Jan Glaubitz
Department of Mathematics
Dartmouth College, NH, USA
Email: jan.glaubitz@dartmouth.edu

Script for

Math 46 (Spring 2020)
Introduction to Applied Mathematics

Version: May 31, 2020

Abstract

This course provides an introduction into the field of Applied Mathematics. In particular, the emphasis is upon mathematical tools to describe and analyze real world phenomena which play a central role, for instance, in the applied and natural sciences. However, the focus will lie 'only' on models resulting in *ordinary* differential and integral equations. Still, this course requires you (the student or any other reader) to already be familiar with basic concepts from ordinary differential equations as well as linear algebra. I would like to point out that this script as well as the corresponding lectures are mainly based on the first four chapters of the third edition of Logan's monograph [Log13]. Moreover, for the second chapter of this script (Dimensional Analysis and Scaling), I also draw inspiration from the lecture notes [Her19] of my former colleague Michael Herrmann from TU Braunschweig in Germany.

Table of Contents

Chapter 1

Introduction - What is Applied Mathematics?

Applied mathematics is a broad subject area dealing with the description, analysis, and prediction of real-world phenomena. It is more than a set of methods that can be used to solve equations that come from physics, engineering, and other applied sciences. In fact, applied mathematics also is about mathematical modeling and an entire process that intertwines with the physical reality. In this course, by a **mathematical model** (sometimes just called a **model**) we usually refer to an equation that describes some physical problem or phenomenon. Moreover, by **mathematical modeling** we mean the process of formulating, analyzing, and potentially refining a mathematical model. In particular, this process includes the following steps:

1. Introducing the relevant quantities or variables in the model
2. Solving the model by some (analytical or numerical) method
3. Comparing the solution of the model to real-world data and interpreting the results
4. If necessary, revising the model until it describes the underlying physical problem sufficiently accurate

Hence, mathematical modeling involves physical intuition, formulation of equations, solution methods, and analysis. Please note that solution methods can be both, of analytical or numerical nature. In this course, however, we will restrict ourselves to analytical methods (the ones involving the use of pen and paper) rather than using numerical methods (the ones involving the use of computers). Moreover, you (the student) are expected to already be familiar with basic solution methods for ordinary differential equations, which will be the class of mathematical models we will focus on in this course. Finally, let us — at least loosely — agree on what we demand from a 'good' model:

- It should be as simple as possible and as complex as necessary[1]
- It should apply to many situations
- It should be predictive

Here, the last criterion means that a model should not just reflect all known features of a real-world process but it should also be able to predict previously unknown mechanisms or phenomena. A prominent example for this are, for instance, gravitational waves, which have already been predicted in 1916 by Einstein [Ein05, Ein18] on the basis of his general theory of relativity (a mathematical model) and directly observed only in 2016 by the LIGO and Virgo Scientific Collaboration (resulting in the

[1] Of course, the model needs to be sufficiently complex to include all relevant mechanisms of the underlying real-world process. Yet, when we have two different models at hand which both describe the same process equally well, we should always prefer the simpler model. This principle is often referred to as **Ockham's razor** [Ari76]

2017 Nobel Prize in Physics). Another example of current significance is the spread of diseases, which can often be modeled by reaction-diffusion equations. In particular, applied mathematics can therefore help us determine which parameters and processes are important (e. g. washing hands) and which are unimportant (e. g. buying toilet paper).

Chapter 2

Dimensional Analysis and Scaling

2.1 Physical Dimensions and Units

Almost all physical quantities we will deal with in this course have a physical (or some other) dimension. Furthermore, this dimension is usually measured by using certain units. It is worth listing and getting familiar some of the more important dimensions and their corresponding units right from the start of this course. This will save us some trouble later on. Let us start by noting that there are two different classes of physical dimensions:

1. **Basis dimensions** (also called **primary dimensions**)
2. **Derived dimensions** (also called **secondary dimensions**)

As the name indicates, basis dimensions are defined independent or fundamental dimensions, from which other dimensions can be obtained. Table 2.1 lists the seven basis dimensions as well as their corresponding symbols and units. We should note that there can be many possible units for measuring a certain dimension. For instance, length can be measured not just by the unit *meter* but also by *millimeter, kilometer, inches, foot, mile*, and many more. In this course, however, we will only use **SI units** (more commonly referred to as **metric units**) from the International System of Units (also *Le Systeme International d'Unites*) [oWMTT01]. The SI units corresponding to the seven basis dimensions can be found in Table 2.1 as well.

All other dimensions can be derived as combinations (powers and products) of these seven basis dimensions and are therefore referred to as derived dimensions. By way of example, all of you will already know that speed is measured, for instance, by dividing a certain number of meters (or any other unit for length, e. g. miles) by a certain number of seconds (or any other unit for time, e. g. hours). In the same manner, many more dimensions can be derived from the seven basis dimensions. Some of the more important ones, together with their corresponding SI units and symbols, can be found in 2.1. Table 2.1 also lists how the respective dimension is derived from the basis dimensions or some other derived dimensions.

Henceforth, given a quantity x, we denote

$$[x] \text{ for the dimension of } x.$$

Note that often a mathematical model, $x = y$, only makes sense if both sides of the equation have the same dimensions, that is $[x] = [y]$. Speaking plainly, we might not want to accidentally compare appels and oranges!

Basis dimensions

dimension	SI unit	symbol
mass	kg (kilogram)	$\mathbf{m}$ (sometimes $\mathbf{M}$)
length	m (meter)	$\mathbf{L}$ (sometimes $\mathbf{l}$)
time	s (second)	$\mathbf{t}$ (sometimes $\mathbf{T}$)
temperature	K (Kelvin)	$\mathbf{T}$ (sometimes $\mathbf{q}$)
electric current	A (ampere)	$\mathbf{I}$ (sometimes $\mathbf{I}$)
amount of light (luminous intensity)	c (candela)	$\mathbf{C}$ (sometimes $\mathbf{I}$)
amount of matter	mol (mole)	$\mathbf{n}$ (sometimes $\boldsymbol{\mu}$)
real numbers	none	$\mathbf{1}$ (dimensionless)

Table 2.1: The seven basis dimensions as well as their corresponding SI units and symbols. Note that real numbers are dimensionless.

Derived dimensions

dimension	SI unit	symbol
speed	$\frac{\mathrm{m}}{\mathrm{s}}$ (meter per second)	$\mathbf{s} = \frac{\mathbf{L}}{\mathbf{t}}$
acceleration	$\frac{\mathrm{m}}{\mathrm{s}^2}$ (meter per square second)	$\mathbf{a} = \frac{\mathbf{L}}{\mathbf{t}^2}$
force	N (Newton)	$\mathbf{F} = \frac{\mathbf{m} \cdot \mathbf{L}}{\mathbf{t}^2}$
pressure	Pa (Pascal)	$\mathbf{p} = \frac{\mathbf{F}}{\mathbf{L}^2}$
energy	J (Joule)	$\mathbf{E} = \mathbf{F} \cdot \mathbf{L}$
power	W (watt)	$\mathbf{P} = \frac{\mathbf{E}}{\mathbf{t}}$
density	$\frac{\mathrm{kg}}{\mathrm{m}^3}$ (mass per volume)	$\boldsymbol{\rho} = \frac{\mathbf{m}}{\mathbf{L}^3}$

Table 2.2: Some derived dimensions as well as their corresponding SI units and symbols. This list might grow throughout the course!

2.2 Dimensional Analysis

Example 2.2.1: Atomic explosions and the Taylor–Sedov formula

To demonstrate the flavor of the concepts discussed in this section, let us consider a calculation made by the British applied mathematician Taylor in the late 1940s. After the first atomic bomb and viewing photographs of the spread of its fireball, Taylor wanted to compute its yield (energy released). From photographs it becomes clear that such an explosion results in a spherical blast wave front. Taylor then argued that there is a relation between the radius r of this blast wave, time t, the initial air density ρ, and the energy released E. Hence, he assumed that there should be a simple physical law of the form

$$r^\alpha t^\beta \rho^\gamma E^\delta = C, \tag{2.1}$$

for some $C \in \mathbb{R}$. All quantities as well as their corresponding dimensions can be found in Table 2.3. Note that we have a dimensionless quantity at the right hand side of (2.1) (remember that $[C] = \mathbf{1}$)! Thus, for (2.1) to make sense, also the dimensions of the quantities on the left hand side have to cancel each other out, i. e.

$$\left[r^\alpha t^\beta \rho^\gamma E^\delta \right] = \mathbf{1}.$$

Consulting tables 2.1 and 2.2 from Chapter 2.1, this results in

$$(\mathbf{L})^\alpha (\mathbf{t})^\beta \left(\mathbf{m}\mathbf{L}^{-3} \right)^\gamma \left(\mathbf{m}\mathbf{L}^2\mathbf{t}^{-2} \right)^\delta = \mathbf{1}.$$

Next, since all dimensions have to cancel out, this provides us with a system of linear equations,

$$\begin{aligned} \alpha - 3\gamma + 2\delta &= 0 \qquad &&(\text{for } \mathbf{L}), \\ \beta - 2\delta &= 0 \qquad &&(\text{for } \mathbf{t}), \\ \gamma = \delta &= 0 \qquad &&(\text{for } \mathbf{m}), \end{aligned}$$

with the general solution

$$\alpha = -5\delta, \quad \beta = 2\delta, \quad \gamma = -\delta.$$

Independent of the choice of δ, this implies the **Taylor–Sedov formula** given by

$$\frac{t^2 E}{r^5 \rho} = C, \tag{2.2}$$

for some $C \in \mathbb{R}$. In particular, we get

$$r = C \left(\frac{Et^2}{\rho} \right)^{1/5}. \tag{2.3}$$

That is, just from dimensional reasoning (and the physical assumption (2.1)), it can be shown that the radius of the blast wave depends on the two-fifth power of time. It should be pointed out that in this example the constant C depends on the dimensionless ratio of specific heats.[a] Moreover, also ρ is usually known. Finally, the initial energy yield can be calculated by fitting the curve (2.3) to experimental data of r versus t.

[a] Taylor used the value $C = 1$

quantity	dimension	symbol
r (radius)	length	$\mathbf{L}$
t (time after detonation)	time	$\mathbf{t}$
ρ (initial air density)	density	$\boldsymbol{\rho} = \frac{\mathbf{m}}{\mathbf{L}^3}$
E (energy release)	energy	$\mathbf{E} = \frac{\mathbf{mL}^2}{\mathbf{t}^2}$

Table 2.3: The quantities and corresponding dimensions used in Example 2.2.1

Definition 2.2.2: Independent and dependent dimensions

We call an m-tuple $(\mathbf{E}_1, \dots, \mathbf{E}_m)$ of dimension symbols **independent** if there exists no non-trivial exponential vector $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_m) \in \mathbb{R}^m \setminus \{0\}$ such that

$$\mathbf{E}_1^{\alpha_1} \cdots \mathbf{E}_m^{\alpha_m} = \mathbf{1}$$

holds. Otherwise, we refer to the m-tuple $(\mathbf{E}_1, \dots, \mathbf{E}_m)$ as **dependent**.

Examples 2.2.3

(a) Let us consider the 3-tupel $(\mathbf{m}, \mathbf{L}, \mathbf{s})$ containing the dimension symbols for mass, length, and speed. Consulting Table 2.2 we can note that

$$\mathbf{s} = \mathbf{L} \cdot \mathbf{t}^{-1}$$

and therefore

$$\mathbf{m}^{\alpha_1}\mathbf{L}^{\alpha_2}\mathbf{s}^{\alpha_3} = \mathbf{m}^{\alpha_1}\mathbf{L}^{\alpha_2+\alpha_3}\mathbf{t}^{-\alpha_3}.$$

Obviously, only the trivial vector $\boldsymbol{\alpha} = (0, 0, 0)$ satisfies

$$\mathbf{m}^{\alpha_1}\mathbf{L}^{\alpha_2}\mathbf{s}^{\alpha_3} = \mathbf{1}.$$

Hence, the 3-tupel $(\mathbf{m}, \mathbf{L}, \mathbf{s})$ is independent.

(b) Next, let us consider the 3-tupel $(\mathbf{m}, \mathbf{L}, \boldsymbol{\rho})$. This time we have

$$\boldsymbol{\rho} = \mathbf{m} \cdot \mathbf{L}^{-3}$$

and therefore

$$\mathbf{m}^{\alpha_1}\mathbf{L}^{\alpha_2}\boldsymbol{\rho}^{\alpha_3} = \mathbf{m}^{\alpha_1+\alpha_3}\mathbf{L}^{\alpha_2-3\alpha_3}.$$

It is easy to note that there are non-trivial vectors $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_3)$ satisfying

$$\mathbf{m}^{\alpha_1}\mathbf{L}^{\alpha_2}\boldsymbol{\rho}^{\alpha_3} = \mathbf{1}.$$

These are given by the infinite many solutions $\boldsymbol{\alpha} = (-\alpha_3, 3\alpha_3, \alpha_3)$ with $\alpha_3 \in \mathbb{R}$ of the system of linear equations

$$\begin{aligned} \alpha_1 + \alpha_3 &= 0 \qquad (\text{for } \mathbf{m}), \\ \alpha_2 - 3\alpha_3 &= 0 \qquad (\text{for } \mathbf{L}). \end{aligned}$$

Hence, the 3-tupel $(\mathbf{m}, \mathbf{L}, \boldsymbol{\rho})$ is dependent.

Definition 2.2.4: Representation of quantities

We say that the quantity q can be **represented** with respect to the vector of dimension symbols $(\mathbf{E}_1, \dots, \mathbf{E}_m)$ if

$$[q] = \mathbf{E}_1^{\alpha_1} \cdots \mathbf{E}_m^{\alpha_m}$$

for some vector $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_m) \in \mathbb{R}^m$. The vector $\boldsymbol{\alpha}$ is often called **dimension vector** of q with respect to $(\mathbf{E}_1, \dots, \mathbf{E}_m)$.

Example 2.2.5

We note that there can be different representations of quantities. Let us consider the quantity q measuring energy, i. e., $[q] = \mathbf{E}$. Consulting Table 2.2, q can be represented with respect to $(\mathbf{F}, \mathbf{L})$, since

$$[q] = \mathbf{E} = \mathbf{F} \cdot \mathbf{L} \qquad \text{(force times length).} \tag{2.4}$$

The corresponding dimension vector is given by $\boldsymbol{\alpha} = (1, 1)$. Yet, at the same time q can be represented with respect to $(\mathbf{m}, \mathbf{t}, \mathbf{L})$. Note that force can be derived by

$$\mathbf{F} = \mathbf{m} \cdot \mathbf{L} \cdot \mathbf{t}^{-2}$$

and we can therefore rewrite (2.4) as

$$[q] = \mathbf{E} = \mathbf{m} \cdot \mathbf{t}^{-2} \cdot \mathbf{L}^2.$$

This time, the corresponding dimension vector is given by $\boldsymbol{\alpha} = (1, -2, 2)$.

Remark 2.2.6: Uniqueness of the dimension vector

Let q be a quantity which can be represented with respect to the vector of dimension symbols $(\mathbf{E}_1, \dots, \mathbf{E}_m)$. The dimension $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_m)$ vector is unique if and only if $(\mathbf{E}_1, \dots, \mathbf{E}_m)$ is independent.

Definition 2.2.7: Dimension matrix

Let $(q_1, \dots, q_n)$ be a vector of quantities which can be represented with respect to the vector of independent dimension symbols $(\mathbf{E}_1, \dots, \mathbf{E}_m)$. That is, there are unique dimension vectors such that

$$[q_i] = \mathbf{E}_1^{\alpha_{i,1}} \cdots \mathbf{E}_m^{\alpha_{i,m}}$$

for $i = 1, \dots, n$. The matrix containing the dimensional vectors

$$A := \begin{pmatrix} \alpha_{1,1} & \cdots & \alpha_{1,m} \\ \vdots & & \vdots \\ \alpha_{n,1} & \cdots & \alpha_{n,m} \end{pmatrix} \in \mathbb{R}^{n \times m}$$

is called the **dimension matrix** of $(q_1, \dots, q_n)$ with respect to $(\mathbf{E}_1, \dots, \mathbf{E}_m)$. The rank of A, $\operatorname{rank}(A) =$ maximal number of independent columns, is referred to as the **dimension rank**.

Example 2.2.8

Revisiting Example 2.2.1, let us consider the vector of quantities (ρ, E) containing the density ρ and the energy E. Consulting Table 2.3, these can be represented with respect to the independent dimension symbols $(\mathbf{m}, \mathbf{L}, \mathbf{t})$, denoting the dimensions mass, length, and time:

$$\begin{aligned}[\rho] &= \boldsymbol{\rho} = \mathbf{m}^1 \cdot \mathbf{L}^{-3} \cdot \mathbf{t}^0 \\ [E] &= \mathbf{E} = \mathbf{m}^1 \cdot \mathbf{L}^2 \cdot \mathbf{t}^{-2}.\end{aligned}$$

Hence, the dimension matrix of the vector of quantities (ρ, E) with respect to $(\mathbf{m}, \mathbf{L}, \mathbf{t})$ is given by

$$A = \begin{pmatrix} 1 & -3 & 0 \\ 1 & 2 & -2 \end{pmatrix}$$

and the dimension rank is $\operatorname{rank}(A) = 2$.

Definition 2.2.9: Generalized monomials

A monomial is a polynomial containing only a single summand. In particular, if $f : \mathbb{R}^n \to \mathbb{R}$ is a monomial it has the form

$$f(x_1, \dots, x_n) = x_1^{\beta_1} \cdots x_n^{\beta_n},$$

where $\boldsymbol{\beta} = (\beta_1, \dots, \beta_n) \in \mathbb{N}^n$ is an exponential vector *only containing integer-valued* exponents. Henceforth, we refer to f as **generalized monomial** if the exponents are allowed to also be real-valued, that is, $\boldsymbol{\beta} \in \mathbb{R}^n$.

Lemma 2.2.10: The dimension of monomials

Let $(q_1, \dots, q_n)$ be an n-dimensional vector of quantities and let $f : \mathbb{R}^n \to \mathbb{R}$ be a generalized monomial with exponential vector $\boldsymbol{\beta} = (\beta_1, \dots, \beta_n)$. Moreover, let $(q_1, \dots, q_n)$ be represented with respect to the independent dimensional symbols $(\mathbf{E}_1, \dots, \mathbf{E}_m)$ with corresponding dimensional matrix $A \in \mathbb{R}^{n \times m}$. The quantity y, given by

$$y := f(q_1, \dots, q_n) = q_1^{\beta_1} \cdots q_n^{\beta_n},$$

has the dimensional vector $\boldsymbol{\beta} A$. In particular, y is dimensionless ($[y] = \mathbf{1}$) if and only if

$$\beta_1 \alpha_{1,m} + \cdots + \beta_n \alpha_{n,m} = 0$$

for all $m = 1, \dots, n$.

Proof. We just have to note that

$$\begin{aligned}[y] &= [q_1^{\beta_1} \cdots q_n^{\beta_n}] \\ &= (\mathbf{E}_1^{\alpha_{1,1}} \cdots \mathbf{E}_m^{\alpha_{1,m}})^{\beta_1} \cdots (\mathbf{E}_1^{\alpha_{n,1}} \cdots \mathbf{E}_m^{\alpha_{n,m}})^{\beta_n} \\ &= \mathbf{E}_1^{\alpha_{1,1}\beta_1 + \cdots + \alpha_{n,1}\beta_n} \cdots \mathbf{E}_m^{\alpha_{1,m}\beta_1 + \cdots + \alpha_{n,m}\beta_n}.\end{aligned}$$

Hence, denoting the dimension vector of y with respect to $(\mathbf{E}_1, \dots, \mathbf{E}_m)$ by $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_m)$, we immediately get

$$\boldsymbol{\gamma} = (\beta_1, \dots, \beta_n) \begin{pmatrix} \alpha_{1,1} & \cdots & \alpha_{1,m} \\ \vdots & & \vdots \\ \alpha_{n,1} & \cdots & \alpha_{n,m} \end{pmatrix} \in \mathbb{R}^{n \times m} = \boldsymbol{\beta} A$$

and therefore the assertion. □

Despite its simplicity, Lemma 2.2.10 comes with some strong consequences. In particular, the dimension of the dimension marix' kernel tells us how many dimensionless quantities can be constructed as generalized monomials of the physical quantities $(q_1, \dots, q_n)$. Moreover, this also tells us how many constants, which we need to be determined by actual measurements, will be in a physical law. This is summarized in the following theorem.

Theorem 2.2.11: The Pi theorem

Let $(q_1, \dots, q_n)$ be an n-dimensional vector of quantities and let $f : \mathbb{R}^n \to \mathbb{R}$ be a generalized monomial with exponential vector $\boldsymbol{\beta} = (\beta_1, \dots, \beta_n)$. Moreover, let $(q_1, \dots, q_n)$ be represented with respect to the independent dimensional symbols $(\mathbf{E}_1, \dots, \mathbf{E}_m)$ with corresponding dimensional matrix $A \in \mathbb{R}^{n \times m}$, $r := \operatorname{rank}(A)$, and $k := n - r$. Then, there exist

(a) r indices $(i_1, \dots, i_r)$,

(b) k dimensionless quantities $(\pi_1, \dots, \pi_k)$, and

(c) two generalized monomials $g : \mathbb{R}^r \to \mathbb{R}$ and $h : \mathbb{R}^k \to \mathbb{R}$

such that

$$f(q_1, \dots, q_n) = g(q_{i_1}, \dots, q_{i_r}) \cdot h(\pi_1, \dots, \pi_k) \tag{2.5}$$

holds.

Proof. Without loss of generality, let $0 < r < n$ and $0 < k < n$. The cases $r = 0$ $(k = n)$ and $r = n$ $(k = 0)$ can be treated analogously. The proof is done in three steps:

1. From linear algebra we know that for $A : \mathbb{R}^m \to \mathbb{R}^n$ the image $\operatorname{im}(A)$ is an r-dimensional subspace of $\mathbb{R}^n$. Thus, there are r indices $n_1, \dots, n_r$ such that

$$\operatorname{im}(A) = \operatorname{span}\{\boldsymbol{\alpha}_{n_1}, \dots, \boldsymbol{\alpha}_{n_r}\},$$

where $\boldsymbol{\alpha}_i = (\alpha_{i,1}, \dots, \alpha_{i,m})$ denotes the dimension vector of q_i with respect to $(\mathbf{E}_1, \dots, \mathbf{E}_m)$. Without loss of generality, let us assume that

$$n_1 = 1, \ \dots, \ n_r = r.$$

Otherwise, we could just change the order of the q_i.

2. Next, note that every dimension vector $\boldsymbol{\alpha}_{r+j}$, $j = 1, \dots, k$, can be written as a linear combination of $\boldsymbol{\alpha}_1, \dots, \boldsymbol{\alpha}_r$:

$$\boldsymbol{\alpha}_{r+j} = \gamma_{j,1}\boldsymbol{\alpha}_1 + \cdots + \gamma_{j,r}\boldsymbol{\alpha}_r$$

In particular, this implies

$$\begin{aligned} [q_{r+j}] &= \mathbf{E}_1^{\alpha_{r+j,1}} \cdots \mathbf{E}_m^{\alpha_{r+j,m}} \\ &= \mathbf{E}_1^{\gamma_{1,1}\alpha_{1,1} + \cdots + \gamma_{j,r}\alpha_{r,1}} \cdots \mathbf{E}_m^{\gamma_{1,1}\alpha_{1,m} + \cdots + \gamma_{j,r}\alpha_{r,m}} \\ &= (\mathbf{E}_1^{\alpha_{1,1}} \cdots \mathbf{E}_m^{\alpha_{1,m}})^{\gamma_{j,1}} \cdots (\mathbf{E}_1^{\alpha_{r,1}} \cdots \mathbf{E}_m^{\alpha_{r,m}})^{\gamma_{j,r}} \\ &= [q_1^{\gamma_{j,1}} \cdots q_r^{\gamma_{j,r}}]. \end{aligned}$$

Thus, the quantity π_j defined by

$$\pi_j = \frac{q_{r+j}}{q_1^{\gamma_{j,1}} \cdots q_r^{\gamma_{j,r}}}, \qquad j = 1, \dots, k,$$

is dimensionless. In fact, these are the k dimensionless quantities predicted by (b) in Theorem 2.2.11.

3. Moreover, the generalized monomials $g : \mathbb{R}^r \to \mathbb{R}$ and $h : \mathbb{R}^k \to \mathbb{R}$, described in (c) of Theorem 2.2.11, can be derived from the following observation. Since $f : \mathbb{R}^n \to \mathbb{R}$ is a generalized monomial it satisfies

$$f(q_1, \dots, q_r, q_{r+1}, \dots, q_n) = (q_1^{\beta_1} \cdots q_r^{\beta_r}) \cdot f(1, \dots, 1, q_{r+1}, \dots, q_n).$$

Moreover, we have (remember that $k = n - r$)

$$\begin{aligned} f(1, \dots, 1, \pi_1, \dots, \pi_n) &= \pi_1^{\beta_{r+1}} \cdots \pi^{\beta_{r+k}} \\ &= \frac{q_{r+1}^{\beta_{r+1}} \cdots q_{r+k}^{\beta_{r+k}}}{\left(q_1^{\gamma_{1,1}} \cdots q_r^{\gamma_{1,r}}\right)^{\beta_{r+1}} \cdots \left(q_1^{\gamma_{k,1}} \cdots q_r^{\gamma_{k,r}}\right)^{\beta_{r+k}}} \\ &= \frac{f(1, \dots, 1, q_{r+1}, \dots, q_n)}{q_1^{(\gamma_{1,1}\beta_{r+1} + \cdots + \gamma_{k,1}\beta_{r+k})} \cdots q_r^{(\gamma_{1,r}\beta_{r+1} + \cdots + \gamma_{k,r}\beta_{r+k})}}. \end{aligned}$$

Combining the two equations above results in

$$f(q_1, \dots, q_n) = \left(q_1^{\delta_1} \cdots q_r^{\delta_r}\right) \cdot f(1, \dots, 1, \pi_1, \dots, \pi_k),$$

where $\delta_i = \beta_i + (\gamma_{1,i}\beta_{r+1} + \cdots + \gamma_{k,i}\beta_{r+k})$ for $i = 1, \dots, r$. Finally, the generalized monomials $g : \mathbb{R}^r \to \mathbb{R}$ and $h : \mathbb{R}^k \to \mathbb{R}$ are therefore given by

$$\begin{aligned} g(q_1, \dots, q_r) &:= q_1^{\delta_1} \cdots q_r^{\delta_r}, \\ h(\pi_1, \dots, \pi_k) &:= f(1, \dots, 1, \pi_1, \dots, \pi k). \end{aligned}$$

The assertion follows immediately!

□

Remark 2.2.12

(a) The indices $(i_1, \dots, i_r)$ and the dimensionless quantities $(\pi_1, \dots, \pi_k)$ might not be unique. However, their numbers, rank(A) and $k = n - r$, are.

(b) For fixed $(i_1, \dots, i_r)$ and $(\pi_1, \dots, \pi_k)$, the generalized monomials $g : \mathbb{R}^r \to \mathbb{R}$ and $h : \mathbb{R}^k \to \mathbb{R}$ are uniquely determined up to a multiplicative constant.

The dimensional arguments occurring in the proof of Theorem 2.2.11 are often used in the natural engineering sciences. There, they are not just applied to general monomials f, however, but also to much more general physical laws. In particular, we note the following principle.

Theorem 2.2.13: Buckingham's principle (version 2 of the Pi theorem)

Let $(q_1, \dots, q_n)$ be an n-dimensional vector of quantities and let $f : \mathbb{R}^n \to \mathbb{R}$ be some function. Moreover, let $(q_1, \dots, q_n)$ be represented with respect to the independent dimensional symbols

$(\mathbf{E}_1, \dots, \mathbf{E}_m)$ with corresponding dimensional matrix $A \in \mathbb{R}^{n \times m}$, $r := \operatorname{rank}(A)$, and $k := n - r$. Then, there exist

(a) k dimensionless quantities $(\pi_1, \dots, \pi_k)$ that can be formed from $(q_1, \dots, q_n)$ and

(b) a function $F : \mathbb{R}^k \to \mathbb{R}$

such that the unit free physical law

$$f(q_1, \dots, q_n) = 0,$$

that relates the dimensional quantities $q_1, \dots, q_n$, is equivalent to the equation

$$F(\pi_1, \dots, \pi_k) = 0,$$

expressed only in terms of the dimensionless quantities $\pi_1, \dots, \pi_k$.

Proof. The proof of Theorem 2.2.13 does not only involve mathematical but also physical arguments and is omitted. □

Like every good physical principle, Buckingham's principle is both, easy to understand as well as powerful. To demonstrate this, let us consider an example.

Example 2.2.14: Heat transfer

Let us consider a very thin and one-sided endless metal bar with constant temperature zero. At time $t_0 = 0$ an amount of temperature U is assumed to be concentrated at the finite end of the bar, let's say at $x = 0$ with $x \in [0, \infty)$. You could imagine the bar being pushed against a heater. Then, the problem is to determine the temperature u as a function of the distance to the source x, time t, the source's temperature U, and a constant κ describing the thermal diffusivity of the metal bar. All involved quantities can be found in Table 2.4. Of course, the problem could be formulated as a partial differential equation, namely the heat equation $\partial_t u - \kappa \partial_x^2 u = 0$ with zero initial condition and boundary condition $u = U$ at $x = 0$. Yet, we want to use this example to demonstrate that already dimensional arguments, in particular the Pi theorem, can provide us with surprisingly deep insights into the above described physical process. We conjecture a physical law of the form

$$f(u, t, x, U, \kappa) = 0, \tag{2.6}$$

which relates the five physical quantities u, t, U, and κ. Next, for our dimensional analysis of (2.6), we proceed in three steps:

1. Find a minimal set of independent dimensions $(\mathbf{E}_1, \dots, \mathbf{E}_m)$ by which the physical quantity can be represented.

2. Determine a set of dimensionless quantities $\pi_1, \dots, \pi_k$ predicted by Buckingham's principle.

3. Go over to an equation $F(\pi_1, \dots, \pi_k) = 0$, that relates the dimensionless quantities.

<u>On 1.</u> Consulting Table 2.4, a suitable choice are the three dimensions temperature $(\mathbf{T})$, time $(\mathbf{t})$, and length $(\mathbf{L})$. For this choice, $(\mathbf{E}_1, \dots, \mathbf{E}_m) = (\mathbf{T}, \mathbf{t}, \mathbf{L})$, we have

$$[u] = \mathbf{T}, \quad [t] = \mathbf{t}, \quad [x] = \mathbf{L}, \quad [U] = \mathbf{T}, \quad [\kappa] = \mathbf{t}^{-1}\mathbf{L}^2,$$

yielding the corresponding dimension matrix

$$A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & -1 & 2 \end{pmatrix}$$

with $\operatorname{rank}(A) = 3$ and $k = n - r = 2$.

On 2. Buckingham's principle (Theorem 2.2.13) tells us that $k = 2$ dimensionless quantities (π_1, π_2) can be formed from the $n = 5$ physical quantities (u, t, x, U, κ). In general, such a dimensionless quantity π has the form

$$\pi = u^{\alpha_1} t^{\alpha_2} x^{\alpha_3} U^{\alpha_4} \kappa^{\alpha_5}$$

and satisfies $[\pi] = 1$. This observation yields

$$\begin{aligned} \mathbf{1} &= [u^{\alpha_1} t^{\alpha_2} x^{\alpha_3} U^{\alpha_4} \kappa^{\alpha_5}] \\ &= \mathbf{T}^{\alpha_1} \mathbf{t}^{\alpha_2} \mathbf{L}^{\alpha_3} \mathbf{T}^{\alpha_4} (\mathbf{t}^{-1} \mathbf{L}^2)^{\alpha_5} \\ &= \mathbf{T}^{\alpha_1 + \alpha_4} \mathbf{t}^{\alpha_2 - \alpha_5} \mathbf{L}^{\alpha_3 - 2\alpha_5}. \end{aligned}$$

Next, since the exponents must vanish, we end up with a system of linear equations,

$$\begin{aligned} \alpha_1 + \alpha_4 &= 0, \\ \alpha_2 - \alpha_5 &= 0, \\ \alpha_3 + 2\alpha_5 &= 0, \end{aligned}$$

with general solution

$$\alpha_1 = -\alpha_4, \quad \alpha_2 = \alpha_5, \quad \alpha_3 = -2\alpha_5, \quad \alpha_4, \alpha_5 \in \mathbb{R}.$$

Two linear independent solutions are given by

$$\begin{aligned} \alpha &= (1, 0, 0, -1, 0) && (\text{for } \alpha_4 = -1,\ \alpha_5 = 0), \\ \alpha &= (0, -1, 2, 0, -1) && (\text{for } \alpha_4 = 0,\ \alpha_5 = -1), \end{aligned}$$

and respectively result in the dimensionless quantities

$$\pi_1 = \frac{u}{U} \quad \text{and} \quad \pi_2 = \frac{x^2}{t\kappa}.$$

On 3. Finally, Buckingham's principle (Theorem 2.2.13) also tells us that the original physical law (2.6) is equivalent to an equation

$$F(\pi_1, \pi_2) = 0,$$

which relates the two dimensionless quantities π_1 and π_2. Hence, solving for π_1 (which contains u), yields

$$\pi_1 = g(\pi_2),$$

for some function g, and therefore

$$u = U g\left(\frac{x^2}{t\kappa}\right).$$

In particular, without solving any differential equations, we are able via dimension analysis to argue that the temperature u depends linearly on the source U.

quantity	dimension	symbol
u (temperature)	temperature	$\mathbf{T}$
t (time)	time	$\mathbf{t}$
x (distance to source)	length	$\mathbf{L}$
V (source temperature)	temperature	$\mathbf{T}$
κ (thermal diffusivity)	thermal diffusivity	$\boldsymbol{\kappa} = \frac{\mathbf{L}^2}{\mathbf{t}}$

Table 2.4: The quantities and corresponding dimensions used in Example 2.2.14

2.3 Scaling

So far, dimensional analysis — especially Buckingham's principle — allowed us to reduce mathematical models, potentially involving many physical quantities $q_1, \dots, q_n$ to an equivalent model relating a usually smaller number of dimensionless quantities $\pi_1, \dots, \pi_k$. Essentially, this tells us which quantities to use in a model. Scaling (also referred to as non-dimensionalization) is of a similar flavor. The goal is to find appropriate dimensionless scales for the involved variables.

Example 2.3.1: Motivation

Suppose that time t is a quantity in a given model. If this model describes, for instance, the motion of a glacier, clearly the unit of seconds is too fast. Significant changes in the glacier could not be observed on the order of seconds. On the other hand, if the problem involved a nuclear reaction, the unit of seconds would be too slow. This time, all of the important actions would be over before the first second has ticked.

Evidently, every problem has an intrinsic time **scale**, or **characteristic quantity** t_c, which is the shortest time for which discernible changes can be observed in the physical quantities. Some processes might even have multiple time scales. Once a characteristic time t_c has been identified, at least for a part of the process, a new dimensionless variable $\bar{t}$ can be defined by

$$\bar{t} := \frac{t}{t_c}.$$

If t_c is chosen correctly, the dimensionless time $\bar{t}$ neither is too large nor too small. After characteristic quantities have been chosen the model can then be reformulated in terms of the new dimensionless quantities. The result will be a model in dimensionless form, where all the variables and parameters are dimensionless and roughly of the same magnitude. Moreover, the number of parameters is usually reduced. This process is usually called **scaling** or **non-dimensionalization**.

Remark 2.3.2

Scaling is not a strict mathematical theory but rather a technique, which can be learned from exercise. Besides technical skills (application of integral and differential transformations), it is important to have a physical understanding of the intrinsic scales of a problem, e. g. characteristic times, lengths, and so on.

In what follows, we consider two examples (population growth as well as the projectile problem) which are supposed to demonstrate the idea and steps behind scaling in greater detail. At the end of this chapter, you should try to apply scaling to some other problems by yourself (see the corresponding exercises).

Example 2.3.3: Population growth

Let $p = p(t)$ be the population of an animal species located at a fixed region at time t. The simplest model to describe population growth is the classical **Malthus model**

$$\frac{\mathrm{d}p}{\mathrm{d}t} = rp, \quad p(0) = p_0;$$

see [Mal72, MWJ92]. Here, $r > 0$ is a parameter called the *growth rate*, given in dimensions of inverse-time. In particular, the Malthus model indicates $p(t) = p_0 e^{rt}$, that is, exponential growth. Unfortunately, this model does not capture the important effect of *competition*. As a population grows, intraspecific competition for food, living space, and natural resources limit the growth. Yet, this is reflected in the **logistics model**

$$\frac{\mathrm{d}p}{\mathrm{d}t} = rp\left(1 - \frac{p}{K}\right), \quad p(0) = p_0, \tag{2.7}$$

which was introduced in a series of three papers by the Belgian mathematician Verhulst between 1838 and 1847; see [Cra04]. Here, $K > 0$ is an additional parameter referred to as the *carrying capacity*. This is the number of individuals that the ecosystem can sustain. Note that the logistics model 2.7 has a total number of two variables (t and p) and three parameters (p_0, r, and K), also see Table 2.5. We now demonstrate how scaling can be used to reduce the number of parameters in a model. This is done by introducing new dimensionless variables for time and population, formed from the parameters p_0, r, and K. Of these, only r contains dimension of time ($[r] = \mathbf{t}^{-1}$). Hence, we use $t_c = r^{-1}$ as the characteristic quality for time and introduce

$$\bar{t} = rt.$$

For the characteristic quantity of the population, p_c, there are two choices, K as well as p_0. Either will work, but here we choose $p_c = K$, yielding

$$\overline{p} = \frac{p}{K}.$$

Reformulating (2.7) with respect to the new dimensionless variables $\bar{t}$ and $\overline{p}$, we get

$$\frac{\mathrm{d}\overline{p}}{\mathrm{d}\bar{t}} = \overline{p}(1 - \overline{p}), \quad \overline{p}(0) = \alpha, \tag{2.8}$$

with $\alpha := p_0/K$. It should be stressed heavily that the scaled model (2.8) only contains a single parameter α. This is a significant simplification over the original logistics model (2.7), which relied on three parameters. Finally, (2.8) can be solved by separating variables to obtain

$$\overline{p}(\bar{t}) = \frac{\alpha}{\alpha + (1-\alpha)e^{-\bar{t}}}.$$

It is clear that $\overline{p}(\bar{t}) \to 1$ for $\bar{t} \to \infty$ if $\alpha > 0$, confirming that the limiting population p is equal to the carrying capacity K.

The following example describes the motion of a projectile thrust vertically upward from the surface of the earth. It was first pointed out by Lin and Segel [LS88] and demonstrates the importance of choosing correct scales (characteristic quantities), especially when it is desired to make simplifications

variable	meaning	dimension
t	time	$\mathbf{t}$
p	population	$\mathbf{1}$
parameter	meaning	dimension
p_0	initial population	$\mathbf{1}$
r	growth rate	$\mathbf{t}^{-1}$
K	carrying capacity	$\mathbf{1}$

Table 2.5: Variables, parameters, and corresponding dimensions used in Example 2.3.3

by neglecting small quantities.

Example 2.3.4: The projectile problem

At time $t = 0$ on the surface of the earth, with radius R and mass M, an object of mass m is given a vertical upward velocity of magnitude V. Here, we want to determine the height $h = h(t)$ above the earth's surface that the object reaches at time t. Forces on the object are the gravitational force and the force due to air resistance. Yet, we assume that the force due to air resistance can be neglected in this particular problem. Then, Newton's second law provides us with the mathematical model

$$\frac{\mathrm{d}^2 h}{\mathrm{d}t^2} = -g\frac{R^2}{(h+R)^2}, \quad h(0) = 0, \quad \frac{\mathrm{d}h}{\mathrm{d}t}(0) = V, \tag{2.9}$$

where g denotes the local acceleration of free fall and is assumed to be constant (since $h/R \approx 0$). All involved quantities can be found in Table 2.6. Next, to scale model (2.9), let us introduce new dimensionless variables for time and the object height:

$$\bar{t} = \frac{t}{t_c} \quad \text{and} \quad \bar{h} = \frac{h}{h_c}.$$

Of course, the question remains of how to choose the characteristic quantities t_c and h_c. We start by noting that t_c and h_c are formed by taking combinations of the parameters in the problem, i. e.

$$t_c = R^{\alpha_1} V^{\alpha_2} g^{\alpha_3}, \quad h_c = R^{\beta_1} V^{\beta_2} g^{\beta_3}.$$

At the same time, for $\bar{t}$ and $\bar{h}$ to be dimensionless,

$$[t] = [t_c], \quad [h] = [h_c]$$

needs to hold. This yields

$$\begin{aligned} t &= \mathbf{L}^{\alpha_1}(\mathbf{L}\mathbf{t}^{-1})^{\alpha_2}(\mathbf{L}\mathbf{t}^{-2})^{\alpha_3} = \mathbf{L}^{\alpha_1+\alpha_2+\alpha_3}\mathbf{t}^{-\alpha_2-2\alpha_3}, \\ h &= \mathbf{L}^{\beta_1}(\mathbf{L}\mathbf{t}^{-1})^{\beta_2}(\mathbf{L}\mathbf{t}^{-2})^{\beta_3} = \mathbf{L}^{\beta_1+\beta_2+\beta_3}\mathbf{t}^{-\beta_2-2\beta_3} \end{aligned}$$

and therefore the two systems of linear equations

$$\begin{aligned} \alpha_1 + \alpha_2 + \alpha_3 &= 0, \\ -\alpha_2 - 2\alpha_3 &= 1, \end{aligned}$$

and

$$\begin{aligned} \beta_1 + \beta_2 + \beta_3 &= 1, \\ -\beta_2 - 2\beta_3 &= 0. \end{aligned}$$

The general solutions are given by

$$\begin{aligned} \alpha &= (1+\alpha_3, -1-2\alpha_3, \alpha_3), \quad \alpha_3 \in \mathbb{R}, \\ \beta &= (1+\beta_3, -2\beta_3, \beta_3), \quad \beta_3 \in \mathbb{R}. \end{aligned}$$

Unfortunately, not all choices for α and β will result in equal success later on. This is where we also need to weight in some physical intuition. To demonstrate this, let us consider different choices:

$$\begin{aligned} \bar{t} &= \frac{t}{RV^{-1}}, \quad \bar{h} = \frac{h}{R}, \quad (\text{for } \alpha_3 = 0,\ \beta_3 = 0) \\ \bar{t} &= \frac{t}{\sqrt{Rg^{-1}}}, \quad \bar{h} = \frac{h}{R}, \quad (\text{for } \alpha_3 = -0.5,\ \beta_3 = 0) \\ \bar{t} &= \frac{t}{Vg^{-1}}, \quad \bar{h} = \frac{h}{V^2 g^{-1}}, \quad (\text{for } \alpha_3 = -1,\ \beta_3 = -1), \end{aligned}$$

which respectively yield the scaled models

$$\varepsilon \frac{\mathrm{d}^2\bar{h}}{\mathrm{d}\bar{t}^2} = -\frac{1}{(1+\bar{h})^2}, \quad \bar{h}(0) = 0, \quad \frac{\mathrm{d}\bar{h}}{\mathrm{d}\bar{t}}(0) = 1, \tag{2.10}$$

$$\frac{\mathrm{d}^2\bar{h}}{\mathrm{d}\bar{t}^2} = -\frac{1}{(1+\bar{h})^2}, \quad \bar{h}(0) = 0, \quad \frac{\mathrm{d}\bar{h}}{\mathrm{d}\bar{t}}(0) = \sqrt{\varepsilon}, \tag{2.11}$$

$$\frac{\mathrm{d}^2\bar{h}}{\mathrm{d}\bar{t}^2} = -\frac{1}{(1+\varepsilon\bar{h})^2}, \quad \bar{h}(0) = 0, \quad \frac{\mathrm{d}\bar{h}}{\mathrm{d}\bar{t}}(0) = 1. \tag{2.12}$$

Here, ε is a dimensionless parameter defined by

$$\varepsilon = \frac{V^2}{gR}.$$

To illustrate how a clumsy choice of the characteristic quatities t_c and h_c can yield to difficulties, let us modify the scaled models (2.10), (2.11), and (2.12) in the case that ε is known to be a small parameter, which can be neglected. Then, the scaled model (2.10) becomes

$$(1+\bar{h})^{-2} = 0, \quad \bar{h}(0) = 0, \quad \frac{\mathrm{d}\bar{h}}{\mathrm{d}\bar{t}}(0) = 1,$$

which has no solution. At the same time, the scaled model (2.11) becomes

$$\frac{\mathrm{d}^2\bar{h}}{\mathrm{d}\bar{t}^2} = -\frac{1}{(1+\bar{h})^2}, \quad \bar{h}(0) = 0, \quad \frac{\mathrm{d}\bar{h}}{\mathrm{d}\bar{t}}(0) = 0.$$

This model, in fact, has a solution. Unfortunately this solution cannot be considered as physical reasonable, since $\bar{h}(\bar{t}) \leq 0$. Hence, in the scaled models (2.10) and (2.11) it is not possible to neglect small parameters, which is unfortunate, since this kind of technique is common practice in making approximations in applied problems. What went wrong is that (2.10) and (2.11) represent incorrectly scaled models. In these, terms that may appear small may not in fact be small. For instance, in the term $\varepsilon\frac{\mathrm{d}^2\overline{h}}{\mathrm{d}\overline{t}^2}$, the parameter ε may be small but $\frac{\mathrm{d}^2\overline{h}}{\mathrm{d}\overline{t}^2}$ may be large, and therefore the whole term may not be negligible. If, on the other hand, the term $\varepsilon\overline{h}$ is neglected in the last scaled model (2.12), we get

$$\frac{\mathrm{d}^2\overline{h}}{\mathrm{d}\overline{t}^2} = -1, \quad \overline{h}(0) = 0, \quad \frac{\mathrm{d}\overline{h}}{\mathrm{d}\overline{t}}(0) = 1,$$

and therefore

$$\overline{h}(\overline{t}) = \overline{t} - \frac{\overline{t}}{2},$$

or

$$h(t) = -\frac{1}{2}gt^2 + VT.$$

Hence, in fact, we have obtained an approximate solution that is consistent with our experience with falling bodies close to the earth. In this case, we were able to neglect the small term ε and obtain a valid approximation because the scaling is correct. Actually, the choices

$$t_c = Vg^{-1}, \quad h_c = V^2g^{-1},$$

can be argued physically. If V is small, then the body will be acted on by a constant gravitational field. Thus, launched with speed V, it will uniformly decelerate and reach its maximum height at time V/g, which therefore is the characteristic time. Moreover, the body will travel a distance of about V/g times its average velocity $V/2$. Hence, V^2/g is revealed as a good choice for the characteristic height. In contrast, measuring the height relative to the radius of the earth, as in (2.10) and (2.11), is not a good choice!

variable	meaning	dimension
t	time	$\mathbf{t}$
h	object height	$\mathbf{L}$
parameter	meaning	dimension
R	earth radius	$\mathbf{L}$
V	initial velocity	$\mathbf{Lt}^{-1}$
g	earth acceleration	$\mathbf{Lt}^{-2}$

Table 2.6: Variables, parameters, and corresponding dimensions used in Example 2.3.4

In general, if a correct scaling is chosen, terms in the equation that appear small are indeed small and may therefore be safely neglected.

Chapter 3

Asymptotic analysis

In the latter chapters of this course we will often try to find approximate solutions to otherwise unsolvable mathematical models. Yet, before digging deeper into these techniques, it is convenient to collect some fundamental definitions and results from asymptotic analysis first. Asymptotic analysis is concerned with the behavior of functions $f(\varepsilon)$ for $\varepsilon \to 0$.

3.1 The Bachmann-Landau notation

The heart of asymptotic analysis is the following definition usually referred to as **Bachmann-Landau notation** (also called asymptotic notation or big/little O notation).

Definition 3.1.1: Bachmann-Landau notation

Let $(X, \|\cdot\|)$ be a normed linear space (vector space) over $\mathbb{R}$, $\varepsilon > 0$, $x_\varepsilon = x(\varepsilon)$ be an X-valued function, and $p_\varepsilon = p(\varepsilon)$ be a nonnegative real-valued function. We write

(a) $x_\varepsilon = \mathcal{O}_{\|\cdot\|}(p_\varepsilon) :\iff \limsup_{\varepsilon\to 0} q_\varepsilon < \infty$,

(b) $x_\varepsilon = \mathbf{o}_{\|\cdot\|}(p_\varepsilon) :\iff \lim_{\varepsilon\to 0} q_\varepsilon = 0$,

where $q_\varepsilon := \frac{\|x_\varepsilon\|}{p_\varepsilon}$. Sometimes, p_ε is referred to as a **gauge function**. When it is clear from the context which norm we are using, we just write $x_\varepsilon = \mathcal{O}(p_\varepsilon)$ and $x_\varepsilon = \mathbf{o}(p_\varepsilon)$ instead of $\mathcal{O}_{\|\cdot\|}(p_\varepsilon)$ and $\mathbf{o}_{\|\cdot\|}(p_\varepsilon)$.

Strictly speaking, $x_\varepsilon = x(\varepsilon)$ is a fixed function value of the function

$$x : \mathbb{R}^+ \to X, \quad \varepsilon \mapsto x_\varepsilon$$

for the argument $\varepsilon \in \mathbb{R}^+$. Yet, it is a common convention in asymptotic analysis to also denote the whole function x —and not just specific function values— by x_ε. We follow this convention here.

Remark 3.1.2

Note that if X is a finite-dimensional linear space all norms on X are equivalent. Hence, Definition 3.1.1 is independent of the norm $\|\cdot\|$ in this case. However, if X is not a finite-dimensional linear space there will be norms which are not equivalent to each other.[a] Thus, in this case, Definition 3.1.1 *does depend* on the chosen norm $\|\cdot\|$.

[a] In fact, it is a well-known result in functional analysis that a linear space has finite dimensions if and only if all norms on this space are equivalent.

The following lemma provides some useful calculation rules for the Bachmann-Landau notation.

Lemma 3.1.3: Calculation rules

Let $(X, \|\cdot\|)$ be a normed linear space over $\mathbb{R}$. Then, the following calculation rules hold:

(a.1) $\|x(\cdot)\| \equiv \text{const} \implies x_\varepsilon = \mathcal{O}(1)$

(a.2) $\|x(\cdot)\| \equiv \text{const} \implies x_\varepsilon = \mathbf{o}(\varepsilon^\alpha) \quad \forall \alpha < 0$

(b) $x_\varepsilon = \mathcal{O}(p_\varepsilon),\ C \in \mathbb{R} \implies Cx_\varepsilon = \mathcal{O}(p_\varepsilon)$

(c) $x_\varepsilon = \mathcal{O}(p_\varepsilon),\ y_\varepsilon = \mathcal{O}(r_\varepsilon) \implies x_\varepsilon + y_\varepsilon = \mathcal{O}(\max\{p_\varepsilon, r_\varepsilon\})$

(d) $x_\varepsilon = \mathcal{O}(p_\varepsilon),\ p_\varepsilon \leq r_\varepsilon \implies x_\varepsilon = \mathcal{O}(r_\varepsilon)$

(e) $y_\varepsilon = \mathcal{O}(r_\varepsilon),\ \|x_\varepsilon\| \leq \|y_\varepsilon\|\ \forall \varepsilon \geq 0 \implies x_\varepsilon = \mathcal{O}(r_\varepsilon)$

Moreover, if $(X, \langle \cdot, \cdot \rangle)$ is an inner product space, we have

(f) $x_\varepsilon = \mathcal{O}(p_\varepsilon),\ y_\varepsilon = \mathcal{O}(r_\varepsilon) \implies \langle x_\varepsilon, y_\varepsilon \rangle = \mathcal{O}(p_\varepsilon r_\varepsilon)$

in the sense of the euclidean space $(\mathbb{R}, \|\cdot\|)$. All of the above calculation rules, except for (a), also hold true if we replace big $\mathcal{O}$ by little $\mathbf{o}$.

Proof. All of the above statements are more or less trivial.

$$\begin{aligned}
&\text{(a.1)} && \|x(\cdot)\| \equiv \text{const} \\
& && \implies \limsup_{\varepsilon \to 0} \|x_\varepsilon\| = \text{const} \\
& && \implies x_\varepsilon = \mathcal{O}(1) \\
&\text{(a.2)} && \|x(\cdot)\| \equiv \text{const},\ \alpha < 0 \\
& && \implies \lim_{\varepsilon \to 0} \frac{\|x_\varepsilon\|}{\epsilon^\alpha} = \lim_{\varepsilon \to 0} \epsilon^{-\alpha} \text{const} = 0 \\
& && \implies x_\varepsilon = \mathbf{o}(\varepsilon^\alpha) \quad \forall \alpha < 0 \\
&\text{(b.1)} && x_\varepsilon = \mathcal{O}(p_\varepsilon),\ C \in \mathbb{R} \\
& && \implies \limsup_{\varepsilon \to 0} \frac{\|Cx_\varepsilon\|}{p_\varepsilon} = |C| \limsup_{\varepsilon \to 0} \frac{\|x_\varepsilon\|}{p_\varepsilon} < \infty \\
& && \implies Cx_\varepsilon = \mathcal{O}(p_\varepsilon) \\
&\text{(b.2)} && x_\varepsilon = \mathbf{o}(p_\varepsilon),\ C \in \mathbb{R} \\
& && \implies \lim_{\varepsilon \to 0} \frac{\|Cx_\varepsilon\|}{p_\varepsilon} = |C| \lim_{\varepsilon \to 0} \frac{\|x_\varepsilon\|}{p_\varepsilon} = 0 \\
& && \implies Cx_\varepsilon = \mathbf{o}(p_\varepsilon) \\
&\text{(c.1)} && x_\varepsilon = \mathcal{O}(p_\varepsilon),\ y_\varepsilon = \mathcal{O}(r_\varepsilon) \\
& && \implies \limsup_{\varepsilon \to 0} \frac{\|x_\varepsilon + y_\varepsilon\|}{\max\{p_\varepsilon, r_\varepsilon\}} \leq \left(\limsup_{\varepsilon \to 0} \frac{\|x_\varepsilon\|}{p_\varepsilon} \right) + \left(\limsup_{\varepsilon \to 0} \frac{\|y_\varepsilon\|}{r_\varepsilon} \right) < \infty \\
& && \implies x_\varepsilon + y_\varepsilon = \mathcal{O}(\max\{p_\varepsilon, r_\varepsilon\}) \\
&\text{(c.2)} && x_\varepsilon = \mathbf{o}(p_\varepsilon),\ y_\varepsilon = \mathbf{o}(r_\varepsilon) \\
& && \implies 0 \leq \lim_{\varepsilon \to 0} \frac{\|x_\varepsilon + y_\varepsilon\|}{\max\{p_\varepsilon, r_\varepsilon\}} \leq \left(\lim_{\varepsilon \to 0} \frac{\|x_\varepsilon\|}{p_\varepsilon} \right) + \left(\lim_{\varepsilon \to 0} \frac{\|y_\varepsilon\|}{r_\varepsilon} \right) = 0
\end{aligned}$$

$$
\begin{aligned}
& && \Longrightarrow \lim_{\varepsilon \to 0} \frac{\|x_\varepsilon + y_\varepsilon\|}{\max\{p_\varepsilon, r_\varepsilon\}} = 0 \\
& && \Longrightarrow x_\varepsilon + y_\varepsilon = \mathcal{O}(\max\{p_\varepsilon, r_\varepsilon\}) \\
&\text{(d.1)} && \qquad x_\varepsilon = \mathcal{O}(p_\varepsilon),\ p_\varepsilon \le r_\varepsilon \\
& && \Longrightarrow \limsup_{\varepsilon \to 0} \frac{\|x_\varepsilon\|}{r_\varepsilon} \le \limsup_{\varepsilon \to 0} \frac{\|x_\varepsilon\|}{p_\varepsilon} < \infty \\
& && \Longrightarrow x_\varepsilon = \mathcal{O}(r_\varepsilon) \\
&\text{(d.2)} && \qquad x_\varepsilon = \mathbf{o}(p_\varepsilon),\ p_\varepsilon \le r_\varepsilon \\
& && \Longrightarrow 0 \le \lim_{\varepsilon \to 0} \frac{\|x_\varepsilon\|}{r_\varepsilon} \le \limsup_{\varepsilon \to 0} \frac{\|x_\varepsilon\|}{p_\varepsilon} < \infty \\
& && \Longrightarrow x_\varepsilon = \mathbf{o}(r_\varepsilon) \\
&\text{(e.1)} && \qquad y_\varepsilon = \mathcal{O}(r_\varepsilon),\ \|x_\varepsilon\| \le \|y_\varepsilon\|\ \forall \varepsilon \ge 0 \\
& && \Longrightarrow \limsup_{\varepsilon \to 0} \frac{\|x_\varepsilon\|}{r_\varepsilon} \le \limsup_{\varepsilon \to 0} \frac{\|y_\varepsilon\|}{r_\varepsilon} < \infty \\
& && \Longrightarrow x_\varepsilon = \mathcal{O}(r_\varepsilon) \\
&\text{(e.2)} && \qquad y_\varepsilon = \mathbf{o}(r_\varepsilon),\ \|x_\varepsilon\| \le \|y_\varepsilon\|\ \forall \varepsilon \ge 0 \\
& && \Longrightarrow 0 \le \lim_{\varepsilon \to 0} \frac{\|x_\varepsilon\|}{r_\varepsilon} \le \lim_{\varepsilon \to 0} \frac{\|y_\varepsilon\|}{r_\varepsilon} = 0 \\
& && \Longrightarrow x_\varepsilon = \mathcal{O}(r_\varepsilon) \\
&\text{(f.1)} && \qquad x_\varepsilon = \mathcal{O}(p_\varepsilon),\ y_\varepsilon = \mathcal{O}(r_\varepsilon) \\
& && \Longrightarrow \limsup_{\varepsilon \to 0} \frac{|\langle x_\varepsilon, y_\varepsilon \rangle|}{p_\varepsilon r_\varepsilon} \le \left(\limsup_{\varepsilon \to 0} \frac{\|x_\varepsilon\|}{p_\varepsilon} \right) \left(\limsup_{\varepsilon \to 0} \frac{\|y_\varepsilon\|}{r_\varepsilon} \right) < \infty \\
& && \Longrightarrow \langle x_\varepsilon, y_\varepsilon \rangle = \mathcal{O}(p_\varepsilon r_\varepsilon) \\
&\text{(f.2)} && \qquad x_\varepsilon = \mathbf{o}(p_\varepsilon),\ y_\varepsilon = \mathbf{o}(r_\varepsilon) \\
& && \Longrightarrow 0 \le \lim_{\varepsilon \to 0} \frac{|\langle x_\varepsilon, y_\varepsilon \rangle|}{p_\varepsilon r_\varepsilon} \le \left(\lim_{\varepsilon \to 0} \frac{\|x_\varepsilon\|}{p_\varepsilon} \right) \left(\lim_{\varepsilon \to 0} \frac{\|y_\varepsilon\|}{r_\varepsilon} \right) < \infty \\
& && \Longrightarrow \langle x_\varepsilon, y_\varepsilon \rangle = \mathbf{o}(p_\varepsilon r_\varepsilon)
\end{aligned}
$$

□

Example 3.1.4

Let $X = \mathbb{R}$ and $\|\cdot\| = |\cdot|$.

(a) First, we consider the case $x_\varepsilon := \varepsilon^2 \ln(\varepsilon)$ and verify that

$$\varepsilon^2 \ln(\varepsilon) = \mathbf{o}(\varepsilon).$$

This can be shown by de L'Hospital's rule:

$$\lim_{\varepsilon \to 0} \frac{|\varepsilon^2 \ln(\varepsilon)|}{\varepsilon} = -\lim_{\varepsilon \to 0} \frac{\ln(\varepsilon)}{\varepsilon^{-1}} \overset{[\frac{\infty}{\infty}]}{=} \lim_{\varepsilon \to 0} \frac{\varepsilon^{-1}}{\varepsilon^{-2}} = \lim_{\varepsilon \to 0} \varepsilon = 0$$

(b) Sometimes also the mean value theorem is of practical use in asymptotic analysis. This is demonstrated by considering $x_\varepsilon = \sin(\varepsilon)$. This time, we want to verify that

$$\sin(\varepsilon) = \mathcal{O}(\varepsilon).$$

Here, the mean value theorem provides us with an ξ between 0 and ε such that

$$\frac{\sin(\varepsilon) - \sin(0)}{\varepsilon - 0} = \cos(\xi).$$

Thus, we have

$$\limsup_{\varepsilon \to 0} \frac{|\sin(\varepsilon)|}{\varepsilon} = \limsup_{\varepsilon \to 0} \cos(\xi) = 1 < \infty,$$

since $\xi \to 0$ for $\varepsilon \to 0$. At the same time, we have $\sin(\varepsilon) \neq \mathbf{o}(\varepsilon)$ but $\sin(\varepsilon) = \mathbf{o}(\varepsilon^\alpha)$ for all $\alpha < 1$.

Note that in Example 3.1.4(a) we have shown that $\varepsilon^2 \ln(\varepsilon) = \mathbf{o}(\varepsilon)$. Obviously, $\varepsilon^2 \ln(\varepsilon) = \mathcal{O}(\varepsilon)$ holds as well. In fact, this observation is true in general.

Lemma 3.1.5: $\mathbf{o}(p_\varepsilon) \subset \mathcal{O}(p_\varepsilon)$

Let $(X, \| \cdot \|)$ be a normed linear space, $\varepsilon > 0$, x_ε be an X-valued function, and $p_\varepsilon \geq 0$ be a real-valued gauge function. Then,

$$x_\varepsilon = \mathbf{o}(p_\varepsilon) \implies x_\varepsilon = \mathcal{O}(p_\varepsilon)$$

holds.

At the same time, the reversed statement ($\mathcal{O}(p_\varepsilon) \subset \mathbf{o}(p_\varepsilon)$) is not true, in general. This has already been demonstrated in Example 3.1.4(b). We end this excursion into the world of asymptotic analysis by an example in which, for fixed ε, x_ε is a function itself and not just a real number.

Example 3.1.6

Let $X = C([0,1])$ and $\|x\| := \max_{0 \leq t \leq 1} |x(t)|$, and

$$x_\varepsilon(t) := \exp\left(-\frac{t}{\varepsilon}\right), \quad \varepsilon > 0.$$

In particular, we have

$$x_\varepsilon = \mathcal{O}(1) \quad \text{and} \quad x_\varepsilon \neq \mathbf{o}(1).$$

Again, this demonstrates that $x_\varepsilon = \mathbf{o}(p_\varepsilon)$ is a stronger property than $x_\varepsilon = \mathcal{O}(p_\varepsilon)$.

Chapter 4

Perturbation Methods

Often, mathematical models cannot be solved in exact form and their solution has therefore to be approximated by a numerical method. Perturbation methods are one such approximation technique, in the case that the model includes very small terms. Such terms arise when the underlying physical process has only small effects. For instance, let us consider an ordinary differential equation

$$F(t,y,y',y'',\varepsilon)=0, \quad t\in I, \tag{4.1}$$

where t denotes time and I the time interval, $y=y(t)$ the dependent variable, and ε a small parameter. Perturbation methods try to find an approximation to the original problem, in this case (4.1), by starting from the exact solution of a simplified version of the problem, in this case

$$F(t,y,y',y'',0)=0, \quad t\in I. \tag{4.2}$$

Usually, we refer to (4.1) as the **perturbed problem** and to (4.2) as the **unperturbed problem**. The approximation of the perturbed problem is then expressed in terms of a formal power series in the small parameter ε,

$$p_\varepsilon(t)=\sum_{k=0}^{\infty}\varepsilon^k y_k(t), \tag{4.3}$$

called **perturbation series**. Thereby, the leading term y_0 is the solution of the (exactly solvable) unperturbed problem (4.2). Further terms describe the deviation in the solution of the perturbed problem due to the deviation from the unperturbed problem. In particular, we therefore would like to know if the perturbation series (4.3) converges to the solution of the perturbed problem (4.1). Such investigations are not just important in the context of ordinary differential equations but also for partial differential equations, algebraic equations, integral equations, and many other types of equations we encounter in Applied Mathematics. In fact, perturbation methods are closely related to numerical analysis and their earliest use includes the otherwise unsolvable mathematical problems of celestial mechanics. In particular, perturbation methods were used to describe and predict the orbit of the moon, which moves noticeable differently from a simple Keplerian ellipse, because of the competing gravitation of the Earth and the Sun.[1]

4.1 Regular Perturbation

The basis idea of **regular perturbation methods** is to assume that a solution of the perturbed problem (4.1) is given by the perturbation series (4.3), where the functions y_k, $k\in\mathbb{N}_0$, are found by substituting the perturbation series into the perturbed problem (4.1). If the method is successful, the **leading term**

[1] This is the oldest of the, by now famous, three-body problems; see [Gut98].

y_0 will be the solution of the unperturbed problem (4.2). The subsequent terms $\varepsilon y_1, \varepsilon^2 y_2$ are therefore regarded as higher-order correction terms that are expected to be small. Often, no more than the first two or three terms are used. The resulting truncated perturbation series,

$$p_{\varepsilon,K}(t) = \sum_{k=0}^{K} \varepsilon^k y_k(t),$$

is called a **perturbation approximation**. In what follows, we illustrate the idea of regular perturbation methods by some examples.

Example 4.1.1

Let us consider the quadratic algebraic equation

$$F(x,\varepsilon) := x^2 + 2\varepsilon x - 3 = 0 \tag{4.4}$$

as a perturbed problem, where $\varepsilon > 0$ is a small parameter. Note that the corresponding unperturbed problem $F(x,0) = 0$ is given by

$$x^2 - 3 = 0.$$

We now assume a solution of (4.4) in the form of the perturbation series

$$p_\varepsilon = \sum_{k=0}^{\infty} \varepsilon^k x_k = x_0 + \varepsilon x_1 + \varepsilon^2 x_2 + \mathcal{O}\left(\varepsilon^3\right).$$

Here, $\mathcal{O}(\varepsilon^3)$ denotes a term $r = r(\varepsilon)$ with $r = \mathcal{O}(\varepsilon^3)$. Substituting the perturbation series p_ε into the perturbed problem (4.4) gives us

$$\left(x_0 + \varepsilon x_1 + \varepsilon^2 x_2 + \mathcal{O}\left(\varepsilon^3\right)\right)^2 + 2\varepsilon\left(x_0 + \varepsilon x_1 + \varepsilon^2 x_2 + \mathcal{O}\left(\varepsilon^3\right)\right) - 3 = 0.$$

Expanding out and collecting the terms in different orders of ε results in

$$\left(x_0^2 - 3\right) + \varepsilon\left(2x_0(x_1+1)\right) + \varepsilon^2\left(x_1^2 + 2x_0x_2 + 2x_1\right) + \mathcal{O}\left(\varepsilon^3\right) = 0.$$

Hence, comparison of the coefficients yield the (nonlinear) system of equations

$$\begin{aligned} x_0^2 - 3 &= 0, \\ 2x_0(x_1+1) &= 0, \\ x_1^2 + 2x_0x_2 + 2x_1 &= 0. \end{aligned}$$

This system of equations can be solved successive, yielding

$$x_0 = \pm\sqrt{3}, \quad x_1 = -1, \quad x_2 = \pm\frac{1}{2\sqrt{3}},$$

and therefore the two perturbation series

$$\begin{aligned} p_\varepsilon &= \sqrt{3} - \varepsilon + \frac{\varepsilon^2}{2\sqrt{3}} + \mathcal{O}\left(\varepsilon^3\right), \\ \tilde{p}_\varepsilon &= -\sqrt{3} - \varepsilon - \frac{\varepsilon^2}{2\sqrt{3}} + \mathcal{O}\left(\varepsilon^3\right). \end{aligned}$$

Thus, we found the two approximate solutions

$$p_{\varepsilon,2} = \sqrt{3} - \varepsilon + \frac{\varepsilon^2}{2\sqrt{3}},$$
$$\tilde{p}_{\varepsilon,2} = -\sqrt{3} - \varepsilon - \frac{\varepsilon^2}{2\sqrt{3}}.$$

In fact, the exact solutions of (4.4) are given by

$$x = -\varepsilon + \sqrt{3+\varepsilon^2}, \quad \tilde{x} = -\varepsilon - \sqrt{3+\varepsilon^2}$$

or

$$x = -\varepsilon + \sqrt{3}\left(1 + \frac{\varepsilon^2}{2\sqrt{3}} + \mathcal{O}\left(\varepsilon^4\right)\right), \quad \tilde{x} = -\varepsilon - \sqrt{3}\left(1 + \frac{\varepsilon^2}{2\sqrt{3}} + \mathcal{O}\left(\varepsilon^4\right)\right)$$

by using the generalized binomial theorem. Hence, by comparing our perturbation approximations with the exact solutions,

$$x - p_{\varepsilon,2} = \mathcal{O}\left(\varepsilon^4\right), \quad \tilde{x} - \tilde{p}_{\varepsilon,2} = \mathcal{O}\left(\varepsilon^4\right),$$

we see that both of them are of fourth order in ε.

In the last example, the regular perturbation method led to a satisfactory result. In the next example, the procedure is the same, but the result does not turn out favorable. This is the first signal that we are in need to modify the regular method later on.

Example 4.1.2

Let us consider a spring-mass oscillator described by the mathematical model

$$m\frac{\mathrm{d}^2y}{\mathrm{d}s^2} = -ky - ay^3, \quad y(0) = A, \quad \frac{\mathrm{d}y}{\mathrm{d}s}(0) = 0,$$

for $s > 0$. Here, $y = y(s)$ describes the displacement of an object with mass m at the time s, k and a are constants characterizing the stiffness properties of the spring, and A is the initial displacement of the object. Moreover, let us assume that $a \ll k$. Then, using the dimensionless variables

$$t = \frac{s}{\sqrt{mk^{-1}}}, \quad u = \frac{y}{A},$$

scaling provides us with the simplified model

$$\ddot{u} = -u - \varepsilon u^3, \quad u(0) = 1, \quad \dot{u}(0) = 0, \tag{4.5}$$

where

$$\varepsilon := \frac{aA^2}{k} \ll 1$$

is a small dimensionless parameter. Furthermore, $\ddot{u}$ denotes the second derivative of u with respect to t, i. e. $\ddot{u} = \mathrm{d}^2u/\mathrm{d}t^2$. Equation (4.5) is known as the *Duffing equation.*

Next, let us consider the regular perturbation method in the context of the Duffing equation (4.5). In this case, the perturbation series looks like

$$p_\varepsilon(t) = \sum_{k=0}^{\infty} \varepsilon^k u_k(t).$$

For sake of simplicity, however, we just focus on the perturbation approximation including the first two terms,

$$p_{\varepsilon,1}(t) = u_0(t) + \varepsilon u_1(t).$$

Once more, the coefficients u_0 and u_1 are determined by substituting $p_{\varepsilon,1}$ into the perturbed problem (4.5), resulting in

$$\begin{aligned} \ddot{u}_0 + \varepsilon \ddot{u}_1 &= -(u_0 + \varepsilon u_1) - \varepsilon (u_0 + \varepsilon u_1)^3, \\ u_0(0) + \varepsilon u_1(0) &= 1, \\ \dot{u}_0(0) + \varepsilon \dot{u}_1(0) &= 0. \end{aligned}$$

Expanding out, collecting the terms in the different orders of ε, and comparing their coefficients yields the following sequence of linear initial value problems:

$$\ddot{u}_0 = -u_0, \quad u_0(0) = 1, \quad \dot{u}_0(0) = 0, \tag{4.6}$$

$$\ddot{u}_1 = -u_1 - u_0^3, \quad u_1(0) = 1, \quad \dot{u}_1(0) = 0 \tag{4.7}$$

The solution u_0 of (4.6) is clearly given by

$$u_0(t) = \cos t.$$

Thus, (4.7) becomes

$$\ddot{u}_1 = -u_1 - (\cos t)^3, \quad u_1(0) = 1, \quad \dot{u}_1(0) = 0,$$

with solution

$$u_1 = \frac{1}{32}(\cos 3t - \cos t) - \frac{3}{8} t \sin t.$$

The perturbation approximation is therefore given by

$$p_{\varepsilon,1}(t) = \cos t + \varepsilon \left[\frac{1}{32}(\cos 3t - \cos t) - \frac{3}{8} t \sin t \right].$$

For a fixed time t the term goes to zero as $\varepsilon \to 0$. Yet, if t itself is of order ε^{-1} or larger as $\varepsilon \to 0$, then the term $\varepsilon \frac{3}{8} t \sin t$ has an increasingly large amplitude. Such terms in a perturbation approximation, which are not ensured to converge to zero as $\varepsilon \to 0$, are called **secular terms**. Unfortunately, this behavior is not consistent with the physical situation of the underlying perturbation problem (4.5).

We summarize that for this example the correction term cannot be made arbitrarily small for $t \in [0, \infty)$, by choosing ε small enough. Note that it is also not possible to repair this by including further correction terms, e. g. of order ε^2 or ε^3. Only if we restrict the time to a finite interval $t \in [0, T]$ for some fixed $T > 0$ the correction term can be made arbitrarily small.

In fact, another way to remedy this type of singular behavior is the **Poincaré–Lindstedt method**. The key idea of this method is to introduce a distorted time scale in the perturbation series. In case of Example 4.1.2, this would yield

$$p_\varepsilon(\tau) = \sum_{k=0}^{\infty} \varepsilon^k u_k(\tau)$$

with $\tau = \omega t$ and

$$\omega = \sum_{k=0}^{K} \varepsilon^k \omega_k.$$

For more details on the Poincaré–Lindstedt method we refer to Chapter 2.1.3 in [Log13].

4.2 Pitfalls of Regular Perturbation

At the end of the last chapter we already observed problems for the regular perturbation method when secular terms arise in the perturbation approximation. In what follows, we demonstrate some more pitfalls of regular perturbation. By doing so we hope to sharpen our understanding of when we can the expect regular perturbation method to succeed and —most notably— when we cannot.

Let us start with two simple examples on algebraic equations.

Example 4.2.1

Given is the quadratic equation

$$\varepsilon x^2 + 2x + 1 = 0 \tag{4.8}$$

with $0 < \varepsilon \ll 1$. Of course, it is not hard to solve this equation exactly and its solutions are given by

$$x_{1,2} = -\frac{1}{\varepsilon} \pm \sqrt{\frac{1}{\varepsilon^2} - \frac{1}{\varepsilon}} = \frac{1}{\varepsilon}\left(-1 \pm \sqrt{1-\varepsilon}\right). \tag{4.9}$$

Note that

$$\lim_{\varepsilon \to 0} x_1 = -\frac{1}{2}, \quad \lim_{\varepsilon \to 0} x_2 = -\infty.$$

Yet, our goal is to illustrate the failure of the regular perturbation method for this example. If we attempt regular perturbation by substituting the perturbation series

$$p_\varepsilon = x_0 + \varepsilon x_1 + \varepsilon^2 x_2 + \dots$$

into the perturbed problem (4.8), we get, after comparing the coefficients of the different orders of ε, the following sequence of equations:

$$\begin{aligned} 2x_0 + 1 &= 0, \\ x_0^2 + 2x_1 &= 0, \\ 2x_1 x_0 + 2x_2 &= 0, \quad \dots \end{aligned}$$

Successively solving these equations provides us with

$$x_0 = -\frac{1}{2}, \quad x_1 = -\frac{1}{8}, \quad x_2 = -\frac{1}{16}, \quad \dots$$

and therefore with the perturbation series

$$p_\varepsilon = -\frac{1}{2} - \frac{1}{8}\varepsilon - \frac{1}{16}\varepsilon^2 - \dots$$

While this perturbation series can be considered consistent with the exact solution x_1 ($\lim_{\varepsilon \to 0} x_1 = -\frac{1}{2} = \lim_{\varepsilon \to 0} p_\varepsilon$), the second solution x_2 is not captured by the perturbation series. We immediately note the following:

- Regular perturbation just gives us a single (approximate) solution, where there should be two.

What went wrong? An obvious problem is that the unperturbed problem

$$2x + 1 = 0,$$

for which the solution is given by x_0, is not a reasonable simplification of the perturbed problem (4.8) for the second solution x_2. Even though ε is a small parameter and might be neglected, the product εx_2^2 cannot, since x_2 increases with decreasing ε; see (4.9). This is a problem we have already encountered in the context of scaling: A term that may appear small, εx_2^2, in fact is not small at all! Yet, this observation also points the way towards a solution of this problem. Motivated by scaling, let us introduce a new variable y of order 1 defined by

$$y = \varepsilon x.$$

With this change of variable the original equation (4.8) becomes

$$y^2 + 2y + \varepsilon = 0.$$

Then, the coefficients of the corresponding perturbation series

$$\tilde{p}_\varepsilon = y_0 + \varepsilon y_1 + \varepsilon^2 y_2 + \dots$$

satisfy

$$\begin{gathered} y_0^2 + 2y_0 = 0, \\ 2y_0 y_1 + 2y_1 + 1 = 0, \quad \dots \end{gathered}$$

and are therefore given by

$$y_0 = -2, \quad y_1 = \frac{1}{2}, \quad \dots$$

Hence, we get

$$\tilde{p}_\varepsilon = -2 + \frac{1}{2}\varepsilon + \dots,$$

or, after resubstitution,

$$\tilde{p}_\varepsilon = -\frac{2}{\varepsilon} + \frac{1}{2} + \dots.$$

Thus, only after appropriately scaling the equation, the regular perturbation method also revealed a perturbation series consistent with the second solution x_2. In summary, since the two solutions x_1 and x_2 were of different orders with respect to ε, one expansion was not able to reveal both.

The approach used in the above example, of considering the perturbed equation as well as an appropriate scaled version, to find perturbation series for solutions of different orders is called **dominant balancing**.

The next example of a second order boundary value problem is supposed to illustrate another pitfall of regular perturbation methods: Sometimes it is not even possible to calculate the leading term. This happens, for instance, when the unperturbed problem is not well defined.

Example 4.2.2

Let us consider the second-order boundary value problem

$$\begin{aligned} \varepsilon y'' + (1+\varepsilon)y' + y = 0, \quad 0 < x < 1, \quad 0 < \varepsilon \ll 1, \\ y(0) = 0, \quad y(1) = 1. \end{aligned} \tag{4.10}$$

Assuming a naive perturbation series of the form

$$p_\varepsilon(x) = y_0(x) + \varepsilon y_1(x) + \varepsilon^2 y_2(x) + \ldots,$$

where the coefficients y_k, $k \in \mathbb{N}_0$, are computed by substituting p_ε into (4.10), the leading term y_0 would be given as the solution of the unperturbed problem

$$\begin{aligned} y' + y = 0, \quad 0 < x < 1, \\ y(0) = 0, \quad y(1) = 1. \end{aligned} \tag{4.11}$$

However, it is easy to note that (4.11) does not have a solution. The general solution of $y' + y = 0$ is given by

$$y(t) = ce^{-x}.$$

Yet, the boundary condition $y(0) = 0$ yields $c = 0$ and therefore

$$y(t) = 0,$$

while the boundary condition $y(1) = 1$ yields $c = e$ and therefore

$$y(t) = e^{1-x}.$$

Generally speaking, it is always a bad sign when the order of an ordinary differential equation is lower than the number of initial/boundary conditions. Once more, we observe the regular perturbation method to fail. In the following list, we summarize some of the several indicators that often suggest failure of the regular perturbation method:

1. When the small parameter is multiplied with the highest derivative in the problem.
2. More generally, when setting the small parameter to zero changes the character of the problem. This includes partial differential equations changing type (e. g., from elliptic to parabolic), or an algebraic equation changing degree.
3. When problems on infinite domains yield secular terms.
4. When the equations that model physical processes have multiple time or spacial scales.

Such problems, resulting in the pitfall of the regular perturbation method, fall in the general class of **singular perturbation problems**. Different techniques to adapt perturbation methods to some of the above singular perturbation problems and therefore overcome their failure in these cases, can be found in *boundary* and *initial layer analysis*; see chapters 2.3 and 2.4 in [Log13]. Another noteworthy approach is the *Wentzel–Kramers–Brillouin perturbation method*. The interested reader may find further information in Chapter 2.5 of [Log13].

Chapter 5

Asymptotic Expansion of Integrals

Many physical quantities can be described by integrals. In particular, the solution of differential equations often yield formulas involving integrals. Unfortunately, in many cases, these integrals cannot be evaluated in closed form. For example, the initial value problem

$$y'' + 2\lambda + y' = 0, \quad y(0) = 0, \quad y'(0) = 1,$$

has the solution

$$y_\lambda(t) = \int_0^t e^{-\lambda s^2} \, \mathrm{d}s.$$

Yet, the solution cannot be calculated, because there is no antiderivative in terms of simple functions, for the integral on the right hand side. For some problems, however, we may want to at least know the behavior for fixed λ and large t or, conversely, for fixed t and large λ. Problems like this are common in applied mathematics and in this chapter we will address some standard techniques to solve them for certain types of integrals.

5.1 Laplace Integrals

We start by investigating integrals of the form

$$I(\lambda) = \int_a^b f(t) e^{-\lambda g(t)} \, \mathrm{d}t \tag{5.1}$$

with $\lambda \gg 1$ and strictly increasing $g \in C^1([a,b])$. A prominent example is the Laplace transform

$$\mathcal{L}\{f\}(\lambda) = \int_0^\infty f(t) e^{-\lambda t} \, \mathrm{d}t,$$

which is an often used tool to transform differential equations into algebraic equations. Note that to study the whole class of integrals 5.1 it actually suffices to consider so-called Laplace integrals.

Definition 5.1.1: Laplace integrals

Let $f : [0, b] \to \mathbb{R}$ be integrable and $\lambda \gg 1$. We call

$$I(\lambda) = \int_0^b f(t) e^{-\lambda t} \, \mathrm{d}t \tag{5.2}$$

a **Laplace integral**.

To observe that every integral of the form (5.1) can be represented as a Laplace integral, we can make a change of variables $s = g(t) - g(a)$. This yields

$$\int_a^b f(t) e^{-\lambda g(t)} \, \mathrm{d}t = e^{-\lambda g(a)} \int_0^{g(b)-g(a)} \frac{f(t(s))}{g'(t(s))} e^{-\lambda s} \, \mathrm{d}s$$

and therefore a Laplace integral. Here, $t = t(s)$ denotes the solution of the equation $s = g(t) - g(a)$.

Unfortunately, in many cases we are not able to compute Laplace integrals exactly. Hence, we will look for approximations to (5.2) next. The fundamental idea we will follow for this is to determine which subintegral gives the dominant contribution to the integral Note that the function $e^{-\lambda t}$ is rapidly decaying for $t > 0$ if $\lambda \gg 1$. Thus, assuming that f does not grow too fast at infinity and is reasonably well behaved at $t = 0$. This approach is often referred to as **Laplace's method** and we start by illustrating it by an example.

Example 5.1.2

Let us consider the Laplace integral

$$I(\lambda) = \int_0^\infty \frac{\sin t}{t} e^{-\lambda t} \, \mathrm{d}t, \quad \lambda \gg 1.$$

It is clear that the integral vanishes rapidly for increasing t and we can therefore expect the main contribution to the integral to occur near $t = 0$. Hence, we partition the interval into $[0, T]$ and (T, ∞), i. e.,

$$I(\lambda) = \underbrace{\int_0^T \frac{\sin t}{t} e^{-\lambda t} \, \mathrm{d}t}_{=:I_1(\lambda,T)} + \underbrace{\int_T^\infty \frac{\sin t}{t} e^{-\lambda t} \, \mathrm{d}t}_{=:I_2(\lambda,T)}$$

for $T > 0$. Note the second integral is an *exponentially small term (EST)*; that is $I_2(\lambda, T) = \mathcal{O}(\lambda^{-1} e^{-\lambda T})$ as $\lambda, T \to \infty$. This can be quickly observed from

$$|I_2(\lambda, T)| \leq \int_T^\infty \left| \frac{\sin t}{t} \right| e^{-\lambda t} \, \mathrm{d}t \leq \int_T^\infty e^{-\lambda t} \, \mathrm{d}t = \lambda^{-1} e^{-\lambda T}.$$

Addressing the first integral next, in the finite interval $[0, T]$ we can replace $\frac{\sin t}{t}$ by its Taylor series around $t = 0$,

$$\frac{\sin t}{t} = 1 - \frac{t^2}{3!} + \frac{t^4}{5!} \mp \dots .$$

This yields

$$I_1(\lambda, T) = \int_0^T \left(1 - \frac{t^2}{3!} + \frac{t^4}{5!} \mp \dots \right) e^{-\lambda t} \, \mathrm{d}t$$

and a change of variable provides us with

$$I_1(\lambda, T) = \frac{1}{\lambda} \int_0^T \left(1 - \frac{u^2}{3! \lambda^2} + \frac{u^4}{5! \lambda^4} \mp \dots \right) e^{-u} \, \mathrm{d}u.$$

But now, the upper limit can be replaced by ∞. The error introduced by doing this is exponentially small as $T \to \infty$. Consequently, we have

$$I_1(\lambda, T) = \frac{1}{\lambda} \int_0^\infty \left(1 - \frac{u^2}{3! \lambda^2} + \frac{u^4}{5! \lambda^4} \mp \dots \right) e^{-u} \, \mathrm{d}u + \mathcal{O}\left(e^{-T} \right).$$

Finally, using the integration formula

$$\int_0^\infty u^m e^{-u} \, \mathrm{d}u = m!, \quad m \in \mathbb{N}_0,$$

we obtain

$$\begin{aligned} I_1(\lambda, T) &= \left(\frac{1}{\lambda} - \frac{3}{\lambda^3} + \frac{5}{\lambda^5} \mp \dots \right) + \mathcal{O}\left(e^{-T}\right) \\ &= \frac{1}{\lambda} - \frac{3}{\lambda^3} + \mathcal{O}\left(\lambda^{-5}\right) + \mathcal{O}\left(e^{-T}\right) \end{aligned}$$

and therefore

$$I(\lambda) = \frac{1}{\lambda} - \frac{3}{\lambda^3} + \mathcal{O}(\lambda^{-5}) + \mathcal{O}\left(e^{-T}\right) + \mathcal{O}(\lambda^{-1}e^{-\lambda T}).$$

Since this equation holds for any $T > 0$, it implies

$$I(\lambda) = \frac{1}{\lambda} - \frac{3}{\lambda^3} + \mathcal{O}(\lambda^{-5}).$$

A general result for Laplace integrals of this flavor can be found in the subsequent Theorem 5.1.6. Yet, before addressing this result, we introduce some additional notation.

Definition 5.1.3: The gamma function

We call the function

$$\Gamma : \mathbb{R}^+ \to \mathbb{R}, \quad \Gamma(x) := \int_0^\infty u^{x-1} e^{-u} \, \mathrm{d}u,$$

gamma function.

The gamma function is considered as a generalization of the factorial function. This is because of the following properties of the gamma function.

Lemma 5.1.4: Properties of the gamma function

Let $x \in \mathbb{R}^+$ and $n \in \mathbb{N}$. The following properties hold for the gamma function Γ:

(a) $\Gamma(x+1) = x\Gamma(x)$

(b) $\Gamma(n) = (n-1)!$

Proof. (a) The assertion follows from Definition 5.1.3 by applying integration by parts.

(b) Calculating $\Gamma(1)$, we have

$$\Gamma(1) = \int_0^\infty e^{-x} \, \mathrm{d}x.$$

Hence, the assertion follows from the observation that

$$\Gamma(1) = 1 \quad \text{and} \quad \Gamma(n+1) = n\Gamma(n),$$

where the latter equation is a special case of (a).

□

Definition 5.1.5: Asymptotic sequences and expansions

Let $(g_n(t,\lambda))_{n\in\mathbb{N}_0}$ be a sequence of gauge functions (see Definition 3.1.1).

(a) We say $(g_n(t,\lambda))_{n\in\mathbb{N}_0}$ is an asymptotic sequence as $\lambda \to \infty$, $t \in I$, if

$$g_{n+1}(t,\lambda) = \mathbf{o}(g_n(t,\lambda)) \quad \text{as } \lambda \to \infty$$

for $n \in \mathbb{N}_0$.

(b) Given a function $y = y(t,\lambda)$, the formal series

$$\sum_{n=0}^{\infty} a_n g_n(t,\lambda) \tag{5.3}$$

is said to be an **asymptotic expansion** of y as $\lambda \to \infty$, if

$$y(t,\lambda) - \sum_{n=0}^{N} a_n g_n(t,\lambda) = \mathbf{o}(g_N(t,\lambda)) \quad \text{as} \quad \lambda \to \infty$$

or

$$y(t,\lambda) - \sum_{n=0}^{N} a_n g_n(t,\lambda) = \mathbf{O}(g_{N+1}(t,\lambda)) \quad \text{as} \quad \lambda \to \infty$$

for every $N \in \mathbb{N}$. We denote 5.3 being an asymptotic expansion of y as $\lambda \to \infty$ by

$$y \sim \sum_{n=0}^{\infty} a_n g_n(t,\lambda), \quad \lambda \to \infty.$$

Note that (a) in Definition 5.1.5 means that every element $g_{n+1}(t,\lambda)$ of the sequence converges faster to zero than its predecessor. Similarly, (b) in Definition 5.1.5 means that the remainder of any partial sum is little oh, $\mathbf{o}$, of the last term. We are now ready to formulate a general result regarding the asymptotic behavior of Laplace integrals.

Theorem 5.1.6: Watson's Lemma

Consider the Laplace integral

$$I(\lambda) = \int_0^b t^\alpha h(t) e^{-\lambda t} \, \mathrm{d}t,$$

where $\alpha > -1$. Let h satisfy $|h(t)| \leq Ce^{kt}$, $0 < t < b$, for some positive constants C and k and let h have a Taylor series expansion around $t = 0$. Then,

$$I(\lambda) \sim \sum_{n=0}^{\infty} \frac{h^{(n)}(0)\Gamma(\alpha+n+1)}{n!\lambda^{\alpha+n+1}}, \quad \lambda \to \infty.$$

Proof. Here, we only outline the proof of Watson's Lemma in a rough manner. A more detailed discussion can be found, for instance, in [Mur12]. The idea behind the proof is to follow the same argument as in Example 5.1.2. Let $T \in (0,b)$ lie in the radius of convergence of the Taylor series of h around $t = 0$.

Then, we split up the integral $I(\lambda)$ to obtain

$$I(\lambda) = \underbrace{\int_0^T t^\alpha h(t) e^{-\lambda t}\,\mathrm{d}t}_{=:I_1(\lambda,T)} + \underbrace{\int_T^b t^\alpha h(t) e^{-\lambda t}\,\mathrm{d}t}_{=:I_2(\lambda,T)}.$$

Note that the condition $|h(t)| \leq Ce^{kT}$, $0 < t < b$, ensures that the second integral satisfies

$$I_2(\lambda, T) = \mathcal{O}(e^{-\lambda})$$

and is therefore an exponentially small term. Next, replacing h by its Taylor series around $t = 0$ in $I_1(\lambda, T)$, we have

$$I_1(\lambda,T) = \int_0^T t^\alpha \left(\sum_{n=0}^\infty h^{(n)}(0) \frac{t^n}{n!} \right) e^{-\lambda t}\,\mathrm{d}t = \sum_{n=0}^\infty \frac{h^{(n)}(0)}{n!} \int_0^T t^{n+\alpha} e^{-\lambda t}\,\mathrm{d}t,$$

where the second equation follows Lebesgue's dominated convergence theorem. Furthermore, making the substitution $u = 2t$ and replacing the upper limit of integration λT by ∞ yields

$$I_1(\lambda,T) = \sum_{n=0}^\infty \frac{h^{(n)}(0)}{n!} \left(\frac{1}{\lambda^{n+\alpha+1}} \int_0^T u^{n+\alpha} e^{-u}\,\mathrm{d}t + \mathcal{O}(e^{-\lambda T}) \right).$$

Note that replacing the upper limit of integration λT by ∞ only introduces an error that is exponentially small as $\lambda \to \infty$, which is reflected by the term $\mathcal{O}(e^{-\lambda T})$. Finally, remembering the definition of the gamma function, we obtain

$$I_1(\lambda,T) = \sum_{n=0}^\infty \frac{h^{(n)}(0)\Gamma(\alpha+n+1)}{n!\lambda^{n+\alpha+1}} + \mathcal{O}(e^{-\lambda T})$$

and therefore the assertion. □

5.2 Integration by Parts

Another useful procedure to derive asymptotic expansions of integrals is integration by parts. We demonstrate this for **Euler's integral**

$$I(\lambda) = \int_0^\infty \frac{e^{-t}}{1 + t/\lambda}\,\mathrm{d}t. \tag{5.4}$$

Using integration by parts, we prove that

$$I(\lambda) \sim \sum_{n=0}^\infty (-1)^n n! \lambda^{-n}. \tag{5.5}$$

Thereby, it suffices to show the following result.

Lemma 5.2.1

For Euler's integral (5.4) the inequality

$$|I(\lambda) - \sum_{n=0}^N (-1)^n n! \lambda^{-n}| \leq (N+1)! \lambda^{-(N+1)}$$

holds for all $N \in \mathbb{N}$.

Proof. Applying integration by parts to Euler's integral yields

$$I(\lambda) = 1 - \lambda^{-1} \int_0^\infty \frac{e^{-t}}{(1+\lambda^{-1}t)^2} \, \mathrm{d}t.$$

In fact, utilizing integration by parts another N times, we find that

$$I(\lambda) = \sum_{n=0}^{N} (-1)^n n! \lambda^{-n} + r_{N+1}(\lambda),$$

where the remainder $r_{N+1}(\lambda)$ is given by

$$r_{N+1}(\lambda) = (-1)^{N+1}(N+1)!\lambda^{-(N+1)} \int_0^\infty \frac{e^{-t}}{(1+\lambda^{-1}t)^{N+2}} \cdot \mathrm{d}t$$

Finally, this yields

$$|I(\lambda) - \sum_{n=0}^{N} (-1)^n n! \lambda^{-n}| \leq (N+1)!\lambda^{-(N+1)} \int_0^\infty e^{-t} \, \mathrm{d}t = (N+1)!\lambda^{-(N+1)}$$

and therefore the assertion. □

Then, (5.5) follows by observing that for the sequence of gauge functions $(g_n(\lambda))_{n \in \mathbb{N}_0}$ with $g_n(\lambda) = \lambda^{-n}$ the relation

$$\lim_{\lambda \to \infty} \frac{|I(\lambda) - \sum_{n=0}^{N} (-1)^n n! \lambda^{-n}|}{g_N(\lambda)} \leq \lim_{\lambda \to \infty} \frac{(N+1)!\lambda^{-(N+1)}}{g_N(\lambda)} = 0$$

holds true and therefore

$$I(\lambda) - \sum_{n=0}^{N} (-1)^n n! g_n(\lambda) = \mathbf{o}(g_N(\lambda)), \quad \lambda \to \infty.$$

Chapter 6

Functional Analysis - A Crash Course

This chapter is supposed to provide you with some basic concepts from functional analysis. Broadly speaking, functional analysis is a branch of mathematical analysis that is concerned with vector spaces which are endowed with a topological structure (e. g. a norm or an inner product) and linear maps between some vector spaces. In particular, such maps can be differential operators mapping functions from the vector space $C^1([0,1])$ to the vector space $C^0([0,1])$ or integral operators. Many of the concepts discussed here will also be useful in the subsequent Chapter 7 on variational methods.

6.1 Normed Vector Spaces

Let us start by adding two prominent topological structures with which vector spaces can be endowed, norms and inner products.

Definition 6.1.1: Norms & normed vector spaces

Let X be a vector space over $\mathbb{R}$. A function

$$X \to \mathbb{R}_0^+, \quad x \mapsto \|x\|$$

is called a **norm** if it satisfies the following properties for all x, $y \in X$ and $\lambda \in \mathbb{R}$:

$$\begin{aligned}
&(\text{N1}) \quad \|x\| = 0 \iff x = 0, && \text{(being positive definite)}\\
&(\text{N2}) \quad \|\lambda x\| = |\lambda|\|x\|, && \text{(being absolutely homogeneous)}\\
&(\text{N3}) \quad \|x+y\| \leq \|x\| + \|y\| && \text{(satisfying the triangle inequality)}
\end{aligned}$$

The pair $(X, \|\cdot\|)$ is referred to as a **normed vector space** then.

Example 6.1.2

Let $X = C([0,1])$. It is easy to verify that

$$\|f\|_\infty := \max_{t \in [0,1]} |f(t)|, \quad f \in X,$$

is a norm on X.

Another famous class of norms are the so-called p-norms.

Lemma 6.1.3

Let $1 \leq p < \infty$. Then

$$\|f\|_p := \left(\int_0^1 |f(t)|^p \, \mathrm{d}t \right)^{1/p}, \quad f \in C([0,1]),$$

is a norm on $C([0,1])$.

Proof. Homework 05. □

Essentially, a norm allows us measure the length of vectors, which often is quite useful. Yet, an even more powerful tool are inner products. These additionally allow us to measure the angle between two vectors.

Definition 6.1.4

Let X be a vector space over $\mathbb{R}$. A *bilinear* function

$$X \times X \to \mathbb{R}, \quad (x,y) \mapsto \langle x, y \rangle$$

is called an **inner product** if it satisfies the following properties for all x, $y \in \mathbb{R}$:

$$\begin{aligned} &\text{(IP1)} \quad \langle x, x \rangle \geq 0, && \text{(being nonnegative)} \\ &\text{(IP2)} \quad \langle x, x \rangle = 0 \iff x = 0, && \text{(being positive definite)} \\ &\text{(IP3)} \quad \langle x, y \rangle = \langle y, x \rangle, && \text{(being symmetric)} \end{aligned}$$

The pair $(X, \langle \cdot, \cdot \rangle)$ is referred to as an **inner product space** (also **pre-Hilbert space**).

In the above definition, $\langle \cdot, \cdot \rangle$ being bilinear means that it is linear in both arguments. That is,

$$\begin{aligned} \langle x + y, z \rangle &= \langle x, z \rangle + \langle y, z \rangle, \\ \langle \lambda x, z \rangle &= \lambda \langle x, z \rangle, \\ \langle x, y + z \rangle &= \langle x, y \rangle + \langle x, z \rangle, \\ \langle x, \lambda z \rangle &= \lambda \langle x, z \rangle \end{aligned}$$

hold for all x, y, $z \in X$ and $\lambda \in \mathbb{R}$.

Remark 6.1.5

Here, we only consider vector spaces over $\mathbb{R}$. While the definition for norms does not change if we go over to vector spaces over $\mathbb{C}$ it should be stressed that the definition of inner products *does*! In the case of X being a vector space over $\mathbb{C}$, the function $\langle \cdot, \cdot \rangle$ is required to be sesquilinear (linear in the first argument and conjugate linear in the second argument) instead of bilinear. Moreover, the property of $\langle \cdot, \cdot \rangle$ being symmetric is replaced by $\langle \cdot, \cdot \rangle$ being conjugate symmetric in this case.

Example 6.1.6

Two of the most prominent inner products are

(a) $\langle x, y \rangle = \sum_{i=1}^n x_i y_i, \quad x = (x_1, \ldots x_n)^\intercal, \; y = (y_1, \ldots, y_n)^\intercal$, on $\mathbb{R}^n$,

(b) $\langle f, g \rangle = \int_0^1 f(t) g(t) \, \mathrm{d}t$ on $C([0,1])$.

It is easy to note that both functions are bilinear and nonnegative. Moreover, both functions are symmetric due to $\mathbb{R}$ being commutative with respect to multiplication ($ab = ba$ for all $a, b \in \mathbb{R}$). Finally, it is also not hard to verify that both functions are positive definite. It is clear that

(a) $x = (0, \dots 0)^\intercal \implies \langle x, x \rangle = \sum_{i=1}^n 0 = 0$,

(b) $f \equiv 0 \implies \langle f, f \rangle = \int_0^1 0 \,\mathrm{d}t = 0$.

On the other hand, the reverse implication can be shown by the contraposition ($A \implies B$ is equivalent to $\neg B \implies \neg A$).

(a) Let $x = (x_1, \dots, x_n) \neq (0, \dots 0)^\intercal$. Then, there exists a $j \in \{1, \dots, n\}$ such that $x_j \neq 0$. Hence,

$$\langle x, x \rangle = \sum_{i=1}^n x_i^2 \geq x_j^2 > 0.$$

(b) Let $f \not\equiv 0$. Then, there exists an $t_0 \in [0, 1]$ such that $f(t_0) \neq 0$. Since $f \in C([0, 1])$, we can find a whole neighborhood $U_\varepsilon(t_0) = (t_0 - \varepsilon, t_0 + \varepsilon)$ such that $f(t) \neq 0$ for all $t \in U_\varepsilon(t_0)$. Thus, we have

$$\langle f, f \rangle = \int_0^1 f^2(t) \,\mathrm{d}t \geq \int_{x_0-\varepsilon}^{x_0+\varepsilon} f^2(t) \,\mathrm{d}t > 0.$$

As noted before, inner products can be considered as a much more powerful tool than norms. In what follows, we summarize some of the most important and far reaching results for inner products. We start by showing that, in particular, every inner product induces a norm.

Lemma 6.1.7

Let $(X, \langle \cdot, \cdot \rangle)$ be an inner product space over $\mathbb{R}$. Then,

$$\|x\| = \sqrt{\langle x, x \rangle}, \quad x \in X,$$

is a norm on X.

Proof. The above defined function $\| \cdot \|$ clearly is nonnegative. Hence, it remains to check the properties (N1)-(N3).

(N1) $\| \cdot \|$ being positive definite is ensured by $\langle \cdot, \cdot \rangle$ being positive definite, due to (IP1).

(N2) $\| \cdot \|$ being absolutely homogeneous is ensured by $\langle \cdot, \cdot \rangle$ being bilinear:

$$\|\lambda x\|^2 = \langle \lambda x, \lambda x \rangle = \lambda^2 \langle x, x \rangle = (|\lambda| \|x\|)^2$$

(N3) Finally, the triangle inequality for $\| \cdot \|$ follows from $\langle \cdot, \cdot \rangle$ being bilinear and symmetric (IP3):

$$\begin{aligned} \|x + y\|^2 &= \langle x + y, x + y \rangle \\ &= \langle x, x \rangle + \langle x, y \rangle + \langle y, x \rangle + \langle y, y \rangle \\ &= \|x\|^2 + 2\langle x, y \rangle + \|y\|^2 \\ &\leq \|x\|^2 + 2|\langle x, y \rangle| + \|y\|^2 \end{aligned}$$

Thus, by utilizing the Cauchy–Schwarz inequality,

$$|\langle x, y \rangle| \leq \|x\| \|y\|, \quad x, y \in X, \tag{6.1}$$

we get

$$\|x+y\|^2 \leq \|x\|^2 + 2\|x\|\|y\| + \|y\|^2 = (\|x\| + \|y\|)^2$$

and therefore the triangle inequality for $\|\cdot\|$.

□

Note that we have utilized the Cauchy-Schwarz inequality (6.1) in the above proof. To not leave a gap in the proof, we need to verify (6.1) next (without using the fact that $\|\cdot\| = \sqrt{\langle\cdot,\cdot\rangle}$ satisfies the triangle inequality; otherwise we would be stuck in a logical loop).

Lemma 6.1.8: The Cauchy–Schwarz inequality

Let $(X, \langle\cdot,\cdot\rangle)$ be an inner product space over $\mathbb{R}$ and let $\|\cdot\| : X \to \mathbb{R}_0^+$ be given by

$$\|x\| = \sqrt{\langle x, x\rangle}, \quad x \in X.$$

Then, the **Cauchy–Schwarz inequality**

$$|\langle x, y\rangle| \leq \|x\|\|y\|,$$

holds for all $x, y \in X$.

Proof. Let $x, y \in X$ and $\lambda \in \mathbb{R}$. We start by noting that

$$\begin{aligned} 0 &\leq \|x - \lambda y\|^2 \\ &= \langle x, x\rangle - \langle x, \lambda y\rangle - \langle \lambda y, x\rangle + \langle \lambda y, \lambda y\rangle \\ &= \|x\|^2 - 2\lambda\langle x, y\rangle + \lambda^2\|y\|^2. \end{aligned}$$

Without loss of generality we can assume that $y \neq 0$. Then, by choosing

$$\lambda = \frac{\langle x, \lambda y\rangle}{\|y\|^2},$$

we get

$$0 \leq \|x\|^2 - 2\frac{|\langle x, \lambda y\rangle|^2}{\|y\|^2} + \frac{|\langle x, \lambda y\rangle|^2}{\|y\|^2} = \|x\|^2 - \frac{|\langle x, \lambda y\rangle|^2}{\|y\|^2}$$

and therefore

$$|\langle x, \lambda y\rangle|^2 \leq \|x\|^2\|y\|^2.$$

□

Another fundamental —yet far reaching— result for inner product spaces is the parallelogram law.

Lemma 6.1.9: The parallelogram law

Let $(X, \langle\cdot,\cdot\rangle)$ be an inner product space over $\mathbb{R}$ and let $\|\cdot\| = \sqrt{\langle\cdot,\cdot\rangle}$ be the norm induced by $\langle\cdot,\cdot\rangle$. Then, the parallelogram law

$$2(\|x\|^2 + \|y\|^2) = \|x+y\|^2 + \|x-y\|^2$$

holds for all $x, y \in X$.

Proof. The parallelogram law is easily established using the properties of the inner product:

$$\begin{aligned} \|x+y\|^2 &= \langle x+y, x+y\rangle = \langle x,x\rangle + \langle x,y\rangle + \langle y,x\rangle + \langle y,y\rangle, \\ \|x-y\|^2 &= \langle x-y, x-y\rangle = \langle x,x\rangle - \langle x,y\rangle - \langle y,x\rangle + \langle y,y\rangle, \\ \implies \|x+y\|^2 + \|x-y\|^2 &= 2\langle x,x\rangle + 2\langle y,y\rangle = 2(\|x\|^2 + \|y\|^2) \end{aligned}$$

□

In particular, the parallelogram law allows us to generalize Pythagoras' theorem to general inner product spaces. This is also related to inner products being able to assign angle-like relations between two vectors. For instance, we can check if two vectors are perpendicular by the concept of orthogonality.

Definition 6.1.10: Orthogonality

Let $(X, \langle\cdot,\cdot\rangle$ be an inner product space. We say that two vectors $x, y \in X$ are **orthogonal** if

$$\langle x, y\rangle = 0$$

holds.

Similarly to Pythagoras' theorem in $\mathbb{R}^2$, we can now formulate the following generalization for orthogonal vectors in inner product spaces.

Lemma 6.1.11: Pythagoras' theorem

Let $(X, \langle\cdot,\cdot\rangle)$ be an inner product space over $\mathbb{R}$, let $x, y \in X$, and let x and y be orthogonal. Then,

$$\|x\|^2 + \|y\|^2 = \|x+y\|^2$$

holds, where $\|\cdot\|$ is the norm induced by $\langle\cdot,\cdot\rangle$.

Proof. The assertion follows immediately from the proof of the parallelogram law when utilizing

$$\langle x, y\rangle = \langle y, x\rangle = 0.$$

□

As we can note from the proof of Pythagoras' theorem, the parallelogram law allows some impressive conclusions. Perhaps most formidable, it even allows to characterize inner product spaces among normed vector spaces.

Theorem 6.1.12: Characterization of inner product spaces

Let $(X, \|\cdot\|)$ be a normed space. Then, there exists an inner product $\langle\cdot,\cdot\rangle$ on X such that $\|x\|^2 = \langle x, x\rangle$ for all $x \in X$ ($\|\cdot\|$ is induced by $\langle\cdot,\cdot\rangle$) if and only if the parallelogram law

$$2\left(\|x\|^2 + \|y\|^2\right) = \|x+y\|^2 + \|x-y\|^2 \tag{6.2}$$

holds for all $x, y \in X$.

Proof. We have already established the parallelogram law to hold in in inner product spaces in Lemma 6.1.9. Hence, it only remains to show that the norm $\|\cdot\|$ is induced by an inner product if the parallelogram law holds in $(X, \|\cdot\|)$. This is discussed in a subsequent lemma. □

Lemma 6.1.13: Polarization identity

Let $(X, \|\cdot\|)$ be a normed space over $\mathbb{R}$. If the parallelogram law (6.2) holds, then there is an inner product $\langle \cdot, \cdot \rangle$ such that $\|x\|^2 = \langle x, x \rangle$ for all $x \in X$. Moreover, the inner product is uniquely given by

$$\langle x, y \rangle = \frac{1}{4}\left(\|x+y\|^2 - \|x-y\|^2 \right)$$

for $x, y \in X$.

Proof. The assertion follows from some lengthy and tiresome (but basic) computations. □

6.2 Convergence and Cauchy Sequences

In the last section, we discussed the topological structures of norms and inner products, with which vector spaces can be endowed. In particular, these also allow us to consider convergence of sequences of vectors and therefore continuity of functions between vector spaces.

Definition 6.2.1: Convergence

Let $(x_n)_{n\in\mathbb{N}}$ be a sequence in a normed vector space $(X, \|\cdot\|)$. We say $(x_n)_{n\in\mathbb{N}}$ **coverges** to a vector $x \in X$ if

$$\forall \varepsilon > 0 \; \exists N \in \mathbb{N} \; \forall n \geq N : \quad \|x - x_n\| < \varepsilon. \tag{6.3}$$

We denote this by

$$\lim_{n\to\infty} x_n = x \quad \text{or} \quad x_n \to x \text{ for } n \to \infty$$

and call x the **limit** of $(x_n)_{n\in\mathbb{N}}$.

Similarly to the usual setting of $X = \mathbb{R}$ or $X = \mathbb{C}$, we can also define Cauchy sequences.

Definition 6.2.2

Let $(x_n)_{n\in\mathbb{N}}$ be a sequence in a normed vector space $(X, \|\cdot\|)$. We call $(x_n)_{n\in\mathbb{N}}$ a **Cauchy sequence** if

$$\forall \varepsilon > 0 \; \exists N \in \mathbb{N} \; \forall n, m \geq N : \quad \|x_m - x_n\| < \varepsilon. \tag{6.4}$$

In many situations, we are used for convergence of a sequence to hold if and only if the sequence is a Cauchy sequence. In this case it is often more convenient to prove the Cauchy property (6.4) instead of directly showing convergence, since (6.4) does not assume any prior knowledge of the potentially existing (or non existing) limit x. Indeed, part of the equivalence between convergence and Cauchy sequences also holds in general normed vector spaces, namely, every convergent series is also a Cauchy sequence. This is noted in the following lemma.

Lemma 6.2.3: Convergence $\implies$ Cauchy sequence

Let $(x_n)_{n\in\mathbb{N}}$ be a sequence in a normed vector space $(X, \|\cdot\|)$. If $(x_n)_{n\in\mathbb{N}}$ converges, then $(x_n)_{n\in\mathbb{N}}$ is a Cauchy sequence.

Proof. Let x denote the limit of the convergent sequence $(x_n)_{n\in\mathbb{N}}$ and let $\varepsilon > 0$. We have to show that there is an $N \in \mathbb{N}$ such that

$$\|x_m - x_n\| < \varepsilon \quad \forall n, m \geq N.$$

Therefor, we note that

$$\|x_m - x_n\| \leq \|x_m - x\| + \|x - x_n\|.$$

Hence, since $x_n \to x$ for $n \to \infty$, there exists an $N \in \mathbb{N}$ such that

$$\|x_m - x\|, \|x - x_n\| < \frac{\varepsilon}{2} \quad \forall n, m \geq N.$$

For this N, we therefore have

$$\|x_m - x_n\| < \varepsilon \quad \forall n, m \geq N,$$

which results in the assertion. □

The above lemma tells us that also in general normed vector spaces, every convergent sequence is a Cauchy sequence. From real analysis ($X = \mathbb{R}$ or $X = \mathbb{R}^n$ or even $X = \mathbb{C}$) we are accustomed to the reversed implication to hold as well. Unfortunately, in general normed vector spaces Cauchy sequences do not necessarily converge. This is, among other things, demonstrated in the subsequent example.

Example 6.2.4

Given is the normed vector space $(C([0,2]), \|\cdot\|_2)$ with norm

$$\|f\|_2 = \left(\int_0^2 |f(t)|^2 \,\mathrm{d}t \right)^{1/2}.$$

Let us consider the sequence $(f_n)_{n\in\mathbb{N}} \subset C([0,2])$ with

$$f_n : [0,2] \to \mathbb{R}, \quad f_n(t) := \begin{cases} t^n & \text{if } 0 \leq t \leq 1, \\ 1 & \text{if } 1 < t \leq 2. \end{cases}$$

- Then, we can note that

$$\begin{aligned} \|f_n - f_m\|_2^2 &= \int_0^2 |t^n - t^m|^2 \,\mathrm{d}t \\ &= \int_0^2 t^{2n} - 2t^{nm} + t^{2m} \,\mathrm{d}t \\ &\leq \int_0^2 t^{2n} + t^{2m} \,\mathrm{d}t \\ &= \frac{1}{2n+1} + \frac{1}{2m+1} \\ &\leq \frac{1}{m} \end{aligned}$$

if $n \geq m$. Hence, $(f_n)_{n\in\mathbb{N}}$ is a Cauchy sequence, because, given $\varepsilon > 0$, with $N \in \mathbb{N}$ such that $N > \frac{1}{\varepsilon^2}$ we have

$$\|f_n - f_m\|_2 \leq \frac{1}{\sqrt{N}} < \varepsilon$$

for all $n, m \geq N$.

- Yet, at the same time, $(f_n)_{n \in \mathbb{N}}$ does *not* converge! To show this, let us assume that $(f_n)_{n \in \mathbb{N}}$ would converge to a limit $f \in C([0,1])$. It is easy to show that

$$f(t) = \begin{cases} 0 & \text{if } 0 \leq t < 1, \\ 0 & \text{if } 1 \leq t \leq 2 \end{cases}$$

needs to hold then. Yet, this would be a contradiction to f being continuous. Hence, there can be no limit in $(C([0,2]), \| \cdot \|_2)$.

We have just seen that while every convergent sequence is a Cauchy sequence in normed vector spaces the reversed statement does not hold in general. In fact, every Cauchy sequence also being convergent is an intrinsic property of a normed or inner product space and has received its own name.

Definition 6.2.5: Completeness

- A normed space $(X, \| \cdot \|)$ in which every Cauchy sequence converges is said to be **complete** and then referred to as a **Banach space**.
- An inner product space $(X, \langle \cdot, \cdot \rangle)$ —also called pre-Hilbert space— in which every Cauchy sequence converges with respecct to the induced norm is also said to be **complete** and referred to as **Hilbert space**.

6.3 Linear Operators and Functionals

After discussing the topological structure of norms (and inner products) on vector spaces and the resulting concept of convergence, let us next focus on maps between normed vector spaces.

Definition 6.3.1: Linear operators and functionals

Let X and Y be two vector spaces over $\mathbb{R}$.

(a) A map $L : X \to Y$ is called an **operator**.

(b) If the map $L : X \to Y$ satisfies

$$\begin{aligned} L(x_1 + x_2) &= L(x_1) + L(x_2), \\ L(\alpha x_1) &= \alpha L(x_1) \end{aligned}$$

for all $x_1, x_2 \in X$ and $\alpha \in \mathbb{R}$ it is referred to as a **linear operator**.

(c) In the case of $Y = \mathbb{R}$ (linear) operators are usually referred to as (linear) **functionals**.

Note that for a linear operator/functional we always have

$$L(0) = 0.$$

We have already seen that many collections of important objects in mathematics can be seen as vector spaces, e.g. different function spaces. For the same reason, (linear) operators are considered such a powerful tool. Many important operations, for instance differentiation and integral operators, can all be interpreted as linear operators between certain function spaces.

Example 6.3.2

(a) The differential operator $\frac{\mathrm{d}}{\mathrm{d}x}$ can be considered as a linear operator L defined by

$$L : C^1(\mathbb{R}) \to C^0(\mathbb{R}), \quad L(f) = f'.$$

(b) Integration of a function f, let's say over $[a, b]$, can be considered as applying a linear functional F given by

$$F : C([a, b]) \to \mathbb{R},\ F(u) = \int_a^b u(t)\,\mathrm{d}t.$$

Next, we note a property for operators which is strongly connected to continuity, yet often much easier to check.

Definition 6.3.3

Let $(X, \|\cdot\|_X)$ and $(Y, \|\cdot\|_Y)$ be normed vector spaces. An operator $L : X \to Y$ is called **bounded** if

$$\exists C > 0 \ \forall x \in X : \ \ \|Lx\|_Y \leq C\|x\|_X$$

holds. The number

$$\|L\| := \sup_{x \in X \setminus \{0\}} \frac{\|Lx\|_Y}{\|x\|_X}$$

is called **norm of** L.

One of the most pleasant properties of linear operators is the following equivalence between their continuity and boundedness. This equivalence allows us to further investigate boundedness of a linear operator, which is often considerably easier, instead of directly addressing continuity.

Theorem 6.3.4

Let $(X, \|\cdot\|_X)$ and $(Y, \|\cdot\|_Y)$ be normed vector spaces and let $L : X \to Y$ be a linear operator. The following properties are equivalent:

(i) L is bounded

(ii) L is continuous

(iii) L is continuous in 0

Proof. We proceed in three steps.

(i) $\Longrightarrow$ (ii): Since L is linear and bounded, we have

$$\|Lx - Ly\|_Y = \|L(x - y)\|_Y \leq \|L\|\|x - y\|_X.$$

Hence, L is Lipschitz continuous and therefore continuous.

(ii) $\Longrightarrow$ (iii): Trivial!

(iii) $\implies$ (i): Since L is continuous in 0, we have

$$\forall \varepsilon > 0 \ \exists \delta > 0 : \ \|x\|_X \leq \delta \implies \|Lx\|_Y \leq \varepsilon.$$

In particular, for $\varepsilon = 1$, there exists an $\delta > 0$ such that

$$\|x\|_X \leq \delta \implies \|Lx\|_Y \leq 1.$$

This yields for $x \neq 0$:

$$\left\| \frac{\delta x}{\|x\|_X} \right\|_X \leq \delta \implies \left\| \frac{\delta L x}{\|x\|_X} \right\|_Y \leq 1$$

Finally, utilizing that $\| \cdot \|_Y$ is absolutely homogeneous, we get

$$\|Lx\|_Y \leq \delta^{-1} \|x\|_X$$

for all $x \in X$, which shows that L is bounded.

□

Example 6.3.5

(a) Let us consider the linear functional

$$F : (C([a,b]), \| \cdot \|_\infty) \to (\mathbb{R}, | \cdot |), \quad F(u) := \int_a^b u(t) \, \mathrm{d}t.$$

This linear functional is bounded since

$$|Fu| \leq \int_a^b |u(t)| \, \mathrm{d}t \leq (b-a) \|u\|_\infty.$$

This observation also suggests that the norm of F is given by $\|F\| = b - a$. It is obvious that $\|F\| \leq b - a$ but we could also have estimated $|Fu| \leq 2(b-a)\|u\|_\infty$ and therefore concluded that $\|F\| \leq 2(b-a)$. Thus, so far, we just know that $b - a$ is an upper bound for $\|F\|$. To prove that not just $\|F\| \leq b - a$ but also $\|F\| \geq b - a$ (and therefore $\|F\| = b - a$), we have to show that there exists a $u \in C([a,b])$ such that $|Fu| \geq b - a$. Therefor, let us consider the simple example $u \equiv 1$, yielding $|Fu| = b - a$. Hence,

$$\|F\| = b - a.$$

Note that by Theorem 6.3.4 the linear functional F is also continuous.

(b) Next, let us consider the linear functional

$$F : (C^1([-1,1]), \| \cdot \|_\infty) \to (\mathbb{R}, | \cdot |), \quad F(u) := u'(0).$$

In fact, this linear functional is not bounded. To show this, let us consider functions of the form

$$u_n(t) = \sin(n \pi t), \ n \in \mathbb{N}.$$

For these, we have

$$\|u\|_\infty = 1, \quad u_n'(t) = n\pi \cos(n \pi t), \quad u_n'(0) = n\pi$$

and therefore

$$|Fu_n| = n\pi \|u\|_\infty.$$

Since this relation holds for every $n \in \mathbb{N}$, there can be no $C > 0$ such that

$$|Fu| \leq C\|u\|_\infty$$

for all $u \in C([-1,1])$.

Chapter 7

Calculus of Variations

The calculus of variations is concerned with optimizing, i. e. minimizing or maximizing functionals over some admissible class of functions.

7.1 Variational Problems

Loosely speaking, by calculus of variations we try to solve variational problems of the following form:

- *Given a (nonlinear) functional $F : A \to \mathbb{R}$, for which $x \in M$ is F minimal or maximal?*

From calculus and real analysis, we are familiar with this question in the case that $M \subset \mathbb{R}$. We would look for local minima or maxima by checking the necessary condition

$$F'(x) = 0$$

then. Yet, in many cases M is not simply given by a subset of the real numbers but, for instance, by a class of functions. A few examples for this —with important applications in mechanics, optics, and quantum mechanics— are provided below:

1. *Dido's problem:* Maximize the area of a surface that is given by a closed curve of fixed length. That is, for
$$A(u) = \int_0^1 u_2(t)u_1'(t)\,\mathrm{d}t, \quad l(u) = \int_0^1 |u'(t)|\,\mathrm{d}t,$$
maximize $A(u)$ under the constraint that $l(u) = 1$.

2. *The Brachistochrone* (curve of fastest decent): Given are two points $A = (0, 0)$ and $B = (B_1, B_2)$ with $B_2 < 0$. Along which curve does an object (frictionless, under the influence of gravity) get from A to B the fastest? That is, for which $u : [0, 1] \to \mathbb{R}^2$ is
$$T(u) = \int_0^1 \frac{\|u'(t)\|_2}{\sqrt{u_2(t)}}\,\mathrm{d}t$$
minimized under the constraint, that $u(0) = A$ and $u(1) = B$?

3. *Geodesics:* Curves representing — in some sense — the shortest path between two points in a surface, or more generally in Riemann manifold.

4. Eigenvalue problems: For instance, the solutions of the *Sturm–Liouville equation*
$$Lu(t) := \frac{\mathrm{d}}{\mathrm{d}t}\left(-p(t)\frac{\mathrm{d}u}{\mathrm{d}t}(t)\right) + q(t)u(t) = \lambda u(t)$$

correspond to solutions of the variational problem to minimize

$$F(u) = \int_a^b p(t)u'(t) + q(t)u(t)^2 \, \mathrm{d}t$$

under the (isoperimetric) constraint

$$\int_a^b u(t)^2 \, \mathrm{d}t = 1.$$

7.2 Necessary conditions for extrema

Considering the functions $f : \mathbb{R} \to \mathbb{R}$ we know from real analysis that $f'(x) = 0$ is a necessary condition for x to be an extrema. In fact, this idea can be generalized for functionals $F : A \to \mathbb{R}$, at least in a certain sense. At the same time we should remember from multidimensional calculus that there can already be different concepts of derivatives for functions $f : \mathbb{R}^n \to \mathbb{R}^m$. Usually, we distinguish between

1. the total derivative (best linear approximation) and
2. the directional derivative (rate of change in a certain direction, partial derivative).

Indeed, both concepts can be generalized for functionals $F : A \to \mathbb{R}$ on general normed vector spaces. The resulting generalization of the total derivative is referred to as *Fréchet derivative* while the generalization of the directional derivative is referred to as *Gâteaux derivative*. In the context of calculus of variations, however, we are only interested in the Gâteaux derivative.

Definition 7.2.1: The Gâteaux derivative

Let $(X, \|\cdot\|_X)$ and $(Y, \|\cdot\|_Y)$ be normed vector spaces and $F : X \to Y$ an operator between them. The operator

$$\delta_X F : X \to Y, \quad h \mapsto \delta_X F(h),$$

is called **Gâteaux differential** on F in x if

$$\lim_{t \to 0} \left\| \frac{F(x+th) - F(x)}{t} - \delta_X F(h) \right\| = 0 \quad \forall h \in X.$$

Moreover, if $\delta_X F$ is linear and continuous (or bounded) it is called the **Gâteaux derivative** of F in x and F is said to be **Gâteaux differentiable** in x.

Remark 7.2.2

Note that if $F : \mathbb{R}^n \to \mathbb{R}^m$ is Gâteaux differentiable in x, F is also partially differentiable in x with

$$\delta_x F(h) = J_F(x)h \quad \forall h \in \mathbb{R}^n,$$

where $J_F(x) \in \mathbb{R}^{m \times n}$ denotes the Jacobi matrix of F in x.

Here, we are only interested in functionals and therefore in the case $Y = \mathbb{R}$. Note that in this case, for a fixed $h \in X$, the Gâteaux derivative $\delta_x F(h)$ essentially is the derivative of the real valued function

$$\mathbb{R} \to \mathbb{R}, \quad t \mapsto F(x+th).$$

This observation yields the definition of variations.

Definition 7.2.3: Variations of functionals

Let $(X, \|\cdot\|)$ be a normed vector space, $A \subseteq X$, and $F : A \to \mathbb{R}$ a functional on X. Let $x \in A$ and $h \in X$ such that

$$x + th \in A$$

for all sufficiently small t. Then, the **n-th (Gâteaux-) variation** of F at x in the direction of h is defined as

$$\delta_x^n F(h) = \left[\frac{\mathrm{d}^n}{\mathrm{d}t^n} F(x+th) \right]_{t=0}, \tag{7.1}$$

provided that the derivative exists. Directions h for which (7.1) exists are called **admissible variations** of F at x and are denoted by $\operatorname{adm}(F,x)$.

Note that the variation of a functional can be considered as an approximation of the local behavior of that functional. If $\operatorname{adm}(F,x) = X$, the function $\delta_x(h)$ is exactly the Gâteaux derivative of F at x. Hence, variations of F can be interpreted as special directional derivatives. Next, let us summarize some elemental properties of the first (1-st) variation.

Lemma 7.2.4: Linearity of the first variation

Let $F : X \supset A \to \mathbb{R}$ be a functional, $x \in X$, and $h \in \operatorname{adm}(F,x)$.

(a) The first variation is homogeneous:

$$\delta_x F(\lambda h) = \lambda \delta_x F(h) \quad \forall \lambda \in \mathbb{R}.$$

(b) Let $G : A \to \mathbb{R}$ be a functional such that $\delta_x G(h)$ exists. The first variation is additive:

$$\delta_x (F+G)(h) = \delta_x F(h) + \delta_x G(h).$$

We can now devote our attention to extrema of functions and how these can be found using the first variation.

Definition 7.2.5: Extrema

Let $(X, \|\cdot\|)$ be a normed vector space. The functional $F : X \supset A \to \mathbb{R}$ is said to have a **(local) minimum** at $x \in A$ if there exists a neighborhood U around x such that

$$F(\xi) \geq F(x)$$

for all $\xi \in U$. We say that F has a **(local) maximum** at $x \in A$ if x is a minimum of $-F$. Finally, we refer to $x \in A$ as a **(local) extremum** of F if x is a minimum or maximum of F.

Remark 7.2.6

If a point x is an extremum of F, in particular, depends on which sets $U \subset X$ are neighborhoods around x ($x \in U$ and U is open). In general, this will also depend on the norm $\|\cdot\|$ on X.

Following a well-known argument from real analysis, we note that if F has an extremum at x, for $h \in \operatorname{adm}(F,x)$, the real-valued function

$$\mathbb{R} \to \mathbb{R}, \quad t \mapsto F(x+th)$$

has an extremum in $t = 0$ then. This already yields the following necessary condition for extrema of functionals.

Theorem 7.2.7: A necessary condition for extrema

Let $F : X \supset A \to \mathbb{R}$ be a functional, $x \in A$, and $h \in \operatorname{adm}(F, x)$. If x is an extremum of F, then

$$\delta_x F(h) = 0$$

holds.

Remark 7.2.8

The converse statement of Theorem 7.2.7 does not hold. That is, $\delta_x F(h) = 0$ is not a sufficient condition for x being an extremum.

Definition 7.2.9

Let $F : X \supset A \to \mathbb{R}$ be a functional, $x \in A$, and $V \subset \operatorname{adm}(F, x)$. If $\delta_x F(h) = 0$ holds for all $h \in V$, we call x an **extremal** of F (with respect to V).

Note that since Theorem 7.2.7 only provides a necessary condition, we are not guaranteed that extremals will actually provide extrema. Yet, we consider extremals as the candidates for extrema. To actually find extrema of a variational problem it is often convinient to follow the following strategy:

- Determine extremals and investigate which of these are actually extrema.

Example 7.2.10

Given is the functional

$$F(x) = \int_0^1 1 + x'(s)^2 \, \mathrm{d}s$$

with $x \in C^1([0,1])$ satisfying $x(0) = 0$ and $x(1) = 1$. Hence, we have $F : A \to \mathbb{R}$ with

$$A := \{\, x \in C^1([0,1]) \mid x(0) = 0, x(1) = 1 \,\} \subset C^1([0,1]).$$

Let $x \in A$ and $h \in C^1([0,1])$ be given by

$$x(s) = s \quad \text{and} \quad h(s) = s(1-s).$$

We would like to determine the first variation of F in x_0 in the direction of h:

$$\begin{aligned} F(x + th) &= \int_0^1 1 + \left[x'(s) + th'(s)\right]^2 \, \mathrm{d}s \\ &= \int_0^1 1 + \left[1 + t(1-2s)\right]^2 \, \mathrm{d}s \\ &= \int_0^1 2 + 2t(1-2s) + t^2(1-2s)^2 \, \mathrm{d}s \\ &= 2 + \frac{t^2}{3} \end{aligned}$$

Hence, we have

$$\frac{\mathrm{d}}{\mathrm{d}t} F(x + th) = \frac{2}{3} t$$

and therefore

$$\delta_x F(h) = \left[\frac{\mathrm{d}}{\mathrm{d}t} F(x+th)\right]_{t=0} = 0.$$

This shows that $x(s) = s$ is an extremal of F.

7.3 The Euler–Lagrange equation

Perhaps the simplest problem in the calculus of variations is to find a (local) minimum of functionals of the form

$$F(x) = \int_a^b L(s, x, x')\,\mathrm{d}s, \tag{7.2}$$

where $x \in C^2([a,b])$ with $x(a) = x_a$ and $x(b) = x_b$. Here, $L : [a,b] \times \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is a given, twice differentiable function, called the **Lagrangian**. Our goal is to reformulate the necessary condition in Theorem 7.2.7 in a more practical manner for the specific functional (7.2):

Let x be a local minimum and $h \in C^2([a,b])$ such that $h(a) = h(b) = 0$. Then, $x + th \in C^2([a,b])$ satisfies

$$(x+th)(a) = x_a, \quad (x+th)(b) = x_b$$

and is therefore an admissible function. Moreover, we have

$$F(x+th) = \int_a^b L(s, x+th, x'+th')\,\mathrm{d}s$$

and therefore

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t} F(x+th) &= \int_a^b \frac{\partial}{\partial t} L(s, x+th, x'+th')\,\mathrm{d}s \\ &= \int_a^b L_x(s, x+th, x'+th')h + L_{x'}(s, x+th, x'+th')h'\,\mathrm{d}s,\end{aligned}$$

where L_x denotes the partial derivative $\frac{\partial}{\partial x} L(s,x,x')$ and $L_{x'}$ the partial derivative $\frac{\partial}{\partial x'} L(s,x,x')$. Thus, the first variation of F at x in the direction of h is given by

$$\begin{aligned}\delta_x F(h) &= \left[\frac{\mathrm{d}}{\mathrm{d}t} F(x+th)\right]_{t=0} \\ &= \int_a^b L_x(s,x,x')h + L_{x'}(s,x,x')h'\,\mathrm{d}s.\end{aligned}$$

This shows the following.

Lemma 7.3.1

A necessary condition for x being a local minimum (or maximum) of F given by (7.2) is

$$\int_a^b L_x(s,x,x')h + L_{x'}(s,x,x')h'\,\mathrm{d}s = 0 \tag{7.3}$$

to hold for all $h \in C^2([a,b])$ with $h(a) = h(b) = 0$.

Unfortunately, condition (7.3) is just of limited use for determining x. Yet, using the fact that it has to hold for all h, we can simplify the condition. Note that applying integration by parts to (7.3) yields

$$\int_a^b \left[L_x(s,x,x') - \frac{\mathrm{d}}{\mathrm{d}s} L_{x'}(s,x,x') \right] h \, \mathrm{d}s + \left[L_{x'}(s,x,x') h \right]_{s=a}^b = 0$$

and utilizing the fact, that $h(a) = h(b) = 0$, we get

$$\int_a^b \left[L_x(s,x,x') - \frac{\mathrm{d}}{\mathrm{d}s} L_{x'}(s,x,x') \right] h \, \mathrm{d}s = 0 \tag{7.4}$$

to hold for all $h \in C^2([a,b])$ with $h(a) = h(b) = 0$ as a necessary condition for x to be a local minimum of F. The following lemma, which goes back to Lagrange, provides the final step in deriving a practical necessary condition.

Lemma 7.3.2

Let $f \in C([a,b])$. If

$$\int_a^b f(s) h(s) \, \mathrm{d}s = 0$$

holds for all $h \in C^2([a,b])$ with $h(a) = h(b) = 0$, then $f(s) = 0$ for all $s \in [a,b]$.

Proof. We prove this by contradiction. Assume there exists an $s_0 \in (a,b)$ such that $f(s_0) \neq 0$, where we assume $f(s_0) > 0$ without loss of generality. Since f is continuous, we can find an $\varepsilon > 0$ such that $f(s) > \varepsilon$ for $s_0 - \delta \leq s \leq s_0 + \delta$ then. Hence, for $s_1 = s_0 - \delta$, $s_2 = s_0 + \delta$, and

$$h(s) = \begin{cases} (s - s_1)^3 (s_2 - s)^3 & \text{if } s_1 \leq s \leq s_2, \\ 0 & \text{otherwise,} \end{cases}$$

we have

$$\int_a^b f(s) h(s) \, \mathrm{d}s \geq \varepsilon \int_{s_1}^{s_2} h(s) \, \mathrm{d}s > 0,$$

since $h(s) \geq 0$ for all $s \in [a,b]$. □

In fact, the above lemma is just a special version of what is called the *fundamental lemma* of the calculus of variations (CoV).

Theorem 7.3.3: Fundamental lemma of the CoV

Let $\Omega \subset \mathbb{R}^n$ and $f \in C(\Omega)$. If

$$\int_\Omega f(s) \varphi(s) \, \mathrm{d}s = 0$$

for all $\varphi \in C_C^\infty(\Omega)$, then $f(s) = 0$ for all $s \in \Omega$. Here, $\varphi \in C_C^\infty(\Omega)$ means that φ is smooth and compactly supported on Ω.

Finally, applying Lemma 7.3.2 to (7.4) yields the following result.

Theorem 7.3.4: The Euler–Lagrange equation

Let $L : [a,b] \times \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ be a twice continuously differentiable function. If a function $x \in C^2([a,b])$

with $x(a) = x_a$ and $x(b) = x_b$ provides a local minimum (or maximum) for the functional

$$F(x) = \int_a^b L(s, x, x')\,\mathrm{d}s,$$

then x must satisfy the differential equation

$$L_x(s, x, x') = \frac{\mathrm{d}}{\mathrm{d}s} L_{x'}(s, x, x'), \quad x \in (a, b). \tag{7.5}$$

Equation (7.5) is called the **Euler–Lagrange equation**.

Note that the Euler–Lagrange equation is a second order equation and potentially nonlinear. This can be seen by writing the derivatives in (7.5) using the chain rule:

$$L_x(s, x, x') - L_{x's}(s, x, x') - L_{x'x}(s, x, x')x' - L_{x'x'}(s, x, x')x'' = 0.$$

It represents a necessary condition for a local minimum and it is analogous to the derivative condition $f'(x) = 0$ in differential calculus. Thus, its solutions are not necessarily local minima. Yet, the solutions of the Euler–Lagrange equation (7.5) are extremals, i. e. candidates for local minima. Next, let us consider two examples to demonstrate the application of the Euler–Lagrange equation.

Example 7.3.5

Given the functional

$$F : A \to \mathbb{R}, \quad F(x) = \int_0^1 x'(s)^2 + 3x(s) + 2s\,\mathrm{d}s$$

with

$$A = \left\{ x \in C^2([0, 1]) \mid x(0) = 0, x(1) = 1 \right\}.$$

We would like to find the extremals of F. Therefore, we can utilize Theorem 7.3.4, which tells us that the extremals of F are given as solutions of the corresponding Euler–Lagrange equation (7.5). Adapting the notation of Theorem 7.3.4, we have

$$\begin{aligned} L(s, x, x') &= (x')^2 + 3x + 2s, \\ L_x(s, x, x') &= 3, \\ L_{x'}(s, x, x') &= 2x' \end{aligned}$$

and the corresponding Euler–Lagrange equation is therefore given by

$$3 \overset{!}{=} \frac{\mathrm{d}}{\mathrm{d}s}\left(2x'(s)\right) = 2x''(s), \quad x \in (0, 1).$$

Integrating twice gives

$$x(s) = \frac{3}{4}s^2 + C_1 s + C_2$$

and incorporating the condition $x(0) = 0$ and $x(1) = 1$ yields $C_2 = 0$ and $C_1 = \frac{1}{4}$. Hence, the single extremal of F in A (i. e. they satisfy the restrictions $x(0) = 0$ and $x(1) = 1$) is

$$x(s) = \frac{3}{4}s^2 + \frac{1}{4}s.$$

Note once more that x is only a candidate for a local extremum so far. Further calculations are required to determine whether x maximizes or minimizes F.

Example 7.3.6

Next, let us consider the *arclength functional*

$$F : A \to \mathbb{R}, \quad F(x) = \int_a^b \sqrt{1 + x'(s)^2}\, \mathrm{d}s,$$

where we

$$A = \left\{ x \in C^2([a,b]) \mid x(a) = x_a, x(b) = x_b \right\}.$$

Again, applying Theorem 7.3.4 and adapting its notation, we have

$$\begin{aligned} L(s,x,x') &= \sqrt{1 + x'(s)^2}, \\ L_x(s,x,x') &= 0, \\ L_{x'}(s,x,x') &= \frac{x'}{\sqrt{1 + x'(s)^2}} \end{aligned}$$

and a necessary condition for $x \in A$ being a local minimum (or maximum) of F is given by the corresponding Euler–Lagrange equation

$$\frac{\mathrm{d}}{\mathrm{d}s} \left(\frac{x'(s)}{\sqrt{1 + x'(s)^2}} \right) = 0.$$

Integration yields

$$\frac{x'(s)}{\sqrt{1 + x'(s)^2}} = C$$

for some constant $C \in \mathbb{R}$. Next, note that

$$\begin{aligned} \frac{x'(s)}{\sqrt{1 + x'(s)^2}} = C \iff x'(s)^2 &= C + Cx'(s)^2 \\ \iff x'(s)^2 &= \frac{C}{1-C} \\ \iff x'(s) &= B \end{aligned}$$

with constant $B = \sqrt{\frac{C}{1-C}} \in \mathbb{R}$. Hence, integration yields

$$x(s) = sB + \tilde{B}$$

and the restrictions $x(a) = x_a$ and $x(b) = x_b$ give us

$$B = \frac{x_b - x_a}{b - a}, \quad \tilde{B} = \frac{bx_a - ax_b}{b - a}$$

and therefore the unique extremal

$$x(s) = \frac{1}{b-a} \left[s(x_b - x_a) + (bx_a - ax_b) \right].$$

Note that this is a straight line connecting (a, x_a) and (b, x_b). This can be considered as a proof of the well-known geometric fact that the shortest connection between two points is a straight line.

7.4 Some Special Cases

The Euler–Lagrange equation (7.5) is a second order ordinary equation. Yet, in special cases, when the Lagrangian $L = L(s, x, x')$ does not explicitly depend on one of its variables, the Euler–Lagrange equation (7.5) becomes significantly simpler.

Lemma 7.4.1

Given is a functional F of the form (7.2) with Lagrangian $L = L(s, x, x')$ and the corresponding Euler–Lagrange equation (7.5).

(a) If $L = L(s, x)$, then the Euler–Lagrange equation becomes

$$L_x(s, x) = 0,$$

which is an algebraic equation.

(b) If $L = L(s, x')$, then the Euler–Lagrange equation becomes

$$L_{x'}(s, x') = C, \tag{7.6}$$

where $C \in \mathbb{R}$ is an arbitrary constant.

(c) If $L = L(x, x')$, then the Euler–Lagrange equation becomes

$$L(x, x') - x' L_{x'}(x, x') = C, \tag{7.7}$$

where $C \in \mathbb{R}$ is an arbitrary constant.

Proof. Remember that the Euler–Lagrange equation (7.5) is given by

$$L_x(s, x, x') = \frac{\mathrm{d}}{\mathrm{d}s} L_{x'}(s, x, x').$$

(a) If $L = L(s, x)$, we have

$$\begin{aligned} L_x(s, x, x') &= L_x(s, x), \\ L_{x'}(s, x, x') &= 0, \\ \frac{\mathrm{d}}{\mathrm{d}s} L_{x'}(s, x, x') &= 0 \end{aligned}$$

and therefore

$$L_x(s, x) = 0.$$

(b) If $L = L(s, x')$, we have

$$\begin{aligned} L_x(s, x, x') &= 0, \\ L_{x'}(s, x, x') &= L_{x'}(s, x'), \\ \frac{\mathrm{d}}{\mathrm{d}s} L_{x'}(s, x, x') &= \frac{\mathrm{d}}{\mathrm{d}s} L_{x'}(s, x') \end{aligned}$$

and therefore

$$0 = \frac{\mathrm{d}}{\mathrm{d}s} L_{x'}(s, x').$$

Hence, integrating with respect to s yields

$$L_{x'}(s, x') = C$$

for $C \in \mathbb{R}$.

(c) If $L = L(x, x')$, we have

$$\begin{aligned} L_x(s, x, x') &= L_x(x, x'), \\ L_{x'}(s, x, x') &= L_{x'}(x, x'), \\ \frac{\mathrm{d}}{\mathrm{d}s} L_{x'}(s, x, x') &= \frac{\mathrm{d}}{\mathrm{d}s} L_{x'}(x, x') \end{aligned}$$

and therefore

$$L_{x'}(x, x') = \frac{\mathrm{d}}{\mathrm{d}s} L_{x'}(x, x').$$

This yields

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}s}\left[L(x, x') - x'L(x, x')\right] &= \frac{\mathrm{d}}{\mathrm{d}s} L(x, x') - x' \frac{\mathrm{d}}{\mathrm{d}s} L_{x'}(x, x') - x'' L_{x'}(x, x') \\ &= x' L_x(x, x') + x'' L_{x'}(x, x') - x' \frac{\mathrm{d}}{\mathrm{d}s} L_{x'}(x, x') - x'' L_{x'}(x, x') \\ &= x' \left[L_x(x, x') - \frac{\mathrm{d}}{\mathrm{d}s} L_{x'}(x, x') \right] \\ &= 0 \end{aligned}$$

and integrating with respect to s provides us with the assertion.

□

Example 7.4.2: The Brachistochrome

In this example, we determine the extremals of the Brachistochrome (curve of the fastest descent) problem.

Problem statement

A bead of mass m with initial velocity zero slides with no friction under the force of gravity g from a point $(0, b)$ to a point $(a, 0)$ along a wire defined by a curve $y = y(x)$ in the xy-plane. We would like to figure out which curve connecting $(0, b)$ and $(a, 0)$ yields the fastest time of descent. To formulate this problem analytically, first, we compute the time of descent T for a fixed curve y:

$$T(y) = \int_a^b \frac{\sqrt{1 + y'(x)^2}}{v(x)} \,\mathrm{d}x.$$

Here, $v = v(x)$ denotes the velocity of the bead along the curve y. Using the fact that the energy is conserved, we have

$$\begin{aligned} &(\text{kinetic energy at } t > 0) + (\text{potential energy at } t > 0) \\ =\ &(\text{kinetic energy at } t = 0) + (\text{potential energy at } t = 0). \end{aligned}$$

Expressed in our notation, this means

$$\frac{1}{2} m v^2 + mgy = 0 + mgb,$$

which yields

$$v(x) = \sqrt{2g(b - y(x))}.$$

Thus, the time required for the bead to descend is

$$T(y) = \int_a^b \frac{\sqrt{1 + y'(x)^2}}{\sqrt{2g(b - y(x))}} \, \mathrm{d}x. \tag{7.8}$$

Note that we can now reformulate the Brachistochrome problem as finding the minimum of the functional

$$T : A \to \mathbb{R}_0^+, \quad y \mapsto T(y)$$

under the restriction $y(0) = b$ and $y(a) = 0$; That is,

$$A = \left\{ y \in C^2([0, a]) \mid y(0) = b, y(a) = 0 \right\}.$$

Finding extremals

Next, let us determine the extremals of the function $T : A \to \mathbb{R}_0^+$ given by (7.8). Once more, note that by Theorem 7.3.4 the extremals are given as solutions of the Euler–Lagrange equation (7.5) with Lagrangian

$$L(x, y, y') = \frac{\sqrt{1 + y'(x)^2}}{\sqrt{2g(b - y(x))}}.$$

This Lagrangian does not depend on x, i. e. $L = L(y, y')$, and Lemma 7.4.1 therefore tell us that the corresponding Euler–Lagrange equation reduces to

$$\begin{aligned} C &\overset{!}{=} L(y, y') - y' L_{y'}(y, y') \\ &= \frac{\sqrt{1 + y'(x)^2}}{2g(b - y(x))} - y'(x) \frac{1}{\sqrt{2g(b - y'(x))}} \frac{y'(x)}{\sqrt{1 + y'(x)^2}}. \end{aligned}$$

The above equation is equivalent to

$$(y')^2 = \frac{1 - \tilde{C}(b - y)}{\tilde{C}(b - y)}$$

with constant $\tilde{C} \in \mathbb{R}$. Taking the square root of both sides and separating variables gives

$$\mathrm{d}x = -\frac{\sqrt{b - y}}{\sqrt{C_1 - (b - y)}} \mathrm{d}y, \quad C_1 = \tilde{C}^{-1},$$

where the minus sign is taken because $\frac{\mathrm{d}y}{\mathrm{d}x} < 0$. The last equation can be integrated by making the trigonometric substitution

$$b - y = C_1 \sin^2\left(\frac{\phi}{2}\right). \tag{7.9}$$

Then, one obtains

$$\mathrm{d}x = C_1 \sin^2\left(\frac{\phi}{2}\right) \mathrm{d}\phi = \frac{C_1}{2} (1 - \cos\phi) \, \mathrm{d}\phi,$$

which yields

$$x(\phi) = \frac{C_1}{2}(\phi - \sin\phi) + C_2. \tag{7.10}$$

Equations (7.9) and (7.10) are parametric equations for a cycloid. Here, in contrast to the problem of finding the curve of shortest length between two points (see Example 7.3.6), it is not clear that the cycloids just obtained actually minimize the functional $T : A \to \mathbb{R}$. Further calculations would be required for confirmation.

7.5 Outlook on Possible Generalizations

We end this chapter by addressing some possible generalizations of the theory presented in Chapter 7.3 concerning functionals of the form

$$F(x) = \int_a^b L(s, x, x') \, \mathrm{d}s$$

with end point conditions $x(a) = x_a$ and $x(b) = x_b$. These generalizations include

- higher derivatives,
- several functions, and
- natural boundary conditions.

In particular, we investigate how the Euler–Lagrange, as a means to determine extremals, change in these cases.

7.5.1 Higher Derivatives

So far, we have considered Lagrangians of the form $L = L(s, x, x')$. An obvious generalization is to include higher derivatives in the Lagrangian. For instance, let us consider the second order problem

$$F : A \to \mathbb{R}, \ F(x) = \int_a^b L(s, x, x', x'') \, \mathrm{d}s,$$

where

$$A = \left\{ x \in C^4([a,b]) \mid x(a) = A_1, x'(a) = A_2, x(b) = B_1, x'(b) = B_2 \right\}.$$

Moreover, the function $L : [a,b] \times \mathbb{R}^3 \to \mathbb{R}$ is assumed to be twice continuously differentiable in each of its arguments. Even though we have now included the second derivative x'' in the Lagrangian, we can proceed quite similar to Chapter 7.3. Again, our goal is to reformulate the necessary condition $\delta_x F(h) = 0$ for all $h \in \operatorname{adm}(F, x)$ provided by Theorem 7.2.7 in a more practical manner:

Let x be a local extremum and $h \in C^4([a,b])$ such that

$$h(a) = h'(a) = h(b) = h'(b) = 0. \tag{7.11}$$

Then, $x + th \in A$ and

$$F(x + th) = \int_a^b L(s, x + th, x' + th', x'' + th'') \, \mathrm{d}s.$$

Hence, we have

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}s} F(x + th) &= \int_a^b \frac{\partial}{\partial t} L(s, x + th, x' + th', x'' + th'') \, \mathrm{d}s \\ &= \int_a^b L_x(s, x + th, x' + th', x'' + th'') h + L_{x'}(s, x + th, x' + th', x'' + th'') h' \\ &\quad + L_{x''}(s, x + th, x' + th', x'' + th'') h'' \, \mathrm{d}s \end{aligned}$$

and the first variation of F at x in the direction of h is therefore given by

$$\begin{aligned} \delta_x F(h) &= \left[\frac{\mathrm{d}}{\mathrm{d}t} F(x + th) \right]_{t=0} \\ &= \int_a^b L_x(s, x, x', x'') h + L_{x'}(s, x, x', x'') h' + L_{x''}(s, x, x', x'') h'' \, \mathrm{d}s. \end{aligned}$$

Next, applying integration by parts, we get

$$\delta_x F(h) = \int_a^b \left[L_x - \frac{\mathrm{d}}{\mathrm{d}s} L_{x'} + \frac{\mathrm{d}^2}{\mathrm{d}s^2} L_{x''} \right] h \, \mathrm{d}s,$$

since $h(a) = h'(a) = h(b) = h'(b) = 0$. Thus, by Theorem 7.2.7, we can note that

$$\int_a^b \left[L_x - \frac{\mathrm{d}}{\mathrm{d}s} L_{x'} + \frac{\mathrm{d}^2}{\mathrm{d}s^2} L_{x''} \right] h \, \mathrm{d}s = 0$$

holds for all $h \in C^4([a,b])$ satisfying (7.11) if x is a (local) extremum. The fundamental lemma of the CoV (Theorem 7.3.3) therefore yields the following result.

Theorem 7.5.1

Let $L : [a,b] \times \mathbb{R}^3 \to \mathbb{R}$ be a twice continuously differentiable function. If a function $x \in C^4([a,b])$ with

$$x(a) = A_1, \;\; x'(a) = A_2, \;\; x(b) = B_1, \;\; x'(b) = B_2$$

provides a local extremum for the functional

$$F(x) = \int_a^b L(s, x, x', x'') \, \mathrm{d}s,$$

then x must satisfy the (generalized) Euler–Lagrange equation

$$L_x(s, x, x', x'') - \frac{\mathrm{d}}{\mathrm{d}s} L_{x'}(s, x, x', x'') + \frac{\mathrm{d}^2}{\mathrm{d}s^2} L_{x''}(s, x, x', x'') = 0.$$

It is now easy to verify that the n-th order variational problem

$$F(x) = \int_a^b L(s, x, x', \dots, x^{(n)}) \, \mathrm{d}s$$

with $x \in C^{2n}([a,b])$ satisfying

$$\begin{aligned} x(a) &= A_1, \quad & x'(a) &= A_2, \quad & \dots, \quad & x^{(n)}(a) = A_n, \\ x(b) &= B_1, \quad & x'(b) &= B_2, \quad & \dots, \quad & x^{(n)}(b) = B_n, \end{aligned}$$

results in the (generalized) Euler–Lagrange equation

$$L_x - \frac{\mathrm{d}}{\mathrm{d}s} L_{x'} \pm \dots + (-1)^n \frac{\mathrm{d}^n}{\mathrm{d}s^n} L_{x^{(n)}}. \tag{7.12}$$

Thus, (7.12) is a necessary condition for x being a (local) extremum of F.

7.5.2 Several Functions

Another possible generalization is to consider Lagrangians which depend not on a single function x and its derivatives but on several functions $x_1, \dots, x_n$ and their derivatives. To illustrate this, let $n = 2$ and let us focus on functionals of the form

$$F(x_1, x_2) = \int_a^b L(s, x_1, x_2, x_1', x_2') \, \mathrm{d}s,$$

where $x_1, x_2 \in C^2([a,b])$ satisfying the boundary conditions

$$\begin{aligned} x_1(a) &= A_1, \quad x_1(b) = B_1, \\ x_2(a) &= A_2, \quad x_2(b) = B_2. \end{aligned}$$

Once more, we can derive some kind of Euler–Lagrange equation which solutions are the extremals of F by following similar arguments as presented in Chapter 7.2 and 7.4:

Let (x_1, x_2) be a local extremum of F and let $h_1, h_2 \in C^2([a,b])$ such that

$$h_1(a) = h_2(b) = h_1(a) = h_2(b) = 0. \tag{7.13}$$

Then, the pair of functions $(x_1 + th_1, x_2 + th_2)$ is an admissible argument of F and we have

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} F(x_1 + th_1, x_2 + th_2) &= \int_a^b \frac{\partial}{\partial t} L(s, x_1 + th_1, x_2 + th_2, x_1' + th_1', x_2 + th_2') \,\mathrm{d}s \\ &= \int_a^b L_{x_1}(s, x_1 + th_1, x_2 + th_2, x_1' + th_1', x_2' + th_2') h_1 \\ &\qquad + L_{x_2}(s, x_1 + th_1, x_2 + th_2, x_1' + th_1', x_2' + th_2') h_2 \\ &\qquad + L_{x_1'}(s, x_1 + th_1, x_2 + th_2, x_1' th_1', x_2' + th_2') h_1' \\ &\qquad + L_{x_2'}(s, x_1 + th_1, x_2 + th_2, x_1' + th_1', x_2' + th_2') h_2' \,\mathrm{d}s \end{aligned}$$

and therefore

$$\begin{aligned} \delta_{x_1,x_2} F(h) &= \left[\frac{\mathrm{d}}{\mathrm{d}t} F(x_1 + th_1, x_2 + th_2) \right]_{t=0} \\ &= \int_a^b L_{x_1} h_1 + L_{x_2} h_2 + L_{x_1'} h_1' + L_{x_2'} h_2' \,\mathrm{d}s. \end{aligned}$$

Hence, integration by parts yields

$$\delta_{(x_1,x_2)} F(h) = \int_a^b \left[L_{x_1} - \frac{\mathrm{d}}{\mathrm{d}s} L_{x_1'} \right] h_1 + \left[L_{x_2} - \frac{\mathrm{d}}{\mathrm{d}s} L_{x_2'} \right] \mathrm{d}s.$$

Thus, if (x_1, x_2) is a (local) extrema of F, by Theorem 7.2.7,

$$\int_a^b \left[L_{x_1} - \frac{\mathrm{d}}{\mathrm{d}s} L_{x_1'} \right] h_1 + \left[L_{x_2} - \frac{\mathrm{d}}{\mathrm{d}s} L_{x_2'} \right] h_2 \,\mathrm{d}s = 0$$

has to hold for all $h_1, h_2 \in C^2([a,b])$ satisfying (7.13). The fundamental theorem of the CoV (Theorem 7.3.3) therefore yields the following result.

Theorem 7.5.2

Let $L : [a,b] \times \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$ be a twice continuously differentiable function. If a pair of functions $x_1, x_2 \in C^2([a,b])$ with

$$x_1(a) = A_1, \quad x_2(a) = A_2, \quad x_1(b) = B_1, \quad x_2(b) = B_2$$

provides a local extremum for the functional

$$F(x_1, x_2) = \int_a^b L(s, x_1, x_2, x_1', x_2') \,\mathrm{d}s,$$

then x_1 and x_2 must satisfy the (system of) Euler–Lagrange equations

$$\begin{aligned} L_{x_1}(s, x_1, x_2, x_1', x_2') &= \frac{\mathrm{d}}{\mathrm{d}s} L_{x_1'}(s, x_1, x_2, x_1', x_2'), \\ L_{x_2}(s, x_1, x_2, x_1', x_2') &= \frac{\mathrm{d}}{\mathrm{d}s} L_{x_2'}(s, x_1, x_2, x_1', x_2'). \end{aligned}$$

In general, if F depends on n functions,

$$F(x_1, \dots, x_n) = \int_a^b L(s, x_1, \dots, x_n, x_1', \dots, x_n') \, \mathrm{d}s,$$

where $x_i \in C^2([a,b])$ and $x_i(a) = A_i$, $x_i(b) = B_i$ for all $i = 1, \dots, n$, a necessary condition for the n-tupel $(x_1, \dots, x_n)$ to provide a local extremum of F is given by the system of Euler–Lagrange equations

$$L_{x_i} = \frac{\mathrm{d}}{\mathrm{d}s} L_{x_i'}$$

for $i = 1, \dots, n$.

7.5.3 Natural Boundary Condition

Finally, let us address natural boundary conditions instead of fixed boundary conditions, which we have considered so far. These can be motivated, for instance, by the following problem.

Example 7.5.3: Motivation for natural boundary conditions

A river with parallel straight banks b units apart has a stream velocity given by

$$\boldsymbol{v}(x, y) = \begin{pmatrix} 0 \\ v(x) \end{pmatrix}.$$

Assuming that one of the banks is the y-axis and that the point $(0,0)$ is the point of departure, what route should a boat take to reach the opposite bank in the shortest possible time? Assume that the speed of the boat in still water is $c \in \mathbb{R}^+$ with $c > v(x)$ for all x. This problem differs from those in earlier sections in that right-hand endpoint, the point of arrival on the line $x = b$, is not specified. Instead, it must be determined as part of the solution. It can be shown that the time required for the boat to cross the river along a given path $y = y(x)$ is

$$F(y) = \int_a^b \frac{\sqrt{c^2[1 + y'(x)^2] - v(x)^2} - v(x)y'(x)}{c^2 - v(x)^2} \, \mathrm{d}x.$$

Thus, the variational problem is to minimize F subject to the conditions

$$y(0) = 0, \;\; y(b) \text{ free}.$$

Such a problem is referred to as a **free endpoint problem**, and if y is an extremal, then a certain condition must hold at $x = b$.

Generally speaking, conditions of the above type are called **natural boundary conditions**. In fact, just as common as free endpoint problems are problems where both, the startpoint and endpoint, are unspecified. To outline the treatment of such variational problems, let us consider the problem

$$F(y) = \int_a^b L(x, y, y') \, \mathrm{d}x,$$

where $y \in C^2([a, b])$ with

$$y(a) = y_a, \quad y(b) \text{ free.}$$

Let y be a local extremum of F and let $h \in C^2([a, b])$ such that

$$h(a) = 0. \tag{7.14}$$

Then, $y + th$ is an admissible function and we have

$$\begin{aligned} \delta_y F(h) &= \left[\frac{\mathrm{d}}{\mathrm{d}t} F(x, y + th, y' + th') \right]_{t=0} \\ &= \int_a^b L_y h + L_{y'} h' \, \mathrm{d}x. \end{aligned}$$

Note that this time we only have $h(a) = 0$ and integration by parts therefore yields

$$\delta_y F(h) = \int_a^b \left[L_y - \frac{\mathrm{d}}{\mathrm{d}s} L_{y'} \right] h \, \mathrm{d}x + L_{y'}(b, y(b), y'(b)) h(b).$$

Thus, if y is a local extremum of F, then

$$\int_a^b \left[L_y - \frac{\mathrm{d}}{\mathrm{d}s} L_{y'} \right] h \, \mathrm{d}x + L_{y'}(b, y(b), y'(b)) h(b) = 0 \tag{7.15}$$

has to hold for all $h \in C^2([a, b])$ with $h(a) = 0$. Yet, in particular, (7.15) has to hold for all $h \in C^2([a, b])$ with $h(a) = h(b) = 0$. Hence, by the fundamental lemma of the CoV, y must satisfy the Euler–Lagrange equation

$$L_y(x, y, y') = \frac{\mathrm{d}}{\mathrm{d}x} L_{y'}(x, y, y').$$

In addition, however,

$$L_{y'}(b, y(b), y'(b)) = 0$$

has to hold as well. Then, (7.15) is ensured to hold for all $h \in C^2([a, b])$ with $h(a) = 0$. This observation is summarized in the following theorem.

Theorem 7.5.4

Let $L : [a, b] \times \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ be a twice continuously differentiable function. If a function $y \in C^2([a, b])$ with

$$y(a) = y_a, \quad y(b) \text{ free}$$

provides a local extremum for the functional

$$F(x) = \int_a^b L(x, y, y') \, \mathrm{d}s,$$

then x must satisfy the equations

$$\begin{aligned} L_y(x, y, y') &= \frac{\mathrm{d}}{\mathrm{d}x} L_{y'}(x, y, y'), \qquad (7.16) \\ L_{y'}(b, y(b), y'(b)) &= 0. \qquad (7.17) \end{aligned}$$

While (7.16) is the usual Euler–Lagrange equation, (7.17) is called the **natural boundary condition**. By similar arguments, if the left endpoint $y(a)$ is unspecified, then the natural boundary condition is given by $L_{y'}(a, y(a), y'(a)) = 0$.

Example 7.5.5

Getting back to the initial Example 7.5.3, we have

$$L_{y'}(x,y,y') = \frac{1}{c^2 - v(x)^2}\left(\frac{c^2 y'}{c^2[1+(y')^2] - v(x)^2} - v(x)\right).$$

Hence, the natural boundary condition (7.17) becomes

$$\frac{c^2 y'(b)}{c^2[1+y'(b)^2] - v(b)^2} - v(b) = 0,$$

which can be simplefied to

$$y'(b) = \frac{v(b)}{c}.$$

Thus, the slope with which the boat enters the bank at $x = b$ is the ratio of the water speed at the bank to the boat's velocity in still water.

Chapter 8

Orthogonal Expansions

In many situations, it is desired to approximate a given function f by a simpler function. A familiar example for this, if f has sufficiently many derivatives at $x = x_0$, is the N-th Taylor polynomial of f around $x = x_0$,

$$T_N[f](x) = \sum_{n=0}^{N} c_n (x - x_0)^n$$

with Taylor coefficients

$$c_n = \frac{f^{(n)}(x_0)}{n!}.$$

Another type of approximations are (truncated) Fourier series. If f is integrable on the interval $[-\pi, \pi]$, then its **truncated Fourier series** is given by

$$F_N[f](x) = \frac{a_0}{2} + \sum_{n=1}^{N} a_n \cos(nx) + b_n \sin(nx)$$

with **Fourier coefficients**

$$\begin{aligned} a_n &= \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \cos(nx) \, \mathrm{d}x, \quad n = 0, 1, 2, \dots, \\ b_n &= \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \sin(nx) \, \mathrm{d}x, \quad n = 1, 2, \dots. \end{aligned}$$

Both approximations have their own advantages. While the Taylor polynomial can provide a good approximation to f locally near $x = x_0$, the Fourier series often is able to provide a good global approximation, at least if f is periodic. In this chapter, by exposing the underlying theory behind Fourier series, we will be able to also extend this idea to nonperiodic and many other functions.

8.1 Best Approximations in Inner Product Spaces

The potentially excellent approximation properties of the Fourier series might be best understood by noting that the (truncated) Fourier series can be characterized as the best approximation in certain inner product spaces. Getting to the heart of the underlying theory — presented in this and the next section — will also allow us to formulate generalized Fourier series, therefore extending the concept of classical Fourier series to a much broader class of function spaces.

Definition 8.1.1: Best approximations

Let $(X, \|\cdot\|)$ be a normed linear space, $f \in X$, and $V \subset X$ be a linear subspace. An element $v^* \in V$ is called **best approximation** of f from V with respect to $\|\cdot\|$ if

$$\|f - v^*\| \leq \|f - v\|$$

holds for all $v \in V$.

This means that there is no element $v \in V$ which is closer to f that v^*.

Example 8.1.2: $(\mathbb{R}^2, \|\cdot\|_\infty)$

Given is the linear space $(\mathbb{R}^2, \|\cdot\|_\infty)$ with $\|(x,y)^\intercal\|_\infty = \max\{|x|, |y|\}$, the element $f = (0,1)^\intercal$, and the linear subspace $V = \mathbb{R} \times \{0\}$ which corresponds to the x-axis of the xy-plane. Then, every element

$$v_r^* = (r, 0)^\intercal \in V, \quad r \in [-1, 1],$$

is a best approximation of $f = (0,1)^\intercal$ from V with respect to $\|\cdot\|_\infty$. This can be noted by observing that

$$\|f - v_r^*\|_\infty = \|(-r, 1)^\intercal\|_\infty = \max\{|r|, 1\} = 1$$

for $r \in [-1, 1]$ and

$$\|f - v\|_\infty > 1$$

for all other elements from V.

While providing an illustration of best approximations in a well-known setting, Example 8.1.2 also shows that, in general, best approximations are not unique. These changes, however, if we restrict ourselves to inner product spaces. Moreover, it is convenient to assume V to be a *finite dimensional* linear subspace.[1]

Theorem 8.1.3: Existence and uniqueness of best approximations

Let $(X, \langle\cdot,\cdot\rangle)$ be an inner product space and $V \subset X$ be a finite dimensional linear subspace. Then, for every $f \in X$, there exists a unique best approximation $v^* \in V$ of f with respect to the induced norm $\|x\| = \sqrt{\langle x, x\rangle}$.

Proof. The details can be found, for instance, in [Gla20]. Yet, we note that the uniqueness follows from $\|\cdot\|$ being strictly convex, which, in turn, follows from the parallelogram law. □

Hence, we can establish uniqueness of the best approximation if we restrict ourselves to norms which are induced by an inner product. This is also demonstrated in the following example.

Example 8.1.4: $(\mathbb{R}^2, \|\cdot\|_2)$

Again, let us consider the linear space $\mathbb{R}^2$ with linear subspace $V = \mathbb{R} \times \{0\}$ and $f = (0,1)^\intercal$. This

[1] Generally speaking, uniqueness is already ensured once we restrict ourselves to to finite dimensional linear subspaces V and strictly convex norms. Hence, also the usual p-norms with $1 < p < \infty$ would yield uniqueness for finite dimensional V. See, for instance, [Gla20, Chapter 3.1.1] and references therein.

time, however, we equip the linear space with the Euclidean norm

$$\|(x,y)^\intercal\|_2 = \sqrt{x^2+y^2},$$

which is induced by the usual inner product

$$\langle (x,y)^\intercal, (\tilde{x},\tilde{y})^\intercal \rangle = x\tilde{x} + y\tilde{y}.$$

Then, in accordance with Theorem 8.1.3, there is only a single best approximation:

$$v^* = (0,0)^\intercal$$

This can be noted by observing that

$$\|f - v^*\|_2^2 = \|(0,1)^\intercal\|_2^2 = 1$$

and

$$\|f - v\|_2^2 = \|(x,1)\|_2^2 = x^2 + 1 > 1$$

for all $v = (x,0)^\intercal \in V$ with $x \neq 0$.

We have observed that the inner product spaces provide us with a unique best approximation from a finite dimension linear subspace to any element. Yet, the actual advantage of inner product spaces is the concept of orthogonality. Looking back at our earlier crash course in functional analysis — more precisely Definition 6.1.10 — we call two vectors **orthogonal** if

$$\langle x, y \rangle = 0.$$

Utilizing the concept of orthogonality, we can note the following characterization of the best approximation, which can be used for explicit and practical computations.

Theorem 8.1.5: Characterization of the best approximation

Let $(X, \langle \cdot, \cdot \rangle)$ be an inner product space, $f \in X$, and $V \subset X$ be a finite dimensional linear subspace. Then, $v^* \in V$ is the best approximation of f from V if and only if

$$\langle f - v^*, v \rangle = 0 \quad \forall v \in V \tag{8.1}$$

holds.

Proof. Let $v^* \in V$. First, we show that (8.1) is a necessary condition for v^* being a best approximation. Afterwards, we prove that (8.1) also is a sufficient condition.

$\Longrightarrow$: Let us assume that there exists a $v \in V$ (with $v \neq 0$) such that (8.1) is violated:

$$\alpha := \langle f - v^*, v \rangle \neq 0$$

Then,

$$\|f - (v^* + \lambda v)\|^2 = \|f - v^*\|^2 - 2\lambda \langle f - v^*, v \rangle + \lambda^2 \|v\|^2$$

and by choosing $\lambda = \frac{\alpha}{\|v\|^2}$, we get

$$\|f - (v^* + \lambda v)\|^2 = \|f - v^*\| - \frac{\alpha^2}{\|v\|^2} < \|f - v^*\|^2.$$

Hence, v^* cannot be the best approximation.

$\Longleftarrow$: Let (8.1) hold for v^* and let $v \in V$. Then, we have

$$\begin{aligned} \|f - v^*\|^2 &= \langle f - v^*, f - v + v - v^* \rangle \\ &= \langle f - v^*, f - v \rangle + \langle f - v^*, v - v^* \rangle. \end{aligned}$$

Since $v - v^* \in V$, condition (8.1) yields $\langle f - v, v - v^* \rangle = 0$ and we get

$$\|f - v^*\|^2 = \langle f - v^*, f - v \rangle \leq \|f - v^*\| \|f - v\|$$

by the Cauchy–Schwarz inequality (Lemma 6.1.8). Thus, $\|f - v^*\| \leq \|f - v\|$ for all $v \in V$, which means that v^* is the best approximation of f from V.

□

Note that the characterization (8.1) essentially means that the error between f and its best approximation v^* is orthogonal to the linear subspace V; that is, $f - v^* \perp V$. This shows that the best approximation v^* is obtained by the orthogonal projection of f onto V.

8.2 The Generalized Fourier Series

In the previous section, we have introduced best approximations and were able to characterize these in inner product spaces by (8.1). In this section, we build up on this observation and show that the (truncated) Fourier series, in fact, is such a best approximation. This also explains many of the favorable approximation properties of the (truncated) Fourier series.

Remark 8.2.1

The characterization of the best approximation provided by Theorem 8.1.5 allows — at least formally — an easy computation of the best approximation. The procedure is described below.

Let $\{v_k\}_{k=1}^N$ be a basis of the finite dimensional linear subspace $V \subset X$. Then, the best approximation v^* of $f \in X$ from V has a unique representation as a linear combination of the basis elements:

$$v^* = \sum_{k=1}^{N} \alpha_k v_k \tag{8.2}$$

Here, the scalar coefficients α_k are elements of the underlying field, in this case $\mathbb{R}$. These coefficients can be determined by consulting (8.1). Note that, because of the sesquilinearity of the inner product, it is sufficient to check (8.1) for the basis elements $v_1, \dots, v_N$. Thus, using the representation (8.2), we get a system of linear equations,

$$\sum_{k=1}^{N} \alpha_k \langle v_k, v_l \rangle = \langle f, v_l \rangle, \quad l = 1, \dots, N.$$

By denoting $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_N)^\intercal$, $\mathbf{b} = (\langle f, v_1 \rangle, \dots, \langle f, v_N \rangle)^\intercal$, and

$$G = \begin{pmatrix} \langle v_1, v_1 \rangle & \cdots & \langle v_N, v_1 \rangle \\ \vdots & & \vdots \\ \langle v_1, v_N \rangle & \cdots & \langle v_N, v_N \rangle \end{pmatrix},$$

the above system of linear equations can be written in matrix vector notation as

$$G\boldsymbol{\alpha} = \mathbf{b}. \tag{8.3}$$

The matrix G is called a *Gram matrix*; see [HJ12, Chapter 7.2]. It is symmetric ($G = G^\intercal$) and positive definite, since

$$\begin{aligned} \beta^\intercal G \beta &= \sum_{k,l=1}^{N} \beta_k \beta_l \langle v_k, v_l \rangle \\ &= \left\langle \sum_{k=1}^{N} \beta_k v_k, \sum_{l=1}^{N} v_l \right\rangle \\ &= \left\| \sum_{k=1}^{N} \beta_k v_k \right\|^2 \\ &> 0 \end{aligned}$$

for $\beta \in \mathbb{R}^N$ with $\beta \neq 0$. Note that the last estimate follows from the basis elements $v_1, \dots, v_N$ being linearly independent. In particular, the Gram matrix G is therefore regular and (8.3) can be solved uniquely for the coefficients $\boldsymbol{\alpha}$ of the best approximation v^*. This can be done, for instance, by Gaussian elimination [TBI97, Lecture 20]. Yet, the real beauty of best approximations in inner product spaces is revealed when we choose $\{v_k\}_{k=1}^N$ to be an orthogonal basis.

Definition 8.2.2: Orthogonal bases

Let $\{v_k\}_{k=1}^N$ be a basis of the linear space V. We say that $\{v_k\}_{k=1}^N$ is an **orthogonal basis** if

$$\langle v_k, v_l \rangle = \delta_{kl} \|v_k\|^2$$

holds for all $k, l = 1, \dots, N$. Here, δ_{kl} denotes the usual *Kronecker delta* defined by

$$\delta_{kl} = \begin{cases} 1, & \text{if } k = l, \\ 0, & \text{if } k \neq l. \end{cases}$$

If the basis elements are, in addition to being orthogonal, also normalized, i. e.

$$\|v_k\| = 1$$

for all $k = 1, \dots, N$, we call $\{v_k\}_{k=1}^N$ an **orthonormal basis**.

When using such an orthogonal basis, the Gram matrix G reduces to a diagonal matrix and (8.3) consists of N independent equations

$$\alpha_k \|v_k\|^2 = \langle f, v_k \rangle, \quad k = 1, \dots, N.$$

Hence, the coefficients are simply given by

$$\alpha_k = \frac{\langle f, v_k \rangle}{\|v_k\|^2}, \quad k = 1, \dots, N,$$

in this case. Note that, when using an orthonormal basis, the Gram matrix even reduces to the identity matrix and we get

$$\alpha_k = \langle f, v_k \rangle, \quad k = 1, \dots, N,$$

for the coefficients. We summarize our above observations in the following theorem.

Theorem 8.2.3: Best approximations for orthogonal bases

Let $(X, \langle \cdot, \cdot \rangle)$ be an inner product space, $V \subset X$ be a finite dimensional subspace, and $\{v_k\}_{k=1}^N$ be an orthogonal basis of V. Then, the best approximation v^* of f from V with respect to $\| \cdot \| = \sqrt{\langle \cdot, \cdot \rangle}$ is uniquely given by

$$v^* = \sum_{k=1}^N \frac{\langle f, v_k \rangle}{\|v_k\|^2} v_k.$$

The above form of a best approximation already is what is usually referred to as truncated generalized Fourier series, at least in the case of $X = L^2([a,b])$. Here, $L^2([a,b])$ denotes the linear space of all square integrable real-valued functions.

Definition 8.2.4: Square integrable functions

A function $f : [a,b] \to \mathbb{R}$ is said to be square **square integrable** on $[a,b]$, for which we write $f \in L^2([a,b])$, if

$$\int_a^b |f(x)|^2 \, \mathrm{d}x < \infty$$

holds.

Equipped with the inner product

$$\langle f, g \rangle = \int_a^b f(x) g(x) \, \mathrm{d}x, \tag{8.4}$$

the function space $L^2([a,b])$ is a (complete) inner product space.[2] To the induced norm,

$$\|f\|^2 = \int_a^b |f(x)|^2 \, \mathrm{d}x, \tag{8.5}$$

we usually refer to either as the L^2 **norm** or the **mean-square norm**. In this setting, we define the generalized Fourier series as follows.

Definition 8.2.5: Generalized Fourier series

Let $\mathcal{F} = \{f_k\}_{k=0}^\infty$ with $f_k \not\equiv 0$ be a set of pairwise orthogonal functions in $(L^2([a,b]), \langle \cdot, \cdot \rangle)$. The **(generalized) Fourier series** of $f \in L^2([a,b])$ with respect to $\mathcal{F}$ is given by

$$F[f](x) = \sum_{k=0}^\infty c_k f_k(x) \tag{8.6}$$

with **(generalized) Fourier coefficients**

$$c_k = \frac{\langle f, f_n \rangle}{\|f_k\|^2}. \tag{8.7}$$

Sometimes we write $F[f] \sim f$ to denote that $F[f]$ is the (generalized) Fourier series of f. Moreover,

$$F_N[f](x) = \sum_{k=0}^N c_k f_k(x) \tag{8.8}$$

is referred to as the N**-th truncated (generalized) Fourier series**.

[2] Strictly speaking, we have to partition the square integrable functions into equivalence classes first. Otherwise, it is not even positive definite. We say that two square integrable functions are equivalent if they are equal almost everywhere.

Example 8.2.6

Given is the set $\mathcal{F} = \{f_n\}_{n=0}^{\infty}$ with $f_n(x) = \cos(n\pi x)$ of pairwise orthogonal functions in $(L^2([0,1]), \langle\cdot,\cdot\rangle)$. Let $f(x) = 1 - x$, then the Fourier coefficients c_n of the generalized Fourier series

$$F_N[f](x) = \sum_{n=0}^{N} c_n \cos(n\pi x)$$

can be calculated as follows: First, we note that

$$\begin{aligned} \|f_n\|^2 &= \int_0^1 \cos(n\pi x)^2 \,\mathrm{d}x \overset{\text{IBP}}{=} \int_0^1 \sin(n\pi x)^2 \,\mathrm{d}x \\ &= \int_0^1 1 - \cos(n\pi x)^2 \,\mathrm{d}x = 1 - \|f_n\|^2 \end{aligned}$$

and therefore

$$\|f_n\|^2 = \frac{1}{2}.$$

Next, we observe that

$$\begin{aligned} \langle f, f_n\rangle &= \int_0^1 (1-x)\cos(n\pi x)\,\mathrm{d}x \\ &= -\int_0^1 x\cos(n\pi x)\,\mathrm{d}x \\ &\overset{\text{IBP}}{=} -\frac{1}{n\pi} x\sin(n\pi x)\Big|_0^1 + \frac{1}{n\pi}\int_0^1 \sin(n\pi x)\,\mathrm{d}x \\ &= -\frac{1}{(n\pi)^2}\left[\cos(n\pi x)\right]_0^1 \\ &= -\frac{1}{(n\pi)^2}\left(\cos(n\pi) - 1\right). \end{aligned}$$

Thus, since

$$\cos(n\pi) = \begin{cases} +1 & \text{if } n \text{ is even,} \\ -1 & \text{if } n \text{ is odd,} \end{cases}$$

holds, we have

$$\langle f, f_n\rangle = \begin{cases} 0 & \text{if } n \text{ is even,} \\ \frac{2}{(n\pi)^2} & \text{if } n \text{ is odd.} \end{cases}$$

We therefore get

$$c_n = \frac{\langle f, f_n\rangle}{\|f_n\|^2} = \begin{cases} 0 & \text{if } n \text{ is even,} \\ \frac{4}{(n\pi)^2} & \text{if } n \text{ is odd,} \end{cases}$$

for the Fourier coefficients of f.

In the previous discussion, we have observed that the truncated Fourier series, $F_N[f]$, is exactly the best approximation in $(L^2([a,b]), \langle\cdot,\cdot\rangle)$. This is summarized in the following theorem.

Theorem 8.2.7

Let $\mathcal{F} = \{f_k\}_{k=0}^{\infty}$ with $f_k \not\equiv 0$ be a set of pairwise orthogonal functions in $(L^2([a,b]), \langle \cdot, \cdot \rangle)$ and let $V = \text{span } \mathcal{F}$ (the linear subspace spanned by $\mathcal{F}$). For every $f \in L^2([a,b])$, the truncated Fourier series $F_N[f]$ is the best approximation of f from V; that is,

$$\|f - F_N[f]\| \leq \|f - v\|$$

for all $v \in V$.

Hence, the truncated Fourier series is constructed in such a way that there can be no function generated from the same orthogonal system $\mathcal{F}$ which is able to yield a better approximation to f that $F_N[f]$.

8.3 Mean-Square Convergence

In the last section, we noted that the truncated Fourier series with respect to $\mathcal{F}$ is the best approximation from span($\mathcal{F}$) with respect to the mean-square norm. In this section we show that — under certain assumptions on the orthogonal system $\mathcal{F}$ — this yields the Fourier series $F_N[f]$ to actually converge to f. Henceforth, $\|\cdot\|$ always denotes the L^2/mean-square norm (8.5). We start our investigation by noting the following identity for the mean-square error of a truncated Fourier series.

Lemma 8.3.1

Let $f \in L^2([a,b])$ and $\mathcal{F} = \{f_k\}_{k=0}^{N}$ be a set of pairwise orthonormal functions with $f_k \not\equiv 0$. Then,

$$\|f - F_N[f]\|^2 = \|f\|^2 - \sum_{k=0}^{N} c_k^2. \tag{8.9}$$

Proof. Note that

$$\begin{aligned} \|f - F_N[f]\|^2 &= \langle f - F_N[f], f - F_N[f] \rangle \\ &= \|f\|^2 - \langle f, F_N[f] \rangle + \|F_N[f]\|^2. \end{aligned}$$

Next, we observe that

$$\langle f, F_N[f] \rangle = \sum_{k=0}^{N} c_k \langle f, f_k \rangle = \sum_{k=0}^{N} c_k^2 \|f_k\|^2 = \sum_{k=0}^{N} c_k^2,$$

by Pythagora's theorem (Lemma 6.1.11). Also note that the f_k are orthonormal and therefore satisfy $\|f_k\|^2 = 1$. Finally, all of this yields

$$\begin{aligned} \|f - F_N[f]\|^2 &= \|f\|^2 - 2\sum_{k=0}^{N} c_k^2 + \sum_{k=0}^{N} c_k^2 \\ &= \|f\|^2 - \sum_{k=0}^{N} c_k^2. \end{aligned}$$

□

From Lemma 8.3.1 we can immediately note **Bessel's inequality**.

Theorem 8.3.2: Bessel's inequality

Let $f \in L^2([a,b])$ and $\mathcal{F} = \{f_k\}_{k=0}^{\infty}$ be a set of pairwise orthonormal functions with $f_k \not\equiv 0$. Then,

$$\sum_{k=0}^{\infty} c_k^2 \leq \|f\|^2.$$

holds, where the c_k's are the Fourier coefficients given by (8.7).

Proof. Let $N \in \mathbb{N}$. Then, the LHS is nonnegative and we therefore have

$$\sum_{k=0}^{N} c_k^2 \leq \|f\|^2$$

Since this inequality holds for every $N \in \mathbb{N}$, we can conclude the assertion. □

Furthermore, from Lemma 8.3.1 it immediately becomes clear that the Fourier series $F[f]$ converges to $f \in L^2([a,b])$, i. e.

$$F[f] = f,$$

if and only if

$$\sum_{k=0}^{N} c_k^2 = \|f\|^2 \tag{8.10}$$

holds for every $f \in L^2([a,b])$. Equation (8.10) is known as **Parseval's equality**. This identity, in turn, holds if and only if we restrict ourselves to complete orthonormal systems $\mathcal{F}$.

Definition 8.3.3: Complete orthonormal systems

An orthonormal system $\mathcal{F} \subset L^2([a,b])$ is said to be **complete** if

$$\langle f, f_k \rangle = 0 \ \forall f_k \in \mathcal{F} \implies f \equiv 0$$

holds for all $f \in L^2([a,b])$.

Thus, an orthonormal system $\mathcal{F} = \{f_k\}_{k=0}^{\infty}$ is complete if and only if the only function having all its Fourier coefficients vanish is the zero function. Sometimes it is difficult to show completeness and we shall usually just state whether a given orthonormal system is complete. Next, it is proven that completeness is equivalent to strict equality holding in Bessel's inequality, resulting in Parseval's equality.

Theorem 8.3.4: Parseval's equality

Let $\mathcal{F} = \{f_k\}_{k=0}^{\infty}$ be an othonormal system in $L^2([a,b])$. $\mathcal{F}$ is complete if and only if Parseval's equality,

$$\sum_{k=0}^{\infty} c_k^2 = \|f\|^2,$$

holds for all $L^2([a,b])$. Here, the c_k's are the Fourier coefficients of f.

Proof. First, we show that $\mathcal{F}$ being complete is sufficient for Parseval's equality to hold and afterwards that it is also a necessary condition.

"$\Longrightarrow$": Let $\mathcal{F}$ be a complete orthonormal system and let us denote

$$\tilde{f} = \sum_{k=0}^{\infty} \langle f, f_k \rangle f_k.$$

Then, we have

$$\begin{aligned}
\langle f - \tilde{f}, f_j \rangle &= \langle f, f_j \rangle - \langle \tilde{f}, f_j \rangle \\
&= \langle f, f_j \rangle - \sum_{k=0}^{\infty} \langle f, f_k \rangle \langle f_k, f_k \rangle \\
&= \langle f, f_j \rangle - \langle f, f_j \rangle \\
&= 0
\end{aligned}$$

for all $f_j \in \mathcal{F}$. Hence, since $\mathcal{F}$ is complete, this implies $f - \tilde{f} \equiv 0$ and therefore

$$\|f\|^2 - \|\tilde{f}\|^2 \leq \|f - \tilde{f}\|^2 = 0.$$

Finally, this yields

$$\sum_{k=0}^{N} c_k^2 = \|\tilde{f}\|^2 \geq \|f\|^2$$

and, in combination with Bessel's inequality, we get

$$\sum_{k=0}^{N} c_k^2 = \|f\|^2.$$

"$\Longleftarrow$": Let us assume that $\mathcal{F}$ is not complete. That is, there exists an $f \in L^2([a,b])$ with $f \not\equiv 0$ such that

$$\langle f, f_k \rangle = 0 \quad \forall f_k \in \mathcal{F}.$$

Then, by Parseval's equality, we have

$$\|f\|^2 = \sum_{k=0}^{\infty} |\langle f, f_k \rangle|^2 = 0.$$

This, however, implies that $f \equiv 0$ which is a contradiction to our initial assumption. Hence, Parseval's equality implies completeness.

□

Finally, using Theorem 8.3.4, we can note that the orthonormal system $\mathcal{F}$ being complete is also equivalent to the Fourier series converging to $f \in L^2([a,b])$ in the mean-square sense,

$$\lim_{N \to \infty} F_N[f] = f \quad \text{or} \quad F[f] = f \text{ in } \left(L^2([a,b]), \| \cdot \| \right).$$

Thus is summarized in the following theorem.

Theorem 8.3.5: Mean-square convergence of the Fourier series

Let $\mathcal{F} = \{f_k\}_{k=0}^{\infty}$ be an orthonormal system in $L^2([a,b])$ and let $f \in L^2([a,b])$. The Fourier series of f with respect to $\mathcal{F}$ converges to f in the mean-square sense if and only if $\mathcal{F}$ is complete.

Proof. The assertion follows from Lemma 8.3.1 and Theorem 8.3.4. □

8.4 Classical Fourier Series

So far, we have discussed generalized Fourier series and their mean-square approximation and convergence properties. We end this chapter by specifically addressing classical Fourier series in more detail. Classical Fourier series were introduced in the early 1800's by Joseph Fourier to solve partial differential equations in heat transfer and arise as a special case of the generalized Fourier series discussed so far. Let us consider the orthogonal system

$$\mathcal{F} = \left\{ 1, \, \cos\left(\frac{n\pi x}{L}\right), \, \sin\left(\frac{n\pi x}{L}\right) \mid n \in \mathbb{N} \right\} \tag{8.11}$$

in $L^2([-L,L])$. Then, the **classical Fourier series** of $f \in L^2([-L,L])$ is given by

$$F[f](x) = \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos\left(\frac{n\pi x}{L}\right) + b_n \sin\left(\frac{n\pi x}{L}\right), \tag{8.12}$$

where the coefficients are

$$\begin{aligned} a_n &= \frac{1}{L} \int_{-L}^{L} f(x) \cos\left(\frac{n\pi x}{L}\right) \, \mathrm{d}x, \quad n = 0, 1, 2, \dots, \\ b_n &= \frac{1}{L} \int_{-L}^{L} f(x) \sin\left(\frac{n\pi x}{L}\right) \, \mathrm{d}x, \quad n = 1, 2, \dots. \end{aligned}$$

It can be shown — which we will not do here, however — that the trigonometric functions (8.11) form a complete orthogonal system. Hence, for $f \in L^2([-L,L])$, the classical Fourier series converges to f in the mean-square sense.

Example 8.4.1

Given is the square-integrable function $f(x) = x$ on $[-1,1]$. The Fourier coefficients of f's Fourier series are easily computed to be

$$\begin{aligned} a_n &= \int_{-1}^{1} f(x) \cos(n\pi x) \, \mathrm{d}x = 0, \quad n = 0, 1, 2, \dots, \\ b_n &= \int_{-1}^{1} f(x) \sin(n\pi x) \, \mathrm{d}x = (-1)^{n+1} \frac{2}{n\pi}, \quad n = 1, 2, \dots. \end{aligned}$$

Thus, the Fourier series of f is given by

$$F[f] = \frac{2}{\pi} \sum_{n=1}^{\infty} (-1)^{n+1} \frac{1}{n} \sin(n\pi x)$$

and Theorem 8.3.5 tells us that $F[f]$ converges to f in the mean-square sense. That is,

$$F[f] = f \text{ in } \left(L^2([-1,1]), \|\cdot\|\right) \quad \text{or} \quad \|F[f] - f\| = 0.$$

It should be stressed heavily that this equality only holds in the mean-square sense and, for instance, not necessarily in the pointwise sense ($F[f](x) = f(x) \ \forall x \in [-1,1]$)! In fact, we can note that

$$F[f](\pm 1) = 0 \neq \pm 1 = f(\pm 1).$$

Hence, the Fourier series does not converge at the boundary points. It does, however, converge pointwise to f in $(-1,1)$.

The above example shows that, while we can expect the Fourier series to converge in the mean-square sense for every $f \in L^2([a,b])$, pointwise convergence does not hold in general. Yet, we can establish such a result if we restrict ourselves to piecewise smooth periodic functions.

Definition 8.4.2

Let $f : [a,b] \to \mathbb{R}$ be a function. We say that f is

(a) **piecewise continuous** if it is continuous except possibly at a finite number of points in $[a,b]$ where it has simple jump discontinuities. That is, there exists

$$a =: x_0 < x_1 < \cdots < x_{N+1} := b$$

such that

$$f|_{(x_n,x_{n+1})} : (x_n, x_{n+1}) \to \mathbb{R} \text{ is continuous} \quad \forall n = 0, \dots, N$$

and

$$f(x_n^+) := \lim_{x \searrow x_n} f(x), \quad f(x_n^-) := \lim_{x \nearrow x_n} f(x) \in \mathbb{R} \quad \forall n = 1, \dots, N.$$

The second condition means that both one-sided limits of f exist and are finite.

(b) **piecewise smooth** if f and f' are piecewise continuous.

(c) **periodic** if $f(a) = f(b)$.

For this class of functions, indeed, we can note the following convergence result.

Theorem 8.4.3: Pointwise convergence of the classical Fourier series

Let $f : [-L, L]) \to \mathbb{R}$ be periodic and piecewise smooth. Then, its classical Fourier series (8.12) converges pointwise for all $x \in [-L, L]$ to the value $f(x)$ if f is continuous at x and to the average value of its left and right limits at x, namely $\frac{1}{2}[f(x^+) + f(x^-)]$, if f is discontinuous at x.

To get stronger convergence results, such as uniform convergence, additional smoothness conditions on f are required. Observe that continuity of f is not enough to guarantee pointwise convergence.

Chapter 9

Sturm–Liouville Problem

In Chapter 8, we have discussed generalized Fourier series to represent or at least approximate square integrable functions. These assume a (complete) orthogonal system, however. If we already have a basis of $L^2([a,b])$ at hand we can construct a complete orthogonal system by the Gram–Schmidt procedure. Here, we discuss another technique to construct complete orthogonal systems without prior knowledge of a basis. This technique builds up on ordinary differential equations (ODEs) of a Sturm–Liouville type.

Definition 9.0.1: Sturm–Liouville type ODEs

We say that an ODE is of a **Sturm–Liouville type** if it has the form

$$-\left[p(x)y'\right]' + q(x)y = \lambda y, \quad a < x < b, \tag{9.1}$$

where λ is a constant and p, q are functions.

For the bounded interval $[a,b]$, equation (9.1) is usually accompanied by boundary conditions of the form

$$\alpha_1 y(a) + \alpha_2 y'(a) = 0, \quad \beta_1 y(b) + \beta_2 y'(b) = 0. \tag{9.2}$$

Here, the pairs of constants α_1, α_2 and β_1, β_2 respectively are not allowed to both be zero. Otherwise, the boundary condition would collapse at that boundary ("0=0"). Two important special cases of (9.2) are **(homogenous) Dirichlet boundary conditions**

$$y(a) = 0, \quad y(b) = 0$$

and **(homogenous) Neumann boundary conditions**

$$y'(a) = 0, \quad y'(b) = 0.$$

It is convenient to also distinguish between certain cases regarding the functions p and q in (9.1).

Definition 9.0.2: Regular and singular SLPs

We call the ODE (9.1) together with a boundary condition (9.2) a **Sturm–Liouville problem (SLP)**. Furthermore, if the interval $[a,b]$ is bounded, $p \in C^1([a,b])$, $q \in C^0([a,b])$ and p is never zero in $[a,b]$, we say that the SLP is **regular**. Otherwise, it is reffered to as **singular**.

It should be stressed that a regular SLP might not have a nontrivial solution ($y \not\equiv 0$) for every value of the constant λ.

Remark 9.0.3

The SLP (9.1), (9.2) can be interpreted as an eigenvalue problem for certain linear operators. Let us consider the differential operator

$$Ly = -[p(x)y']' + q(x)y,$$

then the SLP is equivalent to the eigenvalue problem

$$Ly = \lambda y,$$

where we are only interested in eigenvectors $y \not\equiv 0$ (also called **eigenfunctions**) that additionally satisfy the boundary conditions (9.2). From this interpretation we can note that any constant of a solution (eigenfunction) y gives another — but not independent — solution (eigenfunction) cy.

The really interesting fact about regular SBPs is that they have an infinite number of eigenvalues λ and the corresponding eigenfunctions form a complete orthogonal system of $L^2([a,b])$. This, again, allows us to work with orthogonal expansions, a key idea in many applications.

Example 9.0.4

Given is a regular SLP

$$\begin{aligned} -y'' &= \lambda y, \quad 0 < x < \pi, \\ y(0) &= y(\pi) = 0. \end{aligned} \tag{9.3}$$

The SLP has eigenvalues and corresponding eigenfunctions (solutions)

$$\lambda_k = k^2, \quad y_k(x) = \sin(kx), \quad k \in \mathbb{N}.$$

One way to find these is to separately consider the cases $\lambda = 0$, $\lambda < 0$, and $\lambda > 0$. (We show later that the eigenvalue λ cannot be complex.)

$\lambda = 0$: Then, the ODE (9.3) becomes

$$y'' = 0$$

with general solution

$$y(x) = ax + b.$$

Yet, the BCs yield

$$y \equiv 0,$$

which is not an admissible eigenfunction. Note that eigenvectors (eigenfunctions) are not allowed to be zero. Otherwise, every constant would be an eigenvalue of this eigenvector. Hence, $\lambda = 0$ cannot be an eigenvalue.

$\lambda < 0$: Let us express λ as $\lambda = -k^2$ for a suitable $k \in \mathbb{R}^+$. Then, the ODE (9.3) becomes

$$y'' - k^2 y = 0$$

with general solution

$$y(x) = ae^{kx} + be^{-kx}.$$

The BCs yield $a = b = 0$ and therefore

$$y \equiv 0.$$

Thus, there can be no negative eigenvalue.

$\lambda > 0$: For $\lambda = k^2$ with $k \in \mathbb{R}^+$ the ODE becomes

$$y'' + k^2 y = 0$$

with general solution

$$y(x) = a\cos(kx) + b\sin(kx).$$

This time, the BCs require

$$\begin{aligned} y(0) &= a \overset{!}{=} 0, \\ y(\pi) &= b\sin(k\pi) \overset{!}{=} 0. \end{aligned} \tag{9.4}$$

Hence, $a = 0$ and b has to satisfy (9.4). Note that if $k \notin \mathbb{N}$ we have $b = 0$ and therefore $y \equiv 0$. Once more, $\lambda = k^2$ could not be an eigenvalue. Yet, for $k \in \mathbb{N}$, yielding

$$\lambda_k = k^2, \quad k \in \mathbb{N}, \tag{9.5}$$

equation (9.4) does not pose any restriction on b and we obtain a family of solutions

$$y_k(x) = b\sin(kx), \quad k \in \mathbb{N}.$$

Thus, the SLP (9.3) has eigenvalues and corresponding eigenfunctions

$$\lambda_k = k^2, \quad y_k(x) = \sin(kx), \quad k \in \mathbb{N},$$

where we have arbitrarily chosen $b = 1$.

In Example 9.0.4, by investigating different cases, we saw that only positive real eigenvalues occured and that the corresponding eigenfunctions form a complete orthogonal system of $L^2([a, b])$. The following theorem shows that, in fact, this holds for regular SLPs in general.

Theorem 9.0.5

The regular SLP (9.1), (9.2) has infinitely many eigenvalues λ_k, $k \in \mathbb{N}$. The eigenvalues are real and $\lim_{k\to\infty} |\lambda_k| = \infty$. Furthermore, the corresponding eigenfunctions y_k form a complete orthogonal system in $L^2([a, b])$. That is, every $f \in L^2([a, b])$ can be expanded as

$$f(x) = \sum_{k=1}^{\infty} c_k f_k(x)$$

in the mean-square sense.

Proof. We only sketch the proof of Theorem 9.0.5. For instance, regarding the existence of the eigenvalues, however, we refer to the literature [BR78]. Here we just show that the eigenfunctions corresponding to distinct eigenvalues are orthogonal. Let $\lambda_1 \neq \lambda_2$ be two eigenvalues with corresponding eigenfunctions y_1 and y_2. In particular, these satisfy the ODEs

$$\begin{aligned} -[py_1']' + qy_1 &= \lambda_1 y_1, \\ -[py_2']' + qy_2 &= \lambda_2 y_2, \end{aligned}$$

then. Multiplying the first equation by y_2, multiplying the second equation by y_1, substracting, and integrating the result over the interval $[a,b]$ yields

$$(\lambda_1-\lambda_2)\langle y_1,y_2\rangle = (\lambda_1-\lambda_2)\int_a^b y_1y_2\,\mathrm{d}x = \int_a^b -[py_1']'y_2 + [py_2']'y_1\,\mathrm{d}x.$$

Next, we observe that

$$-[py_1']'y_2 + [py_2']'y_1 = \frac{\mathrm{d}}{\mathrm{d}x}[p(y_1y_2' - y_2y_1')]$$

and therefore

$$(\lambda_1-\lambda_2)\langle y_1,y_2\rangle = p(y_1y_2'-y_2y_1')|_a^b$$

Note that y_1 and y_2 satisfy the same boundary conditions. Thus, $y_1(x)y_2'(x) = y_2(x)y_1'(x)$ for $x = a, b$ and

$$(\lambda_1-\lambda_2)\langle y_1,y_2\rangle = 0.$$

Finally, since $\lambda_1 \neq \lambda_2$, this proves $\langle y_1,y_2\rangle = 0$ (orthogonality of y_1 and y_2). □

Another issue concerns the sign of eigenvalues. The following **energy argument** is sometimes useful in showing that the eigenvalues are all of the same sign.

Example 9.0.6: Energy argument

Given is the regular SLP (of Schrödinger-type)

$$\begin{aligned} &-y'' + q(x)y = \lambda y, \quad a<x<b,\\ &y(a) = y(b) = 0, \end{aligned}$$

where q is a positive continuous function. Multiplying the ODE by y and integrating it over the interval $[a,b]$ gives

$$-\int_a^b y''y\,\mathrm{d}x + \int_a^b qy^2\,\mathrm{d}x = \lambda\int_a^b y^2\,\mathrm{d}x.$$

The first integral can be treated by integration by parts, yielding

$$-yy'|_a^b + \int_a^b (y')^2\,\mathrm{d}x + \int_a^b qy^2\,\mathrm{d}x = \lambda\int_a^b y^2\,\mathrm{d}x.$$

Next, the homogenous Dirichlet boundary conditions result in the boundary terms to vanish. Hence, we have

$$\lambda\int_a^b y^2\,\mathrm{d}x = \int_a^b y^2 + q(y')^2\,\mathrm{d}x.$$

Note that the integrals on both sides are positive. Thus, also the eigenvalues λ have to be positive.

Bibliography

[Ari76] R. Ariew. Ockham's razor: A historical and philosophical analysis of Ockham's principle of parsimony. 1976.

[BR78] G. Birkhoff and G. Rota. *Ordinary Differential Equations.* Wiley, 1978.

[Cra04] J. S. Cramer. The early origins of the logit model. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, 35(4):613–626, 2004.

[Ein18] A. Einstein. Über Gravitationswellen. *Sitzungsberichte der Preußischen Akademie der Wissenschaften Berlin (Math. Phys.)*, 1918:154–167, 1918.

[Ein05] A. Einstein. Näherungsweise Integration der Feldgleichungen der Gravitation. *Albert Einstein: Akademie-Vorträge: Sitzungsberichte der Preußischen Akademie der Wissenschaften 1914–1932*, pages 99–108, 2005.

[Gla20] J. Glaubitz. *Shock capturing and high-order methods for hyperbolic conservation laws.* Logos Verlag Berlin, 2020.

[Gut98] M. C. Gutzwiller. Moon-Earth-Sun: The oldest three-body problem. *Reviews of Modern Physics*, 70(2):589, 1998.

[Her19] M. Hermmann. *Angewandte Analysis.* Lecture notes, TU Braunschweig, Germany, 2019.

[HJ12] R. A. Horn and C. R. Johnson. *Matrix analysis.* Cambridge university press, 2012.

[Log13] J. D. Logan. *Applied mathematics.* John Wiley & Sons, 2013.

[LS88] C.-C. Lin and L. A. Segel. *Mathematics applied to deterministic problems in the natural sciences*, volume 1. Siam, 1988.

[Mal72] T. R. Malthus. *An Essay on the Principle of Population..* 1872.

[Mur12] J. D. Murray. *Asymptotic analysis*, volume 48. Springer Science & Business Media, 2012.

[MWJ92] T. R. Malthus, D. Winch, and P. James. *Malthus:'An Essay on the Principle of Population'.* Cambridge University Press, 1992.

[oWMTT01] I. B. of Weights, Measures, B. N. Taylor, and A. Thompson. *The international system of units (SI).* US Department of Commerce, Technology Administration, National Institute of . . . , 2001.

[TBI97] L. N. Trefethen and D. Bau III. *Numerical linear algebra*, volume 50. Siam, 1997.