51). In this case (5.50) can be written as

$$\begin{aligned} f(x) =& \frac{u_b(z,x)}{W(z)}\Big(c_1 + \int_a^x u_a(z,t) g(t)\, r(t) dt\Big) \\ &+ \frac{u_a(z,x)}{W(z)}\Big(c_2 + \int_x^b u_b(z,t) g(t)\, r(t) dt\Big), \end{aligned} \tag{5.59}$$

implying

$$\begin{aligned} f'(x) =& \frac{u_b'(z,x)}{W(z)}\Big(c_1 + \int_a^x u_a(z,t) g(t)\, r(t) dt\Big) \\ &+ \frac{u_a'(z,x)}{W(z)}\Big(c_2 + \int_x^b u_b(z,t) g(t)\, r(t) dt\Big). \end{aligned} \tag{5.60}$$

Here we have abbreviated

$$W(z) = W(u_b(z), u_a(z)) \tag{5.61}$$

which is independent of x as noted in the previous section.

Now let us choose $c_1 = 0$. Then $f(a) = cu_a(z,a)$ and $f'(a) = cu_a'(z,a)$ (where $c = \frac{c_2+\langle u_b(z)^*,g\rangle}{W(z)}$). So choosing $u_a(z,x)$ such that $BC_a(u_a(z)) = 0$, we infer $BC_a(f) = 0$. Similarly, choosing $c_2 = 0$ and $u_b(z,x)$ such that $BC_b(u_b(z)) = 0$, we infer $BC_b(f) = 0$. But can we always do this? Well, using the initial conditions

$$\begin{aligned} u_a(z,a) &= \sin(\alpha), \quad & p(a)u_a'(z,a) &= \cos(\alpha),\\ u_b(z,b) &= \sin(\beta), \quad & p(b)u_b'(z,b) &= \cos(\beta), \end{aligned} \tag{5.62}$$

we have two solutions of the required type except for the fact that the Wronskian $W(z)$ might vanish. Now what is so special about the zeros of this Wronskian?

Lemma 5.9. *The Wronskian $W(z) = W(u_b(z), u_a(z))$ is an entire function which vanishes precisely at the eigenvalues of L.*

Proof. First of all, $W(z)$ is entire since both $u_a(z,x)$ and $u_b(z,x)$ (as well as their x derivatives) are by Lemma 5.7. Moreover, $W(z) = 0$ implies that $u_b(z)$ and $u_a(z)$ are linearly dependent, that is, $u_b(z,x) = c(z)u_a(z,x)$. Hence $BC_a(u_b(z)) = c(z)BC_a(u_a(z)) = 0$ shows that z is an eigenvalue with corresponding eigenfunction $u_b(z,x)$. □

In particular, all zeros of $W(z)$ must be real and since the zeros of an entire function can have no finite accumulation point (by the identity theorem from complex analysis), the eigenvalues of L are discrete.

Note (Problem 5.20)

$$u_a(z,x)^* = u_a(z^*,x), \qquad u_b(z,x)^* = u_b(z^*,x) \tag{5.63}$$

implying $W(z)^* = W(z^*)$. In particular both solutions are real-valued for $z \in \mathbb{R}$.

Now let us rewrite (5.59) in the operator form $f(x) = R_L(z)g(x)$ by introducing the operator (the **resolvent** of L)

$$R_L(z)g(x) = \int_a^b G(z,x,t)g(t)\,r(t)dt, \qquad g \in \mathfrak{H}_0, \tag{5.64}$$

where

$$G(z,x,t) = \frac{1}{W(z)}\begin{cases} u_b(z,x)u_a(z,t), & x \ge t,\\ u_b(z,t)u_a(z,x), & x \le t, \end{cases} \tag{5.65}$$

is called the **Green function** of L. Note that $G(z,x,t)$ is meromorphic with respect to $z \in \mathbb{C}$ with poles precisely at the zeros of $W(z)$ and satisfies (cf. (5.63))

$$G(z,x,t)^* = G(z^*,x,t), \qquad G(z,x,t) = G(z,t,x). \tag{5.66}$$

Then, by construction we have $R_L(z) : \mathfrak{H}_0 \to \mathfrak{D}(L)$ and

$$(L-z)R_L(z)g = g, \qquad g \in \mathfrak{H}_0. \tag{5.67}$$

Similarly we can verify

$$R_L(z)(L-z)f = f, \qquad f \in \mathfrak{D}(L), \tag{5.68}$$

which shows $\operatorname{Ran}(R_L(z)) = \mathfrak{D}(L)$. To see this we proceed as in the proof of the Lagrange identity

$$\begin{aligned} R_L(z)(L-z)f(x) = & \int_a^b G(z,x,t)((L-z)f(t))r(t)dt \\ = & \frac{u_b(z,x)}{W(z)} \int_a^x u_a(z,t)((L-z)f(t))r(t)dt \\ & + \frac{u_a(z,x)}{W(z)} \int_x^b u_b(z,t)((L-z)f(t))r(t)dt \\ = & \frac{u_a(z,x)W_x(u_b(z),f) - u_b(z,x)W_x(u_a(z),f)}{W(z)} \\ = & f(x). \end{aligned} \tag{5.69}$$

Here we have used the Lagrange identity (5.56) and $W_a(u_a,f) = -BC_a(f) = 0$, $W_b(u_b,f) = -BC_b(f) = 0$ in the third step.

In other words, $R_L(z)$ is the inverse of $L-z$. Our next lemma shows that $R_L(z)$ is compact.

Lemma 5.10. *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 $|G(z,y,t) - G(z,x,t)| \le \varepsilon$ whenever $|y-x| \le \delta$. Let $g(x) = R_L(z)f(x)$. Then

$$\begin{aligned} |g(x) - g(y)| \le & \int_a^b |G(z,y,t) - G(z,x,t)|\, |f(t)|\, r(t)dt \\ \le & \varepsilon \int_a^b |f(t)|\, r(t)dt \le \varepsilon \|1\|\, \|f\|, \end{aligned}$$

whenever $|y-x| \le \delta$. (Here we have used the Cauchy–Schwarz inequality in the last step.) 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 2.18). But a uniformly convergent sequence is also convergent in the norm induced by the scalar product since

$$\|f\| = \sqrt{\int_a^b |f(t)|^2 r(t)dt} \le \sqrt{\sup_{x\in[a,b]} |f(x)|^2 \int_a^b r(t)dt} = \|1\| \sup_{x\in[a,b]} |f(x)|.$$

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:

$$\begin{aligned}\langle g, R_L(\lambda) f\rangle &= \int_a^b g(x)^* \left(\int_a^b G(\lambda, x, t) f(t) r(t) dt\right) r(x) dx\\ &= \int_a^b \left(\int_a^b g(x)^* G(\lambda, x, t) r(x) dx\right) f(t) r(t) dt\\ &= \int_a^b \left(\int_a^b G(\lambda, t, x) g(x) r(x) dx\right)^* f(t) r(t) dt = \langle R_L(\lambda) g, f\rangle.\end{aligned}$$

This finishes the proof. □

As a consequence we can apply Theorem 5.6 to obtain

Theorem 5.11. *The regular Sturm–Liouville problem has a countable number of discrete and simple eigenvalues E_n which accumulate only at ∞. The corresponding normalized eigenfunctions u_n can be chosen real-valued and form an orthonormal basis for $\mathfrak{H}_0$, that is, every $f \in \mathfrak{H}_0$ can be written as*

$$f(x) = \sum_{n=0}^{\infty} \langle u_n, f\rangle u_n(x). \tag{5.70}$$

Moreover, for $f \in \mathfrak{D}(L)$ this series is uniformly convergent.

Proof. Pick a value $\lambda \in \mathbb{R}$ such that $R_L(\lambda)$ exists. By Theorem 5.6 $R_L(\lambda)$ has a countable number of eigenvalues $\alpha_n \to 0$ plus a corresponding orthonormal system of eigenfunctions u_n. Moreover, since $\mathrm{Ran}(R_L(\lambda)) = \mathfrak{D}(L)$ is dense, the eigenfunctions form an orthonormal basis.

Moreover, $R_L(\lambda) u_n = \alpha_n u_n$ is equivalent to $L u_n = (\lambda + \frac{1}{\alpha_n}) u_n$, which shows that $E_n = \lambda + \frac{1}{\alpha_n}$ are eigenvalues of L with corresponding eigenfunctions u_n.

Hence the first two claims follow except for the fact 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. In particular, $u_n(x)$ is a multiple of $u_a(E_n, x)$ and hence can be chosen real-valued.

To show that (5.70) converges uniformly if $f \in \mathfrak{D}(L)$ we begin by writing $f = R_L(\lambda) g$, $g \in \mathfrak{H}_0$, implying

$$\sum_{n=0}^{\infty} \langle u_n, f\rangle u_n(x) = \sum_{n=0}^{\infty} \alpha_n \langle u_n, g\rangle u_n(x)$$

Moreover, the Cauchy–Schwarz inequality shows

$$\left|\sum_{j=m}^{n} \alpha_j \langle u_j, g\rangle u_j(x)\right|^2 \le \sum_{j=m}^{n} |\langle u_j, g\rangle|^2 \sum_{j=m}^{n} |\alpha_j u_j(x)|^2.$$

Now, by (5.32), $\sum_{j=0}^{\infty} |\langle u_j, g\rangle|^2 = \|g\|^2$ and hence the first term is part of a convergent series. Similarly, the second term can be estimated independent of x since

$$\alpha_n u_n(x) = R_L(\lambda) u_n(x) = \int_a^b G(\lambda, x, t) u_n(t) r(t) dt = \langle u_n, G(\lambda, x, .)\rangle$$

implies

$$\sum_{j=m}^{n} |\alpha_j u_j(x)|^2 \le \sum_{j=0}^{\infty} |\langle u_j, G(\lambda, x, .)\rangle|^2 = \int_a^b |G(\lambda, x, t)|^2 r(t) dt \le M(\lambda)^2 \|1\|,$$

where $M(\lambda) = \max_{x,t \in [a,b]} |G(\lambda, x, t)|$, again by (5.32). □

Moreover, it is even possible to weaken our assumptions for uniform convergence. To this end we introduce the **quadratic form** associated with L:

$$\begin{aligned} Q(f, g) = &\int_a^b \big(p(x) f'(x)^* g'(x) + q(x) f(x)^* g(x)\big)\, dx \\ &+ Q_{\alpha,a}(f, g) - Q_{\beta,b}(f, g), \qquad f, g \in C^1([a, b], \mathbb{C}), \end{aligned} \tag{5.71}$$

where

$$Q_{\gamma,c}(f, g) = \begin{cases} 0, & \gamma = 0, \\ \cot(\gamma) f(c)^* g(c), & \gamma \ne 0. \end{cases} \tag{5.72}$$

We will set $Q(f) = Q(f, f)$. An integration by parts shows

$$Q(f, g) = \langle f, Lg\rangle \tag{5.73}$$

provided $g \in \mathfrak{D}(L)$ and f satisfied a possible Dirichlet boundary condition at the endpoints. In fact, the above formula continues to hold for f in a slightly larger class of functions,

$$\mathfrak{Q}(L) = \{f \in C_p^1[a, b] | f(a) = 0 \text{ if } \alpha = 0,\ f(b) = 0 \text{ if } \beta = 0\} \supseteq \mathfrak{D}(L), \tag{5.74}$$

which we call the **form domain** of L. Here $C_p^1[a, b]$ denotes the set of piecewise continuously differentiable functions f in the sense that f is continuously differentiable except for a finite number of points at which it is continuous and the derivative has limits form the left and right. In fact, any class of functions for which the partial integration needed to obtain (5.73) can be justified would be good enough (e.g. the set of absolutely continuous functions).

Lemma 5.12. *The eigenvalues of a regular Sturm–Liouville problem are bounded from below and can hence be ordered as follows:*

$$E_0 < E_1 < \cdots . \tag{5.75}$$

Moreover, we have the **Rayleigh–Ritz principle**

$$E_0 = \min_{f\in\mathfrak{D}(L):\|f\|=1} Q(f) = \min_{f\in\mathfrak{D}(L):\|f\|=1} \langle f, Lf\rangle \tag{5.76}$$

with equality if and only if $f = u_0$. In particular, for $0 \le \alpha \le \frac{\pi}{2}$ and $\frac{\pi}{2} \le \beta \le \pi$ we obtain

$$\min_{x\in[a,b]} q(x) \le E_0. \tag{5.77}$$

Proof. We first assume $0 \le \alpha \le \frac{\pi}{2}$ and $\frac{\pi}{2} \le \beta \le \pi$ such that the boundary terms in (5.71) are non-negative. Then we have $Q(f) \ge \min_{x\in[a,b]} q(x)\|f\|^2$ and hence (5.73) implies $Q(u_j) = E_j \ge \min_{x\in[a,b]} q(x)$. In particular, we can order the eigenvalues as indicated. The second claim now follows using $f = \sum_{j=0}^\infty \langle u_j, f\rangle u_j$ implying

$$\langle f, Lf\rangle = \sum_{j=0}^\infty |\langle u_j, f\rangle|^2 E_j$$

and the equality $Q(f) = \langle f, Lf\rangle$ for $f \in \mathfrak{D}(L)$.

If one of the boundary terms is negative, it can still be controlled in terms of the integral using Problem 5.23. Details are left as an exercise. □

Lemma 5.13. *For a regular Sturm–Liouville problem* (5.70) *converges uniformly provided $f \in \mathfrak{Q}(L)$.*

Proof. We first assume $0 \le \alpha \le \frac{\pi}{2}$ and $\frac{\pi}{2} \le \beta \le \pi$ such that the boundary terms in (5.71) are non-negative.

By replacing $L \to L - q_0$ for $q_0 > \min_{x\in[a,b]} q(x)$ we can assume $q(x) > 0$ without loss of generality (this will shift the eigenvalues $E_n \to E_n - q_0$ and leave the eigenvectors unchanged). In particular, we have $Q(f) > 0$ after this change. By (5.73) we also have $E_j = \langle u_j, Lu_j\rangle = Q(u_j) > 0$.

Now let $f \in \mathfrak{Q}(L)$ and consider (5.70). Then, using that $Q(f,g)$ is a symmetric sesquilinear form (after our shift it is even a scalar product) plus

(5.73) one obtains

$$\begin{aligned}
0 \le& Q\big(f - \sum_{j=m}^{n} \langle u_j, f\rangle u_j\big)\\
=& Q(f) - \sum_{j=m}^{n} \langle u_j, f\rangle Q(f, u_j) - \sum_{j=m}^{n} \langle u_j, f\rangle^* Q(u_j, f)\\
&+ \sum_{j,k=m}^{n} \langle u_j, f\rangle^* \langle u_k, f\rangle Q(u_j, u_k)\\
=& Q(f) - \sum_{j=m}^{n} E_j |\langle u_j, f\rangle|^2
\end{aligned}$$

which implies

$$\sum_{j=m}^{n} E_j |\langle u_j, f\rangle|^2 \le Q(f).$$

In particular, note that this estimate applies to $f(y) = G(\lambda, x, y)$. Now we can proceed as in the proof of the previous theorem (with $\lambda = 0$ and $\alpha_j = E_j^{-1}$)

$$\begin{aligned}
\sum_{j=m}^{n} |\langle u_j, f\rangle u_j(x)| =& \sum_{j=m}^{n} E_j |\langle u_j, f\rangle \langle u_j, G(0, x, .)\rangle|\\
\le& \left(\sum_{j=m}^{n} E_j |\langle u_j, f\rangle|^2 \sum_{j=m}^{n} E_j |\langle u_j, G(0, x, .)\rangle|^2 \right)^{1/2}\\
<& Q(f)^{1/2} Q(G(0, x, .))^{1/2},
\end{aligned}$$

where we have used the Cauchy–Schwarz inequality for the weighted scalar product $(f_j, g_j) \mapsto \sum_j f_j^* g_j E_j$. Finally note that $Q(G(0, x, .))$ is continuous with respect to x and hence can be estimated by its maximum over $[a, b]$.

Finally, if one of the boundary terms is negative, it can still be controlled in terms of the integral using Problem 5.23. Details are again left as an exercise. □

Example. Let us look at the Sturm–Liouville problem which arose in Section 5.1,

$$L = -\frac{d^2}{dx^2}, \qquad \mathfrak{D}(L) = \{f \in C^2([0, 1], \mathbb{C}) | f(0) = f(1) = 0\}.$$

with underlying inner product space and scalar product given by

$$\mathfrak{H}_0 = C([0, 1], \mathbb{C}), \qquad \langle f, g\rangle = \int_0^1 f(x)^* g(x) dx.$$

The corresponding eigenvalues and normalized eigenfunctions are

$$\lambda_n = (\pi n)^2, \quad u_n(x) = \sqrt{2}\sin(n\pi x), \qquad n \in \mathbb{N}.$$

Moreover, every function $f \in \mathfrak{H}_0$ can be expanded into a Fourier sine series

$$f(x) = \sum_{n=1}^{\infty} f_n u_n(x), \qquad f_n = \int_0^1 u_n(x) f(x) dx,$$

which is convergent with respect to our scalar product. If we assume f piecewise continuously differentiable with $f(0) = f(1) = 0$ the series will even converge uniformly. See also Problem 5.22 for a direct proof. ◇

At first sight it might look like Theorem 5.11 answers all our questions concerning Sturm–Liouville problems. Unfortunately this is not true since the assumptions we have imposed on the coefficients are too restrictive for some important applications! 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 4.14) 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 (5.55) makes no sense. However, in this case it is still possible to find two solutions $u_a(z_0, x)$, $u_b(z_0, x)$ (at least for $z_0 \in \mathbb{C}\backslash\mathbb{R}$) which are square integrable near a, b and satisfy $\lim_{x\downarrow a} W_x(u_a(z_0)^*, u_a(z_0)) = 0$, $\lim_{x\uparrow b} W_x(u_b(z_0)^*, u_b(z_0)) = 0$, respectively. Introducing the boundary conditions

$$\begin{array}{rcl} BC_a(f) & = & \lim_{x\downarrow a} W_x(u_a(z_0), f) = 0 \\ BC_b(f) & = & \lim_{x\uparrow b} W_x(u_b(z_0), f) = 0 \end{array} \tag{5.78}$$

one obtains again a symmetric operator. The inverse $R_L(z)$ can be computed as before, however, the solutions $u_a(z, x)$ and $u_b(z, x)$ might not exist for $z \in \mathbb{R}$ and consequently might not be analytic in the entire complex plane.

It can be shown that Lemma 5.10 (and thus the first part of Theorem 5.11) still holds if

$$\int_a^b \int_a^b |G(z, x, y)|^2 r(x) r(y)\, dx\, dy < \infty. \tag{5.79}$$

Integral operators satisfying this estimate are known as **Hilbert–Schmidt operators**. This estimate can for example be verified in the case of Legendre's equation using the explicit behavior of solutions near the singular points ±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 led to the Fourier transform on $\mathbb{R}$.

Problem 5.15. *Compute the eigenvalues and eigenfunctions of*

$$L = -\frac{d^2}{dx^2}, \qquad \mathfrak{D}(L) = \{f \in C^2([0,1],\mathbb{C}) | f'(0) = f'(1) = 0\}.$$

Problem 5.16. *Show directly that $L = -\frac{d^2}{dx^2}$ on $I = (0, \pi)$ with Dirichlet boundary conditions is unbounded. (Hint: Consider $f_n(x) = \sin(nx)$.)*

Problem 5.17. *Show that $\mathfrak{D}(L)$ is a linear subspace invariant under complex conjugation.*

Problem 5.18. *Show that the set of infinitely differentiable functions with compact support $C_c^\infty((a,b),\mathbb{C})$ is dense in $\mathfrak{H}_0$. (Hint: Replace $P(x)$ in the proof of Lemma 5.8 by $\int_0^x \exp((y(y-1))^{-1})dy / \int_0^1 \exp((y(y-1))^{-1})dy$.)*

Problem 5.19. *Show that if f and g both satisfy $BC_a(f) = BC_a(g) = 0$, then $W_a(f, g) = 0$.*

Problem 5.20. *Show* (5.63).

Problem 5.21 (Periodic boundary conditions)**.** *Show that L defined on*

$$\mathfrak{D}(L) = \{f \in C^2([a,b],\mathbb{C}) | f(a) = f(b), p(a)f'(a) = p(b)f'(b)\} \tag{5.80}$$

is symmetric.

Problem 5.22. *Consider the Fourier sine*

$$f(x) = \sum_{n=1}^{\infty} s_n(f) \sin(n\pi x), \qquad s_n(f) = 2\int_0^1 \sin(n\pi x) f(x) dx,$$

and Fourier cosine series

$$f(x) = \sum_{n=0}^{\infty} c_n(f) \cos(n\pi x), \qquad c_n(f) = (2 - \delta_{0,n}) \int_0^1 \cos(n\pi x) f(x) dx,$$

obtained from $L = -\frac{d^2}{dx^2}$ on $[0, 1]$ with Dirichlet and Neumann boundary conditions, respectively.

For given $k \in \mathbb{N}_0$, show that

$$\sum_{n=1}^{\infty} n^k |s_n(f)| < \infty$$

if $f \in C^{k+1}([0,1],\mathbb{C})$ with $f^{(2j)}(0) = f^{(2j)}(1) = 0$ for $0 \le j \le \frac{k}{2}$. (Hint: Use integration by parts to show

$$(1+\delta_{0,n})c_n(f') = 2((-1)^n f(1) - f(0)) + n\pi s_n(f)$$

and

$$s_n(f') = -n\pi(1+\delta_{0,n})c_n(f).$$

Now observe that for $g \in C([0,1],\mathbb{C})$, both $s_n(g)$ and $c_n(g)$ are square summable (by the Bessel inequality). Moreover, the sequence n^{-1} is also square summable and the product of two square summable is (absolutely) summable by the Cauchy–Schwarz inequality.)

Problem 5.23. *Suppose $f \in C_p^1[a,b]$. Show that for every $\varepsilon > 0$*

$$|f(x)|^2 \le \varepsilon \int_a^b |f'(x)|^2\, dx + \left(\frac{1}{\varepsilon} + \frac{1}{b-a}\right) \int_a^b |f(x)|^2\, dx.$$

(Hint: $\frac{d}{dx}|f(x)|^2 = 2\mathrm{Re}\big(f(x)f'(x)\big) \le 2|f(x)f'(x)|$.)

5.5. Oscillation theory

In this section we want to gain further insight by looking at the zeros of the eigenfunctions of a Sturm–Liouville equation. If you look at the simplest Sturm–Liouville equation $r = p = 1$ and $q = 0$, the solutions are trigonometric functions for $\lambda > 0$ and if you plot the solution in phase space, that is, the solutions of the underlying first order system (5.44) given by $(u(x), p(x)u'(x)) \in \mathbb{R}^2$, they will rotate around the origin. It turns out that this behavior is quite common for Sturm–Liouville equations and in order to investigate this further we introduce polar coordinates in phase space which are known as **Prüfer variables**:

$$u(x) = \rho_u(x)\sin(\theta_u(x)), \qquad p(x)u'(x) = \rho_u(x)\cos(\theta_u(x)). \tag{5.81}$$

Clearly the Prüfer radius is given by

$$\rho_u(x) = \sqrt{u(x)^2 + (p(x)u'(x))^2} \tag{5.82}$$

and the Prüfer angle is

$$\theta_u(x) = \mathrm{atan2}(p(x)u'(x), u(x)) \mod 2\pi, \tag{5.83}$$

where

$$\mathrm{atan2}(x,y) = \begin{cases} \arccos(\frac{x}{\sqrt{x^2+y^2}}), & y \ge 0, \\ -\arccos(\frac{x}{\sqrt{x^2+y^2}}), & y < 0. \end{cases} \tag{5.84}$$

For (5.83) to make sense we of course need to assume $\rho_u(x) \neq 0$ but if $\rho_u(x_0) = 0$ we have $u(x_0) = p(x_0)u'(x_0) = 0$ and hence $u \equiv 0$ by uniqueness. Since the trivial solution $u \equiv 0$ is of no interest we will exclude this and assume that u is a non-trivial solution from now on. Moreover, we will also assume that all solutions are real-valued.

Moreover, the angle $\theta_u(x)$ is defined only up to multiples of 2π and if we restrict it to $(-\pi, \pi]$, as usual, it will jump from $+\pi$ to $-\pi$ at a zero of u which crosses the negative x axis from above. Since we do not want this behavior, we will choose the unknown multiple of 2π such that θ_u remains continuous. This makes θ_u unique for $x \in (a, b)$ once an initial value at some point c has been chosen.

That u satisfies $Lu = \lambda u$ is now equivalent to the system (Problem 5.24)

$$\begin{aligned} \theta_u' &= \frac{\cos(\theta_u)^2}{p} + (\lambda r - q)\sin(\theta_u)^2, \\ \rho_u' &= \rho_u\,(\frac{1}{p} + q - \lambda r)\frac{\sin(2\theta_u)}{2}. \end{aligned} \tag{5.85}$$

Observe that the equation for θ_u does not involve ρ_u and that the equation for ρ_u can be solved once θ_u is known:

$$\rho_u(x) = \rho_u(c)\exp\left(\frac{1}{2}\int_c^x (p^{-1}(t) + q(t) - \lambda r(t))\sin(2\theta_u(t))dt\right). \tag{5.86}$$

Hence we have effectively reduced our second order equation to a first order one. However, this does not come without a price: the equation for θ_u is no longer linear! Moreover, note that if we compute θ_u by solving the system (5.85), this will automatically give us the required continuous representative. Finally, note that if $\theta_u(x)$ is a solution of (5.85), then the same is true for $\theta_u(x)+n\pi$, $n \in \mathbb{Z}$, in fact, this is a Prüfer angle corresponding to $(-1)^n u(x)$.

Now, if we look at the right-hand side of the equation for θ_u we see that it will be positive if $\lambda r - q > 0$, which will always hold for sufficiently large λ. In particular, we expect θ_u to increase as λ increases and hence the solution to oscillate faster. We will come back to this in a moment, but for now observe that at a zero of u the Prüfer angle always increases:

$$u(x_0) = 0 \quad \Leftrightarrow \quad \theta_0(x_0) = 0 \mod \pi \quad \Rightarrow \quad \theta_u'(x_0) = p(x_0)^{-1} > 0. \tag{5.87}$$

In particular, the Prüfer angel can cross an integer multiple of π only from below and hence will always increase by π between two consecutive zeros. Hence we can use the integer part of θ_u/π to count the number of zeros:

Lemma 5.14. *Let u be a solution of $Lu = \lambda u$ and denote by $\#(u)$ the number of zeros of u inside (a, b). Then*

$$\#(u) = \lceil \theta_u(b)/\pi \rceil - \lfloor \theta_u(a)/\pi \rfloor - 1, \tag{5.88}$$

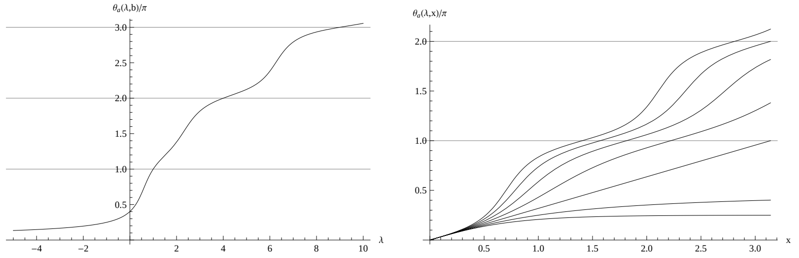


Figure 5.1. Prüfer angle $\theta_a(\lambda, x)/\pi$ as a function of λ for $x = b$ (right) and a function of x for various λ (left).

where $\lfloor x \rfloor = \max\{n \in \mathbb{Z} | n \leq x\}$, $\lceil x \rceil = \min\{n \in \mathbb{Z} | n \geq x\}$ *denote the floor, ceiling functions, respectively.*

Next we want to return to our previous observation that θ_u should increase with λ. So we consider solutions $u(\lambda, x)$ of $Lu = \lambda u$ and denote the associated Prüfer variables by $\rho_u(\lambda, x)$, $\theta_u(\lambda, x)$. In fact, note that if $u(\lambda, x)$ solves $Lu = \lambda u$ and $\lambda_1 > \lambda_0$, then $\theta_u(\lambda_1, x) > \theta_u(\lambda_0, x)$ for $x > c$ provided $\theta_u(\lambda_1, c) \geq \theta_u(\lambda_0, c)$ by Theorem 1.3. For $x < c$ the inequalities have to be reversed.

Now things get particularly interesting if we apply these findings to the solutions $u(x) = u_{a,b}(\lambda, x)$ defined in (5.62), for which we can fix the Prüfer angles by setting

$$\theta_a(\lambda, a) = \alpha \in [0, \pi), \quad -\theta_b(\lambda, b) = \pi - \beta \in [0, \pi). \tag{5.89}$$

By our findings $\theta_a(., x)$ is increasing and bounded below $\theta_a(., x) > 0$. Similarly, $\theta_b(., x)$ is decreasing and bounded above $\theta_b(., x) < 0$, or equivalently $-\theta_b(., x)$ is increasing and bounded below $-\theta_b(., x) > 0$.

The situation for $\theta_a(\lambda, x)$ is illustrated in Figure 5.1 which shows the Prüfer angle as a function of λ (for fixed $x = b$) and as a function of x for some fixed values of λ. Note that for the picture on the right, the crossings with the grid lines correspond to the case where θ_a is an integer multiple of π and hence to the zeros of $u_a(\lambda)$. Since $\theta_a(\lambda)$ increases as λ increases the zeros must move to the left and a new one will enter the interval (a, b) precisely when $u_a(\lambda, b)$ vanishes.

As $\lambda \to -\infty$ the picture seems to indicate that $\theta_a(\lambda, x)$ tends to zero. That this is indeed always the case will be shown in the following lemma.

Lemma 5.15. *We have*

$$\lim_{\lambda \downarrow -\infty} \theta_b(\lambda, x) = 0, \ x \in [a, b), \qquad \lim_{\lambda \downarrow -\infty} \theta_a(\lambda, x) = 0, \ x \in (a, b]. \tag{5.90}$$

Proof. We only do the proof for $\theta_a(x) = \lim_{\lambda \downarrow -\infty} \theta_a(\lambda, x)$. By monotonicity and $\theta_a(\lambda, x) > 0$ the limit exists and satisfies $\theta_a(x) \geq 0$.

Fix $x_0 \in (a,b]$ and consider $w(x) = \pi - (\pi - \varepsilon)\frac{x-a}{x_0-a}$ for $\varepsilon > 0$ small. Abbreviate $p_0 = \inf_{x\in[a,b]} p(x)$ and $q_0 = \inf_{x\in[a,b]} q(x)$. Then, for $\lambda < q_0 - (p_0^{-1} + \frac{\pi-\varepsilon}{x_0-a})\sin(\varepsilon)^{-2}$, we have

$$\frac{1}{p}\cos(w)^2 - (q-\lambda)\sin(w)^2 < \frac{1}{p_0} - (q_0-\lambda)\sin(\varepsilon)^2 < -\frac{\pi-\varepsilon}{x_0-a} = w'$$

for $x \in [a, x_0]$ which shows that w is a super solution. Hence by Lemma 1.2 we infer $0 \le \theta_a(x_0) \le \varepsilon$ for any ε. □

After these preparations we can now easily establish several beautiful and important results. To this end recall that $u_a(\lambda)$ is an eigenfunction if and only if it satisfies the boundary condition at b, that is, if and only if $\theta_a(\lambda, b) = \beta \mod \pi$.

First of all, Lemma 5.15 says that $\theta_a(\lambda, b)$ converges to 0 from above as $\lambda \to -\infty$ and thus will eventually drop below $\beta \in (0, \pi]$ after which it can no longer satisfy the boundary condition at b. Hence there is a lowest eigenvalue E_0 determined by the condition $\theta_a(E_0, b) = \beta$. Now as λ further increases we will hit the second eigenvalue E_1 precisely when $\theta_a(\lambda, b) = \beta+\pi$ and continuing like this we obtain

Lemma 5.16. *We have*

$$\#_{(-\infty,\lambda)}(L) = \left\lceil \frac{\theta_a(\lambda,b)-\beta}{\pi} \right\rceil = \left\lfloor \frac{\alpha - \theta_b(\lambda,a)}{\pi} \right\rfloor, \tag{5.91}$$

where $\#_{(\lambda_0,\lambda_1)}(L)$ denotes the number of eigenvalues of L inside (λ_0, λ_1).

In particular,

$$\theta_a(E_n, b) = \beta + n\pi,\ \beta \in (0,\pi], \quad \theta_b(E_n, a) = \alpha-(n+1)\pi,\ \alpha \in [0,\pi), \tag{5.92}$$

where E_n are the eigenvalues ordered in increasing size.

Moreover, in combination with Lemma 5.14 this even shows that the n'th eigenfunction has precisely n zeros. In summary we have shown:

Theorem 5.17. *The regular Sturm–Liouville problem has a lowest eigenvalue and the eigenvalues can be ordered according to $E_0 < E_1 < \cdots$. In this case the eigenfunction u_n corresponding to E_n has precisely n zeros in the interval (a, b).*

Furthermore,

Theorem 5.18. *Suppose L has a Dirichlet boundary condition at b. Then we have*

$$\#_{(-\infty,\lambda)}(L) = \#(u_a(\lambda)), \tag{5.93}$$

where $\#(u)$ is the number of zeros of u inside (a,b) and as before $\#_{(\lambda_0,\lambda_1)}(L)$ is the number of eigenvalues of L inside (λ_0,λ_1). Likewise, suppose L has a Dirichlet boundary condition at a. Then we have

$$\#_{(-\infty,\lambda)}(L) = \#(u_b(\lambda)). \tag{5.94}$$

Proof. In the first case we have $\beta = \pi$ and $\lfloor \theta_a(\lambda,a)/\pi \rfloor = \lfloor \alpha/\pi \rfloor = 0$. Hence the claim follows by combining Lemma 5.16 with Lemma 5.14. For the second claim note $\lceil \theta_b(\lambda,b)/\pi \rceil = \lceil \beta/\pi \rceil = 1$ and $\lfloor -x \rfloor = -\lceil x \rceil$. □

Up to this point we have only looked at one Sturm–Liouville operator L. However, our key to success was to look at the behavior of solutions of $(L-\lambda)u = 0$ as we vary λ. Hence one might also try to vary not only the spectral parameter λ but the entire operator. Hence we will consider two operators L_0 and L_1 associated with coefficients p_0, q_0, r_0 and p_1, q_1, r_1, respectively. We will consider solutions u_j of $L_j u_j = \lambda_j u_j$ and use the shorthand notation $\rho_j = \rho_{u_j}$, $\theta_j = \theta_{u_j}$ for the corresponding Prüfer variables.

First of all we establish monotonicity of θ with respect to the coefficients.

Lemma 5.19. *Let L_j, $j = 0,1$, be two operators associated with p_j, q_j, r_j and let u_j be solutions of $L_j u_j = \lambda_j u_j$. Suppose $p_1 \le p_0$ and $\lambda_0 r_0 - q_0 \le \lambda_1 r_1 - q_1$.*

If $\theta_1(c) \ge \theta_0(c)$ for some $c \in (a,b)$, then $\theta_1(x) \ge \theta_0(x)$ for all $x \in (c,b)$. If the inequality becomes strict at some $x \in [c,b)$ it remains so.

Moreover, if $\theta_1(c) = \theta_0(c)$ for some $c \in (a,b)$ and $\theta_1(d) = \theta_0(d)$ for some $d \in (c,b)$, then $p_1 = p_0$ and $\lambda_0 r_0 - q_0 = \lambda_1 r_1 - q_1$ on (c,d).

Proof. The first part is immediate from Theorem 1.3. Moreover, by the first part $\theta_1(c) = \theta_0(c)$ and $\theta_1(d) = \theta_0(d)$ can only happen if $\theta_1(x) = \theta_0(x)$ for all $x \in [c,d]$ and the claim follows by subtracting the corresponding differential equations for the Prüfer angles from (5.85). □

With the help of this lemma we now come to the famous **Sturm's comparison** theorem.

Theorem 5.20 (Sturm). *Let L_j, $j = 0,1$, be two operators associated with p_j, q_j, r_j and let u_j be solutions of $L_j u_j = \lambda_j u_j$. Suppose $p_1 \le p_0$ and $\lambda_0 r_0 - q_0 \le \lambda_1 r_1 - q_1$.*

If at each end of $(c,d) \subseteq (a,b)$ either $W(u_1,u_0) = 0$ or $u_0 = 0$, then u_1 must vanish in (c,d) unless u_1 and u_0 are equal up to a constant. (The latter case can only happen if $p_1 = p_0$ and $\lambda_0 r_0 - q_0 = \lambda_1 r_1 - q_1$ on (c,d).)

Proof. Without loss (and perhaps after flipping signs of u_0 and u_1) we can assume $\theta_0(c), \theta_1(c) \in [0,\pi)$. Since by assumption either $\theta_0(c) = 0$ or $\theta_0(c) = \theta_1(c)$ (cf. (5.97) below), we have $\theta_0(c) \le \theta_1(c)$. Hence Lemma 5.19

implies $\theta_0(d) < \theta_1(d)$ unless u_1 and u_0 are equal up to a constant. Now either $\theta_0(d) = 0 \mod \pi$ and thus $\pi \le \theta_0(d) < \theta_1(d)$ by (5.87) or otherwise $\theta_0(d) = \theta_1(d) \mod \pi$ and hence $\theta_0(d) + \pi \le \theta_1(d)$. □

For example, this shows that the zeros of consecutive eigenfunctions must be interlacing:

Lemma 5.21. *Let u_n be the eigenfunctions of a regular Sturm–Liouville problem ordered according to the size of the eigenvalues. Then 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*

$$a < x_{n+1,1} < x_{n,1} < x_{n+1,2} < \cdots < x_{n+1,n+1} < b. \tag{5.95}$$

Our next aim is to generalize Theorem 5.18. It will turn out hat the key object for this generalization will be the Wronski determinant of two solutions u_j of $L_j u_j = \lambda_j u_j$, $j = 0, 1$ defined as

$$W_x(u_0, u_1) = u_0(x) p_1(x) u_1'(x) - p_0(x) u_0'(x) u_1(x). \tag{5.96}$$

The connection with Prüfer angles is given by

$$W_x(u_0, u_1) = \rho_0(x)\rho_1(x)\sin(\theta_0(x) - \theta_1(x)), \tag{5.97}$$

which is straightforward to verify using the trigonometric addition formula $\sin(x-y) = \sin(x)\cos(y) - \cos(x)\sin(y)$. In particular, this last equation shows that the Wronskian will vanish $W_{x_0}(u_0, u_1) = 0$ if and only if $\theta_1(x_0) = \theta_0(x_0) \mod \pi$ which is the case if and only if both u_0 and u_1 satisfy the same boundary condition at x_0.

Of course it is tempting to relate the relative Prüfer angle

$$\Delta_{1,0} = \theta_1(x) - \theta_0(x) \tag{5.98}$$

with the numbers of zeros of the Wronskian as we did for Prüfer angles in Lemma 5.14. However, this turns out impossible. First of all the zeros of the Wronskian are not simple and could vanish on an entire interval (e.g. if both equations agree on an interval) and, even worse, $\Delta_{1,0}$ can clearly cross integer multiples of π from both sides (this reflects the fact that we can always reverse the roles of u_0 and u_1). Nevertheless we simply define

$$\#(u_0, u_1) = \lceil \Delta_{1,0}(b)/\pi \rceil - \lfloor \Delta_{1,0}(a)/\pi \rfloor - 1 \tag{5.99}$$

and call it the weighted number of sign flips of the Wronskian. In other words, we count a sign flip as $+1$ if the relative Prüfer angle crosses an integer multiple from below and as -1 if it crosses from above. Sign flips where the relative Prüfer angel does not cross but just turns around are not counted at all.

If $p_1 \le p_0$ and $\lambda_0 r_0 - q_0 \le \lambda_1 r_1 - q_1$, then Theorem 1.3 implies that $\Delta_{1,0}$ is increasing at a sign flip and hence all sign flips are counted as $+1$. This happens for example if $L_0 = L_1$ and $\lambda_1 > \lambda_0$.

As in the case of Theorem 5.18 one proves

Theorem 5.22. *Suppose L_0 and L_1 are two regular Sturm–Liouville operators associated with the same boundary conditions at a and b. Then*

$$\begin{aligned} \#_{(-\infty,\lambda_1)}(L_1) - \#_{(-\infty,\lambda_0]}(L_0) &= \#(u_{0,a}(\lambda_0), u_{1,b}(\lambda_1)) \\ &= \#(u_{0,b}(\lambda_0), u_{1,a}(\lambda_1)), \end{aligned} \tag{5.100}$$

where $\#(u_0,u_1)$ is the number of weighted sign flips of $W(u_0,u_1)$ inside (a,b) and $\#_{(-\infty,\lambda_j)}(L_j)$ is the number of eigenvalues of L_j inside $(-\infty,\lambda_j)$.

In the special case where we have only one operator L with different spectral parameters the result reads:

Corollary 5.23. *Let L be a regular Sturm–Liouville operator and $\lambda_0 < \lambda_1$. Then*

$$\#_{(\lambda_0,\lambda_1)}(L) = \#(u_a(\lambda_0), u_b(\lambda_1)) = \#(u_b(\lambda_0), u_a(\lambda_1)), \tag{5.101}$$

where $\#(u_a(\lambda_0), u_b(\lambda_1))$ is the number of sign flips of $W(u_a(\lambda_0), u_b(\lambda_1))$.

Finally, we note that given a positive differentiable function h one can modify the Prüfer variables according to

$$u(x) = \frac{\tilde\rho_u(x)}{\sqrt{h(x)}}\sin(\tilde\theta_u(x)), \qquad p(x)u'(x) = \sqrt{h(x)}\tilde\rho_u(x)\cos(\tilde\theta_u(x)). \tag{5.102}$$

That is, they are the Prüfer variables for $(\sqrt{h(x)}u(x), p(x)u'(x)/\sqrt{h(x)})$ and hence have the same properties. In particular,

$$\tilde\rho_u(x) = \sqrt{h(x)u(x)^2 + h(x)^{-1}(p(x)u'(x))^2} \tag{5.103}$$

is positive and

$$\tilde\theta_u(x) = \mathrm{atan2}(p(x)u'(x)/\sqrt{h(x)}, \sqrt{h(x)}u(x)) \mod 2\pi \tag{5.104}$$

is uniquely determined once a value of $\tilde\theta_u(c)$ is chosen by requiring $\tilde\theta_u$ to be continuous. In the special case $h \equiv 1$ we recover our original Prüfer variables, and since the modified Prüfer angle equals the original one at every zero of u as well as at every zero of u', they can differ by at most $\pi/2$:

$$\lfloor \frac{2\theta_u}{\pi} \rfloor = \lfloor \frac{2\tilde\theta_u}{\pi} \rfloor. \tag{5.105}$$

That u satisfies $Lu = \lambda u$ is now equivalent to the system

$$\begin{aligned}\tilde{\theta}_u' &= \frac{h}{p} - \frac{p^{-1}h^2 + q - \lambda r}{h} \sin(\tilde{\theta}_u)^2 + \sin(2\tilde{\theta}_u)\frac{h'}{2h},\\ \tilde{\rho}_u' &= \tilde{\rho}_u \left(\frac{p^{-1}h^2 + q - \lambda r}{2h} \sin(2\tilde{\theta}_u) + \cos(2\tilde{\theta}_u)\frac{h'}{2h}\right).\end{aligned} \tag{5.106}$$

Making appropriate choices for h one can read off the asymptotic behavior of θ_u.

Lemma 5.24. *Suppose $p\,r \in C^1$. Then the Prüfer angles $\theta_{a,b}(\lambda, x)$ satisfy*

$$\theta_a(\lambda, x) = \sqrt{\lambda}\int_a^x \sqrt{\frac{r(t)}{p(t)}}dt + O(1), \quad \theta_b(\lambda, x) = -\sqrt{\lambda}\int_x^b \sqrt{\frac{r(t)}{p(t)}}dt + O(1) \tag{5.107}$$

as $\lambda \to \infty$.

Proof. We only consider the case of θ_a. Without loss of generality we can replace the original Prüfer angles by modified ones with $h(x) = \sqrt{\lambda r(x) p(x)}$ (assuming $\lambda > 0$). Then the differential equation for $\tilde{\theta}_a$ reads

$$\tilde{\theta}_a' = \sqrt{\lambda}\sqrt{\frac{r}{p}} - \frac{q}{\sqrt{\lambda p r}} \sin(\tilde{\theta}_a)^2 + \sin(2\tilde{\theta}_a)\frac{(pr)'}{(pr)}$$

and the claim follows after integrating both sides observing $|\sin(\tilde{\theta}_a)| \le 1$. □

As a simple corollary we now obtain from (5.92) a famous result of Weyl:

Theorem 5.25 (Weyl asymptotics)**.** *Suppose $p\,r \in C^1$. Then the eigenvalues satisfy*

$$E_n = \pi^2 \left(\int_a^b \sqrt{\frac{r(t)}{p(t)}}dt \right)^{-2} n^2 + O(n). \tag{5.108}$$

For another application of the modified Prüfer transformation to obtain asymptotics for large x see Problem 5.31.

We conclude this section by mentioning that the results presented here can be extended to Sturm–Liouville equations on unbounded intervals. Again one can show that there is a connection between the number of eigenvalues and zeros of solutions. Once the interval is unbounded, it can happen that a solution of the equation $(L - \lambda)u = 0$ has an infinite number of zeros and $L - \lambda$ is called **oscillating** in this case. Theorem 5.20 implies that this is then true for all solutions. For example, if we consider the differential equation $(L_0 - \lambda)u = -u'' - \lambda u = 0$ on $I = (0, \infty)$, then it will be oscillatory for $\lambda > 0$ and non-oscillatory for $\lambda \le 0$. In particular, the borderline case $\lambda = 0$ in combination with Sturm's comparison theorem implies that any perturbation $L = L_0 + q$ will be (non-)oscillatory if $q(x)$ is eventually positive (negative). If $q(x) \to 0$ as $x \to \infty$ the following refinement can be applied.

Theorem 5.26 (Kneser). *Consider the differential equation $Lu(x) = -u''(x) + q(x)u(x)$ on $(0,\infty)$. Then*

$$\liminf_{x\to\infty} \big(x^2 q(x)\big) > -\frac{1}{4} \text{ implies nonoscillation of } L \tag{5.109}$$

and

$$\limsup_{x\to\infty} \big(x^2 q(x)\big) < -\frac{1}{4} \text{ implies oscillation of } L. \tag{5.110}$$

Proof. The key idea is that the equation

$$L_\mu u(x) = -u''(x) + \frac{\mu}{x^2} u(x) = 0$$

is of Euler type. Hence it is explicitly solvable with a fundamental system given by

$$u_\pm(x) = x^{\frac{1}{2}\pm\sqrt{\mu+\frac{1}{4}}}.$$

There are two cases to distinguish. If $\mu \ge -1/4$, all solutions are nonoscillatory. If $\mu < -1/4$, one has to take real/imaginary parts and all solutions are oscillatory. Hence a straightforward application of Sturm's comparison theorem between L_μ and L yields the result. □

Problem 5.24. *Prove equation* (5.85).

Problem 5.25. *Prove Lemma 5.21.*

Problem 5.26. *Consider the equation $-u'' + qu = 0$ with $q > 0$. Show that every nontrivial solution has at most one zero.*

Problem 5.27. *Consider the equation $-u'' + qu = 0$ and suppose $q_0 \le q(x) \le q_1 < 0$. Show that for two consecutive zeros x_k and x_{k+1} of $u(x)$ we have*

$$\frac{\pi}{\sqrt{-q_0}} \le x_{k+1} - x_k \le \frac{\pi}{\sqrt{-q_1}}.$$

Problem 5.28. *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*

$$|u(x_k)| \le |u(x_{k+1})| \qquad \text{if} \quad (pq)' \ge 0.$$

Hint: Consider

$$u^2 - \frac{1}{pq}(pu')^2.$$

Problem 5.29. *Consider the ordered eigenvalues $E_n(\alpha)$ of our Sturm–Liouville problem as a function of the boundary parameter α. Show that the eigenvalues corresponding to different parameters are interlacing. That is, suppose $0 < \alpha_1 < \alpha_2 \le \pi$ and show $E_n(\alpha_1) < E_n(\alpha_2) < E_{n+1}(\alpha_1)$.*

Problem 5.30. *Show that the derivative of the Wronskian of two solutions* u_j *of* $L_j u_j = \lambda_j u_j$ *is given by*

$$W'(u_0, u_1) = (q_1 - \lambda_1 - q_0 + \lambda_0) u_0 u_1 + \left(\frac{1}{p_0} - \frac{1}{p_1} \right) p_0 u_0' p_1 u_1'. \qquad (5.111)$$

Problem 5.31. *Show that solutions of the Bessel equation* (4.59) *have the asymptotics*

$$u(x) = a \left(\frac{\sin(x+b)}{x^{1/2}} - \frac{(1/4 - \nu^2)\cos(x+b)}{2x^{3/2}} + O(x^{-5/2}) \right),$$
$$u'(x) = a \left(\frac{\cos(x+b)}{x^{1/2}} + \frac{(1/4 - \nu^2)\sin(x+b)}{2x^{3/2}} + O(x^{-5/2}) \right).$$

(Hint: Show that after the transformation $v(x) = \sqrt{x} u(x)$ *the Bessel equation reads (cf. Problem 4.13)*

$$-v''(x) + q(x) v(x) = 0, \qquad q(x) = -1 - \frac{1/4 - \nu^2}{x^2}.$$

Now use a modified Prüfer transform with $h(x) = \sqrt{-q(x)}$ *(set* $p(x) = 1$, $r(x) = 0$*) and verify*

$$\tilde{\theta}_v'(x) = 1 + \frac{1/4 - \nu^2}{2x^2} + O(x^{-3}), \quad \frac{\tilde{\rho}_v'(x)}{\tilde{\rho}_v(x)} = O(x^{-3}),$$

as $x \to \infty$*.)*

Problem 5.32. *Solve the initial value problem*

$$y' = \cos(y)^2 + \frac{1}{3x^2} \sin(y)^2, \qquad y(1) = 0,$$

numerically and make a guess what the limit $\lim_{x\to\infty} y(x)$ *will be.*

Observe that the above equation is the differential equation for the Prüfer angle of a Sturm–Liouville operator which can be explicitly solved. Show that $\lim_{x\to\infty} y(x) = \infty$.

How does your numerical analysis compare to the analytic result?

5.6. Periodic Sturm–Liouville equations

In Section 3.6 we have investigated periodic differential equations and discussed Hill's equation as a particular example. Of course Hill's equation is a special case of a Sturm–Liouville equation for which $r(x) = p(x) = 1$. Moreover, it will not come as a surprise that our analysis can be generalized to arbitrary Sturm–Liouville equations in a straightforward manner. In fact, using the results of the previous sections we will be able to say much more about the spectrum of Hill's equation.

We will now suppose that $r(x)$, $p(x)$, and $q(x)$ are ℓ-periodic functions throughout this section.

Let us begin by recalling what we already know. Denote by

$$\Pi(z,x,x_0) = \begin{pmatrix} c(z,x,x_0) & s(z,x,x_0) \\ p(x)c'(z,x,x_0) & p(x)s'(z,x,x_0) \end{pmatrix} \tag{5.112}$$

the principal matrix solution of the underlying system (5.44) introduced in (5.46). Since the base point will not play an important role we will just set it equal to $x_0 = 0$ and write $c(z,x) = c(z,x,0)$, $s(z,x) = s(z,x,0)$. Moreover, recall that the determinant of $\Pi(z,x,0)$ equals one, (5.48).

In Section 3.6 we have introduced the **monodromy matrix** $M(z) = \Pi(z,\ell,0)$ and its eigenvalues, the **Floquet multipliers**,

$$\rho_\pm(z) = \Delta(z) \pm \sqrt{\Delta(z)^2 - 1}, \qquad \rho_+(z)\rho_-(z) = 1. \tag{5.113}$$

We will choose the branch of the square root such that $|\rho_+(z)| \le 1$. Here the **Floquet discriminant** is given by

$$\Delta(z) = \frac{\operatorname{tr}(M(z))}{2} = \frac{c(z,\ell) + p(\ell)s'(z,\ell)}{2}. \tag{5.114}$$

Moreover, we have found two solutions

$$u_\pm(z,x) = c(z,x) + m_\pm(z)s(z,x), \tag{5.115}$$

the **Floquet solutions**, satisfying

$$\begin{pmatrix} u_\pm(z,\ell) \\ p(\ell)u'_\pm(z,\ell) \end{pmatrix} = \rho_\pm(z) \begin{pmatrix} u_\pm(z,0) \\ p(0)u'_\pm(z,0) \end{pmatrix} = \rho_\pm(z) \begin{pmatrix} 1 \\ m_\pm(z) \end{pmatrix}. \tag{5.116}$$

Here

$$m_\pm(z) = \frac{\rho_\pm(z) - c(z,\ell)}{s(z,\ell)} = \frac{p(\ell)c'(z,\ell)}{\rho_\pm(z) - p(\ell)s'(z,\ell)} \tag{5.117}$$

are the **Weyl–Titchmarsh m-functions**. Note that at a point z with $s(z,\ell) = 0$, the functions $m_\pm(z)$ and hence also $u_\pm(z,x)$ are not well defined. This is due to our normalization $u_\pm(z,0) = 1$ which is not possible if the first component of the eigenvector of the monodromy matrix vanishes.

Lemma 5.27. *The zeros of $s(z,\ell)$ are all real and only occur when $|\Delta| \ge 1$.*

Proof. Since the zeros are eigenvalues of a Strum–Liouville problem with two Dirichlet boundary conditions, they are real. Moreover, at such a zero $s(z,x)$ is a Floquet solution corresponding to the real Floquet multiplier $p(\ell)s'(z,\ell)$. □

The Wronskian of u_+ and u_- is given by

$$W(u_-(z), u_+(z)) = m_+(z) - m_-(z) = \frac{2\sqrt{\Delta(z)^2 - 1}}{s(z,\ell)} \tag{5.118}$$

and hence they are linearly independent if $\Delta(z) \neq \pm 1$. If $\Delta(z) = \pm 1$ both functions are clearly equal.

The functions $u_\pm(z, x)$ are exponentially decaying as $x \to \pm\infty$ if $|\rho_+(z)| < 1$, that is, $|\Delta(z)| > 1$, and are bounded if $|\rho_+(z)| = 1$, that is, $|\Delta(z)| \leq 1$. The stability set

$$\Sigma = \{\lambda \in \mathbb{R} \,|\, |\Delta(\lambda)| \leq 1\} \tag{5.119}$$

is also called the spectrum. Our goal is to understand the stability set. A critical role is given by the points with $\Delta(\lambda) = \pm 1$ which are precisely the spectra of the **periodic L_+ and antiperiodic L_- operators** associated with (5.43) and the following domains

$$\mathfrak{D}(L_\pm) = \{f \in C^2([0, \ell], \mathbb{C}) | f(\ell) = \pm f(0), p(\ell) f'(\ell) = \pm p(0) f'(0)\}. \tag{5.120}$$

Theorem 5.28. *The spectrum of $L_\pm$ is given by*

$$\sigma(L_\pm) = \{\lambda \in \mathbb{R} \,|\, \Delta(\lambda) = \pm 1\} \tag{5.121}$$

and consist of a sequence of real eigenvalues with no finite accumulation point.

Proof. By definition of the boundary conditions for $\mathfrak{D}(L_\pm)$ we see that $z \in \mathbb{C}$ is an eigenvalue of $L_\pm$ if and only if ± 1 is an eigenvalue of the monodromy matrix, that is, if and only if $\Delta(z) = \pm 1$. As in Section 5.4 one can show that $L_\pm$ is a symmetric operator with compact resolvent (Problem 5.33) and hence the claim follows. □

Note that an eigenvalue of $L_\pm$ is simple if the monodromy matrix has just one eigenvector and twice degenerate if the monodromy matrix has two linearly independent eigenvectors.

First of all note, that there are no eigenvalues of $L_\pm$ below some λ_0.

Lemma 5.29. *We have $\Delta(\lambda) > 1$ for $\lambda < \lambda_0$, where*

$$\lambda_0 = \min_{x \in [0,\ell]} \frac{q(x)}{r(x)}. \tag{5.122}$$

Proof. Let $\lambda < \lambda_0$. Then $q - \lambda r > 0$ and any solution u of (5.43) with $z = \lambda$ satisfies $(pu')' = q - \lambda r > 0$. Hence pu' is increasing, that is, $p(x)u'(x) > p(0)u'(0)$ for $x > 0$. Moreover, if $p(0)u'(0) \geq 0$, then u is also increasing, that is, $u(x) > u(0)$ for $x > 0$. In particular, $c(\lambda, x) > c(\lambda, 0) = 1$ and $p(x)s'(\lambda, x) > p(0)s'(\lambda, 0) = 1$ for $x > 0$. □

To investigate the derivative of $\Delta(z)$ at a point where $\Delta(z) = \pm 1$ we first derive a practical formula for $\dot{\Delta}(z)$.

Lemma 5.30. *We have*

$$\begin{aligned}\dot{\Delta}(z) =& -\frac{s(z,\ell)}{2}\int_0^\ell u_+(z,x)u_-(z,x)r(x)dx\\ =&\frac{1}{2}\int_0^\ell \Big(p(\ell)c'(z,\ell)s(z,x)^2 + \big(c(z,\ell)-p(\ell)s'(z,\ell)\big)s(z,x)c(z,x)\\ & - s(z,\ell)c(z,x)^2\Big)r(x)dx,\end{aligned} \tag{5.123}$$

where the dot denotes a derivative with respect to z.

Proof. Let $u(z,x)$, $v(z,x)$ be two solutions of $Lu = zu$, which are smooth. Then integrating

$$W_x'(u(z_0),v(z)) = (z_0 - z)r(x)u(z_0,x)v(z,x)$$

from 0 to ℓ, dividing by $z_0 - z$ and taking $z_0 \to z$ gives

$$W_\ell(\dot{u}(z),v(z)) - W_0(\dot{u}(z),v(z)) = \int_0^\ell u(z,t)v(z,t)r(t)dt. \tag{5.124}$$

(Use constancy of the Wronskian $W_\ell(u(z),v(z)) - W_0(u(z),v(z)) = 0$ to see that the left-hand side is in fact a differential quotient). Now choose $u(z) = u_+(z)$ and $v(z) = u_-(z)$ in (5.124) and evaluate the Wronskians

$$\begin{aligned}W_\ell(\dot{u}_+(z),u_-(z)) - W_0(\dot{u}_+(z),u_-(z)) &= \dot{\rho}_+(z)\rho_-(z)W(u_+(z),u_-(z))\\ &= -\frac{\dot{\Delta}(z)}{\sqrt{\Delta(z)^2-1}}W(u_-(z),u_+(z))\end{aligned}$$

to obtain the first formula. The second follows using (5.115) plus constancy of the Wronskian $c(z,\ell)p(\ell)s'(z,\ell) - p(\ell)c'(z,\ell)s(z,\ell) = 1$. □

Corollary 5.31. *For $\lambda \in \Sigma$ with $s(\lambda,\ell) \neq 0$ we have*

$$\dot{\Delta}(\lambda) = -\frac{s(\lambda,\ell)}{2}\int_0^\ell |u_\pm(\lambda,x)|^2 r(x)dx. \tag{5.125}$$

In particular, $\dot{\Delta}(\lambda)$ is nonzero in the interior of Σ.

Proof. For $\lambda \in \Sigma$ we have $\rho_-(\lambda) = \overline{\rho_+(\lambda)}$ and consequently also $u_-(\lambda,x) = \overline{u_+(\lambda,x)}$. □

Lemma 5.32. *At a point $E \in \mathbb{R}$ with $\Delta(E) = \pm 1$ we have $\dot{\Delta}(E) = 0$ if and only if $s(E,\ell) = p(\ell)c'(E,\ell) = 0$, that is, if and only if $M(E) = \pm\mathbb{I}$. Moreover, in this case we have $\Delta(E)\ddot{\Delta}(E) < 0$.*

Proof. Suppose $\Delta(E) = \pm 1$. First of all $s(E,\ell) = p(\ell)c'(E,\ell) = 0$ is clearly equivalent to $M(E) = \pm\mathbb{I}$. Moreover, in this case the second part of (5.123) shows $\dot{\Delta}(E) = 0$ (note that we cannot use the first part since $u_\pm(E,x)$ are not well-defined if $s(E,\ell) = 0$).

Conversely, suppose $\Delta(E) = \pm 1$ and $\dot{\Delta}(E) = 0$ but $s(E,\ell) \neq 0$. Then Corollary 5.31 yields a contradiction. Thus $s(E,\ell) = 0$ as well as $c(E,\ell) = p(\ell)s'(\ell) = \Delta$ and the second part of (5.123) shows $p(\ell)c'(E,\ell) = 0$.

For the remaining part we will not display $z = E$ for notational convenience. Differentiating (5.123) and evaluating at point $z = E$ with $M(E) = \pm\mathbb{I}$ shows

$$\ddot{\Delta} = \frac{1}{2}\int_0^\ell \Big(p(\ell)\dot{c}'(\ell)s(x)^2 + (\dot{c}(\ell) - p(\ell)\dot{s}'(\ell))s(x)c(x) - \dot{s}(\ell)c(x)^2\Big)r(x)dx.$$

Furthermore, choosing $u = v = s$ in (5.124) shows

$$W_\ell(\dot{s}, s) = \int_0^\ell s(x)^2 r(x)dx \tag{5.126}$$

and by $s(\ell) = 0$, $p(\ell)s'(\ell) = \pm 1$ we have

$$\dot{s}(\ell) = \pm\int_0^\ell s(x)^2 r(x)dx.$$

Similarly, we obtain

$$p(\ell)\dot{c}'(\ell) = \mp\int_0^\ell c(x)^2 r(x)dx, \quad \dot{c}(\ell) = -p(\ell)\dot{s}'(\ell) = \pm\int_0^\ell s(x)c(x)r(x)dx.$$

Hence

$$\Delta\ddot{\Delta} = \Big(\int_0^\ell s(x)c(x)r(x)dx\Big)^2 - \Big(\int_0^\ell s(x)^2 r(x)dx\Big)\Big(\int_0^\ell c(x)^2 r(x)dx\Big)$$

and since equality in the Cauchy–Schwarz inequality can only occur if $c(x)$ and $s(x)$ were linearly dependent, the right-hand side is strictly negative. □

In summary, these results establish the following behavior of $\Delta(z)$: By Lemma 5.29 $\Delta(\lambda)$ will first hit $+1$ at some point E_0. At this point we must have $\dot{\Delta}(E) < 0$. In fact, $\dot{\Delta}(E) = 0$ would imply $\ddot{\Delta}(E) < 0$ by Lemma 5.32, contradicting the fact that we intersect the line $+1$ from above. By Corollary 5.31 $\Delta(\lambda)$ cannot turn around until it hits -1 at some point $E_1 > E_0$. Now it can either cross ($\dot{\Delta}(E_1) < 0$) or turn around ($\dot{\Delta}(E_1) = 0$, $\ddot{\Delta}(E_1) > 0$). In the first case it will hit -1 again at some later point E_2, in the latter case we can just set $E_2 = E_1$ (in this case $E_1 = E_2$ is a twice degenerate eigenvalue of L_-). Since there is an infinite number of periodic (antiperiodic) eigenvalues (Problem 5.33), this process can never stop and we obtain:

Theorem 5.33. *There is a sequence of real numbers*

$$E_0 < E_1 \leq E_2 < E_3 \leq E_4 \cdots \tag{5.127}$$

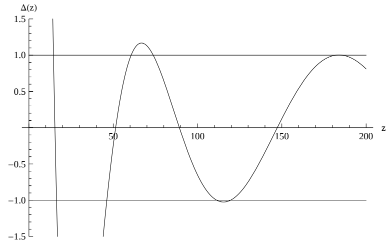


Figure 5.2. A Floquet discriminant $\Delta(z)$.

tending to ∞ *such that*

$$\Sigma = \bigcup_{n=0}^{\infty} [E_{2n}, E_{2n+1}]. \tag{5.128}$$

Moreover,

$$\sigma(L_+) = \{E_0 < E_3 \le E_4 < E_7 \le E_8 < \cdots\} \tag{5.129}$$

and

$$\sigma(L_-) = \{E_1 \le E_2 < E_5 \le E_6 < \cdots\}. \tag{5.130}$$

That is, the spectrum consist of an infinite sequence of bands, some of which might touch. In fact, if $q = 0$ we get $\Delta(z) = \cos(\sqrt{z})$ and all bands touch, so $\Sigma = [0, \infty)$. A prototypical discriminant is depicted in Figure 5.2.

There are even further connections with the spectra of the operators associated with (5.43) and the domains

$$\begin{aligned}\mathfrak{D}(L_\alpha) = \{f \in C^2([0,\ell], \mathbb{C}) |\, & \cos(\alpha) f(0) - \sin(\alpha) p(0) f'(0) = \\ & \cos(\alpha) f(\ell) - \sin(\alpha) p(\ell) f'(\ell) = 0\}. \end{aligned} \tag{5.131}$$

As a preparation we show

Lemma 5.34. *All singularities of* $m_\pm(z)$ *are one the real line where* $|\Delta(z)| \ge 1$*. If a singularity occurs at a point with* $\Delta(z) = \pm 1$ *then* $c(z,\ell) = p(\ell) s'(z,\ell) = \pm 1$ *and both* $m_\pm(z)$ *have a square root type singularity. Otherwise, if a singularity occurs at a point with* $|\Delta(z)| > 1$ *then precisely one of the functions* $m_\pm(z)$ *has a first order pole.*

Proof. By (5.117) singularities of $m_\pm(z)$ can only occur at zeros of $s(\ell, z)$ and the first claim follows from Lemma 5.27. So let μ be a zero of $s(z,\ell)$.

Then, by (5.126) we see that

$$\dot{s}(\mu,\ell) = \frac{1}{p(\ell) s'(\mu,\ell)} \int_0^\ell s(\mu,x)^2 r(x) dx \neq 0$$

and thus μ is a first order zero.

If $\Delta(\mu) = \pm 1$ then we must have $\rho_\pm(\mu) = \pm 1$ and $c(\mu, \ell) = p(\ell)s'(\mu, \ell) = \pm 1$. Hence both $m_\pm(z)$ have a square root type singularity by (5.117).

If $|\Delta(\mu)| > 1$ then the numerator of the first equation in (5.117) can vanish for at most one sign and hence there is at least one pole. Similarly, the denominator of the second equation in (5.117) can vanish for at most one sign and thus there is at most one pole. □

Lemma 5.35. *We have*

$$\dot{m}_\pm(z) = \int_0^{\pm\infty} u_\pm(z,x)^2 r(x) dx, \qquad z \in \mathbb{C}\backslash\Sigma. \tag{5.132}$$

Proof. Integrate

$$W_x'(u_\pm(z_0), u_\pm(z)) = (z_0 - z) r(x) u_\pm(z_0, x) u_\pm(z, x)$$

from 0 to ±∞:

$$-W_0(u_\pm(z_0), u_\pm(z)) = (z_0 - z) \int_0^{\pm\infty} u_\pm(z_0, x) u_\pm(z, x) r(x) dx.$$

Now divide by $z_0 - z$,

$$\begin{aligned}\frac{1}{z_0 - z} W_0(u_\pm(z_0), u_\pm(z)) &= W_0(\frac{u_\pm(z_0) - u_\pm(z)}{z_0 - z}, u_\pm(z)) \\ &= -\int_0^{\pm\infty} u_\pm(z_0, x) u_\pm(z, x) r(x) dx,\end{aligned}$$

and let $z_0 \to z$ (use dominated convergence, c.f. Theorem 9.13),

$$W_0(\dot{u}_\pm(z), u_\pm(z)) = -\int_0^{\pm\infty} u_\pm(z,x)^2 r(x) dx.$$

Finally $W_0(\dot{u}_\pm(z), u_\pm(z)) = -\dot{m}_\pm(z)$ finishes the proof. □

Theorem 5.36. *Denote the spectrum of L_α, $\alpha \in [0, \pi)$, by*

$$\sigma(L_\alpha) = \{\lambda_0(\alpha) < \lambda_1(\alpha) < \cdots\}, \qquad \alpha \neq 0 \tag{5.133}$$

and

$$\sigma(L_0) = \{\lambda_1(0) < \lambda_2(0) < \cdots\}. \tag{5.134}$$

Then there is a one-to-one correspondence between $(-\infty, E_0]$ and $\bigcup_{\alpha\in(0,\pi)} \lambda_0(\alpha)$ respectively $[E_{2j-1}, E_{2j}]$ and $\bigcup_{\alpha\in[0,\pi)} \lambda_j(\alpha)$ for $j \in \mathbb{N}$ with $E_{2j-1} < E_{2j}$. If $E_{2j-1} = E_{2j}$ we have $\lambda_j(\alpha) = E_{2j-1} = E_{2j}$ for all $\alpha \in [0, \pi)$.

Proof. First of all note that $\lambda \in \sigma(L_\alpha)$ if and only if $m_+(\lambda) = \cot(\alpha)$ or $m_-(\lambda) = \cot(\alpha)$ since the corresponding eigenfunction will give rise to an eigenvector of the monodromy matrix and vice versa.

Hence it suffices to show that $m_-(\lambda)$ and $m_+(\lambda)$ traverse all values in $\mathbb{R} \cup \{\infty\}$ when λ runs from E_{2j-1} to E_{2j}. Essentially this follows from monotonicity of $m_\pm(\lambda)$ in these regions (Lemma 5.35) plus the fact that they

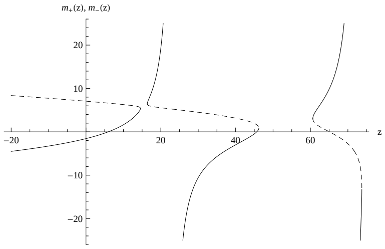


Figure 5.3. The Weyl functions $m_+(z)$ (solid) and $m_-(z)$ (dashed) inside the first three gaps.

must be equal at the boundary points. While this looks like a contradiction at first sight (if both start at the same point and one is increasing, one is decreasing, they can never meet again), it turns out to be none since $m_\pm$ can (and will) have poles! The prototypical situation is depicted in Figure 5.3. Our aim is to prove that the picture always looks like this.

We start with $\lambda \in (-\infty, E_0)$. For $\lambda < \lambda_0$ the proof of Lemma 5.29 shows that $s(\lambda, \ell) > 0$ which together with $\rho_+(\lambda) < 1 < \rho_-(\lambda)$ implies $m_+(\lambda) < m_-(\lambda)$. Now as λ increases, $m_+(\lambda)$ increases and $m_-(\lambda)$ decreases. Since they cannot cross before $\lambda = E_0$ (by linear independence of $u_+(\lambda, x)$ and $u_-(\lambda, x)$), they will meet precisely at $\lambda = E_0$. To see that $m_\pm(\lambda) \to \mp\infty$ as $\lambda \to -\infty$ one observes that $m_\pm(\lambda) = \cot(\theta_\pm(\lambda, 0))$, where $\theta_\pm(\lambda, x)$ is the Prüfer angle of $u_\pm(\lambda, x)$. As in the proof of Lemma 5.15 one shows that $\theta_\pm(\lambda, x)$ converges to a multiple of π and this finishes the case $\lambda \in (-\infty, E_0)$.

If $E_{2j-1} = E_{2j}$ all solutions are (anti)periodic and hence any solution satisfying the boundary condition at 0, also satisfies the same boundary at ℓ. In other words, $\lambda_j(\alpha) = E_{2j-1} = E_{2j}$ for all $\alpha \in [0, \pi)$.

If $\lambda \in (E_{2j-1}, E_{2j})$ there are two cases, either $m_-(\lambda) < m_+(\lambda)$ or $m_-(\lambda) > m_+(\lambda)$. The case $s(\lambda, \ell) = 0$ can always be avoided by moving λ a little. We only do the first case $m_-(\lambda) < m_+(\lambda)$, since the second is completely analogous. As λ increases, $m_-(\lambda)$ decreases and $m_+(\lambda)$ increases, and both will hit at E_{2j}. As λ decreases, $m_-(\lambda)$ increases and $m_+(\lambda)$ decreases. Now if there is no pole in (E_{2j-1}, λ), they cannot meet at a finite value $m_-(E_{2j-1}) = m_+(E_{2j-1})$ and thus $m_\pm(E_{2j-1}) = \infty$, that is $s(E_{2j-1}, \ell) = 0$. Otherwise precisely one of them has a pole and after this pole we have $m_+ > m_-$. Since they cannot cross, there cannot be another pole and they must hit at some finite value at E_{2j-1}. □

As a simple consequence we obtain

Theorem 5.37. *The lowest periodic eigenvalue E_0 is simple and the corresponding eigenfunction $u(E_0, x)$ has no zeros. The antiperiodic eigenfunctions $u(E_{4j-3}, x)$, $u(E_{4j-2}, x)$ have $2j-1$ zeros in $[0,\ell)$ and the periodic eigenfunctions $u(E_{4j-1}, x)$, $u(E_{4j}, x)$ have $2j$ zeros in $[0,\ell)$.*

Proof. First of all note that a periodic eigenfunction must have an even number of zeros and an antiperiodic eigenfunction must have an odd number of zeros (why?). Moreover, by Theorem 5.18 the eigenfunction corresponding to $\lambda_j(0)$ has precisely $j-1$ zeros.

Sturm's comparison theorem (Theorem 5.20) implies that any solution $u(\lambda, x)$ with $\lambda_j(0) \le \lambda \le \lambda_{j+1}(0)$ has at least $j-1$ and at most j zeros. Since $\lambda_j(0) \le E_{2j-1} < E_{2j} \le \lambda_{j+1}(0)$ the claim on the number of zeros follows.

If E_0 is twice degenerate, we could take a linear combination of two linearly independent eigenfunctions, to obtain an eigenfunction which vanishes at 0. By periodicity this function must also vanish at ℓ. Hence it is an eigenfunction of L_0 implying $\lambda_1(0) = E_0$ contradicting $E_0 < E_1 \le \lambda_1(0)$. □

Problem 5.33 (Periodic and antiperiodic spectra)**.**

(i) *Show that $L_\pm$ are symmetric.*

(ii) *Show that the corresponding Green function is given by*

$$G_\pm(z,x,y) = \begin{cases} \frac{1}{1\mp\rho_+(z)} u_+(z,x)u_-(z,y) + \frac{1}{1\mp\rho_-(z)} u_-(z,x)u_+(z,y), & y < x, \\ \frac{\rho_+(z)}{1\mp\rho_+(z)} u_+(z,y)u_-(z,x) + \frac{\rho_-(z)}{1\mp\rho_-(z)} u_-(z,y)u_+(z,x), & y > x. \end{cases}$$

Conclude that the periodic and antiperiodic eigenvalues form a sequence of real numbers which converge to ∞.

(iii) *Show by example that the periodic, antiperiodic eigenvalues are not necessarily simple. (Hint: $r = p = 1$ and $q = 0$.)*

Problem 5.34. *Show:*

$$m_+(z) + m_-(z) = \frac{p(\ell)s'(z,\ell) - c(z,\ell)}{s(z,\ell)},$$

$$m_+(z)m_-(z) = -\frac{p(\ell)c'(z,\ell)}{s(z,\ell)},$$

$$u_+(z,x)u_-(z,x) = \frac{s(z,x+\ell,x)}{s(z,\ell)}.$$

(Hint: $s(z,x,x_0) = c(z,x_0)s(z,x) - c(z,x)s(z,x_0)$ and $\Pi(z,x+\ell,0) = \Pi(z,x+\ell,\ell)M(z) = \Pi(z,x,0)M(z)$.)

Problem 5.35 (Reflection symmetry)**.** *Suppose q is periodic $q(t+\ell) = q(t)$ and symmetric $q(-x) = q(x)$ (and set $r(x) = p(x) = 1$). Prove*

(i) $c(z,-x) = c(z,x)$ *and* $s(z,-x) = -s(z,x)$,

(ii) $c(z, x \pm \ell) = c(z,\ell)c(z,x) \pm c'(z,\ell)s(z,x)$ *and*
$s(z, x \pm \ell) = \pm s(z,\ell)c(z,x) + s'(z,\ell)s(z,x)$,

(iii) $c(z,\ell) = \dot{s}(z,\ell)$.

Problem 5.36. *A simple quantum mechanical model for an electron in a crystal leads to the investigation of*

$$-u'' + q(x)u = \lambda u, \quad \textit{where} \quad q(x+1) = q(x).$$

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 $\Sigma = \{\lambda \in \mathbb{R} | |\Delta(\lambda)| \le 1\}$ is called the spectrum of the crystal. Since $\Delta(\lambda)$ is continuous with respect to λ, the spectrum consists of bands with gaps in between.

Consider the explicit case

$$q(x) = q_0, \quad 0 \le x < \frac{1}{2}, \qquad q(x) = 0, \quad \frac{1}{2} \le x < 1.$$

Show that there are no spectral bands below a certain value of λ. Show that there is an infinite number of gaps if $q_0 \ne 0$. 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 $a \to \infty$.)

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, such as the motion of a few planets under the influence of their respective gravitational forces. Usually you want to know the fate of the system for long times, for instance, 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, for example, 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.

We begin with the definition: A **dynamical system** is a semigroup G with identity element e acting on a set 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

$$T_g \circ T_h = T_{g \circ h}, \qquad T_e = \mathbb{I}. \tag{6.2}$$

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.

Example. The prototypical example of a discrete dynamical system is an iterated map. Let f map an interval I into itself and consider

$$T_n = f^n = f \circ f^{n-1} = \underbrace{f \circ \cdots \circ f}_{n \text{ times}}, \qquad G = \mathbb{N}_0. \tag{6.5}$$

Clearly, if f is invertible, so is the dynamical system if we extend this definition for $n \in \mathbb{Z}$ 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 systems bear a rich mathematical structure with lots of unresolved problems. ◇

Example. 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

$$\dot{x} = f(x), \qquad x(0) = x_0. \tag{6.7}$$

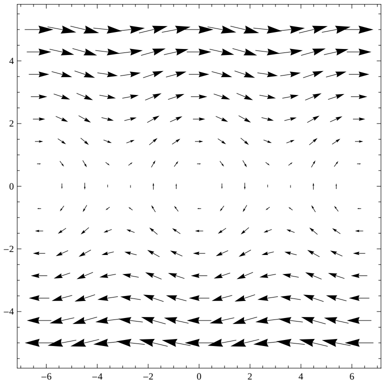
Throughout the rest of this book we will assume $f \in C^k(M, \mathbb{R}^n)$, $k \geq 1$, where M is an open subset of $\mathbb{R}^n$.

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 what the solutions look like, we can simply plot the corresponding vector field.

Example. Using *Mathematica* the vector field of the mathematical pendulum, $f(x, y) = (y, -\sin(x))$, can be plotted as follows.

```
In[1]:= VectorPlot[{y, −Sin[x]}, {x, −2π, 2π}, {y, −5, 5}]
```

Out[1]=



We will return to this example in Section 6.7. ◇

In particular, solutions of the IVP (6.7) are also called **integral curves** or **trajectories**. We will say that ϕ is an integral curve at x_0 if it satisfies $\phi(0) = x_0$.

By Theorem 2.13 there is a (unique) **maximal integral curve** ϕ_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 the maximal integral curve at x, then $\phi(. + s)$ is the maximal integral curve at $y = \phi(s)$ and in particular $I_x = s + I_y$. 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 $t \in I_{\Phi(s,x)} = I_x - s$.

Theorem 6.1. *Suppose $f \in C^k(M, \mathbb{R}^n)$. 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,*

$$\begin{aligned} \Phi(0, x) &= x, \\ \Phi(t + s, x) &= \Phi(t, \Phi(s, x)), \quad x \in M,\ s, t + s \in I_x. \end{aligned} \tag{6.11}$$

Proof. It remains 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 2.10

there is an open neighborhood $(-\varepsilon(x), \varepsilon(x)) \times U(x)$ of $(0, x)$ around each point $x \in \gamma$ such that Φ is defined and C^k on this neighborhood. Since γ is compact, finitely many of the neighborhoods $U(x)$ cover γ and hence we can find an $\varepsilon > 0$ and an open neighborhood U_0 of γ such that Φ is defined on $(-\varepsilon, \varepsilon) \times U_0$. Next, pick $m \in \mathbb{N}$ so large that $\frac{t_0}{m} < \varepsilon$ such that $K \in C^k(U_0, M)$, where $K : U_0 \to M$, $K(x) = \Phi_{\frac{t_0}{m}}$. Furthermore, $K^j \in C^k(U_j, M)$ for any $0 \le j \le m$, where $U_j = K^{-j}(U_0) \subseteq U_0$ is open. Since $x_0 = K^{-j}(\Phi(\frac{j}{m} t_0, x_0))$ we even have $x_0 \in U_j$, that is, U_j is nonempty. In particular,

$$\Phi(t, x) = \Phi(t - t_0, \Phi(t_0, x)) = \Phi(t - t_0, K^m(x))$$

is defined and C^k for all $(t, x) \in (t_0 - \varepsilon, t_0 + \varepsilon) \times U_m$. □

In particular, choosing $s = -t$ respectively $t = -s$ in (6.11) shows that $\Phi_t(.) = \Phi(t, .)$ is a local diffeomorphism with inverse $\Phi_{-t}(.)$. Note also that if we replace $f \to -f$, then $\Phi(t, x) \to \Phi(-t, x)$.

Example. Let $M = \mathbb{R}$ and $f(x) = x^3$. Then $W = \{(t, x) | 2tx^2 < 1\}$ and $\Phi(t, x) = \frac{x}{\sqrt{1-2x^2 t}}$. $T_-(x) = -\infty$ and $T_+(x) = 1/(2x^2)$. ⋄

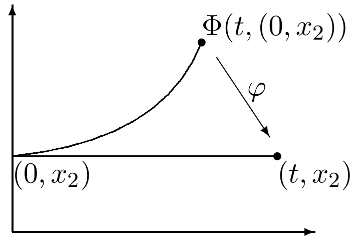
A point x_0 with $f(x_0) = 0$ is called a fixed point. Away from such points all vector fields look locally the same.

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*

$$\dot{y} = (1, 0, \dots, 0). \tag{6.12}$$

Proof. Abbreviate $\delta_1 = (1, 0, \dots, 0)$. It is no restriction to assume $x_0 = 0$. After a linear transformation we see that it is also no restriction to assume $f(0) = \delta_1$.

Consider all points starting on the plane $x_1 = 0$. Then the transform φ we are looking for should map the point $\Phi(t, (0, x_2, \dots, x_n))$ to $(0, x_2, \dots, x_n) + t(1, 0, \dots, 0) = (t, x_2, \dots, x_n)$.



Hence φ should be the inverse of

$$\psi((x_1, \dots, x_n)) = \Phi(x_1, (0, x_2, \dots, x_n)),$$

which is well defined in a neighborhood of 0. The Jacobian determinant at 0 is given by

$$\det\left(\frac{\partial\psi}{\partial x}\right)\Big|_{x=0} = \det\left(\frac{\partial\Phi}{\partial t}, \frac{\partial\Phi}{\partial x_2}, \dots, \frac{\partial\Phi}{\partial x_n}\right)\Big|_{t=0,x=0} = \det \mathbb{I} = 1$$

since $\partial\Phi/\partial x|_{t=0,x=0} = \mathbb{I}_n$ and $\partial\Phi/\partial t|_{t=0,x=0} = f(0) = \delta_1$ by assumption. So by the inverse function theorem we can assume that ψ is a local diffeomorphism and we can consider new coordinates $y = \psi^{-1}(x)$. Since $(\partial\psi/\partial x)\delta_1 = \partial\psi/\partial x_1 = f(\psi(x))$ our system reads in the new coordinates

$$\dot{y} = \left(\frac{\partial\psi}{\partial x}\right)^{-1}\Big|_{y=\psi^{-1}(x)} f(x) = \delta_1,$$

which is the required form. □

Problem 6.1. *Can*

$$\phi(t) = \begin{pmatrix} \sin(t) \\ \sin(2t) \end{pmatrix}$$

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 6.2. *Compute the flow for $f(x) = x^2$ defined on $M = \mathbb{R}$.*

Problem 6.3. *Find a transformation which straightens out the flow $\dot{x} = x$ defined on $M = \mathbb{R}$.*

Problem 6.4. *Show that $\Phi(t,x) = \mathrm{e}^t(1+x) - 1$ is a flow (i.e., it satisfies* (6.11)*). Can you find an autonomous system corresponding to this flow?*

Problem 6.5 (One-parameter Lie groups)**.** *Suppose $\Phi(t,x)$ is differentiable and satisfies* (6.11)*. Then the family $\Phi_t(x)$ is known as a local* **one-parameter Lie group of transformations** *(the term local is omitted if $W = \mathbb{R} \times M$).*

Show that Φ is the flow of the vector field

$$f(x) = \dot{\Phi}(0,x).$$

The differential operator

$$X = f(x) \cdot \mathrm{grad} = \sum_{j=1}^{n} f_j(x) \frac{\partial}{\partial x_j} \qquad (6.13)$$

is known as the **infinitesimal generator** *of $\Phi_t(x)$.*

Suppose $f(x)$ is analytic in x and recall from Theorem 4.1 that $\Phi(t,x)$ is analytic in t. Show that Φ can be recovered from X via its **Lie series**

$$\phi(t,x) = \exp(tX)x = \sum_{j=0}^{\infty} \frac{t^j}{j!} X^j x. \qquad (6.14)$$

Here the right-hand side is to be understood as the definition of $\exp(tX)x$. *(Hint: The Taylor coefficients are the derivatives which can be obtained by differentiating the differential equation.)*

Problem 6.6. *Show that* $T_+(x)$ *is lower semi-continuous:* $\liminf_{x\to x_0} T(x) \geq T(x_0)$. *Similarly,* $T_-(x)$ *is upper semi-continuous:* $\limsup_{x\to x_0} T(x) \leq T(x_0)$.

6.3. Orbits and invariant sets

The **orbit** of x is defined as

$$\gamma(x) = \Phi(I_x \times \{x\}) \subseteq M. \tag{6.15}$$

Note that $y \in \gamma(x)$ implies $y = \Phi(t, x)$ and hence $\gamma(x) = \gamma(y)$ by (6.11). 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 Φ. Otherwise x is called **regular** and $\Phi(., x) : I_x \hookrightarrow M$ is injective.

Similarly we introduce the **forward** and **backward orbits**

$$\gamma_\pm(x) = \Phi((0, T_\pm(x)), x). \tag{6.16}$$

Clearly $\gamma(x) = \gamma_-(x) \cup \{x\} \cup \gamma_+(x)$. One says that $x \in M$ is a **periodic point** of Φ 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 Φ 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 a **periodic orbit** if one (and hence all) point of the orbit is periodic.

It is not hard to see (Problem 6.9) 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 the positive (resp. negative) **lifetime** of x. A point $x \in M$ is called σ complete, $\sigma \in \{\pm\}$, if $T_\sigma(x) = \sigma\infty$ and complete if it is both $+$ and $-$ complete (i.e., if $I_x = \mathbb{R}$).

Corollary 2.15 gives us a useful criterion when a point $x \in M$ is σ 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 being complete means that Φ is globally defined, that is, $W = \mathbb{R} \times M$.

A set $U \subseteq M$ is called σ **invariant**, $\sigma \in \{\pm\}$, if

$$\gamma_\sigma(x) \subseteq U, \qquad \forall x \in U, \tag{6.17}$$

and invariant if it is both $\pm$ invariant, that is, if $\gamma(x) \subseteq U$.

If $C \subseteq M$ is a compact σ invariant set, then Lemma 6.3 implies that all points in C are σ complete.

Lemma 6.4. *(i). Arbitrary intersections and unions of σ invariant sets are σ invariant. Moreover, the closure of a σ invariant set is again σ invariant.*

(ii). If U and V are invariant, so is the complement $U \backslash V$.

Proof. Only the last statement of (i) is nontrivial. Let U be σ invariant and recall that $x \in \overline{U}$ 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 $t \in I_{x_n}$, $n \geq N$, and $\Phi(t, x) = \lim_{n\to\infty} \Phi(t, x_n) \in \overline{U}$.

Concerning *(ii)* let $x \in U \backslash V$. Then, if $\gamma(x) \cap V$ contains some point y, we must have $\gamma(y) = \gamma(x) \subseteq V$ contradicting our assumption $x \notin V$. Thus $\gamma(x) \subseteq U \backslash V$. □

One of our main aims will be to describe the long-time asymptotics of solutions. For this we next introduce the set where an orbit eventually accumulates:

The **$\boldsymbol{\omega_\pm}$-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 $\Phi(t_n, y) = \Phi(t_n, \Phi(t, x)) = \Phi(t_n + t, x)$). Hence $\omega_\pm(x)$ depends only on the orbit $\gamma(x)$. Moreover,

Lemma 6.5. *The set $\omega_\pm(x)$ is a closed invariant set.*

Proof. To see that $\omega_\pm(x)$ is closed, let y be in its closure and choose $y_n \in \omega_\pm(x)$ such that $|y - y_n| < (2n)^{-1}$ and $t_n \to \pm\infty$ such that $|\Phi(t_n, x) - y_n| < (2n)^{-1}$. Then $|\Phi(t_n, x) - y| < n^{-1}$ and thus $y \in \omega_\pm(x)$.

The set $\omega_\pm(x)$ is invariant since if $\Phi(t_n, x) \to y$ we have $\Phi(t_n + t, x) = \Phi(t, \Phi(t_n, x)) \to \Phi(t, y)$ for every $t \in I_y$. □

Example. For the equation $\dot{x} = -x$ we have $\omega_+(x) = \{0\}$ for every $x \in \mathbb{R}$, since every solution converges to 0 as $t \to \infty$. Moreover, $\omega_-(x) = \emptyset$ for $x \neq 0$ and $\omega_-(0) = \{0\}$. ◇

In particular, even for complete x the set $\omega_\pm(x)$ can be empty and we need some further assumptions in order to guarantee that this does not happen.

Lemma 6.6. *If $\gamma_\sigma(x)$ is contained in a compact set C, then $\omega_\sigma(x)$ is nonempty, compact, and connected.*

Proof. By Lemma 6.3, x is σ complete and we can choose a sequence $\Phi(t_n, x)$ with $t_n \to \sigma\infty$. By compactness we can extract a convergent subsequence and hence $\omega_\sigma(x)$ is nonempty and compact (since closed subsets of compact sets are again compact). If $\omega_\sigma(x)$ is disconnected, we can split it into two disjoint closed sets $\omega_{1,2}$. Let $\delta = \inf_{y_1\in\omega_1, y_2\in\omega_2} |y_1 - y_2| > 0$ be the distance between ω_1 and ω_2. Taking all points which are at most $\frac{\delta}{2}$ away from $\omega_{1,2}$, we obtain two disjoint neighborhoods $U_{1,2}$ of $\omega_{1,2}$, respectively. Now choose a strictly monotone sequence $t_n \to \sigma\infty$ 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. □

Under the same assumptions we can also show that the trajectory converges to its $\omega_\pm$-limit set. To this end recall that the distance between a point $x \in \mathbb{R}^n$ and a set $A \subseteq \mathbb{R}^n$ is defined by

$$d(x, A) = \inf_{y\in A} |x - y|. \tag{6.18}$$

Lemma 6.7. *Suppose $\gamma_\sigma(x)$ is contained in a compact set. Then we have $\lim_{t\to\sigma\infty} d(\Phi(t, x), \omega_\sigma(x)) = 0$.*

Proof. It suffices to show that every sequence $t_n \to \sigma\infty$ has a subsequence for which the corresponding points on the orbit converge to a point on $\omega_\sigma(x)$. But for every sequence $t_n \to \sigma\infty$ we can find a subsequence such that the corresponding points on the orbit converge to some point y by our compactness assumption. By definition of $\omega_\sigma(x)$ we must have $y \in \omega_\sigma(x)$ as required. □

Now let us consider an example which shows that the compactness requirement is indeed necessary.

Example. Let $M = \mathbb{R}^2$ and consider the vector field

$$f(x) = \begin{pmatrix} \cos(x_1)^2(\sin(x_1) - x_2\cos(x_1)) \\ \sin(x_1) + x_2\cos(x_1) \end{pmatrix}, \tag{6.19}$$

Since f is bounded it is complete by Theorem 2.17. The singularities are given by $(\mathbb{Z}\pi/2, 0)$. One further verifies that for $x \in (-\pi/2, \pi/2) \times \mathbb{R}$ we have

$$\Phi(t, x) = \begin{pmatrix} \arctan(r\mathrm{e}^{\tau(t)} \cos(\tau(t) + \theta)) \\ r\mathrm{e}^{\tau(t)} \sin(\tau(t) + \theta) \end{pmatrix}, \tag{6.20}$$

where (r, θ) are the polar coordinates of $(\tan(x_1), x_2)$ and

$$\dot{\tau}(t) = \frac{1}{\sqrt{1 + r^2 \mathrm{e}^{2\tau(t)} \cos(\tau(t))^2}}, \quad \tau(0) = 0. \tag{6.21}$$

Clearly, $\tau \in C^\infty(\mathbb{R}, \mathbb{R})$ is a diffeomorphism and hence $\omega_-(x) = (0, 0)$ and $\omega_+(x) = \{\pm\frac{\pi}{2}\} \times \mathbb{R}$ if $x \neq (0, 0)$. Moreover,

$$\Phi(t, (\pm\frac{\pi}{2}, x_2)) = \begin{pmatrix} \pm\frac{\pi}{2} \\ x_2 \pm t \end{pmatrix} \tag{6.22}$$

and hence $\omega_-((\pm\frac{\pi}{2}, 0)) = \omega_+((\pm\frac{\pi}{2}, 0)) = \emptyset$.

Thus far Φ 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 $(t, x_1, x_2) \to (-t, x_1 \pm \pi, x_2)$. ◇

A nonempty, compact, σ invariant set C is called **minimal** if it contains no proper σ invariant subset possessing these three properties. Note that for such a minimal set we have $C = \omega_+(x) = \omega_-(x)$ for every $x \in C$. One example of such a minimal set is a periodic orbit. In fact, in two dimensions this is the only example (Corollary 7.12). However, in three or more dimensions orbits can be dense on a compact hypersurface and in such a case the hypersurface cannot be split into smaller closed invariant sets.

Lemma 6.8. *Every nonempty, compact (σ) invariant set $C \subseteq M$ contains a minimal (σ) invariant set.*

If in addition C is homeomorphic to a closed m-dimensional disc (where m is not necessarily the dimension of M), it contains a fixed point.

Proof. The first part is a standard argument from general topology (cf., e.g., [**25**]). Consider the family $\mathcal{F}$ of all compact (σ) invariant subsets of C partially ordered by inclusion $\subseteq$. Every chain in $\mathcal{F}$ has a minimal element by the finite intersection property of compact sets. So by Zorn's lemma there exists a minimal member of $\mathcal{F}$.

Now let C be homeomorphic to a disc and fix $\sigma = +$ for simplicity. Pick a sequence $T_j \to 0$. 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 $0 \le t - n_j T_j < T_j$. Then

$$\Phi(t, x) = \lim_{j\to\infty} \Phi(n_j T_j, x_j) = \lim_{j\to\infty} x_j = x$$

and x is fixed. □

Problem 6.7. *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 6.8. *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 fixed point and $\lim_{t\to\infty} \dot{\phi}(t) = 0$.*

Problem 6.9 (Periodic points)**.** *Let Φ be the flow of some first-order autonomous system.*

(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 fixed 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 6.10. *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 6.11. *Denote by $d(x,A) = \inf_{y\in A} |x - y|$ the distance between a point $x \in \mathbb{R}^n$ and a set $A \subseteq \mathbb{R}^n$. Show*

$$|d(x,A) - d(z,A)| \leq |x - z|.$$

In particular, $x \mapsto d(x,A)$ is continuous.

6.4. The Poincaré map

Recall the Poincaré map used successfully in Section 1.6 for differential equations $\dot{x} = f(t,x)$, where f is periodic with respect to t, say $f(t+1,x) = f(t,x)$. To fit this equation into our current framework we consider the corresponding autonomous equation

$$\dot{y}_1 = 1,\ \dot{y}_2 = f_1(y_1, y_2, \dots, y_{n+1}),\ \dots,\ \dot{y}_{n+1} = f_n(y_1, y_2, \dots, y_{n+1}).$$

Then the idea was to look at the fate of an initial point after one period, that is we start at some initial point y and ask when it hits the plane $\Sigma = \{y | y_1 = 1\}$. This intersection was precisely our Poincaré map

$$P(y) = \Phi(1, y)$$

up to the fact that we dropped the first component $P_1(y) = \Phi_1(1, y) = y_1 + 1$ which carries no useful information and fixed $y_1 = 0$.

Our present goal is to generalize this concept for later use. To this end, 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

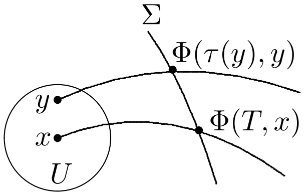
$$\Sigma = \{x \in U | S(x) = 0\}, \tag{6.23}$$

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 Σ 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 6.9. *Suppose $x \in M$ and $T \in I_x$. Let Σ be a 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*

$$\Phi(\tau(y), y) \in \Sigma \tag{6.24}$$

for all $y \in U$.



Proof. Consider the equation $S(\Phi(t, y)) = 0$ which holds for (T, x). Since

$$\frac{\partial}{\partial t} S(\Phi(t, y)) = \frac{\partial S}{\partial x}(\Phi(t, y)) f(\Phi(t, y)) \neq 0$$

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 $T = T(x)$ is its period, then

$$P_\Sigma(y) = \Phi(\tau(y), y) \tag{6.25}$$

is called **Poincaré map**. It maps Σ into itself and every fixed point corresponds to a periodic orbit. It will turn out to be an important tool in the investigation of periodic orbits.

Problem 6.12. *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.*

Problem 6.13. *At what points is* $\Sigma = \{(x,y) \in \mathbb{R}^2 | x^2 + y^2 = 1\}$ *transversal to the vector field* $f(x,y) = (y,-2x)$*?*

Problem 6.14. *The vector field* $f(x,y) = (-y,x)$ *has the periodic solution* $(\cos(t),\sin(t))$*. Compute the Poincaré map corresponding to* $\Sigma = \{(x,y) \in \mathbb{R}^2 | x > 0,\ y = 0\}$

6.5. 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 make the following definition:

A fixed point x_0 of $f(x)$ is called (Liapunov) **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 $t \geq 0$. In this respect recall that a solution remaining in a compact set exists for all positive times by Lemma 6.3. A fixed point which is not stable will be called **unstable**.

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

$$\lim_{t\to\infty} |\phi(t,x) - x_0| = 0 \quad \text{for all } x \in U(x_0). \tag{6.26}$$

Note that (6.26) does not automatically imply stability (Problem 6.16).

Finally, a fixed point x_0 of $f(x)$ is called **exponentially stable** if there are constants $\alpha, \delta, C > 0$ such that

$$|\phi(t,x) - x_0| \leq C \mathrm{e}^{-\alpha t} |x - x_0|, \qquad |x - x_0| \leq \delta. \tag{6.27}$$

Clearly (6.27) implies stability as well as (6.26).

Example. Consider $\dot{x} = ax$ in $\mathbb{R}^1$. Then $x_0 = 0$ is stable if and only if $a \leq 0$ and exponentially stable if and only if $a < 0$. ⋄

Example. The definition above of course agrees with the definition of stability for linear systems $\dot{x} = Ax$ we have introduced in Section 3.2. In particular, by Corollary 3.5 the origin is stable if and only if all eigenvalues α_j of A satisfy $\mathrm{Re}(\alpha_j) \le 0$ and for all eigenvalues with $\mathrm{Re}(\alpha_j) = 0$ the corresponding algebraic and geometric multiplicities are equal. Similarly, by Corollary 3.6 the origin is exponentially stable if and only if all eigenvalues satisfy $\mathrm{Re}(\alpha_j) < 0$. ⋄

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 (compare Section 1.5) that x_0 is stable if

$$\frac{f(x) - f(x_0)}{x - x_0} \le 0, \qquad x \in U(x_0)\backslash\{x_0\} \tag{6.28}$$

for some neighborhood $U(x_0)$ and asymptotically stable if strict inequality holds. It will be exponentially stable if

$$\frac{f(x) - f(x_0)}{x - x_0} \le -\alpha, \qquad 0 < |x - x_0| \le \delta. \tag{6.29}$$

In fact, (6.27) with $C = 1$ follows from a straightforward sub/super solution argument by comparing with solutions of the linear equation $\dot{y} = -\alpha y$.

In particular, if $f'(x_0) \ne 0$ the stability can be read of from the derivative of f at x_0 alone (cf. Problem 6.15). Moreover, Corollary 3.27 implies that a fixed point is exponentially stable if this is true for the corresponding linearized system.

Theorem 6.10 (Exponential stability via linearization)**.** *Suppose $f \in C^1$ has a fixed point x_0 and suppose that all eigenvalues of the Jacobian matrix at x_0 have negative real part. Then x_0 is exponentially stable.*

However, if $f'(x_0) = 0$, no information on the stability of the nonlinear system can be read off from the linearized one as can be seen from the following example.

Example. The equation

$$\dot{x} = \mu x^3 \tag{6.30}$$

is asymptotically stable for $\mu < 0$, stable for $\mu \le 0$, and unstable for $\mu > 0$. ⋄

In $\mathbb{R}^n$, $n > 1$, the equation $\dot{x} = f(x)$ 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 μ. By Theorem 2.11 the flow depends smoothly on the parameter μ (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

$$\dot{x} = \mu x - x^3 \tag{6.31}$$

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

$$\dot{x} = \mu x - x^2 \tag{6.32}$$

has two stable fixed points for $\mu \neq 0$ which collide and exchange stability at $\mu = 0$. This is known as **transcritical bifurcation**. The system

$$\dot{x} = \mu + x^2 \tag{6.33}$$

has one stable and one unstable 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, μ_0) if $f(x_0, \mu_0) = 0$ and $\frac{\partial f}{\partial x}(x_0, \mu_0) = 0$.

Problem 6.15. *Suppose $f \in C^1(\mathbb{R})$. Show directly that a fixed point x_0 is exponentially stable if $f'(x_0) < 0$ and unstable if $f'(x_0) > 0$.*

Problem 6.16. *Consider the system*

$$\begin{aligned} \dot{x} &= x - y - x(x^2 + y^2) + \frac{xy}{\sqrt{x^2 + y^2}}, \\ \dot{y} &= x + y - y(x^2 + y^2) - \frac{x^2}{\sqrt{x^2 + y^2}}. \end{aligned} \tag{6.34}$$

Show that $(1, 0)$ is not stable even though it satisfies (6.26). (Hint: Show that in polar coordinates the system is given by $\dot{r} = r(1 - r^2)$, $\dot{\theta} = 2\sin(\theta/2)^2$.)

Problem 6.17. *Draw phase plots as a function of μ for the three systems from above and prove all statements made above.*

6.6. 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

$$L : U(x_0) \to \mathbb{R} \tag{6.35}$$

which is zero at x_0, positive for $x \neq x_0$, and satisfies

$$L(\phi(t_0)) \ge L(\phi(t_1)), \quad t_0 < t_1, \qquad \phi(t_j) \in U(x_0)\backslash\{x_0\}, \tag{6.36}$$

for any solution $\phi(t)$. It is called a **strict Liapunov function** if equality in (6.36) 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 sublevel sets of L to be positively invariant. Let S_δ be the connected component of $\{x \in U(x_0) | L(x) \le \delta\}$ containing x_0. Note that in general S_δ might not be closed since it can have a common boundary with $U(x_0)$. In such a case orbits can escape through this part of the boundary and in order to avoid this, we need to assume that S_δ is closed.

Lemma 6.11. *If S_δ is closed, then it is positively invariant.*

Proof. Suppose $\phi(t)$ leaves S_δ at t_0 and let $x = \phi(t_0)$. Since S_δ is closed, $x \in S_\delta \subset U(x_0)$ and 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 $L(\phi(t_0+\varepsilon)) > \delta = L(x)$ for some ε since otherwise S_δ could not be the full connected component (we could extend it by adding $\phi([t_0, t_0+\varepsilon])$). This contradicts (6.36). □

Moreover, S_δ is a neighborhood of x_0 which shrinks to a point as $\delta \to 0$.

Lemma 6.12. *For every $\delta > 0$ there is an $\varepsilon > 0$ such that*

$$S_\varepsilon \subseteq B_\delta(x_0) \quad \textit{and} \quad B_\varepsilon(x_0) \subseteq S_\delta. \tag{6.37}$$

Proof. Assume that the first claim in (6.37) is false. Then for every $n \in \mathbb{N}$, there is an $x_n \in S_{1/n}$ such that $|x_n - x_0| \ge \delta$. 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 (6.37) were false, we could find a sequence x_n such that $|x_n - x_0| \le 1/n$ and $L(x_n) \ge \delta$. But then $\delta \le \lim_{n\to\infty} L(x_n) = L(x_0) = 0$, again a contradiction. □

Hence, given any neighborhood $V(x_0)$, we can find an ε such that $S_\varepsilon \subseteq V(x_0)$ is positively invariant. In other words, x_0 is stable.

Theorem 6.13 (Liapunov)**.** *Suppose x_0 is a fixed point of f. If there is a Liapunov function L, then x_0 is stable.*

But we can say even more. For every x with $\phi(t,x) \in U(x_0)$, $t \ge 0$, the limit

$$\lim_{t\to\infty} L(\phi(t,x)) = L_0(x) \tag{6.38}$$

exists by monotonicity. Moreover, for every $y \in \omega_+(x)$ we have some sequence $t_n \to \infty$ such that $\phi(t_n,x) \to y$ and thus $L(y) = \lim_{n\to\infty} L(\phi(t_n,x)) = L_0(x)$. 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

Theorem 6.14 (Krasovskii–LaSalle principle)**.** *Suppose x_0 is a fixed point of f. If there is a Liapunov function L which 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. Moreover, every orbit lying entirely in $U(x_0)$ converges to x_0.*

The same proof also shows

Theorem 6.15. *Let $L : U \to \mathbb{R}$ be continuous and bounded from below. If for some x we have $\gamma_+(x) \subset U$ and*

$$L(\phi(t_0, x)) \geq L(\phi(t_1, x)), \quad t_0 < t_1, \tag{6.39}$$

then L is constant on $\omega_+(x) \cap U$.

Most Liapunov functions will in fact be differentiable. In this case (6.36) 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

$$\mathrm{grad}(L)(x) \cdot f(x) \tag{6.41}$$

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**.

Theorem 6.15 implies that all $\omega_\pm$-limit sets are contained in the set where the Lie derivative of L vanishes.

Example. Consider the system

$$\dot{x} = y, \qquad \dot{y} = -x$$

together with the function $L(x,y) = x^2 + y^2$. The Lie derivative is given by

$$\mathrm{grad}(L)(x) \cdot f(x) = \begin{pmatrix} 2x \\ 2y \end{pmatrix} \begin{pmatrix} y \\ -x \end{pmatrix} = 2xy - 2yx = 0$$

and hence L is a Liapunov function; in fact, even a constant of motion. In particular, the origin is stable. Every level set $L(x,y) = \delta$ corresponds to an orbit and the system is not asymptotically stable. ◇

Problem 6.18. *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 6.19 (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 the Lie derivative of V.)*

Problem 6.20. *Show Theorem 6.15.*

Problem 6.21. *Suppose $L \in C^1(M, \mathbb{R})$. Show that the level set $L(x) = c$ is invariant under the flow if and only if the Lie derivative of L along the vector field vanishes on this level set.*

6.7. Newton's equation in one dimension

Finally, let us look at a specific example which will illustrate the results from this chapter.

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 2.2 then says that through every phase point there passes precisely one phase curve.

The **kinetic energy** is the quadratic form

$$T(\dot{x}) = \frac{\dot{x}^2}{2} \tag{6.43}$$

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

$$E = T(\dot{x}) + U(x). \tag{6.45}$$

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, solving (6.45) for $\dot{x}$, the solution corresponding to the initial conditions $x(0) = x_0$, $\dot{x}(0) = x_1$ can be given implicitly as

$$\mathrm{sign}(x_1) \int_{x_0}^{x} \frac{d\xi}{\sqrt{2(E - U(\xi))}} = t, \qquad E = \frac{\dot{x_1}^2}{2} + U(x_0). \tag{6.47}$$

If $x_1 = 0$ then $\operatorname{sign}(x_1)$ has to be replaced by $-\operatorname{sign}(U'(x_0))$. Fixed points of the equation of motion (6.42) are the solutions of $\dot{x} = 0$, $U'(x) = f(x) = 0$ and hence correspond to extremal points of the potential. Moreover, if $U(x)$ has a local minimum at x_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 x_0. In summary,

Theorem 6.16. *Newton's equation has 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

$$\ddot{x} = -\sin(x). \tag{6.48}$$

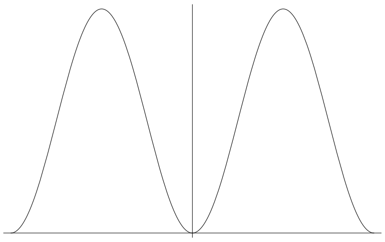
Here x describes the displacement angle from the position at rest ($x = 0$). In particular, x should be understood modulo 2π. The potential is given by $U(x) = 1 - \cos(x)$. To get a better understanding of this system we will look at some solutions corresponding to various initial conditions. This is 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[2]:= PhasePlot[f_, ic_, tmax_, opts___] :=
          Block[{i, n = Length[ic], ff, ivp, sol, phaseplot},
            ff = f /. {x → x[t], y → y[t]};
            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}, ]
              , {i, 1, n}];
            Show[Table[phaseplot[i], {i, 1, n}], opts]
          ];
```

Next, let us define the potential

```
In[3]:= U[x_] = 1 − Cos[x];
        Plot[U[x], {x, −2π, 2π}, Ticks → False]
```

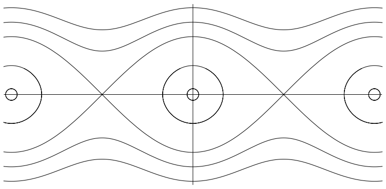
Out[3]=



and plot the phase portrait

```
In[4]:= 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]
```

Out[4]=



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 π and two non-closed orbits connecting $-\pi$ and π. It is usually referred to as **separatrix**. For $E > 2$ the level sets are again closed orbits (since we regard everything modulo 2π).

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. More generally, if

$$U'(x_0) = 0, \qquad U''(x_0) = \frac{\omega^2}{2} > 0, \tag{6.50}$$

our system should be approximated by

$$\ddot{y} = -\omega^2 y, \qquad y(t) = x(t) - x_0. \tag{6.51}$$

Clearly this equation can be transformed to (6.49) by scaling time according to $t \to \frac{t}{\omega}$.

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 7.10)

$$\dot{q} = \frac{\partial H(p,q)}{\partial p}, \qquad \dot{p} = -\frac{\partial H(p,q)}{\partial q}. \tag{6.53}$$

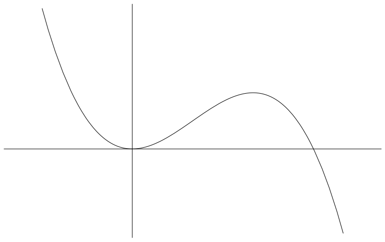
This formalism is called **Hamilton mechanics** and it is also useful for systems with more than one degree of freedom. We will return to this point of view in Section 8.3.

Problem 6.22. *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 6.23. *Investigate the potential $U(x) = \frac{x^2}{2} - \frac{x^3}{3}$.*

```
In[5]:= U[x_] = x^2/2 - x^3/3; Plot[U[x], {x, −1, 2}, Ticks → False]
```

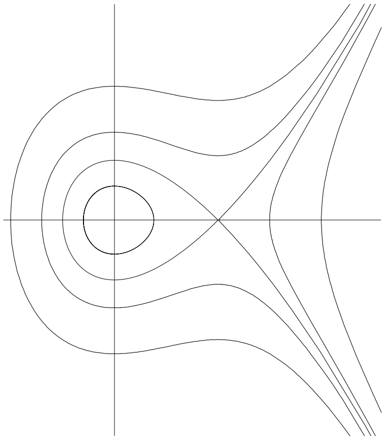
Out[5]=



Here are some interesting phase curves to get you started.

```
In[6]:= PhasePlot[{y, −U'[x]}, {{−1, 0}, {−0.7, 0}, {−0.5, 0}, {−0.3, 0},
          {1.05, 0.05}, {1.5, 0}, {2, 0}}, 8,
          PlotRange → {{−1, 2.5}, {−2, 2}}, Ticks → False]
```

Out[6]=



Problem 6.24 (Korteweg–de Vries equation)**.** *The* **Korteweg–de Vries equation**

$$\frac{\partial}{\partial t}u(t,x) + \frac{\partial^3}{\partial x^3}u(t,x) + 6u(t,x)\frac{\partial}{\partial x}u(t,x)$$

is a model for shallow water waves. One of its outstanding features is the existence of so-called **solitons***, that is, waves which travel in time without changing their shape.*

To find the one soliton solution make the **traveling wave ansatz** $u(x,t) = v(x - ct)$*,* $c \in \mathbb{R}$*, which yields*

$$-cv' + v''' + 6vv' = 0.$$

This equation can be integrated once

$$v'' - cv + 3v^2 + a = 0$$

such that one obtains an equation of Newton type with a cubic potential $U(v) = v^3 - \frac{c}{2}v^2 - av$*. Physicists are interested in solutions which satisfy* $\lim_{x\to\pm\infty} v(x) = 0$*. How does this limit the admissible parameters* a, c*? Find the corresponding shape* $v(x)$*.*

Note that if we eliminate the $-cv'$ *term via the transformation* $v(x) = -2w(x) + \frac{c}{6}$*, we obtain the differential equation*

$$w''' = 12w'w$$

for the Weierstraß elliptic function $\wp(x)$*. The function* $v(x) = -2\wp(x) + \frac{c}{6}$ *corresponds to a periodic solution of the Newton equation.*

Problem 6.25. *Show that all solutions are periodic if* $\lim_{|x|\to\infty} U(x) = +\infty$*.*

Problem 6.26. *The mathematical pendulum with friction is described by*

$$\ddot{x} = -\eta\dot{x} - \sin(x), \qquad \eta > 0.$$

Is the energy still conserved in this case? Show that the energy can be used as a Liapunov function and prove that the fixed point $(\dot{x}, x) = (0, 0)$ *is (asymptotically) stable. How does the phase portrait change?*

Problem 6.27. *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 7.11).*

(ii) *Show that minima of* $U(x)$ *are asymptotically stable.*

Chapter 7

Planar dynamical systems

7.1. 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 1.15). If predators are present, we assume that the growth rate is reduced proportional to the number of predators, that is,

$$\dot{x} = (A - By)x, \qquad A, B > 0. \tag{7.1}$$

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

$$\dot{y} = (Cx - D)y, \qquad C, D > 0. \tag{7.2}$$

Scaling x, y, and t we arrive at the system

$$\begin{array}{l} \dot{x} = (1 - y)x \\ \dot{y} = \alpha(x - 1)y \end{array}, \qquad \alpha > 0, \tag{7.3}$$

which are the predator-prey equations of **Volterra and Lotka**.

There are two fixed points. First of all, there is $(0, 0)$ and the lines $x = 0$ and $y = 0$ are invariant under the flow:

$$\Phi(t, (0, y)) = (0, y\mathrm{e}^{-\alpha t}), \qquad \Phi(t, (x, 0)) = (x\mathrm{e}^{t}, 0). \tag{7.4}$$

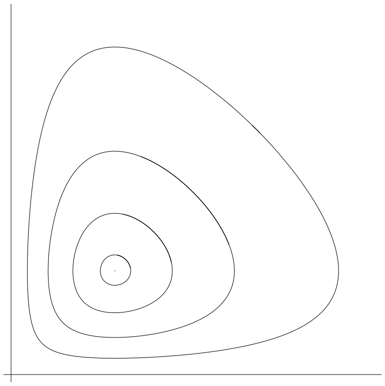


Figure 7.1. Phase portrait of the Volterra–Lotka system.

In particular, since no other solution can cross these lines, the first quadrant $Q = \{(x, y)|x > 0, y > 0\}$ is invariant. This is the region we are interested in. The second fixed point is $(1, 1)$.

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

$$\frac{dy}{dx} = \frac{dy}{dt}\left(\frac{dx}{dt}\right)^{-1} = \alpha\frac{(x-1)y}{(1-y)x}. \tag{7.5}$$

This equation is separable and solving it shows that the orbits are given implicitly by

$$L(x, y) = f(y) + \alpha f(x) = const, \qquad f(x) = x - 1 - \log(x). \tag{7.6}$$

The function f cannot be inverted in terms of elementary functions. However, $f(x)$ is convex with its global minimum at $x = 1$ and tends to ∞ as $x \to 0$ and $x \to \infty$. Hence the level sets are compact and each orbit is periodic surrounding the fixed point $(1, 1)$.

Theorem 7.1. *All orbits of the Volterra–Lotka equations* (7.3) *in* Q *are closed and encircle the only fixed point* $(1, 1)$.

The phase portrait is depicted in Figure 7.1.

Next, let us refine this model by assuming limited grow for both species (compare again Problem 1.15). The corresponding system is given by

$$\begin{array}{l}\dot{x} = (1 - y - \lambda x)x\\ \dot{y} = \alpha(x - 1 - \mu y)y\end{array}, \qquad \alpha, \lambda, \mu > 0. \tag{7.7}$$

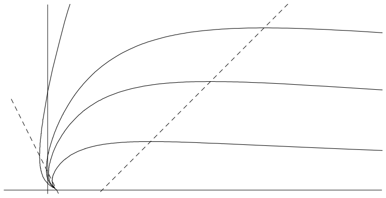


Figure 7.2. Phase portrait of a predator prey model with limiting growth.

As before the lines $x = 0$ and $y = 0$ are invariant but now there are four fixed points

$$(0,0), \quad (\lambda^{-1},0), \quad (0,-\mu^{-1}), \quad \left(\frac{1+\mu}{1+\mu\lambda}, \frac{1-\lambda}{1+\mu\lambda}\right). \tag{7.8}$$

The third one is outside of $\overline{Q}$ and so will be the last one if $\lambda > 1$.

We first look at the case where $\lambda \geq 1$ such that 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 7.1).

Lemma 7.2. *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 $(x_0, y_0) \in \overline{U}$.*

Now let us see how this applies to our case. The regions where $\dot{x}$ and $\dot{y}$ have definite signs are separated by the two lines

$$L_1 = \{(x,y) | y = 1 - \lambda x\}, \qquad L_2 = \{(x,y) | \mu y = x - 1\}. \tag{7.9}$$

A typical situation for $\alpha = \mu = 1$, $\lambda = 2$ is depicted in Figure 7.2.

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) Q_1, Q_2, and Q_3. Observe that the lines L_2 and L_1 are transversal and thus can only be crossed in the direction from $Q_3 \to Q_2$ and $Q_2 \to Q_1$, respectively.

Suppose we start at a point $(x_0, y_0) \in Q_3$. Then, adding to Q_3 the constraint $x \leq x_0$, we can apply Lemma 7.2 to conclude that the trajectory enters Q_2 through L_2 or converges to a fixed point in $\overline{Q_3}$. The last case is only possible if $(\lambda^{-1}, 0) \in \overline{Q_3}$, that is, if $\lambda = 1$. Similarly, starting in Q_2 the

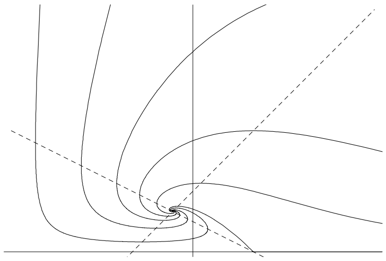


Figure 7.3. Phase portrait of a predator prey model with limiting growth.

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 $\lambda \geq 1$ every trajectory in Q converges to $(\lambda^{-1}, 0)$.

Now consider the remaining case $0 < \lambda < 1$ such that there is a third fixed point $(\frac{1+\mu}{1+\mu\lambda}, \frac{1-\lambda}{1+\mu\lambda})$. A phase portrait for $\alpha = \mu = 1$, $\lambda = \frac{1}{2}$ is shown in Figure 7.3.

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 one of the fixed points. 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 (7.6) we will try to scale x and y such that the minimum is at the fixed point $(x_0, y_0) = (\frac{1+\mu}{1+\mu\lambda}, \frac{1-\lambda}{1+\mu\lambda})$. We introduce

$$L(x, y) = \gamma_1 f(\frac{y}{y_0}) + \alpha\, \gamma_2 f(\frac{x}{x_0}), \qquad f(x) = x - 1 - \log(x), \tag{7.10}$$

where the constants $\gamma_1, \gamma_2 > 0$ are to be determined. Using

$$\dot{x} = (-\bar{y} - \lambda\bar{x})x, \quad \dot{y} = \alpha(\bar{x} - \mu\bar{y})y, \qquad \bar{x} = x - x_0, \bar{y} = y - y_0 \tag{7.11}$$

we compute

$$\dot{L} = \frac{\partial L}{\partial x}\dot{x} + \frac{\partial L}{\partial y}\dot{y} = -\alpha\left(\frac{\lambda\gamma_2}{x_0}\bar{x}^2 + \frac{\mu\gamma_1}{y_0}\bar{y}^2 + (\frac{\gamma_2}{x_0} - \frac{\gamma_1}{y_0})\bar{x}\bar{y}\right). \tag{7.12}$$

The right-hand side will be negative if we choose $\gamma_1 = y_0$ and $\gamma_2 = x_0$ such that the third term vanishes. Hence we again see that all orbits starting in Q converge to the fixed point (x_0, y_0).

Theorem 7.3. *Suppose $\lambda \geq 1$. Then there is no fixed point of the equations (7.7) in Q and all trajectories in Q converge to the point $(\lambda^{-1}, 0)$.*

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 in Q 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 λ^{-1} of the prey species.

Similarly on could consider systems of competing or cooperating species

$$\dot{x} = \alpha(x,y)x, \qquad \dot{y} = \beta(x,y)y. \tag{7.13}$$

Here we will call two species **cooperative** if the growth of one species increases the growth rate of the other and vice versa, that is,

$$\frac{\partial}{\partial y}\alpha(x,y) \geq 0 \quad \text{and} \quad \frac{\partial}{\partial x}\beta(x,y) \geq 0, \qquad (x,y) \in Q. \tag{7.14}$$

Similarly we will call two species **competitive** if the growth of one species decreases the growth rate of the other and vice versa, that is,

$$\frac{\partial}{\partial y}\alpha(x,y) \leq 0 \quad \text{and} \quad \frac{\partial}{\partial x}\beta(x,y) \leq 0, \qquad (x,y) \in Q. \tag{7.15}$$

It turns out that in this situation the analysis is much simpler. Moreover, we can even be slightly more general.

Theorem 7.4. *Suppose the system*

$$\dot{x} = f(x,y), \qquad \dot{y} = g(x,y), \qquad (x,y) \in M \subseteq \mathbb{R}^2 \tag{7.16}$$

is either strictly **cooperative**,

$$\frac{\partial}{\partial y}f(x,y) > 0 \quad \textit{and} \quad \frac{\partial}{\partial x}g(x,y) > 0, \qquad (x,y) \in M, \tag{7.17}$$

or strictly **competitive**,

$$\frac{\partial}{\partial y}f(x,y) < 0 \quad \textit{and} \quad \frac{\partial}{\partial x}g(x,y) < 0, \qquad (x,y) \in M. \tag{7.18}$$

Then all orbits converge either to a fixed point or to a boundary point (including ∞) of M.

Proof. We assume that our system is cooperative and denote the quadrants of $\mathbb{R}^2$ by $Q_1 = \{(x,y)|x,y > 0\}$, $Q_2 = \{(x,y)| -x,y > 0\}$, $Q_3 = \{(x,y)|x,-y > 0\}$, and $Q_4 = \{(x,y)| -x,-y > 0\}$. The competitive case can be handled analogously.

We first note that if $(\dot{x},\dot{y}) \in \overline{Q_1}$ at some time $t = t_0$, then $(\dot{x},\dot{y}) \in Q_1$ for all $t > t_0$. In fact, if we should have $\dot{x}(t_1) = 0$, then

$$\ddot{x} = \frac{\partial f}{\partial x}(x,y)\dot{x} + \frac{\partial f}{\partial y}(x,y)\dot{y}$$

is positive at such a point (if $\dot{y}(t_1) = 0$ vanishes as well, we already are at a fixed point). An analogous argument rules out $\dot{y}(t_1) = 0$. Similarly for Q_3. Finally, if $(\dot{x}, \dot{y}) \in Q_2 \cup Q_4$ it either remains there or enters $\overline{Q_1 \cup Q_3}$. Hence the sign of $\dot{x}(t)$ as well as $\dot{y}(t)$ can change at most once and thus both components are eventually monotone. □

Note that time reversal maps a cooperative system in a competitive one and vice versa.

In particular, if we assume limited growth, that is, $\alpha(x, y)$ becomes eventually negative as $x \to \infty$ and $\beta(x, y)$ becomes eventually negative as $y \to \infty$, then every solution converges to a fixed point.

Problem 7.1. *Prove Lemma 7.2.*

Problem 7.2 (Volterra principle)**.** *Show that for any orbit of the Volterra–Lotka system* (7.3)*, 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 $\frac{d}{dt}\log(x(t))$ over one period.)

Problem 7.3. *Show that the change of coordinates $x = \exp(q)$, $y = \exp(p)$ transforms the Volterra–Lotka system* (7.3) *into a Hamiltonian system with Hamiltonian $H(p, q) = L(\exp(q), \exp(p))$.*

Moreover, use the same change of coordinates to transform (7.7)*. Then use Bendixson's criterion (Problem 7.11) to show that there are no periodic orbits.*

Problem 7.4. *Show that* (7.7) *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

$$\frac{1+\mu}{1+\mu\lambda} < x_0 < \frac{1}{\lambda}, \qquad y_0 = 1 - \lambda x_0. \tag{7.19}$$

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 hits 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

$$(1+\mu)(1-\mu\lambda) \geq (1 - (\mu\lambda)^2)x_0, \tag{7.20}$$

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 7.5 (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 7.6 (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.

7.2. Examples from electrical engineering

In this section we want to come back to electrical circuits, which we already considered in Section 3.3. We will again look at the case of one inductor, one capacitor, and one resistor arranged in a loop. However, this time we want to consider a resistor with arbitrary characteristic

$$V_R = R(I_R). \tag{7.21}$$

Since there is no potential difference if there is no current, we must have $R(0) = 0$. For a classical resistor we have $R(I) = R\,I$, where the resistance R is a constant (Ohm's law), but for sophisticated elements like semiconductors the relation is more complicated. For example, the characteristic of a diode is given by

$$V = \frac{kT}{q} \log(1 + \frac{I}{I_L}), \tag{7.22}$$

where I_L is the leakage current, q the charge of an electron, k the Boltzmann constant and T the absolute temperature.

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 pass in only one direction.

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

$$\begin{array}{l} L\dot{I}_L = -V_C - R(I_L) \\ C\dot{V}_C = I_L \end{array}, \qquad R(0) = 0, \quad L, C > 0. \tag{7.23}$$

In addition, note that the change of energy in each element is given by $I\,V$. By Kirchhoff's laws we have

$$I_L V_L + I_C V_C + I_R V_R = 0, \tag{7.24}$$

which can be rewritten as

$$\frac{d}{dt}\left(\frac{L}{2}I_L^2 + \frac{C}{2}V_C^2\right) = -I_R R(I_R). \tag{7.25}$$

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 7.7)

$$\begin{array}{l} \dot{x} = y - f(x) \\ \dot{y} = -x \end{array}, \qquad f(0) = 0. \tag{7.26}$$

Equation (7.25) now reads

$$\frac{d}{dt}W(x,y) = -xf(x), \qquad W(x,y) = \frac{x^2+y^2}{2}. \tag{7.27}$$

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 7.5. *Suppose $xf(x) \ge 0$ for all $x \in \mathbb{R}$ and $xf(x) > 0$ for $0 < |x| < \varepsilon$. Then every trajectory of Liénard's equation* (7.26) *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 $\dot{W} = -xf(x) = 0$ for $0 \le |x| \le R$ and the result follows from the Krasovskii–LaSalle principle (Theorem 6.14). □

Conversely, note that $(0,0)$ is unstable if $xf(x) < 0$ for $0 < |x| < \varepsilon$. In fact, the above argument shows that within this region the distance to the fixed point will increase.

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. Suppose f is differentiable such that

(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 $x > \alpha$ (i.e., $\alpha = \beta$).

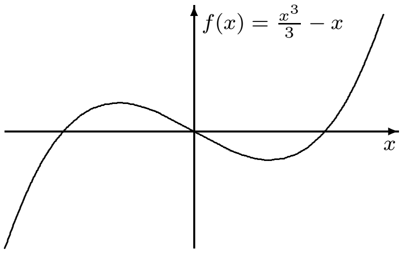


Figure 7.4. Typical f for Liénard's equation.

A prototypical f is depicted in Figure 7.4.

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 leaves the differential equation (7.26) invariant if f is odd.

As a first observation we note that

Lemma 7.6. *Every trajectory of Liénard's equation* (7.26) *in* Q_+ *can cross the graph of* f *at most once.*

Proof. A quick calculation shows that for $\Delta = y - f(x)$ we have $\dot{\Delta} = -x - f'(x)\Delta$ and thus $\dot{\Delta}(t_0) > 0$ whenever $\Delta(t_0) = 0$ and $(x,y) \in Q_+$. □

Next we show

Lemma 7.7. *Suppose* f *satisfies the requirements (ii) and (iii). Then, every trajectory starting in* $\overline{Q_+}$ *above the graph of* f *will eventually hit the graph at a finite positive time. Similarly, every trajectory starting in* Q_+ *on or below the graph of* f *will hit* L_- *at a finite positive time. Finally, every trajectory starting on the graph will hit* L_+ *at a finite negative time.*

Proof. Suppose we start at some point (x_0, y_0) with $x_0 \geq 0$. Choose some $C < \min f(x)$ and consider $\Delta = x^2 + (y-C)^2$. Then $\dot{\Delta} = 2(x\dot{x} + (y-C)\dot{y}) = 2x(C - f(x)) < 0$ for $(x,y) \in Q_+$. Hence starting in the region bounded by L_+, the graph of f and a circle $\Delta = R^2$ we stay inside this region until we hit the graph of f by Lemma 7.2 (we cannot converge to the only fixed point $(0,0)$ since it is unstable). This shows the first claim. The second follows similarly. For the last one use $\Delta = x^2 + (y-M)^2$, where $M > \max_{x \in [0, x_0+\varepsilon]} f(x)$ and consider the region bounded by the graph of f, the vertical line $y = y_0 + \varepsilon$, and a circle $\Delta = R^2$ containing (x_0, y_0). □

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 $(x(T), y(T)) =$

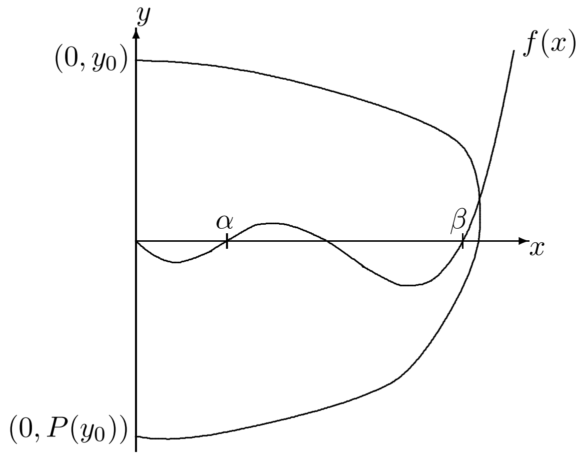


Figure 7.5. Definition of the Poincaré map $P(y_0)$ for Liénard's equation.

$(0, P(y_0))$ (cf. Figure 7.5). 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

$$\Delta(y_0) = W(0, P(y_0)) - W(0, y_0) = -\int_0^T x(t) f(x(t)) dt. \qquad (7.28)$$

Now what can we say about this function? Clearly, for $y_0 < \alpha$ we have $\Delta(y_0) > 0$. Hence it suffices to show that $\Delta(y_0)$ becomes negative as $y_0 \to \infty$.

By the last part of Lemma 7.7 there is a number $r > 0$ such that the trajectory starting at $(0, r)$ intersects the graph of f at $(\beta, 0)$. 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 $0 \le t \le t_1$ and $y(t) < f(x(t))$ for $t_2 \le t \le T$. Now let us split Δ into three parts by splitting the integral at t_1 and t_2.

For the first part we obtain

$$\Delta_1(y_0) = -\int_0^{t_1} x(t) f(x(t)) dt = \int_0^\beta \frac{-x f(x)}{y(x) - f(x)} dx, \qquad (7.29)$$

where only $y(x)$ depends on y_0 in the last expression. 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 $y(t_1) \to \infty$ as $y_0 \to \infty$ we have $\lim_{y_0 \to \infty} \Delta_1(y_0) = 0$.

The second part is

$$\Delta_2(y_0) = -\int_{t_1}^{t_2} x(t) f(x(t)) dt = -\int_{y(t_2)}^{y(t_1)} f(x(y)) dy < 0. \qquad (7.30)$$

By (iii) this part cannot tend to 0.

Finally,

$$\Delta_3(y_0) = -\int_{t_2}^{T} x(t)f(x(t))dt = \int_0^{\beta} \frac{-xf(x)}{f(x)-y(x)}dx \qquad (7.31)$$

also decreases, with a similar argument as for Δ_1.

Moreover, I claim that $\Delta(y_0)$ eventually becomes negative. If $y(t_2) \to -\infty$ then $\Delta_3(y_0) \to 0$ as in the case of Δ_1 and the claim holds. Otherwise, if $y(t_2) \to y_2 < 0$, then every orbit passing through (β, y) with $y \le y_2$ must stay below f for all negative times by Lemma 7.7. Consequently we must have $f(x) \to \infty$ (since it must stay above any such 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 7.8. *Suppose f satisfies the requirements (i)–(iii). Then Liénard's equation* (7.26) *has at least one periodic orbit encircling* $(0,0)$.

If in addition (iv) holds, this periodic orbit is unique and every trajectory (except $(0,0)$*) converges to this orbit as $t \to \infty$.*

Proof. It remains to show that all orbits except $(0,0)$ converge to the unique periodic orbit determined by $\overline{y} = -P(\overline{y})$. Since any initial condition reaches L_+ by Lemma 7.7, we can restrict our attention to orbits starting on L_+. By symmetry a solution starting at $(0,-y) \in L_-$ will hit L_+ at $-P(y)$ and we thus set $P(y) = -P(-y)$ for $y < 0$. Fix $y_0 > 0$ and consider the sequence of points $y_n = P^n(y_0)$ (i.e., $(0, y_{2m+1})$ is the sequence of intersections with L_- and $(0, y_{2m})$ is the sequence of intersections with L_+). Since $\Delta(y)$ is positive for $y < \overline{y}$ and negative for $\overline{y} < y$ the sequence $(-1)^n y_n$ is strictly decreasing for $y_0 > \overline{y}$ and strictly increasing for $y_0 < \overline{y}$ and hence converges to the only fixed point $\overline{y}$. By continuity of the flow the points on the orbit between y_n and y_{n+1} must also converge to $\gamma(\overline{y})$. □

The classical application is **van der Pol's equation**

$$\ddot{x} - \mu(1-x^2)\dot{x} + x = 0, \qquad \mu > 0, \qquad (7.32)$$

which models a triode circuit. By Problem 7.7 it is equivalent to Liénard's equation with $f(x) = \mu(\frac{x^3}{3} - x)$. All requirements of Theorem 7.8 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 in Figure 7.6.

It is also interesting to consider the family of Liénard's equations with $f_\mu(x) = x^3 - \mu x$. For $\mu \le 0$ it has a stable fixed point at $(0,0)$ which is

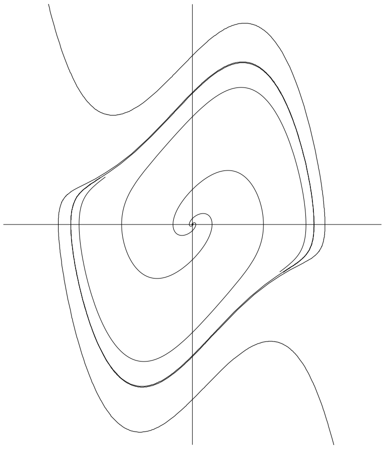


Figure 7.6. Phase portrait of the van der Pol equation.

globally attracting by Theorem 7.5. For $\mu > 0$ this fixed point becomes unstable and a unique globally attracting periodic orbit emerges. This is the prototypical example of a **Poincaré–Andronov–Hopf bifurcation**.

Problem 7.7. *The equation*

$$\ddot{x} + g(x)\dot{x} + x = 0$$

is also often called Liénard's equation. Show that it is equivalent to (7.26) *if we set* $y = \dot{x} + f(x)$*, where* $f(x) = \int_0^x g(t)dt$.

Problem 7.8. *Show that*

$$\dot{z} = z(\mu - (\alpha + \mathrm{i}\beta)|z|^2), \qquad \mu \in \mathbb{R}, \alpha, \beta > 0,$$

where $z(t) = x(t) + \mathrm{i}y(t)$*, exhibits a Hopf bifurcation at* $\mu = 0$*. (Hint: Use polar coordinates* $z = r\mathrm{e}^{\mathrm{i}\varphi}$*.)*

7.3. The Poincaré–Bendixson theorem

In all our examples from the previous sections the solutions behaved quite regular and would either converge to a fixed point or to a periodic orbit. It turns out that this behavior is typical and it is the purpose of the present section to classify the possible omega limit sets for planar systems. What makes $\mathbb{R}^2$ different from $\mathbb{R}^n$, $n \geq 3$, in this respect is the validity 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. We will only use the special case where the curve is piecewise smooth. A proof for this case can be found (e.g.) in [**39**].

So let $M \subseteq \mathbb{R}^2$ and $f \in C^1(M, \mathbb{R}^2)$ be given. By an **arc** $\Sigma \subset \mathbb{R}^2$ we mean a submanifold of dimension one given by a smooth map $t \to s(t)$ with

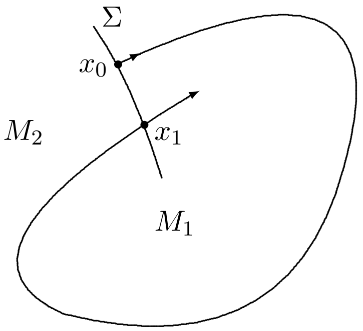


Figure 7.7. Proof of Lemma 7.9

$\dot{s} \neq 0$. Using this map the points of Σ can be ordered. Moreover, for each regular $x \in M$ (i.e., $f(x) \neq 0$), we can find an arc Σ containing x which is transversal to f (i.e., $\dot{s}_1(t)f_2(s(t)) - \dot{s}_2(t)f_1(s(t)) \neq 0$).

Given a regular point $x_0 \in \Sigma$ we can define the points of subsequent intersections of $\gamma_\sigma(x_0)$ with Σ by $x_n = \Phi(t_n, x_0)$. Of course this set may be finite or even empty in general. However, if it is infinite we must have $t_n \to T_\sigma(x_0)$. In fact, if $t_n \to T \neq T_\sigma(x_0)$ then the limit $y = \lim_{t\to T} \Phi(t, x_0) \in M$ exists and must be a regular point. Hence we can straighten out the flow near y, which shows that the difference between two consecutive intersection times cannot converge to 0 and hence contradicts our assumption.

Lemma 7.9. *Let $x_0 \in M$ be a regular point and Σ a transversal arc containing x_0. Denote by $x_n = \Phi(t_n, x_0)$, $n \geq 1$, the (maybe finite) ordered (according to t_n) sequence of intersections of $\gamma_\sigma(x_0)$ with Σ. Then x_n is monotone (with respect to the order of Σ).*

Proof. We only consider the $\sigma = +$ case. 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 Σ. 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 $\tilde{\Sigma} \subset \Sigma$ be the arc from x_0 to x_1 along Σ. Then f always points either in the direction of M_1 or M_2 along $\tilde{\Sigma}$ since it cannot change direction by transversality of Σ. So $\gamma_+(x_1)$ enters either M_1 or M_2 and then is trapped since it can neither exit through $\tilde{\Sigma}$ (as the vector field points in the wrong direction) nor cross the orbit from x_0 to x_1 (compare Figure 7.7). 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. □

Next, observe that if $y \in \Sigma \cap \omega_\sigma(x)$, we can approximate y by a sequence $x_n \in \Sigma \cap \gamma_\sigma(x)$. In fact, choose $t_n \to \sigma\infty$ such that $x_n = \Phi(t_n, x) \to y$. Then, by Lemma 6.9 (with $x = y$ and $T = 0$), we can use $\tilde{t}_n = t_n + \tau(x_n)$ to obtain a sequence $\Phi(\tilde{t}_n, x) \to y$ of the required type.

Corollary 7.10. *Let Σ be a transversal arc. Then $\omega_\sigma(x)$ intersects Σ in at most one point.*

Proof. Suppose there are two points of intersections y_1 and y_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 since both are subsequence of the monotone sequence x_n from Lemma 7.9. □

Corollary 7.11. *Suppose $\omega_\sigma(x) \cap \gamma_\sigma(x) \neq \emptyset$. Then x is periodic and hence $\omega_+(x) = \omega_-(x) = \gamma(x)$.*

Proof. By assumption there is some $y \in \omega_\sigma(x) \cap \gamma_\sigma(x)$. Moreover, by invariance of $\omega_\sigma(x)$ we must even have $\gamma(x) = \gamma(y) \subseteq \omega_\sigma(x)$. If y is fixed we have $\gamma_\sigma(x) = \{y\}$ and there is nothing to do. So we can assume that y is not fixed and pick a transversal arc Σ containing y plus a sequence $x_n \in \Sigma \cap \gamma_\sigma(x) \subseteq \Sigma \cap \omega_\sigma(x)$ converging to y. By the previous corollary we must have $x_n = y$ and hence $\gamma(y) = \gamma(x)$ is periodic. □

Corollary 7.12. *A minimal compact σ invariant set C is a periodic orbit.*

Proof. Pick $x \in C$. Then $\omega_\sigma(x) = C$ by minimality 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 7.13 (Poincaré–Bendixson theorem)**.** *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 Σ 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 7.10, we conclude $y_n = z$ and hence $\omega_\sigma(x)$ is a regular periodic orbit. □

Lemma 7.14. *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 Σ containing $\tilde{y}$. By Lemma 6.9 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 7.10) and hence $z \in \gamma(y)$ contradicting our assumption. □

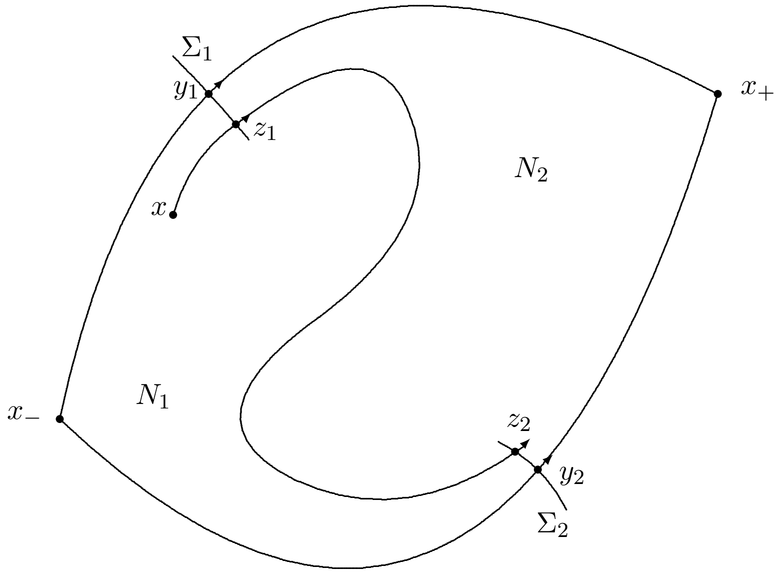


Figure 7.8. Proof of Lemma 7.15

Lemma 7.15. *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 7.11). Pick two transversal arcs $\Sigma_{1,2}$ containing $y_{1,2}$ respectively (compare Figure 7.8). Then $\gamma_\sigma(x)$ intersects $\Sigma_{1,2}$ in some points $z_{1,2}$ respectively. Without loss we can assume that there are no further intersections with Σ_1 and Σ_2 of $\gamma(x)$ between z_1 and z_2. 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 Σ_1, $\gamma_\sigma(x)$, Σ_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 7.9). 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 7.16 (generalized 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:*

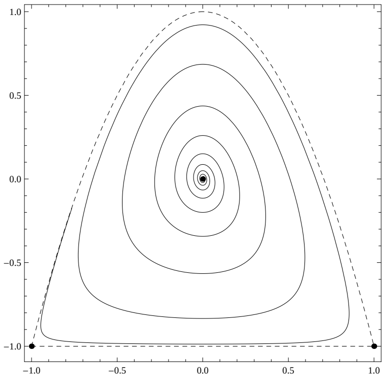


Figure 7.9. Phase portrait of an example where $\omega_+(x)$ consists of two fixed points and two orbits connecting them.

(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 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 7.13. 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 Σ containing z and a sequence $y_n \to z$, $y_n \in \gamma(y) \cap \Sigma$. By Corollary 7.10 $\gamma(y) \subseteq \omega_\sigma(x)$ can intersect Σ only in y. Hence $y_n = z$ and $\gamma(y)$ is regular periodic. Now Lemma 7.14 implies $\gamma(y) = \omega_\sigma(x)$ which is impossible since $\omega_\sigma(x)$ contains fixed points. □

Example. While we have already seen examples for case (i) and (ii) in the Poincaré–Bendixson theorem we have not seen an example for case (iii). Hence we consider the vector field

$$f(x,y) = \begin{pmatrix} y + x^2 - \alpha x(y - 1 + 2x^2) \\ -2(1+y)x \end{pmatrix}.$$

First of all it is easy to check that the curves $y = 1 - 2x^2$ and $y = -1$ are invariant. Moreover, there are four fixed points $(0,0)$, $(-1,-1)$, $(1,-1)$, and $(\frac{1}{2\alpha}, -1)$. We will chose $\alpha = \frac{1}{4}$ such that the last one is outside the region bounded by the two invariant curves. Then a typical orbit starting inside this region is depicted in Figure 7.9: It converges to the unstable fixed point $(0,0)$ as $t \to -\infty$ and spirals towards the boundary as $t \to +\infty$. In

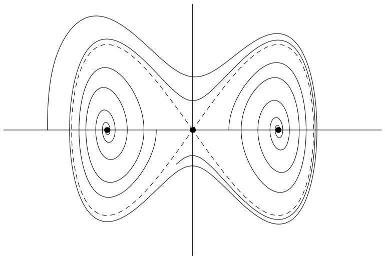


Figure 7.10. Phase portrait of an example where $\omega_+(x)$ consists of two leaves.

particular, its $\omega_+((x_0, y_0))$ limit set consists of three fixed points plus the orbits joining them.

To prove this consider $H(x,y) = x^2(1+y) + \frac{y^2}{2}$ and observe that its change along trajectories

$$\dot{H} = 2\alpha(1 - y - 2x^2)x^2(1+y)$$

is nonnegative inside our region (its boundary is given by $H(x,y) = \frac{1}{2}$). Hence it is straightforward to show that every orbit other than the fixed point $(0,0)$ converges to the boundary. ◇

Note that while Lemma 7.15 allows only one orbit in $\omega_\sigma(x)$ to connect different fixed points in $\omega_\sigma(x)$. There could be more than one (even infinitely many) connecting to the same fixed point as the following example shows.

Example. Consider the vector field

$$f(x,y) = \begin{pmatrix} y \\ -\eta E(x,y)^2 x - U'(x) \end{pmatrix},$$

where

$$E(x,y) = \big(\frac{y^2}{2} + U(x)\big), \quad U(x) = x^2(x^2-1).$$

In the case $\eta = 0$ this is a Newton equation with potential $U(x)$ (cf. Section 6.7). There are two stable fixed points $(\pm\frac{1}{\sqrt{2}}, 0)$ and an unstable one $(0,0)$ plus there are two separatrices

$$\begin{pmatrix} x(t) \\ y(t) \end{pmatrix} = \pm \frac{1}{\cosh(\sqrt{2}t)} \begin{pmatrix} 1 \\ -\sqrt{2}\tanh(\sqrt{2}t) \end{pmatrix},$$

satisfying $E(x(t), y(t)) = 0$. If we consider $\eta > 0$ then the energy $E(x,y)$ will decrease as t increases since $\frac{d}{dt}E = -\eta E^2 y^2$ for all orbits except the two separatrices. In particular, all orbits in the region $E > 0$ will have the set consisting of the two separatrices and the fixed point $(0,0)$ as $\omega_+((x_0, y_0))$. ◇

Let me also remark, that since the domain surrounded by a periodic orbit is invariant, Lemma 6.8 implies

Lemma 7.17. *The interior of every periodic orbit must contain a fixed point.*

Proof. By the Jordan curve theorem the interior is simply connected and thus conformally equivalent to the unit disc by the Riemann mapping theorem. As the boundary is a Jordan curve, this mapping extends to a homeomorphism to the closed unit disc by the Carathéodory theorem. Since orbits starting in the interior cannot escape to the exterior without crossing the boundary, that is our periodic orbit, the interior is also invariant. □

Periodic orbits attracting other orbits are also called **limit cycles** and Hilbert's 16th problem asks for a bound on the number of limit cycles for a planar system with polynomial coefficients.

Note that we can show that every isolated periodic orbit must attract nearby orbits either as $t \to +\infty$ or $t \to -\infty$.

Lemma 7.18. *Let $\gamma(y)$ be an isolated regular periodic orbit (such that there are no other periodic orbits within a neighborhood). Then every orbit $\gamma(x)$ starting sufficiently close to $\gamma(y)$ will have either $\omega_-(x) = \gamma(y)$ or $\omega_+(x) = \gamma(y)$.*

Proof. Choose a neighborhood of $\gamma(y)$ which contains no other periodic orbits and a transversal arc $\Sigma \subset N$ containing y. Now consider a point x_0 on Σ outside of $\gamma(y)$ (the case where it is inside is similar). If this point is sufficiently close to y it will stay inside N and return to Σ at a point $x_1 \neq y$. Moreover, we will assume that x_1 is closer to y (if it is farther away, just reverse time to reduce it to this case). Hence the picture will look as in Figure 7.7 with $\gamma(y)$ inside M_1. Now the semi-orbit $\gamma_+(x_1)$ remains in $M_1 \backslash M_3 \subset N$, where M_3 is the interior of $\gamma(y)$, and the same must be true for $\omega_+(x_0)$. Since this set contains only one periodic orbit $\gamma(y)$ we must have $\omega_+(x_0) = \gamma(y)$. □

Example. Consider the system

$$\dot{x} = -y + f(r)x \quad \dot{y} = x + f(r)y, \qquad r = \sqrt{x^2 + y^2},$$

which in polar coordinates $x = (r\cos(\theta), r\sin(\theta))$ reads just

$$\dot{r} = rf(r), \quad \dot{\theta} = 1.$$

Clearly every positive zero r_0 of $f(r)$ will correspond to a periodic orbit which will attract nearby orbits if $\pm f'(r_0) < 0$ for $t \to \pm\infty$. If we consider a double zero we can obtain an example where solutions on one side are attracted as $t \to +\infty$ and on the other side as $t \to -\infty$. Finally, note that the system will be polynomial if f is a polynomial in r^2. ⋄

Problem 7.9. *Find and prove a "Poincaré–Bendixson theorem" in* $\mathbb{R}^1$.

Problem 7.10. *Suppose* $\mathrm{div} f = 0$ *in some simply connected domain. 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 7.11 (Bendixson's criterion)**.** *Suppose* $\mathrm{div} f$ *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

$$\ddot{x} + p(x)\dot{x} + q(x) = 0, \qquad x \in \mathbb{R},$$

has no regular periodic solutions if $p(x) > 0$.

Problem 7.12 (Dulac's criterion)**.** *Show the following generalization of Bendixson's criterion. Suppose there is a scalar function* $\alpha(x)$ *such that* $\mathrm{div}(\alpha f)$ *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 7.13. *If the intersection* $\omega_+(x) \cap \omega_-(x) \neq \emptyset$ *contains a regular point, then* x *is periodic.*

Chapter 8

Higher dimensional dynamical systems

8.1. 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 σ 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 σ invariant set by considering

$$\gamma_\pm(X) = \bigcup_{\pm t \geq 0} \Phi(t, X) = \bigcup_{x \in X} \gamma_\pm(x). \tag{8.1}$$

Taking the closure $\overline{\gamma_\sigma(X)}$ we even obtain a closed σ invariant set by Lemma 6.4. 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 $\Phi(t_n, x_n) \to y$.

Note that we have

$$\bigcup_{x \in X} \omega_+(x) \subseteq \omega_+(X) \tag{8.2}$$

but equality will not hold in general as the following example shows.

Example. Consider

$$\dot{x} = x(1 - x^2), \qquad \dot{y} = -y. \tag{8.3}$$

The x-direction has two stable $x = \pm 1$ and one unstable $x = 0$ fixed points. Similarly, the y-direction has the only stable fixed point $y = 0$. Hence it is not hard to see that

$$\omega_+(B_r(0)) = [-1,1] \times \{0\}, \qquad r > 0. \tag{8.4}$$

On the other hand,

$$\bigcup_{x \in B_r(0)} \omega_+(x) = \{(-1,0),(0,0),(1,0)\}. \tag{8.5}$$

In particular $\omega_+(B_r(0))$ contains the three fixed points plus their connecting orbits. That is, all orbits which lie entirely in $B_r(0)$. This is also true in general as we will see in Theorem 8.3 below. ◇

The following two lemmas are the analogs of Lemma 6.5 and Lemma 6.6.

Lemma 8.1. *The set $\omega_\sigma(X)$ is a closed invariant set given by*

$$\omega_\sigma(X) = \bigcap_{\sigma t \ge 0} \Phi(t, \overline{\gamma_\sigma(X)}) = \bigcap_{\sigma t \ge 0} \overline{\bigcup_{\sigma(s-t)\ge 0} \Phi(s,X)}. \tag{8.6}$$

Proof. The intersection of closed σ-invariant sets is again a closed σ-invariant set by Lemma 6.4 and invariance follows literally as in Lemma 6.5. Hence it suffices to show (8.6).

We only prove the $\sigma = +$ case. First of all note that since $\Phi(t,.)$ is a diffeomorphism we have

$$\Phi(t, \overline{\gamma_+(X)}) = \overline{\Phi(t,\gamma_+(X))} = \overline{\bigcup_{s \ge t} \Phi(s,X)}.$$

To see $\bigcap_{t\ge 0} \Phi(t,\overline{\gamma_+(X)}) \subseteq \omega_+(X)$ choose some $y \in \bigcap_{t \ge 0} \overline{\Phi(t,\gamma_+(X))}$. Then, for every $n \in \mathbb{N}$ we can find some $y_n = \Phi(n+s_n, x_n) \in \Phi(n, \gamma_+(X))$ such that $|y - y_n| < \frac{1}{n}$. Setting $t_n = n + s_n$ we have found a sequence $t_n \to \infty$ and points $x_n \in X$ such that $\Phi(t_n, x_n) \to y$, that is, $y \in \omega_\sigma(X)$.

Conversely, to show $\omega_+(X) \subseteq \bigcap_{t \ge 0} \Phi(t, \overline{\gamma_+(X)})$ choose some $y \in \omega_+(X)$. Then there exists $t_n \to \infty$ and $x_n \in X$ such that $y_n = \Phi(t_n,x_n) \to y$. This implies $y_n \in \Phi(t,\gamma_+(X))$ for $t_n > t$ and thus $y \in \overline{\Phi(t,\gamma_+(X))}$ for every $t \ge 0$. □

We will only consider the case $\sigma = +$ from now on for notational simplicity. Since by the last equality in (8.6) the sets $\Phi(t,\overline{\gamma_\sigma(X)})$ are decreasing, we see

$$\omega_+(X) = \bigcap_{t \ge t_0} \Phi(t, \overline{\gamma_+(X)}) = \bigcap_{n \in \mathbb{N}} \Phi(n, \overline{\gamma_+(X)}). \tag{8.7}$$

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 8.2. *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. □

Theorem 8.3. *The set $\omega_\sigma(X)$ is the union over all complete orbits lying entirely in $\overline{\gamma_\sigma(X)}$.*

Proof. Let $\gamma(y)$ be such an orbit. Then $\gamma(y) \subseteq \overline{\gamma_+(X)}$ and invariance of $\gamma(y)$ implies $\gamma(y) \subseteq \Phi(t, \overline{\gamma_+(X)})$ for all t and hence $\gamma(y) \subseteq \omega_+(X)$. Conversely, let $y \in \omega_+(X)$. Then invariance of $\overline{\gamma_+(X)}$ implies $\gamma(y) \subseteq \omega_+(X) \subseteq \overline{\gamma_+(X)}$. □

For a given invariant set $\Lambda \subset M$ the sets

$$W^\pm(\Lambda) = \{x \in M | \lim_{t \to \pm\infty} d(\Phi_t(x), \Lambda) = 0\} \tag{8.8}$$

are the **stable** respectively **unstable sets** of Λ. Here $d(x, A) = \inf\{|x - y| \,| y \in A\}$ denotes the distance between x and $A \subseteq \mathbb{R}^n$ (cf. Problem 6.11).

Example. For the previous example we have $W_+([-1, 1] \times \{0\}) = \mathbb{R}^2$ and $W^+(\{(\pm 1, 0)\}) = \mathbb{R}_\pm \times \mathbb{R}$. ◇

An invariant set Λ is called **attracting** if $W^+(\Lambda)$ is a neighborhood of of Λ. In this case the set $W^+(\Lambda)$ is also called the **domain** or **basin of attraction** for Λ. Moreover, for any positively invariant neighborhood U we have

$$W^+(\Lambda) = \bigcup_{t<0} \Phi_t(U). \tag{8.9}$$

In particular, $W^+(\Lambda)$ is invariant and choosing U open we see that the basin of attraction is also open:

Lemma 8.4. *Let Λ be an invariant attracting set. Then its basin of attraction is invariant and open.*

Note that by Lemma 6.4 the boundary $\partial W_+(\Lambda) = \overline{W_+(\Lambda)} \backslash W_+(\Lambda)$ is invariant as well.

But how can we find such an attracting 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 $\Phi_t(\overline{E}) \subset E$ for all $t > 0$. Note that in this case every orbit starting in $\overline{E}$ is complete. In many cases a trapping region can be found by looking for the region bounded by some surface (e.g. the level set of some function) such that the vector field points inwards on that surface, cf. Problem 8.2.

Lemma 8.5. *Let E be a trapping region. Then*

$$\Lambda = \omega_+(E) = \bigcap_{t\ge 0} \Phi(t, E) \tag{8.10}$$

is a nonempty, invariant, compact, and connected attracting set.

Proof. First of all note that by $\Phi(t+\varepsilon, \overline{E}) \subset \Phi(t, E) \subset \Phi(t, \overline{E})$ we have

$$\bigcap_{t\ge 0} \Phi(t, E) = \bigcap_{t\ge 0} \Phi(t, \overline{E}) = \bigcap_{t\ge 0} \Phi(t, \overline{\gamma_+(E)}) = \omega_+(E).$$

and it remains to show that Λ is attracting.

To see this suppose there were an $x \in E$ and a sequence $t_n \to \infty$ with $d(\Phi(t_n, x), \Lambda) \ge \varepsilon > 0$. Then, since $\Phi(t_n, x)$ remains in the compact set $\overline{E}$, we can assume $\Phi(t_n, x) \to y$ after passing to a subsequence. But $y \in \omega_+(x) \subseteq \omega_+(E)$ by (8.2), a contradiction. □

Unfortunately the definition of an attracting set is not always good enough. In our example (8.3) 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 Λ 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 8.6. *Let E be a trapping region. Then*

$$W^-(x) \subseteq \omega_+(E), \qquad \forall x \in \omega_+(E). \tag{8.11}$$

Proof. Let $y \in W^-(x)$, that is $\lim_{t\to-\infty} \Phi(t, y) = x \in E$. Since E is open there is some t_0 such $\gamma_-(\Phi(t_0, y)) \subset E$. Since E is positive invariant we even obtain $\gamma(y) = \gamma(\Phi(t_0, y)) \subseteq E = \gamma_+(E)$ and the claim follows from Theorem 8.3. □

To exclude such situations, one has to ensure that an attracting set cannot be split into smaller invariant sets. One such possibility is to define

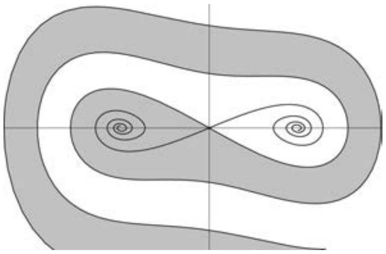


Figure 8.1. Basin of attraction for the fixed point $(-1,0)$ of Duffing's equation.

an **attractor** to be an attracting set which is topologically transitive. Here a closed invariant set Λ 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 Λ is topologically transitive if it contains a dense orbit (Problem 8.1).

This implies that only the sets $\{(-1,0)\}$ or $\{(1,0)\}$ are attractors for the above example. The domains of attraction are $W^+(\{(\pm 1, 0)\}) = \{(x,y) \in \mathbb{R}^2 | \pm x > 0\}$.

Example. As another example let us look at the **Duffing equation**

$$\ddot{x} = -\delta \dot{x} + x - x^3, \quad \delta > 0, \tag{8.12}$$

from Problem 9.5. 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 manifold of the hyperbolic saddle $(0,0)$. The situation for $\delta = 0.3$ is depicted in Figure 8.1. ◇

Example. For the van der Pol equation (7.32) 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 8.1. *Show that a closed invariant set which has a dense orbit is topologically transitive.*

Problem 8.2. *Suppose $L \in C^1(M, \mathbb{R})$. Let $V_R = \{x \in M | L(x) \leq R\}$ be a compact set and suppose the Lie derivative satisfies*

$$\mathrm{grad}(L)(x) \cdot f(x) < 0, \qquad \forall x : L(x) = R.$$

Then every connected component of V_R is a trapping region.

Problem 8.3. *Suppose E is a trapping region and let $\Lambda = \omega_+(E)$. Then*

$$W^+(\Lambda) = \{x \in M | \omega_+(x) \subseteq \Lambda, \, \omega_+(x) \neq \emptyset\}.$$

(Hint: Lemma 6.7.)

8.2. The Lorenz equation

One of the most famous dynamical systems which exhibits chaotic behavior is the **Lorenz equation**

$$\begin{aligned}\dot{x} &= -\sigma(x-y),\\ \dot{y} &= rx - y - xz,\\ \dot{z} &= xy - bz, \end{aligned} \tag{8.13}$$

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 (8.13). 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

$$(x, y, z) \to (-x, -y, z). \tag{8.14}$$

Moreover, the z axis is an invariant manifold since

$$x(t) = 0, \quad y(t) = 0, \quad z(t) = z_0 \mathrm{e}^{-bt} \tag{8.15}$$

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 $r \le 1$. In this case there is only one fixed point of the vector field, namely the origin. The Jacobian matrix at 0 is given by

$$\begin{pmatrix} -\sigma & \sigma & 0 \\ r & -1 & 0 \\ 0 & 0 & -b \end{pmatrix} \tag{8.16}$$

and the corresponding eigenvalues are

$$-b, \qquad -\frac{1}{2}(1 + \sigma \pm \sqrt{(1+\sigma)^2 + 4(r-1)\sigma}). \tag{8.17}$$

Hence the origin is asymptotically stable for $r < 1$ by Corollary 3.27. However, we can even do better. To this end, let us make the ansatz

$$L(x, y, z) = \alpha x^2 + \beta y^2 + \gamma z^2, \qquad \alpha, \beta, \gamma > 0, \tag{8.18}$$

for a Liapunov function. Then a straightforward computation shows

$$\dot{L} = -2\alpha\sigma x^2 + 2(\alpha\sigma + \beta r)xy - 2\beta y^2 - 2\gamma b z^2 + 2(\gamma - \beta)xyz. \tag{8.19}$$

To eliminate the xyz term we choose $\gamma = \beta$. Since no choice of $\alpha, \beta > 0$ will make the xy disappear, we need to absorb it using $2xy = -(x-y)^2 + x^2 + y^2$,

$$\dot{L} = -(\alpha\sigma - \beta r)x^2 - (\alpha\sigma + \beta r)(x-y)^2 - ((2-r)\beta - \alpha\sigma)y^2 - 2\beta b z^2. \tag{8.20}$$

Hence we need to choose $\alpha, \beta > 0$ such that $\alpha\sigma - \beta r \geq 0$ and $(2-r)\beta - \alpha\sigma \geq 0$. For example $\alpha = r$ and $\beta = \sigma$ such that the first term vanishes and the third becomes $2(1-r)\sigma \geq 0$ for $r \leq 1$. In summary, for

$$L(x, y, z) = r x^2 + \sigma y^2 + \sigma z^2 \tag{8.21}$$

we have

$$\dot{L}(x, y, z) = -2\sigma(r(x-y)^2 + (1-r)y^2 + b z^2) \tag{8.22}$$

and the following lemma follows easily from Theorem 6.14 (Problem 8.4).

Lemma 8.7. *Suppose $r \leq 1$. 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

$$(x, y, z) = (\pm\sqrt{b(r-1)}, \pm\sqrt{b(r-1)}, r-1), \tag{8.23}$$

and the linearization is given by

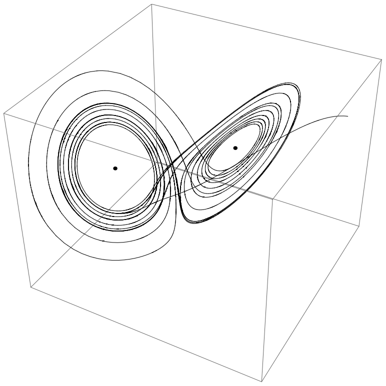
$$\begin{pmatrix} -\sigma & \sigma & 0 \\ 1 & -1 & \mp\sqrt{b(r-1)} \\ \pm\sqrt{b(r-1)} & \pm\sqrt{b(r-1)} & -b \end{pmatrix}. \tag{8.24}$$

One can again compute the eigenvalues but the result would almost fill one page. Note however that by (8.14) the eigenvalues are the same for both points. From (8.17) 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 $1 < r < 470/19 = 24.74$.

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];
```

Out[1]=



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_ε which all trajectories eventually enter and never leave again. To do this, let us consider a small modification of our Liapunov function from above,

$$L(x,y,z) = rx^2 + \sigma y^2 + \sigma(z-2r)^2. \tag{8.25}$$

A quick computation shows

$$\dot{L}(x,y,z) = -2\sigma(rx^2 + y^2 + b(z-r)^2 - br^2). \tag{8.26}$$

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 $E_1 = \{(x,y,z)|L(x,y,z) < M+1\}$. Any point outside E_1 also lies outside E and hence $\dot{L} \leq -\delta < 0$ for such points. That is, for $x \in \mathbb{R}^3\backslash E_1$ the value of L is strictly decreasing along its trajectory and hence it must enter E_1 after some finite time.

Moreover, E_1 is a trapping region for the Lorenz equation and there is a corresponding attracting set

$$\Lambda = \omega_+(E_1), \tag{8.27}$$

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 Λ. Moreover, I even claim that Λ is of Lebesgue measure zero. To see this we need a generalized version of Liouville's formula (3.91).

Lemma 8.8. *Let $\dot{x} = f(x)$ be a dynamical system on $\mathbb{R}^n$ with corresponding flow $\Phi(t,x)$. Let U be a bounded open subset of $\mathbb{R}^n$ and let $V = \int_U dx$ be its volume. Abbreviate $U(t) = \Phi(t,U)$, respectively, $V(t) = \int_{U(t)} dx$. Then*

$$\dot{V}(t) = \int_{U(t)} \operatorname{div}(f(x))\, dx. \tag{8.28}$$

Proof. By the change of variable formula we have

$$V(t) = \int_{U(t)} dx = \int_U \det(d\Phi_t(x))\, dx.$$

Since $\Pi_x(t) = d\Phi_t(x)$ satisfies the first variational equation,

$$\dot{\Pi}_x(t) = df(\Phi_t(x))\Pi_x(t),$$

Liouville's formula (3.91) for linear systems implies

$$\det(d\Phi_t(x)) = \exp\left(\int_0^t \operatorname{div}(f(\Phi_s(x)))ds\right)$$

(recall $\operatorname{tr}(df(x)) = \operatorname{div}(f(x))$). Thus

$$\dot{V}(t) = \int_U \operatorname{div}(f(\Phi_t(x))) \det(d\Phi_t(x))\, dx$$

and a second application of the change of variable formula finishes the proof. □

Applying this lemma to the Lorenz equation we obtain

$$V(t) = V\mathrm{e}^{-(1+\sigma+b)t} \tag{8.29}$$

since

$$\operatorname{div}(f) = -(1+\sigma+b). \tag{8.30}$$

In particular, we see that the measure of $\Phi(t, E_1)$ decreases exponentially, and the measure of Λ 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 Λ 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**

$$\Phi_1 : \Lambda \to \Lambda \tag{8.31}$$

is *chaotic* and if Λ is *fractal*. It is still unknown whether the Lorenz attractor is strange in the sense of this definition. See the book by Sparrow [**38**] 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 11.3 and Section 11.6, 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 8.4. *Prove Lemma 8.7.*

Problem 8.5. *Solve the Lorenz equation for the case $\sigma = 0$.*

Problem 8.6. *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 $\zeta = \sigma(\varepsilon^2 z - \varepsilon)$.*

Show that the resulting system for $\varepsilon = 0$ is given by

$$\xi' = \eta, \quad \eta' = -\xi\zeta, \quad \zeta' = \eta\xi,$$

which has two conserved quantities

$$\xi^2 - 2\zeta = 2\alpha, \quad \eta^2 + \zeta^2 = \beta.$$

Derive the single third order equation $\xi''' = -(\frac{3}{2}\xi^2 - \alpha)\xi'$. Integrate this equation once and observe that the result is of Newton type (see Section 6.7). Now what can you say about the solutions? (Hint: Problem 6.25.)

8.3. 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 $q \in V \subseteq \mathbb{R}^n$. Associated with such a system is a **Lagrange function**

$$L(v, q), \qquad v = \dot{q}, \tag{8.32}$$

and an integral curve $q(t)$ for which the **action integral**

$$\mathcal{I}(q) = \int_{t_0}^{t_1} L(\dot{q}(t), q(t))dt \tag{8.33}$$

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

$$q_\varepsilon(t) = q(t) + \varepsilon r(t), \tag{8.34}$$

we see that a necessary condition for q to be extremal is that

$$\frac{d}{d\varepsilon}\mathcal{I}(q_\varepsilon)\Big|_{\varepsilon=0} = 0. \tag{8.35}$$

Using integration by parts this immediately yields (Problem 8.7) the corresponding **Euler–Lagrange equation**

$$\frac{\partial L}{\partial q} - \frac{d}{dt}\frac{\partial L}{\partial v} = 0. \tag{8.36}$$

In the situation of particles under the influence of some forces we have

$$L(v,q) = \frac{1}{2} v M v - U(q), \tag{8.37}$$

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

$$M\ddot{q} = -\operatorname{grad} U(q). \tag{8.38}$$

If the **momentum**

$$p(v,q) = \frac{\partial L}{\partial v}(v,q) \tag{8.39}$$

is a diffeomorphism for fixed q, and hence

$$\det \frac{\partial^2 L}{\partial v^2} \neq 0, \tag{8.40}$$

then we can consider the **Legendre transform** of L,

$$H(p,q) = pv - L(v,q), \quad v = v(p,q), \tag{8.41}$$

which is known as the **Hamilton function** of the system. The associated variational principle is that the integral

$$\mathcal{I}(p,q) = \int_{t_0}^{t_1} \Big(p(t)\dot{q}(t) - H(p(t),q(t)) \Big) dt \tag{8.42}$$

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

$$\dot{q} = \frac{\partial H(p,q)}{\partial p}, \qquad \dot{p} = -\frac{\partial H(p,q)}{\partial q}. \tag{8.43}$$

This formalism is called **Hamilton mechanics**.

In the special case of some particles we have

$$p = Mv, \qquad H(p,q) = \frac{1}{2} p M^{-1} p + U(q) \tag{8.44}$$

and the Hamiltonian corresponds to the total energy of the system.

Introducing the **symplectic matrix**

$$J = \begin{pmatrix} 0 & \mathbb{I} \\ -\mathbb{I} & 0 \end{pmatrix}, \quad J^{-1} = J^T = -J, \tag{8.45}$$

Hamilton's equation can also be written as

$$\frac{d}{dt} \begin{pmatrix} p \\ q \end{pmatrix} = -\operatorname{grad}_s H(p,q), \tag{8.46}$$

where $\operatorname{grad}_s = -J \operatorname{grad}$ is called the **symplectic gradient**.

A straightforward calculation shows that H is a **constant of motion**, that is,

$$\frac{d}{dt}H(p(t),q(t)) = \frac{\partial H}{\partial p}\dot{p} + \frac{\partial H}{\partial q}\dot{q} = -\frac{\partial H}{\partial p}\frac{\partial H}{\partial q} + \frac{\partial H}{\partial q}\frac{\partial H}{\partial p} = 0. \tag{8.47}$$

More generally, for a differentiable function $I(p,q)$ its change along a trajectory is given by its Lie derivative (compare (6.41))

$$\frac{d}{dt}I(p(t),q(t)) = \{H(p(t),q(t)), I(p(t),q(t))\}, \tag{8.48}$$

where

$$\{H,I\} = \frac{\partial H}{\partial p}\frac{\partial I}{\partial q} - \frac{\partial H}{\partial q}\frac{\partial I}{\partial p} \tag{8.49}$$

is called **Poisson bracket**. (This should be compared with the Heisenberg equation of Problem 3.29.)

A function $I(p,q)$ is called a **first integral** if it is constant along trajectories, that is, if

$$\{H,I\} = 0. \tag{8.50}$$

But how can we find first integrals? One source are symmetries.

Theorem 8.9 (Noether)**.** *Let $\Phi(t,q)$ be the flow generated by $f(q)$. If Φ leaves the Lagrangian invariant, then*

$$I(v,q) = \frac{\partial L(v,q)}{\partial v} f(q) \tag{8.51}$$

is a constant of motion.

Proof. Abbreviate $q^s(t) = \Phi(s,q(t))$. The invariance of $L(v,q)$ implies

$$\begin{aligned} 0 &= \frac{d}{ds}L(\dot{q}^s(t), q^s(t))\Big|_{s=0} \\ &= \frac{\partial L}{\partial v}(\dot{q}(t),q(t))\frac{\partial f}{\partial q}(q(t))\dot{q}(t) + \frac{\partial L}{\partial q}(\dot{q}(t),q(t))f(q(t)) \end{aligned}$$

and hence

$$\begin{aligned} \frac{d}{dt}I(\dot{q}(t),q(t)) &= \left(\frac{d}{dt}\frac{\partial L}{\partial v}(\dot{q},q)\right) f(q) + \frac{\partial L}{\partial v}(\dot{q},q)\frac{\partial f}{\partial q}(q)\dot{q} \\ &= \left(\frac{d}{dt}\frac{\partial L}{\partial v}(\dot{q},q) - \frac{\partial L}{\partial q}(\dot{q},q)\right) f(q) = 0 \end{aligned}$$

by the Euler–Lagrange equation. □

For example, if $L(v,q)$ from (8.37) does not depend on the j'th coordinate q_j (for some fixed j), then it is clearly invariant under $\Phi(s,q) = q + s\delta_j$, where δ_j is the unit vector in the j'th direction. Hence the j'th momentum

$$p_j = \frac{\partial L(v,q)}{\partial v_j} \tag{8.52}$$

is conserved in this case by Noether's theorem. For another example see Problem 8.13.

Another important property of Hamiltonian systems is that they are volume preserving. This follows immediately form Lemma 8.8 since the divergence of a Hamiltonian vector field is zero.

Theorem 8.10 (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 8.11 (Poincaré)**.** *Suppose Φ 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 8.7. *Derive the Euler–Lagrange equation* (8.36) *for $q \in C^2$.*

Problem 8.8. *Consider the Lagrange functions $L_1(q,v) = |v|$ and $L_2(q,v) = \frac{1}{2}|v|^2$ in $\mathbb{R}^n$. What is the corresponding action integral for L_1? What are the extremal curves for L_1 and for L_2? Show that $\mathcal{I}_1(q) \leq \sqrt{2(t_1 - t_0)\mathcal{I}_2(q)}$ with equality if $|\dot{q}| = 1$ (Hint: Cauchy–Schwarz inequality).*

Let $M(q)$ be a positive definite matrix for every $q \in \mathbb{R}^n$. Consider $L_1(q,v) = \sqrt{vM(q)v}$ and $L_2(q,v) = \frac{1}{2}vM(q)v$. The action integral corresponding to L_1 is called the length of the curve q and extremals are called **geodesics**. *Show that the length is independent of reparametrization.*

Problem 8.9 (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 8.10. *Show that the Poisson bracket is a skew-symmetric bilinear form on $C^\infty(V)$ 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 8.11. *Suppose that D is bounded and positively invariant under a volume preserving flow. Then D belongs to the set of nonwandering points. (Hint: Poincaré's recurrence theorem and Problem 6.10.)*

Problem 8.12 (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.

Problem 8.13. *Consider $L(v,q)$ from* (8.37) *in $\mathbb{R}^3$ with $M=m\mathbb{I}_3$ and suppose $U(q)=U(|q|)$ is rotation invariant. Show that the* **angular momentum**

$$l=x\wedge p$$

is conserved in this case. Here $\wedge$ denotes the cross product in $\mathbb{R}^3$.

8.4. 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

$$(P,Q)=\varphi(p,q), \qquad (p,q)=\psi(P,Q), \tag{8.53}$$

we have

$$\begin{pmatrix}\dot{P}\\ \dot{Q}\end{pmatrix}=d\varphi\begin{pmatrix}\dot{p}\\ \dot{q}\end{pmatrix}=-d\varphi J\operatorname{grad}H(p,q)=-(d\varphi J d\varphi^T)\operatorname{grad}K(P,Q), \tag{8.54}$$

where $K=H\circ\varphi$ is the transformed Hamiltonian. Hence, we need to require that the Jacobian matrix of φ is a symplectic matrix, that is,

$$d\varphi\in\mathrm{Sp}(2n)=\{M\in\mathrm{Gl}(2n)|MJM^T=J\}, \tag{8.55}$$

where $\mathrm{Sp}(2n)$ is the **symplectic group**. Such a map is called a **symplectic map**. In this case φ is also called a **canonical transform**. Alternatively they can be characterized as those transformations which leave the **symplectic two form**

$$\omega((p_1,q_1),(p_2,q_2)) = (p_1,q_1)J(p_2,q_2) = p_1q_2 - p_2q_1 \tag{8.56}$$

invariant.

To find canonical transformations, recall that we have derived Hamilton's equations from the variational principle (8.42). Hence, our transform will be canonical if the integrands of (8.42) and

$$\tilde{I}(P,Q) = \int_{t_0}^{t_1} P(t)\dot{Q}(t) - K(P(t),Q(t))dt \tag{8.57}$$

only differ by a total differential. By $H(p,q) = K(P,Q)$ we are led to

$$pdq - PdQ = dS, \tag{8.58}$$

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

$$S = S_1(q,Q) \tag{8.59}$$

and

$$pdq - PdQ = \frac{\partial S_1}{\partial q}dq + \frac{\partial S_1}{\partial Q}dQ \tag{8.60}$$

shows we have

$$p = \frac{\partial S_1}{\partial q}, \qquad P = -\frac{\partial S_1}{\partial Q}, \tag{8.61}$$

since the previous equation must hold for all curves $q(t)$ and $Q(t)$. Moreover, if we require

$$\det \frac{\partial S_1}{\partial q \partial Q} \neq 0, \tag{8.62}$$

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

$$(P,Q) = (\frac{\partial S_1}{\partial Q}(q,Q(p,q)), Q(p,q)). \tag{8.63}$$

Similarly we could choose

$$S = -PQ + S_2(P,q), \tag{8.64}$$

where

$$pdq - PdQ = -QdP - PdQ + \frac{\partial S_2}{\partial P}dP + \frac{\partial S_2}{\partial Q}dQ \tag{8.65}$$

implies

$$Q = \frac{\partial S_2}{\partial P}, \qquad p = \frac{\partial S_2}{\partial q}. \tag{8.66}$$

Again, if we require

$$\det \frac{\partial S_2}{\partial P \partial q} \neq 0, \tag{8.67}$$

we obtain a canonical transformation

$$(P, Q) = (P(p, q), \frac{\partial S_2}{\partial P}(P(p, q), q)). \tag{8.68}$$

The remaining two cases

$$S = qp + S_3(Q, p) \quad \text{and} \quad S = qp - PQ + S_4(P, p) \tag{8.69}$$

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 γ_E which are uniquely determined by the energy E of a point on the orbit. The two intersection points of γ_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

$$I(E) = \frac{1}{2\pi} \int_{\gamma_E} p\, dq = \frac{1}{\pi} \int_{q_-(E)}^{q_+(E)} \sqrt{2(E - U(q))} dq \tag{8.70}$$

is called the **action variable**. Next, by (6.47)

$$I'(E) = \frac{1}{\pi} \int_{q_-(E)}^{q_+(E)} \frac{dq}{\sqrt{2(E - U(q))}} = \frac{T(E)}{2\pi} > 0, \tag{8.71}$$

where $T(E)$ is the period of γ_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

$$S_2(I, q) = \int_{q_-(K(I))}^{q} p dq = \int_{q_-(K(I))}^{q} \sqrt{2(K(I) - U(q))} dq \tag{8.72}$$

and the second variable is

$$\theta = \frac{\partial S_2}{\partial I} = \int_{q_-(E)}^{q} \frac{I'(E)^{-1} dq}{\sqrt{2(E - U(q))}} = \frac{2\pi}{T(E)} t, \tag{8.73}$$

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 θ is called the **angle variable** and is only defined modulo 2π. The equation of motion read

$$\dot{I} = -\frac{\partial K}{\partial \theta} = 0,$$
$$\dot{\theta} = \frac{\partial K}{\partial I} = \Omega(I), \qquad (8.74)$$

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 Γ_c one can define the action variables as before via

$$I_j(c) = \frac{1}{2\pi} \int_{\gamma_j(c)} p\, dq \qquad (8.75)$$

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

$$U(q) = U(q_0) + \frac{1}{2} q W q + o(|q|^2), \qquad (8.76)$$

where W is a positive matrix and $U(q_0)$ can be chosen zero. Neglecting the higher order terms, the resulting model

$$H(p,q) = \frac{1}{2}(pMp + qWq) \qquad (8.77)$$

is known as **harmonic oscillator**. Let V be the (real) orthogonal matrix which transforms the symmetric matrix $M^{-1/2} W M^{-1/2}$ to diagonal form and let ω_j^2 be the eigenvalues. Then the symplectic transform $(P,Q) = (VM^{1/2}p, VM^{-1/2}q)$ (Problem 8.15) gives the decoupled system

$$\dot{Q}_j = P_j, \qquad \dot{P}_j = -\omega_j^2 Q_j, \qquad j = 1, \dots, n. \qquad (8.78)$$

In particular,

$$K(P,Q) = \sum_{j=1}^n K_j, \qquad K_j = \frac{1}{2}(P_j^2 + Q_j^2), \qquad (8.79)$$

where the K_j's are n first integrals in involution (check this). The corresponding action-angle variables are given by (Problem 8.17)

$$I_j = \frac{1}{2}(\frac{P_j^2}{\omega_j} + \omega_j Q_j^2), \qquad \theta_j = \operatorname{arccot} \frac{P_j}{\omega_j Q_j}. \tag{8.80}$$

For example, consider the following Hamiltonian

$$H(p,q) = \sum_{j=1}^{n} \frac{p_j}{2m} + U_0\left(q_{j+1} - q_j\right), \quad q_0 = q_{n+1} = 0 \tag{8.81}$$

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 matrix 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 8.18). This is related to the existence of solitons, see for example [**30**].

Problem 8.14 (Symplectic group)**.** *Show that* 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 8.15. *Show that the linear transformation* $(P,Q) = (Up, (U^{-1})^T q)$*, where* U *is an arbitrary matrix, is canonical.*

Problem 8.16. *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 8.17. *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 8.18 (Fermi–Pasta–Ulam experiment)**.** *Consider the Hamiltonian* (8.81) *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 8.19 (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* (8.43) *are equivalent to the* **Lax equation**

$$\dot{L} = [P, L].$$

Show that the quantities

$$\operatorname{tr}(L^j), \qquad 1 \le j \le n,$$

are first integrals (Hint: Compare Problem 3.29).

8.5. 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

$$T(\dot{x}) = \frac{m_1}{2}\dot{x}_1^2 + \frac{m_2}{2}\dot{x}_2^2 \tag{8.82}$$

and the potential energy is

$$U(x) = U(|x_1 - x_2|). \tag{8.83}$$

The Lagrangian is the difference of both

$$L(\dot{x}, x) = T(\dot{x}) - U(x). \tag{8.84}$$

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 8.9 tells us that all three components of the total momentum

$$m_1\dot{x}_1 + m_2\dot{x}_2 \tag{8.85}$$

are first integrals. Hence we will choose new coordinates

$$q_1 = \frac{m_1 x_1 + m_2 x_2}{m_1 + m_2}, \qquad q_2 = x_1 - x_2 \tag{8.86}$$

in which our Lagrangian reads

$$L(\dot{q}, q) = \frac{M}{2}\dot{q}_1^2 + \frac{\mu}{2}\dot{q}_2^2 - U(q_2), \qquad M = m_1 + m_2,\ \mu = \frac{m_1 m_2}{M}. \tag{8.87}$$

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 8.9, we infer that the **angular momentum**

$$l = \mu q_2 \wedge \dot{q}_2 \tag{8.88}$$

is another first integral. Hence we have found three first integrals and we suspect that our system is integrable. However, since

$$\{l_1, l_2\} = l_3, \quad \{l_1, l_3\} = -l_2, \quad \{l_2, l_3\} = l_1 \tag{8.89}$$

they are not in involution. But using $\{l, |l|^2\} = 0$ it is not hard to see

Theorem 8.12. *The two body problem is completely integrable. A full set of first integrals which are functionally independent and in involution is given by*

$$p_{11}, \quad p_{12}, \quad p_{13}, \quad \frac{\mu}{2}p_2^2 + U(q_2), \quad |l|^2, \quad l_3, \tag{8.90}$$

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, φ) in this plane. The angular momentum now reads

$$l_0 = |l| = \mu r^2 \dot{\varphi} \tag{8.91}$$

and conservation of energy implies

$$\frac{\mu}{2}\left(\dot{r}^2 + \frac{l_0^2}{\mu^2 r^2}\right) + U(r) = E. \tag{8.92}$$

Hence, $r(t)$ follows (implicitly) from

$$\dot{r} = \sqrt{\frac{2(E - U(r))}{\mu} - \frac{l_0^2}{\mu^2 r^2}} \tag{8.93}$$

via separation of variables. In case of the Kepler problem (gravitational force)

$$U(r) = -\frac{\gamma}{r} \tag{8.94}$$

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

$$\frac{1}{r^2}\frac{dr}{d\varphi} = \sqrt{\frac{2\mu(E - U(r))}{l_0^2} - \frac{1}{r^2}}. \tag{8.95}$$

The solution is given by (Problem 8.20)

$$r(\varphi) = \frac{p}{1 - \varepsilon\cos(\varphi - \varphi_0)}, \qquad p = \frac{l_0^2}{\gamma\mu},\ \varepsilon = \sqrt{1 + \frac{2El_0^2}{\mu\gamma^2}} \tag{8.96}$$

Thus the orbit is an ellipsis if $\varepsilon < 1$, a parabola if $\varepsilon = 1$, and a hyperbola if $\varepsilon > 1$.

In the case of an ellipsis the motion is periodic and the period T is given by bringing the square root in (8.93) to the left and integrating from the smallest radius r_- to the largest r_+:

$$\frac{T}{2} = \frac{l_0}{\mu}\int_{r_-}^{r_+} \left(\left(\frac{1}{r} - \frac{1}{r_+}\right)\left(\frac{1}{r_-} - \frac{1}{r}\right)\right)^{-1/2} dr = \pi\sqrt{\frac{\mu}{\gamma}}\left(\frac{p}{1-\varepsilon^2}\right)^{3/2}, \tag{8.97}$$

where $r_\pm = \frac{p}{1\mp\varepsilon}$.

Equations (8.96), (8.91), and (8.97) establish **Kepler's first, second, and third law for planetary motion**:

(i) The orbit of every planet is an ellipse with the Sun at one focus.

(ii) A line segment joining a planet and the Sun sweeps out equal areas during equal time intervals.

(iii) The square of the orbital period of a planet is directly proportional to the cube of the semi-major axis of its orbit.

Problem 8.20. *Solve* (8.95). *(Hint: Use the transformation $\rho = r^{-1}$.)*

8.6. 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

$$T(\dot{x}) = \sum_{j=1}^{N} \frac{m_j}{2}\dot{x}_j^2 \tag{8.98}$$

and the potential energy is

$$U(x) = \sum_{1\le j<k\le N} U_{jk}(|x_j - x_k|). \tag{8.99}$$

In case of the gravitational force one has

$$U_{jk}(|x_j - x_k|) = \frac{m_j m_k}{|x_j - x_k|}. \tag{8.100}$$

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 $N \geq 3$. 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

$$H(p,q) = H_0(p,q) + \varepsilon\, H_1(p,q), \tag{8.101}$$

where H_0 is completely integrable and ε is *small.* Since H_0 is integrable, we can choose corresponding action angle variables (I,θ) and it hence suffices to consider systems of the type

$$H(I,\theta) = H_0(I) + \varepsilon\, H_1(I,\theta), \tag{8.102}$$

where $I \in \mathbb{R}^n$ and all components of θ have to be taken modulo 2π, that is, θ lives on the torus $\mathbb{T}^n$.

By (8.74) the unperturbed motion for $\varepsilon = 0$ is given by

$$I(t) = I_0, \qquad \theta(t) = \theta_0 + \Omega(I_0)t. \tag{8.103}$$

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**,

$$k\Omega(I_0) \neq 0 \quad \text{for all } k \in \mathbb{Z}^n\backslash\{0\}, \tag{8.104}$$

then each orbit is dense. On the other hand, if the frequencies $\Omega(I_0)$ are **resonant**,

$$k\Omega(I_0) = 0 \quad \text{for some } k \in \mathbb{Z}^n\backslash\{0\}, \tag{8.105}$$

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,

$$\Omega(I_0) = k\omega \quad \text{for some } k \in \mathbb{Z}^n,\ \omega \in \mathbb{R}. \tag{8.106}$$

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 Ω satisfying the following **diophantine condition**

$$|k\Omega| \geq \frac{\alpha}{|k|^n} \quad \text{for all } k \in \mathbb{Z}^n \backslash \{0\}. \tag{8.107}$$

Then the following famous result by Kolmogorov, Arnold, and Moser holds

Theorem 8.13 (KAM). *Suppose H_0, H_1 are analytic on $D \times \mathbb{T}^n$ and H_0 is nondegenerate, that is,*

$$\det\left(\frac{\partial H_0}{\partial I}\right) \neq 0. \tag{8.108}$$

Then there exists a constant $\delta > 0$ such that for

$$|\varepsilon| < \delta\alpha^2 \tag{8.109}$$

all Kronecker tori Γ_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 book. However, we will at least consider a simpler toy problem which illustrates some of the ideas and, in particular, explains where the diophantine condition (8.107) comes from. See the books by Arnold [**2**] or Moser [**28**] for further details and references.

But now we come to our toy problem. We begin with the system

$$\dot{x} = Ax, \qquad A = \begin{pmatrix} \mathrm{i}\omega_1 & & \\ & \ddots & \\ & & \mathrm{i}\omega_n \end{pmatrix}, \quad \omega_j \in \mathbb{R}, \tag{8.110}$$

where the solution is quasi-periodic and given by

$$x_j(t) = (\mathrm{e}^{At}c)_j = c_j \mathrm{e}^{\mathrm{i}\omega_j t}. \tag{8.111}$$

Next we perturb this system according to

$$\dot{x} = Ax + g(x), \tag{8.112}$$

where $g(x)$ has a convergent power series

$$g(x) = \sum_{|k|\geq 2} g_k x^k, \qquad k \in \mathbb{N}_0^n, \tag{8.113}$$

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

$$x(t) = \sum_{|k|\geq 1} c_k \mathrm{e}^{\mathrm{i}\omega k\, t} \tag{8.114}$$

or equivalently

$$x(t) = u(\mathrm{e}^{At}c), \tag{8.115}$$

where

$$u(x) = x + \sum_{|k|\geq 2} u_k x^k. \tag{8.116}$$

Inserting this ansatz into (8.112) gives

$$\frac{\partial u}{\partial x}(x)Ax = Au(x) + g(u(x)), \tag{8.117}$$

that is,

$$\sum_{|k|\geq 2} (\omega k - A)u_k x^k = g(x + \sum_{|k|\geq 2} u_k x^k). \tag{8.118}$$

Comparing coefficients of x^k shows that

$$(\mathrm{i}\omega k - A)u_k = \text{terms involving } u_\ell \text{ for } |\ell| < |k|. \tag{8.119}$$

Hence the coefficients u_k can be determined recursively provided

$$\omega k - \omega_j \neq 0 \text{ for all } |k| \geq 2, 1 \leq j \leq n. \tag{8.120}$$

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 δ, τ such that

$$|\omega k - \omega_j| \geq \frac{\delta}{|k|^\tau} \tag{8.121}$$

holds. Moreover, it can be shown that the set of frequencies ω satisfying (8.121) 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 8.21.

Problem 8.21. *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.)

Chapter 9

Local behavior near fixed points

9.1. Stability of linear systems

Our aim in this chapter is to show that a lot of information on the stability of a flow near a fixed point can be read off by linearizing the system around the fixed point. As a preparation we recall the stability discussion for linear systems

$$\dot{x} = Ax \tag{9.1}$$

from Section 3.2. Clearly, our definition of stability in Section 6.5 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, recall that, by virtue of the explicit form (3.42) 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,

$$E^{\pm}(\mathrm{e}^A) = \bigoplus_{\pm \mathrm{Re}(\alpha_j)<0} \mathrm{Ker}(A-\alpha_j)^{a_j}. \tag{9.2}$$

Similarly one can define the **center manifold** $E^0(\mathrm{e}^A)$ corresponding to the eigenvalues with zero real part. However, since the center manifold is generally not stable under small perturbations, one often assumes that it is empty. Hence we will give a system where all eigenvalues have nonzero real part a special name. They are called **hyperbolic** systems.

If all eigenvalues have negative real part we have the following result from Section 3.2 which summarizes Corollary 3.5 and Corollary 3.6.

Theorem 9.1. *Denote the eigenvalues of A by α_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

$$\|\exp(tA)\| \le C\mathrm{e}^{-t\alpha}, \qquad t \ge 0. \tag{9.3}$$

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 9.2. *The linear stable and unstable manifolds $E^\pm = E^\pm(\mathrm{e}^A)$ 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*

$$|\exp(tA)x_\pm| \le C\mathrm{e}^{\mp t\alpha}|x_\pm|, \quad \pm t \ge 0, \quad x_\pm \in E^\pm, \tag{9.4}$$

for any $\alpha < \min\{|\mathrm{Re}(\alpha_j)|\,|\alpha_j \in \sigma(A), \pm\mathrm{Re}(\alpha_j) < 0\}$ and some $C > 0$ depending on α.

For our further investigations, it is also useful to introduce the space spanned by all generalized eigenvectors of A corresponding to eigenvalues with real part less/bigger than $\mp\alpha$,

$$E^{\pm,\alpha}(\mathrm{e}^A) = \bigoplus_{\mp\mathrm{Re}(\alpha_j)>\alpha} \mathrm{Ker}(A-\alpha_j)^{a_j} = E^\pm(\mathrm{e}^{A\pm\alpha}). \tag{9.5}$$

Equivalently,

$$E^{\pm,\alpha}(\mathrm{e}^A) = \{x | \lim_{t\to\pm\infty} \mathrm{e}^{\pm\alpha t}|\exp(tA)x| = 0\}, \tag{9.6}$$

is the space spanned by all initial conditions which converge to 0 with some given exponential rate $\alpha > 0$. Note that $E^{\pm,\alpha}$ is piecewise constant and will jump at those values of α which are equal to the real part of some eigenvalue of A.

Problem 9.1. *For the matrices in Problem 3.9. Determine the stability of the origin and, if the system is hyperbolic, find the corresponding stable and unstable manifolds.*

Problem 9.2. *Let A be a real-valued 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 real-valued 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 $TM - D = (\alpha_1 + \alpha_2)(\alpha_1 + \alpha_3)(\alpha_2 + \alpha_3)$.)

9.2. 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,

$$W^{\pm}(x_0) = \{x \in M | \lim_{t\to\pm\infty} |\Phi(t, x) - x_0| = 0\}. \tag{9.7}$$

Both sets are obviously invariant under the flow. Our goal in this section is to investigate these sets.

Any function $f \in C^1$ vanishing at $x_0 \in M$ can be decomposed as

$$f(x) = A\,(x - x_0) + g(x), \tag{9.8}$$

where A is the Jacobian matrix of f at x_0 and $g(x) = o(|x - x_0|)$. Clearly, in a sufficiently small neighborhood of x_0 we expect the solutions to be described by the solutions of the linearized equation. This is true for small t by Theorem 2.8, but what about $|t| \to \infty$? In Section 6.5 we saw that for $n = 1$ stability can be read off from $A = f'(x_0)$ alone as long as $f'(x_0) \neq 0$. 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.

Since our result is of a local nature we fix a neighborhood $U(x_0)$ of x_0 and define

$$M^{\pm,\alpha}(x_0) = \{x | \gamma_{\pm}(x) \subseteq U(x_0) \text{ and } \sup_{\pm t\geq 0} \mathrm{e}^{\pm\alpha t}|\Phi(t, x) - x_0| < \infty\} \tag{9.9}$$

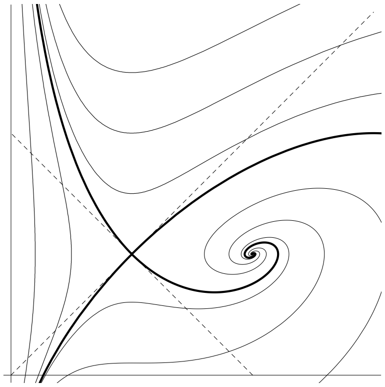


Figure 9.1. Phase portrait for a planar system with a hyperbolic fixed point $(1,1)$ together with the stable/unstable manifold (thick) and their linear counterparts (dashed).

to be the set of all points which converge to x_0 with some exponential rate $\alpha > 0$ as $t \to \pm\infty$. This is the counterpart of $E^{\pm,\alpha}$, the space spanned by all eigenvectors of A corresponding to eigenvalues with real part less/bigger than $\mp\alpha$. Now we define the local **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,

$$M^{\pm}(x_0) = \bigcup_{\alpha>0} M^{\pm,\alpha}(x_0). \tag{9.10}$$

Both sets are ± invariant under the flow by construction.

In the linear case we clearly have $M^{\pm}(0) = E^{\pm}$. Our goal is to show, as a generalization of Theorem 9.2, that the sets $M^{\pm}(x_0)$ are indeed manifolds (smooth) and that $E^{\pm}$ is tangent to $M^{\pm}(x_0)$ at x_0, as illustrated in Figure 9.1. Finally, we will show that $M^{\pm}(x_0) = W^{\pm}(x_0)$ in the hyperbolic case.

For notational convenience we will assume that $x_0 = 0$ is our hyperbolic fixed point. The key idea is again to reformulate 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 (3.97) to rewrite our equation as

$$x(t) = \mathrm{e}^{tA}x(0) + \int_0^t \mathrm{e}^{(t-r)A} g(x(r))dr. \tag{9.11}$$

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_- as a function of x_+: $x_- = h^+(x_+)$. To this end we project out the unstable part of our integral equation and solve for x_-:

$$x_- = \mathrm{e}^{-tA} x_-(t) - \int_0^t \mathrm{e}^{-rA} g_-(x(r)) dr. \tag{9.12}$$

Here $x_\pm(t) = P^\pm x(t)$. If we suppose that $|x(t)|$ is bounded for $t \geq 0$, we can let $t \to \infty$,

$$x_- = -\int_0^\infty \mathrm{e}^{-rA} g_-(x(r)) dr, \tag{9.13}$$

where the integral converges absolutely since the integrand decays exponentially. Plugging this back into our equation we see

$$x(t) = \mathrm{e}^{tA} x_+ + \int_0^t \mathrm{e}^{(t-r)A} g_+(x(r)) dr - \int_t^\infty \mathrm{e}^{(t-r)A} g_-(x(r)) dr. \tag{9.14}$$

Introducing $P(t) = P^+$, $t > 0$, respectively $P(t) = -P^-$, $t \leq 0$, this can be written more compactly as

$$x(t) = K(x)(t), \quad K(x)(t) = \mathrm{e}^{tA} x_+ + \int_0^\infty \mathrm{e}^{(t-r)A} P(t-r) g(x(r)) dr. \tag{9.15}$$

In summary, if A is hyperbolic, then every bounded solution solves (9.15) and we can establish existence of solutions using similar fixed point techniques as in Section 2.1. This will prove 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 $t \to -t$.

In fact, we can do even a little better.

Theorem 9.3. *Suppose $f \in C^k$, $k \geq 1$, has a fixed point x_0 with corresponding Jacobian matrix A. Then, if $\alpha > 0$ and $A + \alpha\mathbb{I}$ is hyperbolic, there is a neighborhood $U(x_0) = x_0 + U$ and a function $h^{+,\alpha} \in C^k(E^{+,\alpha} \cap U, E^{-,\alpha})$ such that*

$$M^{+,\alpha}(x_0) \cap U(x_0) = \{x_0 + a + h^{+,\alpha}(a) | a \in E^{+,\alpha} \cap U\}. \tag{9.16}$$

Both $h^{+,\alpha}$ and its Jacobian matrix vanish at 0, that is, $M^{+,\alpha}(x_0)$ is tangent to its linear counterpart $E^{+,\alpha}$ at x_0.

We have $M^{+,\alpha_2}(x_0) \subseteq M^{+,\alpha_1}(x_0)$ for $\alpha_1 \leq \alpha_2$ and $M^{+,\alpha_2}(x_0) = M^{+,\alpha_1}(x_0)$ whenever $E^{+,\alpha_2} = E^{+,\alpha_1}$.

Proof. We suppose $x_0 = 0$ and begin by assuming that A is hyperbolic such that we can choose $\alpha = 0$. Our underlying Banach space will be $C_b([0,\infty), \mathbb{R}^n)$ equipped with the sup norm

$$\|x\| = \sup_{t \geq 0} |x(t)|.$$

To solve (9.15) by iteration, suppose $|x(t)| \le \delta$. Then, since the Jacobian matrix of g at 0 vanishes, we have

$$|g(x(t)) - g(y(t))| \le \varepsilon\, |x(t) - y(t)|, \tag{9.17}$$

where ε can be made arbitrarily small by choosing δ sufficiently small. Moreover, for $\alpha_0 < \min\{|\mathrm{Re}(\alpha)|\, |\alpha \in \sigma(A)\}$ we have

$$\|\mathrm{e}^{(t-r)A} P(t-r)\| \le C\mathrm{e}^{-\alpha_0|t-r|}$$

by (9.4). Combining this with (9.17) we obtain

$$\begin{aligned}
\|K(x) - K(y)\| &= \sup_{t\ge 0} \left| \int_0^\infty \mathrm{e}^{(t-r)A} P(t-r)\big(g(x(r)) - \ g(y(r))\big)\, dr \right| \\
&\le C \sup_{t\ge 0} \int_0^\infty \mathrm{e}^{-\alpha_0|t-r|} \big|g(x(r)) - \ g(y(r))\big| dr \\
&\le C\varepsilon \|x - y\| \sup_{t\ge 0} \int_0^\infty \mathrm{e}^{-\alpha_0|t-r|} dr = \frac{2C\varepsilon}{\alpha_0} \|x - y\|.
\end{aligned}$$

Hence, for $\varepsilon < \alpha_0/(2C)$ existence of a unique solution $\psi(t, x_+)$ can be established by the contraction principle (Theorem 2.1). However, by Theorem 9.18 (see Section 9.4 below) we even get $\psi(t, x_+)$ is C^k with respect to x_+ if f is.

Clearly we have $\psi(t, 0) = 0$. Introducing the function $h^+(a) = P^-\psi(0, a)$ we obtain $M^+(0) \cap U = \{a + h^+(a) | a \in E^+ \cap U\}$ for the stable manifold of the nonlinear system in a neighborhood U of 0.

Moreover, I claim that $M^+(0)$ is tangent to E^+ at 0. From the proof of Theorem 9.18 it follows that $\varphi(t, x_+) = \frac{\partial}{\partial x_+}\psi(t, x_+)$ satisfies

$$\varphi(t, x_+) = \mathrm{e}^{tA} P^+ + \int_0^\infty \mathrm{e}^{(t-r)A} P(t-r) g_x(\psi(r, x_+))\varphi(r, x_+) dr. \tag{9.18}$$

Evaluating this equation at $(t, x_+) = (0, 0)$ we see $\varphi(0, 0) = P^+$ which is equivalent to

$$\frac{\partial}{\partial a} h^+(a)\Big|_{a=0} = 0, \tag{9.19}$$

that is, $M^+(0)$ is tangent to the linear stable manifold E^+ at 0.

To see the general case, 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 $\sup_{t\ge 0} |\tilde{x}(t)| \le \delta$, that $\sup_{t\ge 0} |x(t)| \le \delta \exp(-\alpha\, t)$. By uniqueness of the solution of our integral equation in a sufficiently small neighborhood of x_0 we obtain (9.16).

For the last claim let $x \in M^{+,\alpha_2}(x_0) \cap U(x_0) \subseteq M^{+,\alpha_1}(x_0) \cap U(x_0)$, then $x = x_0 + a + h^{+,\alpha_2}(a) = x_0 + a + h^{+,\alpha_1}(a)$ for $a \in E^{+,\alpha_1} = E^{+,\alpha_2}$ implies $h^{+,\alpha_2}(a) = h^{+,\alpha_1}(a)$. From this the claim follows. □

As a 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 $\varepsilon > 0$ small such that $E^+ = E^{+,\varepsilon}$.

Theorem 9.4 (Stable manifold)**.** *Suppose $f \in C^k$, $k \geq 1$, has a fixed point x_0 with corresponding Jacobian matrix A. Then, there is a neighborhood $U(x_0) = x_0 + U$ and functions $h^\pm \in C^k(E^\pm \cap U, E^\mp)$ such that*

$$M^\pm(x_0) \cap U(x_0) = \{x_0 + a + h^\pm(a) | a \in E^\pm \cap U\}. \tag{9.20}$$

Both $h^\pm$ and their Jacobian matrices vanish at x_0, that is, $M^\pm(x_0)$ are tangent to their respective linear counterpart $E^\pm$ at x_0. Moreover,

$$|\Phi(t,x) - x_0| \leq C \mathrm{e}^{\mp t\alpha}, \pm t \geq 0, x \in M^\pm \tag{9.21}$$

for any $\alpha < \min\{|\mathrm{Re}(\alpha_j)|\, |\alpha_j \in \sigma(A), \mathrm{Re}(\alpha_j) \neq 0\}$ and some $C > 0$ depending on α.

It can be shown that even a nonlinear counterpart of the center subspace E^0 exists. However, such a center manifold might not be unique (Problem 9.9).

In the hyperbolic case we can even say a little more.

Theorem 9.5. *Suppose $f \in C^k$, $k \geq 1$, 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 $x \in M^\pm(x_0) \cap U(x_0)$. In particular,*

$$W^\pm(x_0) = M^\pm(x_0). \tag{9.22}$$

Proof. This follows since we have shown that any solution staying sufficiently close to x_0 solves (9.14). Hence uniqueness of the solution (in a sufficiently small neighborhood of x_0) implies that the initial value must lie in $M^+(x_0)$. □

Example. Consider the vector field

$$f(x) = (-x_1 + x_2 + 3x_2^2, x_2). \tag{9.23}$$

Then it is not hard to check (start with the second equation) that its flow is given by

$$\Phi(t,x) = (x_1 \mathrm{e}^{-t} + x_2 \sinh(t) + x_2^2(\mathrm{e}^{2t} - \mathrm{e}^{-t}), x_2 \mathrm{e}^t). \tag{9.24}$$

Moreover, there is only one fixed point $x_0 = 0$ and the corresponding stable and unstable manifolds are given by

$$W^+(0) = \{x | x_2 = 0\}, \qquad W^-(0) = \{x | x_1 = \frac{x_2}{2} + x_2^2\}. \tag{9.25}$$

The linearization is given by

$$A = \begin{pmatrix} -1 & 1 \\ 0 & 1 \end{pmatrix} \tag{9.26}$$

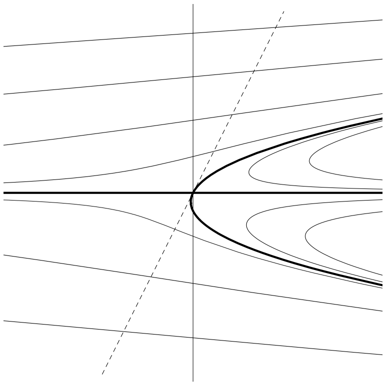


Figure 9.2. Phase portrait for a planar system with a hyperbolic fixed point $(0,0)$ together with the stable/unstable manifold (thick) and their linear counterparts (dashed).

and both $W^+(0)$ and $W^-(0)$ are tangent to their linear counterparts

$$E^+ = \{x|x_2 = 0\}, \qquad E^- = \{x|x_1 = \frac{x_2}{2}\}. \tag{9.27}$$

The system is depicted in Figure 9.2. ◇

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 another proof of Theorem 6.10. 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 λ sufficiently small. In particular we have

$$f(x,\lambda) = A(\lambda)(x - x_0(\lambda)) + g(x,\lambda), \tag{9.28}$$

where $A(\lambda)$ is the Jacobian matrix of $f(.,\lambda)$ at $x_0(\lambda)$. By Problem 3.47, the projectors $P^\pm(\lambda) = P^\pm(A(\lambda))$ vary smoothly with respect to λ and we can proceed as before to obtain (compare Problem 9.12)

Theorem 9.6. *Suppose $f \in C^k$, $k \geq 1$, and let $x_0(\lambda)$ be as above. Then, there is a neighborhood $U(x_0) = x_0 + U$ and functions $h^\pm \in C^k(E^\pm \cap U \times \Lambda, E^\mp)$ such that*

$$M^\pm(x_0(\lambda)) \cap U(x_0) = \{x_0(\lambda) + P^\pm(\lambda)a + h^\pm(a,\lambda) | a \in E^\pm \cap U\}. \tag{9.29}$$

Problem 9.3. *Find the subspaces* $E^{\pm,\alpha}$ *for*

$$A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -2 \end{pmatrix}.$$

Compute the projections $P^{\pm}$.

Problem 9.4. *Find the linearization of*

$$f(x) = (x_2, -\sin(x_1)).$$

and determine the stability of $x = 0$ *if possible.*

Problem 9.5 (Duffing equation)**.** *Investigate the Duffing equation*

$$\ddot{x} = -\delta\dot{x} + x - x^3, \quad \delta \ge 0.$$

Determine the stability of the fixed points by linearization. Find the stable and unstable manifolds of the origin in the case $\delta = 0$.

Problem 9.6. *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 9.7 (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 9.8 (Homoclinic orbit)**.** *Determine the stability of the fixed points of the system in Problem 6.23 by linearization. Find the stable and unstable manifolds. Find a homoclinic orbit.*

Problem 9.9. *Consider*

$$\dot{x} = -x, \qquad \dot{y} = y^2.$$

Find all invariant smooth manifolds of the form $\{(h(a), a) | a \in \mathbb{R}\}$ *which are tangent to* E^0.

Problem 9.10. *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 (9.14)*. Plot the phase portrait (numerically) and compare it to the linearization.*

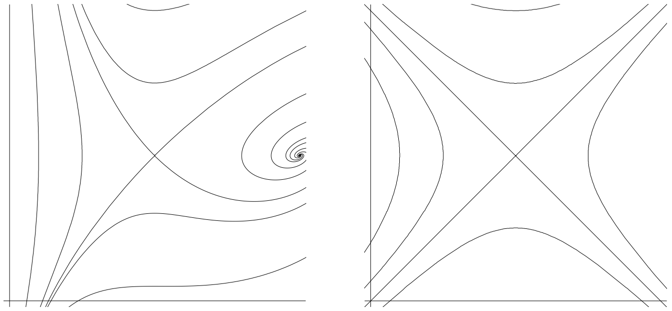


Figure 9.3. Phase portrait for a planar system with a hyperbolic fixed point $(1,1)$ together with the phase portrait of its linearization.

Problem 9.11. *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 which value of r does the number of fixed points change?

Problem 9.12. *Suppose $A(\lambda)$ is a matrix which is C^k with respect to λ in some compact set. Suppose there is an $0 < \alpha_0 < \min\{|\mathrm{Re}(\alpha)|\,|\alpha \in \sigma(A(\lambda))\}$. Then*

$$\| \left(\frac{d}{d\lambda}\right)^n \mathrm{e}^{tA(\lambda)} P(\lambda,t)\| \le C_n(1+|t|^n)\mathrm{e}^{-\alpha_0|t|}, \quad n \le k.$$

(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 ...)

9.3. 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. This is illustrated in Figure 9.3.

If we assume that A has no eigenvalues on the unit circle, we can use $\mathbb{R}^n = E^-(A) \oplus E^+(A)$ to split it into a contracting and expanding part $A = A_- \oplus A_+$, where $A_\pm = A\big|_{E^\pm(A)}$. By construction, all eigenvalues of A_+ are inside the unit circle and all eigenvalues of A_- are outside the unit circle. Hence, by Problem 3.48 we can find a norm such that $\|A_+\| < 1$.

We begin with a lemma for maps.

Lemma 9.7. *Suppose A is an invertible matrix with no eigenvalues on the unit circle and choose a norm such that $\alpha = \max(\|A_-^{-1}\|, \|A_+\|) < 1$ (set*

$\|A_-^{-1}\| = 0$ *if there are no eigenvalues outside the unit circle). Then for every bounded g satisfying*

$$|g(x) - g(y)| \le \varepsilon |x - y|, \qquad \varepsilon < \frac{1-\alpha}{2}, \tag{9.30}$$

there is a unique continuous map $\varphi(x) = x + h(x)$ with h bounded such that

$$\varphi \circ A = f \circ \varphi, \qquad f = A + g. \tag{9.31}$$

If f is invertible (e.g. if $\varepsilon\|A^{-1}\| < 1$), then φ is a homeomorphism and if in addition $g(0) = 0$ then $\varphi(0) = 0$.

Proof. We will assume that A has eigenvalues both inside and outside the unit circle. The modifications for the two remaining cases are straightforward.

The requirement (9.31) is equivalent to

$$h(Ax) - Ah(x) = g(x + h(x)). \tag{9.32}$$

We will investigate this equation in the Banach space of continuous functions $C(\mathbb{R}^n, \mathbb{R}^n)$ with the sup norm. First of all note that the linear operator $U : C(\mathbb{R}^n, \mathbb{R}^n) \to C(\mathbb{R}^n, \mathbb{R}^n)$ given by $(Uh)(x) = h(Ax)$ is invertible (since A is) and norm preserving. Clearly we can also regard A as a linear operator $A : C(\mathbb{R}^n, \mathbb{R}^n) \to C(\mathbb{R}^n, \mathbb{R}^n)$ given by $(Ah)(x) = Ah(x)$.

Introducing $L = U - A$ we can write (9.32) as $Lh(x) = g(x + h(x))$. To obtain a fixed point equation we need to invert L. By splitting $C(\mathbb{R}^n, \mathbb{R}^n) = C(\mathbb{R}^n, E^-(A)) \oplus C(\mathbb{R}^n, E^+(A))$ we obtain corresponding splittings $A = A_- \oplus A_+$, $U = U_- \oplus U_+$, and $L = L_- \oplus L_+$ (note that both A and U leave these spaces invariant).

By $L_- = -A_-(\mathbb{I} - A_-^{-1}U_-)$ we see that $L_-^{-1} = -(\mathbb{I} - A_-^{-1}U_-)^{-1}A_-^{-1}$, where $(\mathbb{I} - A_-^{-1}U_-)$ is invertible with inverse given by the Neumann series (Problem 9.13)

$$(\mathbb{I} - A_-^{-1}U_-)^{-1} = \sum_{n=0}^{\infty} (A_-^{-1}U_-)^n$$

since $\|A_-^{-1}U_-\| \le \alpha$. In particular, $\|L_-^{-1}\| \le \frac{1}{1-\alpha}$. Similarly, $L_+^{-1} = (\mathbb{I} - U_+^{-1}A_+)^{-1}U_+^{-1}$ with $\|L_+^{-1}\| \le \frac{1}{1-\alpha}$.

In summary, $L^{-1} = (U_- - A_-)^{-1} \oplus (U_+ - A_+)^{-1}$ exists and $\|L^{-1}\| \le \frac{2}{1-\alpha}$. Hence it remains to solve the fixed point equation

$$h(x) = L^{-1}g(x + h(x)).$$

Since the operator on the right is a contraction,

$$\begin{aligned}\|L^{-1}g(x+h_1(x)) - L^{-1}g(x+h_2(x))\| \\ \le \frac{2}{1-\alpha}\|g(x+h_1(x)) - g(x+h_2(x))\| \\ \le \frac{2\varepsilon}{1-\alpha}\|h_1 - h_2\|,\end{aligned}$$

the contraction principle (Theorem 2.1) guarantees existence of a unique solution.

Now suppose f is invertible. Then there is a map $\vartheta(x) = x + k(x)$ such that $A \circ \vartheta = \vartheta \circ f$. In fact, defining L as before but with $U(k)(x) = k(f(x))$ we see that this last equation is equivalent to $L(k)(x) = -g(x)$ and since the same argument as above shows that that L is invertible, we obtain $k(x) = -L^{-1}(g)(x)$. 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 φ 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 9.8. *Let A be as in the previous lemma and f arbitrary. Suppose there is a homeomorphism $\varphi(x) = x + h(x)$ with h bounded such that*

$$\varphi \circ A = f \circ \varphi, \tag{9.33}$$

then φ is unique.

Proof. Suppose there are two such maps φ_1 and φ_2 and note that the inverses φ_j^{-1} are of the same type (Problem 9.14). Then $f = \varphi_1 A \varphi_1^{-1} = \varphi_2 A \varphi_2^{-1}$ implies $A(\varphi_1^{-1}\varphi_2) = (\varphi_1^{-1}\varphi_2)A$ which shows that $\varphi_1^{-1}\varphi_2 = \mathbb{I}$ by our above lemma in the case $g \equiv 0$. □

Now we are able to prove the anticipated result.

Theorem 9.9 (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 matrix of f at 0. Then there is a homeomorphism $\varphi(x) = x + h(x)$ with h bounded such that*

$$\varphi \circ \mathrm{e}^{tA} = \Phi_t \circ \varphi \tag{9.34}$$

in a sufficiently small neighborhood of 0.

Proof. Our strategy is to apply Lemma 9.7 to find a φ which works for one fixed t, say $t = 1$, and then verify that it works in fact for all t.

First of all we will need to control

$$\Pi(t,x) = \frac{\partial}{\partial x}\Phi(t,x).$$

From

$$\dot{\Phi}(t,x) = f(\Phi(t,x)), \qquad \Phi(0,x) = x,$$

we obtain

$$\dot{\Pi}(t,x) = \frac{\partial f}{\partial x}(\Phi(t,x))\Pi(t,x), \qquad \Pi(0,x) = \mathbb{I}, \tag{9.35}$$

and, setting $x = 0$,

$$\Pi(t,0) = \mathrm{e}^{tA}.$$

Thus

$$\Phi_1(x) = \mathrm{e}^A x + G(x),$$

where (9.30) holds at least when we are sufficiently close to our fixed point. To make sure it always holds we will modify f.

Let $\phi : [0,\infty) \to \mathbb{R}$ be a smooth bump function such that $\phi(x) = 0$ for $0 \le x \le 1$ and $\phi(x) = 1$ for $x \ge 2$. Replacing $f(x) = Ax + g(x)$ by the function $\tilde{f}(x) = Ax + (1 - \phi(|x|/\delta))g(x)$, it is no restriction to consider the global problem with $f = A$ for $|x| \ge 2\delta$. Note that (show this!)

$$\left|\frac{\partial \tilde{g}}{\partial x}(x)\right| \le C \sup_{|x|\le 2\delta}\left|\frac{\partial g}{\partial x}(x)\right|$$

can be made arbitrarily small by choosing δ small. Moreover, note that $\tilde{G}(x)$ will be 0 for $|x|$ sufficiently large (e.g., for $|x| \ge 2\delta\mathrm{e}^{\alpha}$, where α is some nonnegative number which satisfies $\alpha \ge -\mathrm{Re}(\alpha_j)$ for all eigenvalues α_j of A). We will use $\tilde{f}$ from now on and drop the tilde for notational simplicity.

To be able to apply Lemma 9.7 we need to show that $z(1,x)$, defined by

$$\Pi(t,x) = \mathrm{e}^{tA} + z(t,x),$$

can be made arbitrarily small by choosing δ small. Plugging this into (9.35) we obtain

$$z(t,x) = \int_0^t \frac{\partial g}{\partial x}(\Phi(s,x))\mathrm{e}^{sA}ds + \int_0^t \frac{\partial f}{\partial x}(\Phi(s,x))z(s,x)ds$$

and the claim follows from Gronwall's inequality using that $\frac{\partial g}{\partial x}$ can be made arbitrarily small by choosing δ small as noted above.

Hence, there is a φ such that (9.34) holds at least for $t = 1$. Furthermore, the map $\varphi_s = \Phi_s \circ \varphi \circ \mathrm{e}^{-sA}$ also satisfies (9.34) for $t = 1$:

$$\varphi_s \circ \mathrm{e}^A = \Phi_s \circ \varphi \circ \mathrm{e}^A \circ \mathrm{e}^{-sA} = \Phi_s \circ \Phi_1 \circ \varphi \circ \mathrm{e}^{-sA} = \Phi_1 \circ \varphi_s.$$

Hence, if we can show that $\varphi_t(x) = x + h_t(x)$ with h_t bounded, then Corollary 9.8 will tell us $\varphi = \varphi_t$ which is precisely (9.34). Now observe

$$h_t = \Phi_t \circ \varphi \circ \mathrm{e}^{-tA} - \mathbb{I} = (\Phi_t - \mathrm{e}^{tA}) \circ \mathrm{e}^{-tA} + \Phi_t \circ h \circ \mathrm{e}^{-tA},$$

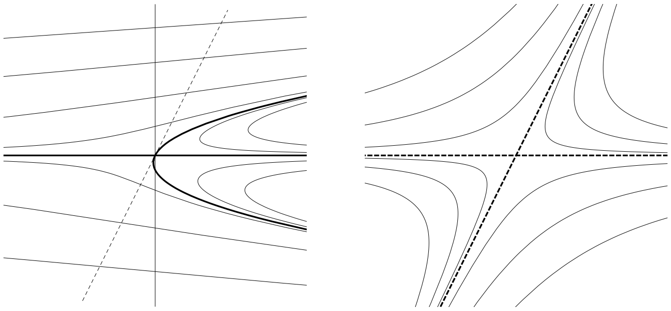


Figure 9.4. Phase portrait for a planar system with a hyperbolic fixed point $(0,0)$ together with the phase portrait of its linearization.

where the first term is bounded since $\Phi_t(x) = \mathrm{e}^{tA}x$ for sufficiently large x (say $|x| \ge 2\delta\mathrm{e}^{t\alpha}$ as pointed out before) and the second is since h is. □

Example. Consider again the vector field

$$f(x) = (-x_1 + x_2 + 3x_2^2, x_2). \tag{9.36}$$

Then one can verify that its flow (9.24) is mapped to its linear counterpart

$$\mathrm{e}^{tA} = \begin{pmatrix} \mathrm{e}^{-t} & \sinh(t) \\ 0 & \mathrm{e}^{t} \end{pmatrix} \tag{9.37}$$

by virtue of

$$\varphi(x) = (x_1 - x_2^2, x_2). \tag{9.38}$$

The system together with its linearization is depicted in Figure 9.4. ◇

Two systems with vector fields f, g and respective flows Φ_f, Φ_g are said to be **topologically conjugate** if there is a homeomorphism φ such that

$$\varphi \circ \Phi_{f,t} = \Phi_{g,t} \circ \varphi. \tag{9.39}$$

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 an even 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 9.10. *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 conjugation). 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 $t \ge 0$ and some small $\alpha > 0$ (Problem 3.48). Replacing $t \to -t$ and $x \to \exp(tA)x$ we also obtain $|\exp(tA)x|_A \ge \exp(-t\alpha)|x|_A$ for all $t \le 0$. Thus

$$\begin{aligned}\frac{d}{dt}|x(t)|_A &= \lim_{s\to 0} \frac{|\exp(sA)x(t)|_A - |x(t)|_A}{s} \\ &\le \lim_{s\to 0} \frac{\exp(-s\alpha) - 1}{s}|x(t)|_A = -\alpha|x(t)|_A\end{aligned}$$

for $t \ge 0$ and there is a unique time $\tau_A(x) > 0$ such that $|\exp(\tau(x)A)x|_A = 1$ for $|x(t)|_A > 1$. Similarly, $\frac{d}{dt}|x(t)|_A \ge -\alpha|x(t)|_A$ for $t \le 0$ and there is also a unique time $\tau_A(x) < 0$ such that $|\exp(\tau(x)A)x|_A = 1$ for $0 < |x(t)|_A \le 1$. Moreover, the unit sphere $|x|_A = 1$ is transversal and hence τ_A is a smooth function by Lemma 6.9. 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

$$h(x) = \exp(-\tau_A(x)B)h_{AB}(\exp(\tau_A(x)A)x), \qquad x \ne 0,$$

which is a homeomorphism from $\mathbb{R}^n\backslash\{0\}$ to itself. In fact its inverse is given by

$$h^{-1}(x) = \exp(-\tau_B(x)A)h_{BA}(\exp(\tau_B(x)B)x), \qquad x \ne 0,$$

which follows since $\tau_B(y) = \tau_A(x)$ if $y = h(x)$. Furthermore, since $\tau(x) \to -\infty$ as $x \to 0$ we have $|h(x)| \le c\|\exp(-\tau_A(x)B)\| \to 0$ as $x \to 0$. Thus we can extend h to a homeomorphism from $\mathbb{R}^n$ to itself by setting $h(0) = 0$.

Finally, h is a topological conjugation since

$$\begin{aligned}h(\exp(tA)x) &= \exp((t - \tau_A(x))B)h_{AB}(\exp((\tau_A(x) - t)A)\exp(tA)x) \\ &= \exp(tB)h(x),\end{aligned}$$

where we have used $\tau_A(\exp(tA)x) = \tau_A(x) - t$. □

Problem 9.13. *Let X be a Banach space and let $A : X \to X$ be a linear operator. Set*

$$\|A\| = \sup_{\|x\|=1} \|Ax\|.$$

Show that this defines a norm. Moreover, show that

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

and that $\mathbb{I}+A$ *is invertible if* $\|A\| < 1$*, with inverse given by the* **Neumann series**

$$(\mathbb{I}-A)^{-1} = \sum_{n=0}^{\infty} A^n.$$

Furthermore, $\|(\mathbb{I}-A)^{-1}\| \le (1-\|A\|)^{-1}$*.*

Problem 9.14. *Let* $\varphi : \mathbb{R}^n \to \mathbb{R}^n$ *be a homeomorphism of the form* $\varphi(x) = x + h(x)$ *with bounded* h*. Show that* $\varphi^{-1}(x) = x + k(x)$*, where* $k(x)$ *is again bounded (with the same bound).*

Problem 9.15. *Let*

$$A = \begin{pmatrix} -\alpha & \beta \\ -\beta & -\alpha \end{pmatrix}, \quad B = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}, \qquad \alpha > 0.$$

Explicitly compute the conjugacy found in the proof of Theorem 9.10.

9.4. Appendix: 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. In this appendix we will prove a few somewhat technical results which can be omitted on first reading.

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 Λ is a subset of some Banach space.

Theorem 9.11 (Uniform contraction principle)**.** *Let* C *be a (nonempty) closed subset of a Banach space* X *and* Λ *a subset of another Banach space. Suppose* $K_\lambda : C \to C$ *is a uniform contraction, that is,*

$$\|K_\lambda(x) - K_\lambda(y)\| \le \theta\|x-y\|, \qquad x, y \in C,\ \lambda \in \Lambda, \tag{9.40}$$

for some $\theta \in [0,1)$*, and* $K_\lambda(x)$ *is continuous with respect to* $\lambda \in \Lambda$ *for every* $x \in C$*. Then the unique fixed point* $\overline{x}(\lambda)$ *is continuous with respect to* λ*.*

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

$$x_{n+1} = K_{\lambda_n}(x_n) \to \overline{x}(\lambda). \tag{9.41}$$

Proof. Existence of $\overline{x}(\lambda)$ for fixed λ follows from Theorem 2.1. We first show that $\overline{x}(\lambda)$ is continuous. By the triangle inequality we have

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

and hence

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

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.

To see the last claim abbreviate $\Delta_n = \|x_n - \overline{x}(\lambda)\|$, $\varepsilon_n = \|\overline{x}(\lambda_n) - \overline{x}(\lambda)\|$ and observe

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

Hence

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

which converges to 0 since ε_n does (show this). □

There is also a uniform version of Theorem 2.4.

Theorem 9.12. *Let C be a (nonempty) closed subset of a Banach space X and Λ a subset of another Banach space. Suppose $K_\lambda : C \to C$ satisfies*

$$\|K_{\lambda_n}\circ\cdots\circ K_{\lambda_1}(x) - K_{\lambda_n}\circ\cdots\circ K_{\lambda_1}(y)\| \le \theta_n\|x-y\|, \quad x,y\in C, \lambda_j \in \Lambda, \tag{9.42}$$

with $\sum_{n=1}^\infty \theta_n < \infty$, and $K_\lambda(x)$ is continuous with respect to $\lambda \in \Lambda$ for every $x \in C$. Then the unique fixed point $\overline{x}(\lambda)$ is continuous with respect to λ.

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

$$x_{n+1} = K_{\lambda_n}(x_n) \to \overline{x}(\lambda). \tag{9.43}$$

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 $\underline{\lambda} \in \Lambda^n$ for fixed $x \in C$. 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

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

where $\underline{\lambda} = (\lambda_1, \dots, \lambda_{n-1})$ and $\underline{\eta} = (\eta_1, \dots, \eta_{n-1})$.

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 9.11. In particular, choosing $\underline{\lambda}_j = (\lambda_j, \dots, \lambda_{j+n-1})$ we have that $x_{n(j+1)+l} = K_{\underline{\lambda}_{nj+l}}(x_{nj+l})$ 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. We will consider functions $f : U \subseteq \mathbb{R}^m \to \mathbb{R}^n$ and by an integrable function we will mean a Riemann (or Lebesgue) integrable function for which $\int |f(x)|dx$ is finite.

Theorem 9.13 (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*

$$|f_n(x)| \le g(x), \tag{9.44}$$

then

$$\lim_{n\to\infty} \int f_n(x)dx = \int f(x)dx. \tag{9.45}$$

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

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

Corollary 9.14. *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

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

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

Corollary 9.15. *Suppose $f(x, \lambda)$ is integrable with respect to x for any λ and continuously differentiable with respect to λ for any x. If there is a dominating function $g(x)$ such that*

$$|\frac{\partial f}{\partial \lambda}(x, \lambda)| \le g(x), \tag{9.46}$$

then the function

$$F(\lambda) = \int f(x, \lambda)dx \tag{9.47}$$

is continuously differentiable with derivative given by

$$\frac{\partial F}{\partial \lambda}(\lambda) = \int \frac{\partial f}{\partial \lambda}(x, \lambda)dx. \tag{9.48}$$

Proof. Again it suffices to consider one dimension. Since

$$f(x, \lambda + \varepsilon) - f(x, \lambda) = \varepsilon \int_0^1 \frac{\partial f}{\partial \lambda}(x, \lambda + \varepsilon t)dt$$

we have

$$\frac{F(\lambda+\varepsilon)-F(\lambda)}{\varepsilon} = \iint_0^1 \frac{\partial f}{\partial \lambda}(x,\lambda+\varepsilon t)dt\, dx.$$

Moreover, by $|\frac{\partial f}{\partial \lambda}(x,\lambda+\varepsilon t)| \le g(x)$ we have

$$\lim_{\varepsilon\to 0}\int_0^1 \frac{\partial f}{\partial \lambda}(x,\lambda+\varepsilon t)dt = \frac{\partial f}{\partial \lambda}(x,\lambda)$$

by the dominated convergence theorem. Applying dominated convergence again, note $|\int_0^1 \frac{\partial f}{\partial \lambda}(x,\lambda+\varepsilon t)dt| \le g(x)$, the claim follows. □

Now let us turn to integral equations. As in Section 2.2 we will equip the set of continuous functions $C(U,\mathbb{R}^n)$ (where $U \subseteq \mathbb{R}^m$) with the sup norm $\|f\| = \sup_{x\in U}|f(x)|$, which will turn $C(U,\mathbb{R}^n)$ into a Banach space.

Suppose V is a closed subset of $\mathbb{R}^n$ and consider the following (nonlinear) **Volterra integral equation**

$$K_\lambda(x)(t) = k(t,\lambda) + \int_0^t K(s,x(s),\lambda)ds, \tag{9.49}$$

where

$$k \in C(I\times\Lambda, V), \qquad K \in C(I\times V\times\Lambda,\mathbb{R}^n), \tag{9.50}$$

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

$$|K(t,x,\lambda) - K(t,y,\lambda)| \le L|x-y|, \qquad x,y \in V. \tag{9.51}$$

Theorem 9.16. *Let K_λ satisfy the requirements (9.50)–(9.51) from above and let $T_0 = \min(T,\frac{\delta}{M})$, where $\delta > 0$ is such that*

$$C_\delta = \{B_\delta(k(t,\lambda))\,|(t,\lambda)\in[T,T]\times\Lambda\} \subset V \tag{9.52}$$

and

$$M = \sup_{(t,x,\lambda)\in[-T,T]\times B_\delta(0)\times\Lambda} |K(t,k(t,\lambda)+x,\lambda)|. \tag{9.53}$$

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

$$|\overline{x}(t,\lambda) - k(t,\lambda)| \le \mathrm{e}^{LT_0}\sup_{\lambda\in\Lambda}\int_{-T_0}^{T_0}|K(s,k(s,\lambda),\lambda)|ds. \tag{9.54}$$

Moreover, if in addition all partial derivatives of order up to r with respect to λ 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 λ 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 V. Then existence and the bound follows as in the proof of Theorem 2.5. By the dominated convergence theorem $K_\lambda(x)$ is continuous with respect to λ for fixed $x(t)$. Hence by Theorem 9.12 the second term in

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

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

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

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

$$\tilde{K}_\lambda(x,y)(t) = \int_0^t \frac{\partial K}{\partial \lambda}(s,x(s),\lambda)ds + \int_0^t \frac{\partial K}{\partial x}(s,x(s),\lambda)y(s)ds. \tag{9.55}$$

This integral operator is linear with respect to y and by the mean value theorem and (9.51) we have

$$\|\frac{\partial K}{\partial x}(t,x,\lambda)\| \le L.$$

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 λ. 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 9.17), Theorem 9.12 implies $(x_n,y_n) \to (\overline{x}(\lambda), \overline{y}(\lambda))$. Moreover, since (x_n,y_n) is uniformly bounded with respect to λ, we conclude by Corollary 9.14 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 $k_\lambda, \frac{\partial K}{\partial \lambda}, \frac{\partial K}{\partial x} \in C^{r-1}$ we can conclude $y \in C^{r-1}$ and hence $x \in C^r$. □

Corollary 9.17. *If, in addition to the requirements from Theorem 9.16, $k \in C^r(I \times \Lambda, V)$ 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

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

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. □

Next we turn to the following **Hammerstein integral equation** which we encountered in Section 9.2,

$$K_\lambda(x)(t) = k(t,\lambda) + \int_0^\infty \kappa(s-t,\lambda) K(s,x(s),\lambda) ds, \tag{9.56}$$

where

$$k \in C([0,\infty)\times\Lambda,\mathbb{R}^n),\ \kappa \in C(\mathbb{R}\times\Lambda,\mathbb{R}^n),\ K \in C([0,\infty)\times V\times\Lambda,\mathbb{R}^n), \tag{9.57}$$

with $\Lambda \subset \mathbb{R}^n$ compact. Now we are going to show the analog of Theorem 9.16 for this equation, which we used in Section 9.2.

We assume that for every compact set $C \subseteq V$, k and K are uniformly continuous and bounded

$$|k(t,\lambda)| \le m, \quad |K(t,x,\lambda)| \le M, \quad (t,x,\lambda) \in [0,\infty)\times C\times\Lambda, \tag{9.58}$$

and that there is an integrable function $\alpha(s)$ such that

$$|\kappa(s+t,\lambda)| \le \alpha(s) \quad \text{for} \quad |t| \le \varepsilon. \tag{9.59}$$

In addition, suppose

$$|K(s,x,\lambda) - K(s,y,\lambda)| \le L|x-y|, \qquad x,y \in V, \tag{9.60}$$

where L is independent of λ, and that

$$L\int_{-\infty}^\infty |\kappa(s,\lambda)| ds \le \theta < 1. \tag{9.61}$$

Theorem 9.18. *Let K_λ satisfy the requirements (9.57)–(9.61) from above. Then the fixed point equation $K_\lambda(x) = x$ has a unique solution $\overline{x}(t,\lambda) \in C([0,\infty)\times\Lambda,V)$.*

Assume in addition that all partial derivatives of order up to r with respect to λ 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 λ of $\kappa(s,\lambda)$ there are dominating functions as in (9.59) and all partial derivatives of order up to r with respect to λ and x of $K(s,x,\lambda)$ are uniformly continuous and bounded when x is restricted to compacts as in (9.58). Then all partial derivatives of order up to r with respect to λ of $\overline{x}(t,\lambda)$ are continuous.

Proof. As in Theorem 9.16 it is no restriction to assume $k(t,\lambda) \equiv 0$. Choose

$$\delta = (1-\theta)^{-1}\|K_\lambda(0)\|,$$

then $\|x\| \le \delta$ implies

$$\begin{aligned}\|K_\lambda(x)\| &\le \int_0^\infty |\kappa(s-t,\lambda)|(|K(s,0,\lambda)| + |K(s,x(s),\lambda) - K(s,0,\lambda)|)ds\\ &\le \|K_\lambda(0)\| + \theta\|x\| \le \delta\end{aligned}$$

and hence K_λ maps $C([0,\infty), B_\delta(0))$ into itself. Moreover, by assumption K_λ is a contraction with contraction constant θ 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 λ,

$$\begin{aligned} &|K_\lambda(x)(t) - K_\eta(x)(t)| \le \\ &\qquad \int_0^\infty |\kappa(s-t,\lambda)|\, |K(s,x(s),\lambda) - K(s,x(s),\eta)| ds \\ &\qquad \int_0^\infty |\kappa(s-t,\lambda) - \kappa(s-t,\eta)|\, |K(s,x(s),\eta)| ds. \end{aligned}$$

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

$$\|K_\lambda(x)(t) - K_\eta(x)(t)\| \le \frac{\varepsilon\theta}{L} + M \int_{-\infty}^\infty |\kappa(s-t,\lambda) - \kappa(s-t,\eta)| ds.$$

Since the right-hand side can be made arbitrarily small by choosing $|\lambda - \eta|$ small (dominated convergence), the claim follows.

Now we can show that $\overline{x}$ is continuous. By our previous consideration, the first term in

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

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

$$\begin{aligned} &|\overline{x}(t,\eta) - \overline{x}(s,\eta)| \\ &\qquad \le \int_0^\infty |\kappa(r-t,\eta) - \kappa(r-s,\eta)|\, |K(r,\overline{x}(r,\eta),\eta)| dr \\ &\qquad \le M \int_0^\infty |\kappa(r-t,\eta) - \kappa(r-s,\eta)| dr. \end{aligned}$$

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

$$\begin{aligned} \tilde{K}_\lambda(x,y)(t) = &\int_0^\infty \kappa_\lambda(s-t,\lambda) K(s,x(s),\lambda) ds \\ &+ \int_0^\infty \kappa(s-t,\lambda) K_\lambda(s,x(s),\lambda) ds \\ &+ \int_0^\infty \kappa(s-t,\lambda) K_x(s,x(s),\lambda) y(s) ds, \end{aligned}$$

where the subscripts denote partial derivatives. The rest follows as in the proof of the Theorem 9.16. To show that $\tilde{K}_\lambda(x,y)$ depends continuously on x you need to use uniform continuity of K and its derivatives. □

Problem 9.16. *Suppose $K : C \subseteq X \to C$ is a contraction and*

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

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

Problem 9.17. *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 (9.55) is continuous with respect to $x \in C(I, \mathbb{R}^n)$. (Hint: Use the dominated convergence theorem.)

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, most results won't be needed until Chapter 12. 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 or 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 $\kappa = \log(1+k)$.

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. We have seen in Section 1.5, 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

$$L_\mu(x) = \mu x(1 - x) \tag{10.8}$$

we can write the solution as the 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 the behavior of solutions 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 L_μ.

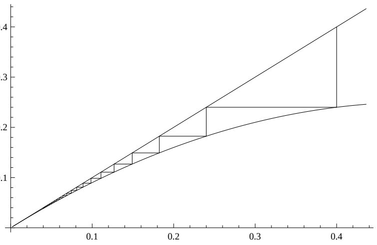
First of all, we will use the following *Mathematica* code

```
In[1]:= ShowWeb[f_, xstart_, nmax_] :=
          Module[{x, xmin, xmax, delta, 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]; delta = 0.1(xmax − xmin)
            graph = Plot[{f[x], x}, {x, xmin − delta, xmax + delta}];
            Show[graph, Graphics[Line[web]]]
          ];
```

to visualize `nmax` iterations of a function $f(x)$ starting at `xstart`. If μ is small, say $\mu = 1$,

```
In[2]:= ShowWeb[1#(1 − #)&, 0.4, 20]
```

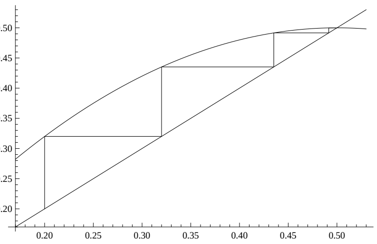
`Out[2]=`



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 μ 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]
```

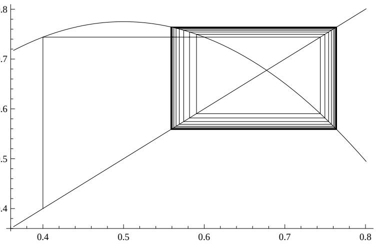
`Out[3]=`



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]
```

`Out[4]=`



Clearly this method of investigating the system gets quite cumbersome. We will return to this problem in Section 11.1.

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**

$$x_{n+m} = F(n, x_n, \dots, x_{n+m-1}). \tag{10.9}$$

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. The set of fixed points of f is denoted by $\mathrm{Fix}(f)$. 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 $1 \leq m < n$. The set of periodic points of f is denoted by $\mathrm{Per}(f)$.

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. Here a set $U \subseteq M$ is called (positively) **invariant**, if $f(U) \subseteq U$. 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 $x_n \in \gamma(x)$, $n < 0$, 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

$$\gamma_+(p) = \{p, f(p), \dots, f^{n-1}(p)\}. \tag{10.15}$$

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. A (forward) orbit of the form (10.15) will be called a **periodic orbit**.

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 $W^+(\gamma(p)) = \bigcup_{q\in\gamma(p)} W^+(q)$ is clearly positively invariant (it is even invariant $f(W^+(\gamma(p))) = W^+(\gamma(p))$ 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_k 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).

Furthermore, 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_μ 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 $f \in C^1(U,U)$, $U \subseteq \mathbb{R}^n$. 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 θ which is larger than the modulus of all eigenvalues (Problem 3.48). 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 by the mean value theorem (cf. Problem 2.5) we have $|f^n(x)-p| = |f^n(x)-f^n(p)| \le \theta|x-p|$ and the first claim follows.

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 $L_\mu(x)$, $x \in \mathbb{R}$, for $\mu = 1$. Show that $W^+(0) = [0,1]$.*

Problem 10.4. *Determine the stability of all fixed points of the logistic map $L_\mu(x)$, $x \in [0,1]$, via linearization for $0 \leq \mu \leq 4$.*

Problem 10.5. *Consider the logistic map L_μ 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.*

Problem 10.7. *Define the set of* **recurrent** *points* $\mathrm{Rec}(f) := \{x \in M \mid$ *for every neighborhood $U(x)$ there is some $n > 0$ with $f^n(x) \in U(x)$*$\}$ *and the set of* **nonwandering** *points* $\mathrm{Nwa}(f) := \{x \in M \mid$ *for every neighborhood $U(x)$ there are $n > 0$ and $y \in U(x)$ with $f^n(y) \in U(x)$*$\}$.

Show:

(i) $\mathrm{Per}(f) \subseteq \mathrm{Rec}(f) \subseteq \mathrm{Nwa}(f)$.

(ii) $\mathrm{Per}(f)$, $\mathrm{Rec}(f)$, *and* $\mathrm{Nwa}(f)$ *are (positively) invariant.*

(iii) $\mathrm{Rec}(f) = \{x \in M \mid$ *there is a sequence* n_k *with* $f^{n_k}(x) \to x\}$.

(iv) $\mathrm{Nwa}(f)$ *is closed.*

(See also Problem 11.9.)

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

$$\Pi(m, m_0) = \prod_{j=m_0}^{m-1} A(j), \quad m \ge m_0. \tag{10.22}$$

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

$$\Pi(m, m_0) = \left(\prod_{j=m}^{m_0-1} A(j) \right)^{-1}, \quad m < m_0 \tag{10.25}$$

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

$$x(m) = \Pi(m, m_0)x_0 - \sum_{j=m-1}^{m_0} \Pi(m, j)g(j), \quad m < m_0. \tag{10.29}$$

Problem 10.8. *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

$$f(x) = Ax + g(x), \tag{10.30}$$

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 9.9).

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*

$$\varphi \circ A = f \circ \varphi, \qquad A = df_0, \tag{10.31}$$

in a sufficiently small neighborhood of 0.

Proof. Let ϕ_δ 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 $g_\delta = (1-\phi_\delta)(f-A)$ satisfies the assumptions of Lemma 9.7 (show this) for δ sufficiently small. Since f and f_δ coincide for $|x| \le \delta$ the homeomorphism for f_δ is also the right one for f for x in the neighborhood $\varphi^{-1}(\{x| \, |x| \le \delta\})$. □

Let me emphasize that the homeomorphism φ 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 φ) 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,

$$M^\pm(p) = \{x \in M | \sup_{\pm m \in \mathbb{N}_0} \alpha^{\mp m} |f^m(x) - p| < \infty \text{ for some } \alpha \in (0,1)\}. \tag{10.32}$$

Both sets are obviously invariant under the flow 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 matrix 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 ε can be made arbitrarily small by choosing δ 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 matrix A. Then, there is a neighborhood $U(p)$ and functions $h^\pm \in C^k(E^\pm(A), E^\mp(A))$ such that*

$$M^\pm(p) \cap U(p) = \{p + a + h^\pm(a) | a \in E^\pm \cap U\}. \tag{10.40}$$

Both $h^\pm$ and their Jacobian matrices 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 α.

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 matrix. 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

Discrete dynamical systems in one dimension

11.1. Period doubling

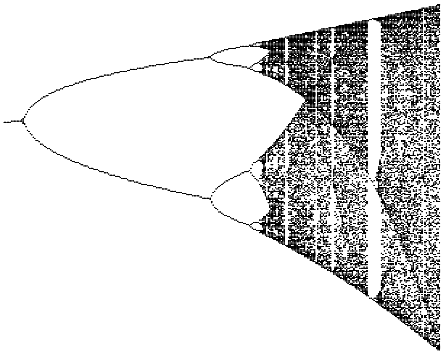
We now return to the logistic equation and the numerical investigation started in Section 10.1. Let us try to get a more complete picture by iterating one given initial condition for different values of μ. Since we are only interested in the asymptotic behavior we first iterate 200 times and then plot 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]
```

Out[2]=



So we see that at certain values of the parameter μ 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 values of μ 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

$$g(x) = f(x) - x. \tag{11.1}$$

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

$$\begin{aligned} L_\mu(x) &= x = \mu x(1-x), \\ L_\mu'(x) &= 1 = \mu(1-2x), \end{aligned} \tag{11.2}$$

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

$$\frac{L_\mu(x) - p}{x - p} = 1 - \mu x. \tag{11.3}$$

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

$$\frac{L_\mu^2(x) - p}{x - p} = (1 - \mu x)(1 - \mu L_\mu(x)). \tag{11.4}$$

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

$$L_\mu^2(x) = x \tag{11.5}$$

shows that, in addition to the fixed points, there is a periodic orbit

$$p_\pm = \frac{1 + \mu \pm \sqrt{(\mu+1)(\mu-3)}}{2\mu} \tag{11.6}$$

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 μ. Suppose that at μ_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 $f([p_-(\mu), p_+(\mu)]) \supseteq [p_-(\mu), p_+(\mu)]$ 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.

Problem 11.1. *Show that for $\mu = 2$ we have*

$$x_n = L_2^n(x_0) = \frac{1}{2}\left(1 - (1-2x_0)^{2^n}\right).$$

11.2. 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 11.1. *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 (Problem 11.2):

(i) If I, J are two compact intervals satisfying $f(J) \supseteq I$, then there is a subinterval J_0 of J such that $f(J_0) = I$.

(ii) 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 $I_0 = [a, b]$, $I_1 = [b, c]$ and observe $f(I_0) \supseteq I_1$, $f(I_1) \supseteq I_0 \cup I_1$.

Set $J_0 = I_1$ and recall $f(J_0) = f(I_1) \supseteq I_1 = J_0$. By (i) we can find a subinterval $J_1 \subseteq J_0$ 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, $k = 0, \dots, n$, such that $f(J_k) = J_{k-1}$. In particular, we have $f^n(J_n) = J_0 \supseteq J_n$ and thus f^n has a fixed point in J_n by (ii). The only problem is, is the prime period of this point n? 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 $f^{n-1}(J_{n-2}) = f(f^{n-2}(J_{n-2})) = f(I_1) \supseteq I_0$. 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, note that x really leaves I_1 in the $(n-1)$-th step since $f^{n-1}(x) \in I_0 \cap I_1 = \{b\}$ contradicts $a = f^{n+1}(x) = f(x) \in I_1$. Consequently 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_μ? 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

$$\{\frac{1}{2}(1+c), 1-c^2, 4c^2(1-c^2)\}, \qquad c = \cos(\frac{\pi}{9}), \tag{11.7}$$

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 ($n = 0$) in decreasing order. That is, denoting the **Sarkovskii ordering** by $\succ$ we have

$$3 \succ 5 \succ \cdots \succ 2\cdot 3 \succ 2 \cdot 5 \succ \cdots \succ 2^m(2n+1) \succ \cdots \succ 2^2 \succ 2 \succ 1. \tag{11.8}$$

With this notation the following claim holds.

Theorem 11.2 (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 11.1 but quite tedious. Hence we omit it here. It can be found (e.g.) in [**33**].

Problem 11.2. *Show items (i) and (ii) from the proof of Lemma 11.1.*

11.3. 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

$$M = (0, \infty), \quad f(x) = (1 + \mu)x, \quad \mu > 0, \tag{11.9}$$

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 forward orbit (Problem 11.3).

A system having both properties is called chaotic in the book by Robinson [**33**]. 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.

$$\begin{tikzcd}
M_1 \arrow[r, "f_1"] \arrow[d, leftrightarrow, "\varphi"'] & M_1 \arrow[d, leftrightarrow, "\varphi"] \\
M_2 \arrow[r, "f_2"] & M_2
\end{tikzcd} \tag{11.10}$$

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

$$M = (0, \infty), \quad f(x) = (1 + \mu)^{-1}x, \quad \mu > 0, \tag{11.11}$$

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 points are dense. If M is finite and the system is 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 11.3. *Suppose $f : M \to M$ is chaotic. Then it exhibits sensitive dependence on initial conditions.*

Proof. First observe that there is a number δ such that for all $x \in M$ there exists a periodic point $q \in M$ whose orbit is of distance at least 4δ from x. In fact, since M is not finite we can pick two periodic points q_1 and q_2 with disjoint orbits. Let 8δ 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 4δ.

Fix $x \in M$ and $\varepsilon > 0$ and let q be a periodic orbit with distance at least 4δ. 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

$$\begin{aligned} d(f^{k+j}(p), f^{k+j}(y)) &\ge d(x, f^j(q)) - d(f^j(q), f^{k+j}(y)) - d(p, x) \\ &> 4\delta - \delta - \delta = 2\delta. \end{aligned}$$

Thus either $d(f^{k+j}(y), f^{k+j}(x)) > \delta$ 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**

$$M = [0,1], \quad T_\mu(x) = \frac{\mu}{2}(1 - |2x - 1|) \tag{11.12}$$

using topological equivalence. The tent map T_2 is equivalent to the logistic map L_4 by virtue of the homeomorphism $\varphi(x) = \sin(\frac{\pi x}{2})^2$ (Problem 11.4). 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

$$T_2(x) = \begin{cases} 2x, & 0 \le x \le \frac{1}{2}, \\ 2 - 2x, & \frac{1}{2} \le x \le 1, \end{cases} \tag{11.13}$$

it is not hard to verify that

$$T_2^n(x) = \left\{ \begin{array}{cc} 2^n x - 2j, & \frac{2j}{2^n} \le x \le \frac{2j+1}{2^n} \\ 2(j+1) - 2^n x, & \frac{2j+1}{2^n} \le x \le \frac{2j+2}{2^n} \end{array} \right\}_{0 \le j \le 2^{n-1}-1}. \tag{11.14}$$

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 11.3. *Show that a closed invariant set which has a dense forward orbit is topologically transitive.*

Problem 11.4. *Show that T_2 and L_4 are topologically equivalent via the map $\varphi(x) = \sin(\frac{\pi x}{2})^2$ (i.e., show that $\varphi : [0,1] \to [0,1]$ is a homeomorphism and that $\varphi \circ T_2 = L_4 \circ \varphi$).*

Problem 11.5. *Find a topological conjugation $\varphi(x) = m\,x + d$ which maps $f(x) = \alpha x^2 + \beta x + \gamma$ to $g(x) = x^2 + c$. Find m, d, and c in terms of α, β, and γ.*

11.4. Cantor sets and the tent map

Now let us further investigate the tent map T_μ for $\mu > 2$. Unfortunately, in this case T_μ does no longer map $[0,1]$ into itself. Hence we must consider it as a map on $\mathbb{R}$,

$$M = \mathbb{R}, \quad T_\mu(x) = \frac{\mu}{2}(1 - |2x - 1|). \tag{11.15}$$

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 Λ_0 under one iteration are given by $(\frac{1}{\mu}\Lambda_0) \cup (1 - \frac{1}{\mu}\Lambda_0)$. Denote this set by

$$\Lambda_1 = [0, \frac{1}{\mu}] \cup [1 - \frac{1}{\mu}, 1]. \tag{11.16}$$

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 Λ_1 under one iteration are given by $(\frac{1}{\mu}\Lambda_1) \cup (1 - \frac{1}{\mu}\Lambda_1)$. Hence the corresponding set

$$\Lambda_2 = [0, \frac{1}{\mu^2}] \cup [\frac{1}{\mu} - \frac{1}{\mu^2}, \frac{1}{\mu}] \cup [1 - \frac{1}{\mu}, 1 - \frac{1}{\mu} + \frac{1}{\mu^2}] \cup [1 - \frac{1}{\mu^2}, 1] \qquad (11.17)$$

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 Λ_n stay in $[0,1]$ for at least n iterations. Moreover, each set Λ_n consists of 2^n closed subintervals of length μ^{-n}.

Now if we want to stay in $[0,1]$ we have to take the intersection of all these sets, that is, we define

$$\Lambda = \bigcap_{n\in\mathbb{N}} \Lambda_n \subset [0,1]. \qquad (11.18)$$

Since the sets Λ_n form a nesting sequence of compact sets, the set Λ is also compact and nonempty. By construction the set Λ is invariant since we have

$$T_\mu(\Lambda) = \Lambda \qquad (11.19)$$

and all points in the open set $\mathbb{R}\backslash\Lambda$ converge to $-\infty$.

Moreover, since the endpoints of the subintervals of Λ_n are just given by $f^{-n}(\{0,1\})$, we see that these points are in Λ. Now the set Λ 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. Secondly, 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 Λ_n for every n. Since the endpoints of these subintervals are in Λ (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 11.4. *The set Λ 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

$$x = \sum_{n\in\mathbb{N}} \frac{x_n}{3^n}, \qquad x_n \in \{0,1,2\}. \qquad (11.20)$$

Recall that this expansion is not unique since we have, for example, $\frac{1}{3} = 0.1 = 0.0\overline{2}$ or $\frac{2}{3} = 0.2\cdots = 0.1\overline{2}\cdots$. Here the $\overline{x}$ implies that the corresponding digit repeats infinitely many times. It will be convenient for us to exclude the expansions which end in 1 or $1\overline{2}$. Then we have $\Lambda_n = \{x | x_j \neq 1,\ 1 \leq j \leq n\}$ (Problem 11.7) and hence

$$\Lambda = \{x | x_j \neq 1, j \in \mathbb{N}\}. \tag{11.21}$$

Moreover, the action of T_3 can also be transparently described using this notation

$$\begin{cases} x_1 = 0 & \Rightarrow \quad T_3(x) = \sum_{n\in\mathbb{N}} \frac{x_{n+1}}{3^n} \\ x_1 = 1 & \Rightarrow \quad T_3(x) \notin [0,1] \\ x_1 = 2 & \Rightarrow \quad T_3(x) = \sum_{n\in\mathbb{N}} \frac{x'_{n+1}}{3^n} \end{cases}, \tag{11.22}$$

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, it is not possible to tell if two points x, y are close by looking at the first n digits 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**

$$\begin{array}{rccl} \varphi: & \Lambda & \to & \Sigma_2 \\ & x & \mapsto & x_n = j \text{ if } T_\mu^n(x) \in I_j \end{array}. \tag{11.23}$$

Then φ is well defined and T_μ 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 (Λ, T_μ) and (Σ_2, σ). But before we can show this, we need some further definitions first.

First of all we need to make sure that (Σ_2, σ) is a dynamical system. Hence we need a metric on Σ_2. We will take the following one

$$d(x,y) = \sum_{n\in\mathbb{N}_0} \frac{|x_n - y_n|}{2^n} \tag{11.24}$$

(prove that this is indeed a metric). Moreover, we need to make sure that σ is continuous. But since

$$d(\sigma(x), \sigma(y)) \leq 2\, d(x,y) \tag{11.25}$$

it is immediate that σ is even uniformly continuous.

So it remains to show that φ is a homeomorphism.

We start by returning to the construction of Λ_n. If we set $I = [0,1]$ we have seen that Λ_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 Λ_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 $\varphi(x)_j = s_j$ for $0 \le j \le n$. Now pick a sequence $s \in \Sigma_2$ and consider the intersection of nesting intervals

$$I_s = \bigcap_{n \in \mathbb{N}_0} I_{s_0,\cdots,s_n}. \tag{11.26}$$

By the finite intersection property of compact sets it is a nonempty interval, hence φ 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, φ is injective.

If x and y are close so are $T_\mu(x)^n$ and $T_\mu(y)^n$ by continuity of T_μ. Hence, for y sufficiently close to x the first n iterates will stay sufficiently close such that $\varphi(x)_j = \varphi(y)_j$ for $0 \le j \le n$. But this implies that $\varphi(x)$ and $\varphi(y)$ are close and hence φ 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 φ^{-1} is continuous.

In summary,

Theorem 11.5. *The two dynamical systems* (Λ, T_μ), $\mu > 2$, *and* (Σ_2, σ) *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 σ on Σ_2. In fact, we will show that (Σ_2, σ), and hence (Λ, T_μ), $\mu > 2$, is chaotic in the next section.

Problem 11.6. *Show that two different ternary expansions define the same number,* $\sum_{n\in\mathbb{N}} x_n 3^{-n} = \sum_{n\in\mathbb{N}} y_n 3^{-n}$, *if and only if there is some* $n_0 \in \mathbb{N}$ *such that* $x_n = y_n$ *for* $n < n_0$, $x_n = y_n \pm 1$ *for* $n = n_0$, *and* $x_n = y_n \mp 2$ *for* $n > n_0$. *Show that every* $x \in [0,1]$ *has a unique expansions if the expansions which end in* 1 *or* $1\overline{2}$ *are excluded.*

Problem 11.7. *Show that for* $\mu = 3$ *we have* $\Lambda_n = \{x | x_j \neq 1,\ 1 \le j \le n\}$, *where* x_j *are the digits in the ternary expansion as in the previous problem.*

11.5. 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**

$$\Sigma_N = \{0, 1, \dots, N-1\}^{\mathbb{N}_0} \tag{11.27}$$

to be the set of sequences taking only the values $0, \dots, N-1$. Note that Σ_N is not countable (why?).

Defining

$$d(x, y) = \sum_{n \in \mathbb{N}_0} \frac{|x_n - y_n|}{N^n}, \tag{11.28}$$

Σ_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 11.6. *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

$$d(x,y) = \sum_{j > n} \frac{|x_j - y_j|}{N^j} \le \frac{1}{N^{n+1}} \sum_{j \ge 0} \frac{N-1}{N^j} = \frac{1}{N^n}.$$

Conversely, if $x_j \ne y_j$ for at least one $j \le n$, we have

$$d(x,y) = \sum_{k \in \mathbb{N}} \frac{|x_k - y_k|}{N^k} \ge \frac{1}{N^j} \ge \frac{1}{N^n}.$$

□

We first show that Σ_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 (Problem 11.8).

Lemma 11.7. *The set Σ_N is a Cantor set.*

Proof. We first prove that Σ_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 Σ_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 Σ_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 $V = \{x | x_{j_0} \neq c\}$. Now let $x, y \in \Sigma_N$, if $x \neq y$ there is a j_0 such that $x_{j_0} \neq y_{j_0}$. Now take $c = x_{j_0}$ then U and V from above are disjoint open sets whose union is Σ_N and which contain x and y respectively. □

On Σ_N we have the **shift map**

$$\begin{array}{rlcl} \sigma: & \Sigma_N & \to & \Sigma_N \\ & (x_0, x_1, \dots) & \mapsto & (x_1, x_2, \dots) \end{array}, \qquad (11.29)$$

which is uniformly continuous since we have

$$d(\sigma(x), \sigma(y)) \leq N d(x, y). \qquad (11.30)$$

Furthermore, it is chaotic as we will prove now. Observe that a point x is periodic for σ if and only if it is a periodic sequence.

Lemma 11.8. *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 11.9. *The shift map has a dense forward orbit.*

Proof. Construct a forward 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 $d(y, \sigma^k(x)) \leq N^{-n}$. Hence the orbit is dense. □

Combining the two lemmas we see that (Σ_N, σ) is chaotic. I leave it as an exercise to show that σ 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 Σ_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 Σ_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

$$\Sigma_N^A = \{x \in \Sigma_N | A_{x_n, x_{n+1}} = 1 \text{ for all } n \in \mathbb{N}_0\}. \tag{11.31}$$

Clearly σ maps Σ_N^A into itself and the dynamical system (Σ_N^A, σ) is called a **subshift of finite type**. It is not hard to see that Σ_N^A is a closed subset of Σ_N and thus compact. Moreover, σ is continuous on Σ_N^A as the restriction of a continuous map. We will denote this restriction by σ_A.

Now let us return to our example. Here we have

$$A = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}. \tag{11.32}$$

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 except $x = (1, 1, 1, \dots)$ are eventually fixed and converge to the fixed point $x = (0, 0, 0, \dots)$. So the system is definitely not chaotic. The same is true for all other possibilities except

$$A = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \tag{11.33}$$

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 **transitive** if there is an integer $l \in \mathbb{N}$ such that $(A^l)_{j,k} \neq 0$ for all $0 \leq j, k \leq N - 1$.

Let A be a transition matrix. We will call $(x_1, \dots, x_k)$ an admissible block of length k if $A_{x_j, x_{j+1}} = 1$ for $1 \leq j \leq k - 1$. The following lemma explains the importance of A^l.

Lemma 11.10. *The (j, k) entry of A^l is equal to the number of admissible blocks $(x_0, \dots, x_l)$ of length $l + 1$ with $x_0 = j$ and $x_l = k$.*

In particular, the number of periodic orbits of length l is equal to $\mathrm{tr}(A^l)$.

Proof. Just observe that the (j, k) entry of A^l is given by

$$(A^l)_{j,k} = \sum_{x_1, \dots, x_{l-1}} A_{j,x_1} A_{x_1,x_2} \cdots A_{x_{l-2},x_{l-1}} A_{x_{l-1},k}$$

and that the above products are 1 if and only if the block $(j, x_1, \dots, x_{l-1}, k)$ is admissible. □

In particular, for A transitive we obtain the following simple consequence which is the key ingredient for our proof that transitive subshifts are chaotic.

Corollary 11.11. *If A is transitive 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 $0 \le j, k \le N - 1$.*

This lemma ensures that, if A is transitive, 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 a first application we prove

Lemma 11.12. *Suppose A is transitive. Then Σ_N^A is a Cantor set.*

Proof. As noted earlier, Σ_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 an admissible block of length $l-1$ from Corollary 11.11 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 an 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 (Σ_N^A, σ) is chaotic.

Lemma 11.13. *Suppose A is transitive. Then the shift map on Σ_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 Corollary 11.11 to the end. Now take this entire block and repeat it periodically. The rest is straightforward. □

Lemma 11.14. *Suppose A is transitive. Then the shift map on Σ_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 Corollary 11.11. □

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

$$\Sigma_N = \{0, 1, \dots, N-1\}^{\mathbb{Z}} \tag{11.34}$$

to be the set of doubly infinite sequences taking only the values $0, \dots, N-1$. Defining

$$d(x,y) = \frac{1}{2} \sum_{n \in \mathbb{N}_0} \frac{|x_n - y_n| + |x_{-n} - y_{-n}|}{N^n}, \tag{11.35}$$

Σ_N becomes a metric space. Again we have

Lemma 11.15. *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 σ is defined as before. However, note that σ is invertible in this case. All other results hold with no further modifications. The details are left to the reader.

Problem 11.8. *Show that the definition of a totally disconnected set given in this section agrees with the one given in the previous section for subsets of $\mathbb{R}$. (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).*

Problem 11.9. *Show that for the shift on two symbols (cf. Problem 10.7): All points are nonwandering, $\mathrm{Nwa}(\sigma) = \Sigma_2$. There are recurrent points which are not periodic and there are nonwandering points which are not recurrent.*

Problem 11.10. *The (Artin-Mazur) zeta function of a discrete dynamical system $f : M \to M$ is defined to be*

$$\zeta_f(z) = \exp\left(\sum_{n=1}^{\infty} \frac{z^n}{n} |\mathrm{Fix}(f^n)| \right),$$

where $|\mathrm{Fix}(f^n)|$ is the cardinality of the set of fixed points of f^n (provided this number is finite for every n). Equivalently, $|\mathrm{Fix}(f^n)|$ is the number of periodic orbits of period n.

Show that

$$\zeta_{\sigma_A}(z) = \frac{1}{\det(\mathbb{I} - zA)}, \quad |z| < \|A\|.$$

(Hint: (3.23)*.)*

11.6. Strange attractors/repellors and fractal sets

A compact invariant set Λ, $f(\Lambda) = \Lambda$, is called **attracting** if there is a neighborhood U of Λ such that $d(f^n(x), \Lambda) \to 0$ as $n \to \infty$ for all $x \in U$. A compact invariant set Λ, $f(\Lambda) = \Lambda$, is called **repelling** if there is a neighborhood U of Λ 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 11.16. *Suppose $f : I \to I$ is continuously differentiable and Λ is a compact invariant set. If there is an $n_0 \in \mathbb{N}$ such that $|d(f^{n_0})_x| < 1$ for all $x \in \Lambda$, then Λ is attracting. Similarly, if there is an $n_0 \in \mathbb{N}$ such that $|d(f^{n_0})_x| > 1$ for all $x \in \Lambda$, then Λ is repelling.*

Proof. We only prove the first claim, the second is similar. Choose α such that $\max_{x\in\Lambda} |d(f^{n_0})_x| < \alpha < 1$. For every y in Λ there is a (nonempty) open interval I_y containing y such that $|d(f^{n_0})_x| \leq \alpha$ 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, $d(f^{n_0}(x), \Lambda) \leq |f^{n_0}(x) - f^{n_0}(y)| \leq \alpha|x - y| = \alpha d(x, \Lambda)$. Hence $d(f^{n_0 n}(x), \Lambda) \to 0$ and by continuity of f and invariance of Λ we also have $d(f^{n_0 n+j}(x), \Lambda) \to 0$ for $0 \leq j \leq n_0$. Thus the claim is proven. □

Repelling, attracting sets as above are called **hyperbolic repelling, attracting sets**, respectively.

An attractor, repellor Λ is called **strange** if the dynamical system (Λ, f) is chaotic and if Λ 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 [**35**]).

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 δ**-cover** if it is countable and if $d(V_j) \leq \delta$ for all j.

For U a subset of $\mathbb{R}^n$ and $\alpha \geq 0$, $\delta > 0$ we define

$$h_\delta^\alpha(U) = \inf \Big\{ \sum_j d(V_j)^\alpha \Big| \{V_j\} \text{ is a } \delta\text{-cover of } U \Big\} \in [0, \infty]. \tag{11.36}$$

As δ decreases the number of admissible covers decreases and hence $h_\delta^\alpha(U)$ is increasing as a function of δ. Thus the limit

$$h^\alpha(U) = \lim_{\delta \downarrow 0} h_\delta^\alpha(U) = \sup_{\delta > 0} h_\delta^\alpha(U) \tag{11.37}$$

exists. Moreover, it is not hard to show that $h^\alpha(U) \leq h^\alpha(V)$ if $U \subseteq V$ and that for countable unions we have

$$h^\alpha(\bigcup_j U_j) \leq \sum_j h^\alpha(U_j). \tag{11.38}$$

Hence h^α is an **outer measure** and the resulting **measure** on the Borel σ-algebra is called the α dimensional **Hausdorff measure**. As any measure it satisfies

$$\begin{aligned} h^\alpha(\emptyset) &= 0, \\ h^\alpha(\bigcup_j U_j) &= \sum_j h^\alpha(U_j), \end{aligned} \tag{11.39}$$

for any countable union of disjoint sets U_j.

For example, consider the case $\alpha = 0$. Suppose $U = \{x, y\}$ consists of two points. Then $h_\delta^0(U) = 1$ for $\delta \geq |x - y|$ and $h_\delta^0(U) = 2$ for $\delta < |x - y|$. In particular, $h^0(U) = 2$. Similarly, it is not hard to see that $h^0(U)$ is just the number of points in U. On the other extreme, it can be shown that $h^n(U) = c_n/2^n|U|$, where $|U|$ denotes the Lebesgue measure of U and $c_n = \pi^{n/2}/\Gamma(n/2 + 1)$ is the volume of the unit ball in $\mathbb{R}^n$.

Using the fact that for $\lambda > 0$ the map $\lambda : x \mapsto \lambda x$ gives rise to a bijection between δ-covers and (δ/λ)-covers, we easily obtain the following scaling property of Hausdorff measures.

Lemma 11.17. *Let $\lambda > 0$ and U be a Borel set of $\mathbb{R}^n$. Then*

$$h^\alpha(\lambda U) = \lambda^\alpha h^\alpha(U). \tag{11.40}$$

Moreover, Hausdorff measures also behave nicely under uniformly Hölder continuous maps.

Lemma 11.18. *Suppose $f : U \to \mathbb{R}^n$ is uniformly Hölder continuous with exponent $\gamma > 0$, that is,*

$$|f(x) - f(y)| \leq c|x - y|^\gamma \quad \text{for all } x, y \in U, \tag{11.41}$$

then

$$h^\alpha(f(U)) \leq c^\alpha h^{\alpha\gamma}(U). \tag{11.42}$$

Proof. A simple consequence of the fact that for every δ-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^α_δ is non increasing with respect to α for $\delta < 1$ and hence the same is true for h^α. 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

$$h^\beta_\delta(U) \le \delta^{\beta-\alpha}\, h^\alpha_\delta(U) \le \delta^{\beta-\alpha}\, h^\alpha(U). \tag{11.43}$$

Thus if $h^\alpha(U)$ is finite, then $h^\beta(U) = 0$ for every $\beta > \alpha$. Hence there must be one value of α where the Hausdorff measure of a set jumps from ∞ to 0. This value is called the **Hausdorff dimension**

$$\dim_H(U) = \inf\{\alpha | h^\alpha(U) = 0\} = \sup\{\alpha | h^\alpha(U) = \infty\}. \tag{11.44}$$

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 $\dim_H(U) \le n$ (Problem 11.12).

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 11.18 it is also straightforward to show

Lemma 11.19. *Suppose $f : U \to \mathbb{R}^n$ is uniformly Hölder continuous with exponent $\gamma > 0$, that is,*

$$|f(x) - f(y)| \le c|x - y|^\gamma \quad \textit{for all } x, y \in U, \tag{11.45}$$

then

$$\dim_H(f(U)) \le \frac{1}{\gamma} \dim_H(U). \tag{11.46}$$

Similarly, if f is bi-Lipschitz, that is,

$$a|x - y| \le |f(x) - f(y)| \le b|x - y| \quad \textit{for all } x, y \in U, \tag{11.47}$$

then

$$\dim_H(f(U)) = \dim_H(U). \tag{11.48}$$

We end this section by computing the Hausdorff dimension of the repellor Λ of the tent map.

Theorem 11.20. *The Hausdorff dimension of the repellor Λ of the tent map T_μ is*

$$\dim_H(\Lambda) = \frac{\log(2)}{\log(\mu)}, \qquad \mu \ge 2. \tag{11.49}$$

In particular, it is a strange repellor.

Proof. Let $\delta = \mu^{-n}$. Using the δ-cover $I_{s_0,\dots,s_{n-1}}$ we see $h^\alpha_\delta(\Lambda) \le (\frac{2}{\mu^\alpha})^n$. Hence for $\alpha = d = \log(2)/\log(\mu)$ we have $h^d_\delta(\Lambda) \le 1$ implying $\dim_H(\Lambda) \le d$.

The reverse inequality is a little harder. Let $\{V_j\}$ be a cover. We suppose $\mu > 2$ (since for $\mu = 2$ we just have $\Lambda = [0,1]$) and $\delta < 1 - 2\mu^{-1}$. It is clearly no restriction to assume that all V_j are open intervals. Moreover, finitely many of these sets cover Λ by compactness. Drop all others and fix j. Furthermore, increase each interval V_j by at most ε

For V_j there is a k such that

$$\frac{1-2\mu^{-1}}{\mu^k} \le |V_j| < \frac{1-2\mu^{-1}}{\mu^{k-1}}.$$

Since the distance of two intervals in Λ_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 $2^{n-k} = 2^n(\mu^{-k})^d \le 2^n(1-2\mu^{-1})^{-d}|V_j|^d$ intervals of Λ_n.

Now choose n larger than all k (for all V_j). Since $\{V_j\}$ covers Λ, we must intersect all 2^n intervals in Λ_n. So we end up with

$$2^n \le \sum_j \frac{2^n}{(1-2\mu^{-1})^d}|V_j|^d,$$

which together with our first estimate yields

$$(1-\frac{2}{\mu})^d \le h^d(\Lambda) \le 1.$$

□

Observe that this result can also formally be derived from the scaling property of the Hausdorff measure by solving the identity

$$\begin{aligned} h^\alpha(\Lambda) &= h^\alpha(\Lambda\cap I_0) + h^\alpha(\Lambda\cap I_1) = 2\,h^\alpha(\Lambda\cap I_0)\\ &= \frac{2}{\mu^\alpha}h^\alpha(T_\mu(\Lambda\cap I_0)) = \frac{2}{\mu^\alpha}h^\alpha(\Lambda) \end{aligned} \tag{11.50}$$

for α. However, this is only possible if we already know that $0 < h^\alpha(\Lambda) < \infty$ for some α.

Problem 11.11. *Let $C = [0,1]\times\{0\} \subseteq \mathbb{R}^2$. Show that $h^1(C) = 1$.*

Problem 11.12. *Show that $\dim_H(U) \le n$ for every $U \subseteq \mathbb{R}^n$. (Hint: It suffices to take for U the unit cube. Now split U into k^n cubes of length $1/k$.)*

11.7. 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

$$\Lambda_1 = I_0 \cup I_1 = [0, G_\mu(1)] \cup [1 - G_\mu(1), 1], \tag{11.51}$$

where

$$G_\mu(x) = \frac{1}{2} - \sqrt{\frac{1}{4} - \frac{x}{\mu}}, \quad L_\mu(G_\mu(x)) = x, \quad 0 \le x \le 1. \tag{11.52}$$

To make our life a little easier we will make the additional assumption that

$$L_\mu'(x) = \mu(1-2x) \ge \alpha > 1 \quad \text{for} \quad x \in I_0. \tag{11.53}$$

Since we have

$$\sqrt{\mu(\mu-4)} = L_\mu'(G_\mu(1)) \le |L_\mu'(x)| \le L_\mu'(0) = \mu, \quad x \in I_0 \cup I_1, \tag{11.54}$$

this implies $\mu > 2 + \sqrt{5} = 4.236$. The general case $\mu > 4$ can be found in the book by Robinson [**33**].

Now proceeding as in the case of the tent map, we see that there is a sequence of nesting sets Λ_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_μ is not (piecewise) linear, we do not know the length of the interval $I_{s_0,\cdots,s_n}$. However, by our assumption (11.53), we know $G_\mu'(x) \le \alpha^{-1}$ and thus $|I_{s_0,\cdots,s_n}| \le G_\mu(1)\alpha^{-n}$. But this is all we have used for the tent map and hence the same proof shows

Theorem 11.21. *Suppose $\mu > 2+\sqrt{5}$. Then the logistic map L_μ leaves the set*

$$\Lambda = \bigcap_{n\in\mathbb{N}} \Lambda_n \subset [0,1] \tag{11.55}$$

invariant. All points $x \in \mathbb{R}\backslash\Lambda$ satisfy $\lim_{n\to\infty} L_\mu^n(x) = -\infty$. The set Λ is a Cantor set and the dynamical system (Λ, L_μ) is topologically equivalent to the shift on two symbols (Σ_2, σ) by virtue of the itinerary map

$$\begin{array}{rccl} \varphi: & \Lambda & \to & \Sigma_2 \\ & x & \mapsto & x_n = j \text{ if } L_\mu^n(x) \in I_j \end{array}. \tag{11.56}$$

In particular, (Λ, L_μ) is chaotic.

Clearly we also want to know whether the repellor Λ of the logistic map is strange.

Theorem 11.22. *The Hausdorff dimension of the repellor Λ of the logistic map satisfies*

$$d(\mu) \le \dim_H(\Lambda) \le \begin{cases} 1, & \mu \le 2+\sqrt{8}, \\ d(\sqrt{\mu(4-\mu)}), & \mu > 2+\sqrt{8}, \end{cases} \qquad d(x) = \frac{\log(2)}{\log(x)}. \tag{11.57}$$

In particular, it is strange if $\mu > 2+\sqrt{8} = 4.828$.

Proof. The proof is analogous to the one of Theorem 11.20. The only difference is that we have to use different estimates for L_μ' from above and below,

$$\sqrt{\mu(4-\mu)} = \alpha \le |L_\mu'(x)| \le \beta = \mu, \quad x \in I_0 \cup I_1.$$

Using the δ-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 Λ_k is at least $\frac{b}{\beta^{k-1}}$, where $b = d(I_0, I_1) = 1 - 2G_\mu(1)$, we obtain

$$b^{d(\beta)} \le h^{d(\beta)}(\Lambda)$$

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 11.23. *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*

$$\Lambda = \{x \in I_0 \cup I_1 | f^n(x) \in I_0 \cup I_1 \text{ for all } n \in \mathbb{N}\} \tag{11.58}$$

and define the itinerary map as

$$\begin{array}{rccl} \varphi: & \Lambda & \to & \Sigma_2 \\ & x & \mapsto & x_n = j \text{ if } f^n(x) \in I_j \end{array}. \tag{11.59}$$

Then the set Λ is a Cantor set and the dynamical system (Λ, f) is topologically equivalent to the shift on two symbols (Σ_2, σ). The Hausdorff dimension of Λ satisfies

$$d(\beta) \le \dim_H(\Lambda) \le d(\alpha), \qquad d(x) = \frac{\log(2)}{\log(x)}, \tag{11.60}$$

and it is strange if $\alpha > 2$.

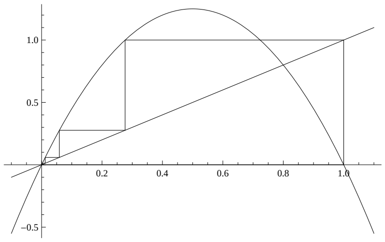
Proof. By assumption, the restricted maps $f : I_0 \to f(I_0)$ and $f : I_1 \to f(I_1)$ are invertible. Denote by $g_0 : f(I_0) \to I_0$ and $g_1 : f(I_1) \to I_1$ the respective inverses. Now proceeding as usual, we see that there is a sequence of nesting sets Λ_n consisting of 2^n subintervals $I_{s_0,\cdots,s_{n-1}}$, $s_j \in \{0, 1\}$, 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 $|I_{s_0,\cdots,s_n}| \leq \alpha^{-n}|I_{s_0}|$ 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 11.23 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 $q = G_\mu(1)$.

$$\texttt{In[3]:=}\ \mu = 5;$$
$$\mathtt{x_0} = \mathtt{Nest}\Big[\Big(\frac{1}{2} - \sqrt{\frac{1}{4} - \frac{\#}{\mu}}\Big)\&, 1., 5\Big];$$
$$\mathtt{ShowWeb}[\mu\#(1-\#)\&, \mathtt{x_0}, 6]$$

Out[3]=



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 ($L'_\mu(0) = \mu > 1$) 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 since f is expanding on W, we have $W \subseteq f(W)$ and 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 11.24. *Suppose $f \in C^1(I, I)$ has a repelling hyperbolic fixed point p and a corresponding nondegenerate homoclinic point q.*

For every sufficiently small neighborhood U of p there is an $n \in \mathbb{N}$ and an f^n invariant Cantor set $\Lambda \subset U$ (i.e., $f^n(\Lambda) = \Lambda$) such that (Λ, f^n) is topologically equivalent to the shift on two symbols (Σ_2, σ).

Proof. We will need to construct two disjoint intervals $I_j \subset U \cap W$, $j = 0, 1$, as in Theorem 11.23 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 $n = m + l$. Next, choose $\tilde{I}_1 \subseteq I_1$ such that $g^m(I_1) \subseteq f^l(\tilde{I}_1)$ but $f^l(\tilde{I}_1) \subseteq g^m(W)$. Then we have $g^m(\tilde{I}_1) \subseteq g^m(I_1) \subseteq f^l(\tilde{I}_1)$ and we can replace I_1 by $\tilde{I}_1$. By construction $f^l(I_1) \subseteq g^m(W)$, that is, $f^n(I_1) \subseteq W$ and thus $|(f^n)'(x)| \geq \varepsilon \alpha^m > 1$ 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. □

Problem 11.13. *Why is the degeneracy condition in Theorem 11.24 necessary? Can you give a counter example?*

Chapter 12

Periodic solutions

12.1. Stability of periodic solutions

In Section 6.5 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

$$\lim_{t\to\infty} d(\Phi(t,x), \gamma(x_0)) = 0 \quad \text{for all } x \in U(x_0). \tag{12.1}$$

Here $d(x,A) = \inf\{|x-y|\,|y \in A\}$ denotes the distance between x and $A \subseteq \mathbb{R}^n$ (cf. Problem 6.11).

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

$$\dot{x} = f(x) \tag{12.2}$$

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

$$A(t) = df_{\Phi(t,x_0)}, \qquad A(t+T) = A(t), \tag{12.3}$$

or problem suggests to investigate the **first variational equation**

$$\dot{y} = A(t)y, \tag{12.4}$$

which we already encountered in (2.49). 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 12.1. *The principal matrix solution of the first variational equation is given by*

$$\Pi_{x_0}(t, t_0) = \frac{\partial \Phi_{t-t_0}}{\partial x}(\Phi(t_0, x_0)). \tag{12.5}$$

Moreover, $f(\Phi(t, x_0))$ *is a solution of the first variational equation*

$$f(\Phi(t, x_0)) = \Pi_{x_0}(t, t_0) f(\Phi(t_0, x_0)). \tag{12.6}$$

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. Finally, (12.6) follows from

$$\begin{aligned} 0 = \frac{\partial}{\partial t_0}\Phi(t, x_0) &= \frac{\partial}{\partial t_0}\Phi(t - t_0, \Phi(t_0, x_0)) \\ &= -f(\Phi(t - t_0, \Phi(t_0, x_0))) + \Pi_{x_0}(t, x_0) f(\Phi(t_0, x_0)). \end{aligned}$$

□

Since $A(t)$ is periodic, all considerations of Section 3.6 apply. In particular, the principal matrix solution is of the form

$$\Pi_{x_0}(t, t_0) = P_{x_0}(t, t_0) \exp((t - t_0) Q_{x_0}(t_0)) \tag{12.7}$$

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

$$M_{x_0}(t_0) f(\Phi(t_0, x_0)) = f(\Phi(t_0, x_0)). \tag{12.8}$$

12.2. The Poincaré map

Let Σ be a transversal submanifold of codimension one containing one value x_0 from our periodic orbit. Recall the Poincaré map

$$P_\Sigma(y) = \Phi(\tau(y), y) \tag{12.9}$$

introduced in Section 6.4. 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_Σ.

Lemma 12.2. *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_Σ.*

Proof. Suppose x_0 is a stable fixed point of P_Σ. 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_Σ 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 Φ 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 12.3. *Suppose $f \in C^k$ has a periodic orbit $\gamma(x_0)$. If all eigenvalues of the derivative of the Poincaré map dP_Σ at x_0 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 12.4. *The eigenvalues of the derivative of the Poincaré map dP_Σ 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 are independent of the base point x_0 and the transversal section Σ.

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 Σ 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

$$\frac{\partial}{\partial x}\Phi(\tau(x), x)\Big|_{x=x_0} = f(x_0) d\tau_{x_0} + \frac{\partial \Phi_T}{\partial x}(x_0)$$

we infer $dP_\Sigma(x_0)_{j,k} = M_{x_0}(t_0)_{j,k}$ for $1 \le j,k \le n-1$ by Lemma 12.1. Moreover, $M_{x_0}(0)f(x_0) = f(x_0)$ and thus

$$M_{x_0}(0) = \begin{pmatrix} dP_\Sigma(x_0) & 0 \\ m & 1 \end{pmatrix}$$

from which the claim is obvious. □

Example. Consider the system

$$\dot{x}_1 = -x_2 + x_1(1 - x_1^2 - x_2^2), \quad \dot{x}_2 = x_1 + x_2(1 - x_1^2 - x_2^2)$$

and observe that a periodic solution is given by $\Phi(t) = (\cos(t), \sin(t))$. Moreover, we have

$$A(t) = \begin{pmatrix} -2\cos(t)^2 & -1 + \sin(2t)^2 \\ 1 - \sin(2t) & -2\sin(t)^2 \end{pmatrix}, \quad f(\Phi(t)) = \begin{pmatrix} -\sin(t) \\ \cos(t) \end{pmatrix}$$

Next,

$$\Pi_{x_0}(t,0) = \begin{pmatrix} \mathrm{e}^{-2t}\cos(t) & -\sin(t) \\ \mathrm{e}^{-2t}\sin(t) & \cos(t) \end{pmatrix} = \begin{pmatrix} \cos(t) & -\sin(t) \\ \sin(t) & \cos(t) \end{pmatrix} \exp\left(t \begin{pmatrix} -2 & 0 \\ 0 & 0 \end{pmatrix} \right)$$

since the second row follows from (12.6) and the first can be obtained using d'Alambert reduction. In particular,

$$M_{x_0}(0) = \begin{pmatrix} \mathrm{e}^{-4\pi} & 0 \\ 0 & 1 \end{pmatrix}$$

and the periodic orbit is stable. Note that the system can be explicitly solved in polar coordinates. ◇

As a consequence we obtain

Corollary 12.5. *The determinants of the derivative of the Poincaré map at x_0 and of the monodromy matrix are equal*

$$\det(dP_\Sigma(x_0)) = \det(M_{x_0}(t_0)). \tag{12.10}$$

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 (3.91) we have

$$\det(M_{x_0}(t_0)) = \exp\left(\int_0^T \mathrm{tr}(A(t))\, dt \right) = \exp\left(\int_0^T \mathrm{div}(f(\Phi(t,x_0))\, dt \right). \tag{12.11}$$

In two dimensions there is only one eigenvalue which is equal to the determinant and hence we obtain

Lemma 12.6. *Suppose f is a planar vector field. Then a periodic point x_0 is asymptotically stable if*

$$\int_0^T \operatorname{div}(f(\Phi(t, x_0))\, dt < 0 \tag{12.12}$$

and unstable if the integral is positive.

Example. In our previous example we have $\operatorname{div}(f(\Phi(t, x_0)) = 2-4(\sin(t)^2+\cos(t)^2) = -2$ and we again get that the periodic solution is asymptotically stable. ⋄

As another application of the use of the Poincaré map we will show that hyperbolic periodic orbits persist under small perturbations.

Lemma 12.7. *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 Σ for $f(x, 0)$ at x_0. That arc is also transversal for $f(x, \lambda)$ with λ 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. □

12.3. 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

$$\dot{y} = \tilde{f}(t, y), \qquad \tilde{f}(t, y) = f(y + \Phi(t, x_0)) - f(\Phi(t, x_0)), \tag{12.13}$$

where $\tilde{f}(t, 0) = 0$ and $\tilde{f}(t + T, x) = \tilde{f}(t, x)$. This equation can be rewritten as

$$\dot{y} = A(t)y + \tilde{g}(t, y) \tag{12.14}$$

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 3.18 the transformation $z(t) = P(t)^{-1}y(t)$ will transform the system to

$$\dot{z} = Qz + g(t, z). \tag{12.15}$$

Hence we can proceed as in Section 9.2 to show the existence of stable and unstable manifolds at x_0 defined as

$$M^{\pm}(x_0) = \{x \in M| \sup_{\pm t \geq 0} \mathrm{e}^{\pm\gamma t}|\Phi(t,x) - \Phi(t,x_0)| < \infty \text{ for some } \gamma > 0\}. \tag{12.16}$$

Making this for different points $\Phi(t_0, x_0)$ in our periodic orbit we set

$$M^{\pm}_{t_0}(x_0) = M^{\pm}(\Phi(t_0, x_0)). \tag{12.17}$$

Note that the linear counterparts are the linear subspaces

$$E^{\pm}(t_0) = \Pi_{x_0}(t_1, 0)E^{\pm}(0) \tag{12.18}$$

corresponding to the stable and unstable subspace of $M_{x_0}(t_0)$ (compare (3.127)).

Theorem 12.8 (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

$$M^{\pm}_{t_0}(x_0) \cap U(\gamma(x_0)) = \{\Phi(t_0, x_0) + a + h^{\pm}(t_0, a)| a \in E^{\pm}(t_0) \cap U\}. \tag{12.19}$$

Both $h^{\pm}(t_0, .)$ and their Jacobian matrices 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,

$$|\Phi(t,x) - \Phi(x_0, t + t_0)| \leq C\mathrm{e}^{\mp t\gamma}, \pm t \geq 0, x \in M^{\pm}_{t_0}(x_0) \tag{12.20}$$

for any $\gamma < \min\{|\mathrm{Re}(\gamma_j)|\}_{j=1}^m$ and some $C > 0$ depending on γ. Here γ_j are the eigenvalues of $Q(t_0)$.

Proof. As already pointed out before, the same proof as in Section 9.2 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 (12.18), 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

$$\begin{aligned} M^{\pm}(\gamma(x_0)) &= \\ &= \{x| \sup_{\pm t \geq 0} \mathrm{e}^{\pm\gamma t}|\Phi(t,x) - \Phi(t + t_0, x_0)| < \infty \text{ for some } t_0, \gamma > 0\} \\ &= \bigcup_{t_0 \in [0,T]} M^{\pm}_{t_0}(x_0). \end{aligned} \tag{12.21}$$

as the **stable and unstable manifold**, respectively. They are clearly invariant under the flow and are locally given by

$$\begin{aligned} &M^{\pm}(\gamma(x_0)) \cap U(\gamma(x_0)) = \\ &\quad \{\Phi(t_0, x_0) + \Pi_{x_0}(t_0, 0)a + h^{\pm}(t_0, \Pi_{x_0}(t_0, 0)a) | \\ &\qquad a \in E^{\pm}(0) \cap U,\ t_0 \in [0, T]\}. \end{aligned} \tag{12.22}$$

The points in $M^{\pm}(\gamma(x_0))$ are said to have an **asymptotic phase**, that is, there is a t_0 such that

$$\Phi(t, x) \to \Phi(t + t_0, x_0) \quad \text{as} \quad t \to \infty \text{ or } t \to -\infty. \tag{12.23}$$

As in the case of a fixed point, the (un)stable manifold coincides with the (un)stable set

$$W^{\pm}(\gamma(x_0)) = \{x | \lim_{t\to\pm\infty} d(\Phi(t, x), \gamma(x_0)) = 0\} \tag{12.24}$$

of $\gamma(x_0)$ if the orbit is hyperbolic.

Theorem 12.9. *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,*

$$W^{\pm}(\gamma(x_0)) = M^{\pm}(\gamma(x_0)). \tag{12.25}$$

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 Σ containing x and consider the corresponding Poincaré map P_Σ. 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 λ, then we already know that a hyperbolic periodic orbit persists under small perturbations and depends smoothly on the parameter by Lemma 12.7. Moreover, the same is true for the stable and unstable manifolds (which can be proven as in Theorem 9.6).

Theorem 12.10. *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*

$$\begin{aligned} M^{\pm}(\gamma(x_0(\lambda))) \cap U(\gamma(x_0(\lambda))) = \{\Phi(t_0, x_0(\lambda), \lambda) + a(\lambda) + h^{\pm}(t_0, a(\lambda)) | \\ a \in E^{\pm}(0) \cap U,\ t_0 \in [0, T]\}, \end{aligned} \tag{12.26}$$

where $a(\lambda) = \Pi_{x_0(\lambda)}(t_0, 0, \lambda)P^{\pm}(\lambda)a$, $h^{\pm} \in C^k$.

Problem 12.1 (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 μ for $\sigma = 1$ and $\sigma = -1$. Compute the stable and unstable manifolds in each case. (Hint: Use polar coordinates.)

12.4. Melnikov's method for autonomous perturbations

In Lemma 12.7 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 6.7 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 6.27, 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 (7.32) is also Hamiltonian for $\mu = 0$ and in Theorem 7.8 we have shown that one of the periodic orbits persists for $\mu > 0$.

So let us consider a Hamiltonian system

$$H(p, q) = \frac{p^2}{2} + U(q), \tag{12.27}$$

with corresponding equation of motions

$$\dot{p} = -U'(q), \quad \dot{q} = p. \tag{12.28}$$

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

$$\dot{p} = -U'(q) + \varepsilon f(p, q), \quad \dot{q} = p + \varepsilon g(p, q), \tag{12.29}$$

which is not necessarily Hamiltonian. Choosing the section $\Sigma = \{(0, q) | q > 0\}$, the corresponding Poincaré map is given by

$$P_\Sigma((0, q), \varepsilon) = \Phi(\tau(q, \varepsilon), (0, q), \varepsilon), \tag{12.30}$$

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

$$\Delta(q, \varepsilon) = \Phi_1(\tau(q, \varepsilon), (0, q), \varepsilon) - q. \tag{12.31}$$

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

$$(p(t,\varepsilon), q(t,\varepsilon)) = \Phi(t,(0,q),\varepsilon), \tag{12.32}$$

then

$$\begin{aligned}\frac{\partial}{\partial\varepsilon}\Delta(q,\varepsilon)\Big|_{\varepsilon=0} &= \frac{\partial}{\partial\varepsilon}q(\tau(q,\varepsilon),\varepsilon)\Big|_{\varepsilon=0} = \dot{q}(T(q),0)\tau_\varepsilon(q,0) + q_\varepsilon(T(q),0)\\ &= p(T(q),0)\tau_\varepsilon(q,0) + q_\varepsilon(T(q),0) = q_\varepsilon(T(q),0),\end{aligned} \tag{12.33}$$

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

$$\dot{p}_\varepsilon(t) = -U''(q_\varepsilon(t))q_\varepsilon(t) + f(p(t),q(t)), \quad \dot{q}_\varepsilon(t) = p_\varepsilon(t) + g(p(t),q(t)) \tag{12.34}$$

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

$$\begin{pmatrix} p_\varepsilon(t)\\ q_\varepsilon(t)\end{pmatrix} = \int_0^t \Pi_q(t,s)\begin{pmatrix} f(p(s),q(s))\\ g(p(s),q(s))\end{pmatrix} ds. \tag{12.35}$$

We are only interested in the value at $t = T(q)$, where

$$\Pi_q(T(q),s) = \Pi_q(T(q),0)\Pi_q(0,s) = \Pi_q(T(q),0)\Pi_q(s,0)^{-1}. \tag{12.36}$$

Furthermore, using Lemma 12.1,

$$\Pi_q(t,0)\begin{pmatrix} -U'(q)\\ 0\end{pmatrix} = \begin{pmatrix} -U'(q(t))\\ p(t)\end{pmatrix} \tag{12.37}$$

and we infer

$$\Pi_q(t,0) = \frac{1}{U'(q)}\begin{pmatrix} U'(q(t)) & -\alpha(t)U'(q(t)) + \beta(t)p(t)\\ -p(t) & \alpha(t)p(t) + \beta(t)U'(q(t))\end{pmatrix}, \tag{12.38}$$

where $\alpha(t)$ and $\beta(t)$ are given by

$$\Pi_q(t,0)\begin{pmatrix} 0\\ U'(q)\end{pmatrix} = \alpha(t)\begin{pmatrix} -U'(q(t))\\ p(t)\end{pmatrix} + \beta(t)\begin{pmatrix} p(t)\\ U'(q(t))\end{pmatrix}. \tag{12.39}$$

Moreover, by Liouville's formula we have $\det \Pi_q(t,s) = 1$ and hence

$$\beta(t) = \frac{U'(q)^2}{U'(q(t))^2 + p(t)^2} \det \Pi_q(t,0) = \frac{U'(q)^2}{U'(q(t))^2 + p(t)^2}. \tag{12.40}$$

Now putting everything together we obtain

$$\Delta_\varepsilon(q,0) = \frac{1}{U'(q)} \int_0^{T(q)} \big(p(s) f(p(s),q(s)) + U'(q(s)) g(p(s),q(s))\big)\, ds. \tag{12.41}$$

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 (7.32). 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

$$\Delta_\varepsilon(q,0) = q \int_0^{2\pi} \cos(s)^2 (\frac{\cos(s)^2}{3q^2} - 1) ds = \frac{\pi q}{4}(q^2 - 4) \tag{12.42}$$

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

$$\dot{x} = f(x) + \varepsilon\, g(x,\varepsilon). \tag{12.43}$$

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 (12.43) and abbreviate $\Phi(t,x) = \Phi(t,x,0)$. Using the orthogonal vector field

$$f^\perp(x) = J f(x), \qquad J = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}. \tag{12.44}$$

we can make the following ansatz for the principal matrix solution of the first variational equation of the unperturbed system

$$\begin{aligned} \Pi_{x_0}(t,0) f(x_0) &= f(x(t)), \\ \Pi_{x_0}(t,0) f^\perp(x_0) &= \alpha_{x_0}(t) f(x(t)) + \beta_{x_0}(t) f^\perp(x(t)), \end{aligned} \tag{12.45}$$

where $x(t) = \Phi(t,x_0)$.

Lemma 12.11. *The coefficients $\alpha_{x_0}(t)$ and $\beta_{x_0}(t)$ are given by*

$$\begin{aligned}\beta_{x_0}(t) &= \frac{|f(x_0)|^2}{|f(x(t))|^2}\, \mathrm{e}^{\int_0^t \operatorname{div}(f(x(s)))ds}\\ \alpha_{x_0}(t) &= \int_0^t \frac{\beta_{x_0}(s)}{|f(x(s))|^2} f(x(s))[J, A(s)]f(x(s))ds,\end{aligned} \tag{12.46}$$

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 (12.45) with respect to t shows

$$\dot{\alpha}(t)f(x(t)) + \dot{\beta}(t)f^{\perp}(x(t)) = \beta(t)(A(t)f^{\perp}(x(t)) - (A(t)f(x(t)))^{\perp})$$

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

$$x(u, v) = \Phi(u, \Psi(v, x_0)). \tag{12.47}$$

Abbreviate $T(v) = T(x(u, v))$ and differentiate $\Phi(T(v), x(u, v)) - x(u, v) = 0$ with respect to v producing

$$\dot{\Phi}(T(v), x(u, v))\frac{\partial T}{\partial v}(v) + \frac{\partial \Phi}{\partial x}(T(v), x(u, v))\frac{\partial x}{\partial v}(u, v) = \frac{\partial x}{\partial v}(u, v). \tag{12.48}$$

Evaluating at $(u, v) = (0, 0)$ gives

$$\Pi_{x_0}(T(x_0), 0)f^{\perp}(x_0) + \frac{\partial T}{\partial v}(0)f(x_0) = f^{\perp}(x_0). \tag{12.49}$$

Using (12.45) we obtain

$$(\alpha_{x_0}(T(x_0)) - \frac{\partial T}{\partial v}(0))f(x_0) = (1 - \beta_{x_0}(T(x_0)))f^{\perp}(x_0) \tag{12.50}$$

or equivalently

$$\alpha_{x_0}(T(x_0)) = \frac{\partial T}{\partial v}(0) = \frac{\partial T}{\partial x}(x_0)f^{\perp}(x_0), \qquad \beta_{x_0}(T(x_0)) = 1. \tag{12.51}$$

After these preparations, let us consider the Poincaré map

$$P_{\Sigma}(x, \varepsilon) = \Phi(\tau(x, \varepsilon), x, \varepsilon), \qquad x \in \Sigma, \tag{12.52}$$

corresponding to some section Σ (to be specified later). Since we expect the ε derivative to be of importance, we fix $x_0 \in \Sigma$ and compute

$$\begin{aligned}\frac{\partial}{\partial\varepsilon}\Phi(\tau(x_0,\varepsilon),x_0,\varepsilon) - x_0\Big|_{\varepsilon=0}&\\ &= \dot\Phi(T(x_0),x_0)\frac{\partial\tau}{\partial\varepsilon}(x_0,0) + \frac{\partial}{\partial\varepsilon}\Phi(T(x_0),x_0,\varepsilon)\Big|_{\varepsilon=0}\\ &= \frac{\partial\tau}{\partial\varepsilon}(x_0,0)f(x_0) + x_\varepsilon(T(x_0)),\end{aligned} \tag{12.53}$$

where $x_\varepsilon(t)$ is the solution of the variational equation

$$\dot x_\varepsilon(t) = A(t)x_\varepsilon(t) + g(x(t),0) \tag{12.54}$$

corresponding to the initial condition $x_\varepsilon(0) = 0$. Splitting g according to

$$g(x(s),0) = \frac{f(x(s))g(x(s),0)}{|f(x(s))|^2}f(x(s)) + \frac{f(x(s))\wedge g(x(s),0)}{|f(x(s))|^2}f^\perp(x(s)) \tag{12.55}$$

and invoking (12.45) we obtain after a little calculation

$$\begin{aligned}x_\varepsilon(T(x_0)) &= \int_0^{T(x_0)}\Pi_{x_0}(T(x_0),s)g(x(s),0)ds\\ &= (N(x_0) + \alpha_{x_0}(T(x_0))M(x_0))f(x_0) + M(x_0)f^\perp(x_0),\end{aligned} \tag{12.56}$$

where

$$M(x_0) = \int_0^{T(x_0)}\frac{f(x(s))\wedge g(x(s),0)}{\beta_{x_0}(s)|f(x(s))|^2}ds \tag{12.57}$$

and

$$\begin{aligned}N(x_0) =& \int_0^{T(x_0)}\frac{f(x(s))g(x(s),0)}{|f(x(s))|^2}ds\\ &- \int_0^{T(x_0)}\alpha_{x_0}(s)\frac{f(x(s))\wedge g(x(s),0)}{\beta_{x_0}(s)|f(x(s))|^2}ds.\end{aligned} \tag{12.58}$$

Putting everything together we have

$$\begin{aligned}\frac{\partial}{\partial\varepsilon}\Phi(\tau(x,\varepsilon),x,\varepsilon) - x\Big|_{\varepsilon=0}&\\ &= (\frac{\partial\tau}{\partial\varepsilon}(x,0) + N(x) + \alpha_x(T(x))M(x))f(x) + M(x)f^\perp(x)\end{aligned} \tag{12.59}$$

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

$$\Delta(v,\varepsilon) = (\Phi(\tau(x,\varepsilon),x,\varepsilon) - x)f^\perp(x_0), \quad x = x_0 + f(x_0)^\perp v, \tag{12.60}$$

and

$$\frac{\partial\Delta}{\partial\varepsilon}(0,0) = |f^\perp(x_0)|^2M(x_0). \tag{12.61}$$

Moreover, since $\Phi(\tau(x_0,\varepsilon),x_0,\varepsilon)\in\Sigma$ we have

$$\frac{\partial\tau}{\partial\varepsilon}(x_0,0)+N(x_0)+\alpha_{x_0}(T(x_0))=0 \tag{12.62}$$

and, if $M(x_0)=0$,

$$\frac{\partial^2\Delta}{\partial\varepsilon\partial v}(0,0)=|f^\perp(x_0)|^2\frac{\partial M}{\partial x}(x_0)f^\perp(x_0). \tag{12.63}$$

Theorem 12.12. *Suppose* (12.43) *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 ε.*

Note that we have

$$M(x(t))=\beta_{x_0}(t)M(x_0). \tag{12.64}$$

Problem 12.2. *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

$$\beta_{x(s)}(T(x_0))=1,\quad \alpha_{x(s)}(T(x_0))=\frac{\alpha_{x_0}(T(x_0))}{\beta_{x_0}(s)}.$$

12.5. Melnikov's method for nonautonomous perturbations

Now let us consider the more general case of nonautonomous perturbations. We consider the nonautonomous system

$$\dot{x}(t)=f(x(t))+\varepsilon\, g(t,x(t),\varepsilon) \tag{12.65}$$

ore equivalently the extended autonomous one

$$\dot{x}=f(x)+\varepsilon\, g(\tau,x,\varepsilon),\qquad \dot{\tau}=1. \tag{12.66}$$

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

$$mT=nT(x_0),\qquad n,m\in\mathbb{N}, \tag{12.67}$$

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

$$P_\Sigma(x,\varepsilon) = \Phi(mT,(x,t_0),\varepsilon) \tag{12.68}$$

and the displacement function is

$$\Delta(x,\varepsilon) = x(mT,\varepsilon) - x, \tag{12.69}$$

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 ε 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 Δ in more suitable coordinates $x(u,v)$ from (12.47). Using the results from the previous section we have

$$\frac{\partial\Delta}{\partial u}(x_0,0) = 0, \qquad \frac{\partial\Delta}{\partial v}(x_0,0) = n\,\alpha_{x_0}(T(x_0))f(x_0) \tag{12.70}$$

and

$$\begin{aligned}\frac{\partial\Delta}{\partial\varepsilon}(x_0,0) = x_\varepsilon(mT) =& (N(t_0,x_0) + n\,\alpha_{x_0}(T(x_0))M(t_0,x_0))f(x_0)\\ &+ M(t_0,x_0)f^\perp(x_0),\end{aligned} \tag{12.71}$$

where

$$M(t_0,x_0) = \int_0^{nT(x_0)} \frac{f(x(s)) \wedge g(s+t_0,x(s),0)}{\beta_{x_0}(s)|f(x(s))|^2} ds \tag{12.72}$$

and

$$\begin{aligned}N(t_0,x_0) =& \int_0^{nT(x_0)} \frac{f(x(s))g(s+t_0,x(s),0)}{|f(x(s))|^2} ds\\ &- \int_0^{nT(x_0)} \alpha_{x_0}(s)\frac{f(x(s)) \wedge g(s+t_0,x(s),0)}{\beta_{x_0}(s)|f(x(s))|^2} ds.\end{aligned} \tag{12.73}$$

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 12.13. *Suppose* (12.65) *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 ε.*

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 (12.47))

$$\Delta(x(u, v), \varepsilon) = \Delta_1(u, v, \varepsilon) f(x_0) + \Delta_2(u, v, \varepsilon) f^{\perp}(x_0). \tag{12.74}$$

Then

$$\frac{\partial \Delta_1}{\partial v}(0, 0, 0) = n\, \alpha_{x_0}(T(x_0)) \neq 0 \tag{12.75}$$

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

$$\tilde{\Delta}_2(u, \varepsilon) = \Delta_2(u, v(u, \varepsilon), \varepsilon). \tag{12.76}$$

Since $\tilde{\Delta}_2(u, 0) = \Delta_2(u, 0, 0) = 0$, we can divide by ε and apply the implicit function theorem as before.

Now using

$$\frac{\partial \tilde{\Delta}_2}{\partial \varepsilon}(0, 0) = M(t_0, x_0). \tag{12.77}$$

and, if $M(t_0, x_0) = 0$,

$$\frac{\partial^2 \tilde{\Delta}_2}{\partial \varepsilon \partial u}(0, 0) = \frac{\partial M}{\partial x}(t_0, x_0) f(x_0) \tag{12.78}$$

we obtain the following result.

Theorem 12.14. *Suppose* (12.65) *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 ε.*

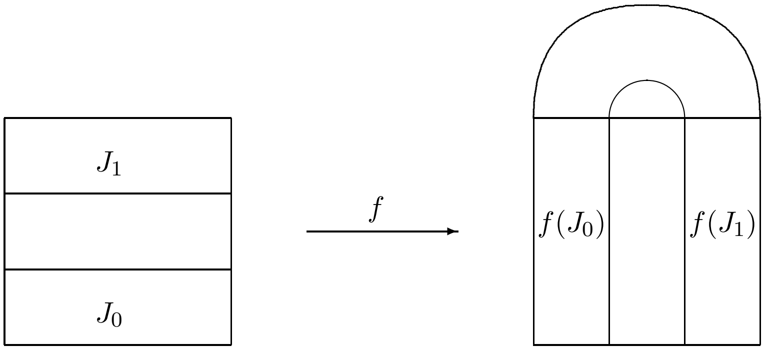
Chapter 13

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_μ. 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 Λ 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 Λ can uniquely defined by a sequence y_n, $n \in \mathbb{N}_0$. Similarly, the x coordinate of every point in Λ 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 Λ and doubly infinite sequences on two symbols. Hence we have found again an itinerary map

$$\begin{array}{rccl} \varphi : & \Lambda & \to & \Sigma_2 \\ & (x,y) & \mapsto & s_n = \left\{ \begin{array}{ll} y_n & n \geq 0 \\ x_{-n-1} & n < 0 \end{array} \right. , \end{array} \tag{13.9}$$

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 φ is continuous (exercise). Now what about the action of $\sigma = \varphi \circ f \circ \varphi^{-1}$? By construction, σ shifts y_n to the left, $\sigma(s)_n = y_{n+1}$, $n \geq 0$, and σ^{-1} shifts x_n to the left, $\sigma^{-1}(s)_n = x_{-n-1}$, $n < 0$. Hence σ 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 σ just shifts s_n to the left, $\sigma(s)_n = s_{n+1}$. In summary, we have shown

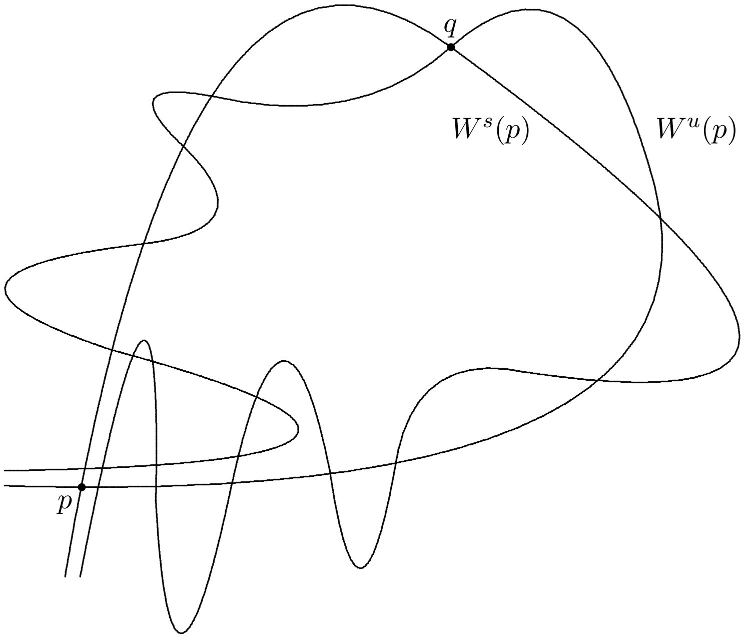
Theorem 13.1. *The Smale horseshoe map has an invariant Cantor set Λ on which the dynamics is equivalent to the double sided shift on two symbols. In particular it is chaotic.*

13.2. The Smale–Birkhoff homoclinic theorem

In this section I will present the higher dimensional analog of Theorem 11.24.

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 Λ 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 [**33**].

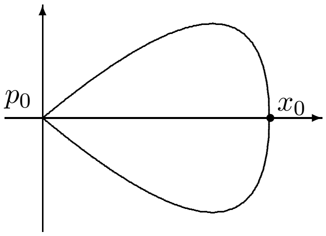
13.3. Melnikov's method for homoclinic orbits

Finally we want to combine the Smale–Birkhoff theorem from the previous section with Melnikov's method from Section 12.5 to obtain a criterion for chaos in ordinary differential equations.

Again we will start with a planar system

$$\dot{x} = f(x) \tag{13.10}$$

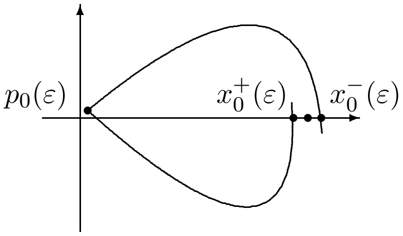
which has a homoclinic orbit $\gamma(x_0)$ at a fixed point p_0. For example, we could take Duffing's equation from Problem 9.5 (with $\delta = 0$). The typical situation for the unperturbed system is depicted below.



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 ε 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 τ, 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 τ 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 Σ as the section of a corresponding Poincaré map P_Σ, these intersections are just the stable and unstable manifold of the fixed point $p_0(\varepsilon) = p_0(t_0, \varepsilon)$ of P_Σ. 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, τ) 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 12.4 to conclude that $\Delta(z_0, \varepsilon)$ has a zero for small ε 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 12.4. 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 12.10 a point in $W^\pm(\gamma(p_0(\varepsilon)))$ can locally be written as

$$(p_0 + h_0^\pm(\tau, a) + h_1^\pm(\tau, a)\varepsilon + o(\varepsilon), \tau).$$

Moreover, fixing $x_0 \in W(p_0)$ there is a unique $s = s(\tau, a)$ such that

$$p_0 + h_0^\pm(\tau, a, 0) = \Phi(s, x_0)$$

and hence we can choose $h^\pm(\tau, s) = h_1^\pm(\tau, a(\tau, s))$. □

Hence we even have

$$y^{\pm}(t) = \int_{\pm\infty}^{t} \mathrm{e}^{\int_s^t \mathrm{tr}(A(r))dr} f(x_0(s)) \wedge g(s-t_0, x_0(s), 0)\, ds \tag{13.20}$$

and thus finally

$$\frac{\partial \Delta}{\partial \varepsilon}(z_0, 0) = M_{x_0}(t_0), \tag{13.21}$$

where $M_{x_0}(t_0)$ is the **homoclinic Melnikov integral**

$$M_{x_0}(t) = \int_{-\infty}^{\infty} \mathrm{e}^{-\int_0^s \mathrm{div}(f(\Phi(r,x_0)))dr} f(\Phi(s,x_0)) \wedge g(s-t, \Phi(s,x_0), 0)\, ds. \tag{13.22}$$

Note that the base point x_0 on the homoclinic orbit is not essential since we have (Problem 13.2)

$$M_{\Phi(t,x_0)}(t_0) = \mathrm{e}^{\int_0^t \mathrm{div}(f(\Phi(r,x_0)))dr} M_{x_0}(t+t_0). \tag{13.23}$$

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_Σ has a transversal homoclinic orbit for sufficiently small $\varepsilon \neq 0$.*

For example, consider the forced Duffing equation (compare Problem 9.5)

$$\dot{q} = p, \quad \dot{p} = q - q^3 - \varepsilon(\delta p + \gamma \cos(\omega\tau)), \quad \dot{\tau} = 1. \tag{13.24}$$

The homoclinic orbit is given by

$$q_0(t) = \sqrt{2}\,\mathrm{sech}(t), \quad p_0(t) = -\sqrt{2}\tanh(t)\mathrm{sech}(t) \tag{13.25}$$

and hence

$$\begin{aligned} M(t) &= \int_{-\infty}^{\infty} q_0(s)\left(\delta p_0(s) + \gamma\cos(\omega(s-t))\right) ds \\ &= \frac{4\delta}{3} - \sqrt{2}\pi\gamma\omega \mathrm{sech}(\frac{\pi\omega}{2}) \sin(\omega t) \end{aligned} \tag{13.26}$$

Thus the Duffing equation is chaotic for δ, γ sufficiently small provided

$$\left|\frac{\delta}{\gamma}\right| < \frac{3\sqrt{2}\pi|\omega|}{4} \mathrm{sech}(\frac{\pi\omega}{2}). \tag{13.27}$$

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* (13.23).

Problem 13.3. *Apply the Melnikov method to the forced mathematical pendulum (compare Section 6.7)*

$$\dot{q} = p, \qquad \dot{q} = -\sin(q) + \varepsilon \sin(t).$$

The End

Bibliographical notes

The aim of this section is not to give a comprehensive guide to the literature, but to document the sources from which I have learned the materials and which I have used during the preparation of this text. In addition, I will point out some standard references for further reading.

Chapter 2: Initial value problems

The material in this section is of course classical. Classical references are Coddington and Levinson [**6**], Hartman [**13**], Hale [**12**], Ince [**23**], or Walter [**42**]. More modern introductions are Arnold [**3**], Hirsch, Smale, and Devaney [**18**], Robinson [**34**], Verhulst [**41**], or Wiggins [**46**].

Further uniqueness results can be found in the book by Walter [**42**] (see the supplement to §12). There you can also find further technical improvements, in particular, for the case alluded to in the remark after Corollary 2.6 (see the second supplement to §10).

More on Mathematica in general can be found in the standard documentation [**47**] and in connections with differential equations in [**10**], [**37**].

General purpose references are the handbooks by Kamke [**24**] and Zwillinger [**48**].

Chapter 3: Linear equations

Again this material is mostly standard and the same references as for the previous chapter apply. More information in particular on n'th order equations can be found in Coddington and Levinson [**6**], Hartman [**13**], Ince [**23**].

Chapter 4: Differential equations in the complex domain

Classical references with more information on this topic include Coddington and Levinson [**6**], Hille [**17**], or Ince [**23**]. For a more modern point of view see Ilyashenko and Yakovenko [**21**]. The topics here are also closely connected with the theory of special functions, see Beals and Wong [**4**] for a modern introduction.

Chapter 5: Boundary value problems

Classical references include Coddington and Levinson [**6**], Hartman [**13**]. A nice informal treatment (although in German) can be found in Jänich [**22**]. More on Hill's equation can be found in Magnus and Winkler [**27**]. For a modern introduction to singular Sturm–Liouville problems see the books by Weidmann [**43**], [**44**], my textbook [**40**], or the book by Levitan and Sargsjan [**26**]. A reference with more applications and numerical methods is by Hastings and McLeod [**16**].

Chapter 6: Dynamical systems

Classical references include Chicone [**5**], Guckenheimer and Holmes [**11**], Hasselblat and Katok [**14**],[**15**], Hirsch, Smale, and Devaney [**18**], Palis and de Melo [**31**], Perko [**32**], Robinson [**33**], [**34**], Ruelle [**36**], Verhulst [**41**], and Wiggins [**45**], [**46**]. In particular, [**14**], [**15**] has emphasis on ergodic theory which is not covered here.

More on the connections with Lie groups and symmetries of differential equations briefly mentioned in Problem 6.5 can be found in the monograph by Olver [**29**].

Chapter 7: Planar dynamical systems

The proof of the Poincaré–Bendixson theorem follows Palis and de Melo [**31**]. More on ecological models can be found in Hofbauer and Sigmund [**19**]. Hirsch, Smale, and Devaney [**18**], Robinson [**34**] also cover these topics nicely.

Chapter 8: Higher dimensional dynamical systems

More on the Lorenz equation can be found in the monograph by Sparrow [**38**]. The classical reference for Hamiltonian systems is of course Arnold's book [**2**] (see also [**3**]) as well as the monograph by Abraham, Marsden, and Ratiu [**1**], which also contains extensions to infinite dimensional systems. Other references are and the notes by Moser [**28**] and the monograph by Wiggins [**45**]. A brief overview can be found in Verhulst [**41**].

Chapter 9: Local behavior near fixed points

The classical reference here is Hartman [**13**]. See also Coddington and Levinson [**6**], Hale [**12**], Robinson [**33**], or Ruelle [**36**].

Chapter 10: Discrete dynamical systems

One of the classical reference is the book by Devaney [**7**]. A nice introduction is by Holmgren citehol. Furhter references are Hasselblat and Katok [**14**], [**15**], Robinson [**34**].

Chapter 11: Discrete dynamical systems in one dimension

The classical reference here is Devaney [**7**]. More on the Hausdorff measure can be found in Falconer [**8**]. See also Holmgren [**20**], Robinson [**34**].

Chapter 12: Periodic solutions

For more information see Chicone [**5**], Robinson [**33**], [**34**], Wiggins [**45**].

Chapter 13: Chaos in higher dimensional systems

A proof of the Smale–Birkhoff theorem can be found in Robinson [**33**]. See also Chicone [**5**], Guckenheimer and Holmes [**11**], Wiggins [**45**].

Bibliography

[1] R. Abraham, J. E. Marsden, and T. Ratiu, *Manifolds, Tensor Analysis, and Applications*, 2^{nd} edition, Springer, New York, 1983.

[2] V.I. Arnold, *Mathematical methods of classical mechanics*, 2^{nd} ed., Springer, New York, 1989.

[3] V.I. Arnold, *Ordinary Differential Equations*, Springer, Berlin, 1992.

[4] R. Beals and R. Wong, *Special Functions*, Cambridge University Press, Cambridge, 2010.

[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^{nd} ed., Addison-Wesley, Redwood City, 1989.

[8] K. Falconer, *Fractal Geometry*, Benjamin/Clummings Publishing, Menlo Park, 1986.

[9] F. R. Gantmacher, *Applications of the Theory of Matrices*, Interscience, New York, 1959.

[10] A. Gray, M. Mezzino, and M. A. Pinsky, *Introduction to Ordinary Differential Equations with Mathematica*, Springer, New York, 1997.

[11] J. Guckenheimer and P. Holmes, *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*, Springer, New York, 1983.

[12] J. Hale, *Ordinary Differential Equations*, Krieger, Malabar, 1980.

[13] P. Hartman, *Ordinary Differential Equations*, 2nd ed., SIAM, Philadelphia, 2002.

[14] B. Hasselblatt and A. Katok, *Introduction to the Modern Theory of Dynamical Systems*, Cambridge UP, Cambridge 1995.

[15] B. Hasselblatt and A. Katok, *A First Course in Dynamics*, Cambridge UP, Cambridge 2003.

[16] S. P. Hastings and J. B. McLeod *Classical Methods in Ordinary Differential Equations: With Applications to Boundary Value Problems*, Amer. Math. Soc., Rhode Island, 2011.

[17] E. Hille, *Ordinary Differential Equations in the Complex Domain*, Dover 1997.

[18] M. W. Hirsch, S. Smale, and R. L. Devaney *Differential Equations, Dynamical Systems, and an Introduction to Chaos*, Elsevier, Amsterdam, 2004.

[19] J. Hofbauer and K. Sigmund, *Evolutionary Games and Replicator Dynamics*, Cambridge University Press, Cambridge, 1998.

[20] R.A. Holmgren, *A First Course in Discrete Dynamical Systems*, 2nd ed., Springer, New York, 1996.

[21] Y. Ilyashenko and S. Yakovenko, *Lectures on Analytic Differential Equations*, Graduate Studies in Mathematics **86**, Amer. Math. Soc., Rhode Island, 2008.

[22] K. Jänich, *Analysis*, 2^{nd} ed., Springer, Berlin, 1990.

[23] E.L. Ince, *Ordinary Differential Equations*, Dover Publ., New York, 1956.

[24] E. Kamke, *Differentialgleichungen, I. Gewöhnliche Differentialgleichungen*, Springer, New York, 1997.

[25] J.L. Kelly, *General Topology*, Springer, New York, 1955.

[26] B.M. Levitan and I.S. Sargsjan, *Introduction to Spectral Theory*, Amer. Math. Soc., Providence, 1975.

[27] W. Magnus and S. Winkler, *Hill's Equation*, Dover, Minolea, 2004.

[28] J. Moser, *Stable and Random Motions in Dynamical Systems: With Special Emphasis on Celestial Mechanics*, Princeton University Press, Princeton, 2001.

[29] P. J. Olver, *Applications of Lie Groups to Differential Equations*, 2nd ed., Springer, New York, 1993.

[30] R.S. Palais, *The symmetries of solitons*, Bull. Amer. Math. Soc., **34**, 339–403 (1997).

[31] J. Palis and W. de Melo, *Geometric Theory of Dynamical Systems*, Springer, New York, 1982.

[32] L. Perko, *Differential Equations and Dynamical Systems*, 2^{nd} ed., Springer, New York, 1996.

[33] C. Robinson, *Dynamical Systems: Stability, Symbolic Dynamics, and Chaos*, CRC Press, Boca Raton, 1995.

[34] C. Robinson, *Introduction to Dynamical Systems: Discrete and Continuous*, Prentice Hall, New York, 2004.

[35] C.A. Rogers, *Hausdorff Measures*, Cambridge University Press, Cambridge, 1970.

[36] D. Ruelle, *Elements of Differentiable Dynamics and Bifurcation Theory*, Academic Press, San Diego, 1988.

[37] D. Schwalbe and S. Wagon, *VisualDSolve. Visualizing Differential Equations with Mathematica*, Springer, New York, 1997.

[38] C. Sparrow, *The Lorenz Equation, Bifurcations, Chaos and Strange Attractors*, Springer, New York, 1982.

[39] E. Stein and R. Shakarchi, *Complex Analysis*, Princeton UP, Princeton, 2003.

[40] G. Teschl, *Mathematical Methods in Quantum Mechanics; With Applications to Schrödinger Operators*, Amer. Math. Soc., Rhode Island, 2009.

[41] F. Verhulst, *Nonlinear Differential Equations and Dynamical Systems*, 2nd ed., Springer, Berlin, 2000.

[42] W. Walter, *Ordinary Differential Equations*, Springer, New York, 1998.

[43] J. Weidmann, *Linear Operators in Hilbert Spaces*, Springer, New York, 1980.

[44] J. Weidmann, *Spectral Theory of Ordinary Differential Operators*, Lecture Notes in Mathematics **1258**, Springer, Berlin, 1987.

[45] S. Wiggins, *Global Bifurcations and Chaos*, 2^{nd} ed., Springer, New York, 1988.

[46] S. Wiggins, *Introduction to Applied Nonlinear Dynamical Systems and Chaos*, 2nd ed., Springer, New York, 2003.

[47] S. Wolfram, *The Mathematica Book*, 4th ed., Wolfram Media/Cambridge University Press, Champaign/Cambridge, 1999.

[48] D. Zwillinger, *Handbook of Differential Equations*, 3rd ed., Academic Press, San Diego, 1997.

Glossary of notation

$A_\pm$	. . . matrix A restricted to $E^\pm(A)$
$B_r(x)$	. . . open ball of radius r centered at x
$C(U,V)$	. . . set of continuous functions from U to V
$C_b(U,V)$	. . . set of bounded 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
χ_A	. . . Characteristic polynomial of A, 103
$d(U)$	. . . diameter of U, 307
$d(x,y)$	. . . distance in a metric space
$d(x,A)$	. . . distance between a point x and a set A, 196
df_x	$= \frac{\partial f}{\partial x}$ Jacobian matrix of a differentiable mapping f at x
$\delta_{j,k}$	. . . Kronecker delta: $\delta_{j,j} = 1$ and $\delta_{j,k} = 0$ if $j \neq k$
$E^0(A)$	. . . center subspace of a matrix, 109
$E^\pm(A)$	. . . (un)stable subspace of a matrix, 109
$\mathrm{Fix}(f)$	$= \{x \vert f(x) = x\}$ set of fixed points of f, 282
$\gamma(x)$	. . . orbit of x, 192
$\gamma_\pm(x)$	. . . forward, backward orbit of x, 192
$\Gamma(z)$	. . . Gamma function, 126
$\mathfrak{H}_0$	. . . inner product space, 146
$\mathbb{I}$	. . . identity matrix
I_x	$= (T_-(x), T_+(x))$ maximal interval of existence, 189
$\mathrm{Ker}(A)$	. . . kernel of a matrix
L_μ	. . . logistic map, 280
Λ	. . . a compact invariant set
$M^\pm$	. . . (un)stable manifold, 256, 320

$\mathbb{N}$	$= \{1, 2, 3, \dots\}$ the set of positive integers
$\mathbb{N}_0$	$= \mathbb{N} \cup \{0\}$
$o(.)$	$\dots$ Landau symbol
$O(.)$	$\dots$ Landau symbol
$\Omega(f)$	$\dots$ set of nonwandering points, 196
$P_\Sigma(y)$	$\dots$ Poincaré map, 197
$\mathrm{Per}(f)$	$= \{x \vert f(x) = x\}$ set of periodic points of f, 282
$\Phi(t, x_0)$	$\dots$ flow of a dynamical system, 189
$\Pi(t, t_0)$	$\dots$ principal matrix of a linear system, 81
$\mathbb{R}$	$\dots$ the set of reals
$\mathrm{Ran}(A)$	$\dots$ range of a matrix
σ	$\dots$ shift map on Σ_N, 303
$\sigma(A)$	$\dots$ spectrum (set of eigenvalues) of a matrix
Σ_N	$\dots$ sequence space over N symbols, 302
$\mathrm{sign}(x)$	$\dots +1$ for $x > 0$ and -1 for $x < 0$; sign function
$T_\pm(x)$	$\dots$ positive, negative lifetime of x, 192
$T(x)$	$\dots$ period of x (if x is periodic), 192
T_μ	$\dots$ tent map, 297
$\omega_\pm(x)$	$\dots$ positive, negative ω-limit set of x, 193
$W^\pm$	$\dots$ (un)stable set, 255, 231, 282
$\mathbb{Z}$	$= \{\dots, -2, -1, 0, 1, 2, \dots\}$ the set of integers
z	$\dots$ a complex number
$\sqrt{z}$	$\dots$ square root of z with branch cut along $(-\infty, 0)$
z^*	$\dots$ complex conjugation
$\Vert.\Vert$	$\dots$ norm in a Banach space
$\vert.\vert$	$\dots$ Euclidean norm in $\mathbb{R}^n$ respectively $\mathbb{C}^n$
$\langle ., .. \rangle$	$\dots$ scalar product in $\mathfrak{H}_0$, 146
(λ_1, λ_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
$\lfloor x \rfloor$	$= \max\{n \in \mathbb{Z} \vert n \le x\}$, floor function
$\lceil x \rceil$	$= \min\{n \in \mathbb{Z} \vert n \ge x\}$, ceiling function
$a \wedge b$	$=$ cross product in $\mathbb{R}^3$

Index