Basics of Mathematics in Machine Learning II

Toni Karvonen

Exactum B326 — University of Helsinki
toni.karvonen@helsinki.fi

April 19, 2024

Contents

1 Introduction

These are lecture notes for the course *Basics of Mathematics in Machine Learning II* (MAT11015) lectured in Spring 2024. Th notes approximately contain the material in Chapter 5 and Sections 6.1–6.4 and 7.1 of *Mathematics for Machine Learning* by Deisenroth, Faisal and Oong (freely available at https://mml-book.github.io/). We cover the following three topics:

I *Univariate calculus* — Section 2

II *Vector calculus* — Section 3

III *Discrete probability* — Section 4

Topics **I** and **II** cover the tools and results from mathematical analysis that are essential for machine learning. The primary learning objective of these topics is to (a) understand the basics of mathematical optimisation and (b) be able to implement some optimisation methods. Our approach to calculus is not completely rigorous, as we do not properly define some of the notions that we encounter or use and occasionally sweep certain finer points under the rug. A rigorous treatment would require *much* more time than we have. Those interested should consider some of the following courses:

- *Raja-arvot* (MAT11003)
- *Differentiaalilaskenta* (MAT11004)
- *Integraalilaskenta* (MAT11005)
- *Calculus IA: Limits and differentiation* (MAT11006)
- *Calculus IB: Integration* (MAT11007)
- *Advanced calculus* (MAT11008)
- *Vektorianalyysi I* (MAT21003)
- *Vektorianalyysi II* (MAT21020)

Topic **III** covers the basics definitions, concepts and results of discrete probability theory. The objective of this topic is to provide the essentials upon which to build on subsequent courses on statistical machine learning and data science. We have time to cover only the bare minimum about probability, and you should seriously consider taking a course dedicated to probability theory, such as

- *Todennäköisyyslaskenta I* (MAT12003)

1.1 Motivation

Let us begin by motivating the these topics with three simple examples.

1.1.1 Ordinary Least Squares

Suppose that we have observations (or outputs) $y_1,\dots,y_n \in \mathbb{R}$ at some distinct locations (or inputs) $x_1,\dots,x_n \in \mathbb{R}$. This is our *training data* (suom. *opetusjoukko*). To name a few examples, the locations could describe spatial coordinates of some physical sensors, they could be time instances [in which case with a *time series* (suom. *aikasarja*)], or they could be different parameters that an experiment is performed with. Given these data, we want to predict what the observation might be at a location that is not included in the dataset. To do this, we can postulate that observations have linear relationship to the locations:

$$y_i = a + bx_i + \varepsilon_i \quad \text{for} \quad i = 1,\dots n, \tag{1.1}$$

where a and b determine the relationship between x_i and y_i and the residual ε_i accounts for observation noise or influence on y_i of sources other than x_i. Our task is to find a "good" or

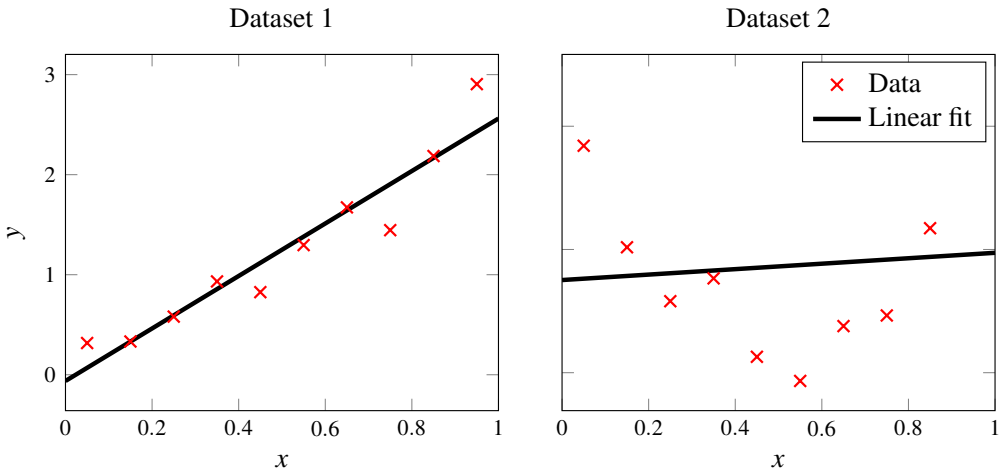

Figure 1: The linear fits $f(x) = a^* + bx^*$ for two different datasets. Here a^* and b^* are computed from (1.3). The fit is quite good for the first dataset but completely useless for the second.

"optimal" values for a and b. This is called *linear regression* (suom. *lineaarinen regressio*). One way to achieve this is to select a and b that minimise the *sum of squared residuals*

$$L(a,b) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2. \tag{1.2}$$

The function L is an important example of a *loss function* (suom. *tappiofunktio*), or *cost function*. The selection of a loss function is arbitrary (e.g., we could alternatively try to minimise $\sum_{i=1}^{n} |\varepsilon_i|$), but the *quadratic* loss function in (1.2) is mathematically extremely convenient, yielding the *ordinary least squares* method (suom. *pienimmän neliösumman menetelmä*). During the course we will learn how to minimise L and that (as long as $n \geq 2$)

$$\begin{bmatrix} a^* \\ b^* \end{bmatrix} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{y}, \quad \text{where} \quad \mathbf{X} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \in \mathbb{R}^{n\times 2} \quad \text{and} \quad \mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \in \mathbb{R}^n, \tag{1.3}$$

solve this minimisation problem *uniquely*, in that $L(a^*,b^*) < L(a,b)$ for any other pair (a,b). The *linear fits* $f(x) = a^* + bx^*$ produced by linear regression are depicted in Figure 1 for two rather different datasets. The figure shows that the linear fit can be quite useful if the data indeed exhibit a linear trend, as is the case for Dataset 1. That is, for an input x_0 not contained in the training data it seems plausible that $f(x_0)$ would be close to the corresponding output, y_0. However, the fit is useless if no such trend is present, as is the case for Dataset 2. This is because the relationship (1.1) between x_i and y_i that we have postulated is not particularly expressive, having only two parameters, a and b.

1.1.2 Neural Networks

At their core, *neural networks* (suom. *neuroverkko*) consist of nothing more than (i) a flexible postulate like (1.1) for the relationship between the inputs and outputs; (ii) selection of some parameters by minimising a loss function; and (iii) computation of a fit that is then used to predict the output at unobserved inputs.[1] Let us look at a simple two-layer network based on the *sigmoid*

[1]The following article gives a decent introduction to neural networks: HIGHAM & HIGHAM (2019). Deep learning: An introduction for applied mathematicians. *SIAM Review* 61(4):860–891.

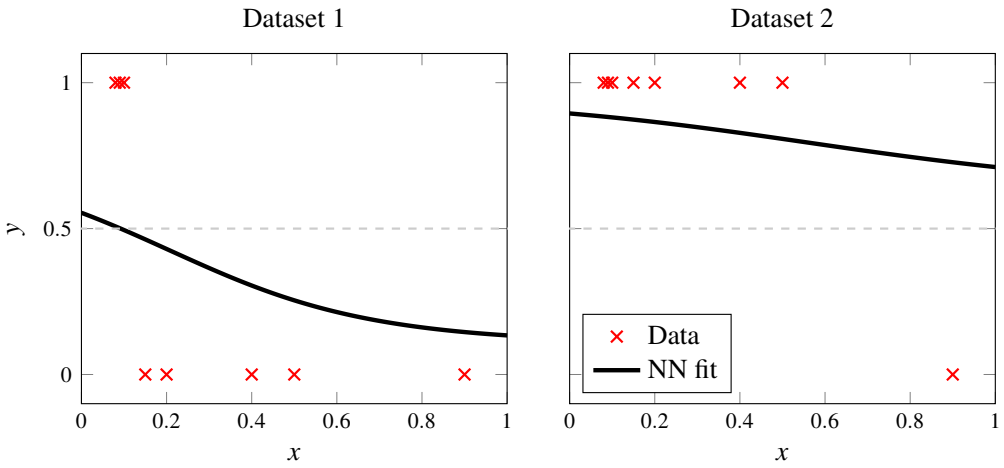

Figure 2: The neural network (NN) fit $f(x) = F(x \mid w_1^*, w_2^*, b_1^*, b_2^*) = \sigma(w_2^*\sigma(w_1^* x + b_1^*) + b_2^*)$ for two different datasets. Here w_1^*, w_2^*, b_1^* and b_2^* have been computed using gradient descent. The datasets consist of $n = 8$ *labels* $y \in \{0, 1\}$ (i.e., points labelled $y = 1$ are in a category A and points those labelled $y = 0$ in category B). After computing the neural network fit f, we may classify a new point by, for example, saying it is in category A if $f(x) > \frac{1}{2}$ and in category B if $f(x) \leq \frac{1}{2}$. This *decision boundary* (suom. *päätöspinta*) is plotted as a dashed grey line. Such classification works for Dataset 1 (at least barely), but not for Dataset 2. This is understandable, as the neural network we have used is extremely simple.

activation function (suom. *aktivaatiofunktio*)

$$\sigma(x) = \frac{1}{1+e^{-x}}. \tag{1.4}$$

Let $w_1, w_2 \in \mathbb{R}$ be *weights* (suom. *paino*) and $b_1, b_2 \in \mathbb{R}$ *biases* (suom. *vakiotermi*). Inspired by (1.1), we postulate that the training outputs $0 \leq y_i \leq 1$ are related to the inputs x_i via the equation

$$y_i = \sigma(w_2\sigma(w_1 x_i + b_1) + b_2) + \varepsilon_i \quad \text{for} \quad i = 1, \ldots, n. \tag{1.5}$$

Define

$$F(x \mid w_1, w_2, b_1, b_2) = \sigma(w_2\sigma(w_1 x + b_1) + b_2), \tag{1.6}$$

so that (1.5) becomes $y_i = F(x_i \mid w_1, w_2, b_1, b_2) + \varepsilon_i$. As in the case of linear regression, we may define a quadratic loss function

$$L(w_1, w_2, b_1, b_2) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - F(x_i \mid w_1, w_2, b_1, b_2))^2 \tag{1.7}$$

and select the weights w_1, w_2 and biases b_1, b_2 by minimising this loss function. But we hit a roadblock. For, unlike in the case of ordinary least squares method that provides the simple expression (1.3) for the minimisers of the loss function (1.2), no such nice formula is available now. Instead, we have to device an *optimisation method* to minimise (1.7). Because in real problems there are thousands of weights and biases and the training data sets are huge, it is essential that this method be efficient, in that it should call the loss function as few times as possible. During the course we will learn how to use *derivatives* and *gradients* to describe the direction and rate of change of functions and how to implement the *gradient descent* (suom. *gradienttimenetelmä*) algorithm, the workhorse of machine learning. Results of selecting the parameters w_1, w_2, b_1 and b_2 using gradient descent are displayed in Figure 2.

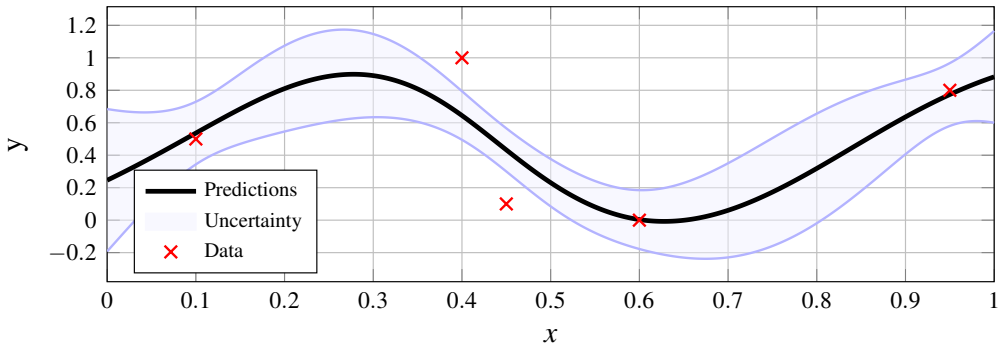

Figure 3: Statistical inference and prediction based on *Gaussian processes* (suom. *gaussinen prosessi*). Given some data, we construct a certain *statistical model* (suom. *tilastollinen malli*) that attempts to represent the process that generated the data (i.e., some relationship between x and y). Using this model and statistical assumptions about noise present in the data we can perform predictive inference. The predictions are uncertain, as represented by the shaded region. While our best prediction for the data value at, say, $x = 0.7$ is $y \approx 0.05$, we acknowledge that there is significant uncertainty in this prediction, considering "it very likely" that $y \in [-0.2, 0.4]$.

1.1.3 Statistical Inference and Probability

Much of machine learning and data analysis is based on *statistical inference* (suom. *tilastollinen päättely*). That is, given some data one seeks to infer properties of an underlying probability distribution and exploit these properties to make predictions. The foundations of statistical inference are in the *theory of probability* (suom. *todennäköisyysteoria*), which provides a mathematical language for decision-making and uncertainty. Figure 3 illustrates a certain statistical inference and prediction method popular in machine learning and spatial statistics. During the course we will learn the basic definitions, concepts and results of probability theory.

1.2 Notational Conventions

Number sets. The sets $\mathbb{R}$, $\mathbb{N}$, and $\mathbb{Z}$ of real numbers, natural numbers, and integers are

$$\mathbb{R} = (-\infty, \infty), \quad \mathbb{N} = \{1, 2, \ldots\}, \quad \text{and} \quad \mathbb{Z} = \{\ldots, -2, -1, 0, 1, 2, \ldots\},$$

respectively. Note that this definition excludes zero from $\mathbb{N}$ and that the infinities $-\infty$ and ∞ are not included in $\mathbb{R}$.

Scalars, vectors and matrices. When talking about scalars, vectors, and matrices, we will use plain font (i.e., x or X) for scalars, bold lowercase font for vectors (i.e., $\mathbf{x}$), and bold uppercase font for matrices (i.e., $\mathbf{X}$). Note that, depending on the dimensions, $\mathbf{x}$ may happen to be a scalar and $\mathbf{X}$ a vector (or even a scalar). But x will never be a vector or a matrix.

Constants and functions. The constant $e \approx 2.718$ is *Euler's constant*, or *Napier's constant* (suom. *Neperin luku*). Moreover, $\exp(x) = e^x$. Euler's constant is the base of the natural logarithm. The natural logarithm of x is always denoted $\log(x)$ in these notes. It is the inverse function of e^x, satisfying $\log(e^x) = x \log(e) = x$. We do not use logarithms with other bases.

1.3 Preliminaries on Functions

Let us recall some basic notation and concepts that we will use throughout the course.

Definition 1.1 (FUNCTION). Let X and Y be sets. A *function* (suom. *funktio*) f from X to Y, denoted

$$f\colon X \to Y, \tag{1.8}$$

assigns to every element $x \in X$ exactly one element $f(x) \in Y$. The function f *maps* (suom. *kuvaa*) the *input* x to the *output* $f(x)$. This input-output relationship is often written as

$$x \mapsto f(x). \tag{1.9}$$

While there can be only one output $f(x)$ for each $x \in X$, each output does not have to correspond to a unique input. That is, different inputs can yield the same output [i.e., there can be $x_1 \neq x_2$ such that $f(x_1) = f(x_2)$]. Examples of such functions include any constant function from $\mathbb{R}$ to $\mathbb{R}$ given by $f(x) = c$ for some fixed $c \in \mathbb{R}$ or the quadratic function $f(x) = x^2$ from $\mathbb{R}$ to $\mathbb{R}$, which satisfies $f(1) = f(-1) = 1$.

In this course we mostly focus on functions $f\colon \mathbb{R}^d \to \mathbb{R}$ (i.e., $X = \mathbb{R}^d$ and $Y = \mathbb{R}$) that map vectors $\mathbf{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d$ to real numbers $f(\mathbf{x}) \in \mathbb{R}$. Examples of such functions are the bivariate polynomial $f(\mathbf{x}) = f(x_1, x_2) = x_1^2 x_2^3 + x_1 - 3$ (i.e., $f\colon \mathbb{R}^2 \to \mathbb{R}$) and the *ReLU activation function*

$$f(\mathbf{x}) = \max\{0, a + \mathbf{w}\cdot\mathbf{x}\} = \max\{0, a + \textstyle\sum_{i=1}^d w_i x_i\} = \begin{cases} 0 & \text{if} \quad a + \mathbf{w}\cdot\mathbf{x} < 0, \\ a + \mathbf{w}\cdot\mathbf{x} & \text{otherwise,} \end{cases} \tag{1.10}$$

where $a \in \mathbb{R}$ and $\mathbf{w} \in \mathbb{R}^d$ are fixed (i.e., $f\colon \mathbb{R}^d \to \mathbb{R}$). However, a function can be something much more complicated: A computer program that maps user input, such as text or parameter values, to some output, such as a text response or the result of a physics simulation, is also a function.

1.4 Function Composition

Complicated functions are often formed by *composing* simple functions.

Definition 1.2 (FUNCTION COMPOSITION). Let $g\colon X \to Y$ and $h\colon Y \to Z$ be functions. Their *composite function* (suom. *yhdistetty funktio*) $f = h \circ g$ is a function from X to Z given by

$$f(x) = h(g(x)). \tag{1.11}$$

To compute $f(x) = h(g(x))$ we (i) first map x through g and (ii) after this map the output $g(x)$ through h. We can also compose more than two functions, so that

$$f = r \circ h \circ g \quad \text{is given by} \quad f(x) = (r \circ h \circ g)(x) = r(h(g(x))). \tag{1.12}$$

One can think recursively:

$$f = r \circ h \circ g = r \circ f_1, \quad \text{where} \quad f_1 = h \circ g. \tag{1.13}$$

Example 1.3. Define the functions

$$g(x) = 3\pi, \quad h(x) = \log(1+x), \quad r(x) = 2 + x^{-3} \quad \text{and} \quad l(x) = x, \tag{1.14}$$

all from $\mathbb{R}$ to $\mathbb{R}$. Then

$$f(x) = (h \circ g)(x) = h(g(x)) = h(3\pi) = \log(1+3\pi), \tag{1.15a}$$
$$f(x) = (g \circ h)(x) = g(h(x)) = g(\log(1+x)) = 3\pi, \tag{1.15b}$$
$$f(x) = (h \circ r)(x) = h(r(x)) = h(2+x^{-3}) = \log(3+x^{-3}), \tag{1.15c}$$
$$f(x) = (r \circ l)(x) = r(l(x)) = r(x) = 2+x^{-3}, \tag{1.15d}$$
$$f(x) = (r \circ l \circ h)(x) = r(l(h(x))) = r(h(x)) = r(\log(1+x)) = 2+\log(1+x)^{-3}. \tag{1.15e}$$

Note that (1.15a) and (1.15b) are simply the *constant functions* $x \mapsto \log(1+3\pi)$ and $x \mapsto 3\pi$, which map every $x \in \mathbb{R}$ to the constants $\log(1+3\pi)$ and 3π, respectively.

Example 1.4. We can write the ReLU activation function (1.10) as the composition

$$f = h \circ g, \tag{1.16}$$

where

$$g(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} \quad \text{and} \quad h(x) = \max\{0, a+x\} \tag{1.17}$$

are functions from $\mathbb{R}^d$ to $\mathbb{R}$ and from $\mathbb{R}$ to $\mathbb{R}$, respectively. The composition is *not* unique. We could have alternatively selected $g(\mathbf{x}) = a + \mathbf{w} \cdot \mathbf{x}$ and $h(x) = \max\{0, x\}$.

Example 1.5. Consider the following idiotic piece of pseudocode:

```
1: procedure FUNC(scalar x, integer n ≥ 1)
2:     if n ≥ 1 then
3:         return 2 × FUNC(x, n − 1)
4:     else
5:         return x
6:     end if
7: end procedure
```

Given an integer constant $n \geq 1$, this procedure computes $f_n(x) = \text{FUNC}(x, n) = 2^n x$ by using the n-fold composition

$$f_n(x) = \underbrace{(g \circ \cdots \circ g)}_{n \text{ times}}(x), \quad \text{where} \quad g(x) = 2x. \tag{1.18}$$

For example, when $n = 3$ we have $f_n(x) = (g \circ g \circ g)(x) = g(g(g(x))) = 2(2(2x)) = 2^3 x = 8x$.

1.5 Piecewise defined functions

We will occasionally encounter *piecewise defined functions* (suom. *paloittain määritelty funktion*), such as the ReLU activation function (1.10). Typical piecewise defined functions include the *step function* (suom. *porrasfunktio*)

$$f(x) = \begin{cases} 0 & \text{if } x < 0, \\ 1 & \text{if } x \geq 0, \end{cases} \tag{1.19}$$

the *minimum* and *maximum*

$$f(x) = \min\{0, x\} = \begin{cases} x & \text{if } x < 0, \\ 0 & \text{if } x \geq 0 \end{cases} \quad \text{and} \quad f(x) = \max\{0, x\} = \begin{cases} 0 & \text{if } x < 0, \\ x & \text{if } x \geq 0, \end{cases} \tag{1.20}$$

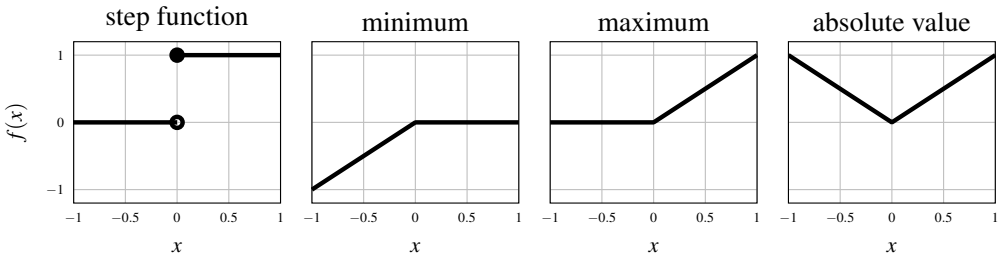

Figure 4: Four piecewise defined functions. From left to right: The step function in (1.19), the minimum and maximum functions in (1.20), and the absolute value function in (1.21).

and the *absolute value* (suom. *itseisarvo*)

$$f(x) = |x| = \begin{cases} -x & \text{if } x < 0, \\ x & \text{if } x \geq 0. \end{cases} \tag{1.21}$$

These four functions are plotted in Figure 4.

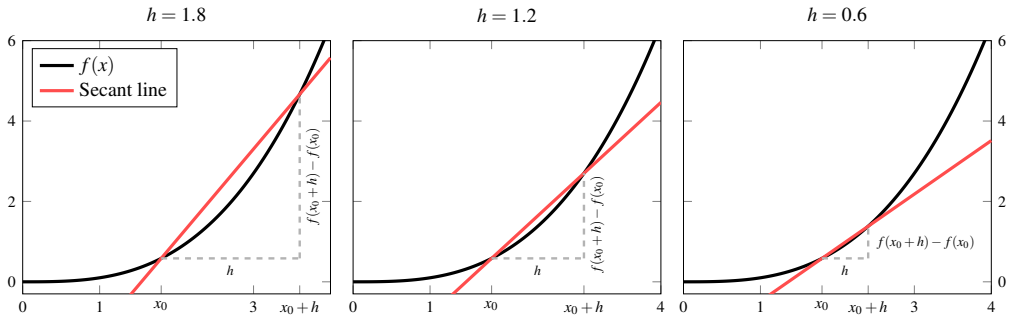

Figure 5: Secant lines for the function $f(x) = \frac{1}{10}x^3$ and $h = 1.8$, $h = 1.2$, and $h = 0.6$ at the point $x_0 = 1.8$. Observe how the secant line "barely touches" the graph of f when $h = 0.6$.

2 Calculus

In this section we consider *univariate* functions $f\colon \mathbb{R} \to \mathbb{R}$, which map reals to reals. *Multivariate* functions will be considered in Section 3. Our treatment of differentiation is by necessity very brief, cursory, and non-rigorous. If you need more practice with differentiation, you may wish to look at, for example, Chapter 3 in *Calculus: Volume 1* by Strang and Herman (freely available at https://openstax.org/details/books/calculus-volume-1). The book used on the courses taught in Finnish on differentiation and related topics at the Department of Mathematics and Statistics is *Analyysiä reaaliluvuilla* by Harjulehto, Klén, and Koskenoja. This book contains a mathematically rigorous and complete treatment of most of the topics covered in the present section.

2.1 Differentiation

We are interested in how to describe and compute the *rate of change* of a function. Having access to a quantity describing how a function changes is very useful in optimisation: When we want to minimise a function, it is natural to move towards the direction in which the function is decreasing.

Definition 2.1 (Difference quotient). Let $h \in \mathbb{R}$. The expression

$$\frac{f(x+h)-f(x)}{h} \tag{2.1}$$

is called the *difference quotient* (suom. *erotusosamäärä*) of the function f at point x.

The difference quotient measures the *slope* (suom. *kulmakerroin*) of the *secant line* (suom. *sekantti*) passing through the points $f(x+h)$ and $f(x)$. As h tends to zero, denoted $h \to 0$, this secant line tends to the *tangent* (suom. *tangentti*), the line that "barely touches" the curve defined by f at $(x, f(x))$; see Figure 5. The slope of the tangent line is the derivative of f at x.

Definition 2.2 (Derivative). The *derivative* (suom. *derivaatta*) of a function f at point x is

$$f'(x) = \frac{\mathrm{d}f}{\mathrm{d}x}(x) = \frac{\mathrm{d}}{\mathrm{d}x}f(x) = \lim_{h\to 0}\frac{f(x+h)-f(x)}{h}. \tag{2.2}$$

Computation of the derivative $f'(x)$ is called *differentiation* (suom. *derivointi*). The derivative is not necessarily well-defined at every $x \in \mathbb{R}$; we shall discuss this in Remark 2.6.

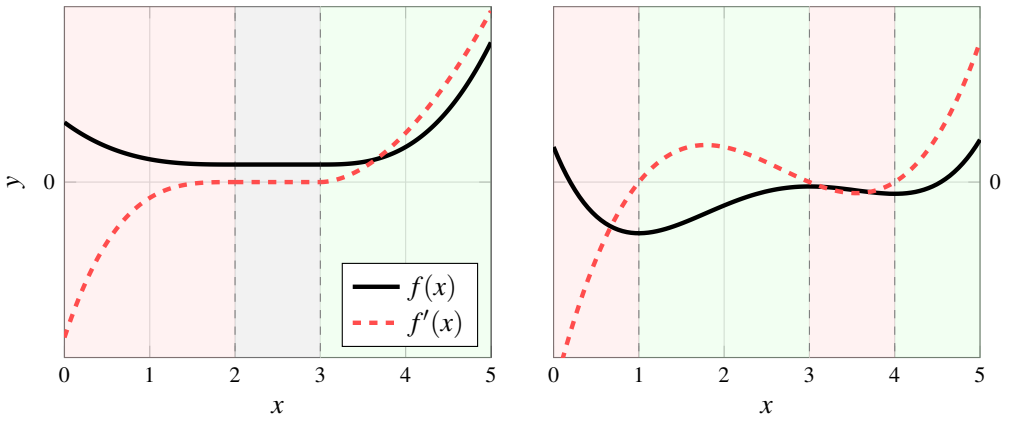

Figure 6: Two functions and their derivatives. The functions are decreasing on the red intervals ($[0,2]$ for the first function and $[0,1]$ and $[3,4]$ for the second), constant on the grey interval ($[2,3]$ for the first function), and increasing on the green intervals ($[3,5]$ for the first functino and $[1,3]$ and $[4,5]$ for the second). When the functions decrease, their derivatives are negative; when they are constant, the derivatives are zero; and when they increase, the derivatives are positive.

While there is a proper mathematical definition[2] for what $\lim_{h\to 0}$ means in (2.2), we will be content with the following informal definition: Let g be a function and x_0 a point. Then

$$g(x) \to c \text{ as } x \to x_0, \text{ or } \lim_{x\to x_0} g(x) = c, \text{ or } \text{“}g(x) \text{ tends to } c \text{ as } x \text{ tends to } x_0\text{”}, \tag{2.3}$$

means, *very* roughly speaking, that $g(x)$ can be made arbitrarily close (or even equal) to c by taking x sufficiently close to x_0. In the context of Equation (2.2) this means that $c = f'(x)$ if the difference quotient $g(h) = [f(x+h) - f(x)]/h$ can be made arbitrarily close to c by taking h sufficiently close to 0. Note that "h is close to 0" permits h to be negative. That is, we are considering $h \in \mathbb{R}$ such that the absolute value $|h|$ is close to zero.

Working with limits is intuitive in many cases. For example, it should be clear that $\lim_{h\to 0} h^n = 0$ for any $n \in \mathbb{N}$ since a number close to zero raised to a positive power remains close to zero; we will see this principle in action in Examples 2.3 and 2.5. Similarly,

$$\lim_{h\to 0} \frac{1+h^n}{e^h} = 1 \tag{2.4}$$

because $1+h^n$ tends to 1 and e^h tends to $e^0 = 1$. However, sometimes such reasoning fails.

Leibniz's notation $\mathrm{d}f/\mathrm{d}x$ for derivative is suggestive: To compute derivative we divide a small difference, $\mathrm{d}f$, in the values of f [i.e., $f(x+h) - f(h)$] by a small difference, $\mathrm{d}x$, in x [i.e., $h = (x+h) - x$]. While Leibniz's notation can be treacherous, its formal manipulation allows one to easily recall many basic differentiation formulae.

Example 2.3. Let $a \in \mathbb{R}$. For the *constant function* $f(x) = a$ we have

$$f'(x) = \lim_{h\to 0} \frac{f(x+h) - f(x)}{h} = \lim_{h\to 0} \frac{a-a}{h} = \lim_{h\to 0} 0 = 0. \tag{2.5}$$

[2]The mathematically rigorous definition of the limit, the so-called ε-δ definition by Bolzano, Cauchy and Weierstrass from the 1800s, goes as follows: A function $g\colon \mathbb{R} \to \mathbb{R}$ has the limit c at a point $x_0 \in \mathbb{R}$, denoted $\lim_{x\to x_0} g(x) = c$, if for every $\varepsilon > 0$ one can find $\delta > 0$ (which can depend on ε) such that $|g(x) - c| < \varepsilon$ for all x such that $|x - x_0| < \delta$. You can learn this type of rigorous calculus on the courses *Raja-arvot* (MAT11003) and *Calculus IA: Limits and Differentiation* (MAT11006).

Because a constant function does not change, its derivative (which describes the rate of change) should indeed be zero. For the *linear function* $f(x) = ax$ we have

$$f'(x) = \lim_{h\to 0} \frac{f(x+h)-f(x)}{h} = \lim_{h\to 0} \frac{a(x+h)-ax}{h} = \lim_{h\to 0} \frac{ah}{h} = a. \tag{2.6}$$

For the *quadratic function* $f(x) = ax^2$ we have

$$f'(x) = \lim_{h\to 0} \frac{a(x+h)^2-ax^2}{h} = \lim_{h\to 0} \frac{ax^2+2ahx+ah^2-ax^2}{h} = \lim_{h\to 0}(2ax+ah) = 2ax. \tag{2.7}$$

Derivative describes both the *rate* and *direction* of change. If $f'(x)$ is positive, the function f is increasing at x; if the derivative is negative, the function is decreasing. If $f'(x) = 0$, the function "does not change", though what this precisely means depends on the situation (we shall return to this in Section 2.5). Figure 6 illustrates each of these cases.

Result 2.4 (DIFFERENTIATION OF ELEMENTARY FUNCTIONS). Let $a \in \mathbb{R}$ be a constant.

$$f(x) = ax^n \implies f'(x) = anx^{n-1} \quad [n \in \mathbb{R}] \tag{2.8}$$
$$f(x) = c \implies f'(x) = 0 \quad [c \in \mathbb{R}] \tag{2.9}$$
$$f(x) = a\log(x) \implies f'(x) = ax^{-1} \quad [x > 0] \tag{2.10}$$
$$f(x) = ae^x \implies f'(x) = ae^x \tag{2.11}$$
$$f(x) = a\sin(x) \implies f'(x) = a\cos(x) \tag{2.12}$$
$$f(x) = a\cos(x) \implies f'(x) = -a\sin(x) \tag{2.13}$$

In Example 2.3 we derived (2.8) for $n \in \{0,1,2\}$. The proof is similar for general $n \in \mathbb{N}\cup\{0\}$.

Example 2.5. Let $f(x) = x^n$ for $n \in \mathbb{N}$. First, observe that

$$\begin{aligned}(x+h)(x+h)(x+h) = (x+h)(x^2+2xh+h^2) &= x(x^2+2xh+h^2)+h(x^2+2xh+h^2)\\ &= x^3+2x^2h+xh^2+hx^2+h(2xh+h^2)\\ &= x^3+3x^2h+h^2a_3,\end{aligned}$$

where $a_3 = 3x+h$. Multiplying this with $x+h$ gives us

$$\begin{aligned}(x+h)^4 = x(x^3+3x^2h+h^2a)+h(x^3+3x^2h+h^2a) &= x^4+3x^3h+xh^2a+hx^3+3x^2h^2+h^3a\\ &= x^4+4x^3h+h^2a_4,\end{aligned}$$

where $a_4 = xa+3x^2+ah$. Iterating this procedure gives us $(x+h)^n = x^n+nx^{n-1}h+h^2a_n$. Therefore

$$\begin{aligned}f'(x) = \lim_{h\to 0}\frac{f(x+h)-f(x)}{h} = \lim_{h\to 0}\frac{(x+h)^n-x^n}{h} &= \lim_{h\to 0}\frac{x^n+nx^{n-1}h+h^2a_n-x^n}{h}\\ &= \lim_{h\to 0}(nx^{n-1}+ha_n)\\ &= nx^{n-1},\end{aligned}$$

which is (2.8). The general form of $(x+h)^n$ is given by the *binomial theorem* (suom.

binomikaava), which states that

$$(x+h)^n = \sum_{k=0}^{n} \binom{n}{k} x^{n-k} h^n, \quad \text{where} \quad \binom{n}{k} = \frac{n!}{k!(n-k)!}. \tag{2.14}$$

Here $n! = 1 \cdot 2 \cdots (n-1) \cdot n$ is the *factorial* (suom. *kertoma*) and we define $0! = 1$. The term $nx^{n-1}h$ that played a crucial role in computation of the derivative is the $k = 1$ term in the binomial theorem since $1! = 1$ and $n!/(n-1)! = n$.

Above we have implicitly assumed that f is *differentiable* (suom. *derivoituva*) at x. That is, that the limit in (2.2) in fact exists. The following remark shows that this need not be the case.

Remark 2.6. Consider the absolute value function $f(x) = |x|$ plotted in Figure 4. Let us try to differentiate it at $x = 0$. For any $h > 0$ the difference quotient in (2.1) is

$$\frac{f(0+h) - f(0)}{h} = \frac{|h| - |0|}{h} = \frac{h}{h} = 1. \tag{2.15}$$

But for $h < 0$ we get (since in this case $|h| = -h$)

$$\frac{f(0+h) - f(0)}{h} = \frac{|h| - |0|}{h} = \frac{-h}{h} = -1. \tag{2.16}$$

This means that there is no single number c to which the difference quotient would tend to as $|h|$ is taken closer and closer to zero. The other three functions in Figure 4 behave similarly at $x = 0$. It should be clear from the figure that the rate of change of any of these functions at $x = 0$ cannot be defined unambiguously, as the rate depends on whether one approaches from left or right. Note however that a function being piecewise defined does not imply that it is not differentiable at the changepoint. For example, the function

$$f(x) = \begin{cases} 0 & \text{if } x < 0, \\ x^2 & \text{if } x \geq 0. \end{cases} \tag{2.17}$$

is differentiable at $x = 0$. By (2.8), its derivative is $f'(x) = 0$ if $x < 0$ and $f'(x) = 2x$ if $x > 0$. Since its derivatives to the left and right of 0 are equal at 0, the function is differentiable also at $x = 0$ and we may write its piecewise defined derivative as

$$f'(x) = \begin{cases} 0 & \text{if } x < 0, \\ 2x & \text{if } x \geq 0. \end{cases} \tag{2.18}$$

Figure 7 shows the functions discussed in this example along with their derivatives.

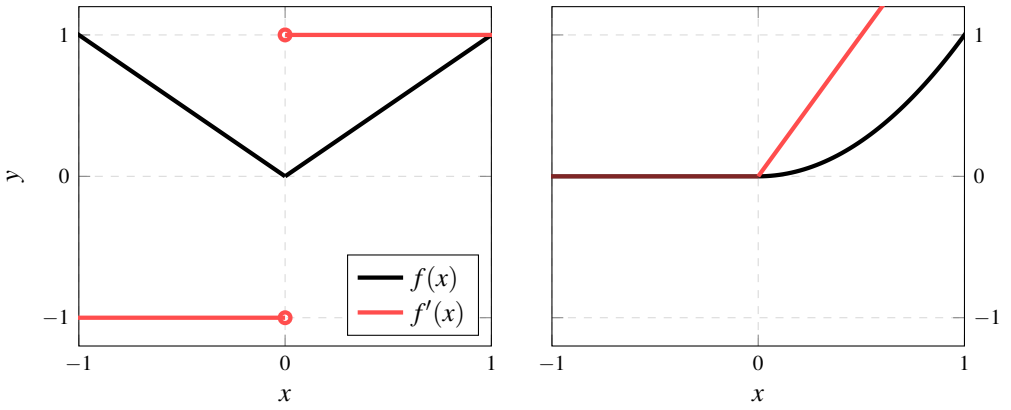

Figure 7: *Left*: The absolute value function $f(x) = |x|$. *Right*: The function in (2.17). We see that the derivative of the absolute value function has a discontinuity at $x = 0$ and is not well-defined at this point. The derivative of the piecewise-defined function in (2.17) is well-defined everywhere. However, its *second* derivative (see Definition 2.7) would have a discontinuity at $x = 0$.

The derivative itself is a function that can be differentiated.

Definition 2.7 (HIGHER-ORDER DERIVATIVES). The nth derivative of a function f is defined recursively as

$$f^{(n)}(x) = \frac{\mathrm{d}}{\mathrm{d}x} f^{(n-1)}(x), \quad \text{where } f^{(0)} = f. \tag{2.19}$$

Note that $f = f^{(0)}$ and $f' = f^{(1)}$. The second derivative $f^{(2)}$ is usually denoted f''. The nth derivative is also written as

$$f^{(n)}(x) = \frac{\mathrm{d}^n f}{\mathrm{d}x^n}(x) = \frac{\mathrm{d}^n}{\mathrm{d}x^n} f(x). \tag{2.20}$$

Just as the derivative describes how the function changes, the nth derivative describes how the $(n-1)$th derivative changes.

Example 2.8. Consider the polynomial $f(x) = x^5$. By (2.8) with $n = 5$ and $a = 1$, the derivative of this polynomial is

$$f'(x) = f^{(1)}(x) = 5x^{5-1} = 5x^4. \tag{2.21}$$

Applying (2.8) to f', now with $n = 4$ and $a = 4$, then yields

$$f''(x) = f^{(2)}(x) = \frac{\mathrm{d}}{\mathrm{d}x} f'(x) = 5 \cdot 4x^{4-1} = 5 \cdot 4x^3. \tag{2.22}$$

Carrying on in this manner, we obtain the fifth derivative

$$f^{(5)}(x) = 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 x^0 = 5!, \tag{2.23}$$

where $5! = 120$ is the factorial from the end of Example 2.5. Because the 5th derivative is a constant function, Equation (2.9) implies that $f^{(6)}$ and all subsequent derivatives are zero.

Example 2.9. Consider the function $f(x) = \sin(x)$. From (2.12) we get

$$f'(x) = f^{(1)}(x) = \cos(x). \tag{2.24}$$

Equation (2.13) then yields

$$f''(x) = f^{(2)}(x) = \frac{\mathrm{d}}{\mathrm{d}x} f'(x) = -\sin(x). \tag{2.25}$$

Similarly,

$$f^{(3)}(x) = \frac{\mathrm{d}}{\mathrm{d}x} f''(x) = -\cos(x) \tag{2.26}$$

and

$$f^{(4)}(x) = \frac{\mathrm{d}}{\mathrm{d}x} f^{(3)}(x) = -(-\sin(x)) = \sin(x). \tag{2.27}$$

Here we observe a pattern emerge:

$$f^{(2n)}(x) = (-1)^n \sin(x) \quad \text{and} \quad f^{(2n+1)}(x) = (-1)^n \cos(x) \tag{2.28}$$

for every $n \geq 0$.

2.2 Differentiation Rules

Result 2.4 provided derivatives of elementary functions. The following differentiation rules are used to differentiate more complicated functions formed out of these elementary functions.

Result 2.10 (LINEARITY). Let $f(x) = ag(x) + bh(x)$ for functions g and h and constants $a, b \in \mathbb{R}$. Then

$$f'(x) = ag'(x) + bh'(x). \tag{2.29}$$

Result 2.11 (PRODUCT RULE). Let $f(x) = g(x)h(x)$. The *product rule* (suom. *tulon derivoimissääntö*) states that

$$f'(x) = g'(x)h(x) + g(x)h'(x). \tag{2.30}$$

Example 2.12. The product rule (2.30) and Result 2.4 yield the following derivatives:

- Let $f(x) = xe^x = g(x)h(x)$ for $g(x) = x$ and $h(x) = e^x$. By (2.8) and (2.11), $g'(x) = 1$ and $h'(x) = e^x$. Therefore (2.30) gives

$$f'(x) = g'(x)h(x) + g(x)h'(x) = 1 \cdot e^x + x \cdot e^x = e^x(1+x). \tag{2.31}$$

- Let $f(x) = e^x \sin(x) = g(x)h(x)$ for $g(x) = e^x$ and $h(x) = \sin(x)$. By (2.11) and (2.12), $g'(x) = e^x$ and $h'(x) = \cos(x)$. Therefore (2.30) gives

$$f'(x) = g'(x)h(x) + g(x)h'(x) = e^x \cdot \sin(x) + e^x \cdot \cos(x) = e^x[\sin(x) + \cos(x)]. \tag{2.32}$$

- Let $f(x) = e^{2x} = e^x \cdot e^x = g(x)h(x)$ for $g(x) = e^x$ and $h(x) = e^x$. By (2.11), $g'(x) = e^x$

and $h'(x) = e^x$. Therefore (2.30) gives

$$f'(x) = g'(x)h(x) + g(x)h'(x) = e^x \cdot e^x + e^x \cdot e^x = 2e^{2x}. \tag{2.33}$$

Result 2.13 (QUOTIENT RULE). Let $f(x) = g(x)/h(x)$. The *quotient rule* (suom. *osamäärän derivoimissääntö*) states that

$$f'(x) = \frac{g'(x)h(x) - g(x)h'(x)}{h(x)^2} \quad \text{if} \quad h(x) \neq 0. \tag{2.34}$$

Example 2.14. The quotient rule (2.34) and Results 2.4 and 2.10 yield the following derivatives:

$$f(x) = \frac{x}{e^x} \implies f'(x) = \frac{e^x - xe^x}{e^{2x}} \qquad [g(x) = x \text{ and } h(x) = e^x]$$

$$f(x) = \frac{\cos(x)}{1+x^2} \implies f'(x) = -\frac{(1+x^2)\sin(x) + 2x\cos(x)}{(1+x^2)^2} \qquad [g(x) = \cos(x) \text{ and } h(x) = 1+x^2]$$

$$f(x) = e^{-x} \implies f'(x) = -e^{-x} \qquad [g(x) = 1 \text{ and } h(x) = e^x]$$

2.3 Chain Rule

Recall from Section 1.4 that $f = h \circ g$ denotes the composite function given by $f(x) = h(g(x))$. The *chain rule* is used to differentiate composite functions.

Result 2.15 (CHAIN RULE). Let $f = h \circ g$. The *chain rule* (suom. *ketjusääntö*) states that

$$f'(x) = (h \circ g)'(x) = (h' \circ g)(x)g'(x) = h'(g(x))g'(x). \tag{2.35}$$

The procedure to apply chain rule to compute $f'(x)$ goes as follows:

1. Compute the derivative h' of h.
2. Compute $h'(g(x))$, the value of this derivative at $g(x)$.
3. Multiply by $g'(x)$, the derivative of g at x.

Leibniz's notation provides an easy mnemonic for the chain rule. By treating $\mathrm{d}f$, $\mathrm{d}x$, and other such quantities as numbers, we can formally write

$$f' = \frac{\mathrm{d}f}{\mathrm{d}x} = \frac{\mathrm{d}f}{\mathrm{d}g} \cdot \frac{\mathrm{d}g}{\mathrm{d}x}. \tag{2.36}$$

Here we first differentiate f with respect to g, which means that we treat g as the argument to f. Because $f = h \circ g = h(g)$, this gives us

$$\frac{\mathrm{d}f}{\mathrm{d}g} = \frac{\mathrm{d}h}{\mathrm{d}g}(g) = h'(g) = h' \circ g. \tag{2.37}$$

From (2.36) we then obtain the chain rule in (2.35):

$$f'(x) = \frac{\mathrm{d}f}{\mathrm{d}x}(x) = \frac{\mathrm{d}f}{\mathrm{d}g}(x) \cdot \frac{\mathrm{d}g}{\mathrm{d}x}(x) = (h' \circ g)(x)g'(x) = h'(g(x))g'(x). \tag{2.38}$$

Again, these are purely formal computations (i.e., we treat and manipulate the derivative as if it were the quotient of two numbers, forgetting the presence of a limit in its definition). Although the result is correct, Equation (2.36) is *not* a rigorous mathematical proof of the chain rule.

Example 2.16. Let $h(x) = \log(x)$ and $g(x) = x^n$. By the chain rule and (2.8) and (2.10),

$$f'(x) = (h \circ g)'(x) = h'(g(x))g'(x) = \frac{1}{g(x)} \cdot g'(x) = \frac{1}{x^n} \cdot nx^{n-1} = \frac{n}{x}. \tag{2.39}$$

Note that we obtain the same derivative from $\log(x^n) = n\log(x)$ and Result 2.10.

Example 2.17. Let $f(x) = \exp(-x^2)$. Set $h(x) = \exp(x)$ and $g(x) = -x^2$, so that $f = h \circ g$. Then the chain rule and Results 2.4 and 2.10 yield

$$f'(x) = h'(g(x))g'(x) = \exp(g(x)))g'(x) = \exp(-x^2) \cdot (-2x) = -2x\exp(-x^2). \tag{2.40}$$

Example 2.18. Let $f = h(ax)$ for a function h and a constant $a \in \mathbb{R}$. By setting $g(x) = ax$, we obtain from the chain rule, Equation (2.8) with $n = 1$ and Result 2.10 that

$$f'(x) = h'(g(x))g'(x) = a\,h'(ax). \tag{2.41}$$

Note that this differs from computing $h'(ax)$, the derivative of h at point ax.

Example 2.19. The derivative of the logarithm in (2.10) can be derived from the chain rule and $\mathrm{d}e^x/\mathrm{d}x = e^x$. Let $g(x) = e^x$ and $h(x) = \log(x)$. Then $f(x) = h(g(x)) = \log(e^x) = x$, so that $f'(x) = 1$ by (2.8). But we can alternatively apply the chain rule, which gives

$$f'(x) = h'(g(x))g'(x) = \left(\frac{\mathrm{d}h}{\mathrm{d}x}(e^x)\right)e^x = h'(e^x)\,e^x. \tag{2.42}$$

Since we know that (2.42) equals 1, division by e^x gives us

$$h'(e^x) = e^{-x}. \tag{2.43}$$

Note that this equation says the the derivative of the logarithm h' equals e^{-x} when evaluated at e^x. By selecting $x = \log(z)$, so that $e^{\log z} = z$ and $e^{-x} = e^{-\log z} = z^{-1}$, we obtain

$$h'(z) = z^{-1}, \tag{2.44}$$

which is (2.10).

2.4 Automatic Differentiation

While it is good to have some grasp of the basic differentiation rules, the derivatives involved in applications quickly become practically impossible (or at least extremely tedious) to compute by hand. It would be preferable to have computer take care of differentiation. This is possible in three different ways.

2.4.1 Numerical Differentiation

In *numerical differentiation* the limit $h \to 0$ in the definition (2.2) of derivative is replaced with an h close to zero. That is, the derivative is approximation as the difference quotient

$$f'(x) \approx \frac{f(x+h) - f(x)}{h}, \quad \text{where} \quad h \approx 0. \tag{2.45}$$

This approach is conceptually straightforward and simple to implement (at least in a very naive way). However, numerical differentiation provides only an approximation to the derivative, can introduce round-off errors, and does not scale well to higher-order derivatives and the multivariate setting.

2.4.2 *Symbolic Differentiation*

Symbolic computation packages can be used to perform *symbolic differentiation*, which automates the process of computing symbolic expressions for derivatives. For example, inputting (here sinh and cosh are the *hyperbolic* sine and cosine functions)

```
differentiate sin(log(x^(sqrt(exp(-cos(x)))) + x^2) + cosh(x))
```

to WolframAlpha yields the lovely expression

$$\cos\left(\log(x^{\sqrt{\exp(-\cos(x))}}+x^2)+\cosh(x)\right) \times\left(\frac{x^{\sqrt{\exp(-\cos(x))}}[x^{-1}\sqrt{\exp(-\cos(x))}+\frac{1}{2}\log(x)\sin(x)\sqrt{\exp(-\cos(x))}\,]+2x}{x^{\sqrt{\exp(-\cos(x))}}+x^2}+\sinh(x)\right)$$

for the derivative of the function

$$f(x)=\sin\left(\log(x^{\sqrt{\exp(-\cos(x))}}+x^2)+\cosh(x)\right). \tag{2.46}$$

This is quite impressive and useful if one needs to study how the derivative behaves as a part of some mathematical problem (or when it is part of an exercise to compute a derivative). However, computing such symbolic expression is computationally expensive and *completely unnecessary*. For in practice it is not a symbolic expression such as this that is needed *but values of a derivative at certain points*. To compute $f'(x_0)$ at a given point x_0 (e.g., $x_0=3$) it is not necessary to have access to a full symbolic expression for the derivative.

2.4.3 *Automatic Differentiation*

Automatic differentiation[3] utilises the fact that complicated functions reduce to a sequence of compositions and arithmetic operations involving the elementary functions whose derivatives are available in Result 2.4. The chain rule and other differentiation rules then permit algorithmically straightforward computation of derivatives of arbitrarily complicated functions. For example, consider a function f given by

$$f(x)=h(g(h(x)g(x))+h(x)). \tag{2.47}$$

We may write this function as

$$f(x)=h(r(x)), \tag{2.48}$$

where

$$r(x)=u(x)+h(x),\quad u(x)=g(s(x)),\quad \text{and}\quad s(x)=h(x)g(x). \tag{2.49}$$

By applying the differentiation rules from Sections 2.2 and 2.3 we can write the derivative of f as

$$f'(x)=h'(r(x))r'(x), \tag{2.50}$$

where

$$r'(x)=u'(x)+h'(x),\quad u'(x)=g'(s(x))s'(x)\quad \text{and}\quad s'(x)=h'(x)g(x)+h(x)g'(x). \tag{2.51}$$

[3]The following article gives a good introduction to automatic differentiation: NEIDINGER (2010). Introduction to automatic differentiation and MATLAB object-oriented programming. *SIAM Review* 52(3):545–563.

If h and g are elementary functions appearing in Result 2.4, we can easily compute $s'(x_0)$ at any point x_0 we want. Subsequently we can use $u'(x_0) = g'(s(x_0))s'(x_0)$ to obtain the derivative $u'(x_0)$, which then gives us $r'(x_0)$. Finally, we can use $r'(x_0)$ in (2.50) to compute $f'(x_0)$. To obtain $f'(x_1)$ we simply repeat the process with $x = x_1$. At no point do we require access to a symbolic expression for $f'(x)$ or the intermediate derivatives $r'(x)$, $u'(x)$ and $s'(x)$.

Implementing very basic automatic differentiation is surprisingly simple. The following Python code suffices to differentiate the function in (2.46). This is the so-called *forward mode* of automatic differentiation.

```
import numpy as np

class fDf:
    def __init__(self, val, deriv):
        self.val = val
        self.deriv = deriv
    def __add__(self, u):
        return fDf(self.val + u.val, self.deriv + u.deriv)
    def __mul__(self, u):
        return fDf(self.val * u.val, self.deriv * u.val + self.val * u.deriv)
    def __rmul__(self, c):
        return fDf(c * self.val, c * self.deriv)

def exp(u):
    return fDf(np.exp(u.val), np.exp(u.val) * u.deriv)
def sin(u):
    return fDf(np.sin(u.val), np.cos(u.val) * u.deriv)
def log(u):
    return fDf(np.log(u.val), (1.0 / u.val) * u.deriv)
def sqrt(u):
    return fDf(np.sqrt(u.val), (0.5 / np.sqrt(u.val)) * u.deriv)
def cosh(u):
    return fDf(np.cosh(u.val), np.sinh(u.val) * u.deriv)

x = fDf(3.0, 1.0)
fDfx = sin(log(exp(log(x) * sqrt(exp(-1.0*cos(x)))) + x * x) + cosh(x))
Dfx = fDfx.deriv
```

We first define a class `fDf` which holds the value and derivative of a function at a point of interest. By overloading addition and multiplication we can perform arithmetic on this class. For example, for two instances `f` and `g` of the class `fDf`, which store $f(x)$ and $g(x)$ as well as $f'(x)$ and $g'(x)$ at some point x, the multiplication `f * g` is an instance of `fDf` that stores the function value $f(x)g(x)$ and the derivative $f'(x)g(x) + f(x)g'(x)$ at x that has been computed using the product rule (Result 2.11). We then define a number of functions that operate on instances of `fDf`. Each of these functions calls the corresponding `numpy` to perform function evaluation and implements the chain rule (Result 2.15) for derivative evaluation. For example, `exp` assumes that it is given a function $u(x)$ and its derivative $u'(x)$ at some point x and uses these to compute the function evaluation $e^{u(x)}$ and, via the chain rule, the derivative evaluation $e^{u(x)}u'(x)$. Wit this code we can differentiate any function that involves sums, products and compositions of exp, sin, cos, log, square root and cosh. At the end we specify that we want to differentiate the function in (2.46) at $x_0 = 3$. The line `x = fDf(3.0, 1.0)` does this by defining that the innermost function in automatic differentiation is $u(x) = x$ and that $u(x_0) = 3$ and $u'(x_0) = 1$ since $u'(x) = 1$. Note that in defining the function being differentiated we use $x^{u(x)} = \exp(u(x)\log(x))$ in order to avoid having to define a differentiation rule for $x^{u(x)}$.

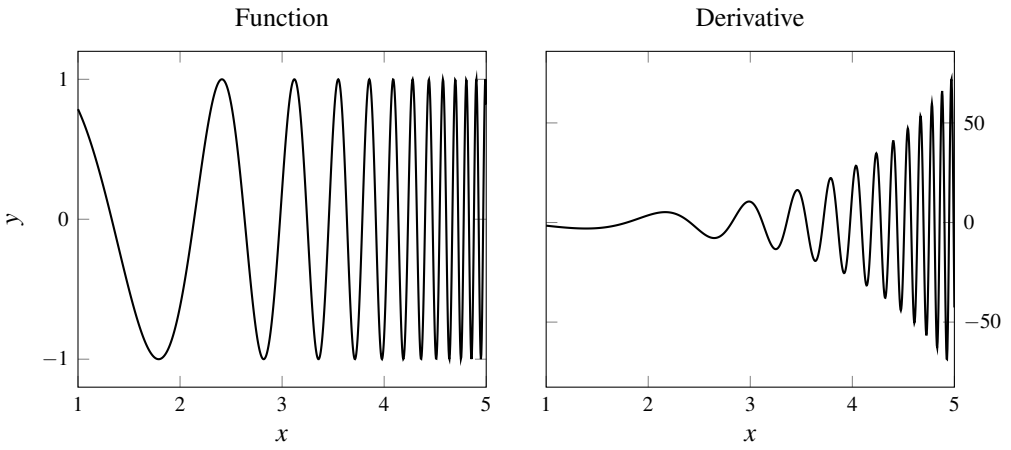

Figure 8: The function in (2.46) and its derivative that has been computed using `JAX`.

What makes automatic differentiation very powerful is the fact that the function being differentiated does not have to a simple algebraic expression. For example, with the above code we could easily differentiate functions that have been defined using loops and conditional statements. Of course, one should not use home-made automatic differentiation routines. One popular option is `JAX` (https://jax.readthedocs.io/), the usage of which to differentiate (2.46) is illustrated by the following code.

```
import jax.numpy as jnp
from jax import grad

def f(x):
    return jnp.sin(jnp.log(jnp.exp(jnp.log(x) *
        jnp.sqrt(jnp.exp(-1.0*jnp.cos(x)))) + x**2) + jnp.cosh(x))

Df = grad(f)
x = 3.0
Dfx = Df(x)
```

The derivative `Dfx = Df(x)` is now equal to the derivative `Dfx = fDfx.deriv` computed by our earlier code. The crucial JAX function `grad` performs automatic differentiation. Its name stands for "gradient", the multivariate generalisation of derivative that we shall learn about in Section 3. We will return to JAX in Section 3.4. Figure 8 shows the function in (2.46) and its derivative that has been computed using `JAX`.

2.5 Univariate Local Optimisation

The purpose of Sections 2.1 to 2.4 has been to review some prerequisites for *local optimisation* (suom. *paikallinen optimointi*) of univariate functions. The goal of optimisation is to find point (or points) x^* at which a function f of interest attains its smallest or largest value. In machine learning f is usually a loss function that is to be minimised (recall the examples in Section 1.1). We therefore focus on minimisation. Note that this comes at no loss of generality because

$$\text{maximising } f \quad \text{is equivalent to} \quad \text{minimising } -f.$$

Definition 2.20 (LOCAL AND GLOBAL MINIMA). Let $f\colon \mathbb{R} \to \mathbb{R}$ be a function. A point $x^* \in \mathbb{R}$ is a

- *global minimum point* of f if $f(x^*) \leq f(x)$ for all $x \in \mathbb{R}$;
- *local minimum point* of f if there exists $\varepsilon > 0$ such that $f(x^*) \leq f(x)$ for every point x in the interval $[x^* - \varepsilon, x^* + \varepsilon]$ of length 2ε centered at x^*.

If x^* is a global minimum point, $f(x^*)$ is called the *global minimum* (suom. *globaali minimi*). If x^* is a local minimum point, $f(x^*)$ is called a *local minimum* (suom. *paikallinen minimi*).

Global and local maxima are defined analogously, with "$\geq$" replacing "$\leq$" in the definitions. Minima and maxima are collectively known as *extrema* (suom. *ääriarvo*) and minimum and maximum points as *extremal points* (suom. *ääriarvopiste*).

A function attains its smallest possible value at a global minimum point. At a local minimum point a function attains a value that is smaller or equal to its "nearby" values. The definition of a local minimum point says simply that there is *some* interval around x^* on which $f(x*) \leq f(x)$. Note that a global minimum point is also a local minimum point. Figure 9 shows global and local minima of a particular function. We focus on local optimisation as opposed to *global optimisation*. That is, we are happy find a local minimum point. Finding a global minimum point is significantly more challenging. Points at which the derivative of a function is zero and local minimum and maximum points are intimately connected.

Definition 2.21 (CRITICAL POINT). A point $x^* \in \mathbb{R}$ is a *critical point* (suom. *kriittinen piste*), or a *stationary point*, of a differentiable function f if $f'(x^*) = 0$.

Example 2.22. Consider the functions $f(x) = \cos(x)$ and $g(x) = \frac{1}{3}x^3 - x$. Equation (2.13) gives $f'(x) = -\sin(x)$. From trigonometry we know that $\sin(x) = 0$ if and only if $x = \pi n$ for n an integer (i.e., $n \in \mathbb{Z}$). Therefore $f'(x) = 0$ if and only if $x = \pi n$ for an integer n. These are thus the critical points of $f(x) = \cos(x)$. That is, the cosine has an infinite number of critical points. Equation (2.8) gives $g'(x) = x^2 - 1$, so that $g'(x) = 0$ if and only if $x = 1$ or $x = -1$. Therefore the polynomial $g(x) = \frac{1}{3}x^3 - x$ has two critical points, $x^* = -1$ and $x^* = 1$.

Result 2.23 (FERMAT'S THEOREM). Every local extremal point x^* of a differentiable function f is a critical point, in that $f'(x^*) = 0$.

Proof. Suppose that x^* is a local extremal point. Assume that it is a local minimum point (the proof is completely analogous for a local maximum point). By the definition of a local minimum point (Definition 2.20), there is $\varepsilon > 0$ such that

$$f(x^*) \leq f(x) \quad \text{for all} \quad x \in [x^* - \varepsilon, x^* + \varepsilon]. \tag{2.52}$$

Recall from Section 2.1 that $f'(x^*)$ is defined as the limit

$$f'(x^*) = \lim_{h \to 0} \frac{f(x^* + h) - f(x^*)}{h}. \tag{2.53}$$

For every $h \in (0, \varepsilon)$ the difference quotient satisfies

$$\frac{f(x^*+h)-f(x^*)}{h} \leq \frac{f(x^*)-f(x^*)}{h} = 0 \tag{2.54}$$

since, in this case, $f(x^*+h) \leq f(x^*)$ by (2.52). For $h \in (-\varepsilon, 0)$ we similarly obtain

$$\frac{f(x^*+h)-f(x^*)}{h} = \frac{f(x^*)-f(x^*+h)}{|h|} \geq \frac{f(x^*)-f(x^*)}{|h|} = 0, \tag{2.55}$$

where we have used the fact that $h = -|h|$ if $h < 0$. Equations (2.54) and (2.55) show that the difference quotient (a) is non-positive if h is positive and close to zero (b) non-negative if h negative and close to zero. Because the derivative is the limit of the difference quotient as $h \to 0$ and the only number that is simultaneously non-positive and non-negative is 0, we conclude that $f(x^*) = 0$. □

We can see Result 2.23 in action in Figure 9: at each extremal point the tangent line would be completely horizontal and the derivative hence zero. This result furnishes the basic principle behind local optimisation methods: find a point at which the derivative is zero. However, Result 2.23 only says that extremal points are critical points, not that every critical point is an extremal point. Whether or not a critical point is an extremal point can be usually determined by examining the derivative between critical points.

Result 2.24 (CLASSIFICATION OF CRITICAL POINTS). Let f be a differentiable function. Suppose that $x_1^* < x_2^* < x_3^*$ are critical points of f.

1. If $f'(x) > 0$ for all $x_1^* < x < x_2^*$ and $f'(x) > 0$ for all $x_2^* < x < x_3^*$, then x_2^* is *not an extremal point*.
2. If $f'(x) < 0$ for all $x_1^* < x < x_2^*$ and $f'(x) < 0$ for all $x_2^* < x < x_3^*$, then x_2^* is *not an extremal point*.
3. If $f'(x) > 0$ for all $x_1^* < x < x_2^*$ and $f'(x) < 0$ for all $x_2^* < x < x_3^*$, then x_2^* is a *local maximum point*.
4. If $f'(x) < 0$ for all $x_1^* < x < x_2^*$ and $f'(x) > 0$ for all $x_2^* < x < x_3^*$, then x_2^* is a *local minimum point*.

Recall that the sign of its derivative tells us whether a function is increasing or decreasing. In Cases 1 and 2 of Result 2.24 the function is increasing (Case 1) or decreasing (Case 2) both to the left and right of the critical point x_2^*. Therefore x_2^* cannot be an extremal point. In Case 3 the function is increasing to the left of x_2^* and decreasing to the right of it, meaning that it must attain a local maximum at x_2^*. In Case 4 the function is decreasing to the left and increasing to the right, so that it must have a local minimum at x_2^*. The following example and Figure 10 illustrate this.

Example 2.25. Let us classify the critical points of the function

$$f(x) = x^3 \exp(-\tfrac{1}{2}x^2). \tag{2.56}$$

Using the differentiation rules in Sections 2.1 to 2.3 we compute

$$f'(x) = (3x^2 - x^4)\exp(-\tfrac{1}{2}x^2). \tag{2.57}$$

Since the function $\exp(-\frac{1}{2}x^2)$ is everywhere positive, $f'(x) = 0$ if and only if $3x^2 - x^4 = 0$.

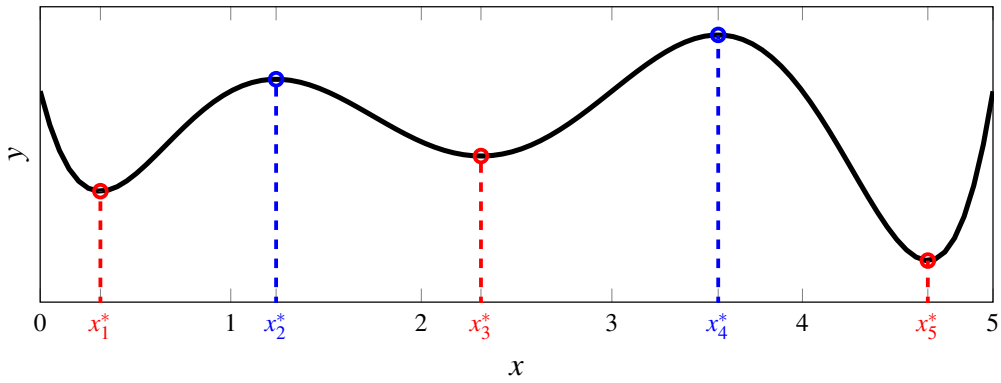

Figure 9: The local minimum points (red; x_1^*, x_3^*, and x_5^*) and maximum points (blue; x_2^* and x_4^*) of a certain function. The local minimum point x_5^* is also a *global* minimum point. Although this cannot be determined from the figure, this function has no global maximum point as it tends to ∞ as $x \to -\infty$ and $x \to \infty$.

The equation

$$3x^2 - x^4 = x^2(3 - x^2) = 0 \tag{2.58}$$

has the three solutions $x = 0$, $x = \sqrt{3}$, and $x = -\sqrt{3}$. The critical points of f are therefore

$$x_1^* = -\sqrt{3}, \quad x_2^* = 0, \quad \text{and} \quad x_3^* = \sqrt{3}. \tag{2.59}$$

Let us begin with x_2^*. The sign of the derivative (2.57) is equal to the sign of $3x^2 - x^4$. By writing this as $x^2(3 - x^2)$, we observe that

$$f'(x) > 0 \quad \text{for all} \quad -\sqrt{3} < x < 0 \quad \text{and} \quad 0 < x < \sqrt{3}. \tag{2.60}$$

Therefore f' does *not* change sign at $x_2^* = 0$ and thus x_2^* is not an extremal point by Case 1 of Result 2.24. For $x_1^* = -\sqrt{3}$ and $x_3^* = \sqrt{3}$ we note that

$$f'(x) < 0 \quad \text{for all} \quad x < -\sqrt{3} \quad \text{and} \; x > \sqrt{3}. \tag{2.61}$$

Therefore f is decreasing to the left of x_1^* and increasing to the right of it, meaning that x_1^* is a local (in fact, global) minimum point. Conversely, f is increasing to the left of x_3^* and increasing to the right of it, meaning that x_3^* is a local (in fact, global) maximum point. Figure 10 attempts to show what is going on.

Result 2.26 (MINIMUM OF A POLYNOMIAL). Let $P(x) = a_0 + a_1 x + \cdots + a_{n-1}x^{n-1} + a_n x^n$ be a polynomial of degree n. Suppose that $a_n \neq 0$. Then P has a global minimum if and only if n is even and $a_n > 0$.

When minimising a function it is natural to move to a direction in which the function decreases. This basic idea is behind the following *gradient descent* (suom. *gradienttimenetelmä*) algorithm. In the multivariate version of this algorithm (Section 3.4) derivatives are replaced by gradients, their generalisations to higher dimensions.

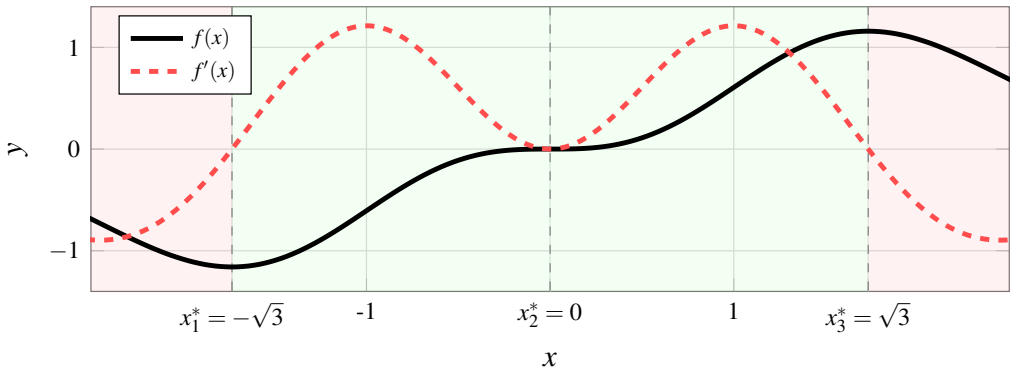

Figure 10: The function $f(x) = x^3 \exp(-\frac{1}{2}x^2)$ and its derivative in (2.57). The function is increasing (i.e., derivative is positive) on the green interval and decreasing (i.e., derivative is negative) on the two red intervals. The derivative vanishes at x_1^*, x_2^*, and x_3^*, which means that these are critical points of f. The critical points x_1^* and x_3^* are extremal points, but the critical point $x_2 = 0$ is not because f is increasing both the left and right of x_2^*.

Algorithm 2.27 (UNIVARIATE GRADIENT DESCENT). The following algorithm attempts to find a local minimum point of a function f with derivative f':

Require: function handle f', initial point x_0, learning rate $\eta > 0$, tolerance $t > 0$, maximum number of iterations $n_{\max} \in \mathbb{N}$

1: $n \leftarrow 0$
2: **while** $|f'(x_n)| \geq t$ **and** $n \leq n_{\max}$ **do**
3: $\quad x_{n+1} \leftarrow x_n - \eta f'(x_n)$
4: $\quad n \leftarrow n + 1$
5: **end while**
6: **return** x_{n+1}

In this very basic form the gradient descent algorithm is quite simple. The loop on lines 2–5 is executed until the derivative becomes smaller than some user-specified tolerance (e.g., $t = 10^{-6}$) or a maximum number of iterations is reached (which is to protect against infinite loops). Line 3 is the heart of the algorithm: at each iteration we move a step towards the direction opposite to the direction the derivative is pointing. For example, if $f(x_n) > 0$, the function is increasing at x_n, which means that we want to move to the left (i.e., take x_{n+1} smaller than x_n). The *learning rate* (suom. *oppimisnopeus*), or *step size*, η controls how large steps we take at each iteration. If η is too large, the algorithm may not work properly, "jumping around" too much. If the function has a local minimum, $f'(x_n)$ will hopefully get closer and closer to zero, so that the steps taken [i.e., $-\eta f'(x_n)$] become smaller and smaller.

Since the derivative is close to zero at the point returned by the algorithm (provided that tolerance t is small), this point is close to a critical point. Of course, a critical point is not necessarily a local extremal point, as we recall from Result 2.24 and Example 2.25. Figure 11 shows gradient descent in action.

Remark 2.28. Critical points and Result 2.23 are useful only when f is differentiable. Consider the absolute value function $f(x) = |x|$ As we saw in Remark 2.6, one cannot differentiate this function at $x = 0$. However, it is clear that f has a global minimum point $x^* = 0$.

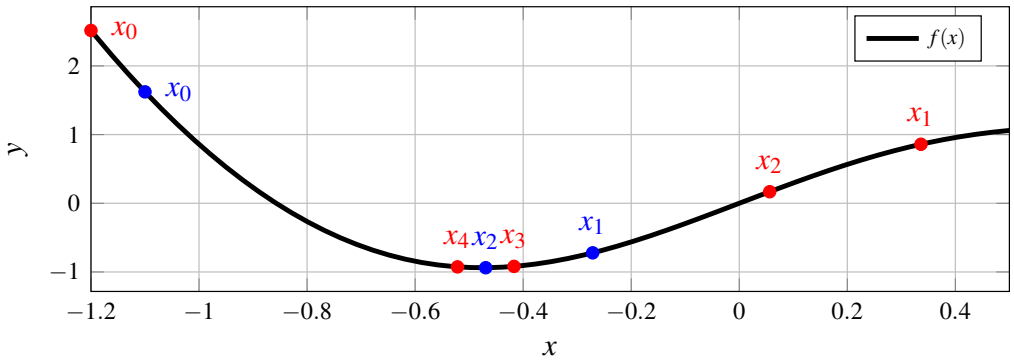

Figure 11: Gradient descent for the function $f(x) = x^4 + \sin(3x)$ with global minimum point $x^* \approx -0.47562$. The sequence of red points was produced by gradient descent with $x_0 = -1.2$ and $\eta = 0.16$; that of blue points with $x_0 = -1.1$ and $\eta = 0.1$. Observe that the point x_{n+1} is always to the right of x_n if f is decreasing at x_n and to the left of x_{n+1} if f is increasing at x_n (e.g., x_2 is to the left of x_1 since f is increasing at x_1).

2.6 Linearisation and Taylor Series

We can use the derivative at a point x_0 to form the linear approximation

$$f(x) \approx L(x) = f(x_0) + f'(x_0)(x - x_0) \tag{2.62}$$

to f around x_0. This approximation can be expected to be accurate only locally. That is, only for x close to x_0 [observe that the right-hand side reduces to $f(x_0)$ when $x = x_0$]. Finding the linear approximation L is called *linearisation* (suom. *linearisaatio*).

Example 2.29. We consider linearisation around $x_0 = 1$.

- Consider the function $f(x) = x^2$. Then $f(x_0) = 1^2 = 1$ and $f'(x_0) = 2 \cdot 1 = 2$ by (2.8). Therefore

$$L(x) = f(x_0) + f'(x_0)(x - x_0) = 1 + 2(x - 1) = 2x - 1. \tag{2.63}$$

- Consider the function $f(x) = \sin(3x)$. Then $f(x_0) = \sin(3 \cdot 1) \approx 0.1411$ and $f'(x_0) = 3\cos(3 \cdot 1) \approx -2.970$ by (2.12) and the chain rule (2.35). Therefore

$$L(x) = f(x_0) + f'(x_0)(x - x_0) = \sin(3) + 3\cos(3)(x - 1). \tag{2.64}$$

- Consider the function $f(x) = \exp(-3x)$. Then $f(x_0) = \exp(-3 \cdot 1) \approx 0.04978$ and $f'(x_0) = -3\exp(-3 \cdot 1) \approx -0.1494$ by (2.11) and the chain rule (2.35). Therefore

$$L(x) = f(x_0) + f'(x_0)(x - x_0) = e^{-3} - 3e^{-3}(x - 1). \tag{2.65}$$

These three functions and their linearisations are displayed in Figure 12.

Taylor polynomials generalise linearisation. Let $n \geq k$. By iterating (2.8), we get

$$\frac{\mathrm{d}^k}{\mathrm{d}x^k} x^n = n \frac{\mathrm{d}^{k-1}}{\mathrm{d}x^{k-1}} x^{n-1} = \cdots = n(n-1)\cdots(n-k+1)x^{n-k} = \frac{n!}{(n-k)!} x^{n-k}, \tag{2.66}$$

where $n! = 1 \cdot 2 \cdots (n-1) \cdot n$ is the factorial (note that $0! = 1$). Here it may be helpful to recall Example 2.8. If $k > n$, the kth derivative of x^n is zero since the nth derivative of x^n is the constant

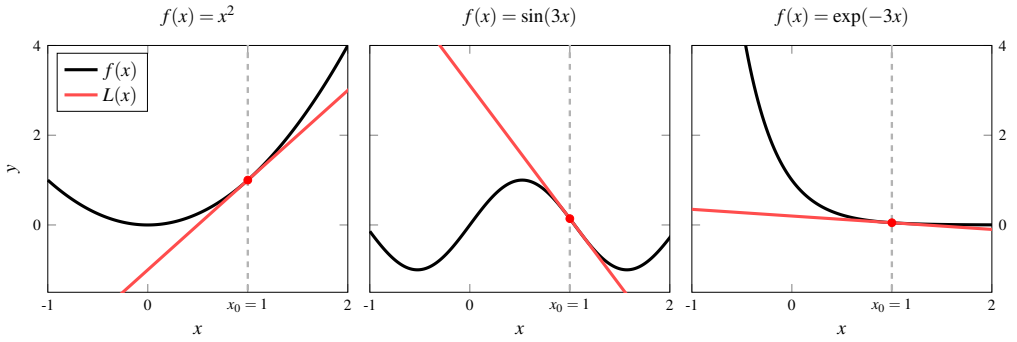

Figure 12: Three functions and their linearisations (2.62) around $x_0 = 1$. Observe how the linearisations are quite accurate [i.e., $L(x) \approx f(x)$] for x close to x_0 but approximate the function extremely badly further away from x_0.

$n!$ [set $k = n$ in (2.66)] and the derivative of a constant is zero. Under the notational convention $0^0 = 1$, we may write any monomial $P(x) = x^n$ as the series[4]

$$P(x) = \sum_{k=0}^{\infty} \frac{P^{(k)}(0)}{k!} x^k = \sum_{k=0}^{n} \frac{n!}{k!(n-k)!} (y^{n-k}\big|_{y=0}) x^k + \sum_{k=n+1}^{\infty} \frac{0}{k!} x^k = x^n, \tag{2.67}$$

where the last equation follows from the fact that $y^{n-k}|_{y=0} = 0$ if $k \neq n$. The expansion (2.67) extends to any polynomial of the form $P(x) = a_0 + a_1 x + \cdots + a_{n-1} x^{n-1} + a_n x^n$ by linearity of differentiation. In fact, this expansion extends to many non-polynomial functions as well.

Definition 2.30 (TAYLOR POLYNOMIAL AND SERIES). The *Taylor polynomial* (suom. *Taylorin polynomi*) of degree n of a function f around a point x_0 is the function T_n given by

$$T_n(x) = \sum_{k=0}^{n} \frac{f^{(k)}(x_0)}{k!} (x - x_0)^k. \tag{2.68}$$

The *Taylor series* (suom. *Taylorin sarja*) of a function f around a point x_0 is the Taylor polynomial of degree $n = \infty$:

$$T_\infty(x) = T_\infty(x) = \sum_{k=0}^{\infty} \frac{f^{(k)}(x_0)}{k!} (x - x_0)^k. \tag{2.69}$$

These definitions assume that f has sufficiently many derivatives. Taylor series around $x_0 = 0$ is sometimes called *Maclaurin series*.

Roughly speaking, an infinitely many times differentiable function is *analytic* (suom. *analyyttinen*) if it is equal to its Taylor series around some point x_0: $f(x) = (T_\infty f)(x)$ for all $x \in \mathbb{R}$.[5] The elementary function that one usually encounters are analytic.

[4] We do not attempt to provide a rigorous definition for the infinite series $\sum_{k=1}^{\infty} a_k$.

[5] The precise definition of an analytic function is somewhat more subtle: An infinitely differentiable function f is analytic on $\mathbb{R}$ if for every $x_0 \in \mathbb{R}$ there is $\varepsilon > 0$ (which may depend on x_0) such that $f(x)$ equals its the Taylor series $T_\infty(x)$ around x_0 for all x such that $|x - x_0| < \varepsilon$. The point is that the point x_0 around which the Taylor series is developed may depend on x if the Taylor series is to equal $f(x)$.

Example 2.31. Let $f(x) = e^x$ be the exponential function, which is analytic. From (2.11) we get $f^{(k)}(x) = e^x$ for every $k \geq 0$. Therefore we have the Maclaurin series (i.e., $x_0 = 0$)

$$e^x = \sum_{k=0}^{\infty} \frac{f^{(k)}(0)}{k!} x^k = \sum_{k=0}^{\infty} \frac{e^0}{k!} x^k = \sum_{k=0}^{\infty} \frac{1}{k!} x^k. \tag{2.70}$$

Example 2.32. Let $f(x) = \sin(x)$ be the sine function, which is analytic. Recall Example 2.9. From (2.12) and (2.13) we get $f^{(2k)}(x) = (-1)^k \sin(x)$ and $f^{(2k+1)}(x) = (-1)^k \cos(x)$. Because $\sin(0) = 0$ and $\cos(0) = 1$, we obtain the Maclaurin series

$$\sin(x) = \sum_{k=0}^{\infty} \frac{f^{(k)}(0)}{k!} x^k = \sum_{k=0}^{\infty} \frac{(-1)^k}{(2k+1)!} x^{2k+1}. \tag{2.71}$$

In a similar manner,

$$\cos(x) = \sum_{k=0}^{\infty} \frac{(-1)^k}{(2k)!} x^{2k}. \tag{2.72}$$

By setting $n = 1$ we obtain the Taylor polynomial

$$T_1(x) = \frac{f^{(0)}(x_0)}{0!}(x - x_0)^0 + \frac{f^{(1)}(x_0)}{1!}(x - x_0)^1 = f(x_0) + f'(x_0)(x - x_0), \tag{2.73}$$

which is precisely the linear approximation L_{x_0} in (2.62). A special case of a more general *Taylor's theorem* now allows us to bound linearisation error.

Result 2.33 (LINEARISATION ERROR). Suppose that f is a twice differentiable function whose second derivative f'' is continuous. Let L be the linear approximation in (2.62). Then

$$|f(x) - L(x)| \leq \frac{C}{2}(x - x_0)^2, \tag{2.74}$$

where C is the maximum of $|f''(y)|$ over all y between x_0 and x.

As long as the second derivative of f is well-defined and not too large (i.e., f is sufficiently "nice"), linearisation error scales (at most) as a square of the distance to the linearisation point x_0.

2.7 Integration

The integral

$$\int_a^b f(x)\,\mathrm{d}x \tag{2.75}$$

gives the area between points a and b and bounded by the graph of f and the x-axis. When f takes negative values, the corresponding area is subtracted.[6] Figure 13 provides a simple illustration. One can think of integral as a "continuous" version of summation. Let $x_k = a + (b-a)k/n$ and $\Delta x = x_{k+1} - x_k = (b-a)/n$. For "nice enough" functions it holds that

$$\sum_{k=1}^{n} f(x_k)\Delta x \approx \int_a^b f(x)\,\mathrm{d}x, \tag{2.76}$$

and the left-hand side converges to the integral as $n \to \infty$. Here we provide a very brief overview of integration. Integration is the reverse of differentiation.

[6] The rigorous definition (or definitions) of integral are outside the scope of this course. Equation (2.76) is the starting point to define the *Riemann integral*, while defining the more general and flexible *Lebesgue integral* requires the machinery of *measure theory* (suom. *mittateoria*).

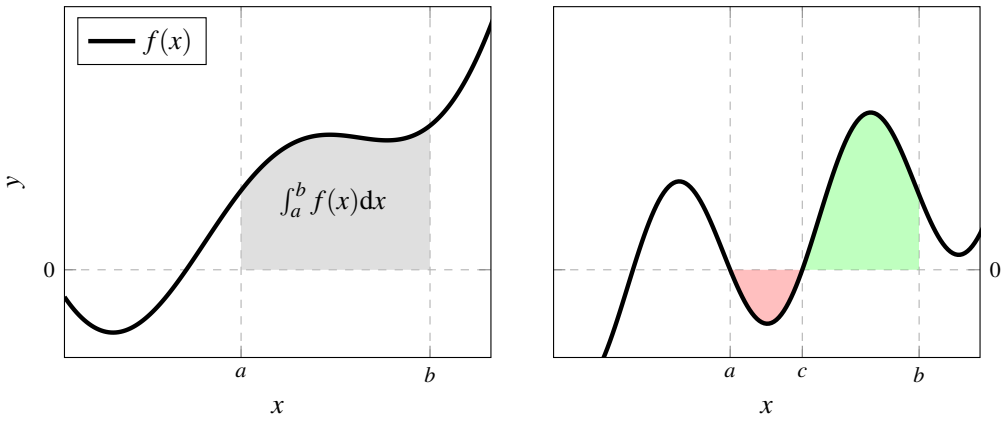

Figure 13: The integral of the function on the left from a to b is simply the area of the region (grey) between a and b bounded by the graph of f and the x-axis. When the function takes negative values on the domain of integration $[a,b]$, as in the right panel, the interal is obtained by subtracting the below x-axis (red) from the area above the axis (green).

Result 2.34 (FUNDAMENTAL THEOREM OF CALCULUS). Suppose that f is a continuous function and F a function such that $F'(x) = f(x)$. The *fundamental theorem of calculus* (suom. *analyysin peruslause*) states that

$$F(b) - F(a) = \int_a^b f(x)\,\mathrm{d}x. \tag{2.77}$$

Note that this result applies also to functions that change sign, such as the function in the second panel of Figure 13.

Result 2.35 (LINEARITY OF INTEGRATION). Let $f(x) = \alpha g(x) + \beta h(x)$ for functions g and h and constants $\alpha, \beta \in \mathbb{R}$. Then

$$\int_a^b f(x)\,\mathrm{d}x = \int_a^b [\alpha g(x) + \beta h(x)]\,\mathrm{d}x = \alpha \int_a^b g(x)\,\mathrm{d}x + \beta \int_a^b h(x)\,\mathrm{d}x. \tag{2.78}$$

The fundamental theorem of calculus provides a way to compute integrals that is conceptually straightforward: find a function F whose derivative equals f and evaluate this function at the end points a and b of integration. Unfortunately, to find this *antiderivative*, or *indefinite integral* (suom. *määräämätön integraali*), is much more difficult than differentiating a function. For example, it is easy to use the chain rule to compute that the derivative of $\exp(-\frac{1}{2}x^2)$ is $-x\exp(-\frac{1}{2}x^2)$, which we essentially did in Example 2.17. However, $\exp(-\frac{1}{2}x^2)$ admits *no* elementary antiderivative. That is, a function F such that $F'(x) = \exp(-\frac{1}{2}x^2)$ cannot be expressed in terms of the functions appearing in Result 2.4 using a finite number of arithmetic operations.

The following result collects antiderivatives of the elementary functions in Result 2.4. If F is an indefinite integral of f, so is $F(x) + C$ for any constant C because the derivative of a constant function is zero.

Result 2.36 (ANTIDERIVATIVES OF ELEMENTARY FUNCTIONS). Let F be a function such that $F'(x) = f(x)$. Then

$$f(x) = x^n \quad \Longrightarrow \quad F(x) = \frac{1}{n+1}x^{n+1} + C \quad [n \in \mathbb{R}, n \neq -1], \tag{2.79}$$
$$f(x) = x^{-1} \quad \Longrightarrow \quad F(x) = \log(x) + C, \tag{2.80}$$
$$f(x) = e^x \quad \Longrightarrow \quad F(x) = e^x + C, \tag{2.81}$$
$$f(x) = \sin(x) \quad \Longrightarrow \quad F(x) = -\cos(x) + C, \tag{2.82}$$
$$f(x) = \cos(x) \quad \Longrightarrow \quad F(x) = \sin(x) + C. \tag{2.83}$$

Example 2.37. Let us verify two of the formulae in Result 2.36. Equation (2.79) follows from (2.8) since

$$\frac{\mathrm{d}}{\mathrm{d}x}F(x) = \frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{1}{n+1}x^{n+1} + C\right) = \frac{1}{n+1} \cdot \frac{\mathrm{d}}{\mathrm{d}x}x^{n+1} = \frac{1}{n+1}(n+1)x^n = x^n. \tag{2.84}$$

Equation (2.82) follows from (2.83) since

$$\frac{\mathrm{d}}{\mathrm{d}x}F(x) = \frac{\mathrm{d}}{\mathrm{d}x}[-\cos(x) + C] = -\frac{\mathrm{d}}{\mathrm{d}x}\cos(x) = -(-\sin(x)) = \sin(x). \tag{2.85}$$

Example 2.38. We can now use Results 2.34 and 2.36 to compute integrals. We may always ignore the constant C in Result 2.36 because it gets cancelled when computing $F(b) - F(a)$ in Result 2.34. For example,

$$\int_a^b x^n \,\mathrm{d}x = \frac{1}{n+1}b^{n+1} - \frac{1}{n+1}a^{n+1} = \frac{1}{n+1}(b^{n+1} - a^{n+1}) \tag{2.86}$$

for any $a < b$ by (2.79) and

$$\int_0^{2\pi} \sin(x) \,\mathrm{d}x = -\cos(2\pi) - (-\cos(0)) = -1 - (-1)) = 0 \tag{2.87}$$

by (2.82) and $\cos(2\pi) = 1$. To understand why the latter integral is zero you may want to examine the graph of $\sin(x)$ and recall that areas below the x-axis are to be subtracted when computing the integral (see also the right panel in Figure 13).

Integration by substitution or *change of variables* (suom. *sijoitusmenetelmä* tai *muuttujanvaihto*) is a powerful technique to compute integrals. Integration by substitution is very useful in probability theory (though during this course we will not get that far enough in probability to witness this).

Result 2.39 (INTEGRATION BY SUBSTITUTION). Let f be continuous and g a differentiable function such that g' is continuous. Then

$$\int_{g(a)}^{g(b)} f(x) \,\mathrm{d}x = \int_a^b f(g(y))g'(y) \,\mathrm{d}y. \tag{2.88}$$

Proof. Let F be an antiderivative of f. The chain rule (2.35) then gives

$$(F \circ g)'(y) = F'(g(y))g'(y) = f(g(y))g'(y). \tag{2.89}$$

Therefore Result 2.34 yields

$$\int_a^b f(g(y))g'(y)\,\mathrm{d}y = (F\circ g)(b) - (F\circ g)(a) = F(g(b)) - F(g(a)), \tag{2.90}$$

which, by proceeding from left to right in (2.77), is equal to the integral

$$\int_{g(a)}^{g(b)} f(x)\,\mathrm{d}x \tag{2.91}$$

because $F'(x) = f(x)$. □

It is not particularly intuitive to use Result 2.39 to perform integration by substitution. The following technique based on Leibniz's notation

$$f'(x) = \frac{\mathrm{d}f}{\mathrm{d}x}(x) \tag{2.92}$$

for differentiation is usually easier to use. Suppose that a function f we want to integrate can be written as $f(x) = h(g(x))g'(x)$ for some functions g and h. Then

$$\int_a^b f(x)\,\mathrm{d}x = \int_a^b h(g(x))g'(x)\,\mathrm{d}x. \tag{2.93}$$

We employ the substitution $u = g(x)$. Following (2.92), we write

$$g'(x) = \frac{\mathrm{d}u}{\mathrm{d}x} \tag{2.94}$$

and subsequently engage in notational abuse by rearranging this equation as

$$\mathrm{d}x = \frac{1}{g'(x)}\,\mathrm{d}u. \tag{2.95}$$

We may then formally substitute $g'(x)^{-1}\,\mathrm{d}u$ for $\mathrm{d}x$ in (2.93). However, we also need to take care to adjust the limits of integration. Since $x = a$ and $x = b$ at the original limits, the substitution $u = g(x)$ results in new integration limits $g(a)$ and $g(b)$. In this way we get

$$\int_a^b f(x)\,\mathrm{d}x = \int_a^b h(g(x))g'(x)\,\mathrm{d}x = \int_{g(a)}^{g(b)} h(u)g'(x)\cdot\frac{1}{g'(x)}\,\mathrm{d}u = \int_{g(a)}^{g(b)} h(u)\,\mathrm{d}u. \tag{2.96}$$

If h is, for example, one of the elementary functions with closed-form antiderivatives in Result 2.36, we can use (2.77) to compute the integral. The following two examples illustrate integration by substitution.

Example 2.40. Suppose that we want to compute

$$\int_1^3 f(x)\,\mathrm{d}x = \int_1^3 x\exp(x^2)\,\mathrm{d}x. \tag{2.97}$$

Make the substitution $u = x^2$, so that

$$\frac{\mathrm{d}u}{\mathrm{d}x} = 2x. \tag{2.98}$$

This gives $\mathrm{d}u = 2x\,\mathrm{d}x$. Because

$$\int_1^3 f(x)\,\mathrm{d}x = \int_1^3 x\exp(x^2)\,\mathrm{d}x = \frac{1}{2}\int_1^3 2x\exp(x^2)\,\mathrm{d}x = \frac{1}{2}\int_1^3 \exp(x^2)\cdot 2x\,\mathrm{d}x, \tag{2.99}$$

we can substitute $u = x^2$ and $\mathrm{d}u = 2x\,\mathrm{d}x$ into this integral. The integration limits $x = 1$ and $x = 3$ become $u = 1^2 = 1$ and $u = 3^2 = 9$ due to the substitution $u = x^2$. This gives

$$\int_1^3 f(x)\,\mathrm{d}x = \frac{1}{2}\int_1^3 \exp(x^2)\cdot 2x\,\mathrm{d}x = \frac{1}{2}\int_1^9 e^u\,\mathrm{d}u. \tag{2.100}$$

By (2.81) and (2.77),

$$\frac{1}{2}\int_1^9 e^u\,\mathrm{d}u = \frac{1}{2}(e^9 - e^1) = \frac{e}{2}(e^8 - 1). \tag{2.101}$$

Thus we have

$$\int_1^3 x\exp(x^2)\,\mathrm{d}x = \frac{e}{2}(e^8 - 1). \tag{2.102}$$

Example 2.41. Suppose that we want to compute

$$\int_{-2}^0 f(x)\,\mathrm{d}x = \int_{-2}^0 \left(5 + \frac{x}{3}\right)^5 \mathrm{d}x. \tag{2.103}$$

In principle, we could expand $(5 + x/3)^5$ and then apply (2.79). But this is tedious and integration by substitution provides an easier way. Make the substitution $u = 5 + x/3$, so that

$$\frac{\mathrm{d}u}{\mathrm{d}x} = \frac{1}{3}. \tag{2.104}$$

This gives $\mathrm{d}x = 3\,\mathrm{d}u$, which we may substitute for $\mathrm{d}x$ in (2.103). The integration limits $x = -2$ and $x = 0$ become $u = 5 + (-2)/3 = \frac{13}{3}$ and $u = 5 + 0/3 = 5$ due to the substitution $u = 5 + x/3$. Thus

$$\begin{aligned}\int_{-2}^0 f(x)\,\mathrm{d}x = \int_{-2}^0 \left(5 + \frac{x}{3}\right)^5 \mathrm{d}x = \int_{\frac{13}{3}}^5 3u^5\,\mathrm{d}u = \frac{3}{5+1}\left(5^{5+1} - \left(\frac{13}{3}\right)^{5+1}\right) \\ = \frac{3\,281\,908}{729}\end{aligned} \tag{2.105}$$

by (2.79) and (2.77) [also recall (2.86)].

3 Vector Calculus

We now move to higher dimensions. In this section we consider *multivariate* functions

$$f\colon \mathbb{R}^d \to \mathbb{R} \quad \text{and} \quad \mathbf{f}\colon \mathbb{R}^d \to \mathbb{R}^m. \tag{3.1}$$

The former maps a d-dimensional vector $\mathbf{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d$ to a real $f(x)$, while the latter maps $\mathbf{x}$ to the m-vector

$$\mathbf{f}(\mathbf{x}) = \begin{bmatrix} f_1(\mathbf{x}) \\ \vdots \\ f_m(\mathbf{x}) \end{bmatrix} \in \mathbb{R}^m, \tag{3.2}$$

where the m functions $f_1, \ldots, f_m\colon \mathbb{R}^d \to \mathbb{R}$ are the *components* of $\mathbf{f}$. To us all vectors are *column vectors*, which means that in computations they are to be treated as $d \times 1$ matrices (written $\mathbb{R}^{d\times 1}$). The transpose $\mathbf{x}^\mathsf{T}$ of a column vector $\mathbf{x} \in \mathbb{R}^d = \mathbb{R}^{d\times 1}$ is an a row vector in $\mathbb{R}^{1\times d}$. Here and elsewhere in these notes

$$\|\mathbf{x}\| = \sqrt{x_1^2 + \cdots + x_d^2} \tag{3.3}$$

is the Euclidean norm of $\mathbf{x} \in \mathbb{R}^d$ and

$$\mathbf{x} \cdot \mathbf{y} = \mathbf{x}^\mathsf{T}\mathbf{y} = x_1 y_1 + \cdots + x_d y_d = \sum_{i=1}^{d} x_i y_i \tag{3.4}$$

the inner product of $\mathbf{x} \in \mathbb{R}^d$ and $\mathbf{y} \in \mathbb{R}^d$. Recall that $\mathbf{x} \cdot \mathbf{x} = \mathbf{x}^\mathsf{T}\mathbf{x} = \|\mathbf{x}\|^2$.

3.1 Partial Derivatives

Partial derivative of a multivariate function is its derivative along a particular coordinate axis.

Definition 3.1 (Partial derivative). Let $f\colon \mathbb{R}^d \to \mathbb{R}$ be a function. Its *partial derivative* (suom. *osittaisderivaatta*) $\partial_i f$ with respect to the ith variable, x_i, is its derivative with respect to this variable when the other variables are kept fixed:

$$\partial_i f(\mathbf{x}) = \frac{\partial f}{\partial x_i}(\mathbf{x}) = \frac{\partial}{\partial x_i} f(\mathbf{x}) = \lim_{h\to 0} \frac{f(x_1, \ldots, x_i + h, x_{i+1}, \ldots, x_d) - f(x_1, \ldots, x_d)}{h}. \tag{3.5}$$

The partial derivative with respect to x_i is also denoted $\partial_{x_i} f(\mathbf{x})$ (and in various other ways).

The symbol ∂ has many different names, one of them being simply "partial". The partial derivative with respect to x_i gives the rate of change of a multivariate function along the ith coordinate axis; see Figure 14. To compute the partial derivative $\partial_i f$, one considers all other coordinates except the ith one constants and applies the differentiation rules from Section 2 with $x = x_i$. For example, suppose that f is a function of two variables, x and y. The partial derivative $\partial_x f(x,y)$ with respect to x is the derivative of the function $g_y(x) = f(x,y)$, where y acts as a fixed constant:

$$\partial_x f(x,y) = \frac{\partial f(x,y)}{\partial x} = \frac{\mathrm{d} g_y(x)}{\mathrm{d}x} = g_y'(x). \tag{3.6}$$

The next two examples illustrate how this works.

Example 3.2. Let $f(\mathbf{x}) = f(x_1, x_2) = x_1 + x_2 - x_1 x_2^3$ (i.e., $d = 2$). To compute $\partial f_1(x_1, x_2)$, we consider x_2 a constant and apply the differentiation rules from Section 2 with $x = x_1$:

$$\partial_1 f(x_1, x_2) = \frac{\partial}{\partial x_1} x_1 + \frac{\partial}{\partial x_1} x_2 - \frac{\partial}{\partial x_1} x_1 x_2^3 = 1 + 0 - x_2^3 \cdot \frac{\partial}{\partial x_1} x_1 = 1 - x_2^3. \tag{3.7}$$

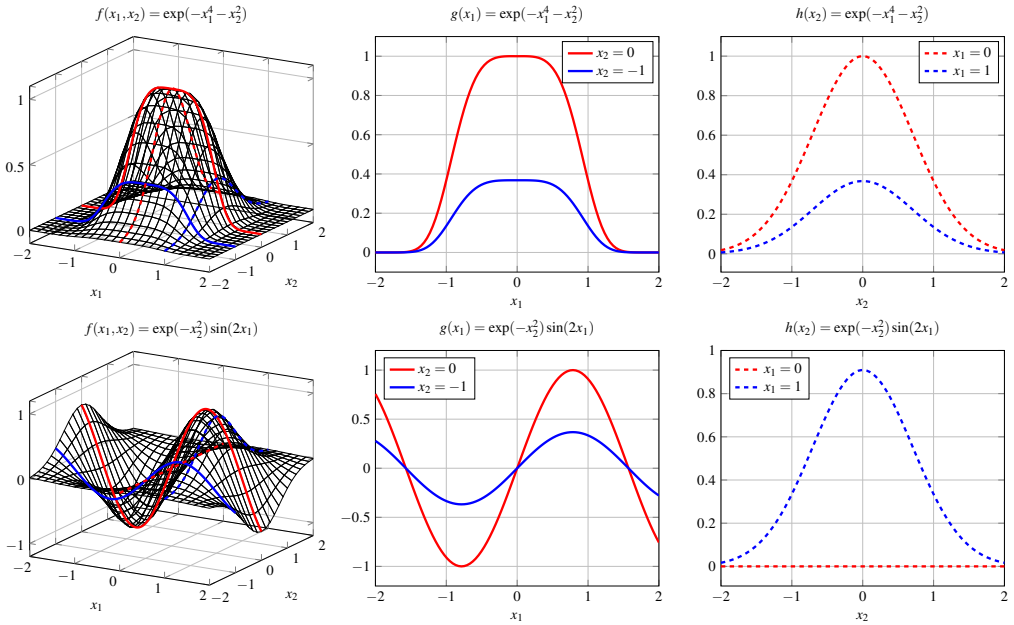

Figure 14: The ith partial derivative is the derivative along the ith coordinate axis. On left we have two functions $f\colon \mathbb{R}^2 \to \mathbb{R}$. On the middle and right we plot f along the two coordinate axes when the other coordinate is fixed. That is, on the middle we have the function $g(x_1) = f(x_1,x_2)$ for fixed $x_2 = 0$ and $x_2 = -1$ and on the right the function $h(x_2) = f(x_1,x_2)$ for fixed $x_1 = 0$ and $x_1 = 1$. The partial derivatives $\partial_1 f(x_1,x_2)$ and $\partial_2 f(x_1,x_2)$ are the derivatives of these two functions: $\partial_1 f(x_1,x_2) = g'(x_1)$ and $\partial_2 f(x_1,x_2) = h'(x_2)$. For example, $\partial_1 f(x_1,-1)$ equals the derivative of $g(x_1)$ with $x_2 = -1$ (solid blue) while $\partial_2 f(0,x_2)$ equals the derivative of $h(x_2)$ with $x_1 = 0$ (dashed red).

The first term, $\partial x_1/\partial x_1$, is computed using (2.8) with $n = 1$ and $a = 1$. Second, $\partial x_2/\partial x_1 = 0$ by (2.9) because x_2 is constant with respect to x_1 [i.e., $f'(x_1) = 0$ if $f(x_1) = x_2$ for all $x_1 \in \mathbb{R}$]. To compute the third term, $\partial x_1 x_2^3/\partial x_1$, we use (2.8) with $n = 1$ and $a = x_2^3$. The partial derivative with respect to x_2 is obtained in a similar manner:

$$\partial_2 f(x_1,x_2) = \frac{\partial}{\partial x_2} x_1 + \frac{\partial}{\partial x_2} x_2 - \frac{\partial}{\partial x_2} x_1 x_2^3 = 0 + 1 - x_1 \cdot \frac{\partial}{\partial x_2} x_2^3 = 1 - 3x_1 x_2^2. \tag{3.8}$$

Example 3.3. Let

$$f(\mathbf{x}) = f(x_1,x_2,x_3) = \exp(-\tfrac{1}{2}\|\mathbf{x}\|^2) = \exp\left(-\frac{x_1^2 + x_2^2 + x_3^2}{2}\right), \tag{3.9}$$

be the *Gaussian function* in dimension $d = 3$. Again, to compute $\partial_1 f(x_1,x_2,x_3)$, we consider x_2 and x_3 constants and apply the differentiation rules from from Section 2 with $x = x_1$.

Using the chain rule with $h(x) = \exp(x)$ and $g(x) = -\frac{1}{2}(x^2 + x_2^2 + x_3^2)$ gives

$$\begin{aligned} \partial_1 f(x_1,x_2,x_3) = \frac{\partial}{\partial x_1} \exp\left(-\frac{x_1^2+x_2^2+x_3^2}{2}\right) &= h'(g(x_1)) \cdot g'(x_1) \\ &= \exp\left(-\frac{x_1^2+x_2^2+x_3^2}{2}\right) \cdot (-x_1) \\ &= -x_1 \exp(-\tfrac{1}{2}\|\mathbf{x}\|^2). \end{aligned} \tag{3.10}$$

Similarly, $\partial_2 f(x_1,x_2,x_3) = -x_2 \exp(-\frac{1}{2}\|\mathbf{x}\|^2)$ and $\partial_3 f(x_1,x_2,x_3) = -x_3 \exp(-\frac{1}{2}\|\mathbf{x}\|^2)$.

In applications one often takes partial derivatives *with respect to parameters*. There is no fundamental difference between "parameters" and "variables": one can always think of parameters as variables and consider functions of parameters. For example, consider the quadratic function $f(x) = ax^2$ with parameter $a \in \mathbb{R}$. If we take a fixed $x = x_0$, we can consider the function $g(a) = g_{x_0}(a) = ax_0^2$ of the parameter a and compute its derivative

$$g'(a) = \frac{\mathrm{d}g(a)}{\mathrm{d}a} = x_0^2. \tag{3.11}$$

The function g describes how changing the parameter a affects the value of the function f at $x = x_0$.

Example 3.4. Let us momentarily return to the ordinary least squares example in Section 1.1. Let

$$L(a,b) = \sum_{i=1}^{n} (y_i - a - bx_i)^2 \tag{3.12}$$

be the loss function in (1.2). The loss function is a function from $\mathbb{R}^2$ to $\mathbb{R}$ whose two variables are the linear regression parameters a and b, while the inputs x_i and the outputs y_i are simply some fixed constants. We may thus compute the partial derivatives of L with respect to a and b. We obtain

$$\partial_a L(a,b) = \frac{\partial}{\partial a} L(a,b) = \sum_{i=1}^{n} \frac{\partial}{\partial a} (y_i - a - bx_i)^2 = -2 \sum_{i=1}^{n} (y_i - a - bx_i), \tag{3.13}$$

where the minus appears because the derivative of $-a$ with respect to a is -1, and

$$\partial_b L(a,b) = \frac{\partial}{\partial b} L(a,b) = \sum_{i=1}^{n} \frac{\partial}{\partial b} (y_i - a - bx_i)^2 = -2 \sum_{i=1}^{n} x_i (y_i - a - bx_i). \tag{3.14}$$

These two partial derivatives tell us how the value of the loss function changes for a fixed set of data.

Example 3.5. Let us momentarily return to the neural network example in Section 1.1. Let

$$L(w_1,w_2,b_1,b_2) = \sum_{i=1}^{n} (y_i - F(x_i \mid w_1,w_2,b_1,b_2))^2 \tag{3.15}$$

be the loss function in (1.7). Recall that here

$$F(x \mid w_1,w_2,b_1,b_2) = \sigma(w_2 \sigma(w_1 x + b_1) + b_2) \quad \text{and} \quad \sigma(x) = \frac{1}{1+e^{-x}}. \tag{3.16}$$

The loss function L is function from $\mathbb{R}^4$ to $\mathbb{R}$ whose four variables are the neural network parameters w_1, w_2, b_1, and b_2. We may thus compute its partial derivatives with respect to each of these parameters. For example, taking partial derivative with respect to w_2 yields

$$\begin{aligned}\partial_{w_2} L(w_1,w_2,b_1,b_2) &= \frac{\partial}{\partial w_2} L(w_1,w_2,b_1,b_2) \\ &= \sum_{i=1}^{n} 2(y_i - F(x_i \mid w_1,w_2,b_1,b_2)) \cdot \frac{\partial}{\partial w_2}(y_i - F(x_i \mid w_1,w_2,b_1,b_2)) \\ &= -2\sum_{i=1}^{n}(y_i - F(x_i \mid w_1,w_2,b_1,b_2)) \cdot \partial_{w_2} F(x_i \mid w_1,w_2,b_1,b_2)).\end{aligned}$$

To compute the partial derivative of $F(x_i \mid w_1,w_2,b_1,b_2)$ we treat x_1, w_1, b_1, and b_2 as constants and compute the derivative of the function

$$f(x) = F(x_i \mid w_1,x,b_1,b_2) = \sigma(x\sigma(w_1x_i+b_1)+b_2) \tag{3.17}$$

at $x = w_2$. Since $\sigma(w_1x_i+b_1)$ is constant with respect to x, the derivative is [this equation is essentially Example 2.18 with $a = \sigma(w_1x_i+b_1)$]

$$f'(x) = \sigma'(x\sigma(w_1x_i+b_1)+b_2)\sigma(w_1x_i+b_1). \tag{3.18}$$

We then obtain $\partial_{w_2} F(x_i \mid w_1,w_2,b_1,b_2)) = \sigma'(x\sigma(w_1x_i+b_1)+b_2)\sigma(w_1x_i+b_1)$ by setting $x = w_2$ in (3.18). We could now compute the derivative of the function σ in (3.16) and insert the resulting equation for $\partial_{w_2} F(x_i \mid w_1,w_2,b_1,b_2))$ in the expression for $\partial_{w_2} L(w_1,w_2,b_1,b_2)$ that we derived above.

3.2 Gradient and Jacobian

The gradient is a vector of partial derivatives.

Definition 3.6 (Gradient). The *gradient* (suom. *gradientti*) of a function $f\colon \mathbb{R}^d \to \mathbb{R}$ is the function $\nabla f\colon \mathbb{R}^d \to \mathbb{R}^d$ given by

$$\nabla f(\mathbf{x}) = \nabla f(x_1,\ldots,x_d) = \nabla_{\mathbf{x}} f(\mathbf{x}) = \frac{\mathrm{d}}{\mathrm{d}\mathbf{x}} f(\mathbf{x}) = \begin{bmatrix}\partial_1 f(\mathbf{x}) \\ \vdots \\ \partial_d f(\mathbf{x})\end{bmatrix} = \begin{bmatrix}\frac{\partial}{\partial x_1} f(\mathbf{x}) \\ \vdots \\ \frac{\partial}{\partial x_d} f(\mathbf{x})\end{bmatrix} \in \mathbb{R}^d. \tag{3.19}$$

The upside-down Δ (delta) in the notation for the gradient in (3.19) is called *nabla*. Here gradient is a column vector, a convention far from universal. Here notations such as $\mathrm{d}f(\mathbf{x})/\mathrm{d}\mathbf{x}$ in which one differentiates with respect to a vector are to be interpreted in a coordinate-wise sense: one forms a vector by differentiating $f(\mathbf{x})$ with respect to each coordinate of $\mathbf{x}$. We consider a function $f\colon \mathbb{R}^d \to \mathbb{R}$ differentiable if all its partial derivative exist, meaning that the gradient exists as well.

Example 3.7. Consider the function $f(\mathbf{x}) = f(x_1,x_2) = x_1 + x_2 - x_1x_2^3$ from Example 3.2. In (3.7) and (3.8) we computed that the partial derivatives of this function are

$$\partial_1 f(x_1,x_2) = 1 - x_2^3 \quad \text{and} \quad \partial_2 f(x_1,x_2) = 1 - 3x_1x_2^2. \tag{3.20}$$

Therefore the gradient is given by

$$\nabla f(x_1,x_2) = \begin{bmatrix} \partial_1 f(x_1,x_2) \\ \partial_2 f(x_1,x_2) \end{bmatrix} = \begin{bmatrix} 1-x_2^3 \\ 1-3x_1x_2^2 \end{bmatrix}. \tag{3.21}$$

Example 3.8. Consider the Gaussian function

$$f(\mathbf{x}) = f(x_1,x_2,x_3) = \exp(-\tfrac{1}{2}\|\mathbf{x}\|^2) = \exp\left(-\frac{x_1^2+x_2^2+x_3^2}{2}\right) \tag{3.22}$$

from Example 3.3. In that example we computed that this function has the partial derivatives

$$\partial_i f(x_1,x_2,x_3) = -x_i \exp(-\tfrac{1}{2}\|\mathbf{x}\|^2) \quad \text{for} \quad i = 1,2,3. \tag{3.23}$$

Consequently,

$$\nabla f(x_1,x_2,x_3) = \begin{bmatrix} \partial_1 f(x_1,x_2,x_3) \\ \partial_2 f(x_1,x_2,x_3) \\ \partial_3 f(x_1,x_2,x_3) \end{bmatrix} = \begin{bmatrix} -x_1 \exp(-\frac{1}{2}\|\mathbf{x}\|^2) \\ -x_2 \exp(-\frac{1}{2}\|\mathbf{x}\|^2) \\ -x_3 \exp(-\frac{1}{2}\|\mathbf{x}\|^2) \end{bmatrix} = -\exp(-\tfrac{1}{2}\|\mathbf{x}\|^2)\,\mathbf{x}. \tag{3.24}$$

Note that in the last expression we multiply the vector $\mathbf{x} \in \mathbb{R}^3$ by the scalar $-\exp(-\frac{1}{2}\|\mathbf{x}\|^2)$.

Example 3.9. It is often convenient to use the operations of linear algebra to express multivariate functions. Let

$$\mathbf{A} = \begin{bmatrix} 1 & -1 & 1 \\ 1 & 0 & 2 \\ 0 & 3 & 7 \end{bmatrix} \in \mathbb{R}^{3\times 3} \quad \text{and} \quad \mathbf{a} = \begin{bmatrix} 8 \\ 1 \\ -3 \end{bmatrix} \in \mathbb{R}^3. \tag{3.25}$$

Consider the function $f\colon \mathbb{R}^3 \to \mathbb{R}$ given by

$$f(\mathbf{x}) = \mathbf{a}^\mathsf{T}\mathbf{A}\mathbf{x}. \tag{3.26}$$

We have

$$\mathbf{a}^\mathsf{T}\mathbf{A} = \begin{bmatrix} 8 & 1 & -3 \end{bmatrix} \begin{bmatrix} 1 & -1 & 1 \\ 1 & 0 & 2 \\ 0 & 3 & 7 \end{bmatrix} = \begin{bmatrix} 9 & -17 & -11 \end{bmatrix}, \tag{3.27}$$

so that

$$f(\mathbf{x}) = \begin{bmatrix} 9 & -17 & -11 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = 9x_1 - 17x_2 - 11x_3. \tag{3.28}$$

The partial derivatives of f are

$$\partial_1 f(\mathbf{x}) = 9, \quad \partial_2 f(\mathbf{x}) = -17, \quad \text{and} \quad \partial_3 f(\mathbf{x}) = -11. \tag{3.29}$$

We shall return to functions that have been defined using linear algebra in Section 3.3.

Recall from Section 2.1 that the derivative gives the direction (if positive, the function increases; if negative, the function decreases) and the rate of change of a function. The gradient

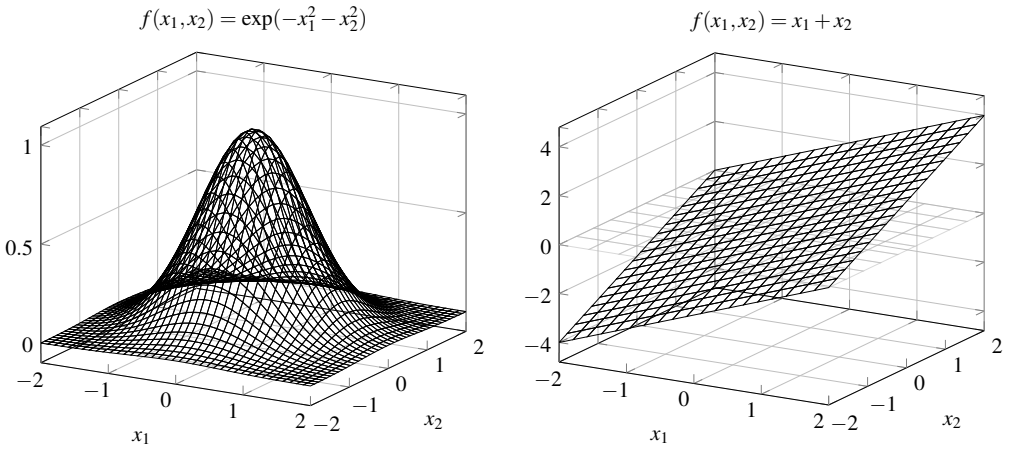

Figure 15: The two functions from $\mathbb{R}^2$ to $\mathbb{R}$ used in Examples 3.10 and 3.11.

has a similar property. Namely,

the gradient $\nabla f(\mathbf{x})$ gives the direction and the rate of fastest increase of f at $\mathbf{x}$.

The following example and Figure 16 illustrate what this means.

Example 3.10. Let us consider the Gaussian function

$$f(\mathbf{x}) = f(x_1,x_2) = \exp(-\tfrac{1}{2}\|\mathbf{x}\|^2) = \exp\left(-\frac{x_1^2+x_2^2}{2}\right) \tag{3.30}$$

of two variables; see Figure 15. We have used the three-dimensional version in Examples 3.3 and 3.8 but two-dimensional functions are somewhat easier to plot than three-dimensional ones. Exactly the same computation as that in Example 3.8 provides us the gradient

$$\nabla f(\mathbf{x}) = \nabla f(x_1,x_2) = -\exp(-\tfrac{1}{2}\|\mathbf{x}\|^2)\,\mathbf{x}. \tag{3.31}$$

The gradient at $\mathbf{x}$ is thus a vector that points towards the origin. Now, observe that the function f depends only on the norm $r = \|\mathbf{x}\|$ (i.e., the distance of $\mathbf{x}$ from the origin). That is, for any vectors $\mathbf{x},\mathbf{z} \in \mathbb{R}^2$ we have $f(\mathbf{x}) = f(\mathbf{z})$ if $\|\mathbf{x}\| = \|\mathbf{z}\|$. Moreover, f is a decreasing function of $r = \|\mathbf{x}\|$. Therefore the direction in which f increases the fastest is the direction in which $\|\mathbf{x}\|$ decreases the fastest. The fastest decrease in $\|\mathbf{x}\|$ is obtained by moving directly towards the origin, which is precisely what the the gradient tells us.

Example 3.11. Consider the function $f(x_1,x_2) = x_1 + x_2$, which defines a plane depicted in Figure 15. The gradient is

$$\nabla f(x_1,x_2) = \begin{bmatrix} \partial_1 f(x_1,x_2) \\ \partial_2 f(x_1,x_2) \end{bmatrix} = \begin{bmatrix} \frac{\partial}{\partial x_1}x_1 + \frac{\partial}{\partial x_1}x_2 \\ \frac{\partial}{\partial x_2}x_1 + \frac{\partial}{\partial x_2}x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}. \tag{3.32}$$

Therefore the fastest increase is obtained by moving diagonally with respect to the two coordinate axes.

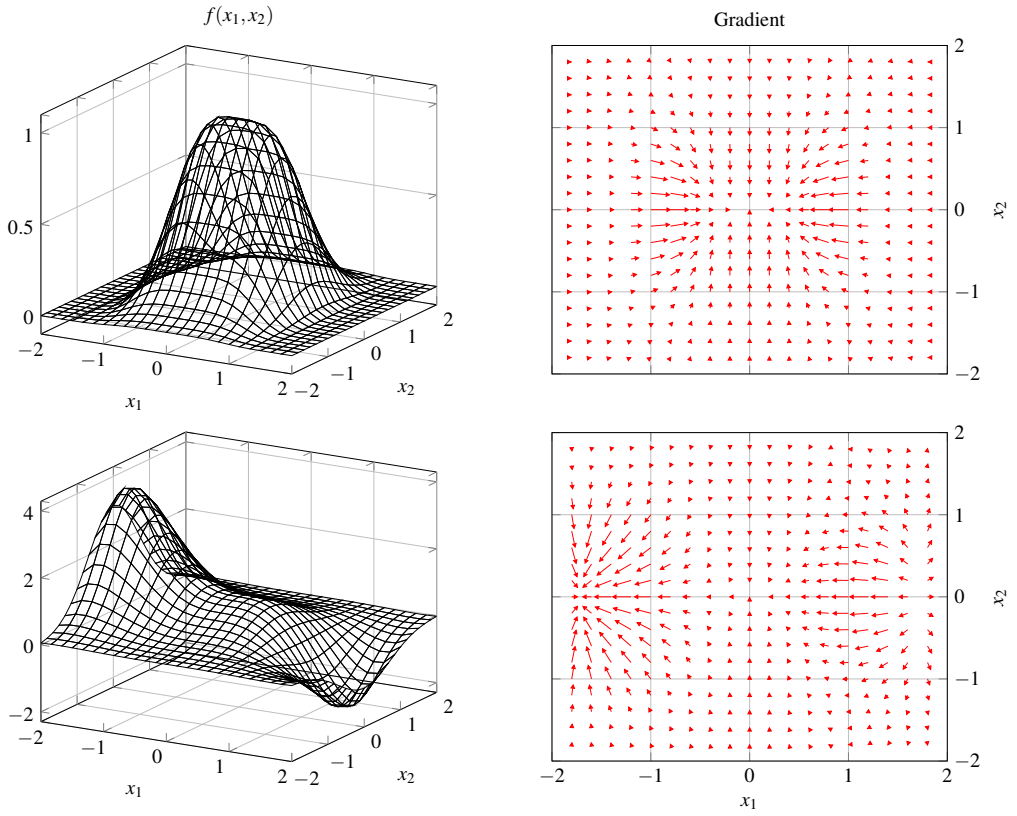

Figure 16: On the left we plot two functions $f\colon \mathbb{R}^2 \to \mathbb{R}$ and on the right their gradients at a number of points are represented as arrows. The direction of each arrow indicates the direction of the gradient at that point. The gradient is a vector that points towards the direction of the fastest increase of the function.

The Jacobian of a vector-valued function $\mathbf{f}$ is the matrix obtained by stacking gradients of the component of $\mathbf{f}$.

Definition 3.12 (JACOBIAN). The *Jacobian* (suom. *Jacobin matriisi*) of a function

$$\mathbf{f} = \begin{bmatrix} f_1 \\ \vdots \\ f_m \end{bmatrix} : \mathbb{R}^d \to \mathbb{R}^m. \tag{3.33}$$

at $\mathbf{x} \in \mathbb{R}^d$ is the $m \times d$ matrix

$$\mathbf{J}_\mathbf{f}(\mathbf{x}) = \begin{bmatrix} \dfrac{\partial \mathbf{f}}{\partial x_1}(\mathbf{x}) & \cdots & \dfrac{\partial \mathbf{f}}{\partial x_d}(\mathbf{x}) \end{bmatrix} = \begin{bmatrix} \nabla f_1(\mathbf{x})^\mathsf{T} \\ \vdots \\ \nabla f_m(\mathbf{x})^\mathsf{T} \end{bmatrix} = \begin{bmatrix} \partial_1 f_1(\mathbf{x}) & \cdots & \partial_d f_1(\mathbf{x}) \\ \vdots & \ddots & \vdots \\ \partial_1 f_m(\mathbf{x}) & \cdots & \partial_d f_m(\mathbf{x}) \end{bmatrix}. \tag{3.34}$$

That is, $[\mathbf{J}_\mathbf{f}(\mathbf{x})]_{i,j} = \partial_j f_i(\mathbf{x})$.

Definition 3.13 (JACOBIAN DETERMINANT). Let $\mathbf{f}\colon \mathbb{R}^d \to \mathbb{R}^d$ be a function as in (3.33) with $m = d$. The determinant

$$\det(\mathbf{J}_\mathbf{f}(\mathbf{x})) = |\mathbf{J}_\mathbf{f}(\mathbf{x})| = \begin{vmatrix} \partial_1 f_1(\mathbf{x}) & \cdots & \partial_d f_1(\mathbf{x}) \\ \vdots & \ddots & \vdots \\ \partial_1 f_m(\mathbf{x}) & \cdots & \partial_d f_m(\mathbf{x}) \end{vmatrix} \tag{3.35}$$

is called the *Jacobian determinant* (suom. *Jacobin determinantti*).

As determinants are defined only for square matrices, we must require that the output dimension m be equal the input dimension d in the definition of Jacobian determinant. The Jacobian determinant will play an important role in multivariate integration by substitution that we shall discuss in Section 3.6. Recall that determinant is a scalar. Also recall that the determinant of a 2×2 matrix has the simple expression

$$\det\left(\begin{bmatrix} a & b \\ c & d \end{bmatrix}\right) = \begin{vmatrix} a & b \\ c & d \end{vmatrix} = ad - bc \tag{3.36}$$

and that the determinant of a diagonal matrix equals the product of the diagonal values:

$$\begin{vmatrix} a_1 & 0 & \cdots & 0 \\ 0 & a_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & a_d \end{vmatrix} = a_1 \cdots a_d = \prod_{i=1}^{d} a_i. \tag{3.37}$$

Example 3.14. Consider the function $\mathbf{f}\colon \mathbb{R}^2 \to \mathbb{R}^2$ given by

$$\mathbf{f}(\mathbf{x}) = \mathbf{f}(x_1, x_2) = \begin{bmatrix} x_1 \sin(x_2) \\ x_1^2 + x_2^2 \end{bmatrix}. \tag{3.38}$$

The components of this function are

$$f_1(x_1, x_2) = x_1 \sin(x_2) \quad \text{and} \quad f_2(x_1, x_2) = x_1^2 + x_2^2. \tag{3.39}$$

Therefore

$$\partial_1 f_1(\mathbf{x}) = \frac{\partial}{\partial x_1} x_1 \sin(x_2) = \sin(x_2) \quad \text{and} \quad \partial_2 f_1(\mathbf{x}) = \frac{\partial}{\partial x_2} x_1 \sin(x_2) = x_1 \cos(x_2) \tag{3.40}$$

and

$$\partial_1 f_2(\mathbf{x}) = \frac{\partial}{\partial x_1}(x_1^2 + x_2^2) = 2x_1 \quad \text{and} \quad \partial_2 f_2(\mathbf{x}) = \frac{\partial}{\partial x_2}(x_1^2 + x_2^2) = 2x_2. \tag{3.41}$$

The Jacobian is then

$$\mathbf{J}_\mathbf{f}(\mathbf{x}) = \begin{bmatrix} \partial_1 f_1(\mathbf{x}) & \partial_2 f_1(\mathbf{x}) \\ \partial_1 f_2(\mathbf{x}) & \partial_2 f_2(\mathbf{x}) \end{bmatrix} = \begin{bmatrix} \sin(x_2) & x_1 \cos(x_2) \\ 2x_1 & 2x_2 \end{bmatrix} \in \mathbb{R}^{2\times 2}. \tag{3.42}$$

By (3.36), the Jacobian determinant is

$$\det(\mathbf{J}_\mathbf{f}(\mathbf{x})) = |\mathbf{J}_\mathbf{f}(\mathbf{x})| = \begin{vmatrix} \sin(x_2) & x_1\cos(x_2) \\ 2x_1 & 2x_2 \end{vmatrix} = 2x_2\sin(x_2) - 2x_1^2\cos(x_2). \tag{3.43}$$

Example 3.15. Let $\mathbf{A} \in \mathbb{R}^{d\times d}$ be a diagonal matrix with the diagonal element $[\mathbf{A}]_{i,i} = i$ for every $i = 1,\dots,d$. Consider the function $\mathbf{f}\colon \mathbb{R}^d \to \mathbb{R}^d$ given by

$$\mathbf{f}(\mathbf{x}) = \mathbf{A}\mathbf{x}. \tag{3.44}$$

The ith component of this function is simply $f_i(\mathbf{x}) = ix_i$. In particular, the ith component depends only on the ith variable x_i. Therefore

$$\partial_i f_i(\mathbf{x}) = i \quad \text{and} \quad \partial_j f_i(\mathbf{x}) = 0 \quad \text{if} \quad j \neq i. \tag{3.45}$$

This means that the Jacobian matrix $\mathbf{J}_\mathbf{f}(\mathbf{x})$ equals the diagonal matrix $\mathbf{A}$. By (3.37), the Jacobian determinant is thus

$$\det(\mathbf{J}_\mathbf{f}(\mathbf{x})) = \det(\mathbf{A}) = 1\cdot 2\cdots(d-1)\cdot d = d!. \tag{3.46}$$

Example 3.16. Consider the function $\mathbf{f}\colon \mathbb{R}^3 \to \mathbb{R}^2$ given by

$$\mathbf{f}(\mathbf{x}) = \mathbf{f}(x_1,x_2,x_3) = \begin{bmatrix} x_1x_2x_3 - e^{x_1+x_2} \\ x_1^7x_3^3 + 1 \end{bmatrix}. \tag{3.47}$$

The components of this function are

$$f_1(x_1,x_2,x_3) = x_1x_2x_3 - e^{x_1+x_2} \quad \text{and} \quad f_2(x_1,x_2,x_3) = x_1^7x_3^3 + 1. \tag{3.48}$$

Therefore

$$\partial_1 f_1(\mathbf{x}) = x_2x_3 - e^{x_1+x_2}, \quad \partial_2 f_1(\mathbf{x}) = x_1x_3 - e^{x_1+x_2}, \quad \text{and} \quad \partial_3 f_1(\mathbf{x}) = x_1x_2 \tag{3.49}$$

and

$$\partial_1 f_2(\mathbf{x}) = 7x_1^6x_3^3, \quad \partial_2 f_2(\mathbf{x}) = 0, \quad \text{and} \quad \partial_3 f_2(\mathbf{x}) = 3x_1^7x_3^2. \tag{3.50}$$

The Jacobian is then

$$\mathbf{J}_\mathbf{f}(\mathbf{x}) = \begin{bmatrix} \partial_1 f_1(\mathbf{x}) & \partial_2 f_1(\mathbf{x}) & \partial_3 f_1(\mathbf{x}) \\ \partial_1 f_2(\mathbf{x}) & \partial_2 f_2(\mathbf{x}) & \partial_3 f_2(\mathbf{x}) \end{bmatrix} = \begin{bmatrix} x_2x_3 - e^{x_1+x_2} & x_1x_3 - e^{x_1+x_2} & x_1x_2 \\ 7x_1^6x_3^3 & 0 & 3x_1^7x_3^2 \end{bmatrix} \in \mathbb{R}^{2\times 3}.$$

The Jacobian determinant is not defined because the Jacobian is not a square matrix.

3.3 Differentiation Rules and Matrix Calculus

So far we have computed gradients and Jacobians by computing each necessary partial derivative individually. In Examples 3.9 and 3.15 we encountered hints of a more efficient approach called *matrix calculus*, which we review in this section. For fuller accounts and much richer collections of formulae you should consult

- *The Matrix Cookbook* by Petersen and Pedersen (https://www2.imm.dtu.dk/pubdb/edoc/imm3274.pdf) and

- the Wikipedia page *Matrix calculus* and in particular its section on identities (https://en.wikipedia.org/wiki/Matrix_calculus).

Note that different conventions (*denominator* and *numerator* layous) exist on how to write gradients and Jacobians; here we follow a hybrid convention. The convention one uses affects where transposes appear in the formulae.

3.3.1 Differentiation Rules

The following results generalise Results 2.10, 2.11, and 2.15 for gradients of multivariate functions $f\colon \mathbb{R}^d \to \mathbb{R}$.

Result 3.17 (Linearity of gradient). Let $f(\mathbf{x}) = ag(\mathbf{x}) + bh(\mathbf{x})$ for functions $g,h\colon \mathbb{R}^d \to \mathbb{R}$ and constants $a,b \in \mathbb{R}$. Then

$$\nabla f(\mathbf{x}) = a\nabla g(\mathbf{x}) + b\nabla h(\mathbf{x}). \tag{3.51}$$

Example 3.18. Consider the function

$$f(x_1,x_2) = 2g(x_1,x_2) - 3h(x_1,x_2), \tag{3.52}$$

where $g(x_1,x_2) = x_1^2 + x_2^2$ and $h(x_1,x_2) = x_1 e^{x_2}$. These functions have the gradients

$$\nabla g(x_1,x_2) = \begin{bmatrix} 2x_1 \\ 2x_2 \end{bmatrix} \quad \text{and} \quad \nabla h(x_1,x_2) = \begin{bmatrix} e^{x_2} \\ x_1 e^{x_2} \end{bmatrix}. \tag{3.53}$$

Therefore Result 3.17 gives

$$\nabla f(x_1,x_2) = 2\nabla g(x_1,x_2) - 3\nabla h(x_1,x_2) = \begin{bmatrix} 4x_1 - 3e^{x_2} \\ 4x_2 - 3x_1 e^{x_2} \end{bmatrix}. \tag{3.54}$$

Result 3.19 (Multivariate product rule). Let $f(\mathbf{x}) = g(\mathbf{x})h(\mathbf{x})$ for functions $g,h\colon \mathbb{R}^d \to \mathbb{R}$. Then

$$\nabla f(\mathbf{x}) = h(\mathbf{x})\nabla g(\mathbf{x}) + g(\mathbf{x})\nabla h(\mathbf{x}). \tag{3.55}$$

Example 3.20. Consider the function

$$f(x_1,x_2) = x_1^2 x_2^3 \cos(x_1 x_2) = g(x_1,x_2)h(x_1,x_2), \tag{3.56}$$

where

$$g(x_1,x_2) = \cos(x_1 x_2) \quad \text{and} \quad h(x_1,x_2) = x_1^2 x_2^3. \tag{3.57}$$

Because

$$\nabla g(x_1,x_2) = \begin{bmatrix} -x_2 \sin(x_1 x_2) \\ -x_1 \sin(x_1 x_2) \end{bmatrix} \quad \text{and} \quad \nabla h(x_1,x_2) = \begin{bmatrix} 2x_1 x_2^3 \\ 3x_1^2 x_2^2 \end{bmatrix}, \tag{3.58}$$

Result 3.19 yields

$$\begin{aligned}\nabla f(x_1,x_2) &= h(x_1,x_2)\nabla g(x_1,x_2) + g(x_1,x_2)\nabla h(x_1,x_2)\\ &= x_1^2x_2^3\begin{bmatrix}-x_2\sin(x_1x_2)\\-x_1\sin(x_1x_2)\end{bmatrix} + \cos(x_1x_2)\begin{bmatrix}2x_1x_2^3\\3x_1^2x_2^2\end{bmatrix}\\ &= \begin{bmatrix}2x_1x_2^3\cos(x_1x_2) - x_1^2x_2^4\sin(x_1x_2)\\3x_1^2x_2^2\cos(x_1x_2) - x_1^3x_2^3\sin(x_1x_2)\end{bmatrix}.\end{aligned} \tag{3.59}$$

We could have obtained the same gradient by computing the partial derivatives $\partial_1 f(x_1,x_2)$ and $\partial_2 f(x_1,x_2)$ with the univariate product rule (2.30).

Result 3.21 (MULTIVARIATE CHAIN RULE). Let $f = h \circ \mathbf{g}\colon \mathbb{R}^d \to \mathbb{R}$ be a composite function given by $f(\mathbf{x}) = h(\mathbf{g}(\mathbf{x}))$ for functions $h\colon \mathbb{R}^m \to \mathbb{R}$ and $\mathbf{g}\colon \mathbb{R}^d \to \mathbb{R}^m$. Then

$$\nabla f(\mathbf{x}) = \mathbf{J}_{\mathbf{g}}(\mathbf{x})^\mathsf{T} \nabla h(\mathbf{g}(\mathbf{x})). \tag{3.60}$$

Recall that the univariate chain rule from Result 2.15 is

$$f'(x) = h'(g(x))g'(x) = g'(x)h'(g(x)) \quad \text{for} \quad f(x) = h(g(x)). \tag{3.61}$$

The multivariate chain rule has precisely the same general form: the derivative (i.e., gradient) of the outer function h is evaluated at $\mathbf{g}(\mathbf{x})$ and the result is multiplied by the derivative of $\mathbf{g}$ (i.e., Jacobian). If $g\colon \mathbb{R}^d \to \mathbb{R}$ and $h\colon \mathbb{R} \to \mathbb{R}$ (i.e., $m = 1$), then the gradient of h is nothing but the derivative h' and the Jacobian of g is the transpose of its gradient. Therefore the multivariate chain rule simplifies to

$$\nabla f(\mathbf{x}) = h'(g(\mathbf{x}))\nabla g(\mathbf{x}) \quad \text{if} \quad f(\mathbf{x}) = h(g(\mathbf{x})) \text{ for } h\colon \mathbb{R} \to \mathbb{R} \text{ and } g\colon \mathbb{R}^d \to \mathbb{R}. \tag{3.62}$$

Example 3.22. Consider the function

$$f(x_1,x_2,x_3) = e^{-\sin(x_1x_2)} + (\sin(x_1x_2))^2 - \sin(x_3)\cos(x_3). \tag{3.63}$$

We may write this function as

$$f(x_1,x_2,x_3) = h(\mathbf{g}(x_1,x_2,x_3)), \tag{3.64}$$

where the functions $\mathbf{g}\colon \mathbb{R}^3 \to \mathbb{R}^2$ and $h\colon \mathbb{R}^2 \to \mathbb{R}$ are given by

$$\mathbf{g}(x_1,x_2,x_3) = \begin{bmatrix}g_1(x_1,x_2,x_3)\\g_2(x_1,x_2,x_3)\end{bmatrix} = \begin{bmatrix}\sin(x_1x_2)\\x_3\end{bmatrix} \quad \text{and} \quad h(z_1,z_2) = e^{-z_1} + z_1^2 - \sin(z_2)\cos(z_2).$$

Here we use z_1 and z_2 for the arguments of h for clarity. That is, to evaluate $h(\mathbf{g}(x_1,x_2,x_3))$ we first compute $\mathbf{g}(x_1,x_2,x_3)$ and the insert this into h:

$$\begin{aligned}h(\mathbf{g}(x_1,x_2,x_3)) &= h(g_1(x_1,x_2,x_3), g_2(x_1,x_2,x_3))\\ &= h(\sin(x_1x_2), x_3)\\ &= h(z_1 = \sin(x_1x_2), z_2 = x_3)\\ &= e^{-\sin(x_1x_2)} + (\sin(x_1x_2))^2 - \sin(x_3)\cos(x_3),\end{aligned} \tag{3.65}$$

which is $f(x_1,x_2,x_3)$. We have the gradient and Jacobian

$$\nabla h(z_1,z_2) = \begin{bmatrix} -e^{z_1}+2z_1 \\ -(\cos(z_2))^2+(\sin(z_2))^2 \end{bmatrix} \in \mathbb{R}^2 \tag{3.66}$$

and

$$\mathbf{J}_{\mathbf{g}}(x_1,x_2,x_2) = \begin{bmatrix} x_2\cos(x_1x_2) & x_1\cos(x_1x_2) & 0 \\ 0 & 0 & 1 \end{bmatrix} \in \mathbb{R}^{2\times 3}. \tag{3.67}$$

Therefore Result 3.21 gives

$$\begin{aligned} \nabla f(x_1,x_2,x_3) &= \mathbf{J}_{\mathbf{g}}(x_1,x_2,x_3)^{\mathsf{T}} \nabla h(\mathbf{g}(x_1,x_2,x_3)) \\ &= \begin{bmatrix} x_2\cos(x_1x_2) & 0 \\ x_1\cos(x_1x_2) & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} -e^{\sin(x_1x_2)}+2\sin(x_1x_2) \\ -(\cos(x_3))^2+(\sin(x_3))^2 \end{bmatrix} \\ &= \begin{bmatrix} x_2\cos(x_1x_2)(2\sin(x_1x_2)-e^{\sin(x_1x_2)}) \\ x_1\cos(x_1x_2)(2\sin(x_1x_2)-e^{\sin(x_1x_2)}) \\ (\sin(x_3))^2-(\cos(x_3))^2 \end{bmatrix} \in \mathbb{R}^3, \end{aligned} \tag{3.68}$$

which we could have also obtained by applying the univariate chain rule to each of the partial derivatives $\partial_1 f(x_1,x_2,x_3)$, $\partial_2 f(x_1,x_2,x_3)$, and $\partial_3 f(x_1,x_2,x_3)$.

Example 3.23. Consider the function

$$f(x_1,x_2) = \log(1+x_1^2+x_2^2), \tag{3.69}$$

which we may write as

$$f(x_1,x_2) = h(g(x_1,x_2)) \tag{3.70}$$

for the functions $g\colon \mathbb{R}^2 \to \mathbb{R}$ and $h\colon \mathbb{R} \to \mathbb{R}$ given by

$$g(x_1,x_2) = x_1^2+x_2^2 \quad \text{and} \quad h(z) = \log(1+z). \tag{3.71}$$

We can now use (3.62) because g is real-valued. The gradient of g and derivative of h are

$$\nabla g(x_1,x_2) = \begin{bmatrix} 2x_1 \\ 2x_2 \end{bmatrix} = \quad \text{and} \quad h'(z) = \frac{1}{1+z}. \tag{3.72}$$

Therefore (3.62) yields the gradient

$$\nabla f(\mathbf{x}) = \nabla f(x_1,x_2) = h'(g(x_1,x_2))\nabla g(x_1,x_2) = \frac{1}{1+x_1^2+x_2^2}\begin{bmatrix} 2x_1 \\ 2x_2 \end{bmatrix} = \frac{2}{1+\|\mathbf{x}\|^2}\mathbf{x}. \tag{3.73}$$

Example 3.24. Let us consider a variant of the neural network example in Section 1.1 in which the inputs are two-dimensional: $\mathbf{x}_i \in \mathbb{R}^2$. Note that here we use the subscript i to denote different points in $\mathbb{R}^2$ rather than the coordinates of a point. Let $\mathbf{w} \in \mathbb{R}^2$ be a weight vector and $b \in \mathbb{R}$ the bias. For $\mathbf{x} \in \mathbb{R}^2$, define [it may be useful to compare this to (1.6)

and (1.10)]

$$F(\mathbf{x} \mid \mathbf{w}, b) = \sigma(\mathbf{w} \cdot \mathbf{x} + b) = \sigma(w_1 x_1 + w_2 x_2 + b), \quad \text{where} \quad \sigma(x) = \frac{1}{1+e^{-x}}. \tag{3.74}$$

Given some outputs $y_i \in \mathbb{R}$ at inputs $\mathbf{x}_i = (x_{i,1}, x_{i,2}) \in \mathbb{R}^2$, the corresponding quadratic loss function [recall (1.7)] would be

$$L(\mathbf{w}, b) = \sum_{i=1}^{n} (y_i - F(\mathbf{x}_i \mid \mathbf{w}, b))^2 = \sum_{i=1}^{n} (y_i - \sigma(\mathbf{w} \cdot \mathbf{x} + b))^2. \tag{3.75}$$

We use the multivariate chain rule (3.60) to compute the gradient of $L(\mathbf{p}) = L(\mathbf{w}, b)$ with respect to the full parameter vector $\mathbf{p} = (\mathbf{w}, b) = (w_1, w_2, b) \in \mathbb{R}^3$. Begin by writing the loss function as

$$L(\mathbf{p}) = \sum_{i=1}^{n} g_i(\mathbf{p})^2 = \|\mathbf{g}(\mathbf{p})\|^2 = h(\mathbf{g}(\mathbf{p})), \tag{3.76}$$

where the function $\mathbf{g}\colon \mathbb{R}^3 \to \mathbb{R}^n$ has the components $g_i(\mathbf{p}) = g_i(\mathbf{w}, b) = y_i - \sigma(\mathbf{w} \cdot \mathbf{x}_i + b)$ and $h\colon \mathbb{R}^n \to \mathbb{R}$ is given by $h(\mathbf{z}) = \|\mathbf{z}\|^2$. To apply Result 3.21 we need the gradient of h and the Jacobian of $\mathbf{g}$. Because

$$h(\mathbf{z}) = \|\mathbf{z}\|^2 = z_1^2 + \cdots + z_n^2, \tag{3.77}$$

the partial derivative with respect to the jth variable is $\partial_j h(\mathbf{z}) = 2z_j$. Thus

$$\nabla h(\mathbf{z}) = \begin{bmatrix} \partial_1 h(\mathbf{z}) \\ \vdots \\ \partial_n h(\mathbf{z}) \end{bmatrix} = \begin{bmatrix} 2z_1 \\ \vdots \\ 2z_n \end{bmatrix} = 2\mathbf{z} \in \mathbb{R}^n. \tag{3.78}$$

The Jacobian of $\mathbf{g}$ is

$$\begin{aligned} \mathbf{J}_{\mathbf{g}}(\mathbf{p}) &= \begin{bmatrix} \partial_1 g_1(\mathbf{p}) & \partial_2 g_1(\mathbf{p}) & \partial_3 g_1(\mathbf{p}) \\ \vdots & \vdots & \vdots \\ \partial_1 g_n(\mathbf{p}) & \partial_2 g_n(\mathbf{p}) & \partial_3 g_n(\mathbf{p}) \end{bmatrix} \\ &= \begin{bmatrix} \partial_{w_1} g_1(w_1, w_2, b) & \partial_{w_2} g_1(w_1, w_2, b) & \partial_b g_1(w_1, w_2, b) \\ \vdots & \vdots & \vdots \\ \partial_{w_1} g_n(w_1, w_2, b) & \partial_{w_2} g_n(w_1, w_2, b) & \partial_b g_n(w_1, w_2, b) \end{bmatrix}. \end{aligned} \tag{3.79}$$

From

$$\sigma'(x) = \frac{\mathrm{d}}{\mathrm{d}x} \frac{1}{1+e^{-x}} = \frac{e^{-x}}{(1+e^{-x})^2} \tag{3.80}$$

and the univariate chain rule we obtain the partial derivatives

$$\partial_{w_1} g_i(w_1, w_2, b) = -\frac{x_{i,1}\, e^{-\mathbf{w} \cdot \mathbf{x}_i - b}}{(1+e^{-\mathbf{w} \cdot \mathbf{x}_i - b})^2}, \tag{3.81}$$

$$\partial_{w_2} g_i(w_1, w_2, b) = -\frac{x_{i,2}\, e^{-\mathbf{w} \cdot \mathbf{x}_i - b}}{(1+e^{-\mathbf{w} \cdot \mathbf{x}_i - b})^2}, \tag{3.82}$$

$$\partial_b g_i(w_1, w_2, b) = -\frac{e^{-\mathbf{w} \cdot \mathbf{x}_i - b}}{(1+e^{-\mathbf{w} \cdot \mathbf{x}_i - b})^2}. \tag{3.83}$$

By inserting these in (3.79) and observing that all elements of the resulting matrix have the factor $-e^{-\mathbf{w}\cdot\mathbf{x}_i-b}/(1+e^{-\mathbf{w}\cdot\mathbf{x}_i-b})^2$ yields the Jacobian

$$\mathbf{J}_\mathbf{g}(\mathbf{p}) = -\frac{e^{-\mathbf{w}\cdot\mathbf{x}_i-b}}{(1+e^{-\mathbf{w}\cdot\mathbf{x}_i-b})^2}\begin{bmatrix} x_{1,1} & x_{1,2} & 1 \\ \vdots & \vdots & \vdots \\ x_{n,1} & x_{n,2} & 1 \end{bmatrix} = -\frac{e^{-\mathbf{w}\cdot\mathbf{x}_i-b}}{(1+e^{-\mathbf{w}\cdot\mathbf{x}_i-b})^2}\begin{bmatrix} \mathbf{x}_1^\mathsf{T} & 1 \\ \vdots & \vdots \\ \mathbf{x}_n^\mathsf{T} & 1 \end{bmatrix} \in \mathbb{R}^{n\times 3}.$$

From the multivariate chain rule (3.60) we then obtain

$$\begin{aligned} \nabla L(\mathbf{w},b) = \nabla L(\mathbf{p}) &= \mathbf{J}_\mathbf{g}(\mathbf{p})^\mathsf{T}\nabla h(\mathbf{g}(\mathbf{p})) \\ &= -\frac{e^{-\mathbf{w}\cdot\mathbf{x}_i-b}}{(1+e^{-\mathbf{w}\cdot\mathbf{x}_i-b})^2}\begin{bmatrix} \mathbf{x}_1 & \cdots & \mathbf{x}_n \\ 1 & \cdots & 1 \end{bmatrix}\cdot 2\begin{bmatrix} y_1-\sigma(\mathbf{w}\cdot\mathbf{x}_1+b) \\ \vdots \\ y_n-\sigma(\mathbf{w}\cdot\mathbf{x}_n+b) \end{bmatrix} \\ &= -\frac{2e^{-\mathbf{w}\cdot\mathbf{x}_i-b}}{(1+e^{-\mathbf{w}\cdot\mathbf{x}_i-b})^2}\begin{bmatrix} \sum_{i=1}^n x_{i,1}(y_i-\sigma(\mathbf{w}\cdot\mathbf{x}_i+b)) \\ \sum_{i=1}^n x_{i,2}(y_i-\sigma(\mathbf{w}\cdot\mathbf{x}_i+b)) \\ \sum_{i=1}^n (y_i-\sigma(\mathbf{w}\cdot\mathbf{x}_i+b)) \end{bmatrix} \\ &= -\frac{2e^{-\mathbf{w}\cdot\mathbf{x}_i-b}}{(1+e^{-\mathbf{w}\cdot\mathbf{x}_i-b})^2}\sum_{i=1}^n (y_i-\sigma(\mathbf{w}\cdot\mathbf{x}_i+b))\begin{bmatrix}\mathbf{x}_i \\ 1\end{bmatrix} \in \mathbb{R}^3. \end{aligned} \tag{3.84}$$

3.3.2 *Matrix Calculus*

The following results provide convenient expressions for gradients and Jacobians of functions that are expressed in terms of linear algebra.

Result 3.25 (GRADIENT FORMULAE). Functions $f\colon \mathbb{R}^d \to \mathbb{R}$ have the following gradients.

Let $\mathbf{a}\in\mathbb{R}^d$:	$f(\mathbf{x}) = \mathbf{a}\cdot\mathbf{x}$	$\Longrightarrow$	$\nabla f(\mathbf{x}) = \mathbf{a}$	(3.85)
Let $\mathbf{A}\in\mathbb{R}^{p\times d}$, $\mathbf{b}\in\mathbb{R}^p$:	$f(\mathbf{x}) = \mathbf{b}^\mathsf{T}\mathbf{A}\mathbf{x}$	$\Longrightarrow$	$\nabla f(\mathbf{x}) = \mathbf{A}^\mathsf{T}\mathbf{b}$	(3.86)
Let $\mathbf{A}\in\mathbb{R}^{d\times d}$:	$f(\mathbf{x}) = \mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{x}$	$\Longrightarrow$	$\nabla f(\mathbf{x}) = (\mathbf{A}+\mathbf{A}^\mathsf{T})\mathbf{x}$	(3.87)
Let $\mathbf{A}\in\mathbb{R}^{d\times d}$ be symmetric :	$f(\mathbf{x}) = \mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{x}$	$\Longrightarrow$	$\nabla f(\mathbf{x}) = 2\mathbf{A}\mathbf{x}$	(3.88)
	$f(\mathbf{x}) = \lVert\mathbf{x}\rVert^2$	$\Longrightarrow$	$\nabla f(\mathbf{x}) = 2\mathbf{x}$	(3.89)
Let $\mathbf{a}\in\mathbb{R}^d$ and $\mathbf{x}\neq\mathbf{a}$:	$f(\mathbf{x}) = \lVert\mathbf{x}-\mathbf{a}\rVert$	$\Longrightarrow$	$\nabla f(\mathbf{x}) = \dfrac{\mathbf{x}-\mathbf{a}}{\lVert\mathbf{x}-\mathbf{a}\rVert}$	(3.90)

Note that (3.88) follows from (3.87) because a symmetric matrix satisfies $\mathbf{A}^\mathsf{T} = \mathbf{A}$, so that $\mathbf{A}+\mathbf{A}^\mathsf{T} = 2\mathbf{A}$. Equation (3.89) follows from (3.88) by selecting $\mathbf{A}$ as the $d\times d$ identity matrix

$$\mathbf{I} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix} \in \mathbb{R}^{d\times d}. \tag{3.91}$$

Then

$$\mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{x} = \mathbf{x}^\mathsf{T}\mathbf{I}\mathbf{x} = \mathbf{x}^\mathsf{T}\mathbf{x} = \lVert\mathbf{x}\rVert^2 \quad \text{and} \quad 2\mathbf{A}\mathbf{x} = 2\mathbf{I}\mathbf{x} = 2\mathbf{x}. \tag{3.92}$$

The formulae in Result 3.25 can be verified by expanding the linear-algebraic expressions and computing each partial derivative. We do this for a few of them in the following examples.

Example 3.26. Let $f(\mathbf{x}) = \mathbf{a} \cdot \mathbf{x}$ for some vector $\mathbf{a} \in \mathbb{R}^d$. Because

$$f(\mathbf{x}) = \mathbf{a} \cdot \mathbf{x} = \mathbf{a}^\mathsf{T}\mathbf{x} = a_1 x_1 + \cdots + a_d x_d = \sum_{i=1}^{d} a_i x_i, \tag{3.93}$$

we compute that the partial derivative with respect to x_j is

$$\partial_j f(\mathbf{x}) = \frac{\partial}{\partial x_j} \sum_{i=1}^{d} a_i x_i = \sum_{i=1}^{d} a_i \frac{\partial}{\partial x_j} x_i = a_j, \tag{3.94}$$

where $\partial x_i / \partial x_j = 0$ if $j \neq i$. This gives the gradient

$$\nabla f(\mathbf{x}) = \begin{bmatrix} \partial_1 f(\mathbf{x}) \\ \vdots \\ \partial_d f(\mathbf{x}) \end{bmatrix} = \begin{bmatrix} a_1 \\ \vdots \\ a_d \end{bmatrix} = \mathbf{a} \in \mathbb{R}^d, \tag{3.95}$$

which is (3.85).

Example 3.27. Let $f(\mathbf{x}) = \|\mathbf{x}\|^2 = x_1^2 + \cdots + x_d^2 = \sum_{i=1}^d x_i^2$. In Equation (3.78) of Example 3.24 we computed that the gradient is $\nabla f(\mathbf{x}) = 2\mathbf{x}$, which is (3.89).

Example 3.28. Let a $f(\mathbf{x}) = \|\mathbf{x} - \mathbf{a}\|$ for $\mathbf{a} \in \mathbb{R}^d$. We may write this function as

$$f(\mathbf{x}) = h(\mathbf{g}(\mathbf{x})) \quad \text{for} \quad h(\mathbf{x}) = \|\mathbf{x}\| \quad \text{and} \quad \mathbf{g}(\mathbf{x}) = \mathbf{x} - \mathbf{a}. \tag{3.96}$$

After computing the Jacobian of $\mathbf{g}$ and the gradient of h we can apply the multivariate chain rule in (3.60). Because $[\mathbf{g}(\mathbf{x})]_i = g_i(\mathbf{x}) = x_i - a_i$, we have $\partial_i g_i(\mathbf{x}) = 1$ and $\partial_j g_i(\mathbf{x}) = 0$ if $j \neq i$. Therefore the Jacobian of $\mathbf{g}$ is nothing but the identity matrix:

$$\mathbf{J}_\mathbf{g}(\mathbf{x}) = \begin{bmatrix} \partial_1 g_1(\mathbf{x}) & \cdots & \partial_d g_1(\mathbf{x}) \\ \vdots & \ddots & \vdots \\ \partial_1 g_d(\mathbf{x}) & \cdots & \partial_d g_d(\mathbf{x}) \end{bmatrix} = \mathbf{I} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix} \in \mathbb{R}^{d \times d}. \tag{3.97}$$

By (2.8) with $n = \frac{1}{2}$ and the univariate chain rule, the partial derivatives of h are

$$\partial_i h(\mathbf{x}) = \frac{\partial}{\partial x_i} \|\mathbf{x}\| = \frac{\partial}{\partial x_i} (x_1^2 + \cdots + x_d^2)^{1/2} = \tfrac{1}{2}(x_1^2 + \cdots + x_d^2)^{-1/2} \cdot 2x_i = \frac{x_i}{\|\mathbf{x}\|} \tag{3.98}$$

for $i = 1, \ldots, d$ when $\mathbf{x} \neq \mathbf{0}$. The gradient is thus

$$\nabla h(\mathbf{x}) = \begin{bmatrix} \partial_1 h(\mathbf{x}) \\ \vdots \\ \partial_d h(\mathbf{x}) \end{bmatrix} = \begin{bmatrix} x_1/\|\mathbf{x}\| \\ \vdots \\ x_d/\|\mathbf{x}\| \end{bmatrix} = \frac{\mathbf{x}}{\|\mathbf{x}\|} \in \mathbb{R}^d \quad \text{for} \quad \mathbf{x} \neq \mathbf{0}. \tag{3.99}$$

The gradient does not exist for $\mathbf{x} = \mathbf{0}$. The norm function $f(\mathbf{x}) = \|\mathbf{x}\|$ is the multivariate version of the absolute value function $f(x) = |x|$ whose differentiability we discussed in Remark 2.6. We can now use the multivariate chain rule, which gives [since the Jacobian is the $d \times d$ identity matrix and $\mathbf{g}(\mathbf{x}) = \mathbf{x} - \mathbf{a}$]

$$\nabla f(\mathbf{x}) = \mathbf{J}_{\mathbf{g}}(\mathbf{x})^{\mathsf{T}} \nabla h(\mathbf{g}(\mathbf{x})) = \mathbf{I} \nabla h(\mathbf{x} - \mathbf{a}) = \nabla h(\mathbf{x} - \mathbf{a}) = \frac{\mathbf{x} - \mathbf{a}}{\|\mathbf{x} - \mathbf{a}\|}, \tag{3.100}$$

which is (3.90).

Using Result 3.25 with multivariate product and chain rules can make gradient computations quite compact.

Example 3.29. Consider the function $f\colon \mathbb{R}^d \to \mathbb{R}$ given by

$$f(\mathbf{x}) = \|\mathbf{x} - \mathbf{a}\| \, \|\mathbf{x}\|^2 \tag{3.101}$$

for a vector $\mathbf{a} \in \mathbb{R}^d$. We can apply the multivariate product rule with $g(\mathbf{x}) = \|\mathbf{x} - \mathbf{a}\|$ and $h(\mathbf{x}) = \|\mathbf{x}\|^2$. Together with (3.89) and (3.90) this yields the gradient

$$\nabla f(\mathbf{x}) = h(\mathbf{x}) \nabla g(\mathbf{x}) + g(\mathbf{x}) \nabla h(\mathbf{x}) = \|\mathbf{x}\|^2 \frac{\mathbf{x} - \mathbf{a}}{\|\mathbf{x} - \mathbf{a}\|} + 2\|\mathbf{x} - \mathbf{a}\| \, \mathbf{x} \in \mathbb{R}^d \tag{3.102}$$

for $\mathbf{x} \neq \mathbf{0}$. Note that using $\mathbf{a} = \mathbf{0}$ would simplify the function to $f(\mathbf{x}) = \|\mathbf{x}\|^3$ and the gradient to

$$\nabla f(\mathbf{x}) = 3\|\mathbf{x}\| \, \mathbf{x}, \tag{3.103}$$

which is analogous to the univariate derivative

$$\frac{\mathrm{d}}{\mathrm{d}x} x^3 = 3x^2. \tag{3.104}$$

Example 3.30. Consider the function $f\colon \mathbb{R}^d \to \mathbb{R}$ given by

$$f(\mathbf{x}) = \log(1 + (\mathbf{a} \cdot \mathbf{x})^2) \tag{3.105}$$

for a vector $\mathbf{a} \in \mathbb{R}^d$. We can write this function as

$$f(\mathbf{x}) = h(g(\mathbf{x})) \quad \text{for} \quad h(z) = \log(1 + z^2) \quad \text{and} \quad g(\mathbf{x}) = \mathbf{a} \cdot \mathbf{x}. \tag{3.106}$$

Note that $h\colon \mathbb{R} \to \mathbb{R}$ and $g\colon \mathbb{R}^d \to \mathbb{R}$. By the univariate chain rule and (3.85),

$$h'(z) = \frac{2z}{1 + z^2} \quad \text{and} \quad \nabla g(\mathbf{x}) = \mathbf{a}. \tag{3.107}$$

The simplified version of the multivariate chain rule in (3.62) thus yields

$$\nabla f(\mathbf{x}) = h'(g(\mathbf{x})) \nabla g(\mathbf{x}) = \left(\frac{2\,\mathbf{a} \cdot \mathbf{x}}{1 + (\mathbf{a} \cdot \mathbf{x})^2} \right) \mathbf{a} \in \mathbb{R}^d. \tag{3.108}$$

Formulae similar to those in Result 3.25 are available for the Jacobian, which should not come as a surprise given than that the Jacobian consists of stacked gradients of the component functions [recall (3.34)].

Result 3.31 (JACOBIAN FORMULAE). Let $\mathbf{g}\colon \mathbb{R}^d \to \mathbb{R}^p$ and $\mathbf{h}\colon \mathbb{R}^d \to \mathbb{R}^q$ be functions and $\mathbf{A} \in \mathbb{R}^{m\times p}$ and $\mathbf{B} \in \mathbb{R}^{m\times q}$ matrices. Then

$$\mathbf{J}_\mathbf{f}(\mathbf{x}) = \mathbf{A}\mathbf{J}_\mathbf{g}(\mathbf{x}) + \mathbf{B}\mathbf{J}_\mathbf{h}(\mathbf{x}) \in \mathbb{R}^{m\times d} \quad \text{if} \quad \mathbf{f}(\mathbf{x}) = \mathbf{A}\mathbf{g}(\mathbf{x}) + \mathbf{B}\mathbf{h}(\mathbf{x}). \tag{3.109}$$

In particular

$$\mathbf{J}_\mathbf{f}(\mathbf{x}) = \mathbf{A} \in \mathbb{R}^{m\times d} \quad \text{if} \quad \mathbf{f}(\mathbf{x}) = \mathbf{A}\mathbf{x} \quad \text{for} \quad \mathbf{A} \in \mathbb{R}^{m\times d}. \tag{3.110}$$

Example 3.32. Let us verify (3.110). Let $\mathbf{a}_1^\mathsf{T}, \dots, \mathbf{a}_d^\mathsf{T} \in \mathbb{R}^{1\times d}$ be the rows of $\mathbf{A} \in \mathbb{R}^{d\times d}$. That is,

$$\mathbf{A} = \begin{bmatrix} \mathbf{a}_1^\mathsf{T} \\ \vdots \\ \mathbf{a}_d^\mathsf{T} \end{bmatrix}. \tag{3.111}$$

Then

$$\mathbf{f}(\mathbf{x}) = \mathbf{A}\mathbf{x} = \begin{bmatrix} \mathbf{a}_1^\mathsf{T} \\ \vdots \\ \mathbf{a}_d^\mathsf{T} \end{bmatrix} \mathbf{x} = \begin{bmatrix} \mathbf{a}_1^\mathsf{T}\mathbf{x} \\ \vdots \\ \mathbf{a}_d^\mathsf{T}\mathbf{x}, \end{bmatrix} \tag{3.112}$$

which means that the ith component of $\mathbf{f}$ is given by $f_i(\mathbf{x}) = \mathbf{a}_i^\mathsf{T}\mathbf{x} = \mathbf{a}_i \cdot \mathbf{x}$. From (3.85) we thus get

$$\nabla f_i(\mathbf{x}) = \mathbf{a}_i \quad \text{for} \quad i = 1, \dots, d. \tag{3.113}$$

From (3.34) we recall that the Jacobian can be obtained by stacking gradients. Consequently,

$$\mathbf{J}_\mathbf{f}(\mathbf{x}) = \begin{bmatrix} \nabla f_1(\mathbf{x})^\mathsf{T} \\ \vdots \\ \nabla f_m(\mathbf{x})^\mathsf{T} \end{bmatrix} = \begin{bmatrix} \mathbf{a}_1^\mathsf{T} \\ \vdots \\ \mathbf{a}_d^\mathsf{T} \end{bmatrix} = \mathbf{A}, \tag{3.114}$$

which is (3.110).

Example 3.33. Let $f\colon \mathbb{R}^d \to \mathbb{R}$ be given by

$$f(\mathbf{x}) = \|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2 \tag{3.115}$$

for a matrix $\mathbf{A} \in \mathbb{R}^{m\times d}$ (so that $\mathbf{A}\mathbf{x} \in \mathbb{R}^m$) and a vector $\mathbf{b} \in \mathbb{R}^m$. We use Results 3.25 and 3.31 and the multivariate chain rule (3.60) to compute the Jacobian of f. Write the function as

$$f(\mathbf{x}) = h(\mathbf{g}(\mathbf{x})) \quad \text{for} \quad h(\mathbf{z}) = \|\mathbf{z}\|^2 \quad \text{and} \quad \mathbf{g}(\mathbf{x}) = \mathbf{A}\mathbf{x} - \mathbf{b}, \tag{3.116}$$

where $h\colon \mathbb{R}^m \to \mathbb{R}$ and $\mathbf{g}\colon \mathbb{R}^d \to \mathbb{R}^m$. By Results 3.25 and 3.31, the gradient of h and Jacobian of $\mathbf{g}$ are

$$\nabla h(\mathbf{z}) = 2\mathbf{z} \quad \text{and} \quad \mathbf{J}_\mathbf{g}(\mathbf{x}) = \mathbf{A}. \tag{3.117}$$

Therefore the multivariate chain rule (3.60) gives

$$\nabla f(\mathbf{x}) = \mathbf{J}_\mathbf{g}(\mathbf{x})^\mathsf{T} \nabla h(\mathbf{g}(\mathbf{x})) = \mathbf{A}^\mathsf{T} \cdot 2(\mathbf{A}\mathbf{x} - \mathbf{b}) = 2(\mathbf{A}^\mathsf{T}\mathbf{A}\mathbf{x} - \mathbf{A}^\mathsf{T}\mathbf{b}). \tag{3.118}$$

3.3.3 *Other Useful Formulae*

There is a multitude of other useful formulae whose existence is good to keep in mind. We mention two that recur in statistics and probability (but are not used on this course). Suppose that

$$\mathbf{A}(x) = \begin{bmatrix} a_{1,1}(x) & \cdots & a_{1,d}(x) \\ \vdots & \ddots & \vdots \\ a_{d,1}(x) & \cdots & a_{d,d}(x) \end{bmatrix} \in \mathbb{R}^{d\times d} \tag{3.119}$$

is a matrix whose elements $a_{i,j}(x)$ are functions of $x \in \mathbb{R}$. It is natural to define the derivative of $\mathbf{A}$ by taking element-wise derivatives:

$$\frac{\mathrm{d}}{\mathrm{d}x}\mathbf{A}(x) = \begin{bmatrix} \frac{\mathrm{d}}{\mathrm{d}x}a_{1,1}(x) & \cdots & \frac{\mathrm{d}}{\mathrm{d}x}a_{1,d}(x) \\ \vdots & \ddots & \vdots \\ \frac{\mathrm{d}}{\mathrm{d}x}a_{d,1}(x) & \cdots & \frac{\mathrm{d}}{\mathrm{d}x}a_{d,d}(x) \end{bmatrix}. \tag{3.120}$$

Result 3.34 (DERIVATIVE OF MATRIX INVERSE). If $\mathbf{A}(x) \in \mathbb{R}^{d\times d}$ is invertible, then

$$\frac{\mathrm{d}}{\mathrm{d}x}\mathbf{A}(x)^{-1} = -\mathbf{A}(x)^{-1}\left(\frac{\mathrm{d}}{\mathrm{d}x}\mathbf{A}(x)\right)\mathbf{A}(x)^{-1}. \tag{3.121}$$

Result 3.35 (DERIVATIVE OF LOGARITHMIC DETERMINANT). Let $\mathbf{A}(x) \in \mathbb{R}^{d\times d}$. Then

$$\frac{\mathrm{d}}{\mathrm{d}x}\log(|\mathbf{A}(x)|) = \frac{\mathrm{d}}{\mathrm{d}x}\log(\det(\mathbf{A}(x))) = \mathrm{tr}\left(\mathbf{A}(x)^{-1}\frac{\mathrm{d}}{\mathrm{d}x}\mathbf{A}(x)\right), \tag{3.122}$$

where tr stands for the trace, the sum of diagonal elements of a matrix.

Note that (3.121) and (3.122) are generalisations of the formulae

$$\frac{\mathrm{d}}{\mathrm{d}x}f(x)^{-1} = \frac{\mathrm{d}}{\mathrm{d}x}\frac{1}{f(x)} = -\frac{f'(x)}{f(x)^2} \quad \text{and} \quad \frac{\mathrm{d}}{\mathrm{d}x}\log(f(x)) = \frac{f'(x)}{f(x)}, \tag{3.123}$$

which one obtains from the univariate chain rule (2.35) by writing (a) $f(x)^{-1} = h(g(x))$ for $h(x) = x^{-1}$ and $g(x) = f(x)$ and (b) $\log(f(x)) = h(g(x))$ for $h(x) = \log(x)$ and $g(x) = f(x)$.

3.4 Local Optimisation

We now turn to local optimisation of a d-variate function $f\colon \mathbb{R}^d \to \mathbb{R}$. The case $d = 1$ was discussed in Section 2.5. We again focus on minimisation as maximisation of f is equivalent to minimisation of $-f$. In this section we discuss two topics:

- minimisation of convex functions by finding the point at which their gradient vanishes and
- multivariate gradient descent whose one-dimensional version was given in Algorithm 2.27.

Let us first present multivariate versions of some definitions and results that are familiar to us from Section 2.5.

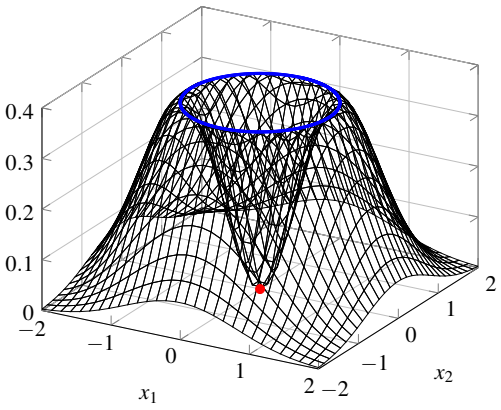

Figure 17: A function with an infinite number of critical points. The function $f(\mathbf{x}) = \exp(-\|\mathbf{x}\|^2)\|\mathbf{x}\|^2$ for $d = 2$ has critical points marked in blue and red. The blue critical points, which satisfy $\|\mathbf{x}^*\|^2 = 1$, are local maximum points while the red critical point $\mathbf{x}^* = \mathbf{0}$ is a local minimum point.

Definition 3.36 (LOCAL AND GLOBAL MINIMA IN $\mathbb{R}^d$). Let $f\colon \mathbb{R}^d \to \mathbb{R}$ be a function. A point $\mathbf{x}^* \in \mathbb{R}^d$ is a

- *global minimum point* of f if $f(\mathbf{x}^*) \leq f(\mathbf{x})$ for all $\mathbf{x} \in \mathbb{R}^d$;
- *local minimum point* of f if there exists $\varepsilon > 0$ such that $f(\mathbf{x}^*) \leq f(\mathbf{x})$ for every $\mathbf{x} \in \mathbb{R}^d$ satisfying $\|\mathbf{x}^* - \mathbf{x}\| \leq \varepsilon$.

The definition of a local minimum requires that $f(\mathbf{x}^*) \leq f(\mathbf{x})$ for all $\mathbf{x}$ in *some* non-empty *closed ball* (suom. *suljettu kuula*) centered at $\mathbf{x}^*$. The closed ball $B(\mathbf{x}^*, \varepsilon)$ of radius $\varepsilon > 0$ centered at $\mathbf{x}^*$ is the set of those $\mathbf{x} \in \mathbb{R}^d$ which are at most ε away from $\mathbf{x}^*$ in the Euclidean distance:

$$B(\mathbf{x}^*, \varepsilon) = \{\mathbf{x} \in \mathbb{R}^d \,:\, \|\mathbf{x}^* - \mathbf{x}\| \leq \varepsilon\}. \tag{3.124}$$

Here the notation reads as "those $\mathbf{x} \in \mathbb{R}^d$ that satisfy the condition to the right of the colon" (i.e., $\|\mathbf{x}^* - \mathbf{x}\| \leq \varepsilon$). Global and local maxima, as well as the concept of extremal points, are defined in the same way as in Section 2.5. In higher dimensions a critical point is a point at which the gradient vanishes [i.e., equals the zero vector $\mathbf{0} = (0, \ldots, 0) \in \mathbb{R}^d$].

Definition 3.37 (CRITICAL POINTS IN $\mathbb{R}^d$). A point $\mathbf{x}^* \in \mathbb{R}^d$ is a *critical point* of a differentiable function $f\colon \mathbb{R}^d \to \mathbb{R}$ if $\nabla f(\mathbf{x}^*) = \mathbf{0}$.

Note that each of the d coordinates of the gradient $\nabla f(\mathbf{x}^*) \in \mathbb{R}^d$ has to be zero if $\mathbf{x}^*$ is to be a critical point.

Example 3.38. Consider the function $f\colon \mathbb{R}^d \to \mathbb{R}$ given by

$$f(\mathbf{x}) = \exp(-\|\mathbf{x}\|^2)\|\mathbf{x}\|^2. \tag{3.125}$$

By the multivariate product and chain rules and Result 3.25 the gradient of this function is

$$\nabla f(\mathbf{x}) = \|\mathbf{x}\|^2 \nabla \exp(-\|\mathbf{x}\|^2) + \exp(-\|\mathbf{x}\|^2)\nabla\|\mathbf{x}\|^2 = 2\exp(-\|\mathbf{x}\|^2)(1 - \|\mathbf{x}\|^2)\,\mathbf{x} \tag{3.126}$$

Since the exponential function is positive, the gradient is zero only if $\mathbf{x} = \mathbf{0}$ or $\|\mathbf{x}\|^2 = 1$. Therefore the origin and every point on the unit sphere

$$\{\mathbf{x} \in \mathbb{R}^d \,:\, \|\mathbf{x}\|^2 = 1\} \tag{3.127}$$

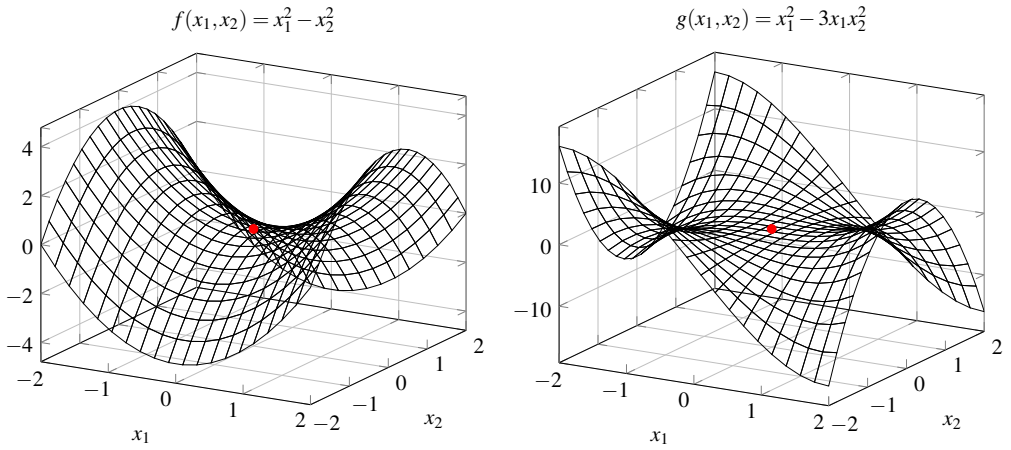

Figure 18: The functions $f(x_1,x_2) = x_1^2 - x_2^2$ and $g(x_1,x_2) = x_1^2 - 3x_1x_2^2$ from Example 3.41 both have the saddle point $\mathbf{x}^* = (0,0)$ (marked in red). These figures may explain why saddle points are called saddle points.

is a critical point of f. That is, f has an infinite number of critical points. From Figure 17 we see that the critical points on the unit sphere are local maximum points while the origin is a local minimum point.

Fermat's theorem (Result 2.23) remains valid in $\mathbb{R}^d$.

Result 3.39 (FERMAT'S THEOREM IN $\mathbb{R}^d$). Every local extremal point $\mathbf{x}^*$ of a differentiable function $f\colon \mathbb{R}^d \to \mathbb{R}$ is a critical point, in that $\nabla f(\mathbf{x}^*) = \mathbf{0}$.

As in the one-dimensional case, there is no implication to the other direction. That is, from $\nabla f(\mathbf{x}^*) = \mathbf{0}$ one cannot deduce that $\mathbf{x}^*$ is a local extremal point. Non-extremal points at which the gradient vanishes are called saddle points.

Definition 3.40 (SADDLE POINT). A point $\mathbf{x}^*$ such that $\nabla f(\mathbf{x}^*) = \mathbf{0}$ is a *saddle point* (suom. *satulapiste*) of $f\colon \mathbb{R}^d \to \mathbb{R}$ if $\mathbf{x}^*$ is *not* a local extremal point of f.

Example 3.41. The functions

$$f(x_1,x_2) = x_1^2 - x_2^2 \quad \text{and} \quad g(x_1,x_2) = x_1^2 - 3x_1x_2^2. \tag{3.128}$$

have the gradients

$$\nabla f(x_1,x_2) = \begin{bmatrix} 2x_1 \\ -2x_2 \end{bmatrix} \quad \text{and} \quad \nabla g(x_1,x_2) = \begin{bmatrix} 2x_1 - 3x_2^2 \\ -6x_1x_2 \end{bmatrix}. \tag{3.129}$$

Both gradients equal $\mathbf{0} = (0,0)$ only when $\mathbf{x} = (x_1,x_2) = \mathbf{x}^* = (0,0)$. Therefore $\mathbf{x}^* = (0,0)$ is the only critical point for both f and g. It turns out that $\mathbf{x}^*$ is a saddle point; see Figure 18.

3.4.1 Convex Functions

Each partial derivative $\partial_j f(\mathbf{x})$ of a function $f\colon \mathbb{R}^d \to \mathbb{R}$ is a function from $\mathbb{R}^d$ to $\mathbb{R}$. We can therefore compute partial derivatives of each $\partial_j f(\mathbf{x})$ to obtain *second-order partial derivatives*

$$\partial_i \partial_j f(\mathbf{x}) = \partial_{x_i} \partial_{x_j} f(\mathbf{x}) = \frac{\partial^2}{\partial x_i \partial x_j} f(\mathbf{x}) = \frac{\partial}{\partial x_i}\left(\frac{\partial}{\partial x_j} f(\mathbf{x})\right) \quad \text{for} \quad i,j = 1,\dots,d. \tag{3.130}$$

If the second partial derivatives are continuous functions, the order of differentiation can be interchanged. That is, we have

$$\partial_i \partial_j f(\mathbf{x}) = \partial_j \partial_i f(\mathbf{x}) \quad \text{or} \quad \frac{\partial}{\partial x_i}\left(\frac{\partial}{\partial x_j} f(\mathbf{x})\right) = \frac{\partial}{\partial x_j}\left(\frac{\partial}{\partial x_i} f(\mathbf{x})\right). \tag{3.131}$$

Example 3.42. Consider the function $f(\mathbf{x}) = f(x_1, x_2) = x_1 x_2 + x_2^3$. The partial derivatives of this function are

$$\partial_1 f(x_1, x_2) = x_2 \quad \text{and} \quad \partial_2 f(x_1, x_2) = x_1 + 3x_2^2. \tag{3.132}$$

There are four second partial derivatives:

$$\partial_1 \partial_1 f(x_1, x_2) = \frac{\partial}{\partial x_1} \partial_1 f(x_1, x_2) = \frac{\partial}{\partial x_1} x_2 = 0, \tag{3.133}$$

$$\partial_1 \partial_2 f(x_1, x_2) = \frac{\partial}{\partial x_1} \partial_2 f(x_1, x_2) = \frac{\partial}{\partial x_1} (x_1 + 3x_2^2) = 1, \tag{3.134}$$

$$\partial_2 \partial_1 f(x_1, x_2) = \frac{\partial}{\partial x_2} \partial_1 f(x_1, x_2) = \frac{\partial}{\partial x_2} x_2 = 1, \tag{3.135}$$

$$\partial_2 \partial_2 f(x_1, x_2) = \frac{\partial}{\partial x_2} \partial_2 f(x_1, x_2) = \frac{\partial}{\partial x_2} (x_1 + 3x_2^2) = 6x_2. \tag{3.136}$$

Note that the *mixed partial derivatives* $\partial_1 \partial_2 f(x_1, x_2)$ and $\partial_2 \partial_1 f(x_1, x_2)$ in (3.134) and (3.135) coincide, as they should by (3.131).

The matrix of that collects all d^2 second partial derivatives of a function $f\colon \mathbb{R}^d \to \mathbb{R}$ is called the Hessian.

Definition 3.43 (HESSIAN). The *Hessian*, or *Hessian matrix*, (suom. *Hessen matriisi*) of a function $f\colon \mathbb{R}^d \to \mathbb{R}$ is the $d \times d$ matrix

$$\mathbf{H}_f(\mathbf{x}) = \nabla^2 f(\mathbf{x}) = \begin{bmatrix} \partial_1 \partial_1 f(\mathbf{x}) & \cdots & \partial_1 \partial_d f(\mathbf{x}) \\ \vdots & \ddots & \vdots \\ \partial_d \partial_1 f(\mathbf{x}) & \cdots & \partial_d \partial_d f(\mathbf{x}) \end{bmatrix} = \begin{bmatrix} \frac{\partial^2}{\partial x_i \partial x_j} f(\mathbf{x}) & \cdots & \frac{\partial^2}{\partial x_i \partial x_j} f(\mathbf{x}) \\ \vdots & \ddots & \vdots \\ \frac{\partial^2}{\partial x_i \partial x_j} f(\mathbf{x}) & \cdots & \frac{\partial^2}{\partial x_i \partial x_j} f(\mathbf{x}) \end{bmatrix}.$$

That is, $[\mathbf{H}_f(\mathbf{x})]_{i,j} = \partial_i \partial_j f(\mathbf{x})$.

The Hessian of a univariate function is simply its second derivative. Because the ith component of the gradient $\nabla f\colon \mathbb{R}^d \to \mathbb{R}^d$ is $\partial_i f$, the Hessian is in fact the transposed Jacobian of the gradient:

$$\mathbf{H}_f(\mathbf{x}) = \mathbf{J}_{\nabla f}(\mathbf{x})^\mathsf{T}. \tag{3.137}$$

Recall that a square matrix $\mathbf{A}$ is symmetric if

$$[\mathbf{A}]_{i,j} = [\mathbf{A}]_{j,i} \quad \text{or equivalently} \quad \mathbf{A} = \mathbf{A}^\mathsf{T}. \tag{3.138}$$

By (3.131) we can interchange the order of differentiation. Therefore the Hessian is a symmetric matrix and we can drop the transpose in (3.137), which yields

$$\mathbf{H}_f(\mathbf{x}) = \mathbf{J}_{\nabla f}(\mathbf{x}). \tag{3.139}$$

Example 3.44. In (3.133)–(3.136) we computed the second partial derivatives of the function $f(x_1,x_2) = x_1x_2 + x_2^3$. The Hessian is formed by placing these partial derivatives in a matrix:

$$\mathbf{H}_f(x_1,x_2) = \begin{bmatrix} \partial_1\partial_1 f(x_1,x_2) & \partial_1\partial_2 f(x_1,x_2) \\ \partial_2\partial_1 f(x_1,x_2) & \partial_2\partial_2 f(x_1,x_2) \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 1 & 6x_2 \end{bmatrix}. \tag{3.140}$$

Note that the Hessian is symmetric.

Example 3.45. Let $f\colon \mathbb{R}^d \to \mathbb{R}$ be given by

$$f(\mathbf{x}) = \|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2 \tag{3.141}$$

for a matrix $\mathbf{A} \in \mathbb{R}^{m\times d}$ (so that $\mathbf{A}\mathbf{x} \in \mathbb{R}^m$) and a vector $\mathbf{b} \in \mathbb{R}^m$. In Example 3.33 we computed that

$$\nabla f(\mathbf{x}) = 2(\mathbf{A}^\mathsf{T}\mathbf{A}\mathbf{x} - \mathbf{A}^\mathsf{T}\mathbf{b}). \tag{3.142}$$

By (3.139) the Hessian is the Jacobian of this gradient. We can apply Result 3.31 to compute the Jacobian:

$$\mathbf{H}_f(\mathbf{x}) = \mathbf{J}_{\nabla f}(\mathbf{x}) = 2\mathbf{A}^\mathsf{T}\mathbf{A}. \tag{3.143}$$

The second derivative of the function $f(x) = (ax-b)^2 = |ax-b|^2$ is

$$f''(x) = \frac{\mathrm{d}}{\mathrm{d}x} f'(x) = \frac{\mathrm{d}}{\mathrm{d}x} 2a(ax-b) = 2a^2. \tag{3.144}$$

Equation (3.143) is just a multivariate generalisation of this formula.

Definition 3.46 (Convex function). A function $f\colon \mathbb{R}^d \to \mathbb{R}$ is *convex* (suom. *konveksi* tai *alaspäin kupera*) if the line segment between any two points on the graph of f lies above the graph:

$$f(t\mathbf{x} + (1-t)\mathbf{y}) \le tf(\mathbf{x}) + (1-t)f(\mathbf{y}) \quad \text{for any } 0 \le t \le 1 \text{ and } \mathbf{x},\mathbf{y} \in \mathbb{R}^d. \tag{3.145}$$

Figure 19 illustrates what the condition that the line segment between any two points should lie above the graph means. A function is *concave* (suom. *konkaavi* tai *ylöspäin kupera*) if the line segment lies *below* the graph, which amounts to having "$\ge$" rather than "$\le$" in (3.145). We do not utilise the line segment definition of convexity in (3.145) but always use the following result that characterises convexity in terms of positive-semidefiniteness of the Hessian. Recall that a symmetric square matrix $\mathbf{A} \in \mathbb{R}^{d\times d}$ with elements $[\mathbf{A}]_{i,j} = a_{i,j}$ is *positive-semidefinite* if

$$\mathbf{z}^\mathsf{T}\mathbf{A}\mathbf{z} = \sum_{i=1}^{d}\sum_{j=1}^{d} z_i z_j a_{i,j} \ge 0 \tag{3.146}$$

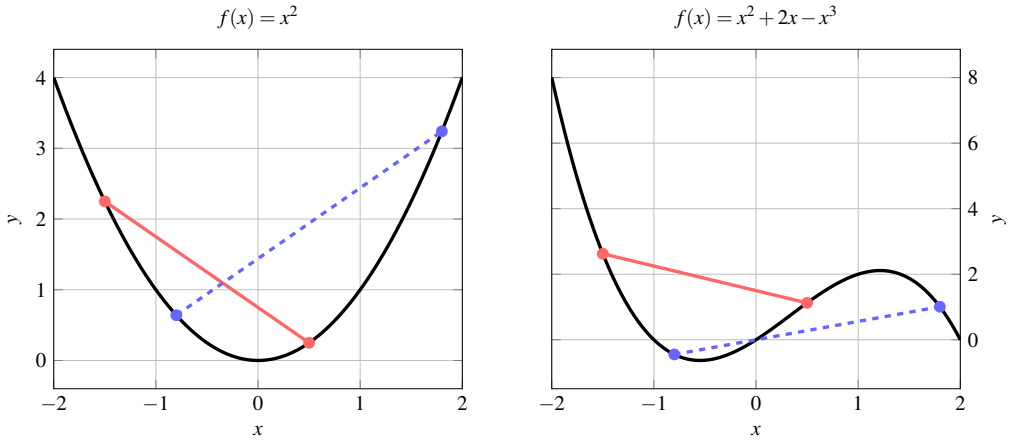

Figure 19: Two functions $f\colon \mathbb{R} \to \mathbb{R}$ and two line segments drawn between points on the graphs of these functions. The function $f(x) = x^2$ on the left is convex because every line segment will lie *above* the graph. The function $f(x) = x^2 + 2x - x^3$ is not convex because there are line segments that lie partially *below* the graph, such as the dashed blue segment.

for every vector $\mathbf{z} \in \mathbb{R}^d$ and *positive-definite* if the inequality in (3.146) is strict (i.e., has ">" instead of "$\geq$") for every $\mathbf{z} \neq \mathbf{0}$.[7] All eigenvalues of a positive-definite matrix are non-negative; the eigenvalues of a positive-definite matrix are positive. All positive-definite matrices are invertible, for otherwise there would be $\mathbf{z} \neq \mathbf{0}$ such that $\mathbf{Az} = \mathbf{0}$, which would violate the requirement that $\mathbf{z}^\mathsf{T}\mathbf{Az} > 0$ for all $\mathbf{z} \neq \mathbf{0}$. Note also that every positive-definite matrix is positive-semidefinite.

Result 3.47 (Convexity via Hessian). A function $f\colon \mathbb{R}^d \to \mathbb{R}$ whose Hessian $\mathbf{H}_f(\mathbf{x})$ exists is convex if and only if $\mathbf{H}_f(\mathbf{x})$ is positive-semidefinite for all $\mathbf{x} \in \mathbb{R}^d$:

$$\mathbf{z}^\mathsf{T}\mathbf{H}_f(\mathbf{x})\,\mathbf{z} \geq 0 \quad \text{for all} \quad \mathbf{z}, \mathbf{x} \in \mathbb{R}^d. \tag{3.147}$$

Example 3.48. The function $f(\mathbf{x}) = \|\mathbf{Ax} - \mathbf{b}\|^2$ from Example 3.45 is convex because in (3.143) we computed that its Hessian is

$$\mathbf{H}_f(\mathbf{x}) = 2\mathbf{A}^\mathsf{T}\mathbf{A}. \tag{3.148}$$

Therefore [recall the tranpose rule $(\mathbf{Ab})^\mathsf{T} = \mathbf{b}^\mathsf{T}\mathbf{A}^\mathsf{T}$]

$$\mathbf{z}^\mathsf{T}\mathbf{H}_f(\mathbf{x})\mathbf{z} = 2\mathbf{z}^\mathsf{T}\mathbf{A}^\mathsf{T}\mathbf{Az} = 2(\mathbf{Az})^\mathsf{T}(\mathbf{Az}) = 2\|\mathbf{Az}\|^2. \tag{3.149}$$

Being the square of a norm, $\|\mathbf{Az}\|^2$ is non-negative for every $\mathbf{z} \in \mathbb{R}^d$. Thus $\mathbf{z}^\mathsf{T}\mathbf{H}_f(\mathbf{x})\mathbf{z} \geq 0$, which means that the Hessian is positive-semidefinite and f convex by Result 3.47.

One of the reasons that convex functions are interesting is that they are easy to minimize.

Result 3.49 (Minimum of a convex function). Every critical point of a differentiable convex function is a global minimum point. That is, if $f\colon \mathbb{R}^d \to \mathbb{R}$ is differentiable and convex and $\mathbf{x}^* \in \mathbb{R}^d$ a point such that $\nabla f(\mathbf{x}^*) = \mathbf{0}$, then $f(\mathbf{x}) \geq f(\mathbf{x}^*)$ for all $\mathbf{x} \in \mathbb{R}^d$.

[7] There is no consistency in terminology: many people use the terms positive-definite and *strictly* positive-definite for positive-semidefinite and positive-definite matrices, respectively.

Result 3.49 implies that convex functions have no saddle points and that all local extremal points are global minimum points. To minimise a convex function it thus suffices to compute the gradient and find a point at which it vanishes. In some cases it is possible to do this analytically.

Result 3.50 (**Minimum of a quadratic function**). Let $f\colon \mathbb{R}^d \to \mathbb{R}$ be a function given by

$$f(\mathbf{x}) = \mathbf{x}^\mathsf{T}\mathbf{B}\mathbf{x} + \mathbf{c}^\mathsf{T}\mathbf{x} + r \quad \text{for } \mathbf{B} \in \mathbb{R}^{d\times d}, \quad \mathbf{c} \in \mathbb{R}^d, \quad \text{and} \quad r \in \mathbb{R}. \tag{3.150}$$

If the matrix $\mathbf{B}$ is symmetric positive-definite, f is convex and has the unique local and global minimum point

$$\mathbf{x}^* = -\tfrac{1}{2}\mathbf{B}^{-1}\mathbf{c}. \tag{3.151}$$

Proof. Let us first compute the Hessian and check that it is positive-semidefinite to verify that f is convex. By (3.85) and (3.88),

$$\nabla f(\mathbf{x}) = 2\mathbf{B}\mathbf{x} + \mathbf{c}. \tag{3.152}$$

From (3.139) and Result 3.31 we get

$$\mathbf{H}_f(\mathbf{x}) = \mathbf{J}_{\nabla f}(\mathbf{x}) = 2\mathbf{B}. \tag{3.153}$$

Because the matrix $\mathbf{B}$ has been assumed positive-definite, the Hessian is positive-definite and thus f is convex by Result 3.47. We may thus use Result 3.49, which states that a point $\mathbf{x}^* \in \mathbb{R}^d$ is a local and global minimum point of f if $\nabla f(\mathbf{x}^*) = \mathbf{0}$. Setting (3.152) to $\mathbf{0}$ gives us the equation

$$2\mathbf{B}\mathbf{x} + \mathbf{c} = \mathbf{0} \tag{3.154}$$

that $\mathbf{x}^*$ is to solve. Equivalently, $\mathbf{B}\mathbf{x}^* = -\frac{1}{2}\mathbf{c}$. As the matrix $\mathbf{B}$ has been assumed positive-definite, it is invertible and we may multiply both sides by $\mathbf{B}^{-1}$ from the left. This gives

$$\mathbf{x}^* = -\tfrac{1}{2}\mathbf{B}^{-1}\mathbf{c}. \tag{3.155}$$

□

Invertibility of $\mathbf{A}$ is crucial in Result 3.50 to ensure uniqueness of the minimum point. If $\mathbf{A}$ were not invertible, Equation (3.154) might have multiple or no solutions. For example, if $\mathbf{A}$ is the zero matrix (positive-semidefinite but not positive-definite), then (3.154) becomes $\mathbf{b} = \mathbf{0}$. If $\mathbf{A}$ is the zero matrix, f is the linear function $f(\mathbf{x}) = \mathbf{b}^\mathsf{T}\mathbf{x} + c$, which has no critical or extremal points unless $\mathbf{b} = \mathbf{0}$; Figure 15 shows this function for $d = 2$, $\mathbf{b} = (1,1) \in \mathbb{R}^2$, and $c = 0$.

Result 3.51 (**Minimum of a norm function**). Let $f\colon \mathbb{R}^d \to \mathbb{R}$ be a function given by

$$f(\mathbf{x}) = \|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2 \tag{3.156}$$

for a matrix $\mathbf{A} \in \mathbb{R}^{m\times d}$ and a vector $\mathbf{b} \in \mathbb{R}^m$. This function convex. Moreover, if $\mathbf{A}$ has linearly independent columns, then f has the unique local and global minimum point

$$\mathbf{x}^* = (\mathbf{A}^\mathsf{T}\mathbf{A})^{-1}\mathbf{A}^\mathsf{T}\mathbf{b}. \tag{3.157}$$

Proof. In Example 3.48 we saw that this function is convex. We may expand the function as

$$\begin{aligned} f(\mathbf{x}) = \|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2 = (\mathbf{A}\mathbf{x} - \mathbf{b})^\mathsf{T}(\mathbf{A}\mathbf{x} - \mathbf{b}) &= \mathbf{x}^\mathsf{T}\mathbf{A}^\mathsf{T}\mathbf{A}\mathbf{x} - \mathbf{x}^\mathsf{T}\mathbf{A}^\mathsf{T}\mathbf{b} - \mathbf{b}^\mathsf{T}\mathbf{A}\mathbf{x} + \mathbf{b}^\mathsf{T}\mathbf{b} \\ &= \mathbf{x}^\mathsf{T}\mathbf{A}^\mathsf{T}\mathbf{A}\mathbf{x} - 2(\mathbf{A}^\mathsf{T}\mathbf{b})^\mathsf{T}\mathbf{x} + \|\mathbf{b}\|^2, \end{aligned} \tag{3.158}$$

where we used the fact that $\mathbf{x}^\mathsf{T}\mathbf{A}^\mathsf{T}\mathbf{b} = (\mathbf{x}^\mathsf{T}\mathbf{A}^\mathsf{T}\mathbf{b})^\mathsf{T} = \mathbf{b}^\mathsf{T}\mathbf{A}\mathbf{x} \in \mathbb{R}$ (since the transpose of a scalar is the scalar itself). Therefore we can apply Result 3.50 with $\mathbf{B} = \mathbf{A}^\mathsf{T}\mathbf{A} \in \mathbb{R}^{d\times d}$ [which is symmetric and, as we verified in (3.149), positive-semidefinite], $\mathbf{c} = -2\mathbf{A}^\mathsf{T}\mathbf{b} \in \mathbb{R}^d$, and $r = \|\mathbf{b}\|^2 \in \mathbb{R}$ to conclude that f has the unique local and global minimum point

$$\mathbf{x}^* = (\mathbf{A}^\mathsf{T}\mathbf{A})^{-1}\mathbf{A}^\mathsf{T}\mathbf{b} \tag{3.159}$$

if $\mathbf{A}^\mathsf{T}\mathbf{A}$ is invertible, which is equivalent to $\mathbf{A}$ having linearly independent columns. □

3.4.2 Least Squares

We now have enough tools to revisit and derive solutions to the ordinary least squares example from Section 1.1 and more general polynomial regression. Suppose that we have outputs $y_1,\dots,y_n \in \mathbb{R}$ corresponding to some pairwise distinct inputs $x_1,\dots,x_n \in \mathbb{R}$. In linear regression we postulate the linear relationship

$$y_i = a + bx_i + \varepsilon_i \quad \text{for} \quad i = 1,\dots,n \tag{3.160}$$

between the outputs and inputs. Here ε_i are residuals that account for noise and other dependencies between the inputs and outputs. To select the two parameters a and b we want to minimize the quadratic loss function

$$L(a,b) = \sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n (y_i - a - bx_i)^2. \tag{3.161}$$

Let us derive the solution to this minimization problem given in (1.3). Define

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \in \mathbb{R}^{n\times 2} \quad \text{and} \quad \mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \in \mathbb{R}^n. \tag{3.162}$$

Let $\mathbf{z} = (a,b)$. We may write the loss function in (3.161) as

$$L(a,b) = L(\mathbf{z}) = \sum_{i=1}^n (y_i - a - bx_i)^2 = \sum_{i=1}^n \left(y_i - \begin{bmatrix} 1 \\ x_i \end{bmatrix}^\mathsf{T} \begin{bmatrix} a \\ b \end{bmatrix} \right)^2 = \|\mathbf{y} - \mathbf{X}\mathbf{z}\|^2 = \|\mathbf{X}\mathbf{z} - \mathbf{y}\|^2.$$

Because the inputs x_i are assumed pairwise distinct (i.e., $x_i \neq x_j$ for all $i \neq j$) the two columns of $\mathbf{X}$ are linearly independent. From this it follows that the matrix $\mathbf{X}^\mathsf{T}\mathbf{X} \in \mathbb{R}^{2\times 2}$ is invertible and positive-definite. We can thus utilise Result 3.51 with $\mathbf{A} = \mathbf{X}$ and $\mathbf{b} = \mathbf{y}$. By (3.157) the minimum point of the loss function is hence

$$\mathbf{z}^* = \begin{bmatrix} a^* \\ b^* \end{bmatrix} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{y}. \tag{3.163}$$

This is called the *normal equation*.

Result 3.52 (LINEAR REGRESSION). The parameters that minimise the quadratic loss function in (3.161) for linear regression are

$$\begin{bmatrix} a^* \\ b^* \end{bmatrix} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{y} \quad \text{with} \quad \mathbf{X} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \in \mathbb{R}^{n\times 2} \quad \text{and} \quad \mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \in \mathbb{R}^n.$$

The resulting linear regression fit is the function

$$F(x) = a^* + b^* x. \tag{3.164}$$

A more flexible approach is to postulate a polynomial relation between the inputs and ouputs:

$$y_i = a_0 + a_1 x_i + \cdots + a_p x_i^p + \varepsilon_i = \sum_{l=0}^{p} (a_l x_i^l + \varepsilon_i) \quad \text{for} \quad i = 1, \ldots, n, \tag{3.165}$$

where $p \in \mathbb{N}$. This is called *polynomial regression* (suom. *polynominen regressio*). There are now $p+1$ parameters $\mathbf{a} = (a_0, a_1, \ldots, a_p) \in \mathbb{R}^{p+1}$ to be selected by minimisation of the loss function

$$L(a_0, a_1, \ldots, a_p) = L(\mathbf{a}) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - (a_0 + a_1 x_i + \cdots + a_p x_i^p))^2. \tag{3.166}$$

This time we introduce the matrix

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^p \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^p \end{bmatrix} \in \mathbb{R}^{n \times (p+1)} \tag{3.167}$$

and write the loss function as

$$L(\mathbf{a}) = \|\mathbf{y} - \mathbf{X}\mathbf{a}\|^2 = \|\mathbf{X}\mathbf{a} - \mathbf{y}\|^2. \tag{3.168}$$

Under the assumption that $\mathbf{X}$ has linearly independent columns we can again make use of Result 3.51 with $\mathbf{A} = \mathbf{X}$ and $\mathbf{b} = \mathbf{y}$ to derive the parameters

$$\mathbf{a}^* = \begin{bmatrix} a_0^* \\ \vdots \\ a_p^* \end{bmatrix} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{y} \in \mathbb{R}^{p+1} \tag{3.169}$$

that minimise the loss function in (3.166).

Result 3.53 (POLYNOMIAL REGRESSION). The parameters that minimise the quadratic loss function in (3.166) for polynomial regression of order $p \in \mathbb{N}$ are

$$\begin{bmatrix} a_0^* \\ \vdots \\ a_p^* \end{bmatrix} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{y} \text{ with } \mathbf{X} = \begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^p \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^p \end{bmatrix} \in \mathbb{R}^{n \times (p+1)} \text{ and } \mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \in \mathbb{R}^n.$$

The resulting polynomial regression fit is the function

$$F_p(x) = a_0^* + a_1^* x + a_2^* x^2 + \cdots + a_p^* x^p. \tag{3.170}$$

Note that the case $p = 1$ in polynomial regression corresponds to linear regression. Linear and polynomial regression fits for a particular dataset are displayed in Figure 20.

3.4.3 Gradient Descent

Let us finally introduce a multivariate version of gradient descent whose univariate implementation was introduced in Algorithm 2.27. Gradient descent exploits the fact that the gradient gives the direction and the rate of fastest increase of a function, so that fastest *decrease* is obtained by heading towards the *negative* gradient.

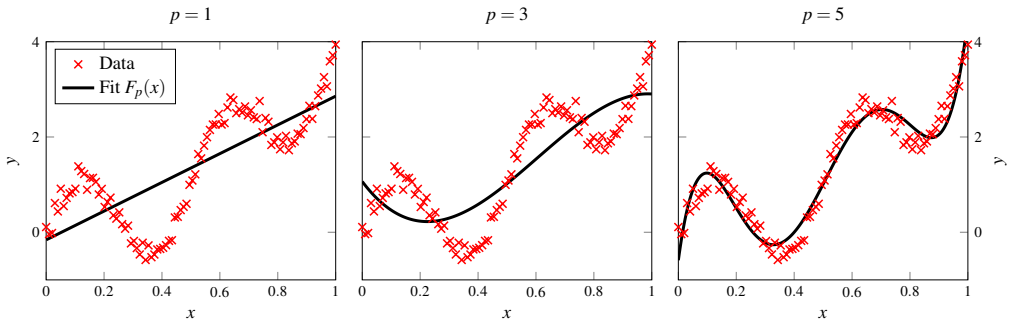

Figure 20: Polynomial regression for a dataset consisting of $n = 100$ pairs of x_i and y_i. The data have a clear increasing and periodic trend. This trend is poorly captured by linear ($p = 1$) and cubic ($p = 3$) regression. Polynomial regression of order $p = 5$ appears to capture most important properties of the dataset. The polynomial regression fit $F_p(x)$ is given in (3.170).

Algorithm 3.54 (GRADIENT DESCENT). The following algorithm attempts to find a local minimum point of a function $f\colon \mathbb{R}^d \to \mathbb{R}$ with gradient $\nabla f\colon \mathbb{R}^d \to \mathbb{R}^d$:

Require: function handle ∇f, initial point $\mathbf{x}_0$, learning rate $\eta > 0$, tolerance $t > 0$, maximum number of iterations $n_{\max} \in \mathbb{N}$

$n \leftarrow 0$
while $\|\nabla f(\mathbf{x}_n)\| \geq t$ **and** $n \leq n_{\max}$ **do**
  $\mathbf{x}_{n+1} \leftarrow \mathbf{x}_n - \eta \nabla f(\mathbf{x}_n)$
  $n \leftarrow n + 1$
end while
return $\mathbf{x}_{n+1}$

That heading towards the negative gradient results in fastest possible decrease is true only *locally*, in the vicinity of the current point $\mathbf{x}_n$. This makes gradient descent (at least in the naive form presented here) a greedy algorithm that at each steps makes a locally optimal choice. However, there is no guarantee that making a sequence of locally optimal choices gives a solution to a minimisation problem that is *globally* good. That is, over multiple steps this may trap gradient descent in a "bad" local minimum point (i.e., one at which f takes a much larger value than at some other local minimum points). Figure 21 demonstrates gradient descent for minimising a bivariate function.

3.5 Multivariate Linearisation

In Section 2.6 we learned about linearisation and Taylor series. While Taylor polynomials and series can be generalised for d-variate functions, their general forms are rarely used (multivariate Taylor polynomials are reviewed at the end of this section for completeness). In this section we focus on multivariate linearisation and second-order Taylor expansion. Recall from (2.62) that the linearisation of a function $f\colon \mathbb{R} \to \mathbb{R}$ around a point x_0 is

$$f(x) \approx L(x) = f(x_0) + f'(x_0)(x - x_0). \tag{3.171}$$

The natural generalisation for multivariate functions $f\colon \mathbb{R}^d \to \mathbb{R}$ is to replace the derivative with the gradient and the product with inner product. This yields the multivariate linear approximation

$$f(\mathbf{x}) \approx L(\mathbf{x}) = f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0) \cdot (\mathbf{x} - \mathbf{x}_0) = f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0)^{\mathsf{T}} (\mathbf{x} - \mathbf{x}_0). \tag{3.172}$$

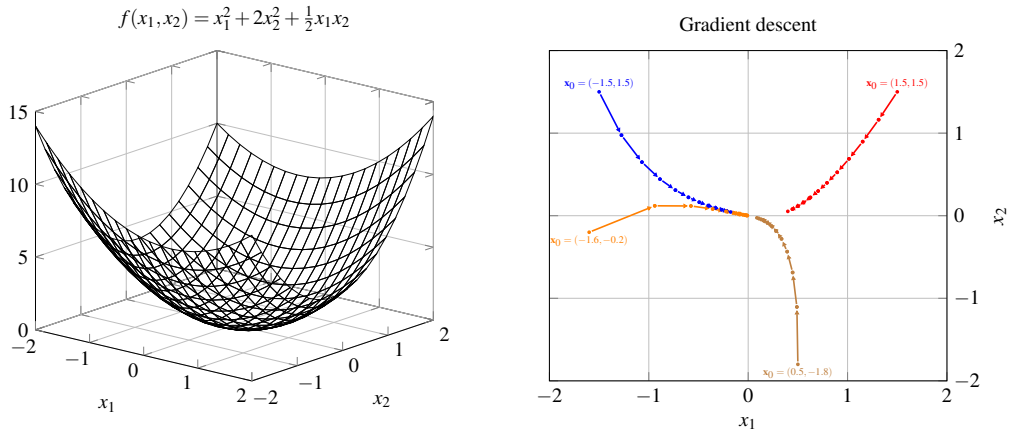

Figure 21: Four gradient descent trajectories (right) for the function $f(x_1,x_2) = x_1^2 + 2x_2^2 + \frac{1}{2}x_1x_2$ (left) with different initial points and learning rates. The function attains its minimum at the origin $\mathbf{x}^* = (0,0)$.

To linearise a function $\mathbf{f}\colon \mathbb{R}^d \to \mathbb{R}^m$ we can linearise each of its m components f_i individually. By recalling that the Jacobian consists of stacked transposed gradients of the component functions, we obtain the linear approximation

$$\mathbf{f}(\mathbf{x}) \approx \mathbf{L}(\mathbf{x}) = \mathbf{f}(\mathbf{x}_0) + \mathbf{J}_\mathbf{f}(\mathbf{x})(\mathbf{x} - \mathbf{x}_0). \tag{3.173}$$

Example 3.55. Consider the function

$$f(x_1,x_2) = \log(1 + x_1^2 + x_2^2) = \log(1 + \|\mathbf{x}\|^2). \tag{3.174}$$

In Example 3.23 we computed that the gradient of this function is

$$\nabla f(\mathbf{x}) = \nabla f(x_1,x_2) = \frac{2}{1 + \|\mathbf{x}\|^2}\,\mathbf{x}. \tag{3.175}$$

The linearisation around a point $\mathbf{x}_0 \in \mathbb{R}^2$ is therefore

$$L(\mathbf{x}) = f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0) \cdot (\mathbf{x} - \mathbf{x}_0) = \log(1 + \|\mathbf{x}_0\|^2) + \frac{2}{1 + \|\mathbf{x}_0\|^2}\,\mathbf{x}_0 \cdot (\mathbf{x} - \mathbf{x}_0). \tag{3.176}$$

Figure 22 shows this linear approximation for $\mathbf{x}_0 = (0.2, 0.4)$.

Example 3.56. Consider the function $\mathbf{f}\colon \mathbb{R}^3 \to \mathbb{R}^2$ given by

$$\mathbf{f}(\mathbf{x}) = \mathbf{f}(x_1,x_2,x_3) = \begin{bmatrix} x_1x_2x_3 - e^{x_1+x_2} \\ x_1^7x_3^3 + 1 \end{bmatrix}. \tag{3.177}$$

In Example 3.16 we computed that the Jacobian of this function is

$$\mathbf{J}_\mathbf{f}(\mathbf{x}) = \begin{bmatrix} x_2x_3 - e^{x_1+x_2} & x_1x_3 - e^{x_1+x_2} & x_1x_2 \\ 7x_1^6x_3^3 & 0 & 3x_1^7x_3^2 \end{bmatrix} \in \mathbb{R}^{2\times 3}. \tag{3.178}$$

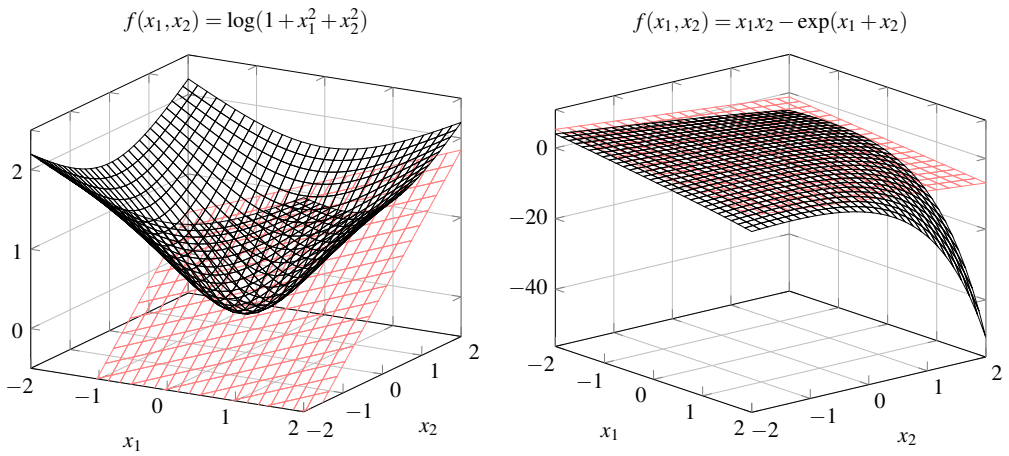

Figure 22: Linear approximations (red) $L(\mathbf{x})$ given in (3.172) for two functions $f\colon \mathbb{R}^2 \to \mathbb{R}$. Linearisation has been done at the point $\mathbf{x}_0 = (0.2, 0.4)$.

The linearisation around a point $\mathbf{x}_0 = (x_{0,1}, x_{0,2}, x_{0,3}) \in \mathbb{R}^3$ is therefore

$$\begin{aligned}\mathbf{L}(\mathbf{x}) &= \mathbf{f}(\mathbf{x}_0) + \mathbf{J}_\mathbf{f}(\mathbf{x})(\mathbf{x}-\mathbf{x}_0) \\ &= \begin{bmatrix} x_1x_2x_3 - e^{x_1+x_2} \\ x_1^7x_3^3 + 1 \end{bmatrix} + \begin{bmatrix} x_{0,2}x_{0,3} - e^{x_{0,1}+x_{0,2}} & x_{0,1}x_{0,3} - e^{x_{0,1}+x_{0,2}} & x_{0,1}x_{0,2} \\ 7x_{0,1}^6x_{0,3}^3 & 0 & 3x_{0,1}^7x_{0,3}^2 \end{bmatrix} (\mathbf{x}-\mathbf{x}_0).\end{aligned}$$

See Figure 22 for a linearisations of a function $f\colon \mathbb{R}^2 \to \mathbb{R}$ closely related to the function in (3.177).

Recall from Definition 2.30 that the Taylor polynomial of degree $n = 2$ for $f\colon \mathbb{R} \to \mathbb{R}$ around a point x_0 is

$$T_2(x) = f(x_0) + f'(x_0)(x-x_0) + \frac{f''(x_0)}{2}(x-x_0)^2. \tag{3.179}$$

The second order Taylor polynomial has a convenient expression for $f\colon \mathbb{R}^d \to \mathbb{R}$ in terms of the gradient ∇f and Hessian $\mathbf{H}_f$. Namely,

$$T_2(\mathbf{x}) = f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0)\cdot(\mathbf{x}-\mathbf{x}_0) + \tfrac{1}{2}(\mathbf{x}-\mathbf{x}_0)^\mathsf{T}\mathbf{H}_f(\mathbf{x}_0)(\mathbf{x}-\mathbf{x}_0) \tag{3.180}$$

is a *quadratic approximation* to f in the vicinity of $\mathbf{x}_0$.

Example 3.57. Consider the function $f(\mathbf{x}) = f(x_1,x_2) = x_1x_2 + x_2^3$. In Examples 3.42 and 3.44 we computed that

$$\nabla f(\mathbf{x}) = \begin{bmatrix} x_2 \\ x_1 + 3x_2^2 \end{bmatrix} \quad \text{and} \quad \mathbf{H}_f(\mathbf{x}) = \begin{bmatrix} 0 & 1 \\ 1 & 6x_2 \end{bmatrix}. \tag{3.181}$$

Therefore the quadratic approximation around $\mathbf{x}_0 = \mathbf{0} = (0,0)$ is

$$\begin{aligned} T_2(\mathbf{x}) &= f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0)\cdot(\mathbf{x}-\mathbf{x}_0) + \tfrac{1}{2}(\mathbf{x}-\mathbf{x}_0)^\mathsf{T}\mathbf{H}_f(\mathbf{x}_0)(\mathbf{x}-\mathbf{x}_0) \\ &= 0 + \mathbf{0}\cdot(\mathbf{x}-\mathbf{0}) + \tfrac{1}{2}(\mathbf{x}-\mathbf{0})^\mathsf{T}\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}(\mathbf{x}-\mathbf{0}) \\ &= x_1x_2. \end{aligned} \tag{3.182}$$

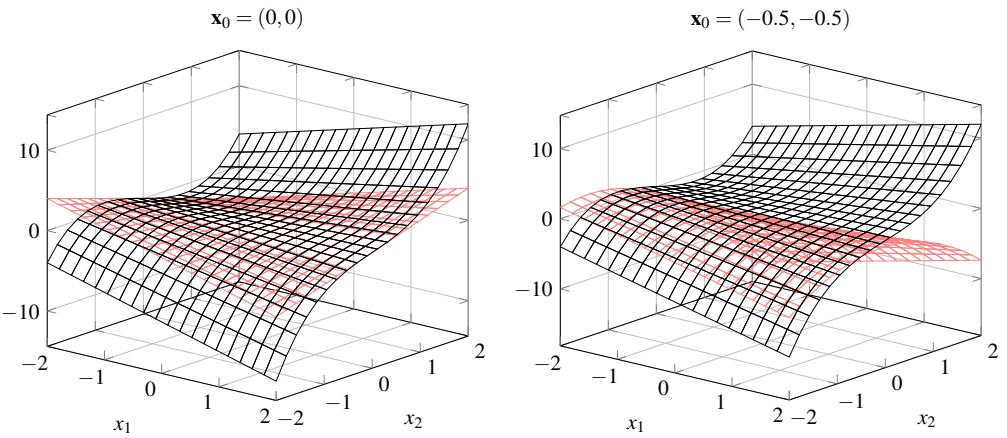

Figure 23: Quadratic approximations (red) $T_2(\mathbf{x})$ given in (3.180) for the function $f(x_1,x_2) = x_1x_2 + x_2^3$ (black). The approximations have been formed at the points $\mathbf{x}_0(0,0)$ and $\mathbf{x}_0 = (-0.5,-0.5)$. Observe that the quadratic approximation around $\mathbf{x}_0 = (0,0)$ is very accurate for $x_2 \approx 0$. This is because in Example 3.57 we saw that $f(\mathbf{x}) = T_2(\mathbf{x}) + x_2^3$. The additional term x_2^3 is close to zero when x_2 is close to zero.

Two quadratic approximations to f are depicted in Figure 23.

For completeness, let us also review multivariate Taylor polynomials in their full generality. With the exception of the linear and quadratic approximation above these polynomials are not used on this course.

Definition 3.58 (MULTI-INDEX). Let $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$ be the set of non-negative integers and $d \in \mathbb{N}$. A *multi-index* (suom. *multi-indeksi*) $\boldsymbol{\alpha}$ is a vector in the d-dimensional set $\mathbb{N}_0^d$:

$$\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_d) \in \mathbb{N}_0^d, \tag{3.183}$$

where $\alpha_i \in \mathbb{N}_0$ for each $i = 1, \ldots, d$.

Definition 3.59 (MULTI-INDEX NOTATION). Let $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{N}_0^d$ be multi-indices and $\mathbf{x} \in \mathbb{R}^d$. The following multi-index notation is used:

$$\boldsymbol{\alpha} \leq \boldsymbol{\beta} \quad \text{if and only if} \quad \alpha_i \leq \beta_i \quad \text{for all} \quad i = 1, \ldots, d, \tag{3.184}$$

$$\boldsymbol{\alpha} \pm \boldsymbol{\beta} = (\alpha_1 \pm \beta_1, \ldots, \alpha_d \pm \beta_d) \in \mathbb{N}_0^d, \tag{3.185}$$

$$|\boldsymbol{\alpha}| = \alpha_1 + \cdots + \alpha_d, \tag{3.186}$$

$$\boldsymbol{\alpha} = \alpha_1! \cdots \alpha_d!, \tag{3.187}$$

$$\mathbf{x}^{\boldsymbol{\alpha}} = x_1^{\alpha_1} \cdots x_d^{\alpha_d}. \tag{3.188}$$

Definition 3.60 (MULTIVARIATE TAYLOR POLYNOMIALS AND SERIES). The Taylor polynomial of degree n of a function $f\colon \mathbb{R}^d \to \mathbb{R}$ around a point $\mathbf{x}_0 \in \mathbb{R}^d$ is the function $T_n\colon \mathbb{R}^d \to \mathbb{R}$ given by

$$T_n(\mathbf{x}) = \sum_{\substack{\boldsymbol{\alpha} \in \mathbb{N}_0^d \\ |\boldsymbol{\alpha}| \leq n}} \frac{\mathrm{D}^{\boldsymbol{\alpha}} f(\mathbf{x}_0)}{\boldsymbol{\alpha}!} (\mathbf{x} - \mathbf{x}_0)^{\boldsymbol{\alpha}}, \tag{3.189}$$

where $\mathrm{D}^{\boldsymbol{\alpha}}$ denotes differentiation α_i times with respect to the ith coordinate:

$$\mathrm{D}^{\boldsymbol{\alpha}} f(\mathbf{x}) = \frac{\partial^{|\boldsymbol{\alpha}|}}{\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}} f(\mathbf{x}). \tag{3.190}$$

The multivariate Taylor series is

$$T_\infty(\mathbf{x}) = \sum_{\boldsymbol{\alpha} \in \mathbb{N}_0^d} \frac{\mathrm{D}^{\boldsymbol{\alpha}} f(\mathbf{x}_0)}{\boldsymbol{\alpha}!} (\mathbf{x} - \mathbf{x}_0)^{\boldsymbol{\alpha}}. \tag{3.191}$$

Observe that the cases $n = 1$ and $n = 2$ reduce to the linear and quadratic approximations in (3.172) and (3.180). For $n = 1$ the Taylor polynomial is

$$\begin{aligned} T_n(\mathbf{x}) = \sum_{\substack{\boldsymbol{\alpha} \in \mathbb{N}_0^d \\ |\boldsymbol{\alpha}| \leq 1}} \frac{\mathrm{D}^{\boldsymbol{\alpha}} f(\mathbf{x}_0)}{\boldsymbol{\alpha}!} (\mathbf{x} - \mathbf{x}_0)^{\boldsymbol{\alpha}} &= \frac{\mathrm{D}^{\mathbf{0}} f(\mathbf{x}_0)}{\mathbf{0}!} (\mathbf{x} - \mathbf{x}_0)^{\mathbf{0}} + \sum_{i=1}^{d} \frac{\partial_i f(\mathbf{x}_0)}{1!} (x_i - x_{0,i})^1 \\ &= f(\mathbf{x}_0) + \sum_{i=1}^{d} [\partial_i f(\mathbf{x}_0)](x_i - x_{0,i}) \\ &= f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0) \cdot (\mathbf{x} - \mathbf{x}_0), \end{aligned} \tag{3.192}$$

which is (3.172). When $n = 2$, the Taylor polynomial has the additional terms

$$\begin{aligned} \sum_{\substack{\boldsymbol{\alpha} \in \mathbb{N}_0^d \\ |\boldsymbol{\alpha}| = 2}} \frac{\mathrm{D}^{\boldsymbol{\alpha}} f(\mathbf{x}_0)}{\boldsymbol{\alpha}!} (\mathbf{x} - \mathbf{x}_0)^{\boldsymbol{\alpha}} &= \sum_{i=1}^{d} \sum_{j=1}^{i-1} \frac{\partial_i \partial_j f(\mathbf{x}_0)}{1!1!} (x_i - x_{0,i})^1 (x_j - x_{0,j})^1 + \sum_{i=1}^{d} \frac{\partial_i \partial_i f(\mathbf{x}_0)}{2!} (x_i - x_{0,i})^2 \\ &= \frac{1}{2} \sum_{i=1}^{d} \sum_{j=1}^{d} [\partial_i \partial_j f(\mathbf{x}_0)](x_i - x_{0,i})(x_j - x_{0,j}) \\ &= \tfrac{1}{2} (\mathbf{x} - \mathbf{x}_0)^{\mathsf{T}} \mathbf{H}_f(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0). \end{aligned}$$

Summing this up with (3.192) gives (3.180).

3.6 Multivariate Integration by Substitution

Recall from Section 2.7 that the integral $\int_a^b f(x)\,\mathrm{d}x$ of a univariate function f represents an area between the graph of f and the x-axis. Similarly, the integral of a function $f\colon \mathbb{R}^d \to \mathbb{R}$ represents the *volume* of the region under the surface defined by f and bounded by the integration domain. If the integration domain is a rectangle $A = [a_1, b_1] \times \cdots \times [a_d, b_d]$ for $a_i < b_i$, the integral can be expressed as a multiple integral:

$$\int_A f(\mathbf{x})\,\mathrm{d}\mathbf{x} = \int_{a_d}^{b_d} \cdots \int_{a_1}^{b_1} f(x_1, \ldots, x_d)\,\mathrm{d}x_1 \cdots \mathrm{d}x_d. \tag{3.193}$$

We focus on multivariate integration by affine substitution. The general multivariate form of integration by substitution (the univariate version of which was introduced in Result 2.39) is given by the following result. In this result $\mathbf{g}(A)$ stands for the *image* (suom. *kuvajoukko* tai *arvojoukko*) of a set A under the function $\mathbf{g}$. The image is the set of points formed by applying $\mathbf{g}$ to points in A:

$$\mathbf{g}(A) = \{\mathbf{g}(\mathbf{x}) : \mathbf{x} \in A\}. \tag{3.194}$$

Result 3.61 (INTEGRATION BY SUBSTITUTION). Let $A \subset \mathbb{R}^d$ be a subset of $\mathbb{R}^d$ and $\mathbf{g}\colon A \to \mathbb{R}^d$ an *injective* function [i.e., $\mathbf{g}(\mathbf{x}) \neq \mathbf{g}(\mathbf{y})$ if $\mathbf{x} \neq \mathbf{y}$] such that $\mathbf{J}_\mathbf{g}(\mathbf{x}) \neq \mathbf{0}_{d\times d}$ for all $\mathbf{x} \in A$. Then

$$\int_{\mathbf{g}(A)} f(\mathbf{x})\,\mathrm{d}\mathbf{x} = \int_A f(\mathbf{g}(\mathbf{y}))|\det(\mathbf{J}_\mathbf{g}(\mathbf{y}))|\,\mathrm{d}\mathbf{y}. \tag{3.195}$$

Observe that integration by substitution involves the Jacobian determinant from Definition 3.13. The affine [i.e., $\mathbf{g}(\mathbf{x}) = \mathbf{A}\mathbf{x} - \mathbf{b}$] version of integration by substitution is simpler, particularly when integration is done over the entire Euclidean space $\mathbb{R}^d$.

Result 3.62 (INTEGRATION BY AFFINE SUBSTITUTION). Let $\mathbf{A} \in \mathbb{R}^{d\times d}$ be an invertible matrix and $\mathbf{b} \in \mathbb{R}^d$ a vector. Then

$$\int_{\mathbb{R}^d} f(\mathbf{A}\mathbf{x} - \mathbf{b})\,\mathrm{d}\mathbf{x} = \frac{1}{|\det(\mathbf{A})|}\int_{\mathbb{R}^d} f(\mathbf{x})\,\mathrm{d}\mathbf{x}. \tag{3.196}$$

Proof. Select $\mathbf{g}(\mathbf{x}) = \mathbf{A}\mathbf{x} - \mathbf{b}$ and $A = \mathbb{R}^d$ in Result 3.61. By Result 3.31, the Jacobian is $\mathbf{J}_\mathbf{g}(\mathbf{y}) = \mathbf{A}$. Because the matrix $\mathbf{A}$ has been assumed invertible, the image $\mathbf{g}(A) = \mathbf{g}(\mathbb{R}^d)$ is $\mathbb{R}^d$. To see this, take any $\mathbf{y} \in \mathbb{R}^d$. By selecting the input $\mathbf{x} = \mathbf{A}^{-1}(\mathbf{y} + \mathbf{b})$ we get

$$\mathbf{g}(\mathbf{x}) = \mathbf{A}\mathbf{A}^{-1}(\mathbf{y} + \mathbf{b}) - \mathbf{b} = \mathbf{y}, \tag{3.197}$$

which means that for each $\mathbf{y} \in \mathbb{R}^d$ there is $\mathbf{x} \in \mathbb{R}^d$ such that $\mathbf{g}(\mathbf{x}) = \mathbf{y}$. Therefore $\mathbf{g}(A) = \mathbf{g}(\mathbb{R}^d) = \mathbb{R}^d$. Equation (3.195) is thus

$$\int_{\mathbb{R}^d} f(\mathbf{x})\,\mathrm{d}\mathbf{x} = \int_{\mathbb{R}^d} f(\mathbf{A}\mathbf{y} - \mathbf{b})|\det(\mathbf{A})|\,\mathrm{d}\mathbf{y} = |\det(\mathbf{A})|\int_{\mathbb{R}^d} f(\mathbf{A}\mathbf{y} - \mathbf{b})\,\mathrm{d}\mathbf{y}. \tag{3.198}$$

Dividing both sides by $|\det(\mathbf{A})|$ gives (3.196). □

Example 3.63. It holds that

$$\int_{\mathbb{R}} \frac{1}{1+x^2}\,\mathrm{d}x = \int_{-\infty}^{\infty} \frac{1}{1+x^2}\,\mathrm{d}x = \pi. \tag{3.199}$$

Suppose that we want to compute the integral

$$\int_{\mathbb{R}^2} f(\mathbf{x})\,\mathrm{d}\mathbf{x} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x_1, x_2)\,\mathrm{d}x_1\,\mathrm{d}x_2 \tag{3.200}$$

of the function

$$f(x_1, x_2) = \frac{1}{1 + (x_1 + 3x_2 - 2)^2} \times \frac{1}{1 + (2x_1 + x_2 - 1)^2}. \tag{3.201}$$

By introducing

$$\mathbf{A} = \begin{bmatrix} 1 & 3 \\ 2 & 1 \end{bmatrix} \in \mathbb{R}^{2\times 2} \quad \text{and} \quad \mathbf{b} = \begin{bmatrix} 2 \\ 1 \end{bmatrix} \in \mathbb{R}^2 \tag{3.202}$$

we can write this function as

$$f(\mathbf{x}) = f(x_1, x_2) = h(\mathbf{A}\mathbf{x} - \mathbf{b}) \quad \text{for} \quad h(\mathbf{x}) = \frac{1}{1+x_1^2} \times \frac{1}{1+x_2^2}. \tag{3.203}$$

Applying Result 3.62 to h then gives

$$\int_{\mathbb{R}^2} f(\mathbf{x})\,\mathrm{d}\mathbf{x} = \int_{\mathbb{R}^2} h(\mathbf{A}\mathbf{x}-\mathbf{b})\,\mathrm{d}\mathbf{x} = \frac{1}{|\det(\mathbf{A})|}\int_{\mathbb{R}^2} h(\mathbf{x})\,\mathrm{d}\mathbf{x}. \tag{3.204}$$

By (3.199), the integral of h is

$$\begin{aligned}\int_{\mathbb{R}^2} h(\mathbf{x})\,\mathrm{d}\mathbf{x} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(x_1,x_2)\,\mathrm{d}x_1\,\mathrm{d}x_2 &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{1}{1+x_1^2}\times\frac{1}{1+x_2^2}\,\mathrm{d}x_1\,\mathrm{d}x_2 \\ &= \int_{-\infty}^{\infty}\left(\int_{-\infty}^{\infty}\frac{1}{1+x_1^2}\,\mathrm{d}x_1\right)\frac{1}{1+x_2^2}\,\mathrm{d}x_2 \\ &= \pi\int_{-\infty}^{\infty}\frac{1}{1+x_2^2}\,\mathrm{d}x_2 \\ &= \pi^2.\end{aligned} \tag{3.205}$$

By the formula (3.36) for the determinant of a 2×2 matrix, the determinant of $\mathbf{A}$ is

$$\det(\mathbf{A}) = 1\cdot 1 - 3\cdot 2 = -5. \tag{3.206}$$

Consequently,

$$\int_{\mathbb{R}^2} f(\mathbf{x})\,\mathrm{d}\mathbf{x} = \frac{1}{|\det(\mathbf{A})|}\int_{\mathbb{R}^2} h(\mathbf{x})\,\mathrm{d}\mathbf{x} = \frac{\pi^2}{|-5|} = \frac{\pi^2}{5}. \tag{3.207}$$

Result 3.64 (Gaussian substitution). Suppose that $\mathbf{\Sigma}\in\mathbb{R}^{d\times d}$ is a symmetric positive-definite matrix and $\boldsymbol{\mu}\in\mathbb{R}^d$ a vector. Then

$$\frac{1}{\det(\mathbf{\Sigma}^{1/2})}\int_{\mathbb{R}^d}\exp(-\tfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathsf{T}}\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}))\,\mathrm{d}\mathbf{x} = \int_{\mathbb{R}^d}\exp(-\tfrac{1}{2}\|\mathbf{x}\|^2)\,\mathrm{d}\mathbf{x}, \tag{3.208}$$

where $\mathbf{\Sigma}^{1/2}\in\mathbb{R}^{d\times d}$ is the *matrix square root* of $\mathbf{\Sigma}$, the unique symmetric positive-definite matrix such that $\mathbf{\Sigma}^{1/2}\mathbf{\Sigma}^{1/2} = \mathbf{\Sigma}$.

Proof. Because $\mathbf{\Sigma}$ is symmetric positive-definite, there is a unique symmetric positive-definite matrix $\mathbf{B} = \mathbf{\Sigma}^{1/2}\in\mathbb{R}^{d\times d}$ such that $\mathbf{\Sigma} = \mathbf{B}\mathbf{B}$. By utilising this matrix and its symmetricity we may write

$$\begin{aligned}(\mathbf{x}-\boldsymbol{\mu})^{\mathsf{T}}\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) = (\mathbf{x}-\boldsymbol{\mu})^{\mathsf{T}}\mathbf{B}^{-1}\mathbf{B}^{-1}(\mathbf{x}-\boldsymbol{\mu}) &= (\mathbf{B}^{-1}(\mathbf{x}-\boldsymbol{\mu}))^{\mathsf{T}}(\mathbf{B}^{-1}(\mathbf{x}-\boldsymbol{\mu})) \\ &= \|\mathbf{B}^{-1}(\mathbf{x}-\boldsymbol{\mu})\|^2,\end{aligned}$$

so that

$$\begin{aligned}\int_{\mathbb{R}^d}\exp(-\tfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathsf{T}}\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}))\,\mathrm{d}\mathbf{x} &= \int_{\mathbb{R}^d}\exp(-\tfrac{1}{2}\|\mathbf{B}^{-1}(\mathbf{x}-\boldsymbol{\mu})\|^2)\,\mathrm{d}\mathbf{x} \\ &= \int_{\mathbb{R}^d}\exp(-\tfrac{1}{2}\|\mathbf{B}^{-1}\mathbf{x}-\mathbf{B}^{-1}\boldsymbol{\mu}\|^2)\,\mathrm{d}\mathbf{x}.\end{aligned} \tag{3.209}$$

Result 3.62 with $f(\mathbf{x}) = \exp(-\frac{1}{2}\|\mathbf{x}\|^2)$, $\mathbf{A} = \mathbf{B}^{-1}$, and $\mathbf{b} = \mathbf{B}^{-1}\boldsymbol{\mu}$ then gives

$$\begin{aligned}\int_{\mathbb{R}^d} \exp(-\tfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}))\,\mathrm{d}\mathbf{x} &= \int_{\mathbb{R}^d} f(\mathbf{A}\mathbf{x}-\mathbf{b})\,\mathrm{d}\mathbf{x} \\ &= \frac{1}{|\det(\mathbf{A})|}\int_{\mathbb{R}^d} f(\mathbf{x})\,\mathrm{d}\mathbf{x} \\ &= \frac{1}{|\det(\mathbf{B}^{-1})|}\int_{\mathbb{R}^d} \exp(-\tfrac{1}{2}\|\mathbf{x}\|^2)\,\mathrm{d}\mathbf{x} \\ &= \det(\boldsymbol{\Sigma}^{1/2})\int_{\mathbb{R}^d} \exp(-\tfrac{1}{2}\|\mathbf{x}\|^2)\,\mathrm{d}\mathbf{x},\end{aligned} \tag{3.210}$$

where on the last line we used the fact that $\det(\mathbf{A}^{-1}) = \det(\mathbf{A})^{-1}$ (i.e., determinant of the inverse equals inverse of the determinant). Dividing by $\det(\boldsymbol{\Sigma}^{1/2})$ yields (3.208). □

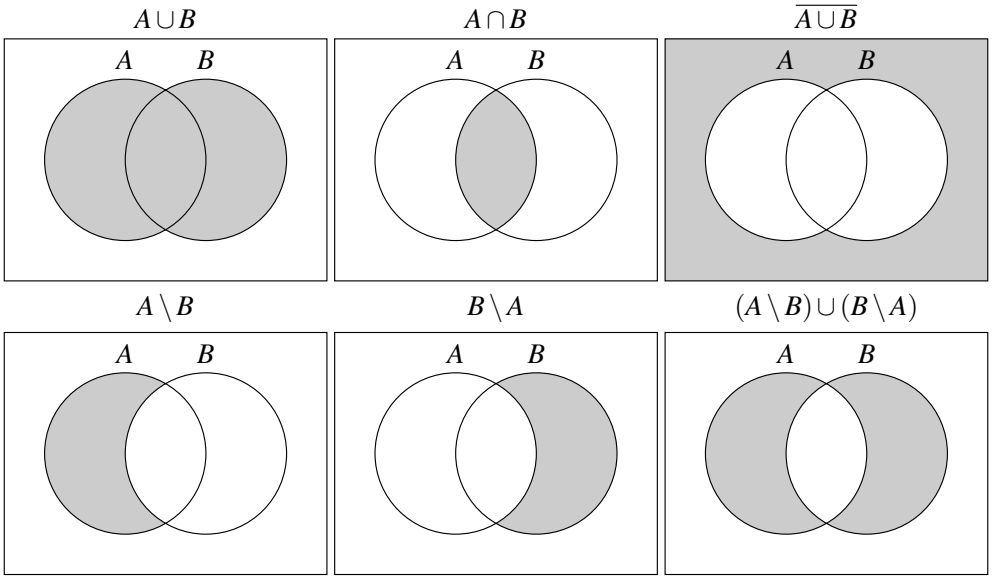

Figure 24: Venn diagrams illustrating six sets (in grey) formed out of sets A and B.

4 Discrete Probability

In this section we learn about basic concepts of *discrete probability theory*. This section is essentially an abridgement of parts of Chapters 1–4 and 6 in the lecture notes *Introduction to Probability* by Ville Hyvönen, Patrik Lauha, and Topias Tolonen (April 2020). These lectures notes are available on Moodle under Week 5 in both English and Finnish.

4.1 A Little Bit of Set Theory

Probability theory is built upon set theory. We therefore begin with a brief recap of some basic notions and notations from set theory.

Definition 4.1 (SET THEORETIC NOTATION). A *set* is a non-ordered collection of some objects called *elements* of the set. The following notations are used:

$x \in A$	means	x is an element of set A
$x \notin A$	means	x is *not* an element of set A
$A \subset B$	means	set A is a *subset* of set B (every element of A is also an element of B)
$A = B$	means	sets A and B are equal (they contain the same elements)
$A \cup B$	means	*union* of sets A and B (set of x such that $x \in A$ *or* $x \in B$)
$A \cap B$	means	*intersection* of sets A and B (set of x such that $x \in A$ *and* $x \in B$)
$A \setminus B$	means	*difference* of sets A and B (set of x such that $x \in A$ but $x \notin B$)
$\overline{A}$	means	*complement* of set A (those elements "under consideration" *not* in A)
$\emptyset$	means	*empty set*, the set which contains no elements

Venn diagrams in Figure 24 illustrate various set operations. The elements of a set can be *anything*, not just real numbers of d-dimensional vectors. Note that we always have

$$\emptyset \subset A \quad \text{and} \quad A \subset A. \tag{4.1}$$

Example 4.2. Let a, b, and c be some distinct objects and denote $A = \{a,b,c\}$. Sets are non-ordered, which means that we could alternatively write $A = \{b,c,a\}$, $A = \{c,b,a\}$, and so on; all these are the same set. Then we have $a,b,c \in A$ but $d \notin A$ if d is not equal to a, b, or c. The sets $\{a,b\}$ and $\{c\}$ are examples of subsets of A. That is, $\{a,b\} \subset A$ and $\{c\} \subset A$. Note that $\{c\}$ is not the *object* c but a *set* whose only element c is. This means that $\emptyset$ and $\{\emptyset\}$ are not the same set; the former is a set that contains no elements while the latter is a set that contains the empty set. Therefore $\emptyset \subset \{\emptyset\}$ (because the empty set is a subset of every set) and $\emptyset \in \{\emptyset\}$.

Repeated elements are counted as one. That is, the sets $A = \{a,b,c\}$ and $B = \{a,a,b,c\}$ are equal.

Result 4.3 (SET OPERATIONS). Any sets A, B, and C have the following properties.

- *Commutativity* (suom. *vaihdannaisuus*):

$$A \cup B = B \cup A \quad \text{and} \quad A \cap B = B \cap A. \tag{4.2}$$

- *Associativity* (suom. *liitännäisyys*):

$$(A \cup B) \cup C = A \cup (B \cup C) = A \cup B \cup C \quad \text{and} \quad (A \cap B) \cap C = A \cap (B \cap C) = A \cap B \cap C.$$

- *Distributive laws* (suom. *osittelulait*):

$$(A \cup B) \cap C = (A \cap C) \cup (B \cap C) \quad \text{and} \quad (A \cap B) \cup C = (A \cup C) \cap (B \cup C). \tag{4.3}$$

- *De Morgan's laws*:

$$\overline{A \cup B} = \overline{A} \cap \overline{B} \quad \text{and} \quad \overline{A \cap B} = \overline{A} \cup \overline{B}. \tag{4.4}$$

  De Morgan's laws hold for any number of sets. For example, $\overline{A \cup B \cup C} = \overline{A} \cap \overline{B} \cap \overline{C}$.

Given an infinite number of sets A_i for $i \in \mathbb{N}$, we write

$$\bigcup_{i=1}^{\infty} A_i \quad \text{and} \quad \bigcap_{i=1}^{\infty} A_i \tag{4.5}$$

for their union and intersection, respectively. The union above is the set of all elements that in *some* A_i, while the intersection is the set of all elements that are in *every* A_i.

Example 4.4. For each $i \in \mathbb{N}$, define the set

$$A_i = \{i, i+1, \dots, \} \subset \mathbb{N} \tag{4.6}$$

consisting of natural numbers larger than or equal to i. Then

$$\bigcup_{i=1}^{\infty} A_i = \mathbb{N} \quad \text{and} \quad \bigcap_{i=1}^{\infty} A_i = \emptyset. \tag{4.7}$$

The union equals $\mathbb{N}$ because $A_1 = \mathbb{N}$ while the intersection is empty because $n \notin A_{n+1}$ for any $n \in \mathbb{N}$.

Definition 4.5 (POWER SET). The *power set* (suom. *potenssijoukko*) $\mathcal{P}(A)$ of a set A is the set of all subsets of A:

$$\mathcal{P}(A) = \{B : B \subset A\}. \tag{4.8}$$

Example 4.6. Power sets are sets of sets. Consider the set $A = \{a, b, c\}$ consisting of three distinct objects a, b, and c. The power set of A is

$$\mathcal{P}(A) = \{\emptyset, \{a\}, \{b\}, \{c\}, \{a, b\}, \{a, c\}, \{b, c\}, \{a, b, c\}\}. \tag{4.9}$$

Note that both the empty set $\emptyset$ and the set itself, $A = \{a, b, c\}$, are elements of the power set $\mathcal{P}(A)$. Also note that the elements a, b, and c of A are *not* elements of $\mathcal{P}(A)$. Rather, it is the *singleton sets* $\{a\}$, $\{b\}$, and $\{c\}$ that are elements of the power set.

Result 4.7 (CARDINALITY OF THE POWER SET). If a set A has n elements (written $|A| = n$), then the power set $\mathcal{P}(A)$ has 2^n elements.

4.2 Probability Space

We can now provide a mathematical definition of probability. The purpose of this mathematical construction is to model randomness occurring in real world. Our goal is to use mathematics to model *random experiments*, such as tossing a coin or rolling dice.

Definition 4.8 (SAMPLE SPACE). The set Ω of all possible outcomes of a random experiment is called the *sample space* (suom. *perusjoukko*). The elements $\omega \in \Omega$ are called *outcomes* (suom. *alkeistapaus*).

Example 4.9. Suppose one rolls a single 6-sided die and records the value on its upper side. The sample space is the set of possible values on the side of the die:

$$\Omega = \{1, 2, 3, 4, 5, 6\}. \tag{4.10}$$

Each of these six values is an outcome, ω.

A set is *countable* (suom. *numeroituva*) if one can list all its elements. That is, a countable set can be written as

$$\Omega = \{\omega_i : i \in \mathbb{N}\} \tag{4.11}$$

for some elements ω_i. A finite set $\Omega = \{\omega_1, \ldots, \omega_n\}$ is countable as we can simply set $\omega_i = \omega_n$ for $i \geq n$ in (4.11). That is,

$$\{\omega_1, \ldots, \omega_n\} = \{\omega_1, \ldots, \omega_{n-1}, \omega_n, \omega_n, \omega_n, \ldots\} = \{\omega_1, \ldots, \omega_{n-1}, \omega_n, \omega_{n+1}, \ldots\}, \tag{4.12}$$

where $\omega_{n+1}, \omega_{n+2}, \ldots$ are defined to be equal to ω_n. The sets in (4.12) are equal because repeating an element does not alter a set. The sets $\mathbb{N}$, $\mathbb{Z}$, and $\mathbb{Q}$ are also countable. However, the set $\mathbb{R}$ of real numbers is not countable, nor are non-empty intervals $[a, b] \subset \mathbb{R}$.

Definition 4.10 (EVENT). The subsets of a countable sample space $\Omega = \{\omega_i : i \in \mathbb{N}\}$ are called *events* (suom. *tapahtuma*). The set of all events is the power set $\mathcal{P}(\Omega)$ of the sample space.

Remark 4.20 describes certain difficulties one encounters when working with uncountable sample spaces.

Example 4.11. If outcomes of a random experiment are rolls of a single 6-sided die, we saw in Example 4.9 that the sample space is $\Omega = \{1,2,3,4,5,6\}$. The set of all events is the power set of Ω. By Result 4.7, there are $2^6 = 64$ different events. We can assign simple meaning to, for example, the following events:

$$\begin{aligned}
\text{“the roll is even”} &\quad \text{is the event} \quad \{2,4,6\} \in \mathcal{P}(\Omega),\\
\text{“the roll is odd”} &\quad \text{is the event} \quad \{1,3,5\} \in \mathcal{P}(\Omega),\\
\text{“the roll is at most 2”} &\quad \text{is the event} \quad \{1,2\} \in \mathcal{P}(\Omega),\\
\text{“the roll is 5”} &\quad \text{is the event} \quad \{5\} \in \mathcal{P}(\Omega),\\
\text{“the roll is 7”} &\quad \text{is the event} \quad \emptyset \in \mathcal{P}(\Omega).
\end{aligned}$$

Example 4.12. Suppose that our random experiment is picking a card from a standard 52-card deck. The sample space Ω is the set consisting of the 52 different cards and the set of all events is the power set $\mathcal{P}(\Omega)$. By Result 4.7, there are 2^{52} possible events.

Example 4.13. Suppose that we roll *two* 6-sided dice and record the rolls. Now the sample space is the set of all *pairs* (i, j) for $i, j = 1,\dots,6$:

$$\Omega = \{(i,j) : i,j = 1,\dots,6\}, \tag{4.13}$$

where i indicates the roll of the first die and j that of the second. Each of these pairs is an outcome. There are a total of $6^2 = 36$ outcomes. By Result 4.7 there are thus 2^{36} events. Let us consider some of these events:

$$\begin{aligned}
\text{“one of the rolls is 4”} &\quad \text{is} \quad A_1 = \{(4,1),(4,2),\dots,(4,6),(1,4),(2,4),\dots,(6,4)\},\\
\text{“both rolls are at most 2”} &\quad \text{is} \quad A_2 = \{(1,1),(1,2),(2,1),(2,2)\},\\
\text{“the sum of rolls is 10”} &\quad \text{is} \quad A_3 = \{(4,6),(5,5),(6,4))\},\\
\text{“both rolls are even”} &\quad \text{is} \quad A_4 = \{(2,2),(2,4),(2,6),(4,2),\dots,(6,6)\}.
\end{aligned}$$

We can express additional events using set theoretic operations:

$$\begin{aligned}
\text{“one of the rolls is 4 and the other is even”} &\quad \text{is} \quad A_1 \cap A_4,\\
\text{“the rolls are even and sum to 10”} &\quad \text{is} \quad A_3 \cap A_4 = \{(4,6),(6,4)\},\\
\text{“the rolls are } \leq 2 \text{ and one of them is 4”} &\quad \text{is} \quad A_2 \cap A_1 = \emptyset,\\
\text{“one of the rolls is 4 or both rolls are even”} &\quad \text{is} \quad A_1 \cup A_4,\\
\text{“the rolls are at most 2 or sum to 10”} &\quad \text{is} \quad A_2 \cup A_3,\\
\text{“the rolls are } \leq 2 \text{ and at least one of them is odd”} &\quad \text{is} \quad A_2 \setminus A_4 = \{(1,1),(1,2),(2,1)\},\\
\text{“both rolls are even but they do not sum to 10”} &\quad \text{is} \quad A_4 \setminus A_3,\\
\text{“at least one of the rolls is odd”} &\quad \text{is} \quad \overline{A}_4 = \Omega \setminus A_4,\\
\text{“neither roll is 4”} &\quad \text{is} \quad \overline{A}_1 = \Omega \setminus A_1.
\end{aligned}$$

Definition 4.14 (PROBABILITY). A function $P\colon \mathcal{P}(\Omega) \to [0,1]$ is a *probability* (suom. *todennäköisyys*) if it satisfies the following *axioms of probability* (suom. *todennäköisyysaksioomat*):

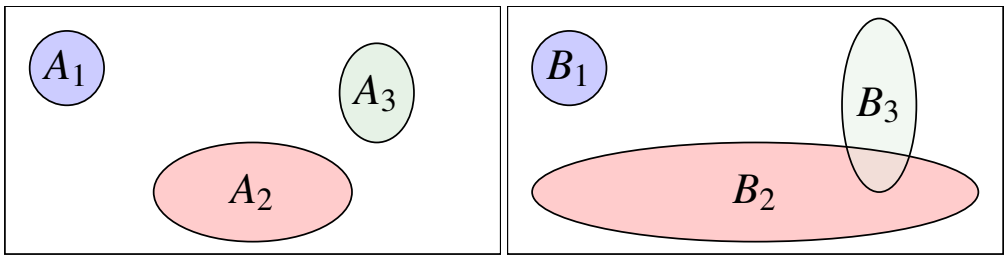

Figure 25: The sets A_1, A_2, and A_3 on the left are disjoint. Thus the probability $P(A_1 \cup A_2 \cup A_3)$, which one can think of as the total area of the coloured regions, equals the sum of the areas of individual coloured regions, $P(A_1) + P(A_2) + P(A_3)$. The sets B_1, B_2, and B_3 on the right are *not* disjoint and therefore the area of the coloured region need not equal the sum of individual areas.

(A1) $P(\Omega) = 1$.

(A2) The function is *additive*, in that

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i) \tag{4.14}$$

for any events such that $A_i \cap A_j = \emptyset$ for all $i \neq j$ (such sets are called *disjoint*).

We say that P is a probability *on* Ω.

One can think of probability as a function that defines the volume of each subset of Ω. Then axiom (A1) states that the volume of the entire sample space should be one and axiom (A2) that the volume of a set can be computed by summing the volumes of disjoint subsets; see Figure 25. Observe that every set is disjoint with the empty set $\emptyset$ (i.e., $A \cap \emptyset = \emptyset$ holds for any set A). By setting $A_i = \emptyset$ for all $i > n$ in (4.14) we therefore obtain the finite additivity property

$$P(A_1 \cup \cdots \cup A_n) = P(A_1) + \cdots + P(A_n) \quad \text{if} \quad A_i \cap A_j = \emptyset \quad \text{for all} \quad i \neq j. \tag{4.15}$$

Note that probability maps *events* to reals on $[0,1]$, not outcomes. However, in a slight abuse of notation we often write $P(\omega) = P(\{\omega\})$ for $\omega \in \Omega$.

Definition 4.15 (DISCRETE PROBABILITY SPACE). If the sample space Ω is countable and $P\colon \mathcal{P}(\Omega) \to [0,1]$ is probability, the pair (Ω, P) is called a *discrete probability space* (suom. *diskreetti todennäköisyysavaruus*).

Discrete probability is fully determined by the probabilities $P(\omega)$.

Result 4.16 (DISCRETE PROBABILITY). Suppose that $\Omega = \{\omega_i : i \in \mathbb{N}\}$ is countable and

$$P(\omega_i) = P(\{\omega_i\}) = a_i \quad \text{for} \quad a_i \geq 0 \quad \text{such that} \quad \sum_{i=1}^{\infty} a_i = 1. \tag{4.16}$$

Then $P\colon \mathcal{P}(\Omega) \to [0,1]$ is a probability if and only if

$$P(A) = \sum_{\omega \in A} P(\omega) \quad \text{for every} \quad A \in \mathcal{P}(\Omega). \tag{4.17}$$

Proof. For the function P satisfying (4.16) and (4.17) it holds that

$$P(\Omega) = \sum_{\omega\in\Omega} P(\omega) = \sum_{i=1}^{\infty} P(\omega_i) = \sum_{i=1}^{\infty} a_i = 1 \tag{4.18}$$

and, if A_i are disjoint,

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{\omega\in\bigcup_{i=1}^{\infty} A_i} P(\omega) = \sum_{i=1}^{\infty} \sum_{\omega\in A_i} P(\omega) = \sum_{i=1}^{\infty} P(A_i), \tag{4.19}$$

which are the probability axioms (A1) and (A2). Therefore P is a probability. Conversely, suppose that P is a probability that satisfies (4.16). We can write any $A \subset \Omega$ as a countable union of singletons:

$$A = \bigcup_{\omega\in A} \{\omega\}. \tag{4.20}$$

These singletons are disjoint, so that the probability axiom (A2) gives

$$P(A) = P\left(\bigcup_{\omega\in A} \{\omega\}\right) = \sum_{\omega\in A} P(\omega), \tag{4.21}$$

which is (4.17). □

Example 4.17. A natural model for rolling a die is that all rolls are equally probable. That is,

$$P(i) = P(\{i\}) = \frac{1}{6} \quad \text{for every} \quad i = 1,\dots,6. \tag{4.22}$$

This results in probability given by

$$P(A) = \frac{|A|}{6} \quad \text{for} \quad A \in \mathcal{P}(\Omega), \tag{4.23}$$

where $|A|$ stands for the number of elements in A.

Example 4.18. If $\Omega = \{\omega_1,\dots,\omega_n\}$ is finite, then the function $P\colon \mathcal{P}(\Omega) \to [0,1]$ given by

$$P(A) = \frac{|A|}{n} \quad \text{for} \quad A \in \mathcal{P}(\Omega) \tag{4.24}$$

is a probability and the pair (Ω, P) is a *finite probability space* (suom. *äärellinen todennäköisyysavaruus*).

Example 4.19. Let $\Omega = \mathbb{N} \cup \{0\}$ be a countably infinite sample space. Let us define

$$P(A) = \frac{1}{e} \sum_{k\in A} \frac{1}{k!} \quad \text{for} \quad A \in \mathcal{P}(\Omega). \tag{4.25}$$

Note that each A is simply some subset of non-negative integers so that each k is a natural number and the factorial $k!$ makes sense. For any disjoint events A and B (i.e., $A \cap B = \emptyset$) we have

$$P(A \cup B) = \frac{1}{e} \sum_{k\in A\cup B} \frac{1}{k!} = \frac{1}{e}\left(\sum_{k\in A} \frac{1}{k!} + \sum_{k\in B} \frac{1}{k!}\right) = P(A) + P(B), \tag{4.26}$$

which shows that the probability axiom (A2) holds. Moreover,

$$P(\Omega) = \frac{1}{e} \sum_{k \in \mathbb{N} \cup \{0\}} \frac{1}{k!} = \frac{1}{e} \sum_{k=0}^{\infty} \frac{1}{k!} = 1, \tag{4.27}$$

where we used the fact that $e = \sum_{k=0}^{\infty} 1/k!$ [recall (2.70)]. This verified the probability axiom (A1). Therefore (Ω, P) is a discrete probability space.

Remark 4.20. Life is more complicated if the sample space is *uncountable* (suom. *ylinumeroituva*), such as $\Omega = [0,1] \subset \mathbb{R}$. It turns out that one cannot assign a natural measure of probability to every event of such a sample space. Namely, if Ω is the interval $[0,1]$, a uniform probability would naturally satisfy

$$P([a,b]) = b - a \quad \text{for any} \quad 0 \le a \le b \le 1. \tag{4.28}$$

However, it is possible to prove that a function $P\colon \mathcal{P}([0,1]) \to [0,1]$ cannot simultaneously satisfy (a) the additivity axiom (A2) in (4.14) for all countable collections of disjoint subsets of $[0,1]$ and (b) Equation (4.28). That is, there are some "bad" subsets of $[0,1]$ that we must disqualify as events in the uncountable case.

Unless otherwise stated, from now on we assume that we are working with a discrete probability space (Ω, P). The following result summarises some properties that a probability has. All these properties follow from the axioms of probability.

Result 4.21 (Basic properties of probability). A probability P has the following properties:

1. $P(\Omega) = 1$ and $P(\emptyset) = 0$.
2. $P(A_1 \cup \cdots \cup A_n) = P(A_1) + \cdots + P(A_n)$ if $A_1, \ldots, A_n$ are disjoint (i.e., $A_i \cap A_j = \emptyset$ for $i \neq j$).
3. $P(\overline{A}) = P(\Omega \setminus A) = 1 - P(A)$.
4. $P(A) \le P(B)$ if $A \subset B$.
5. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Proof. Let us provide a proof for property 5. We can write the set A as $A = (A \setminus B) \cup (A \cap B)$. Because $A \setminus B$ and $A \cap B$ are disjoint, probability axiom (A2) gives

$$P(A) = P(A \setminus B) + P(A \cap B). \tag{4.29}$$

Similarly we obtain $P(B) = P(B \setminus A) + P(A \cap B)$. Consequently,

$$P(A \setminus B) = P(A) - P(A \cap B) \quad \text{and} \quad P(B \setminus A) = P(B) - P(A \cap B). \tag{4.30}$$

We can decompose the union $A \cup B$ as

$$A \cup B = (A \setminus B) \cup (A \cap B) \cup (B \setminus A), \tag{4.31}$$

where the sets $A \setminus B$, $A \cap B$, and $B \setminus A$ are disjoint. From the probability axiom (A2) and (4.30) it thus follows that

$$\begin{aligned} P(A \cup B) &= P(A \setminus B) + P(A \cap B) + P(B \setminus A) \\ &= [P(A) - P(A \cap B)] + P(A \cap B) + [P(B) - P(A \cap B)] \\ &= P(A) + P(B) - P(A \cap B). \end{aligned}$$ □

4.3 Conditional Probability and Independence

In this section we introduce the concepts of conditional probability and independence. We also discuss Bayes' theorem that connects the conditional probability of an event A given event B to that of event B given A.

4.3.1 Conditional Probability

Conditional probability is the probability of event A given that we know that event B has occurred.

Definition 4.22 (CONDITIONAL PROBABILITY). Let A and B be two events such that $P(B) > 0$. The *conditional probability* (suom. *ehdollinen todennäköisyys*), $P(A \mid B)$, of A given B is defined as

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}. \tag{4.32}$$

When computing conditional probability one can think of the event B as the new sample space; the conditional probability $P(A \mid B)$ is then the probability of the portion of A that falls within this new sample space. Observe that conditional probability of A given the entire sample space is simply the probability of A:

$$P(A \mid \Omega) = \frac{P(A \cap \Omega)}{P(\Omega)} = \frac{P(A)}{P(\Omega)} = \frac{P(A)}{1} = P(A). \tag{4.33}$$

We can also verify that the conditional probability is a probability (i.e., it satisfies the two axioms of probability in Definition 4.14):

$$P(\Omega \mid B) = \frac{P(\Omega \cap B)}{P(B)} = \frac{P(B)}{P(B)} = 1 \tag{4.34}$$

and

$$P(\cup_{i=1}^{\infty} A_i \mid B) = \frac{P((\bigcup_{i=1}^{\infty} A_i) \cap B)}{P(B)} = \frac{P(\bigcup_{i=1}^{\infty} (A_i \cap B))}{P(B)} = \frac{\sum_{i=1}^{\infty} P(A_i \cap B)}{P(B)} = \sum_{i=1}^{\infty} P(A_i \mid B) \tag{4.35}$$

for disjoint A_i. Note also that any event A such that $A \cap B = \emptyset$ has conditional probability zero because $P(A \cap B) = P(\emptyset) = 0$ by Result 4.21.

Example 4.23. Consider the setting of Example 4.13 where we rolled two 6-sided dice. The sample space consists of 36 pairs of the form (i, j) for $i, j \in \{1, \ldots, 6\}$ that signify the rolls the two dice. Let B be the event "the sum of the rolls is 10". That is,

$$B = \{(4,6), (5,5), (6,4)\}. \tag{4.36}$$

Let A be the event "both rolls are even":

$$A = \{(2,2), (2,4), (2,6), (4,2), \ldots, (6,6)\}. \tag{4.37}$$

Then $P(A \mid B)$ is the probability that both rolls are even given that their sum was 10.

From (4.32) we compute

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{P(\{(4,6),(6,4)\})}{P(\{(4,6),(5,5),(6,4)\})}. \tag{4.38}$$

If we consider both dice fair, each outcome has equal probability $\frac{1}{36}$. Consequently (recall Example 4.18),

$$P(\{(4,6),(6,4)\}) = \frac{|\{(4,6),(6,4)\}|}{36} = \frac{2}{36} \tag{4.39}$$

and

$$P(\{(4,6),(5,5),(6,4)\}) = \frac{|\{(4,6),(5,5),(6,4)\}|}{36} = \frac{3}{36}. \tag{4.40}$$

Therefore

$$P(A \mid B) = \frac{2/36}{3/36} = \frac{2}{3}. \tag{4.41}$$

That is, two thirds of outcomes in (4.36) are also in (4.37). We can also compute the conditional probability $P(B \mid A)$, which is the probability that the sum of rolls is 10 given that both rolls were even. Now

$$P(B \mid A) = \frac{P(B \cap A)}{P(A)}. \tag{4.42}$$

We have $P(B \cap A) = P(A \cap B) = P(\{(4,6),(6,4)\}) = \frac{2}{36}$ from (4.39). It is straightforward to compute (e.g., by simply enumerating all outcomes) that the event A consists of 9 outcomes, so that

$$P(A) = P(\{(2,2),(2,4),(2,6),(4,2),\ldots,(6,6)\}) = \frac{9}{36}. \tag{4.43}$$

Therefore

$$P(B \mid A) = \frac{P(B \cap A)}{P(A)} = \frac{2/36}{9/36} = \frac{2}{9}. \tag{4.44}$$

Observe that $P(A \mid B) = \frac{2}{3} \neq \frac{2}{9} = P(B \mid A)$.

Example 4.24. Suppose we flip a coin three times. Each outcome is now a triplet of the form abc for $a,b,c \in \{\mathrm{h},\mathrm{t}\}$ (h for "heads" and t for "tails"). Consider the events

$$A = \text{"all flips are heads"} = \{\mathrm{hhh}\} \tag{4.45}$$

and

$$B = \text{"the first flip is tails"} = \{\mathrm{ttt},\mathrm{tth},\mathrm{thh},\mathrm{tht}\}. \tag{4.46}$$

Now $A \cap B = \emptyset$, so that the probability of all flips being heads given that the first flip was tails is

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{P(\emptyset)}{P(B)} = 0. \tag{4.47}$$

This is of course as it should be: if the first flip is tails, all flips cannot obviously be heads!

The *chain rule of probability* (suom. *ketjusääntö*) can be used to compute probabilities of intersections from conditional probabilities. The chain rule is also known as the *product rule* (suom. *tulosääntö*). There is no relation to the chain and product rule for differentiation.

Result 4.25 (CHAIN RULE OF PROBABILITY). Let $A_1, \ldots, A_n$ be events for which it holds that $P(A_1 \cap \cdots \cap A_{n-1}) > 0$. Then

$$P(A_1 \cap \cdots \cap A_n) = P(A_1)P(A_2 \mid A_1)P(A_3 \mid A_1 \cap A_2) \cdots P(A_n \mid A_1 \cap \cdots \cap A_{n-1}). \quad (4.48)$$

In particular,

$$P(A \cap B) = P(A \mid B)P(B) \quad (4.49)$$

if $P(B) > 0$.

Proof. We obtain (4.49) by multiplying both sides in the definition (4.32) of conditional probability with $P(B)$. By setting $A = A_n$ and $B = A_1 \cap \cdots \cap A_{n-1}$ we can then proceed iteratively to obtain (4.48). □

Sets $B_1, \ldots, B_n \subset \Omega$ are said to form a *partition* (suom. *ositus*) of Ω if (a) they are disjoint (i.e., $B_i \cap B_j = \emptyset$ for all $i \neq j$) and (b) $\Omega = B_1 \cup \cdots \cup B_n$.

Result 4.26 (LAW OF TOTAL PROBABILITY). Suppose that events $B_1, \ldots, B_n$ form a partition of the sample space Ω. For any event A we have the *law of total probability* (suom. *kokonaistodennäköisyyden kaava*)

$$P(A) = P(B_1)P(A \mid B_1) + \cdots + P(B_n)P(A \mid B_n). \quad (4.50)$$

Proof. Since the events $B_1, \ldots, B_n$ form a partition, we have $\Omega = B_1 \cup \cdots \cup B_n$ and consequently

$$A = (A \cap B_1) \cup \cdots \cup (A \cap B_n), \quad (4.51)$$

where $A \cap B_i$ are disjoint. Therefore the probability axiom (A2) and the chain rule in (4.49) give

$$P(A) = P(A \cap B_1) + \cdots + P(A \cap B_n) = P(B_1)P(A \mid B_1) + \cdots + P(B_n)P(A \mid B_n). \quad \square$$

The law of total probability can be used when there are multiple different ways that one outcome can occur.

Example 4.27. Suppose that televisions are manufactured by Factories 1, 2, and 3 that supply 10%, 30%, and 60% of televisions available for purchase, respectively. Suppose also that the televisions manufactured by the factories have 2%, 5%, and 10% chances of being defective, respectively. We can use the law of total probability to compute the probability that a purchased television is defective. Let B_i denote the event "television is manufactured by Factory i" and A the event "television is defective". Because every television is manufactured by one (and only one) of the three factories, the events B_1, B_2, and B_3 form a partition of the sample space. Moreover,

$$P(B_1) = 0.1, \quad P(B_2) = 0.3, \quad \text{and} \quad P(B_3) = 0.6. \quad (4.52)$$

Additionally, we have the conditional probabilities for a television manufactured by a given factory being defective:

$$P(A \mid B_1) = 0.02, \quad P(A \mid B_2) = 0.05, \quad \text{and} \quad P(A \mid B_3) = 0.1. \quad (4.53)$$

We can thus use the law of total probability to compute the probability of a purchased

television being defective:

$$\begin{aligned} P(A) &= P(B_1)P(A \mid B_1) + P(B_2)P(A \mid B_2) + P(B_3)P(A \mid B_3) \\ &= 0.1 \cdot 0.02 + 0.3 \cdot 0.05 + 0.6 \cdot 0.1 \\ &= 0.077. \end{aligned} \tag{4.54}$$

That is, the probability of purchasing a defective television is 7.7% (in the sense that 7.7% of all television available for purchase are defective).

4.3.2 Bayes' Theorem

Bayes' theorem is used to compute the conditional probability $P(A \mid B)$ from the conditional probability $P(B \mid A)$ and the probabilities $P(A)$ and $P(B)$.

Result 4.28 (BAYES' THEOREM). Let A and B be events such that $P(B) > 0$. Then

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)} \tag{4.55}$$

Proof. By the definition of conditional probability in (4.32) and the chain rule (4.49),

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B \mid A)P(A)}{P(B)}. \qquad \square$$

Equation (4.55) is known as *Bayes' rule* (suom. *Bayesin kaava*). In itself, Bayes' theorem is nothing more than a straightforward consequence of the definition of conditional probability. The following test sensitivity example is a classical example of Bayes' theorem in action.

Example 4.29. Suppose that 0.5% of population carries a certain disease and that we have a test for the disease that returns a positive result (i.e., indicates that a person has the disease) in 95% of cases if the person *does* carry the disease. However, the test also returns a positive result for 1% of cases if the person *does not* carry the disease. We can use Bayes' theorem to calculate the probability a person who tested positive actually carries the disease. Define the events

$$A = \text{“the person carries the disease”} \quad \text{and} \quad B = \text{“the person tests positive”}. \tag{4.56}$$

We thus want to compute the probability $P(A \mid B)$. By Bayes' theorem,

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}. \tag{4.57}$$

Here $P(A) = 0.005$ is the probability that the person carries the disease and $P(B \mid A) = 0.95$ is the probability that a person carrying the disease tests positive. To compute $P(B)$, the probability that a person tests positive, we can use the law of total probability. For this purpose, let C be the event “the person does not carry the disease”. Then A and C partition the sample space and the law of total probability in (4.50) gives

$$P(B) = P(A)P(B \mid A) + P(C)P(B \mid C). \tag{4.58}$$

Here $P(A) = 0.005$ is the probability of carrying the disease and $P(C) = 0.995$ the probability of not carrying it. The probability that a person *not* carrying the disease tests positive is

$P(B \mid C) = 0.01$. Therefore

$$P(B) = 0.005 \cdot 0.95 + 0.995 \cdot 0.01 = 0.0147. \tag{4.59}$$

Bayes' theorem thus gives

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)} = \frac{0.95 \cdot 0.005}{0.0147} \approx 0.32. \tag{4.60}$$

That is, the probability that a person who tests positive carries the disease is only 32%. This is explained by the fact the test gives a lot of false positives (1% of cases) in comparison to the rarity of the disease (only 0.5% of the population). For example, if 2% of population carried the disease, the conditional probability above would jump to $P(A \mid B) \approx 0.66$.

Example 4.30. The following is one possible formulation of the famous Monty Hall problem: In a game show there are three doors and behind one of them a prize. A contestant picks one door. Then the host, who knows where the prize is, picks uniformly at random one of the remaining doors behind which there is no prize, opens it, and asks the contestant if they would like to switch their choice of a door. Should the contestant do this? We can use Bayes' theorem to solve this problem. Suppose that it was Door 1 that the contestant chose initially and that the host revealed that the prize is not behind Door 3. Let B denote the event that the host opened Door 3, A_1 the event that the prize is behind Door 1, and A_2 the event that the prize is behind Door 2. From Bayes' theorem we obtain the probabilities

$$P(A_1 \mid B) = \frac{P(B \mid A_1)P(A_1)}{P(B)} \quad \text{and} \quad P(A_2 \mid B) = \frac{P(B \mid A_2)P(A_2)}{P(B)} \tag{4.61}$$

for the prize being behind Door 1 and Door 2 given that the host opened Door 3. Here

$$P(A_1) = P(A_2) = \frac{1}{3} \tag{4.62}$$

if we assume that the door behind which the prize is placed is selected uniformly at random. However,

$$P(B \mid A_1) = \frac{1}{2} \quad \text{but} \quad P(B \mid A_2) = 1. \tag{4.63}$$

Because the host can pick either Door 2 or Door 3 when the prize is behind Door 1, we have $P(B \mid A_1) = \frac{1}{2}$. But the host must pick Door 3 if the prize is behind Door 2, for otherwise they would reveal the prize. Therefore $P(B \mid A_2) = 1$. It follows from (4.61)–(4.63) that

$$P(A_2 \mid B) = \frac{P(B \mid A_2)P(A_2)}{P(B)} = \frac{1/3}{P(B)} > \frac{1/2 \cdot 1/3}{P(B)} = \frac{P(B \mid A_1)P(A_1)}{P(B)} = P(A_1 \mid B), \tag{4.64}$$

which means that the contestant should switch to Door 2 as the prize is twice as likely to be behind that that door than Door 1.

In *Bayesian statistics* prevalent in machine learning Bayes' theorem is interpreted as providing the *posterior probability* (suom. *posterioritodennäköisyys*), $P(A \mid B)$, of A after new data or evidence B has been taken into account. In this interpretation $P(A)$ is the *prior probability* (suom. *prioritodennäköisyys*) that expresses one's belief about event A (is A a likely or unlikely event) prior to obtaining any data and $P(B \mid A)$ is the *likelihood* (suom. *uskottavuus*), the probability of data B under the assumption that A is true.

4.3.3 *Independence*

Events A and B are said to be independent if the probability that both of them occur (i.e., the event $A \cap B$ in terms of set theory) equals the product of their probabilities.

Definition 4.31 (INDEPENDENCE). Two events A and B are *independent* (suom. *riippumattomia*) if

$$P(A \cap B) = P(A)P(B). \tag{4.65}$$

Events $A_1, \ldots, A_n$ are independent if

$$P(\cap_{i \in \mathcal{I}} A_i) = \prod_{i \in \mathcal{I}} P(A_i) \tag{4.66}$$

for every index set $\mathcal{I} \subset \{1, \ldots, n\}$.

Note that n events are considered independent only if (4.66) holds for *every* set of indices $\mathcal{I} \subset \{1, \ldots, n\}$. For example, for three events A_1, A_2, and A_3 it is not enough that

$$P(A_1 \cap A_2 \cap A_3) = P(A_1)P(A_2)P(A_3) \tag{4.67}$$

holds but we must also have

$$P(A_1 \cap A_2) = P(A_1)P(A_2), \quad P(A_1 \cap A_3) = P(A_1)P(A_3), \quad \text{and} \quad P(A_2 \cap A_3) = P(A_2)P(A_3).$$

Note also that events being independent is completely different from them being disjoint. If A and B are disjoint, then $A \cap B = \emptyset$, so that $P(A \cap B) = 0$. If A and B are independent, then $P(A \cap B) = P(A)P(B)$, which is zero only if $P(A) = 0$ or $P(B) = 0$.

Example 4.32. Consider the random experiment of rolling two dice that we have already discussed in Examples 4.13 and 4.23. It is intuitively clear that the events

$$A = \text{“the first roll is 3”} = \{(3,1), (3,2), (3,3), (3,4), (3,5), (3,6)\} \tag{4.68}$$

and

$$B = \text{“the second roll is 3”} = \{(1,3), (2,3), (3,3), (4,3), (5,3), (6,3)\} \tag{4.69}$$

should be independent (the first roll cannot affect the second and vice versa). It is straightforward to verify that these events are indeed independent in the sense of Definition 4.31. We have

$$P(A) = \frac{|A|}{36} = \frac{6}{36} = \frac{1}{6} \quad \text{and} \quad P(B) = \frac{|B|}{36} = \frac{6}{36} = \frac{1}{6}. \tag{4.70}$$

The intersection (“both rolls are 3”) is simply $A \cap B = \{(3,3)\}$ and thus

$$P(A \cap B) = \frac{|A \cap B|}{36} = \frac{1}{36} = \frac{1}{6} \cdot \frac{1}{6} = P(A) \cdot P(B), \tag{4.71}$$

which shows that A and B are independent. We can also exhibit events that are *not* independent. Consider the events

$$A = \text{“both rolls are at least 5”} = \{(5,5), (5,6), (6,5), (6,6)\} \tag{4.72}$$

and

$$B = \text{“the sum of rolls is 10”} = \{(4,6), (5,5), (6,4)\}. \tag{4.73}$$

Now we have

$$P(A) = \frac{4}{36} = \frac{1}{9} \quad \text{and} \quad P(B) = \frac{3}{36} = \frac{1}{12} \tag{4.74}$$

but

$$P(A \cap B) = P(\{5,5\}) = \frac{1}{36} \neq \frac{1}{108} = \frac{1}{9} \cdot \frac{1}{12} = P(A) \cdot P(B), \tag{4.75}$$

so that A and B are *not* independent.

4.4 Random Variables and Distributions

In this section we discuss random variables and their distributions.

4.4.1 Random Variables

Let us begin with the definition of a random variable.

Definition 4.33 (RANDOM VARIABLE). Any function $X : \Omega \to \mathbb{R}$ is called a *random variable* (suom. *satunnaismuuttuja*).

Random variables are neither "random" or "variable". That is, a random variable X is simply a function that maps an outcome $\omega \in \Omega$ to a real number $X(\omega)$. For example, if $\Omega = \mathbb{N}$, then the function $X : \mathbb{N} \to \mathbb{R}$ given by

$$X(\omega) = \omega^2 \quad \text{for} \quad n \in \mathbb{N} \tag{4.76}$$

is a random variable. In these notes we focus on discrete probability spaces (Ω, P). Because the sample space Ω of a discrete probability space is countable, so is its image

$$X(\Omega) = \{X(\omega) : \omega \in \Omega\} \tag{4.77}$$

under a random variable $X : \Omega \to \mathbb{R}$.

Definition 4.34 (PROBABILITY DISTRIBUTION). Let $X : \Omega \to \mathbb{R}$ be a random variable. The *distribution* (suom. *jakauma*) of X is the function $P_X : \mathcal{P}(X(\Omega)) \to [0,1]$ given by

$$P_X(A) = P(X \in A) = P(\{\omega \in \Omega : X(\omega) \in A\}) \quad \text{for} \quad A \subset \mathbb{R}. \tag{4.78}$$

The probability $P(X \in A)$ is the probability that the random variable $X : \Omega \to \mathbb{R}$ takes value in the set $A \subset \mathbb{R}$.

Example 4.35. Let $\Omega = \mathbb{N} \cup \{0\}$ and consider the random variable $X : \Omega \to \mathbb{R}$ given by $X(\omega) = \omega^2$. Suppose that the probability P on Ω is given by

$$P(A) = \frac{1}{e} \sum_{k \in A} \frac{1}{k!} \quad \text{for} \quad A \in \mathcal{P}(\Omega) \tag{4.79}$$

as in Example 4.19. Let us compute $P_X([0,a]) = P(X \in [0,a])$ for any interval $[0,a]$. This is the probability that X has value at most a. Since $X(\omega) = \omega^2$, we have $X(\omega) \in [0,a]$ if and only if $\omega \leq \sqrt{a}$. Therefore

$$P_X([0,a]) = P(\{\omega \in \Omega : X(\omega) \in [0,a]\} = P(\{\omega \in \mathbb{N} \cup \{0\} : \omega \leq \sqrt{a}\}). \tag{4.80}$$

By (4.79),

$$P_X([0,a]) = \frac{1}{e} \sum_{k \leq \sqrt{a}} \frac{1}{k!}, \tag{4.81}$$

where summation is over all non-negative integers k such that $k \leq \sqrt{a}$. In particular, for $m \in \mathbb{N}\{0\}$ we have

$$P_X([0,m^2]) = \frac{1}{e}\sum_{k\leq m}\frac{1}{k!} = \frac{1}{e}\sum_{k=1}^{m}\frac{1}{k!}. \tag{4.82}$$

Result 4.36 (DISTRIBUTIONS ARE PROBABILITIES). The distribution P_X of a random variable $X\colon \Omega \to \mathbb{R}$ is a probability on $\mathbb{R}$.

Proof. Because $P_X(A)$ is defined as the probability of a certain subset of Ω for each $A \subset \mathbb{R}$, the distribution takes values in $[0,1]$. It remains to verify the probability axioms (A1) and (A2) of Definition 4.14. First,

$$P_X(\mathbb{R}) = P(\{\omega \in \Omega : X(\omega) \in \mathbb{R}\}) = P(\Omega) = 1, \tag{4.83}$$

which verifies axiom (A1). To verify axiom (A2), let $A_i \subset \mathbb{R}$ for $i \geq 1$ be disjoint. Then

$$\begin{aligned} P_X\left(\bigcup_{i=1}^{\infty} A_i\right) = P(\{\omega \in \Omega : X(\omega) \in \cup_{i=1}^{\infty} A_i\}) &= P\left(\bigcup_{i=1}^{\infty}\{\omega \in \Omega : X(\omega) \in A_i\}\right) \\ &= \sum_{i=1}^{\infty} P(\{\omega \in \Omega : X(\omega) \in A_i\}) \\ &= \sum_{i=1}^{\infty} P_X(A_i). \end{aligned}$$

□

Typically we do not work directly with the underlying sample space Ω but rather with random variables and their distributions. Result 4.36 is useful because it implies that all the definitions and results we have proved above for probabilities remain valid for distributions. Effectively we can think of the image $X(\Omega)$ as the sample space [note that $P_X(A) = 0$ for any $A \subset \mathbb{R}$ such that $A \cap X(\Omega) = \emptyset$], its subsets as events, and the distribution P_X as the probability of interest.

Because we have assumed that the sample space is countable, the random variables we consider are discrete, in that they take only countable many different values. The distribution of a discrete random variable is fully described by its probability mass function.

Definition 4.37 (PROBABILITY MASS FUNCTION). The *probability mass function* (suom. *pistetodennäköisyysfunktio*) of a discrete random variable $X\colon \Omega \to \mathbb{R}$ is the function $p_X\colon \mathbb{R} \to [0,1]$ given by

$$p_X(x) = P(X = x) = P(\{\omega \in \Omega : X(\omega) = x\}). \tag{4.84}$$

For each $x \in \mathbb{R}$, the probability mass function gives the probability that the random variable takes value x. Note that there are only countably many x for which the probability mass function of a discrete random variable satisfies $p_X(x) > 0$.

Example 4.38. Consider the probability P and the random variable given by $X(\omega) = \omega^2$ from Example 4.35. The probability mass function of X is given by

$$p_X(x) = P(X = x) = \begin{cases} P(\sqrt{x}) = e^{-1}\frac{1}{k!} & \text{if} \quad \sqrt{x} = k \in \mathbb{N}\cup\{0\}, \\ 0 & \text{otherwise.} \end{cases} \tag{4.85}$$

Result 4.39 (PROBABILITY MASS DETERMINES DISTRIBUTION). The distribution of a discrete random variable $X\colon \Omega \to \mathbb{R}$ can be written as

$$P_X(A) = P(X \in A) = \sum_{x \in A} p_X(x) \quad \text{for all} \quad A \subset \mathbb{R}. \tag{4.86}$$

Example 4.40. Let us once again consider rolling two dice (see also Examples 4.13, 4.23, and 4.32). Each outcome is a pair $\omega = (i,j)$, where $i,j \in \{1,\ldots,6\}$. Let us consider a random variable X that gives the sum of the rolls:

$$X(\omega) = X(i,j) = i + j. \tag{4.87}$$

The possible values of this random variable, the sets of "favourable" outcomes $\omega = (i,j)$ giving rise to these outcomes, and the cardinalities of these sets are given below:

$$\begin{array}{lll}
X(\omega) = 2 & \omega \in \{(1,1)\} & |\{X=2\}| = 1, \\
X(\omega) = 3 & \omega \in \{(1,2),(2,1)\} & |\{X=3\}| = 2, \\
X(\omega) = 4 & \omega \in \{(1,3),(2,2),(3,1)\} & |\{X=4\}| = 3, \\
X(\omega) = 5 & \omega \in \{(1,4),(2,3),(3,2),(4,1)\} & |\{X=5\}| = 4, \\
X(\omega) = 6 & \omega \in \{(1,5),(2,4),(3,3),(4,2),(5,1)\} & |\{X=6\}| = 5, \\
X(\omega) = 7 & \omega \in \{(1,6),(2,5),(3,4),(4,3),(5,2),(6,1)\} & |\{X=7\}| = 6, \\
X(\omega) = 8 & \omega \in \{(2,6),(3,5),(4,4),(3,5),(6,2)\} & |\{X=8\}| = 5, \\
X(\omega) = 9 & \omega \in \{(3,6),(4,5),(5,4),(6,3)\} & |\{X=9\}| = 4, \\
X(\omega) = 10 & \omega \in \{(4,6),(5,5),(6,4)\} & |\{X=10\}| = 3, \\
X(\omega) = 11 & \omega \in \{(5,6),(6,5)\} & |\{X=11\}| = 2, \\
X(\omega) = 12 & \omega \in \{(6,6)\} & |\{X=12\}| = 1.
\end{array}$$

Because the total number of outcomes is 36, the probability mass function of X is given by

$$p_X(x) = P(X=x) = P(\{X=x\}) = P(\{\omega \in \Omega : X(\omega) = x\}) = \frac{|\{X=x\}|}{36}. \tag{4.88}$$

This yields

$$\begin{array}{lll}
p_X(2) = \frac{1}{36}, & p_X(3) = \frac{2}{36} = \frac{1}{18}, & p_X(4) = \frac{3}{36} = \frac{1}{12}, \\
p_X(5) = \frac{4}{36} = \frac{1}{9}, & p_X(6) = \frac{5}{36}, & p_X(7) = \frac{6}{36} = \frac{1}{6}, \\
p_X(8) = \frac{5}{36}, & p_X(9) = \frac{4}{36} = \frac{1}{9}, & p_X(10) = \frac{3}{36} = \frac{1}{12}, \\
p_X(11) = \frac{2}{36} = \frac{1}{18}, & p_X(12) = \frac{1}{36}. &
\end{array}$$

The probability of a random variable being at most some number is given by the cumulative distribution function.

Definition 4.41 (CUMULATIVE DISTRIBUTION FUNCTION). The *cumulative distribution function* (suom. *kertymäfunktio*) of a random variable $X\colon \Omega \to \mathbb{R}$ is the function $F_X\colon \mathbb{R} \to [0,1]$ given by

$$F_X(x) = P(X \leq x) = P(X \in (-\infty, x]) = P(\{\omega \in \Omega : X(\omega) \leq x\}). \tag{4.89}$$

For discrete random variables that we consider the cumulative distribution function can be expressed using the probability mass function. Let us enumerate the image of X as $X(\Omega) =$

$\{x_1, x_2, \ldots\}$. Then the cumulative distribution function is

$$F_X(x) = P(X \leq x) = \sum_{x_k \leq x} P(X = x_k) = \sum_{x_k \leq x} p_X(x_k). \tag{4.90}$$

In Example 4.35 we essentially computed the cumulative distribution function of the random variable given by $X(\omega) = \omega^2$.

Example 4.42. In Example 4.40 we computed the probability mass function for a random variable X that gives the sum of two dice rolls. Using the expressions for the probability mass function and (4.90) we obtain

$$\begin{aligned}
F_X(x) &= 0 && \text{if} \quad x < 2,\\
F_X(x) &= \tfrac{1}{36} && \text{if} \quad x \in [2,3),\\
F_X(x) &= \tfrac{1}{36} + \tfrac{2}{36} = \tfrac{1}{12} && \text{if} \quad x \in [3,4),\\
F_X(x) &= \tfrac{1}{36} + \cdots + \tfrac{3}{36} = \tfrac{1}{6} && \text{if} \quad x \in [4,5),\\
F_X(x) &= \tfrac{1}{36} + \cdots + \tfrac{4}{36} = \tfrac{5}{18} && \text{if} \quad x \in [5,6),\\
F_X(x) &= \tfrac{1}{36} + \cdots + \tfrac{5}{36} = \tfrac{15}{36} && \text{if} \quad x \in [6,7),\\
F_X(x) &= \tfrac{1}{36} + \cdots + \tfrac{6}{36} = \tfrac{7}{12} && \text{if} \quad x \in [7,8),\\
F_X(x) &= \tfrac{21}{36} + \tfrac{5}{36} = \tfrac{13}{18} && \text{if} \quad x \in [8,9),\\
F_X(x) &= \tfrac{21}{36} + \tfrac{5}{36} + \tfrac{4}{36} = \tfrac{5}{6} && \text{if} \quad x \in [9,10),\\
F_X(x) &= \tfrac{21}{36} + \tfrac{5}{36} + \cdots + \tfrac{3}{36} = \tfrac{11}{12} && \text{if} \quad x \in [10,11),\\
F_X(x) &= \tfrac{21}{36} + \tfrac{5}{36} + \cdots + \tfrac{2}{36} = \tfrac{35}{36} && \text{if} \quad x \in [11,12),\\
F_X(x) &= \tfrac{21}{36} + \tfrac{5}{36} + \cdots + \tfrac{1}{36} = 1 && \text{if} \quad x \geq 12.
\end{aligned}$$

The concept of independence of events from Section 4.3.3 extends to random variables via their distributions.

Definition 4.43 (INDEPENDENT RANDOM VARIABLES). Random variables $X, Y : \Omega \to \mathbb{R}$ are *independent* if

$$P((X \in A) \cap (Y \in A)) = P(X \in A)P(Y \in A) = P_X(A)P_Y(A) \tag{4.91}$$

for all $A \subset \mathbb{R}$. Alternatively, the random variables are independent if the events $X \in A$ and $Y \in A$ are independent.

Note that

$$(X \in A) \cap (Y \in A) = \{\omega \in \Omega : X(\omega) \in A \text{ and } Y(\omega) \in A\}. \tag{4.92}$$

4.4.2 Some Discrete Distributions

Let us next review two important discrete distributions.

Definition 4.44 (UNIFORM DISTRIBUTION). A random variable $X : \Omega \to \mathbb{R}$ whose image $X(\Omega) = \{x_1, \ldots, x_n\}$ is a finite set follows a *discrete uniform distribution* (suom. *diskreetti tasajakauma*) if

$$p_X(x_k) = P(X = x_k) = \frac{1}{n} \quad \text{for every} \quad k = 1, \ldots, n. \tag{4.93}$$

A random variable is uniformly distributed if each of its possible values is equally likely.

Example 4.45. Suppose that the sample space $\Omega = \{1,2,3,4,5,6\}$ is the set of possible rolls of a single die and X is the (rather trivial) random variable $X(\omega) = \omega + 13$. If all rolls are assumed equally likely [i.e., $P(\omega) = 1/6$ for every $\omega \in \Omega$], then the random variable X follows a discrete uniform distribution and

$$p_X(x) = P(X = x) = \frac{1}{6} \quad \text{for} \quad x \in X(\Omega) = \{14, \ldots, 19\}. \tag{4.94}$$

We can use the same sample space and probability P to construct other uniformly distributed random variables. Let $Y : \Omega \to \mathbb{R}$ be the random variable given by

$$Y(\omega) = \begin{cases} 0 & \text{if } \omega \text{ is odd,} \\ 1 & \text{if } \omega \text{ is even.} \end{cases} \tag{4.95}$$

The random variable Y is a binary random variable that indicates parity of the roll. Its image is $Y(\Omega) = \{Y(\omega) : \omega \in \Omega\} = \{0,1\}$ and

$$p_Y(0) = P(\{1,3,5\}) = \frac{3}{6} = \frac{1}{2} \quad \text{and} \quad p_Y(1) = P(\{2,4,6\}) = \frac{3}{6} = \frac{1}{2}. \tag{4.96}$$

Therefore Y follows a discrete uniform distribution.

Definition 4.46 (POISSON DISTRIBUTION). A random variable $X : \Omega \to \mathbb{R}$ follows a *Poisson distribution* (suom. *Poisson-jakauma*) with parameter $\lambda > 0$ if

$$p_X(k) = P(X = k) = e^{-\lambda} \frac{\lambda^k}{k!} \quad \text{for} \quad k \in \mathbb{N} \cup \{0\}. \tag{4.97}$$

In this case we write

$$X \sim \text{Poisson}(\lambda). \tag{4.98}$$

We have already encountered the Poisson distribution (with $\lambda = 1$) in Examples 4.19 and 4.35. The Poisson distribution is used to *model* the number of occurrences of events during a fixed time interval when the events are independent and λ events are expected to occur during the interval.

Example 4.47. Suppose that an average of 4 customers enter a store every hour. Assuming that the customers enter independently of one another, we can model the number of customers as a Poisson distribution with $\lambda = 4$ (we do not attempt to discuss whether or not this model is "good" or "useful" in some sense). Let X be the number of customers entering the store during an hour and assume that

$$X \sim \text{Poisson}(4). \tag{4.99}$$

We can now use (4.90) and (4.97) to, for example, calculate that *under this model* the probability that at most 2 customers enter in an hour is

$$F_X(2) = P(X \leq 2) = p_X(0) + p_X(1) + p_X(2) = e^{-4}\left(1 + \frac{4^1}{1!} + \frac{4^2}{2!}\right) \approx 0.24. \tag{4.100}$$

The probability that at most 8 customers enter is

$$F_X(8) = P(X \leq 8) = \sum_{k=0}^{8} p_X(8) = e^{-4} \sum_{k=0}^{8} \frac{4^k}{k!} \approx 0.98. \tag{4.101}$$

4.5 Expected Value and Variance

Finally, let us define and briefly discuss expected value and variance of a random variable.

Definition 4.48 (**EXPECTED VALUE**). The *expected value* (suom. *odotusarvo*) of a random variable $X : \Omega \to \mathbb{R}$ with image $X(\Omega) = \{x_1, x_2, \ldots\}$ is

$$\mathbb{E}(X) = \sum_{k=1}^{\infty} x_k P(X = x_K) = \sum_{k=1}^{\infty} x_k p_X(x_k). \tag{4.102}$$

Expected value is the weighted average of the values that a random variable can take. Each value x_k is weighted by the probability $P(X = x_k)$ that X takes this particular value.

Example 4.49. Consider again rolling two dice and suppose that we are interested in the expected value of the sum of the rolls. As in Example 4.40, we can introduce the random variable X given by $X(\omega) = X(i,j) = i + j$. From the computations in that example and the definition of the expected value we obtain

$$\begin{aligned}\mathbb{E}(X) = \sum_{k=2}^{12} k p_X(k) &= \tfrac{2\cdot1}{36} + \tfrac{3\cdot2}{36} + \tfrac{4\cdot3}{36} + \tfrac{5\cdot4}{36} + \tfrac{6\cdot5}{36} + \tfrac{7\cdot6}{36} + \tfrac{8\cdot5}{36} + \tfrac{9\cdot4}{36} + \tfrac{10\cdot3}{36} + \tfrac{11\cdot2}{36} + \tfrac{12\cdot1}{36} \\ &= \tfrac{252}{36} \\ &= 7.\end{aligned}$$

Therefore the expected value of the sum of rolls is 7.

Example 4.50. Let us compute the expected value of a random variable $X \sim \text{Poisson}(\lambda)$ that follows a Poisson distribution with parameter $\lambda > 0$. From (4.97) and the definition of expected value we get

$$\begin{aligned}\mathbb{E}(X) = \sum_{k=0}^{\infty} k p_X(k) = \sum_{k=0}^{\infty} k e^{-\lambda} \frac{\lambda^k}{k!} = e^{-\lambda} \sum_{k=0}^{\infty} \frac{k}{k!} \lambda^k &= e^{-\lambda} \sum_{k=1}^{\infty} \frac{k}{k!} \lambda^k \\ &= \lambda e^{-\lambda} \sum_{k=1}^{\infty} \frac{1}{(k-1)!} \lambda^{k-1} \\ &= \lambda e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} \\ &= \lambda,\end{aligned} \tag{4.103}$$

where we used the facts that $k/k! = (k-1)!$ and $\lambda^k = \lambda \cdot \lambda^{k-1}$ as well as the Maclaurin expansion of the exponential function in (2.70). Thus the expected value of $X \sim \text{Poisson}(\lambda)$ is λ.

The following result collects a number of intuitive properties of the expected value.

Result 4.51 (PROPERTIES OF EXPECTED VALUE). Let X and Y be random variables with expected values $\mathbb{E}(X)$ and $\mathbb{E}(Y)$. The expected value has the following properties.

1. *Positivity*: $\mathbb{E}(X) \geq 0$ if $X \geq 0$.
2. *Linearity*: $\mathbb{E}(aX + bY) = a\mathbb{E}(X) + b\mathbb{E}(Y)$ for any $a, b \in \mathbb{R}$.
3. *Monotonicity*: $\mathbb{E}(X) \leq \mathbb{E}(Y)$ if $X \leq Y$.
4. *Constants*: $\mathbb{E}(a) = a$ if $a \in \mathbb{R}$ is a constant (i.e., the random variable that maps each $\omega \in \Omega$ to a).
5. *Independence*: If X and Y are independent, then $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$.

Variance describes how "wide" a distribution is, or how much it differs from its expected value.

Definition 4.52 (VARIANCE). The *variance* (suom. *varianssi*) of a random variable $X : \Omega \to \mathbb{R}$ is

$$\mathrm{Var}(X) = \mathbb{E}[(X - \mathbb{E}(X))^2]. \tag{4.104}$$

The square root of the variance, $\sqrt{\mathrm{Var}(X)}$, is called *standard deviation* (suom. *keskihajonta*) of X.

It is often more convenient to compute variance from equation

$$\mathrm{Var}(X) = \mathbb{E}(X^2) - [\mathbb{E}(X)]^2. \tag{4.105}$$

Let us verify this equation. We can expand (4.104) as

$$\mathrm{Var}(X) = \mathbb{E}[X^2 - 2X\mathbb{E}(X) + (\mathbb{E}(X))^2]. \tag{4.106}$$

The linearity of expectation then yields [note that $\mathbb{E}(X)$ is just a constant]

$$\mathrm{Var}(X) = \mathbb{E}(X^2) - 2\mathbb{E}(X)\mathbb{E}(X) + [\mathbb{E}(X)]^2 = \mathbb{E}(X^2) - [\mathbb{E}(X)]^2. \tag{4.107}$$

Example 4.53. Suppose that a random variable X follows a discrete uniform distribution and $X(\Omega) = \{1, \ldots, n\}$. Then $p_X(k) = \frac{1}{n}$ for each $k \in \{1, \ldots, n\}$. Computing the variance of X from (4.105) requires the expectations $\mathbb{E}(X^2)$ and $\mathbb{E}(X)$. The latter is

$$\mathbb{E}(X) = \sum_{k=1}^{n} k p_X(k) = \sum_{k=1}^{n} \frac{k}{n} = \frac{n+1}{2}, \tag{4.108}$$

where we used the formula $1 + 2 + \cdots + n = \frac{n(n+1)}{2}$ for the sum of an arithmetic sequence. To compute $\mathbb{E}(X^2)$ we need to consider the random variable $Y = X^2$. The probability that this random variable takes value k^2 is $p_Y(k) = p_X(k) = \frac{1}{n}$. Therefore

$$\mathbb{E}(X^2) = \mathbb{E}(Y) = \sum_{k=1}^{n} k^2 p_Y(k) = \sum_{k=1}^{n} \frac{k^2}{n} = \frac{(n+1)(2n+1)}{6}, \tag{4.109}$$

where we used the formula $1^2 + 2^2 + \cdots + n^2 = \frac{n(n+1)(2n+1)}{6}$. Consequently the variance of

X is

$$\mathrm{Var}(X) = \mathbb{E}(X^2) - [\mathbb{E}(X)]^2 = \frac{(n+1)(2n+1)}{6} - \left(\frac{n+1}{2}\right)^2 = \frac{n^2-1}{12}. \tag{4.110}$$

Example 4.54. Let us compute the variance of a random variable $X \sim \mathrm{Poisson}(\lambda)$ that follows a Poisson distribution with parameter $\lambda > 0$. In Example 4.50 we computed that $\mathbb{E}(X) = \lambda$. To use (4.104) we need the expectation $\mathbb{E}(X^2)$ of the random variable $Y = X^2$. Similar to Example 4.53, this random variable takes the value k^2 with probability $p_Y(k) = p_X(k) = e^{-\lambda}\lambda^k/k!$. Therefore

$$\mathbb{E}(X^2) = \mathbb{E}(Y) = \sum_{k=0}^{\infty} k^2 p_Y(k) = e^{-\lambda} \sum_{k=0}^{\infty} k^2 \frac{\lambda^k}{k!}. \tag{4.111}$$

One can show that (we forgo the details) the expression on the right evaluates to $\lambda^2 + \lambda$. Thus

$$\mathrm{Var}(X) = \mathbb{E}(X^2) - [\mathbb{E}(X)]^2 = (\lambda^2 + \lambda) - \lambda^2 = \lambda. \tag{4.112}$$

Result 4.55 (PROPERTIES OF VARIANCE). Let X and Y be random variables with variances $\mathrm{Var}(X)$ and $\mathrm{Var}(Y)$. The variance has the following properties.

1. *Positivity*: $\mathrm{Var}(X) \geq 0$.
2. *Linear transformations*: $\mathrm{Var}(aX + b) = a^2\mathrm{Var}(X)$ for any $a, b \in \mathbb{R}$.
3. *Constants*: $\mathrm{Var}(a) = 0$ if $a \in \mathbb{R}$ is a constant (i.e., the random variable that maps each $\omega \in \Omega$ to a).
4. *Independence*: If X and Y are independent, then $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$.

Proof. Let us verify properties 3 and 4. If the random variable $X = a$ is constant, we obtain from properties 2 and 4 in Result 4.51 and (4.105) that

$$\mathrm{Var}(a) = \mathbb{E}(a^2) - [\mathbb{E}(a)]^2 = \mathbb{E}(a \cdot a) - [\mathbb{E}(a)]^2 = a\mathbb{E}(a) - a^2 = a^2 - a^2 = 0, \tag{4.113}$$

which is property 3. If X and Y are independent, then properties 2 and 5 in Result 4.51 and (4.105) give

$$\begin{aligned}
\mathrm{Var}(X+Y) &= \mathbb{E}[(X+Y)^2] - [\mathbb{E}(X+Y)]^2 \\
&= \mathbb{E}(X^2 + 2XY + Y^2) - [\mathbb{E}(X)]^2 - 2\mathbb{E}(XY) - [\mathbb{E}(Y)]^2 \\
&= \mathbb{E}(X^2) + 2\mathbb{E}(X)\mathbb{E}(Y) + \mathbb{E}(Y^2) - [\mathbb{E}(X)]^2 - 2\mathbb{E}(X)\mathbb{E}(Y) - [\mathbb{E}(Y)]^2 \\
&= \mathbb{E}(X^2) - [\mathbb{E}(X)]^2 + \mathbb{E}(Y^2) - [\mathbb{E}(Y)]^2 \\
&= \mathrm{Var}(X) + \mathrm{Var}(Y).
\end{aligned}$$

□