Lectures on Mathematics

Peter Tallos

CUB, Department of Mathematics

September, 2020

Preface

This book covers the necessary mathematics intended for students with major in *Economics*. This course provides the fundamental mathematical background for studying *Microeconomics, Macroeconomics, Statistics* and other important topics in economics, or probabilistic or stochastic disciplines. The main mathematical topics covered are *Mathematical Analysis (Calculus), Probability Theory,* and *Linear Algebra*.

Most of the time, we avoid the rigorous mathematical proofs of our statements. Instead, we rather present "justifications", which are intuitive, but not necessarily precise. However, emphasis is placed on the correct formulations of definitions.

The text is illustrated by a large number of examples. On the one hand, they help the deeper understanding. On the other hand, they give an idea, how to apply them in practical situations. Therefore, the thorough study of examples is a profoundly important homework assignment. Each chapter covers one week of the semester, on a one week – one chapter basis.

In the end of each chapter references are given to the *Textbook*, which should be interpreted the following way.

Textbook-1: K. Sydsaeter and P. Hammond, *Mathematics for Economic Analysis*, Prentice Hall, 1995, ISBN 0–13–583600–X, or any of the later editions.

Textbook-2 R. E. Walpole, R. H. Myers, S. L. Myers and K. Ye *Probability and Statistics for Engineers and Scientists*, Prentice Hall, 2012, ISBN: 978–0–321–62911–1, or any of the later editions.

These textbooks are widely used at most recognized universities worldwide.

Some of the indicated homework exercises refer to the *Textbook*. Most of the midterm quiz or final exam problems are similar or identical to those exercises. More problems and exercises with solutions are posted on my web site.

Special thanks to my colleagues Csaba Puskás, Éva Ernyes and Balázs Fleiner, who read the manuscript, and their valuable remarks significantly improved the quality of the text. My thanks go to my former students as well, their comments in or outside the classroom were extremely helpful for making the text more understandable.

Budapest, September, 2020.

Peter Tallos

Contents

Part I

First Semester: Differential and Integral Calculus

Chapter 1

Sequences

1.1 Limits of sequences

The function

$$a : \mathbb{N} \to \mathbb{R}$$

defined on the set of natural numbers $\mathbb{N}$ is called a (infinite) sequence.

We use the notation a_n for the n-th element.

Some examples: $a_n = n$, $a_n = \frac{1}{n}$, $a_n = \frac{n+1}{n+2}$.

Definition 1.1 The sequence a_n is said to be *convergent* and tends to A, if for any $\varepsilon > 0$, there exists an index N, such that,

$$|a_n - A| < \varepsilon.$$

whenever $n \geq N$.

If the sequence is convergent then A is called the limit of the sequence a_n and we write

$$\lim_{n\to\infty} a_n = A \ .$$

If there is no such real number A, then the sequence is called *divergent*.

Theorem 1.2 *If the sequences a_n and b_n are convergent and $\lim_{n\to\infty} a_n = A$ and $\lim_{n\to\infty} b_n = B$ then*

- $\lim_{n\to\infty}(a_n \pm b_n) = A \pm B$,
- $\lim_{n\to\infty}(a_n \cdot b_n) = A \cdot B$,
- *if $B \neq 0$, then* $\lim_{n\to\infty} \frac{a_n}{b_n} = \frac{A}{B}$

Example 1.3 Let us consider the sequence $a_n = \frac{1}{n}$. For an arbitrary $\varepsilon > 0$ let N be any integer, greater than $1/\varepsilon$. Then if $n \geq N$

$$\frac{1}{n} < \varepsilon\,,$$

therefore, in view of Definition 1.1

$$\lim_{n\to\infty} \frac{1}{n} = 0\ .$$

Example 1.4 In a similar way we can find the limits of other sequences. Let us consider for example the sequence

$$a_n = \frac{2n^2+5}{n^2-6n+8}.$$

If we divide both the numerator and the denominator by n^2, then we have

$$a_n = \frac{2+5/n^2}{1-6/n+8/n^2},$$

where the limit of the numerator is 2 and the limit of the denominator is 1. Therefore

$$\lim_{n\to\infty} a_n = 2\ .$$

Every irrational number can be written as a limit of a sequence of rational numbers. For example, consider the sequence $a_1 = 1.4$, $a_2 = 1.41$, $a_3 = 1.414$, $a_4 = 1.4142\ldots$ then

$$\lim_{n\to\infty} a_n = \sqrt{2}$$

Indeed, according to Definition 1.1, if $\varepsilon = 10^{-N}$, then $|a_n - \sqrt{2}| < \varepsilon$ for $n \geq N$.

A typical example for a sequence which has no limit is

$$a_n = (-1)^n.$$

1.2 Sequences tending to infinity

Let us investigate the sequence

$$a_n = 2n+5.$$

The terms of this sequence are greater than any given number K if n is large enough. In that case, we say, that the limit of the sequence is infinity. We use the symbol ∞ to denote infinity.

Definition 1.5 We say that the sequence a_n approaches $+\infty$ if for any real number K there exists an index N such that for every $n \geq N$ we have $a_n > K$ This is expressed in the formula

$$\lim_{n\to\infty} a_n = +\infty \ ,$$

In a completely analogous way we can define the fact that a sequence approaches $-\infty$, that is $\lim_{n\to\infty} a_n = -\infty$.

1.3 Squeezing Theorem

Often the limit of a sequence can be determined with the aid of other sequences the limits of which are known. Such a situation is described by the Squeezing Theorem.

Theorem 1.6 **(Squeezing Theorem)** *Let a_n, b_n and c_n be sequences such that for every index n*

$$a_n \leq b_n \leq c_n$$

holds and, moreover, the sequences a_n and c_n converge to the same limit A. Then the sequence b_n is also convergent and $\lim_{n\to\infty} b_n = A$.

Example 1.7 Let $a > 1$ be a real number and consider the sequence $b_n = \sqrt[n]{a}$. Since $a > 1$, the elements of the sequence can be written in the form

$$\sqrt[n]{a} = 1 + h_n \ ,$$

where $h_n > 0$ for every n. By the Binomial Theorem we get

$$a = (1 + h_n)^n > 1 + nh_n \ .$$

where we skipped all other positive terms on the right-hand side. Rearranging the inequality it follows that

$$0 < h_n < \frac{a-1}{n} \ .$$

The expression on the right-hand side tends to zero, hence, by the Squeezing Theorem $h_n \to 0$, that is $\sqrt[n]{a} \to 1$.

Obviously, if $0 < a \leq 1$, then we can carry out the same argument, by taking reciprocals of the elements of the sequence. This shows that our theorem holds for any constant $a > 0$.

1.4 Bounded and monotone sequences

Clearly, the elements of a sequence approaching infinity cannot stay between two real numbers. We introduce the following definition.

Definition 1.8 The sequence a_n is *bounded from above*, if there is a real number K such that $a_n \leq K$ for every index n. If there is a real number K such that $a_n \geq K$ for every index n, the sequence is said to be *bounded from below*. A sequence is called bounded if it is bounded both from above and from below.

Example 1.9 Decide whether the sequence

$$a_n = \frac{2n}{\sqrt{4n^2+5}+8}$$

is bounded or not? Dividing both the numerator and the denominator by $2n$ we get

$$a_n = \frac{1}{\sqrt{1+5/4n^2}+8/2n}\,,$$

hence $0 \leq a_n \leq 1$. Thus the sequence is bounded. It is also clear that the smallest upper bound of the sequence is 1, while 0 is a lower bound, but not the greatest one.

Monotone sequences have special importance.

Definition 1.10 We say that the sequence a_n is *monotone increasing*, if $a_n \leq a_{n+1}$ for every index n. A decreasing sequence is defined similarly. A sequence that is either increasing or decreasing is called monotone.

Example 1.11 Consider the sequence

$$a_n = \frac{2n-1}{n+2}\,.$$

We have

$$a_n = \frac{2n+4-5}{n+2} = 2 - \frac{5}{n+2}\,.$$

The value of the fraction subtracted from 2 decreases if n increases, therefore the sequence a_n is increasing. It is also clear that the sequence is bounded from above and its smallest upper bound is 2. Moreover,

$$\lim_{n\to\infty} a_n = 2\,.$$

Our next theorem states that this property is characteristic for bounded monotone sequences.

Theorem 1.12 *An increasing sequence which is bounded from above is convergent.*

An analogous statement holds for decreasing sequences that are bounded from below.

We do not prove this theorem, but we note it is based on the property of real numbers that we always have a least upper bound (among the infinitly many upper bounds) which turns out to be the limit of the sequence.

Analogous theorem applies for monotone decreasing and bounded from below sequences.

1.5 Euler's number e

In many applications of mathematics the sequence

$$a_n = \left(1 + \frac{1}{n}\right)^n . \tag{1.1}$$

appears frequently. We can show that this sequence is monotone increasing, bounded from above, and consequently convergent.

To verify these properties we exploit the inequality between the arithmetic and geometric means. In particular, if $x_1, \ldots, x_n$ are positive numbers, then

$$x_1 \ldots x_n \leq \left(\frac{x_1 + \ldots + x_n}{n}\right)^n$$

for every integer n. Equality holds if and only if $x_1 = \ldots = x_n$ that is, all the numbers are equal.

Proposition 1.13 *The sequence (1.1) is strictly monotone increasing and bounded from above.*

Proof. Let n be a given integer. Consider the $n + 1$ pieces of positive numbers

$$x_1 = 1 + \frac{1}{n}, \ \ldots, \ x_n = 1 + \frac{1}{n}, \ x_{n+1} = 1$$

which are not all equal. Using the inequality for the arithmetic and geometric means, we have

$$\left(1 + \frac{1}{n}\right)^n < \left(\frac{n+1+1}{n+1}\right)^{n+1} = \left(1 + \frac{1}{n+1}\right)^{n+1}$$

which exactly says that the sequence is strictly monotone increasing.

Second, consider the $n+2$ pieces of positive numbers

$$x_1 = 1 + \frac{1}{n},\ \ldots,\ x_n = 1 + \frac{1}{n},\ x_{n+1} = \frac{1}{2},\ x_{n+2} = \frac{1}{2}$$

which are not all equal. Using the inequality again, we have

$$\frac{1}{4} \cdot \left(1 + \frac{1}{n}\right)^n < \left(\frac{n+1+1}{n+2}\right)^{n+2} = 1$$

Rearranging the inequality we obtain $a_n < 4$, that means that the sequence is bounded from above. Consequently, the sequence (1.1) is convergent. $\square$

We use the notation e for the limit of this sequence. More elaborate computations show that e is irrational, and

$$e = 2.7182...$$

Proposition 1.14 *Let α be an arbitrarily given real number. Then*

$$\lim_{n\to\infty} \left(1 + \frac{\alpha}{n}\right)^n = e^\alpha$$

Example 1.15 Consider the sequence

$$a_n = \left(\frac{2n+1}{2n+3}\right)^n$$

Then, by rewriting the sequence we get

$$a_n = \left(\frac{2n+1}{2n+3}\right)^n = \frac{\left(1 + \frac{1/2}{n}\right)^n}{\left(1 + \frac{3/2}{n}\right)^n} \to \frac{e^{1/2}}{e^{3/2}}$$

and hence $\lim_{n\to\infty} a_n = e^{-1}$.

Study at home:

1. Careful study of Mathematical Analysis Exercises.
2. Study the exercises below.
3. Textbook-1, Chapter 1 and Section 6.4.

Chapter 2

Infinite Series

2.1 Series

Definition 2.1 Let a_k be a real infinite sequence and compose the formal sum

$$\sum_{k=1}^{\infty} a_k \,. \tag{2.1}$$

This symbol is called an *infinite series* (or just simply a series).

The meaning of this expression should be clarified, because only the addition of finitely many real numbers was defined so far.

For any natural number n define the n-th *partial sum* of the series (2.1) as follows:

$$S_n = \sum_{k=1}^{n} a_k \tag{2.2}$$

This way we created a real sequence S_n.

Definition 2.2 The infinite series (2.1) is said to be *convergent* and its sum is S, if the sequence S_n is convergent and its limit is S. In this case we use the notation:

$$S = \sum_{k=1}^{\infty} a_k$$

Otherwise the series is said to be *divergent*.

Please note that the infinite series is divergent if the sequence S_n has no limit or its limit is not finite. For instance, if $a_k = (-1)^k$ for all k then

$$S_n = \sum_{k=1}^{n} (-1)^k = 0 \text{ if } n \text{ is even and } S_n = \sum_{k=1}^{n} (-1)^k = -1 \text{ if } n \text{ is odd}$$

therefore, the sequence S_n has no limit and the series is obviously divergent.

2.2 Geometric series

Example 2.3 **(Geometric series)** Let r be a real number and consider the infinite geometric series with common ratio r:

$$\sum_{k=0}^{\infty} r^k$$

The nth partial sum of this series is

$$S_n = \sum_{k=0}^{n-1} r^k = \begin{cases} \frac{1-r^n}{1-r} & if \quad r \neq 1 \\ n & if \quad r = 1 \end{cases}$$

It is well known about the sequence $a_n = r^n$ that $r^n \to 0$ if $|r| < 1$, $r^n \to 1$ if $r = 1$ and otherwise the sequence is divergent. Therefore, we get that the geometric series is convergent if and only if $|r| < 1$ and then its sum is given by

$$S = \sum_{k=0}^{\infty} r^k = \frac{1}{1-r}$$

2.3 Convergence based on examining the partial sums

Example 2.4 Consider the series

$$\sum_{k=2}^{\infty} \frac{1}{k(k-1)} \quad . \tag{2.3}$$

The terms of this series can be rewritten in this form:

$$\frac{1}{k(k-1)} = \frac{1}{k-1} - \frac{1}{k}$$

Observe that the n-th partial sum will be given like:

$$S_n = (1 - 1/2) + (1/2 - 1/3) + \ldots + (1/(n-1) - 1/n) = 1 - 1/n$$

The limit of this sequence is obviously 1 (the negative and positive identical terms cancel each other) and we conclude that the series is convergent and its sum is $S = 1$.

Example 2.5 Try the apply the above argument for the series

$$\sum_{k=2}^{\infty} \frac{1}{k^3 - k}$$

and by eliminating the terms that cancel each other, find the sum of the series.

2.4 Conditions for convergence

Theorem 2.6 **(Necessary condition for convergence)** *Assume that the series*

$$\sum_{k=1}^{\infty} a_k$$

is convergent. Then $\lim_{k\to\infty} a_k \to 0$.

Example 2.7 This theorem formulates a necessary condition which may not be sufficient. For instance, we can show that the series hogy a

$$\sum_{k=1}^{\infty} \frac{1}{k}$$

fulfills the necessary condition but it is divergent. This series is called the *Harmonic series*.

Indeed, let an integer n be given, and consider the 2^n-th partial sum of the Harmonic series. Rearrange the terms in the following way

$$S_{2^n} = 1 + \frac{1}{2} + \left(\frac{1}{3} + \frac{1}{4}\right) + \left(\frac{1}{5} + \ldots + \frac{1}{8}\right) + \ldots + \left(\frac{1}{2^{n-1}+1} + \ldots + \frac{1}{2^n}\right),$$

where every expression within the parentheses goes up to the next power of 2. The sum of terms inside the parentheses is always bigger than $1/2$, and we have exactly n pairs of parentheses, hence

$$S_{2^n} > 1 + \frac{1}{2}n\,.$$

That tells us that sequence of partial sums is not bounded and therefore, the series is divergent.

Theorem 2.8 **(Sufficient condition for convergence)** *let us suppose that for each index k we have $a_k \geq 0$ and the series*

$$\sum_{k=1}^{\infty} a_k$$

is convergent. If for every index k we have $0 \le b_k \le a_k$, then the series

$$\sum_{k=1}^{\infty} b_k$$

is also convergent.

Indeed, on the one hand the sequence of partial sums $S_n = \sum_{k=1}^{n} b_k$ is monotone increasing, and on the other hand it is also bounded. Consequently, the series is convergent.

In an analogous way we may formulate a sufficient condition for divergence: if all terms of a series are bigger than the nonnegative terms of a divergent series, than it is divergent as well.

Example 2.9 As an application consider the series $\sum_{k=1}^{\infty} 1/k^2$. Since for every $k > 1$

$$\frac{1}{k^2} < \frac{1}{k(k-1)}$$

then for the n-th partial sums we get

$$S_n = \sum_{k=1}^{n} \frac{1}{k^2} < 1 + \sum_{k=2}^{n} \frac{1}{k(k-1)}$$

According to the sufficient condition we conclude that this series is convergent, and for its sum we have $S < 2$.

In general, it can be verified that the series $\sum_{k=1}^{\infty} 1/k^{\alpha}$ is divergent, if $\alpha \le 1$, and it is convergent if $\alpha > 1$ (see more details in Chapter 9).

2.5 Absolute convergence

In this section we examine series that may contain positive and terms as well. Consider the series

$$\sum_{k=1}^{\infty} a_k \tag{2.4}$$

where the terms a_k are not necessarily all nonnegative.

Definition 2.10 We say that the series (2.4) is *absolutely convergent*, if the series

$$\sum_{k=1}^{\infty} |a_k|$$

is convergent.

Theorem 2.11 *If a series is absolutely convergent, then it is convergent as well.*

We do not go into the details of the proof. As a justification we note the following. If S_n denotes the sum of the absolute values of the first n terms, then by our condition it is convergent and

$$\lim \sum_{k=1}^{n} |a_k| = \lim S_n = S\,.$$

Let R_n and T_n denote the sum of the negative and positive terms respectively from the first n terms of the series $\sum_{k=1}^{\infty} a_k$. Then R_n is monotone decresing, while T_n is monotone increasing, and both sequences are bounded, since

$$R_n \geq -S \quad \text{and} \quad T_n \leq S\,.$$

Therefore both sequences are convergent, in notation: $\lim R_n = R$, and $\lim T_n = T$. Thus the limit of of S_n can be given as:

$$\lim S_n = \lim \sum_{k=1}^{n} a_k = \lim(T_n + R_n) = T + R\,,$$

and we deduce that the series is really convergent.

The example below shows that the converse of our previous theorem is not necessarily true.

Example 2.12 Consider the following series with alternating signs:

$$\sum_{k=1}^{\infty} \frac{(-1)^{k-1}}{k}$$

Clearly, this series is not absolutely convergent, since the series with the absolute values of the terms is identical to the Harmonic series, which is divergent.

We show however, that the series above is convergent. Indeed, the sum of the terms with even indeces:

$$\begin{aligned} S_{2n} &= \left(1 - \frac{1}{2}\right) + \left(\frac{1}{3} - \frac{1}{4}\right) + \ldots + \left(\frac{1}{2n-1} - \frac{1}{2n}\right) = \\ &= \frac{1}{2} + \frac{1}{12} + \ldots + \frac{1}{2n(2n-1)}\,. \end{aligned}$$

In view of Example 2.3, this sequence is monotone increasing and bounded from above, because $S_{2n} < 2$. Hence, it is convergent. Denote its limit by

$$\lim S_{2n} = S\,.$$

On the other hand, for the sum of the terms with odd indeces we have

$$S_{2n-1} = S_{2n} + \frac{1}{2n}$$

therefore, $\lim S_{2n-1} = S$, which means that $\lim S_n = S$. This implies that the series is convergent.

2.6 Quotient-test

In this section we formulate a very useful sufficient condition for the convegence or divergence of infinite series. Create the absolute values of the quotients of the consecutive terms of the series

$$\sum_{k=1}^{\infty} a_k$$

and suppose that the limit

$$\lim_{k\to\infty} \left| \frac{a_{k+1}}{a_k} \right| = \alpha$$

exists.

Theorem 2.13 (Quotient-test)

- *If $\alpha < 1$, then the series is absolutely convergent.*
- *If $\alpha > 1$, then the series is divergent.*
- *If $\alpha = 1$, then both cases can occur.*

Proof. If $\alpha < 1$, then choose a real number β with $\alpha < \beta < 1$. Then from a certain index N we have

$$\left| \frac{a_{k+1}}{a_k} \right| < \beta$$

for every $k \geq N$. Then goin step-by-step backward we get

$$|a_{k+1}| < \beta |a_k| < \beta^2 |a_{k-1}| < \ldots < \beta^{k-N+1} |a_N|$$

So, for the $n+1$-th partial sum

$$S_{n+1} = \sum_{k=0}^{n} |a_{k+1}| < \sum_{k=0}^{N-1} |a_{k+1}| + |a_N| \cdot \sum_{k=N}^{n} \beta^{k-N+1}$$

where the last sum is the partial sum of a convergent series (in view of $0 < \beta < 1$), and consequently bounded if $n \to \infty$. This proves the statement.

If $\alpha > 1$, then the the proof can be carried out similarly, with a choice of $1 < \beta < \alpha$ we can come up with an estimate with a divergent geometric series. □

Example 2.14 In this example we demonstrate that in the case of $\alpha = 1$ nothing can be stated about the convergence of the series.

Indeed, if the divergent Harmonic series is considered, then for $a_k = 1/k$ we have

$$\frac{a_{k+1}}{a_k} = \frac{k}{k+1} \to 1 \quad \text{if } k \to \infty\,.$$

However, if we take the convergent series, where $a_k = 1/k^2$, then

$$\frac{a_{k+1}}{a_k} = \left(\frac{k}{k+1}\right)^2 \to 1 \quad \text{if } k \to \infty\,,$$

which demonstrates that both cases can occur.

Example 2.15 Find out if the series

$$\sum_{k=1}^{\infty} \frac{k^2 \cdot 2^k}{k!}$$

is convergent or not. Use the Quotient-rule:

$$\frac{a_{k+1}}{a_k} = \frac{(k+1)^2 2^{k+1}}{(k+1)!} \cdot \frac{k!}{k^2 2^k} = 2\left(\frac{k+1}{k}\right)^2 \cdot \frac{1}{k+1} \to 0$$

Thus $\alpha = 0 < 1$, which tells us that the series is convergent.

Study at home:

1. Review the "Mathematical Analysis Exercises"
2. Additional homework: check the exercises below
3. Textbook-1, Section 6.5.

Chapter 3

Limits and continuity

3.1 Basic concepts

In the subsequent chapter we study the limit of functions $f : \mathbb{R} \to \mathbb{R}$. Let x_0 be a point (possibly equal to $\pm\infty$) for which there exists a sequence x_n in the domain of f such that $x_n \neq x_0$ and $x_n \to x_0$.

Definition 3.1 The limit of the function f at the point x_0 is said to be A (which can be $\pm\infty$) and in notation

$$\lim_{x \to x_0} f(x) = A$$

if for any sequence x_n from the domain of f whenever $x_n \to x_0$, $x_n \neq x_0$, then $f(x_n) \to A$.

ATTENTION!

Please note that the limit of f at x_0 has nothing to do with $f(x_0)$. The function may not even be defined at x_0. However, in some cases the limit may be equal to $f(x_0)$.

Theorem 3.2 *If the functions f and g have limits at x_0 and $\lim_{x \to x_0} f(x) = A$ and $\lim_{x \to x_0} g(x) = B$ then*

- $\lim_{x \to x_0} (f \pm g)(x) = A \pm B$,
- $\lim_{x \to x_0} (f \cdot g)(x) = A \cdot B$,
- *if* $B \neq 0$ *then* $\lim_{x \to x_0} \frac{f}{g}(x) = \frac{A}{B}$,
- *if* $A \neq 0$ *and* $B = 0$ *then* $\lim_{x \to x_0} \frac{f}{g}(x) = \pm\infty$.

Example 3.3 Determine the limit

$$\lim_{x\to 2} \frac{x^2-4}{x-2}.$$

This function is not defined for $x = 2$ but it is equal to $x+2$ at any point $x \neq 2$. Therefore it is easily seen that

$$\lim_{x\to 2} \frac{x^2-4}{x-2} = \lim_{x\to 2}(x+2) = 4\ .$$

Example 3.4 Consider the function $f(x) = 1/x$. This function is not defined at $x = 0$. On the other hand, for any sequence $x_n > 0$, $x_n \to 0$ from the domain we have $f(x_n) \to +\infty$ while $f(-x_n) \to -\infty$. Thus this function has no limit at $x = 0$, that is

$$\lim_{x\to 0} \frac{1}{x}$$

does not exist.

Example 3.5 Consider the following limit:

$$\lim_{x\to +\infty} \frac{2x^4 - 5x^3 + x - 8}{8x^3 - x^2 + 12}$$

Dividing both the numerator and denominator by x^3 we get the expression

$$\frac{2x - 5 + 1/x^2 - 8/x^3}{8 - 1/x + 12/x^3}.$$

Now for any sequence $x_n \to +\infty$ the limit of the numerator is $+\infty$, while the limit of the denominator equals 8, thus the fraction tends to $+\infty$.

Very similarly, we can show that the limit of the fraction is $-\infty$, if $x \to -\infty$.

Example 3.6 Show that

$$\lim_{x\to +\infty} (\sqrt{1+x^2} - x) = 0\ .$$

Indeed,

$$\sqrt{1+x^2} - x = \left(\sqrt{1+x^2} - x\right) \frac{\sqrt{1+x^2}+x}{\sqrt{1+x^2}+x} = \frac{1}{\sqrt{1+x^2}+x}$$

and the expression on the right hand side approaches 0 if $x \to +\infty$.

3.2 Squeezing theorem

In this section we formulate the Squeezing theorem for limits of functions.

Theorem 3.7 **(Squeezing Theorem)** *Let f, g and h be real functions such that for any x*

$$f(x) \leq g(x) \leq h(x)$$

and furthermore, $\lim_{x \to x_0} f(x) = \lim_{x \to x_0} h(x) = A$. Then the limit of the function g at x_0 exists, and

$$\lim_{x \to x_0} g(x) = A\,.$$

Example 3.8 Find the limit

$$\lim_{x \to 0} \frac{\sin x}{x}$$

This is an even function, therefore it is enough to consider positive values of x. A geometric interpretation (open the Figures file!) shows that for all $0 < x < \pi/2$

$$\sin x < x < \tan x\,.$$

Dividing this inequality by $\sin x$, we get

$$1 < \frac{x}{\sin x} < \frac{1}{\cos x}\,.$$

By taking the reciprocals, we obtain

$$\cos x < \frac{\sin x}{x} < 1$$

for every $0 < x < \pi/2$. In view of the Squeezing Theorem we receive

$$\lim_{x \to 0} \frac{\sin x}{x} = 1$$

3.3 One-sided limits

In some situations the limit of a function at a given point does not exist, but we still can speak about a one-sided limit.

Definition 3.9 We say that the *right-hand limit* of f at the point x_0 exists and is equal to

$$\lim_{x \to x_0+} f(x) = A$$

if for any sequence $x_n \to x_0$, $x_n > x_0$ from the domain of f we have $f(x_n) \to A$. The *left-hand* limit is defined analogously.

It is obvious from the definition that if at a point the limit exists, then both one-sided limits exist, and they are equal.

Example 3.10 Consider the function:

$$f(x) = \frac{2x+1}{x-2}$$

It is easy to see that if x_n approaches 2 from the right then $f(x_n) \to +\infty$, while if x_n approaches 2 from the left then $f(x_n) \to -\infty$. Therefore

$$\lim_{x \to 2-} f(x) = -\infty \quad \text{and} \quad \lim_{x \to 2+} f(x) = +\infty \; .$$

We can say that the limit of a function at a point exists if and only if both one-sided limits exist, and they are equal (the common value is the limit).

3.4 Continuity

Definition 3.11 Consider a function f that is defined on an interval. We say that the function f is *continuous* at a point x_0 of its domain if

$$\lim_{x \to x_0} f(x) = f(x_0) \; .$$

If f is not continuous at a point x_0 of its domain, then it is said that the function has a discontinuity there.

A function is simply called continuous, if it is continuous at every point of the domain.

ATTENTION!

Continuity is defined only at points in the domain of the function. For instance the function $f(x) = 1/x$ is continuous at each point of its domain, that is at each $x \neq 0$. The point $x_0 = 0$ is not in the domain of f, so we cannot speak of discontinuity here.

On the other hand, f cannot be defined at $x_0 = 0$ so that it becomes continuous, as the limit of the function does not exist there.

Functions obtained from continuous function by composition or by elementary operations (addition, subtraction, multiplication, division) are also continuous except maybe at points, where the denominator of the fraction equals zero.

Example 3.12 For instance, consider the following function:

$$f(x) = \begin{cases} \frac{1-\cos x}{x^2} & \text{if } x \neq 0 \\ \frac{1}{2} & \text{if } x = 0 \end{cases}$$

It is clear that this function is continuous for all $x \neq 0$, furthermore

$$\frac{1-\cos x}{x^2} = \frac{1-\cos^2 x}{(1+\cos x)x^2} = \left(\frac{\sin x}{x}\right)^2 \cdot \frac{1}{1+\cos x}\,.$$

This shows that the limit of the function at 0 equals $1/2$. Thus, this function is continuous on the whole real line.

We think of a continuous function as one whose graph can be drawn by an unbroken curve (without lifting the pencil from the paper). This is expressed in Bolzano's theorem.

Theorem 3.13 **(Bolzano)** *Let f be a continuous function on the finite interval $[a,b]$, and suppose that $f(a)$ and $f(b)$ have different signs. Then there exists a point $c \in (a,b)$ such that $f(c) = 0$.*

We do not prove the theorem, but note that a simple idea would be bisecting the interval, and selecting the part where f has opposite signs at the endpoints. If we keep doing this infinitely many times, we receive a sequence of intervals, so that each one is the half of the preceding interval. We think that the intersection of the intervals reduces to a single point, which is necessarily a zero of the function.

Example 3.14 *Prove that the equation*

$$2x^5 - 18x^4 + 3x^3 + 20x - 13 = 0$$

has at least one real solution. The expression on the left side of the equation defines a continuous function f for which

$$\lim_{x\to+\infty} f(x) = +\infty \quad \textit{and} \quad \lim_{x\to-\infty} f(x) = -\infty\,.$$

Therefore f is positive for sufficiently large values of x and takes negative values if x is small enough. Therefore, by the Bolzano-theorem the equation has at least one real solution.

The following property of continuous functions is of fundamental importance for extremum problems and optimization.

Theorem 3.15 **(Weierstrass)** *Let f be a continuous function on the finite interval $[a, b]$. Then f takes its maximum and minimum on this interval.*

We do not prove this theorem, but note that the function has to be bounded, and there exists a lowest upper bound. It can be shown that the lowest upper bound is the maximum of the function. A similar argument applies for the minimum.

For example, the function

$$f(x) = \begin{cases} x & \text{if } 0 \leq x \leq 1 \\ 3 - x & \text{if } 1 < x \leq 2 \end{cases}$$

does not reach its maximum value on the interval $[0, 2]$, but as we see, it is not continuous at 1.

Study at home:

- Textbook-1, read Sections 6.1, 6.2, 6.7, 7.1 and 7.2.
- Textbook-1, Exercises on pages 171–172, 177–178, 198, 202 and 205.
- Thourough study of "Mathematical Analysis Exercises" on my web site.

Chapter 4

Differentiation of functions

4.1 The derivative

Let f be a real function defined on an interval, and suppose that x_0 is an interior point of the interval.

Definition 4.1 We say that f is *differentiable* at x_0 if the following limit exists and it is finite:

$$\lim_{h\to 0} \frac{f(x_0+h)-f(x_0)}{h}$$

This limit is called the *derivative* of f at the point x_0, its notation is $f'(x_0)$. We say that the function f is differentiable in an interval, if it is differentiable at every interior point of the interval.

The quotient above is called the *difference quotient* of f at the point x_0.

Example 4.2 Consider the function $f(x) = x^2$ on the real line. The difference quotient at x_0 is:

$$\frac{f(x_0+h)-f(x_0)}{h} = \frac{(x_0+h)^2 - x_0^2}{h} = 2x_0 + h$$

whose limit is $2x_0$, if $h \to 0$. Consequently

$$f'(x_0) = 2x_0 \ .$$

In a very similar way we can show that in the case of $f(x) = x^n$ (where n is an integer),

$$f'(x_0) = nx_0^{n-1} \ .$$

In fact, use the identity

$$(x_0+h)^n - x_0^n = h((x_0+h)^{n-1} + (x_0+h)^{n-2}\cdot x_0 + \ldots + x_0^{n-1})$$

Theorem 4.3 *If f differentiable at x, then it is continuous at x.*

Proof. Let $h_n \to 0$, $h_n \neq 0$ be a sequence, then by the differentiability

$$\lim_{n\to\infty} \frac{f(x+h_n) - f(x)}{h_n} = f'(x) \,,$$

which is finite. This is only possible if $\lim_{n\to\infty}(f(x+h_n) - f(x)) = 0$, that is $\lim_{n\to\infty} f(x+h_n) = f(x)$. This exactly means that f is continuous at x. □

ATTENTION! The converse statement is not true in general, as it is demonstrated by the following example

Example 4.4 Consider the function $f(x) = |x|$ on the real line, and examine its difference quotient at $x_0 = 0$. It is clear that

$$\frac{f(h) - f(0)}{h} = \frac{|h|}{h} = \begin{cases} 1 & \text{if } h > 0 \\ -1 & \text{if } h < 0 \end{cases}$$

and therefore, the limit does not exist when $h \to 0$, since the right-hand limit is $+1$, while the left-hand limit is -1. Thus the function f is not differentiable at $x = 0$

However, f is differentiable at any other point, in particular $f'(x) = 1$, if $x > 0$, and $f'(x) = -1$, if $x < 0$.

4.2 Tangent lines

Geometric interpretation (see Figures.pdf) shows that $f'(x_0)$ is the slope of the tangent line to the graph of f at x_0.

By using this observation, we can give the equation of the tangent line to the graph of f that passes through the point $P(x_0, f(x_0))$:

$$y = f'(x_0)(x - x_0) + f(x_0) \,.$$

For instance, the equation of the tangent line to the graph of $f(x) = x^3$ at $x_0 = 1$ is

$$y = 3(x-1) + 1$$

Example 4.5 Find the equation of the tangent line to the graph of $f(x) = \sin x$ at $x_0 = 0$. On the one hand, the tangent line passes through the origin, on the other hand, the slope is:

$$f'(0) = \lim_{h\to 0} \frac{f(h) - f(0)}{h} = \lim_{h\to 0} \frac{\sin h}{h} = 1 \,.$$

Therefore, the equation is $y = x$ that intersects the graph at the origin.

4.3 Rules of differentiation

Consider the functions f and g, and assume that both are differentiable at x. The rules below follow from the basic properties of limits.

Derivative of a sum If α and β real numbers, then $\alpha f(x) + \beta g(x)$ is differentiable at x and

$$(\alpha f(x) + \beta g(x))' = \alpha f'(x) + \beta g'(x) \ ,$$

Derivative of a product $f(x) \cdot g(x)$ is differentiable at x and

$$(f(x) \cdot g(x))' = f'(x) \cdot g(x) + f(x) \cdot g'(x) \ ,$$

Derivative of a quotient if $g(x) \neq 0$, then $f(x)/g(x)$ is differentiable at x, and

$$\left(\frac{f(x)}{g(x)}\right)' = \frac{f'(x)g(x) - f(x)g'(x)}{g(x)^2} \ .$$

As an example let us see how we can prove the differentiability of the product:

$$\frac{f(x+h) \cdot g(x+h) - f(x) \cdot g(x)}{h} =$$

$$\frac{f(x+h) \cdot g(x+h) - f(x+h) \cdot g(x)}{h} +$$

$$\frac{f(x+h) \cdot g(x) - f(x) \cdot g(x)}{h} =$$

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

Here the limit of the first factor is $f(x)g'(x)$ based on the continuity of f, while the limit of the second factor is $f'(x)g(x)$, if $h \to 0$. That completes the proof. The proofs of the other rules can be carried out in a very similar way.

Example 4.6 The tangent line to the graph of $f(x) = 1/x$ taken at any point encloses a triangle with the coordinate axes. (See Figures.pdf.) Show that the area of this traingle is the same, no matter at what point the tangent line is taken.

Because of the symmetry, it is enough to focus to points $x_0 > 0$. By the Quotient-rule

$$f'(x_0) = -\frac{1}{x_0^2}$$

hence, the equation of the tangent line taken at x_0 is:

$$y = -\frac{1}{x_0^2}(x - x_0) + \frac{1}{x_0}$$

The intersection points with the coordinate axes are:

if $x = 0$, then the intersection point on the y-axis is $b = 2/x_0$

and similarly

if $y = 0$, then the intersection point on the x-axis is $a = 2x_0$.

Thus, the are of the enclosed right triangle is

$$A = \frac{1}{2}ab = \frac{1}{2} \cdot 2x_0 \cdot \frac{2}{x_0} = 2$$

which is independents of the choice of x_0.

4.4 Composition of functions

Let f and g be both $\mathbb{R} \to \mathbb{R}$ functions so that the range of g lies inside (subset) the domain of f. Then the function

$$x \to f(g(x))$$

is called the *composition* of f and g. For this function we use the notetation Jelölése $f \circ g$, that is:

$$f \circ g(x) = f(g(x)) \ .$$

For instance if $f(x) = \sqrt{x}$ and $g(x) = 1 + x^2$, then

$$f \circ g(x) = \sqrt{1 + x^2} \ .$$

Attention, the order is important!

In general $f \circ g \neq g \circ f$. If we consider the example above, then

$$g \circ f(x) = 1 + x$$

but this function is defined only for $x \geq 0$!

It may even turn out that $f \circ g$ is defined on the nonnegative half line, but $g \circ f$ is not defined anywhere. For instance, if

$$f(x) = -1 - x^4 \quad \text{and} \quad g(x) = \sqrt{x} \ ,$$

then $f \circ g(x) = -1 - x^2$, if $x \geq 0$, but $g \circ f(x) = \sqrt{-1 - x^4}$ is not defined for any real number.

4.5 Chain-Rule

Our theorem on the differentiability of composition functions is a very powerful tool for calculating the derivatives of more complicated functions.

Theorem 4.7 **(Chain-Rule)** *Suppose that g is differentiable at x, and f is differentiable at $g(x)$, then $f \circ g$ is differentiable at x, and its derivative is given by*

$$(f \circ g)'(x) = f'(g(x)) \cdot g'(x)$$

If we introduce the notation $k = g(x+h) - g(x)$, then the difference quotient of the composition function $f \circ g$ at x can be written like:

$$\frac{f(g(x+h)) - f(g(x))}{h} =$$

$$\frac{f(g(x)+k) - f(g(x))}{k} \cdot \frac{g(x+h) - g(x)}{h}$$

provided $g(x+h) - g(x) \neq 0$. In the case of $h \to 0$, in view of the continuity of g, we have $k \to 0$, and consequently, the limit of the expression on the right-hand is:

$$f'(g(x)) \cdot g'(x)$$

Unfortunately, this idea does not work when $k = 0$. In that case the proof is somewhat more complicated, we do not go into the details of that situation.

Example 4.8 For example, consider the function

$$F(x) = (1 + 3x - x^2)^6 \ .$$

We can find the derivative without expanding the 6-th power, if we notice that with the notations $f(x) = x^6$ and $g(x) = 1 + 3x - x^2$ we can write $F = f \circ g$. Therefore, by the Chain-Rule:

$$F'(x) = 6(1 + 3x - x^2)^5 \cdot (3 - 2x) \ .$$

Example 4.9 Now find the derivative of

$$F(x) = \left(\frac{2x+3}{5+x^2}\right)^3 \quad x \in \mathbb{R}$$

Then by using the notations

$$g(x) = \frac{2x+3}{5+x^2} \quad \text{and} \quad f(x) = x^3$$

we get $F = f \circ g$. Keep in mind that g is a quotient (use the Qutient-rule!), so we obtain

$$F'(x) = f'(g(x)) \cdot g'(x) = 3\left(\frac{2x+3}{5+x^2}\right)^2 \cdot \frac{2(5+x^2) - 2x(2x+3)}{(5+x^2)^2}$$

that form can still be further simplified if we wish.

Study at home:

1. Review the "Mathematical Analysis Exercises"
2. Review the Exercises below
3. Textbook-1, Chapter 4, Sections 5.2 and 5.6.

Chapter 5

The Mean Value Theorem

5.1 The inverse function

Consider a function $f : \mathbb{R} \to \mathbb{R}$ that is one-to-one on a given interval. In the case of a continuous function this means that it is either strictly monotone decreasing or strictly monotone increasing (in view of Bolzano's theorem, see Theorem 3.13).

Definition 5.1 The *inverse* of f is the function f^{-1} whose domain is the range of f, its range is the domain of f, and further

$$f^{-1} \circ f(x) = x$$

at every point in the domain of f.

This "reverse" correspondence can be obtained by taking the equality

$$y = f(x)$$

and isolate x as the function of y:

$$x = f^{-1}(y) \ .$$

For instance, if $f(x) = (2x + 5)^3$, then we get

$$f^{-1}(y) = \frac{\sqrt[3]{y} - 5}{2} \ .$$

Geometrically this means that the graphs of f^{-1} and of f are symmetric with respect to the staight line $y = x$ (that bisects the right angle at the origin).

5.2 Differentiability of the inverse function

Theorem 5.2 *Assume that f is continuous and strictly monotone on a given interval, and it is differentiable at an interior pont x. Also suppose that $f'(x) \neq 0$. Then f^{-1} is differentiable at $y = f(x)$, and*

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

Roughly, the situation is the following. Consider the difference quotient:

$$\frac{f^{-1}(y+h) - f^{-1}(y)}{h}$$

Let x and $x + k$ be points in the domain of f such that $y = f(x)$ and $y + h = f(x + k)$. Then the difference quotient can be written in the following form:

$$\frac{x+k-x}{f(x+k) - f(x)} = \frac{1}{\frac{f(x+k)-f(x)}{k}}$$

If here $h \to 0$, then $k \to 0$ (ATTENTION, this is not trivial! It means the continuity of f^{-1}.), and hence, the limit of the fraction on the right-hand side is really $1/f'(x)$.

Example 5.3 Find the derivative of the function

$$g(x) = \sqrt[n]{x}$$

at a point $x > 0$. As we see, g is the inverse of the power function $f(x) = x^n$ on the non-negative half line, that is $g(y) = f^{-1}(y)$. Thus,

$$g'(y) = \frac{1}{f'(x)} = \frac{1}{nx^{n-1}} = \frac{1}{n} \cdot y^{\frac{1}{n}-1}$$

since $y = x^n$ and consequently

$$x^{n-1} = y^{\frac{n-1}{n}}$$

In view of this example we conclude that for every rational exponent r the function $F(x) = x^r$ is differentiable at every point $x > 0$, and its derivative is:

$$F'(x) = rx^{r-1} .$$

Example 5.4 Calculate the derivative of the function

$$F(x) = \sqrt{1+x^4}$$

Set $f(x) = \sqrt{x}$ and $g(x) = 1 + x^4$, with these notations we have $F = f \circ g$. Making use of the Chain-Rule we get

$$F'(x) = f'(g(x)) \cdot g'(x) = \frac{4x^3}{2\sqrt{1+x^4}}$$

5.3 The exponential and logarithm functions

Consider the exponential function with base e on the real line, and its inverse, which is the logarithm function with base e (that is denoted by the symbol ln):

$$f(x) = e^x \qquad f^{-1}(x) = \ln x \quad (x > 0) \ .$$

They are called the *natural* exponential function, and the *natural* logarithm function, respectively. Below we find their derivatives. We start with the equality

$$\lim_{x \to \pm\infty} \left(1 + \frac{1}{x}\right)^x = e \ .$$

Find the derivative of the natural logarithm function at $x_0 = 1$.

$$\frac{\ln(1+h) - \ln 1}{h} = \ln(1+h)^{1/h}$$

whose common right-hand limit and left-hand limit at zero is $\ln e$. (Here we supposed the continuity of the logarithm function.) Therefore, the derivative is 1.

The derivative of $f(x) = e^x$ at the point 0 can be determined by exploiting our theorem about the differentiability of the inverse function:

$$f'(0) = \lim_{h \to 0} \frac{e^h - 1}{h} = \frac{1}{(\ln)'(1)} = 1 \ .$$

This enables us to get the derivative of the exponential function at an arbitrary point x:

$$f'(x) = \lim_{h \to 0} \frac{e^{x+h} - e^x}{h} = e^x \cdot \lim_{h \to 0} \frac{e^h - 1}{h} = e^x$$

Using the differentiability of the inverse again, we obtain the derivative of the logarithm function at any given point $x > 0$:

$$(f^{-1})'(x) = \frac{1}{e^{\ln x}} = \frac{1}{x} \ .$$

Example 5.5 As a straightforward application, find the derivative of the function

$$f(x) = x^\alpha$$

at any given point $x > 0$, where α is an arbitrary real exponent. First we write:

$$f(x) = x^\alpha = e^{\alpha \ln x}$$

Then, in view of the Chain-Rule we get

$$f'(x) = \alpha \frac{1}{x} e^{\alpha \ln x} = \alpha \frac{1}{x} x^\alpha = \alpha x^{\alpha - 1}$$

This tells us that the differentiation can be carried out the same way as in the case of rational exponents.

5.4 Necessary condition for an extremum

Consider a function $f : \mathbb{R} \to \mathbb{R}$.

Definition 5.6 We say that a point x_0 in the domain of f is a (global) minimum point, if $f(x_0) \leq f(x)$ for every point $x \neq x_0$ in the domain of f.

We say that a point x_0 in the domain of f is a local minimum point, if there exists a positive number $\varepsilon > 0$ such that $f(x_0) \leq f(x)$ at every point in the domain x with $0 < |x - x_0| < \varepsilon$.

In both cases we strict minimum points if strict inequalities apply.

We can formulate analogous definitions for maximum points.

It is obvious that a global minimum point is also a local minimum point. The converse statement however, is not true in general, as it is shown in the following example. For instance, the function

$$f(x) = \begin{cases} (x+1)^2 & \text{ha } x < 0 \\ (x-1)^2 & \text{ha } x \geq 0 \end{cases}$$

admits a local maximum at $x = 0$ (here the function is continuous, but not differentiable, check it!) but this function does not have a global maximum, since it is not bounded from above.

For differentiable functions we can present the following charcterization of local extreme (minimum or maximum) points.

Theorem 5.7 *Let us suppose that f is defined on an interval, at it is differentiable at an interior point x_0. If x_0 is a local minimum point of f, then $f'(x_0) = 0$.*

Proof. Indeed, consider the difference quotient:

$$\frac{f(x_0+h)-f(x_0)}{h} .$$

If $h>0$, then the difference quotient for small values of h is non-negative, and consequently, the right-hand limit is non-negative. On the other hand, if $h<0$, then similarly, the left-hand limit is non-positive. By the differentiability assumption the difference quotient has a limit when $h \to 0$, which therefore, can only be zero. Thus $f'(x_0)=0$.

This theorem formulates only a necessary condition for minimum, which is not sufficient! For example, the function $f(x)=x^3$ has no extereme point at $x=0$, but $f'(0)=0$.

In the case of a differentiable function, those points x_0 where $f'(x_0)=0$, are called *critical* (or sometimes stationary) points. Using this vocabulary, we may say that the extreme points of a function are critical, the converese statement is not necessarily true.

5.5 Lagrange's Mean Value Theorem

Based on the geometric interpretation, the Mean Value Theorem formulates a very illustrative statement.

Theorem 5.8 *Let f be continuous on the finite closed interval $[a,b]$, and differentiable in the interior of the interval. Then there exists a point $\xi \in (a,b)$ so that*

$$f'(\xi)=\frac{f(b)-f(a)}{b-a}$$

Proof. Introduce the function

$$g(x)=f(x)-f(a)-\frac{f(b)-f(a)}{b-a}(x-a)$$

According to the assumptions, this function is continuous on the interval $[a,b]$, hence, by Weierstrass' Theorem (see Theorem 3.15) it achieves its minimum and maximum in $[a,b]$ intervallumon. At least one of the extreme points (either the minimum, or the maximum) is in the interior of the interval, because

$$g(a)=g(b)=0 .$$

If this interior extreme point is $\xi \in (a,b)$, then by our previous theorem $g'(\xi)=0$. This exactly means that

$$f'(\xi)-\frac{f(b)-f(a)}{b-a}=0 .$$

Please observe, that the continuity assumption in our theorem is vital! Sketch a figure to show that!

5.6 L'Hôpital's Rule

The procedure below makes it possible to compute complicated limits relatively easily.

Let both f and g be differentiable, and their derivatives f' and g' are continuous in a neighborhood of a point x_0, and suppose that $f(x_0) = g(x_0) = 0$. We want to find the limit

$$\lim_{x \to x_0} \frac{f(x)}{g(x)}$$

which is of the form $0/0$ so "undefined".

By the Mean Value Theorem

$$\frac{f(x)}{g(x)} = \frac{\frac{f(x)-f(x_0)}{x-x_0}}{\frac{g(x)-g(x_0)}{x-x_0}} = \frac{f'(\xi)}{g'(\eta)}$$

where ξ and η are points between x and x_0. Now if $x \to x_0$, then both $\xi \to x_0$ and $\eta \to x_0$. Therefore, by the continuity of the derivative functions we get

$$\lim_{x \to x_0} \frac{f(x)}{g(x)} = \lim_{x \to x_0} \frac{f'(x)}{g'(x)}$$

This equality is called L'Hôpital's Rule. If the resulting limit still has the form $0/0$, then apply L'Hôpital's Rule again until a "decent" limit is received.

Example 5.9 Find the following limit by using L'Hôpital's Rule:

$$\lim_{x \to 0} \frac{2 \sin x}{1 - \sqrt{1+x}}$$

Taking the limits of the derivatives, we have:

$$\lim_{x \to 0} \frac{2 \sin x}{1 - \sqrt{1+x}} = \frac{2 \cos 0}{-\frac{1}{2\sqrt{1+0}}} = -4$$

Study at home:

1. Review of the exercises in "Mathematical Analysis Exercises"
2. Textbook-1, Sections 5.1, 5.4, 7.5 and 7.6, Chapter 8.

Chapter 6

Complete analysis of functions

6.1 Monotone functions

Definition 6.1 We say that f is monotone increasing on an interval, if for any two points of the interval with $x_1 < x_2$ we have $f(x_1) \leq f(x_2)$. An analogous definiton applies for monotone decreasing functions.

We say that the function is strictly monotone (in either case), if we have strict inequalities in the definition.

Theorem 6.2 *Let f be continuous on a finite closed interval $[a, b]$, and differentiable in its interior. If we have $f'(x) > 0$ at every interior point of the interval, then f is strictly monotone increasing on $[a, b]$.*

Indeed, if $x_1 < x_2$ are two arbitrary points of the interval $[a, b]$, then by the Lagrange's Mean Value Theorem there exists a point $x_1 < \xi < x_2$, such that

$$f(x_2) - f(x_1) = f'(\xi)(x_2 - x_1) \ .$$

By our assumption the right-hand side is positive, therefore

$$f(x_2) - f(x_1) > 0$$

that means f is strictly monotone increasing on the interval.

Now, let us examine a function that is monotone increasing and differentiable in an interval. For any two different points x and $x + h$ in the interval we have:

$$\frac{f(x+h) - f(x)}{h} \geq 0$$

regardless of $h > 0$ or $h < 0$. Passing to the limit $h \to 0$ we obtain $f'(x) \geq 0$. Thus, we can formulate the following theorem.

Theorem 6.3 *Let f be continuous on the interval $[a, b]$, and differentiable in its interior. Then f is monotone increasing on the interval if and only if $f'(x) \geq 0$ at each interior point of the interval.*

A completely similar statement can be formulated for monotone decreasing functions.

However, the assertion that if f is strictly monotone increasing, then we would have $f'(x) > 0$ for every interior point x is NOT TRUE. For example, the function $f(x) = x^3$ is strictly monotone increasing on the entire real line, but $f'(0) = 0$.

6.2 Finding extreme points

Consider a function $f : \mathbb{R} \to \mathbb{R}$ and pick an interior point x_0 in the domain. Suppose that f is differentiable at x_0.

As we have seen, the necessary condition for x_0 for being an extreme point is $f'(x_0) = 0$. The question is, how we can formulate a sufficient condition for really having an extremum at x_0. It is easy to see that if there exists a positive number $\varepsilon > 0$ so that f is monotone decreasing on the interval $[x_0 - \varepsilon, x_0]$, moreover f is monotone increasing on the interval $[x_0, x_0 + \varepsilon]$, then x_0 is definitely a local minimum point f lokális.

For differentiable functions we can summarize this observation in the following theorem.

Theorem 6.4 *Assume that f is differentiable in an interval, and x_0 is an interior point of the interval. If there exists a positive number $\varepsilon > 0$, so that*

- $f'(x) \leq 0\ ,$ *if* $x \in (x_0 - \varepsilon, x_0)$
- $f'(x) \geq 0\ ,$ *if* $x \in (x_0, x_0 + \varepsilon)$

then x_0 is a local minimum point of f.

Obviously, an analogous statement can be formulated for the case of local maximum as well.

Example 6.5 Find the extreme points and the intervals of monotonicity of the function

$$f(x) = x^2 e^{-x}$$

By the Product-Rule, the derivative is:

$$f'(x) = (2x - x^2)e^{-x}$$

whose sign depends exclusively on the first factor (the second is positive). Consequently:

- If $x \in (-\infty, 0)$, then $f'(x) < 0$, so f is monotone decreasing.
- If $x = 0$, then $f'(0) = 0$, this is a critical point.
- If $x \in (0, 2)$, then $f'(x) > 0$, so f is monotone increasing.
- If $x = 2$, then $f'(2) = 0$, this is another critical point.
- If $x \in (2, +\infty)$, then $f'(x) < 0$, so f is monotone decreasing.

By the changing the signs of f' we can conclude that $x = 0$ is a minimum point (global), while $x = 2$ is a local maximum point.

Example 6.6 Consider the following function on the real line:

$$f(x) = x + \sin x$$

Since $f'(x) = 1 + \cos x$, it is clear that function has critical points at

$$x = (2k+1)\pi \qquad k = 0, \pm 1, \pm 2, \ldots$$

However, none of them is an extremum:

$$x \neq (2k+1)\pi \quad \text{then} \quad f'(x) > 0 \ ,$$

because $\cos x > -1$. This means that the derivative does not change its sign. In fact, this function is strictly monotone increasing on the entire real line.

6.3 Higher order derivatives

If a function f is differentiable in a given interval, then the correspondence $x \to f'(x)$ is called the derivative function of f. If f' is again differentiable at a given point x_0, then we say that f is twice differentiable at this point. Instead of using the complicated notation $(f')'(x_0)$, we use the brief formula

$$f''(x_0)$$

and this is called the second derivative of f at x_0.

In a completely similar way, if n is a given integer, we can define the n-the derivative of the function f at x_0, and its notation is

$$f^{(n)}(x_0)\,.$$

For instance, for the function $f(x) = 1/x$ at any given point $x_0 \neq 0$ we have

$$f''(x_0) = \frac{2}{x_0^3} \qquad \text{and} \qquad f^{(n)}(x_0) = \frac{(-1)^n n!}{x^{n+1}}$$

for every integer n.

Example 6.7 Consider the function $f(x) = \sin x$, and find its derivative function.

$$\begin{aligned} f'(x) &= \lim_{h\to 0} \frac{\sin(x+h) - \sin x}{h} = \lim_{h\to 0} \frac{\sin x \cos h + \cos x \sin h - \sin x}{h} \\ &= \sin x \cdot \lim_{h\to 0} \frac{\cos h - 1}{h} + \cos x \cdot \lim_{h\to 0} \frac{\sin h}{h} \end{aligned}$$

In view of Example 3.12 the first limit is 0, and in view of Example 3.8 the second limit is 1. Therefore,

$$f'(x) = \cos x$$

By using the identity $\cos x = \sin(x + \pi/2)$ and the Chain-Rule, we have

$$(\cos x)' = \cos(x + \pi/2) = -\sin x$$

Therefore, the higher order derivatives of $f(x) = \sin x$ can be given in terms of the divisibility by 4:

$$f^{(n)}(x) = \begin{cases} \cos x & \text{if } n = 4k+1 \\ -\sin x & \text{if } n = 4k+2 \\ -\cos x & \text{if } n = 4k+3 \\ \sin x & \text{if } n \text{ is divisible by } 4 \end{cases}$$

6.4 Second order conditions

It may happen that we analyze a function, where the sign of its derivative is not easy to determine (for instance a higher degree polynomial). In a case like that, the second order (sufficient) condition proves to be useful.

Theorem 6.8 *Let f be differentiable in an interval, and suppose that f is twice differentiable at an interior point x_0.*

If $f'(x_0) = 0$ and $f''(x_0) > 0$, then x_0 is a local minimum point of f.

Proof. Indeed, by examining the different quotient we get

$$\begin{aligned} f''(x_0) &= \lim_{h\to 0} \frac{f'(x_0+h) - f'(x_0)}{h} = \\ &= \lim_{h\to 0} \frac{f'(x_0+h)}{h} > 0 \end{aligned}$$

This means that the quotient $f'(x_0+h)/h$ is positive for $0 < |h| < \varepsilon$ for some $\varepsilon > 0$. This implies that

- if $x \in (x_0 - \varepsilon, x_0)$, then $f'(x) < 0$,
- if $x \in (x_0, x_0 + \varepsilon)$, then $f'(x) > 0$.

Making use of Theorem 6.4 we conclude that x_0 is really a local minimum point.

We can formulate an analogous second order sufficient condition for the case of local naximum.

By using proof by contradiction, we get the second order necessary condition for an extremum point.

Theorem 6.9 *Assume that f is twice differentiable in an interval, and let x_0 be an interior point of the interval.*

- *If x_0 is a local minimum point, then $f'(x_0) = 0$, and $f''(x_0) \geq 0$.*
- *If x_0 is a local maximum point, then $f'(x_0) = 0$, and $f''(x_0) \leq 0$.*

Example 6.10 For $x > 0$ consider the function

$$f(x) = x \ln x$$

Then $f'(x) = 1 + \ln x$, therefore, the only critical point of f is $x = 1/e$. On the other hand $f''(x) = 1/x$, so we have

$$f''(1/e) = e > 0 \ ,$$

Thus, $x = 1/e$ is a local minimum point of f. (It is not hard to verify that this is a global minimum point as well.)

Please observe that our theorems provide no information for a critical point x_0 with

$$f''(x_0) = 0 \ .$$

The reason that in this "marginal" situation anything can happen. For example, examine the behavior of the functions

$$f(x) = x^n \qquad (n \geq 3)$$

at the critical point $x_0 = 0$. On the one hand, here $f'(0) = 0$ and $f''(0) = 0$. On the other hand

- if n is even, then $x_0 = 0$ is (global) minimum point,
- if n is odd, then $x_0 = 0$ is not an extremum point (so-called saddle point).

Very similarly, if n is even, then $x_0 = 0$ is a (global) maximum point of $-f$.

6.5 Convex and concave functions

Definition 6.11 The function f is said to be convex on the interval $[a, b]$, if for any two points x_1 and x_2 from the interval, and for any real number $0 \le \alpha \le 1$

$$f(\alpha x_1 + (1-\alpha)x_2) \le \alpha f(x_1) + (1-\alpha)f(x_2)\ .$$

The geometric meaning of this definition is that any cord to the graph (i.e. a segment that connects two points on the graph) can nowhere be below the graph of the function.

Concave functions are defined by the opposite inequality.

We now give a simple characterization of convexity for twice differentiable functions.

Theorem 6.12 *Assume that f is continuous on a closed interval, and twice differentiable in the interior. The necessary and sufficient condition for the convexity of f is:*

$$f''(x) \ge 0$$

at every interior point of the interval.

In particular, this means that for convex functions the slope of the tangent line (i.e. the derivative) is monotone increasing. Geometrically this can be illustrated by the fact that the graph of the function is nowhere below the tangent line.

Example 6.13 Give a complete analysis of the function

$$f(x) = \frac{x}{1+x^2}$$

First we calculate the derivative:

$$f'(x) = \frac{1-x^2}{(1+x^2)^2}\ .$$

By examining the sign of the derivative, we come up with the following summary:

- f is strictly monotone decreasing on the interval $(-\infty, -1)$
- $x = -1$ is a (global) minimum point
- f is strictly monotone increasing on the interval $(-1, 1)$
- $x = 1$ is a (global) maximum point
- f is strictly monotone decreasing on the interval $(1, +\infty)$.

The convexity is investigated by specifying the sign of the second derivative:

$$f''(x) = \frac{2x^3 - 6x}{(1 + x^2)^3}$$

Obviously, the denominator is positive, so it is enough to find the sign of the numerator:

$$2x^3 - 6x = 2x(x^2 - 3)$$

By examining the factors we come up with the following summary:

- f is concave on the interval $(-\infty, -\sqrt{3})$
- f is convex on the interval $(-\sqrt{3}, 0)$
- f is concave on the interval $(0, \sqrt{3})$
- f is convex on the interval $(\sqrt{3}, +\infty)$.

Please notice that we have $f''(-\sqrt{3}) = f''(0) = f''(\sqrt{3}) = 0$, and the second derivative changes the sign at those points. In other words those points separate the convex and concave segments of the function. Such point are called the *points of inflection* of f. At a point of inflection the tangent line intersects the graph of the function.

Probably the most important property of convex function is that every local minimum point is a global minimum point as well.

Theorem 6.14 *Consider a twice differentiable convex function f on an interval, and let x_0 be an interior point of the interval. If x_0 is a local minimum point, then it is a global minimum point.*

Proof. Indeed, on the one hand $f'(x_0) = 0$, on the other hand f' is monotone increasing. Therefore, at every interior point x_0:

- if $x < x_0$, then $f'(x) \leq 0$, and hence, $f(x) \geq f(x_0)$,
- if $x > x_0$, then $f'(x) \geq 0$, and hence, $f(x) \geq f(x_0)$.

This proves our statement.

A completely analogous theorem can be formulated for concave functions and maximum points.

Example 6.15 Define the function f for $x > 0$ on the positive half line:

$$f(x) = ax + 2\ln x$$

where a is an unspecified parameter. For what value of a will f possess a global maximum point at $x = 6$?

By the necessary condition for an extremum

$$f'(x) = a + \frac{2}{x} = 0$$

that yields $x = -2/a$. By the condition $x = 6$, we get $a = -1/3$. The second derivative of f is:

$$f''(x) = -\frac{1}{x^2} < 0\,,$$

therefore the function is concave on the whole domain. Consequently, for the parameter $a = -1/3$ the function f has a global maximum point at $x = 6$.

Study at home:

1. Careful review of the "Mathematical Analysis Exercises"
2. Textbook-1: Chapter 9.

Chapter 7

Integration

7.1 The indefinite integral

Definition 7.1 Let f be a function defined on an interval I. A differentiable function F defined on I is called the *indefinite integral* of f, or sometimes its *primitive function*, if

$$F'(x) = f(x)$$

for every $x \in I$.

It is clear that taking the indefinite integral is the reverse operation of differentiation. It is important to note that the indefinite integral is not unique! Indeed, if F is the indefinite integral of a function f, then by adding a constant C to F we again have an indefinite integral:

$$(F(x) + C)' = F'(x) = f(x)$$

for every $x \in I$.

We show that this is the only way to create other indefinite integrals.

Theorem 7.2 *If F is an idefinite integral of f on the interval I, then any indefinite integral of f can be given in the form $F + C$, where C is a constant.*

Proof. Indeed, if the differentiable function G is an indefinite integral of f on the interval I, then at every point $x \in I$ we have

$$(F(x) - G(x))' = f(x) - f(x) = 0$$

This means that the derivative of $F - G$ is zero on I. By the Mean Value Theorem we get that $F - G$ is constant on the interval.

In view of our theorem, we use the following notation for indefinite integrals:

$$\int f(x)\,dx = F(x) + C$$

For instance, by simple differentiation we can verify

$$\int \cos x\,dx = \sin x + C$$

or very similarly

$$\int x^\alpha\,dx = \frac{x^{\alpha+1}}{\alpha+1} + C \qquad (\alpha \neq -1)$$

where C is an arbitrary constant. If a function has an indefinite integral on an interval, then there are infinitely many of them.

7.2 Basic integrals

The following rule can be useful for finding indefinite integrals:

Theorem 7.3 $\int (\alpha f(x) + \beta g(x))\,dx = \alpha \int f(x)\,dx + \beta \int g(x)\,dx$

This rule can be extended to any sums with finitely many terms.

ATTENTION: Not all functions have indefinite integrals. For example, a function f with a point of discontinuity, where the one-sided limits exist, they are finite, but not equal, cannot possess an indefinite integral. The following theorem formulates a useful sufficient condition for the existence of the indefinite integral.

Theorem 7.4 *If f is continuous on the interval I, then it has an indefininite integral.*

We can easily create rules for finding indefinite integrals by reversing the differentiation rules. By taking the opposites of differentiation rules for elementary functions, we obtain rules for finding indefinite integrals.

In general, any formula for an indefinite integral can be verified by direct differentiation. For example:

$$\int \sin x\,dx = -\cos x + C$$

$$\int (2x^2 - 5x + 8)\,dx = \frac{2}{3}x^3 - \frac{5}{2}x^2 + 8x + C$$

$$\int e^{2x-1}\,dx = \frac{1}{2}e^{2x-1} + C$$

$$\int \frac{2x}{1+x^2}\,dx = \ln(1+x^2) + C$$

7.3 Initial value problems

As we have seen, a function can have infinitely many indefinite integrals (if any exists), and they differ only in an additive constant. However, if fix a point in the coordinate system, and looking only for a definite integral that passes through the given point, then the solution of the problem may be unique.

Example 7.5 Find the function F for which

$$F'(x) = 2e^{-x} \qquad \text{and} \qquad F(0) = 1$$

In this case we are looking for a specific indefinite integral

$$F(x) = 2\int e^{-x}\,dx = -2e^{-x} + C$$

so that $F(0) = 1$. The condition implies $C = 3$, and this is the only solution.

7.4 Definite integrals

In this section we briefly outline how Berhard Riemann, professor of mathematics at University of Göttingen (Germany) introduced the concept of integration in the 19-th century. The idea is based on the two-sided approximation developed Archimedes, the ancient greek mathematician. This idea is a fundamental element of human thinking, and this is how Archimedes determined the area of the circle in Syracuse, using the areas of approximating polygons from inside and outside.

Let f be a continuous function on the finite interval $[a, b]$, and consider the partition of the interval into n subintervals by using the points

$$a = x_0 < x_1 < \ldots < x_n = b$$

On every subinterval $[x_{k-1}, x_k]$ let m_k denote minimum value of f, and let M_k denote the maximum value of f. Those extreme values exist by virtue of Weierstrass' theorem (see Theorem 3.15). Create the sum

$$s_n = \sum_{k=1}^{n} m_k(x_k - x_{k-1})$$

that we call *lower sum*, and the sum

$$S_n = \sum_{k=1}^{n} M_k(x_k - x_{k-1})$$

that we call *upper sum*. The sum the areas of these rectangles approximate the area below the graph of f from below, and from above, respectively. Check Figures.pdf for details!

We can easily see that by inserting a new node point s_n cannot decrease, and S_n cannot increase. It can be shown that if the density of the partion gets higher then the lowest upper bound of the lower sums coincides with the highest lower bound of the upper sums. Following Riemann's idea, this common value S is called the definite integral of f on the interval $[a, b]$. The notation is:

$$S = \int_a^b f(x)\,dx$$

which means the (signed!) area below the graph of f.

ATTENTION!

The area above the x-axis comes with positive sign, the area below the x-axis comes with negative sign, respectively.

Based on this geometric interpretation, the following properties of the definite integral are intuitively obvious.

Theorem 7.6 *Let f and g be functions that have definite integrals on $[a, b]$. Then*

1. *if $f(x) \le g(x)$ on the interval $[a, b]$, then*

$$\int_a^b f(x)\,dx \le \int_a^b g(x)\,dx$$

2. *in particular, $|\int_a^b f(x)\,dx| \le \int_a^b |f(x)|\,dx$.*

3. *If $f(x) \le M$ on the interval $[a, b]$ (M is a constant), then*

$$\int_a^b f(x)\,dx \le M(b-a)$$

4. *If f is continuous on the interval $[a, b]$, then there exists a point $\overline{x} \in [a, b]$, for which $\int_a^b f(x)\,dx = f(\overline{x})(b-a)$.*

5. *By definition: $\int_b^a f(x)\,dx = -\int_a^b f(x)\,dx$ if $a \le b$.*

6. *$\int_a^c f(x)\,dx = \int_a^b f(x)\,dx + \int_b^c f(x)\,dx$*

Create a picture, and interpret the above statements geometrically!

7.5 Newton-Leibniz-formula

In this section we show how a definite integral can be evaluated by using the indefinite integral (primitive function). Our main result is sometimes called the "Fundamental Theorem of Calculus" (in the English literature).

Theorem 7.7 (Newton-Leibniz-formula) *If F is a primitive function of the continuous function f on the finite interval $[a, b]$, then*

$$\int_a^b f(x)\,dx = F(b) - F(a)$$

Justification (not a proof!): It is easy to see that our statement is independent of the choice of the indefinite integral. Indeed, if G is another primitive function of f, then

$$G(x) = F(x) + C$$

on $[a, b]$ with a constant C (see Theorem 7.2), and therefore,

$$\int_a^b f(x)\,dx = [G(x)]_a^b = G(b) - G(a) = (F(b) + C) - (F(a) + C) = F(b) - F(a)\,.$$

On the other hand, fix a point $x \in [a, b]$ and consider the integral

$$F(x) = \int_a^x f(t)\,dt$$

Then $F(a) = 0$, since the length of the path of integration is zero. It would be enough to show that this F is an indefinite integral of f.

In view of Theorem 7.6 for any $a < x < b$ and $h \neq 0$ with $x + h \in [a, b]$, there exists a point $\overline{x}$ between x and $x + h$ with the following property:

$$\frac{1}{h}(F(x+h) - F(x)) = \frac{1}{h}\int_x^{x+h} f(t)\,dt = \frac{1}{h} f(\overline{x}) \cdot h$$

Now, if we pass to the limit $h \to 0$, then $\overline{x} \to x$, and by the continuity of f we also have $f(\overline{x}) \to f(x)$ that is

$$\lim_{h\to 0} \frac{1}{h}(F(x+h) - F(x)) = F'(x) = \lim_{h\to 0} f(\overline{x}) = f(x)$$

This means that F is really a primitive function of f. □

It was an amazing achievment by Newton and Leibniz, and the mathematics of their time, to find the beautiful relationship between the derivative and the geometry of definite integrals, as it is described in our theorem.

This discovery is so fundamental that it cannot be overestimated. First, it triggered a very rapid developement in physics and chemistry, and somewhat later it gave a massive boost to the evolution of sciences like biology, economics and others. Summing up, we may say today that the theory of differentiation and integration provides the precise scientific language and vocabulary in all branches of sciences.

For convenience, sometimes we use the following notation:

$$\int_a^b f(x)\,dx = [F(x)]_a^b = F(b) - F(a)$$

As a consequence of the Newton-Leibniz-formula, we can formulate the following statement.

Consequence 7.8 *If f is continuous on an interval, then it has a primitive function on that interval.*

Proof. In view of the proof of the Newton-Leibniz-formula, we get that the function

$$F(x) = \int_a^x f(t)\,dt$$

is really a primitive function of f on the given interval. □

Example 7.9 Evaluate the definite integral below.

$$\int_1^2 \left(2x^3 + 1 + \frac{1}{x^2}\right)\,dx = \left[\frac{x^4}{2} + x - \frac{1}{x}\right]_1^2 = 9$$

Some more examples:

$$\int_0^{\pi/2} \sin x\,dx = [-\cos x]_0^{\pi/2} = 1$$

$$\int_0^1 e^x\,dx = [e^x]_0^1 = e - 1$$

$$\int_0^4 \sqrt{x}\,dx = \left[\frac{2}{3} \cdot x^{3/2}\right]_0^4 = \frac{16}{3}$$

Study at home:

1. Careful review of "Mathematical Analysis Exercises"
2. Textbook-1, Chapter 10.

Chapter 8

Methods of integration

8.1 Integration by parts

If f and g are continuously differentiable functions on an interval I, then by the Product-Rule we have:

$$\int f'(x)g(x)\,dx = f(x)g(x) - \int f(x)g'(x)\,dx$$

This formula is called *integration by parts*. For example, consider the integral

$$\int xe^{-x}\,dx$$

then by using the allocation $f'(x) = e^{-x}$ and $g(x) = x$ (could we do it the other way?):

$$\int xe^{-x}\,dx = -xe^{-x} + \int e^{-x}\,dx = -xe^{-x} - e^{-x} + C$$

Example 8.1 Use integration by parts in the integral

$$\int x^n \ln x\,dx$$

(where $n \neq -1$). Introduce the notation $f'(x) = x^n$ and $g(x) = \ln x$, then (what do we get in the opposite way?)

$$\int x^n \ln x\,dx = \frac{x^{n+1}}{n+1}\ln x - \int \frac{x^n}{n+1}\,dx = \frac{x^{n+1}}{n+1}\ln x - \frac{x^{n+1}}{(n+1)^2} + C$$

In particular, for $n = 0$ we have:

$$\int \ln x\,dx = x\ln x - x + C = x(\ln x - 1) + C$$

8.2 Integration by parts in definite integrals

We can use integration by parts in definite integrals in the following way:

$$\int_a^b f'(x)g(x)\,dx = [f(x)g(x)]_a^b - \int_a^b f(x)g'(x)\,dx$$

For istance, by setting $f'(x) = \sin x$ and $g(x) = x$ (would the opposite way successful?):

$$\begin{aligned}\int_0^\pi x\sin x\,dx &= [-x\cos x]_0^\pi + \int_0^\pi \cos x\,dx \\ &= \pi + [\sin x]_0^\pi = \pi\end{aligned}$$

This procedure is faster than first computing the indefinite integral and then substituting the bounds. Further, it may minimize the chance of miscalculation.

Example 8.2 Sometimes we need to carry out integration by parts more times in a row. Consider the integral

$$\int x^2 e^{-\lambda x}\,dx$$

where $\lambda > 0$ is a given parameter. Introduce the notations $f'(x) = e^{-\lambda x}$, and $g(x) = x^2$, then

$$\int x^2 e^{-\lambda x}\,dx = -\frac{1}{\lambda}x^2 e^{-\lambda x} + \frac{1}{\lambda}\int 2x e^{-\lambda x}\,dx$$

The last integral can be evaluated by a repeated integration by parts.

Attention! We stick to the notations $f'(x) = e^{-\lambda x}$ and $g(x) = x$. In the opposite situation we come to an absolutely useless identity. Give it a try!

$$\int x^2 e^{-\lambda x}\,dx = -\frac{1}{\lambda}x^2 e^{-\lambda x} - \frac{2}{\lambda^2}x e^{-\lambda x} - \frac{2}{\lambda^3}e^{-\lambda x} + C$$

Example 8.3 Find the definite integral below:

$$\int_0^\pi e^x \sin x\,dx$$

Apply the setting $f'(x) = e^x$ and $g(x) = \sin x$, the by two consecutive integrations by parts:

$$\begin{aligned}\int_0^\pi e^x\sin x\,dx &= [e^x \sin x]_0^\pi - \int_0^\pi e^x\cos x\,dx \\ &= -[e^x\cos x]_0^\pi - \int_0^\pi e^x \sin x\,dx\end{aligned}$$

Isolate the original integral on the laft-hand side:

$$2\int_0^\pi e^x \sin x\,dx = -[e^x \cos x]_0^\pi$$

which means

$$\int_0^\pi e^x \sin x\,dx = \frac{1}{2}(e^\pi + 1)$$

8.3 Integration by substitution

From the diffentiation of a composition of functions (i.e. the Chain-Rule) we derive the following identity:

$$\int f(g(t))g'(t)\,dt = \int f(x)\,dx$$

where $x = g(t)$ is a continuously differentiable function on an interval. This formula is called the *integration by substitution.*

Example 8.4 Calculate the following indefinite integral:

$$\int 5t^3\sqrt{2+t^4}\,dt$$

Observe that by introducing the substitution $x = g(t) = t^4$, the integral can be rewritten in this form:

$$\int 5t^3\sqrt{2+t^4}\,dt = \frac{5}{4}\int \sqrt{2+x}\,dx = \frac{5}{4}\cdot\frac{2}{3}(2+x)^{3/2} + C$$

By performing the backsubstitution:

$$\int 5t^3\sqrt{2+t^4}\,dt = \frac{5}{6}(2+t^4)^{3/2} + C$$

Example 8.5 Consider an example, where the converse approach is useful:

$$\int e^x\sqrt{1+e^x}\,dx$$

Introduce the substitution $x = g(t) = \ln t$, then $g'(t) = 1/t$, and we obtain:

$$\int e^x\sqrt{1+e^x}\,dx = \int t\sqrt{1+t}\frac{1}{t}\,dt = \frac{2}{3}(1+t)^{3/2} + C$$

By the backsubstitution $t = e^x$ we get:

$$\int e^x\sqrt{1+e^x}\,dx = \frac{2}{3}(1+e^x)^{3/2} + C$$

8.4 Substitution in definite integrals

When substitution is applied in definite integrals, instead of backsubstitution, it is much more efficient to change the bounds of the integral according the substitution:

$$\int_a^b f(g(t))g'(t)\,dt = \int_{g(a)}^{g(b)} f(x)\,dx$$

Example 8.6 In the example below we use the setting $x = g(t) = \cos t$, then $g'(t) = -\sin t$, and

$$\begin{aligned}
\int_0^{\pi/2} \frac{\sin 2t}{1+\cos^2 t}\,dt &= \int_0^{\pi/2} \frac{2\sin t\cos t}{1+\cos^2 t}\,dt \\
&= -\int_1^0 \frac{2x}{1+x^2}\,dx = \int_0^1 \frac{2x}{1+x^2}\,dx \\
&= [\ln(1+x^2)]_0^1 = \ln 2
\end{aligned}$$

Example 8.7 Apply this rule to evaluate the following celebrated integral:

$$\int_0^1 \sqrt{1-x^2}\,dx$$

Introduce the substitution $x = g(t) = \sin t$, then $g'(t) = \cos t$ and (please observe the change of the bounds of the integral!):

$$\begin{aligned}
\int_0^1 \sqrt{1-x^2}\,dx &= \int_0^{\pi/2} \cos^2 t\,dt \\
&= \frac{1}{2}\int_0^{\pi/2} (1+\cos 2t)\,dt = \frac{1}{2}\left[t + \frac{\sin 2t}{2}\right]_0^{\pi/2} = \frac{\pi}{4}
\end{aligned}$$

The geometric interpretation of this example is as follows. We determined the area of the first qudrant of the unit circle with center at the origin!

8.5 Linear differential equations

By a differential equation we mean an equation in which the unknown function and its derivative appear. Several problems and models in micro and macroeconomics lead to such equations. A typical equation like that is the linear differential equation.

Let a and b are given real numbers, and we are looking for the unknown differentiable function y for which

$$\begin{aligned} y' &= ay + b \\ y(0) &= y_0 \end{aligned} \tag{8.1}$$

where y_0 is an "a priori" given real number.

The equality $y(0) = y_0$ is called the initial condition. We say that the differentiable function y is a solution to the above problem, if for any $t \in \mathbb{R}$ we have $y'(t) = ay(t) + b$, moreover $y(0) = y_0$. The question is, how to find the solution of this problem?

Let us suppose that y is a solution. Multiply both sides of the equation by the expression e^{-at}, then after rearranging the terms, we get

$$y'(t)e^{-at} - ay(t)e^{-at} = be^{-at}$$

for every real number t. valós számra. Observe that on the left-hand side we have precisely the derivative of the product $y(t)e^{-at}$. Therefore, by integrating both sides from 0 to t-ig (and changing the variable of the integration from t to s)

$$\int_0^t (y'(s)e^{-as} - ay(s)e^{-as})\, ds = \left[y(s)e^{-as}\right]_0^t = \int_0^t be^{-as}\, ds$$

By plugging in the bounds we receive

$$y(t)e^{-at} - y(0) = \int_0^t be^{-as}\, ds\,.$$

Rearranging and multiplying both sides by the expression e^{at} we can formulate our result in the following theorem.

Theorem 8.8 (Cauchy-formula) *The solution to problem (8.1) is given by*

$$y(t) = e^{at}\left(y_0 + \int_0^t be^{-as}\, ds\right)$$

on the entire real line.

Recall that without prescribing the initial condition $y(0) = y_0$ the linear differential equation (8.1) would possess infinitely many solutions.

Example 8.9 For instance, if we are looking for the solution of the linear differential equation

$$\begin{aligned} y' &= 2y + 5 \\ y(0) &= 3 \end{aligned}$$

then by the Cauchy-formula we conclude that

$$y(t) = e^{2t}\left(3 + \int_0^t 5e^{-2s}\,ds\right) = e^{2t}\left(3 - \frac{5}{2}\left[e^{-2s}\right]_0^t\right) = \frac{11}{2}e^{2t} - \frac{5}{2}$$

for each $t \in \mathbb{R}$.

Verify that this is the correct solution, by direct substitution!

Study at home:

1. Careful review of "Mathematical Analysis Exercises"
2. Textbook-1, Sections 11.1 and 11.2.

Chapter 9

Extension of integration

9.1 Improper integrals

Assume that f is continuous on an infinite interval $[a, +\infty)$. Then for every $b \geq a$ the integral $\int_a^b f(x)\,dx$ exists.

Definition 9.1 We say that the *improper integral* of f exists (or convergent) on the infinite interval $[a, \infty)$, if the limit $\lim_{b\to\infty} \int_a^b f(x)\,dx$ exists and it is finite. The value of the improper integral is defined by

$$\int_a^\infty f(x)\,dx = \lim_{b\to\infty} \int_a^b f(x)\,dx$$

If the limit above is not finite, or does not exist, then we say that the improper integral does not exist (or not convergent).

We define the improper integral

$$\int_{-\infty}^a f(x)\,dx$$

in a completely analogous way.

Example 9.2 Investigate the improper integral

$$\int_1^\infty \frac{1}{x}\,dx$$

By the definition

$$\int_1^b \frac{1}{x}\,dx = [\ln x]_1^b = \ln b$$

Passing to the limit $b \to \infty$ we see that the limit of $\ln b$ is not finite, therefore this improper integral is not convergent.

However, the improper integral

$$\int_1^\infty \frac{1}{x^2}\,dx$$

does exist, since

$$\lim_{b\to\infty} \int_1^b \frac{1}{x^2}\,dx = \lim_{b\to\infty} \left[-\frac{1}{x}\right]_1^b = 1$$

and the value of the improper integral is 1.

By applying the same argument, we see that the improper integral

$$\int_1^\infty \frac{1}{x^\alpha}\,dx$$

is convergent if and only if $\alpha > 1$, and its value is

$$\int_1^\infty \frac{1}{x^\alpha}\,dx = \frac{1}{\alpha - 1} \tag{9.1}$$

since the limit at the upper bound is zero.

Example 9.3 Consider the following important example (density function of the exponential distribution):

$$\int_0^\infty \lambda e^{-\lambda x}\,dx$$

where $\lambda > 0$ is a given constant. Then for any $b > 0$ we have:

$$\int_0^b \lambda e^{-\lambda x}\,dx = \left[-e^{-\lambda x}\right]_0^b = 1 - e^{-\lambda b}$$

Consequently

$$\int_0^\infty \lambda e^{-\lambda x}\,dx = \lim_{b\to\infty} (1 - e^{-\lambda b}) = 1$$

for any given constant $\lambda > 0$.

9.2 Improper integrals on the real line

Definition 9.4 We say that improper integral of f on the real line exists, if the integrals

$$\int_{-\infty}^0 f(x)\,dx \quad \text{and} \quad \int_0^\infty f(x)\,dx$$

are convergent. Then the value of $\int_{-\infty}^{\infty} f(x)\, dx$ is given by the sum of the two integrals.

Example 9.5 For instance, the improper integral

$$\int_{-\infty}^{\infty} \frac{2x}{1+x^2}\, dx$$

does not exist, although for any given $b > 0$ we get

$$\int_{-b}^{b} \frac{2x}{1+x^2}\, dx = 0$$

because the integrand is an odd function. However,

$$\int_{0}^{b} \frac{2x}{1+x^2}\, dx = \ln(1+b^2)$$

and its limit is $+\infty$, when $b \to \infty$ and according to the definition the integral is not convergent. The same can be said about the integral on $(-\infty, 0]$.

Example 9.6 Evaluate the following improper integral:

$$I = \int_{0}^{\infty} x e^{-cx^2}\, dx$$

where $c > 0$ is a given constant. Here for every $b > 0$ we obtain

$$\int_{0}^{b} x e^{-cx^2}\, dx = \left[-\frac{1}{2c} e^{-cx^2} \right]_0^b$$

This implies that $I = 1/2c$. On the other hand, the integrand is an odd function, thus,

$$\int_{-\infty}^{\infty} x e^{-cx^2}\, dx = 0\,.$$

Note that it was important to verify that the integral is convergent!

Example 9.7 **(Gauss-integral)** The following integral is important in probability theory:

$$I = \int_{-\infty}^{\infty} e^{-x^2}\, dx$$

(density function of the normal distribution). The evaluation of this improper integral needs some sophisticated calculations, we skip the details here. The reason why this problem is hard is that the primitive function cannot be given explicitly.

ATTENTION! That does not mean there is no primitive function! The integrand is continuous, which implies that the primitive function exists (see the Chapter 7). The main difficulty is that this primitive function cannot be expressed in terms of elementary functions.

It can be shown that

$$\int_0^{\infty} e^{-x^2}\, dx = \frac{\sqrt{\pi}}{2}$$

and therefore $I = \sqrt{\pi}$, since the integrand is an even function.

By applying the substitution $x = t\sqrt{2}$, we also see that

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{x^2}{2}}\, dx = 1 \tag{9.2}$$

This equality will play an important role in probability theory.

9.3 Integration by parts in improper integrals

In the upcoming examples we use integration by parts in improper integrals. For simplicity, instead of passing to the limit $b \to +\infty$, we briefly indicate the upper bound $+\infty$. (But we should know what it means!)

Example 9.8 Suppose that λ is a positive constant, and evaluate the improper integral

$$\int_0^{\infty} \lambda x e^{-\lambda x}\, dx$$

By setting $f'(x) = \lambda e^{-\lambda x}$ and $g(x) = x$ (this way we make sure that the multiplier x will disappear in the second integral), we get

$$\begin{aligned}
\int_0^{\infty} \lambda x e^{-\lambda x}\, dx &= \left[-x e^{-\lambda x}\right]_0^{\infty} - \int_0^{\infty} -e^{-\lambda x}\, dx \\
&= -\left[\frac{e^{-\lambda x}}{\lambda}\right]_0^{\infty} = \frac{1}{\lambda}\,.
\end{aligned}$$

Observe that the expression within the brackets is zero! It is a consequence of L'Hôpital's Rule.

Example 9.9 Suppose again that λ is a positive constant, and now evaluate the improper integral

$$\int_0^{\infty} \lambda x^2 e^{-\lambda x}\, dx$$

Applying again the setting $f'(x) = \lambda e^{-\lambda x}$ and $g(x) = x^2$ (this way we make sure that the degree of the multiplier x^2 decreases), by two consecutive integrations by parts (with the same setting) we obtain

$$\begin{aligned}
\int_0^\infty \lambda x^2 e^{-\lambda x}\,dx &= \left[-x^2 e^{-\lambda x}\right]_0^\infty - \int_0^\infty -2xe^{-\lambda x}\,dx \\
&= \left[\frac{-2xe^{-\lambda x}}{\lambda}\right]_0^\infty - \int_0^\infty -2\frac{e^{-\lambda x}}{\lambda}\,dx \\
&= \left[-2\frac{e^{-\lambda x}}{\lambda^2}\right]_0^\infty = \frac{2}{\lambda^2}
\end{aligned}$$

In this example we needed two integrations by parts in a row to eliminate the multiplier x^2. In view of the L'Hôpital-Rule, the expressions inside the brackets are zero.

Example 9.10 Use integration by parts to evaluate the improper integral

$$\int_{-\infty}^{\infty} x^2 e^{-x^2/2}\,dx$$

By allocating the roles among the factors in a smart way, we conclude:

$$\int_{-\infty}^{\infty} (-x)\cdot\left(-xe^{-x^2/2}\right)\,dx = \left[-xe^{-x^2/2}\right]_{-\infty}^{\infty} + \int_{-\infty}^{\infty} e^{-x^2/2}\,dx = \sqrt{2\pi}$$

where we relied on formula (9.2). Indeed, making use of L'Hôpital's Rule, we see that both limits of the expression within the brackets are zero, hence

$$\int_{-\infty}^{\infty} x^2 e^{-x^2/2}\,dx = \sqrt{2\pi}\,. \tag{9.3}$$

9.4 Harmonic series revisited

As we have seen in Chapter 2, for a given exponent $\alpha > 0$ the infinite series

$$\sum_{k=1}^{\infty} \frac{1}{k^\alpha} \tag{9.4}$$

is divergent if $\alpha \leq 1$, and it is convergent if $\alpha \geq 2$. However, we were unable to find the answer when $1 < \alpha < 2$. Now we give a complete solution by using improper integrals. Consider n-th partial sum of the series

$$S_n = \sum_{k=1}^{n} \frac{1}{k^\alpha}$$

and sketch the graph of the function

$$f(x) = \frac{1}{x^\alpha}$$

on the positive part of the real line. Take the values of the functions at the integers $1, \ldots, n$, then by examining the graph we can easily see that

$$S_n < 1 + \int_1^n \frac{1}{x^\alpha}\, dx$$

since the function f is strictly monotone decreasing.

ATTENTION! Check Figures.pdf for the details!

On the other hand f is positive, and for $\alpha > 1$ its improper integral on the interval $[1, \infty)$ is convergent, see the equality (9.1). Therefore

$$S_n < 1 + \int_1^n \frac{1}{x^\alpha}\, dx < 1 + \int_1^\infty \frac{1}{x^\alpha}\, dx = 1 + \frac{1}{\alpha - 1} = \frac{\alpha}{\alpha - 1}.$$

We conclude that S_n is bounded from above, and it is clearly strictly monotone increasing, hence it is convergent. We summarize this result in the following theorem.

Theorem 9.11 *The infinite series (9.4) is convergent if and only if $\alpha > 1$, and in this case*

$$\sum_{k=1}^{\infty} \frac{1}{k^\alpha} < \frac{\alpha}{\alpha - 1}$$

Study at home

1. Careful review of "Mathematical Analysis Exercises"
2. Textbook-1, Sections 11.3 and 11.4.

Chapter 10

Power series

10.1 Sum of power series

If $-1 < x < 1$ is a given real number, then the geometric series

$$\frac{1}{1-x} = \sum_{k=0}^{\infty} x^k .$$

is convergent. It is an interesting question if a given function f can be given in the form

$$f(x) = \sum_{k=0}^{\infty} a_k x^k \tag{10.1}$$

with appropriate coefficients a_k. In this case we say that f can be expanded in a power series.

Definition 10.1 The series on the right-hand side of the equality (10.1) is called a *power series*, the function f on the left-hand side is called the *sum* of the power series.

In this chapter we examine two interesting questions.

1. For what values x is the power series convergent, and what is its sum f.

2. Conversely, if a function f is given, how can we find the power series whose sum is precisely f (if possible).

A power series is obviously convergent for $x = 0$ and its sum is a_0. The set of all values of x for which the power series is convergent is called the *set of convergence*.

10.2 Radius of convergence

The set of convergence of a power series is always an interval that is symmetric about the origin. This fact is formulated in the following theorem.

Theorem 10.2 (Cauchy-Hadamard-theorem) *For the power series (10.1) there exists a nonnegative number R (maybe $R = 0$ or infinity) so that the series is convergent in the open interval $-R < x < R$, and it is divergent outside the closed interval $[-R, R]$.*

Proof. We just restrict our attention to the case when the limit

$$\lim_{k\to\infty} \left| \frac{a_{k+1}}{a_k} \right| = r$$

exists. Introduce the notation:

$$R = \begin{cases} 1/r & \text{ha } 0 < r < +\infty \\ +\infty & \text{ha } r = 0 \\ 0 & \text{ha } r = \infty \end{cases}$$

In view of the Quotient Test the seies is convergent, if

$$\lim_{k\to\infty} \left| \frac{a_{k+1}x^{k+1}}{a_k x^k} \right| < 1$$

which exactly means that $|x| < R$.

A completely analogous argument shows that the series is divergent when $|x| > R$. $\square$

ATTENTION!

This theorem says nothing about the boundary of the interval! At $|x| = R$ the series may or may not be convergent. This cannot be decided by our theorem, further analysis is needed.

Definition 10.3 The number R above is called the *radius of convergence* of the power series.

Example 10.4 Consider the power series

$$\sum_{k=0}^{\infty} \frac{x^k}{k!}$$

Here we have

$$\lim_{k\to\infty} \frac{a_{k+1}}{a_k} = \lim_{k\to\infty} \frac{k!}{(k+1)!} = \lim_{k\to\infty} \frac{1}{k+1} = 0$$

and hence $R = \infty$. This means that the power series is convergent on the whole real line.

Another example is the power series

$$\sum_{k=1}^{\infty} \frac{x^k}{k}$$

Then we get

$$\lim_{k \to \infty} \frac{a_{k+1}}{a_k} = \lim_{k \to \infty} \frac{k}{k+1} = 1$$

and hence $R = 1$. We conclude that the series is convergent in the open interval $(-1, 1)$, and it is divergent outside the closed interval $[-1, 1]$.

On the other hand, we see that for $x = 1$ we obtain the divergent harmonic series, and further for $x = -1$ we get a convergent series with alternating signs, see Example 2.12. Thus, the interval of convergence of this power series is the interval

$$[-1, 1)$$

closed from the left and open from the right. Please observe that on the boundary anything can happen!

10.3 Differentiability of power series

Consider a power series whose radius of convergence is $R > 0$ and its sum function is f that is

$$\sum_{k=0}^{\infty} a_k x^k = f(x)$$

for every $-R < x < R$.

Theorem 10.5 *The sum f of the power series is differentiable, in particular*

$$f'(x) = \sum_{k=1}^{\infty} k a_k x^{k-1}$$

in the open interval $(-R, R)$.

We do not prove this theorem (it is technical), just note that it is based on the so-called "uniform convergence" principle. Some consequences however, can easily be derived from this statement.

- The derivative of the sum is obtained from differentiating the power series term by term. This is not obvious, since the sum rule (in general) is not true for infinitely many terms. FIND COUNTEREXAMPLES!

- Observe that the radius of convergence of the derivative power series is still R. VERIFY!
- As we see that f' is the sum of a power series in the same interval, by repeated applications of the theorem, we deduce that f is infinitely many times differentiable in the open interval $(-R, R)$.

Example 10.6 Consider the geometric series in the open interval $-1 < x < 1$

$$\sum_{k=0}^{\infty} x^k = \frac{1}{1-x}$$

Note that the first term is 1, whose derivative is zero. Making use of our theorem

$$\sum_{k=1}^{\infty} kx^{k-1} = \frac{1}{(1-x)^2}$$

for every $-1 < x < 1$.

Example 10.7 Find the function f that is given by the following power series:

$$f(x) = \sum_{k=1}^{\infty} (-1)^{k-1} \frac{x^k}{k}$$

A simple calculation shows that the radius of convergence is $R = 1$. On the one hand $f(0) = 0$, on the other hand, by the differentiability of the power series

$$f'(x) = \sum_{k=1}^{\infty} k \frac{(-x)^{k-1}}{k} = \sum_{k=1}^{\infty} (-x)^{k-1} = \frac{1}{1+x}$$

for each $-1 < x < 1$. This implies

$$f(x) = f(0) + \int_0^x \frac{1}{1+t}\, dt = [\ln(1+t)]_0^x = \ln(1+x)$$

in the open interval $(-1, 1)$. Moreover, by Example 2.12 the original series is convergent at $x = 1$, which leads to the celebrated identity

$$1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \frac{1}{5} - \ldots = \ln 2$$

However, the series is divergent at $x = -1$.

10.4 Finding the coefficients

Suppose that a function f can be given as the sum of a power series in the interval of convergence. Then necessarily f is infinitely many times differentiable in the interval. How could we determine the coefficients of the power series?

By succesively taking the derivatives of both sides of equality (10.1), the coefficients a_k can be computed step by step. Indeed, observe that

$$f(0) = a_0, \quad f'(0) = a_1, \quad f''(0) = 2a_2, \quad \ldots$$

and in general, for any given index k we get:

$$f^{(k)}(0) = k! \cdot a_k$$

If we substitute these expressions for a_k in the power series, then we have

$$f(x) = \sum_{k=0}^{\infty} \frac{f^{(k)}(0)}{k!} x^k$$

This form is called the *Taylor-series* (or Taylor expansion) of f.

10.5 Taylor-series of the exponential function

In this section we consider the exponential function $f(x) = e^x$. If this function is the sum of a power series, then the coefficients can only be

$$a_k = \frac{1}{k!}$$

for every k. Indeed, any derivative of e^x is e^x, which takes the value 1 at $x = 0$. Therefore, the Taylor-series associated with the function e^x is:

$$\sum_{k=0}^{\infty} \frac{x^k}{k!}$$

and we have seen that this series is convergent on the entire real line.

The reason why we did not write equality is that it is not yet clear at the moment that the sum of this series is really e^x.

To overcome this difficulty, consider the function

$$f(x) = \sum_{k=0}^{\infty} \frac{x^k}{k!}$$

on the real line, which is yet to be determined. Clearly $f(0) = 1$. On the other hand, in view of the differentiability theorem:

$$f'(x) = \sum_{k=1}^{\infty} k \frac{x^{k-1}}{k!} = \sum_{k=1}^{\infty} \frac{x^{k-1}}{(k-1)!} = f(x)$$

for every $-\infty < x < \infty$. This is a simple linear differential equation for the unknown f, whose only solution is

$$f(x) = e^x$$

on the real line. As a consequence, we deduce the celebrated identity

$$e = 1 + \frac{1}{1!} + \frac{1}{2!} + \ldots + \frac{1}{n!} + \ldots$$

by substituting $x = 1$.

Study at home

1. Careful review of "Mathematical Analysis Exercises"
2. Textbook-1, Section 6.5.

Chapter 11

Functions of two variables

11.1 Partial derivatives

Consider a function $f : \mathbb{R}^2 \to \mathbb{R}$ of two variables. Fix the coordinate $y = b$ and examine the function

$$x \to f(x, b)$$

of only one variable. Assume that this function is differentiable at a point a, and determine its derivative.

Definition 11.1 The derivative above is called the *partial derivative* of the function f with respect to the variable x at the point (a, b). We denote it by

$$\frac{\partial f}{\partial x}(a, b) = f_1'(a, b)$$

Sometimes the notation $f_x'(a, b)$ is also used.

Example 11.2 Consider for instance the function $f(x, y) = (x + 2y)e^{x+3y-1}$ and find its partial derivative with respect to x at the point $(1, 1)$.

Then $f(x, 1) = (x + 2)e^{x+2}$, whose derivative at any x is

$$f_1'(x, 1) = e^{x+2} + (x + 2)e^{x+2} = (x + 3)e^{x+2}$$

Substituting $x = 1$ we obtain $f_1'(1, 1) = 4e^3$.

Example 11.3 Principally, we could also calculate the partial derivative of the function f with respect to the variable x with an arbitrarily selected and fixed y, and substitute the values $x = a$ and $y = b$. This is of course good, but not always convenient, as shown in the following example. Take

$$f(x, y) = \sqrt{x^2 + y^2 + 5} \cdot e^{-2x+y} \cdot \cos(y + \pi/2)$$

and find the partial derivative with respect to x at the point $(1, 0)$. Then the above way would give you the right answer, but it requires a long calculation and very time consuming. However, if we follow the definition, then we see that

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

for every x, and therefore $f_1'(1, 0) = 0$.

The correspondence

$$x \to \frac{\partial f}{\partial x}(x) , \quad x \in \mathbb{R}$$

is called the *partial derivative function* of f with respect to the variable x.

11.2 Tangent planes

Partial derivatives (similarly to the one variable case) can be given a nice geometric interpretation. Consider a function $f : \mathbb{R}^2 \to \mathbb{R}$ with two variables. The graph of this function is a surface in the three dimensional space. Pick a point

$$P(a, b, f(a, b))$$

on the surface. If this surface is intersected by the plane $y = b$ passing through the point P, then we get a curve lying on the surface. The slope of the tangent line to this curve at P is exactly the partial derivative $f_1'(a, b)$. We can give an analogous interpretation for the slope of the tangent line that lies in the plane $x = a$. The plane spanned by the two tangent lines has the following normal vector (perpendicular):

$$v = (f_1'(a, b), f_2'(a, b), -1)$$

By using the notation $c = f(a, b)$ the equation of this plane is

$$f_1'(a, b)(x - a) + f_2'(a, b)(y - b) - (z - c) = 0 .$$

This plane is called the *tangent plane* to the surface at the point P.

Example 11.4 Find the value of the parameter p if the tangent plane to the function

$$f(x, y) = px\sqrt{x^2 + y^2 + 1} - 7$$

at the point $a = 2$, $b = 2$, $c = f(2, 2)$ passes through the point $Q(2, -1, 6)$.

Simple substitution shows that $f(2,2) = 6p - 7$, which means that we are looking for the equation of the tangent plane at the point $P(2, 2, 6p - 7)$. Calculate the partial derivatives:

$$\frac{\partial f}{\partial x}(2,2) = \frac{13}{3}p \qquad \text{and} \qquad \frac{\partial f}{\partial y}(2,2) = \frac{4}{3}p$$

Hence, the equation of the tangent plane at P is:

$$\frac{13}{3}p(x-2) + \frac{4}{3}p(y-2) - (z - 6p + 7) = 0\,.$$

If the tangent plane passes through the point Q, then its coordinates satisfy the equation of the plane. This gives us the following equation for the unknown parameter p:

$$-4p = 13 - 6p\,.$$

The only solution is $p = 13/2$.

11.3 Chain Rule

Consider now the functions $f : \mathbb{R}^2 \to \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}^2$ where for every $t \in \mathbb{R}$ we use the notation

$$g(t) = (g_1(t), g_2(t))$$

Suppose that the range of g lies in the domain of f. Then we may examine the composition

$$f \circ g : \mathbb{R} \to \mathbb{R}$$

We want to give a condition on the differentiability of $f \circ g$.

Theorem 11.5 **(Chain Rule)** *If both g_1 and g_2 are differentiable at t, and the partial derivative functions of f are continuous at $g(t)$, then $f \circ g$ is differentiable at t, and*

$$(f \circ g)'(t) = \frac{\partial f}{\partial x}(g(t))g_1'(t) + \frac{\partial f}{\partial y}(g(t))g_2'(t)$$

Our theorem is very similar to the Chain Rule with one variable (see Chapter 4). Its proof (skipped) would follow the same ideas, but technically a bit more involved.

Example 11.6 Take for instance $f(x, y) = x^2 - xy + y^2$, and

$$x = g_1(t) = \cos t \qquad y = g_2(t) = \sin t$$

and consider the composition function $F(t) = (f \circ g)(t)$. Making use of the Chain Rule

$$\begin{aligned} F'(t) &= (f \circ g)'(t) = \frac{\partial f}{\partial x}(g(t))g_1'(t) + \frac{\partial f}{\partial y}(g(t))g_2'(t) \\ &= (2\cos t - \sin t)(-\sin t) + (-\cos t + 2\sin t)\cos t = \sin^2 t - \cos^2 t \end{aligned}$$

for every $t \in \mathbb{R}$.

Example 11.7 Consider the function $f : \mathbb{R}^2 \to \mathbb{R}$ and suppose that its partial derivatives exist and are continuous. Take the vector $v = (v_1, v_2) \in \mathbb{R}^2$ in the plane, and let a point $P(a, b) \in \mathbb{R}^2$ be given. Then the equation of the straight line in the direction v and passing through the point $P(a, b)$ is:

$$g(t) = (a, b) + tv = (a + tv_1, b + tv_2)\,.$$

Using these notations we have $g_1'(t) = v_1$, $g_2'(t) = v_2$. Further, take the composition function

$$F(t) = f((a, b) + tv)$$

then by the Chain Rule, its derivative is given by:

$$F'(t) = \frac{\partial f}{\partial x}((a, b) + tv)v_1 + \frac{\partial f}{\partial y}((a, b) + tv)v_2$$

In particular for $t = 0$ we obtain:

$$F'(0) = \frac{\partial f}{\partial x}(a, b)v_1 + \frac{\partial f}{\partial y}(a, b)v_2$$

11.4 Local extrema

The absolute value (or the distance from the origin) of a vector $v = (x, y)$ in the two dimensional plane is defined by:

$$\|v\| = (x^2 + y^2)^{1/2}$$

that is called the *norm* of the vector v.

Definition 11.8 In the plane $\mathbb{R}^2$ the set

$$B = \{v \in \mathbb{R}^2 : \|v\| \le 1\}$$

is called the *unit disk* (with center at the origin and radius equals 1). Consequently, a disk with center at the point $(a,b) \in \mathbb{R}^2$ and radius $r > 0$ is given by

$$(a,b) + rB = \{v \in \mathbb{R}^2 : \|v - (a,b)\| \leq r\}$$

(i.e. the set of points, whose distance from the center is at most r).

Consider a function $f : \mathbb{R}^2 \to \mathbb{R}$. We say that a point $P(a,b)$ in the domain is a local minimum point of f, if there exists a $\varepsilon > 0$ such that

$$f(x,y) \geq f(a,b)$$

for all points (x,y) in the domain of f, where $(x,y) \in (a,b) + \varepsilon B$, that is $\|(x,y) - (a,b)\| \leq \varepsilon$.

The local maximum is defined analogously. For global minimum or maximum the inequality must hold on the entire domain of f.

11.5 First order necessary condition

In this section we suppose that partial derivatives of the function $f : \mathbb{R}^2 \to \mathbb{R}$ exist and are continuous.

Theorem 11.9 *If the point $(a,b) \in \mathbb{R}^2$ is a local minimum point of f, then $f_1'(a,b) = 0$ and $f_2'(a,b) = 0$.*

Proof. Take a non zero vector $v \in \mathbb{R}^n$ arbitrarily, and consider the composition function

$$F(t) = f((a,b) + tv)\,.$$

In vew of our assumption the function F has a local minimum at $t = 0$. On the other hand, F is differentiable, namely

$$F'(t) = \frac{\partial f}{\partial x}((a,b) + tv)v_1 + \frac{\partial f}{\partial y}((a,b) + tv)v_2$$

Applying Theorem 5.7 we get $F'(0) = 0$ for every vector v, in other words

$$\frac{\partial f}{\partial x}((a,b))v_1 + \frac{\partial f}{\partial y}((a,b))v_2 = 0$$

for all real numbers v_1 and v_2. This is only possible if

$$\frac{\partial f}{\partial x}((a,b)) = 0 \qquad \text{and} \qquad \frac{\partial f}{\partial y}((a,b)) = 0$$

and this is exactly that we wanted to prove. □

Analogous theorem applies for the case of local maximum.

This theorem tells us that local extrema can only be at points where both partial derivatives are zero. In other words, local extrema can be only be found in the solution set of the system of equations with both partial derivatives being zero. This is however, just a necessary condition (just like in the one-variable case), and by no means sufficient! For example, in the case of the function

$$f(x,y)=x^3y^2$$

we have the necessary condition $f_1'(x,y) = f_2'(x,y) = 0$. A solution to this system is $(x,y)=(0,0)$, and at this point

$$f(0,0)=0$$

But this is neither a minimum nor a maximum. It is easy to see that the function has both positive and negative values in any disk around the origin (with whatever positive radius). Thus, the origin cannot be a local extreme point.

Example 11.10 Consider the function

$$f(x,y)=\frac{1}{x}+\frac{1}{y}+\frac{xy}{8}$$

on the plane, where $x \neq 0$ and $y \neq 0$, and try to find its local extreme points. Find the zeros of the partial derivatives!

$$\begin{aligned}\frac{\partial f}{\partial x} &= -\frac{1}{x^2}+\frac{y}{8}=0\\ \frac{\partial f}{\partial y} &= -\frac{1}{y^2}+\frac{x}{8}=0\end{aligned}$$

The only solution to the simultaneous equations is

$$x=2 \qquad \text{and} \qquad y=2\,,$$

therefore f can only have a local extremum (minimum or maximum) at this point.

A comprehensive method for deciding whether or not a critical point is a local extremum will be discussed in the Linear Algebra course (third semester, sophomore year). We note here that $P(2,2)$ is in fact a local minimum point of f (see the "Mathematical Analysis Exercises" for more details).

Study at home

1. Careful review of "Mathematical Analysis Exercises"
2. Textbook-1, Sections 15.3, 15.4, 15.6, 16.1 and 16.2.

Chapter 12

Constrained extrema

12.1 Implicit functions

A problem often encountered in microeconomics is the following. If an equation

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

is given, can we uniquely express the variable y from the equation as a function of x? In other words: can we find a unique function $y = g(x)$ such that the identity

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

holds at every point x?

Such a function does not necessarily exist. For example, in the case of the equation

$$F(x, y) = x^2 + y^2 - 1 = 0$$

(equation of the unit circle) the variable y cannot be expressed uniquely as a function of x. Geometrically this means that the set of points on the plane that satisfy the equation $F(x, y) = 0$ cannot be the graph of a function. The reason for this is that some vertical lines (parallel to the y-axis) intersect this curve twice.

It may even happen that the variable y cannot be expressed from the equation by algebraic manipulations. Such an example is the equation

$$F(x, y) = e^{x+y} - 2\cos y + 1 = 0$$

It is easy to see that the point $(x, y) = (0, 0)$ satisfies the equation, but the variable y cannot be isolated on one side.

We also raise the following question. If F is differentiable, then can we express the variable y from the equation as a differentiable function of x? This question is answered by the following theorem.

Tétel 12.1 **(Implicit function theorem)** *Assume that the at the point* (x_0, y_0) *we have*

$$F(x_0, y_0) = 0$$

moreover the partial derivatives of F *are continuous in a neighborhood of this point, and*

$$F_2'(x_0, y_0) \neq 0$$

Then there exists a unique continuously differentiable function g *in a neighborhood of the point* x_0 *such that*

- $g(x_0) = y_0$
- $F(x, g(x)) = 0$ *at every point* x
- $g'(x) = -F_1'(x, g(x))/F_2'(x, g(x))$

We point out that from the continuity of the partial derivatives we get that $F_2'(x, g(x)) \neq 0$ in a neighborhood of the point x_0.

The geometric interpretation of our theorem is that if the tangent line to the planar curve with equation $F(x, y) = 0$ at the point (x_0, y_0) is not parallel to the y-axis (i.e. "the curve cannot turn back"), then y can be expressed (locally) as a differentiable function of x.

Example 12.2 Consider the implicit equation

$$F(x, y) = e^{x+y} + x + y - 1 = 0$$

The point $(0, 0)$ satisfies the equation. On the other hand, at this point

$$F_2'(0, 0) = 2$$

Hence, F fulfills the conditions of the Implicit function theorem: there exist a unique differentiable function $y = g(x)$ with

$$\begin{aligned} g'(x) &= -F_1'(x, g(x))/F_2'(x, g(x)) \\ &= -\frac{1}{e^{x+g(x)} + 1} \cdot (e^{x+g(x)} + 1) = -1 \end{aligned}$$

at every point x. Since $g(0) = 0$, this implies

$$g(x) = -x$$

and this is the only solution.

Példa 12.3 A slightly more complicated example is

$$F(x, y) = e^{x+y} - 2\cos y + 1 = 0$$

The point $(0, 0)$ satisfies the equation. On the other hand, at this point

$$F_2'(0, 0) = 1$$

and hence, the conditions of the Implicit function theorem are fulfilled. We conclude that the equation uniquely determines a differentiable function g so that $F(x, g(x)) = 0$ at every x. However, this function cannot be expressed explicitly by using algebraic manipulations.

12.2 Constrained minima

Consider the functions f and F that are both $\mathbb{R}^2 \to \mathbb{R}$ and suppose that their partial derivatives are continuous. By a *constrained minimum* problem we mean the following problem:

$$\begin{aligned} f(x, y) &\to \min \\ F(x, y) &= c \end{aligned} \tag{12.1}$$

where c is a given real constant. In other words, we look for the minimum (or sometimes maximum) of f on the set keressük a

$$H = \{(x, y) \in \mathbb{R}^2 : F(x, y) = c\}$$

This equality is called the *constraint*.

Definition 12.4 We say that the point $(x_0, y_0) \in H$ is the solution of the constrained minimization problem (12.1) if

$$f(x_0, y_0) \leq f(x, y)$$

for every $(x, y) \in H$ esetén. An analogous definition applies for maximum problems.

Example 12.5 The example below illustrates that for constrained minimization problems the usual necessary conditions for extrema do not work. Consider the constrained minimization problem

$$f(x, y) = x^2 + 2y, \quad F(x, y) = x + y = 0 \quad \text{i.e.} \quad c = 0$$

From the constraint $x+y = 0$ we get $y = -x$, and consequently $f(x, y) = x^2-2x$ on the set H. This function achieves its minimum at the point $x = 1$ and in H this necessarily means $y = -1$. Thus, the constrained minimum is at the point

$$(x_0, y_0) = (1, -1)$$

However, et this point none of the equalities

$$\frac{\partial f}{\partial x} = 0, \quad \frac{\partial f}{\partial y} = 0$$

is true. Verify this!

This example also exhibits that a constrained extremum problem can be transformed into a non-constrained extremum problem by expressing the variable y as a function of x from the constraint $F(x, y) = c$. In more complicated problems this may not be possible by algebraic manipulations. This is the point where we need the Implicit function theorem.

12.3 Lagrange multipliers

Consider the constrained minimization problem (12.1). By using the Implicit function theorem we make sure that the variable y can be expressed from the constraint $F(x, y) = c$, and that way we can solve the problem. This procedure is described below.

Definition 12.6 The *Lagrange-function* (or Lagrangian) of the problem (12.1) is defined by

$$\mathcal{L}(x, y, \lambda) = f(x, y) - \lambda(F(x, y) - c)$$

λ is an arbitrary real number.

Theorem 12.7 (Lagrange-method) *Let us suppose that (x_0, y_0) is the solution of the problem (12.1), and assume that the partial derivatives of f and F are continuous in a neighborhood of this point. If*

$$F_2'(x_0, y_0) \neq 0\,, \tag{12.2}$$

there exists a unique real number λ such that

$$\frac{\partial \mathcal{L}}{\partial x}(x_0, y_0, \lambda) = 0\,, \quad \textit{and} \quad \frac{\partial \mathcal{L}}{\partial y}(x_0, y_0, \lambda) = 0$$

Proof. In view of (12.2) the conditions of the Implicit function theorem are fulfilled. Thus, there exists a unique continuously differentiable function g with

- $g(x_0) = y_0$, and
- $F(x, g(x)) = c$ in a neighborhood of x_0, furthermore

- $g'(x_0) = -F'_1(x_0, y_0)/F'_2(x_0, y_0)$.

If (x_0, y_0) is the solution of problem (12.1), then the function $x \to f(x, g(x))$ achieves its minimum at x_0, therefore, its derivative at this point is zero. Applying the Cahin Rule, the derivative can be given in this form: szerint

$$f'_1(x_0, y_0) + f'_2(x_0, y_0)g'(x_0) = f'_1(x_0, y_0) - \frac{f'_2(x_0, y_0)}{F'_2(x_0, y_0)}F'_1(x_0, y_0) = 0\,.$$

Introduce the notation:

$$\lambda = \frac{f'_2(x_0, y_0)}{F'_2(x_0, y_0)}\,.$$

Using this notation, the above derivative can be rewritten:

$$\frac{\partial \mathcal{L}}{\partial x}(x_0, y_0, \lambda) = f'_1(x_0, y_0) - \lambda F'_1(x_0, y_0) = 0\,.$$

The second equality of the theorem is trivial by simply substituting λ. Indeed:

$$\frac{\partial \mathcal{L}}{\partial y}(x_0, y_0, \lambda) = f'_2(x_0, y_0) - \lambda F'_2(x_0, y_0) = 0\,.\square$$

Our theorem could be formulated analogously for the case of maximum.

12.4 Solving the constrained minimization problem

The procedure of solving the constrained minimization problem (12.1) is as follows.

1. Find the Lagrange-function of the problem.
2. Find the partial derivatives with respect to x and y, and make them equal zero.
3. Take into account that $F(x_0, y_0) = c$.
4. Solve the system of three equations for x, y and λ.

The point (x_0, y_0) obtained that way satisfies the necessary condition for an extremum. The solution λ is called the *Lagrange multiplier* associated with the problem.

Example 12.8 Now solve the constrained minimization problem in Example 12.5 by using the Lagrange-method. The Lagrange-function of the problem is:

$$\mathcal{L}(x, y, \lambda) = x^2 + 2y - \lambda(x + y)\,.$$

The system of equations is of the form:

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial x}(x_0, y_0, \lambda) &= 2x_0 - \lambda = 0 \\
\frac{\partial \mathcal{L}}{\partial y}(x_0, y_0, \lambda) &= 2 - \lambda = 0 \\
\frac{\partial \mathcal{L}}{\partial \lambda}(x_0, y_0, \lambda) &= x_0 + y_0 = 0
\end{aligned}$$

The only solution to this system is $\lambda = 2$, $x_0 = 1$ and $y_0 = -1$.

Példa 12.9 The following type of problem frequently appears in microeconomics. Find the constrained maximum of consumer demand:

$$\begin{aligned}
x^\alpha y^\beta &\to \max \\
px + y &= m
\end{aligned} \tag{12.3}$$

where α, β, p and m are given positive real numbers. In this problem

$$f(x, y) = x^\alpha y^\beta \quad \text{and} \quad F(x, y) = px + y\,,$$

Therefore, the Lagrange-function of the problem is:

$$\mathcal{L}(x, y, \lambda) = x^\alpha y^\beta - \lambda(px + y - m)\,.$$

The system of equation that comes from the Lagrange-method:

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial x}(x_0, y_0, \lambda) &= \alpha \cdot x_0^{\alpha-1} y_0^\beta - \lambda p = 0 \\
\frac{\partial \mathcal{L}}{\partial y}(x_0, y_0, \lambda) &= \beta \cdot x_0^\alpha y_0^{\beta-1} - \lambda = 0 \\
\frac{\partial \mathcal{L}}{\partial \lambda}(x_0, y_0, \lambda) &= px_0 + y_0 - m = 0\,.
\end{aligned}$$

This system admits the following single solution:

$$px_0 = \frac{\alpha}{\alpha + \beta} m \quad \text{and} \quad y_0 = \frac{\alpha}{\alpha + \beta} m\,,$$

The Lagrange multiplier λ can be then calculated from the second equation.

Study at home

1. Careful review of "Mathematical Analysis Exercises"
2. Textbook-1, Sections 16.3, 18.1, 18.2, 18.3, 18.4, 18.5 and 18.6.

Part II

Second Semester: Probability Theory

Chapter 13

Probability

13.1 Experiments

In the sequel we deal with experiments that have chance outcomes. In other words, the experiments have outcomes that cannot be predicted.

1. Toss a playing die and check the number that comes out.
2. Toss a pair of dice.
3. Toss a die, then flip a coin as many times as the number on the die.
4. Keep tossing a die until 6 comes out for the first time.
5. Pick a point randomly on the unit disc (with radius 1).

More complicated examples:

- The number of cars that pass an intersection between 10 am and 11 am.
- The number of calls received by a call center between 8 am and 9 am.
- The length of time period between two successive calls
- The price of a stock at the stock exchange at closing time.
- The waiting time at a customer service desk.

13.2 The sample space

Definition 13.1 Let Ω denote the set of all possible outcomes in an experiment. The set Ω is called the *sample space* associated with the experiment.

Specify the sample spaces that are associated with the previous experiments. Then in the same order:

1. $\Omega = \{1, 2, 3, 4, 5, 6\}$
2. $\Omega = \{(1, 1), (1, 2), (2, 1), (1, 3), \ldots, (6, 6)\}$
3. $\Omega = \{1H, 1T, 2HH, 2HT, 2TH, 2TT, \ldots\}$ (Question: how many elements are in the sample space?)
4. Ω consists of all finite sequences whose last digit is 6, and all previous digits are any of the numbers 1,2,3,4,5.
5. $\Omega = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 \leq 1\}$

13.3 Events

Definition 13.2 The subsets of the sample space are called *events*.

Take some examples in the sample spaces above.

1. Let A denote the event that the outcome is even. Then $A = \{2, 4, 6\}$.
2. Let A denote the event that the sum of the two numbers is 7. Then $A = \{(1, 6), (6, 1), (2, 5), (5, 2), (3, 4), (4, 3)\}$.
3. Let A denote the event that we have no Tail (all of them are Head). Then $A = \{1H, 2HH, 3HHH, 4HHHH, 5HHHHH, 6HHHHHH\}$.
4. Let A denote the event that we needed at most two tosses. Then $A = \{6, 16, 26, 36, 46, 56\}$.
5. Let A denote the event that the distance of the point from the center is less than 1/2. Then $A = \{(x, y) : x^2 + y^2 < 1/4\}$.

13.4 Operations with events

We say that the event $A \subset \Omega$ occurs, if the experiment results in an outcome $\omega \in \Omega$ such that $\omega \in A$.

The impossible event has no elements, notation: $\emptyset$ (empty set). The certain event is: Ω (the whole sample space).

1. $A \cap B$ occurs if and only if both A and B occur. We say that A and B are mutually exclusive, if $A \cap B = \emptyset$.

2. $A \cup B$ occurs if and only if either A or B occurs (or both).
3. $\overline{A}$ (the complement of A) occurs if and only if A does not occur.

We say that A implies B (or B is a consequence of A), if $A \subset B$.

Theorem 13.3 (De Morgan Rules)

1. $\overline{A \cup B} = \overline{A} \cap \overline{B}$
2. $\overline{A \cap B} = \overline{A} \cup \overline{B}$

These identities hold true for an arbitrary number of events as well.

Proof. We demonstrate the first identity. Let $x \in \overline{A \cup B}$ be selected arbitrarily. Then

$$x \in \overline{A \cup B} \Rightarrow x \notin A \cup B \Rightarrow x \notin A \text{ and } x \notin B \Rightarrow x \in \overline{A} \text{ and } x \in \overline{B} \Rightarrow x \in \overline{A} \cap \overline{B}$$

This proves that $\overline{A \cup B} \subset \overline{A} \cap \overline{B}$. The opposite direction (and hence the equality) follows from the fact that each implication can be reversed (i.e. they are equivalences). The second identity can be verified in a completely analogous way. □

When we carry out an experiment, some possible outcomes may not be observable. For instance, if we toss a pair of completely identical (indistinguishable) dice, we cannot decide whether the outcome is $(1,2)$ or $(2,1)$. We can only claim that the event $\{(1,2),(2,1)\}$ occured.

Definition 13.4 Let $\mathcal{A}$ denote the collection of *observable events*. We assume that they possess the following properties.

- If $A \in \mathcal{A}$, then $\overline{A} \in \mathcal{A}$ and $\Omega \in \mathcal{A}$.
- If $A_1, A_2, \ldots \in \mathcal{A}$, then $A_1 \cup A_2 \cup \ldots \in \mathcal{A}$.

Proposition 13.5 *If A and B are observable, then so is $A \cap B$.*

Proof. Indeed, if A and B are observable, then

$$A \cap B = \overline{\overline{A} \cup \overline{B}} \in \mathcal{A}$$

in view of the De Morgan Rules. □

By the De Morgan Rules, this proposition remains true for any countable number of events.

Definition 13.6 In the following, by an *experiment* we mean the couple $\mathcal{K} = (\Omega, \mathcal{A})$.

13.5 Probability space

Suppose that we perform an experiment $\mathcal{K}$ n times in a row, and every time we observe whether or not a given event $A \in \mathcal{A}$ occurs. If A occurs k_n times out of n trials, then the relative frequency of A is:

$$\frac{k_n}{n}$$

Experience shows that by raising n, the relative frequency exhibits a dumping oscillation around a specific number. This number can be regarded as the probability of A.

Instead of using this experimental approach, below we develop an axiomatic introduction of probability. From the axioms we can derive the above experimental fact.

Definition 13.7 **(Axioms of Probability)** Consider an experiment $\mathcal{K} = (\Omega, \mathcal{A})$. By the *probability* we mean a function

$$P : \mathcal{A} \to [0,1]$$

that satisfies the following two axioms:

1. $P(\Omega) = 1$

2. If $A_1, A_2, \ldots \in \mathcal{A}$ are pairwise mutually exclusive events, then

$$P(\bigcup_{k=1}^{\infty} A_k) = \sum_{k=1}^{\infty} P(A_k)$$

In this case the triple $(\Omega, \mathcal{A}, P)$ is called a *probability space*.

This axiomatic approach is due to A. N. Kolmogorov (1933), and this can be regarded as the origin of modern probability theory.

From the axioms we can easily derive the following properties of probability spaces.

Theorem 13.8

1. *For any* $A \in \mathcal{A}$ *we have*

$$P(\overline{A}) = 1 - P(A)$$

and consequently $P(\emptyset) = 0$.

2. *If* $A, B \in \mathcal{A}$ *and* $A \subset B$, *then*

$$P(A) \leq P(B)$$

3. *If* $A, B \in \mathcal{A}$, *then*

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Proof. 1. Since $A \cup \overline{A} = \Omega$, moreover A and $\overline{A}$ are exclusive events, the statement follows immediately from the axioms.

2. If $A \subset B$, then $A \cup (B \cap \overline{A}) = B$, moreover A and $B \cap \overline{A}$ are exclusive events, therefore, by the axioms

$$P(B) = P(A) + P(B \cap \overline{A}) \geq P(A)$$

because $P(B \cap \overline{A}) \geq 0$.

The 3. statement is proven the following way. We divide the event $A \cup B$ into disjoint pieces like this:

$$A \cup B = (A \cap \overline{B}) \cup (\overline{A} \cap B) \cup (A \cap B)\,.$$

Then, using the axioms, we get:

$$\begin{aligned} P(A \cup B) &= P(A \cap \overline{B}) + P(\overline{A} \cap B) + P(A \cap B) \\ &= P(A) - P(A \cap B) + P(B) - P(A \cap B) + P(A \cap B) \end{aligned}$$

and the statement ensues. □

Example 13.9

In a Freshman class the probability that a randomly selected student passed the mathematics exam is 0.72, passed the microeconomics exam is 0.66, and passed both is 0.54. Find the probability that a randomly selected student

(a) passed at least one of those exams,

(b) passed the microeconomics exam, but did not pass the mathematics exam,

(c) passed none of the exams.

Let A denote the event that a randomly selected student passed the mathematics exam, and B is the event that the student passed the microeconomics exam. Then $P(A) = 0.72$, $P(B) = 0.66$ and $P(A \cap B) = 0.54$. Using the events A and B, the desired probabilities can be given the following way.

(a) $P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0.84$

(b) $P(\overline{A} \cap B) = P(B) - P(A \cap B) = 0.12$

(c) $P(\overline{A} \cap \overline{B}) = P(\overline{A \cup B}) = 1 - P(A \cup B) = 0.16$

Recitation and Exercises

1. Reading: Textbook-2, Sections 2.1, 2.2, 2.3, 2.4, 2.5.
2. Homework: Textbook-2, Exercises 2.11, 2.19, 2.32, 2.33, 2.37, 2.38, 2.54, 2.58, 2.59, 2.61, 2.110, 2.112.
3. Review: Highschool Combinatorics and Binomial Theorem (Textbook-2, Section 2.3), and "Probability Exercises"

Chapter 14

Sampling methods

14.1 Classical probability spaces

Definition 14.1 Consider a probability space $(\Omega, \mathcal{A}, P)$. It is called a *classical probability space*, if

- Ω is a finite set,
- for every $\omega \in \Omega$ we have $\{\omega\} \in \mathcal{A}$,
- every singleton subset of Ω has the same probability.

Obviously, if Ω contains exactly n elements, then for every $\omega \in \Omega$ we get

$$P(\{\omega\}) = \frac{1}{n}$$

In particular, if the event $A \subset \Omega$ consists of k elements, then

$$P(A) = \frac{k}{n}$$

This observation can be interpreted as the probability of A can be given like:

$$P(A) = \frac{\text{number of favorable outcomes}}{\text{total number of outcomes}} \tag{14.1}$$

The formula (14.1) will be called the classical formula.

Example 14.2 A regular playing die is tossed twice in a row. What is the probability that the sum of the two numbers is exactly 7?

Let A denote the event that the sum is 7. Clearly, the sample space Ω contains 36 elements (total number of outcomes), while A is a subset of 6 elements containing the pairs $(1,6),(6,1),(2,5),(5,2),(3,4),(4,3)$ (favorable outcomes). Consequently

$$P(A) = \frac{6}{36} = \frac{1}{6}$$

by making use of the classical formula (14.1).

Example 14.3 From a deck of 52 playing cards we draw 5 cards at random. Find the probability that either all 5 cards are clubs, or at least one of them is an Ace?

Introduce the following notations:

$$A = \{\text{all 5 cards are clubs}\} \qquad B = \{\text{at least one of them is Ace}\}$$

Obviously we are looking for $P(A \cup B)$. Since the draws of any 5 cards are equally likely, therefore:

$$P(A) = \frac{\binom{13}{5}}{\binom{52}{5}} \qquad P(B) = 1 - \frac{\binom{48}{5}}{\binom{52}{5}}$$

and further:

$$P(A \cap B) = \frac{\binom{12}{4}}{\binom{52}{5}}$$

By using the additive rule

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)\,.$$

Example 14.4 On a seasonal sale in a supermarket there are 10 different pairs of shoes in a basket. A thief quickly grabs 4 pieces of shoes from the basket at random and runs away. What is the probability that he gets at least 1 complete pair?

Below we outline two approaches, but only one of them is correct.

- First select one pair, the other two pieces of shoes can be taken arbitrarily, another pair, or any two of the remaining shoes, i.e.:

$$\frac{10\binom{18}{2}}{\binom{20}{4}}$$

- Find the probability of not having a complete pair at all. This can be done by selecting a single shoe, and then putting its matching pair aside. Keep in mind that the order of the selection does not count. Then passing to the complement event, we obtain

$$1 - \frac{\frac{20\cdot 18\cdot 16\cdot 14}{4!}}{\binom{20}{4}}$$

Check out that the two probabilities do not coincide! Which one is correct (if any)?

Example 14.5 Keep tossing a die until 6 comes out for the first time. What is the probability that we need an even number of tosses?

Let A stand for the event that we need an even number of tosses and A_k is the event that we need k tosses, respectively. Then we have (verify!)

$$P(A_k) = \left(\frac{5}{6}\right)^{k-1} \cdot \frac{1}{6}$$

for every $k = 1, 2, \ldots$ The event A can be expressed like this:

$$A = A_2 \cup A_4 \cup \ldots = \bigcup_{k=1}^{\infty} A_{2k}$$

On the right hand side the events mutually exclude each other, hence

$$P(A) = \sum_{k=1}^{\infty} P(A_{2k}) = \sum_{k=1}^{\infty} \left(\frac{5}{6}\right)^{2k-1} \cdot \frac{1}{6} = \frac{5}{11}$$

14.2 Sampling without replacement

Consider a set of N objects so that m of them are defective. Select a sample of n objects from the whole set at random, without replacement ($n \leq m$). Denote by A_k the event, that the sample contains exactly k defective objects ($0 \leq k \leq n$). Then

$$P(A_k) = \frac{\binom{m}{k} \cdot \binom{N-m}{n-k}}{\binom{N}{n}}$$

which we call the formula of sampling without replacement.

Example 14.6 From a deck of 52 playing cards we draw 5 cards at random without replacement. Find the probability that we selected exactly 2 diamonds.

Let A denote the given event. Making use of our formula we get

$$P(A) = \frac{\binom{13}{2} \cdot \binom{39}{3}}{\binom{52}{5}}.$$

In this argument the diamonds are the "defective objects".

Example 14.7 Determine the probability that in Hungarian lottery (5 winners out of 90) we have at least 2 winning numbers on a lottery ticket filled in at random.

Denote by A the event that we have 2 winning numbers, and by A_k the event that we have exactly k winning numbers on our ticket. Clearly, the events A_k are mutually exclusive for $k = 2, \ldots, 5$. On the other hand $A = A_2 \cup A_3 \cup A_4 \cup A_5$, and this implies

$$P(A) = \sum_{k=2}^{5} P(A_k) = \sum_{k=2}^{5} \frac{\binom{5}{k} \cdot \binom{85}{5-k}}{\binom{90}{5}}$$

since the probability of the disjoint union is the sum of the probabilities.

Example 14.8 From a deck of 52 playing cards we select 5 cards at random, without replacement. What is the probability that all 4 suits (clubs, diamonds, hearts, spades) are represented in the sample?

Examine the following argument. Let A denote the event that all 4 suits appear in the sample of 5 cards. Since the choice of any 5 cards is equally likely, we deal with a classical probability space.

In order to find out the number of favorable outcomes, take into account that we have 13 options for each suit. Once one card from each suit has been taken, then any card can be chosen from the remaining 48 cards.

The total number of outcomes: as many as the number of selections of 5 cards out of 52. So:

$$P(A) = \frac{13^4 \cdot 48}{\binom{52}{5}}$$

Is this the correct solution? If not, how could it be fixed?

14.3 Sampling with replacement

Consider again the set of N objects so that m of them are defective. Select n objects at random from the whole set, consecutively one after another with replacement. Let A_k denote the event that the sample contains exactly k defective items.

Examine the draws of different orders. Since the selection of k defectives and $n-k$ non-defectives in any order admits the probability

$$\frac{m^k \cdot (N-m)^{n-k}}{N^n} = \left(\frac{m}{N}\right)^k \left(1-\frac{m}{N}\right)^{n-k}$$

and we have exactly $\binom{n}{k}$ options for such selections, moreover they mutually exclude each other, we receive

$$P(A_k) = \binom{n}{k} \left(\frac{m}{N}\right)^k \left(1-\frac{m}{N}\right)^{n-k}$$

This equality is called the formula of sampling with replacement.

Example 14.9 Take 5 cards out of a deck of 52 cards at random, successively with replacement. (The card taken at a time is always put back.) Find the probability that this way

(a) exactly 2 diamonds are selected,

(b) at least 2 diamonds are selected.

Introduce the event A_k which means that exactly k diamonds are selected. Then

$$(a) \quad P(A_2) = \binom{5}{2} \left(\frac{1}{4}\right)^2 \left(1-\frac{1}{4}\right)^3$$

and

$$(b) \quad P(A_2 \cup \ldots \cup A_5) = \sum_{k=2}^{5} \binom{5}{k} \left(\frac{1}{4}\right)^k \left(\frac{3}{4}\right)^{5-k}$$

because the events $A_2, \ldots, A_5$ are mutually exclusive.

14.4 The Bernoulli experiment

The argument above can be generalized the following way. Suppose that the probability of an event A in a given experiment is a specific number $0 \leq p \leq 1$.

Let us assume that we carry out this experiment n times in a row (independently of each other) and every time we observe whether or not A occurs. This procedure is called the Bernoulli experiment.

Let $0 \leq k \leq n$ be a given integer. Denote by A_k the event that A occurs exactly k times out of the n trials.

Following the reasoning, analogous to the previos section, we immediately get

$$P(A_k) = \binom{n}{k} p^k (1-p)^{n-k}$$

for every integer $k = 0, 1, \ldots, n$.

Example 14.10 In the Hungarian lottery we say that a lottery ticket is a winning ticket, if it contains at least two winning numbers. Suppose we purchase 20 tickets and fill in them at random (independently of each other). Find the probability that we will have at least 5 winning tickets.

For just one ticket the probability of being a winning ticket is:

$$p = \sum_{k=2}^{5} \frac{\binom{5}{k} \cdot \binom{85}{5-k}}{\binom{90}{5}}$$

Since this is true for every ticket, and the tickets are filled in independently from each other, this problem can be regarded as a Bernoulli experiment, with the parameter p specified above. Therefore, applying our formula:

$$\sum_{k=5}^{20} \binom{20}{k} p^k (1-p)^{20-k}$$

where p is the probability given above.

Recitation and Exercises

1. Reading: Textbook-2, Sections 2.1, 2.2, 2.3, 2.4, 2.5.
2. Homework: Textbook-2, Exercises 2.20, 2.39, 2.42, 2.48, 2.64, 2.71, 2.72, 2. 113, 2.114, 2.115, 2.116.
3. Review: Highschool Combinatorics and Binomial Theorem (Textbook-2, Section 2.3), and "Probability Exercises"

Chapter 15

Conditional probability and Bayes' Rule

15.1 Conditional probability

In several problems we need to find the probability of the event A under the a priori condition that a certain event B occured. In such problems we take into account only those elements of the sample space, which also belong to B.

This actually means that the sample space Ω is reduced to the subset B, and we calculate the (conditional) probability of A with respect to B.

Definition 15.1 Consider the probability space $(\Omega, \mathcal{A}, P)$ and an event $B \in \mathcal{A}$ so that $P(B) \neq 0$. The conditional probability of the event $A \in \mathcal{A}$ with respect to B (read: probability of A given B) is defined by the equality:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Example 15.2 We toss a pair of dice, but we cannot see the outcome. Someone tells us that one of them is a 5. What is the probability that other one is 6?

ATTENTION! The answer is not 1/6 for the following reason!

Let A and B denote the following events:

$$B = \{\text{one of the tosses is } 5\} \qquad A = \{\text{the other one is } 6\}$$

On the one hand $P(B) = 11/36$ since there are 11 pairs that contain 5. On the other hand $A \cap B = \{(5,6), (6,5)\}$, and hence $P(A \cap B) = 2/36$. Therefore:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{2/36}{11/36} = \frac{2}{11}$$

Example 15.3 We are looking for a friend in the university main building. He can be in 5 rooms equally likely. The probability that he is in fact in the building is $0 < p < 1$. We have checked 4 of the 5 rooms, and he was in none of them. What is the probability that he is in the fifth room?

Let A_k denote the event that our friend is in room number k ($k = 1, \dots 5$), which means $P(A_1 \cup \dots \cup A_5) = p$. Since the events A_k are mutually exclusive, this implies that $P(A_k) = p/5$ for every index k. Therefore, in view of the De Morgan Rule we obtain:

$$\begin{aligned} P(A_5|\overline{A_1} \cap \dots \cap \overline{A_4}) &= P(A_5|\overline{A_1 \cup \dots \cup A_4}) \\ &= \frac{P(A_5 \cap (\overline{A_1 \cup \dots \cup A_4}))}{P(\overline{A_1 \cup \dots \cup A_4})} \end{aligned}$$

Obviously (think about it!):

$$A_5 \subset \overline{A_1 \cup \dots \cup A_4}$$

and hence

$$P(A_5 \cap (\overline{A_1 \cup \dots \cup A_4})) = P(A_5)$$

Consequently, the desired conditional probability is:

$$\begin{aligned} P(A_5|\overline{A_1} \cap \dots \cap \overline{A_4}) &= P(A_5|\overline{A_1 \cup \dots \cup A_4}) \\ &= \frac{P(A_5 \cap (\overline{A_1 \cup \dots \cup A_4}))}{P(\overline{A_1 \cup \dots \cup A_4})} \\ &= \frac{P(A_5)}{P(\overline{A_1 \cup \dots \cup A_4})} = \frac{p/5}{1 - 4p/5} = \frac{p}{5 - 4p} \end{aligned}$$

15.2 Independence

Consider the following simple example. Toss a die twice in a row, and we cannot see the result. Someone tells us that the first outcome is an odd number. Find the probability that the sum of the two numbers is 7.

Introduce the events A and B the following way:

$$A = \{\text{the sum is } 7\} \qquad B = \{\text{the first outcome is odd}\}$$

Then, by the definition of the conditional probability

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{3/36}{18/36} = \frac{1}{6}$$

In view of one of a previous example this means

$$P(A|B) = P(A)$$

that is "the occurance of B has no impact on the probability of A". This fact is expressed like "the event A is independent of the event B".

In case of $P(B) \neq 0$ the condition $P(A|B) = P(A)$ is equivalent to the equality:

$$P(A \cap B) = P(A) \cdot P(B) \tag{15.1}$$

Since we figure that independence is a symmetric relation (i.e. if A is independent of B, then B is also independent of A) and the above equality is visibly symmetric, relation (15.1) can serve as a comfortable definition for independence.

Definition 15.4 Let $(\Omega, \mathcal{A}, P)$ be a probability space, and $A, B \in \mathcal{A}$ are observable events. We say that A and B are *independent*, if they fulfill the condition (15.1).

Example 15.5 From a deck of 52 cards we draw 2 cards in succession with replacement. Find the probability that the first draw is a diamond, and the second draw is an Ace.

Introduce the following events:

$$A = \{\text{first draw is a diamond}\} \qquad B = \{\text{second draw is an Ace}\}$$

Then

$$P(A \cap B) = \frac{13 \cdot 4}{52^2} = \frac{13}{52} \cdot \frac{4}{52} = P(A) \cdot P(B)$$

that tells us that the events A and B are independent.

ATTENTION! We NEVER argue like: since the events A and B are "visibly" independent, therefore $P(A \cap B) = P(A) \cdot P(B)$. On the contrary: we conclude the independence of events by verifying this equality!

15.3 Theorem of Total Probability

Example 15.6 There are 3 identical envelopes on our desk,

1. the first contains 2 of 1000 Ft bills and 3 of 2000 Ft bills (banknotes),
2. the second contains 5 of 1000 Ft bills and 2 of 2000 Ft bills,
3. the third contains 5 of 2000 Ft bills.

We select one of the envelopes at random and draw one of the bills from the envelope. What is the probability that we take a 2000 Ft bill?

Let A denote the event that we draw a 2000 Ft bill. The probability $P(A)$ would be easy to determine if we knew, which envelope is selected. In particular, if B_k stands for the event that envelope k is selected, then the conditional probabilities $P(A|B_k)$ are $3/5$, $2/7$ and 1 respectively.

This observation immediately gives an idea of how to solve the problem. The events B_k are mutually exclusive and their union is the certain event. Thus:

$$A = A \cap \Omega = A \cap (B_1 \cup B_2 \cup B_3) = (A \cap B_1) \cup (A \cap B_2) \cup (A \cap B_3)$$

Since the events on the right-hand side are exclusive:

$$\begin{aligned} P(A) &= P(A \cap B_1) + P(A \cap B_2) + P(A \cap B_3) \\ &= P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + P(A|B_3)P(B_3) \\ &= \frac{3}{5} \cdot \frac{1}{3} + \frac{2}{7} \cdot \frac{1}{3} + 1 \cdot \frac{1}{3} \end{aligned}$$

The argument above can be exteded to an arbitrary number of events B_k. This leads us to the following definition.

Definition 15.7 We say that the observable events $B_1, B_2, \ldots \in \mathcal{A}$ form a *partition* of the sample space, if none of them has probability zero, and further

1. they are mutually exclusive, i.e. $B_i \cap B_j = \emptyset$ if $i \neq j$,
2. one of them occurs, i.e. $B_1 \cup B_2 \cup \ldots = \Omega$.

Following the analogous argument of Example 15.6 for an arbitrary number of events B_k, we come up with the following theorem.

Theorem 15.8 (Theorem of Total Probability) *Let us suppose that in the probability space $(\Omega, \mathcal{A}, P)$ the events $B_1, B_2, \ldots$ form a partition of the sample space. Then for any event $A \in \mathcal{A}$ we have*

$$P(A) = P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + \ldots$$

Proof. Indeed, if the events B_k form a partition of the sample space, then

$$A = (A \cap B_1) \cup (A \cap B_2) \cup (A \cap B_3) \cup \ldots$$

where the terms of the union are mutually exclusive. Thus:

$$P(A) = P(A \cap B_1) + P(A \cap B_2) + P(A \cap B_3) + \dots$$

By the very definition of the conditional probability, for every index k

$$P(A \cap B_k) = P(A|B_k) \cdot P(B_k)$$

and the theorem ensues. $\square$

Example 15.9 If the probability that the number of incoming calls to a call center is n on a given day is given by $0 < q_n < 1$, and every call is a wrong number with probability $0 < p < 1$ (independently of each other), find the probability that the number of wrong calls is exactly k on that day.

Introduce the following notations. Let A be the event that the center receives k wrong calls, and B_n is the event that the total number of incoming calls is n. In this case the events B_n form a partition of the sample space, hence by the theorem of total probability

$$P(A) = \sum_{n=1}^{\infty} P(A|B_n) \cdot P(B_n) = \sum_{n=k}^{\infty} q_n \binom{n}{k} p^k (1-p)^{n-k}$$

In fact, for $n \geq k$ the number of wrong calls can be regarded as the outcome of a Bernoulli experiment: how many wrong calls do we have out of n incoming calls. Keep in mind that we have $P(A|B_n) = 0$, for $n < k$.

15.4 Bayes' Rule

Let us return to Example 15.6. Assume that someone has performed the draw (we did not see it) and tells us that the draw is a 2000 Ft bill. What is the probability that the bill was taken from the first envelope?

Using our former notations, we need to find the conditional probability $P(B_1|A)$.

$$P(B_1|A) = \frac{P(A \cap B_1)}{P(A)} = \frac{P(A|B_1)P(B_1)}{P(A)}$$

The denominator of the fraction on the right-hand side can be evaluated by the theorem of total probability:

$$\begin{aligned} P(B_1|A) &= \frac{P(A|B_1)P(B_1)}{P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + P(A|B_3)P(B_3)} \\ &= \frac{\frac{3}{5} \cdot \frac{1}{3}}{\frac{3}{5} \cdot \frac{1}{3} + \frac{2}{7} \cdot \frac{1}{3} + 1 \cdot \frac{1}{3}} \end{aligned}$$

This argument can be extended to any partition of the sample space.

Theorem 15.10 **(Bayes' Rule)** *Let us suppose that in the probability space* $(\Omega, \mathcal{A}, P)$ *the events* $B_1, B_2, \ldots$ *form a partition of the sample space. Then for any event* $A \in \mathcal{A}$, $P(A) \neq 0$ *and any index* i *we have*

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + \ldots}$$

Proof. Indeed, by the definition of the conditional probability

$$P(B_i|A) = \frac{P(A \cap B_i)}{P(A)} = \frac{P(A|B_i)P(B_i)}{P(A)},$$

and our statement is proven by applying the theorem of total probability. □

Example 15.11 For instance, in our call center Example 15.9 the probability that the number of incoming calls on a given day is i provided that exactly k wrong calls have been registered is

$$P(B_i|A) = \frac{q_i \binom{i}{k} p^k (1-p)^{i-k}}{\sum_{n=k}^{\infty} q_n \binom{n}{k} p^k (1-p)^{n-k}}$$

for $i \geq k$, while this probability is 0, for $i < k$.

Recitation and Exercises

1. Reading: Textbook-2, Sections 2.6 and 2.7
2. Homework: Textbook-2, Exercises 2.80, 2.81, 2.87, 2.95, 2.97, 2.100, 2.109, 2.118
3. Review: Highschool Combinatorics and Binomial Theorem (Textbook-2, Section 2.3), and "Probability Exercises"

Chapter 16

Random variables and distributions

16.1 Random variables

Definition 16.1 Consider a probability space $(\Omega, \mathcal{A}, P)$. The function

$$X : \Omega \to \mathbb{R}$$

is called *random variable*, if for any $x \in \mathbb{R}$

$$\{X < x\} = \{\omega \in \Omega : X(\omega) < x\} \in \mathcal{A}$$

that is all level sets are observable (and hence possess a probability).

In the examples below specify the range R of the given random variables!

Example 16.2

1. Toss a pair of dice. Let X denote the sum of the numbers. Then $R = \{2, 3, \ldots, 12\}$
2. Let X be the least winning number in Hungarian lottery. Then $R = \{1, 2, \ldots, 86\}$
3. Keep tossing a die until 6 comes out for the first time. Denote by X the number of tosses. Then $R = \mathbb{N}$.
4. Pick a point arbitrarily on the unit disc (with center at the origin and radius 1). Let X denote the distance of the point from the origin. Then $R = [0, 1]$.

Definition 16.3 We say that a random variable is *discrete*, if its range is a countable set (finite or infinite). That is the elements of the range can be arranged in a finite or infinite sequence.

In our examples the first three random variables are discrete, but the fourth is not.

16.2 Distribution of discrete variables

Definition 16.4 Let X be a discrete random variable, whose range is $R = \{x_1, x_2, \ldots\}$. The sequence

$$p_k = P(X = x_k)\,, \qquad k = 1, 2, \ldots$$

is called the *distribution* of X.

Example 16.5 Consider our introductory examples for random variables

1. If X means the sum of the numbers when a pair of dice tossed, then the distribution can be given by the following *chart*:

x_k	2	3	4	$\ldots$	12
p_k	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\ldots$	$\frac{1}{36}$

2. If X means the least winning number in lottery, then the distribution can be given by the following *formula*:

$$p_k = \frac{\binom{90-k}{4}}{\binom{90}{5}} \qquad k = 1, 2, \ldots 86$$

3. If X means the number of tosses needed to get the first 6, the distribution of X is:

$$p_k = \left(\frac{5}{6}\right)^{k-1} \cdot \frac{1}{6} \qquad k = 1, 2, \ldots$$

   Unlike in the previous two examples, this distribution is an infinite sequence.

The most important properties of distributions are summed up in the following theorem.

Theorem 16.6 *Consider a discrete random variable X with range $R = \{x_1, x_2, \ldots\}$ and distribution $p_k = P(X = x_k)$, $k = 1, 2, \ldots$. Then*

- *$0 \le p_k \le 1$ for all indeces $k = 1, 2, \ldots$.*
- *$p_1 + p_2 + \ldots = 1$.*
- *If $a < b$ any real numbers, then*

$$P(a < X < b) = \sum_{a < x_k < b} p_k$$

where the sum is taken for all indeces k such that the inequality $a < x_k < b$ holds true. The last statement remains true if instead of the strict inequalities, the signs $\le$ are inserted simultaneously on both sides.

16.3 The cumulative distribution function

Definition 16.7 Consider a probability space $(\Omega, \mathcal{A}, P)$, and a random variable $X : \Omega \to \mathbb{R}$. For every $x \in \mathbb{R}$ set

$$F(x) = P(X < x)\,.$$

The function $F : \mathbb{R} \to [0, 1]$ is called the *cumulative distribution function* of X. (Or sometimes briefly distribution function.)

Example 16.8 It is easy to see that the distribution function of the random variable X defined in the introductory example 4, is

$$F(x) = \begin{cases} 0 & \text{if } x \le 0 \\ x^2 & \text{if } 0 < x \le 1 \\ 1 & \text{if } x > 1 \end{cases} \tag{16.1}$$

In fact we mean that the probability that the randomly picked point belongs to a given subset of the unit disc is proportional to the area of the subset. In particular, for instance $P(0 \le X < 1/2) = 1/4$.

In several problems in probability and statistics, and their applications we need to find a a probability of the form $P(a \le X < b)$. This probability can be expressed in term of the distribution function. The basic properties of the distribution function are summarized in the theorem below.

Theorem 16.9 *Let X be a random variable and consider its distribution function F.*

- *For every $x \in \mathbb{R}$ we have $0 \le F(x) \le 1$.*

- *F is monotone increasing and at every point continuous from the left.*
- $\lim_{x \to -\infty} F(x) = 0$ *and* $\lim_{x \to +\infty} F(x) = 1$.
- *For any real numbers $a < b$ we have*

$$P(a \leq X < b) = F(b) - F(a)\,.$$

If the range of a discrete random variable X is given by $R = \{x_1, x_2, \ldots\}$, where $x_1 < x_2 < \ldots$, and X takes these values with the probabilities $p_1, p_2, \ldots$ respectively, then the distribution function of X has the form:

$$F(x) = \begin{cases} 0 & \text{if } x \leq x_1 \\ p_1 + \ldots + p_k & \text{if } x_k < x \leq x_{k+1} \end{cases}$$

for each $k = 1, 2, \ldots$. Sketch the graph!

This tells us that in this case the distribution function is piecewise constant. Instead of using the formula $P(a \leq X < b) = F(b) - F(a)$, it is reasonable to collect all elements of the range of X that are in the open interval (a, b). In particular, if $P(X = x_k) = p_k$ for every k, then

$$P(a \leq X < b) = \sum_{a \leq x_k < b} p_k$$

On the right-hand side only the probabilities $P(X = x_k)$ appear, therefore, it is more convenient to rely on the distribution X.

16.4 The density function

Definition 16.10 We say that X is *continuously distributed*, if there exists an integrable function f on the real line with

$$F(x) = \int_{-\infty}^{x} f(t)\,dt$$

for every $x \in \mathbb{R}$. In this case the function f is called the *density function* of X.

For instance, in the example (16.1) we can easily verify that

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

If the random variable X is continuously distributed, then the distribution function F is continuous. Moreover, at every point x where the density function f is continuous, the distribution function F is differentiable, namely

$$F'(x) = f(x)$$

Theorem 16.11 *If X is continuously distributed and f is its density function, then for any real numbers $a < b$*

$$P(a \leq X < b) = \int_a^b f(t)\,dt$$

What can we say about the probability that the random variable X takes a single point? Let $a \in \mathbb{R}$ be any real number, then we conclude that

$$\begin{array}{rcl} P(X = a) & = & P(\bigcap_{n=1}^{\infty} \{a \leq X < a + \frac{1}{n}\}) = \lim_{n\to\infty} P(a \leq X < a + \frac{1}{n}) \\ & = & \lim_{n\to\infty} (F(a + \frac{1}{n}) - F(a)) = \lim_{x\to a+} F(x) - F(a) \end{array}$$

Consequently $P(X = a)$ equals the "jump" of F at the point a. **ATTENTION:** Why can we pass to the limit in the first line of the array formula?

A simple consequence of the previous argument is that $P(X = a) = 0$ if and only if F is continuous at the point a. In particular, if X is continuously distributed, then F is continuous on the whole real line, hence for any real numbers $a < b$ we get

$$P(a < X < b) = P(a \leq X \leq b)$$

We sum up the basic properties of density functions.

Theorem 16.12 *If f is the density function of the random variable X, then*

1. *$f(x) \geq 0$ for every $x \in \mathbb{R}$,*
2. $$\int_{-\infty}^{+\infty} f(x)\,dx = 1\,,$$
3. *if $a < b$ are any real numbers, then*
$$P(a < X < b) = P(a \leq X \leq b) = \int_a^b f(x)\,dx\,.$$

Example 16.13 Let us suppose that the density function of X is given by

$$f(x) = \begin{cases} x & \text{if } 0 < x \leq 1 \\ 2 - x & \text{if } 1 < x < 2 \\ 0 & \text{elsewhere} \end{cases}$$

ATTENTION! Verify that f fulfills all conditions of the previous theorem, so it is in fact a density function.

Then, for instance

$$\begin{aligned} P(0 \leq X \leq 3/2) &= P(0 < X < 3/2) = \int_0^{3/2} f(x)\,dx \\ &= \int_0^1 x\,dx + \int_1^{3/2} (2-x)\,dx \\ &= 1 - \int_{3/2}^2 (2-x)\,dx = \frac{7}{8} \end{aligned}$$

Recitation and Exercises

1. Reading: Textbook-2, Sections 3.1, 3.2 and 3.3
2. Homework: Textbook-2, Exercises 3.7, 3.9, 3.11, 3.14, 3.21, 3.22, 3.25, 3.26, 3.32 and 3.36
3. Review: Calculus, integration and infinite series and "Probability Exercises"

Chapter 17

Mean and variance

In everyday language by the mean (or expected value) of a random variable we think of the weighted average, by the standard deviation we think of the average deviation from the mean. Precise definitions will follow below.

17.1 Mean of discrete distributions

Definition 17.1 Consider a discrete random variable X whose distribution is given by

$$P(X = x_k) = p_k \qquad k = 1, 2, \ldots$$

We say that X has a mean (or expected value) if the series $\sum_{k=1}^{\infty} |x_k| \cdot p_k$ is convergent. In this case the sum

$$E(X) = \sum_{k=1}^{\infty} x_k \cdot p_k$$

is called the *mean* (or *expected value*) of X.

Remark that the convergence of the series $\sum_{k=1}^{\infty} |x_k| \cdot p_k$ is an important condition, because otherwise the sum $E(X)$ might depend on the rearrangement of the terms.

Example 17.2 Toss a pair of playing dice. Find the expected value of the sum of the two numbers.

Let X denote the sum of the two numbers, then the distribution of X is given in Example 16.5. Therefore, the mean of the sum is:

$$E(X) = \sum_{k=2}^{12} k p_k = 2 \cdot \frac{1}{36} + 3 \cdot \frac{2}{36} + \ldots + 12 \cdot \frac{1}{36} = 7\,.$$

Example 17.3 Take a sample of 5 cards from a deck of 52 playing cards at random. Find the expected number of diamonds in the sample.

Denote by X the number of diamonds in the sample. By using sampling without replacement, the distribution of X is given by:

$$P(X=k)=\frac{\binom{13}{k}\cdot\binom{39}{5-k}}{\binom{52}{5}} \qquad k=0,\ldots,5$$

Hence, the expected value is:

$$\begin{aligned} E(X) &= \sum_{k=0}^{5} kP(X=k)=\sum_{k=0}^{5} k\frac{\binom{13}{k}\cdot\binom{39}{5-k}}{\binom{52}{5}} \\ &= \frac{13}{\binom{52}{2}}\sum_{k=1}^{5}\binom{12}{k-1}\binom{39}{4-(k-1)}=\frac{13}{\binom{52}{5}}\cdot\binom{51}{4}=\frac{5}{4}. \end{aligned}$$

Example 17.4 Consider the Bernoulli experiment that we discussed in Section 14.4. and determine the expected number of occurances of the event A out of n trials.

Let X denote the number of times A occurs, then the distribution of X is:

$$P(X=k)=\binom{n}{k}p^k(1-p)^{n-k} \qquad k=0,1,\ldots,n$$

By virtue of the binomial theorem, the mean of X is:

$$\begin{aligned} E(X) &= \sum_{k=0}^{n} k\binom{n}{k}p^k(1-p)^{n-k} \\ &= np\sum_{k=1}^{n}\frac{(n-1)!}{(k-1)!(n-k)!}p^{k-1}(1-p)^{n-k} \\ &= np\sum_{k=1}^{n}\binom{n-1}{k-1}p^{k-1}(1-p)^{n-k}=np \end{aligned}$$

17.2 Mean of infinite distributions

In this section we investigate discrete random variables with infinite range.

Example 17.5 We keep tossing a die until 6 comes out for the first time. What is the expected number of tosses?

If X means the number of tosses, then the distribution of X is given by

$$P(X = k) = \left(\frac{5}{6}\right)^{k-1} \cdot \frac{1}{6} \qquad k = 1, 2, \ldots$$

Thus the expected value is

$$E(X) = \sum_{k=1}^{\infty} k \cdot \left(\frac{5}{6}\right)^{k-1} \cdot \frac{1}{6} = \frac{1}{6} \cdot \frac{1}{(1 - 5/6)^2} = 6$$

Example 17.6 Let λ be a given positive number, and consider a random variable X with the following distribution

$$P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda} \qquad k = 0, 1, 2, \ldots$$

In view of the power series of the exponential function, the mean of X is:

$$\begin{aligned} E(X) &= \sum_{k=0}^{\infty} k \frac{\lambda^k}{k!} e^{-\lambda} = \lambda \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} e^{-\lambda} \\ &= \lambda e^{-\lambda} \sum_{i=0}^{\infty} \frac{\lambda^i}{i!} = \lambda e^{-\lambda} e^{\lambda} = \lambda \end{aligned}$$

Example 17.7 In a box there is a black and a white ball. We take one ball at random. If it is black, we put it back, and add another black ball. We continue this process until the white ball is selected. Find the expected number of draws.

If X stands for the number of draws, then the distribution of X can be given like $P(X = 1) = 1/2$, and:

$$P(X = k) = \frac{1}{2} \cdot \frac{2}{3} \cdot \frac{3}{4} \cdots \frac{k-1}{k} \cdot \frac{1}{k+1} = \frac{1}{k(k+1)}, \quad k = 2, 3, \ldots$$

Therefore, for the mean of X we obtain the following infinite series:

$$E(X) = \sum_{k=1}^{\infty} k P(X = k) = \sum_{k=1}^{\infty} k \frac{1}{k(k+1)} = \sum_{k=1}^{\infty} \frac{1}{k+1}$$

Apart from the first term, this series exactly coincides with the harmonic series, which is divergent. Consequently, this random variable does not have a mean.

17.3 Mean of continuous distributions

Definition 17.8 Let X be a continuously distributed random variable with density function f. We say that X has a mean if the improper integral $\int_{-\infty}^{\infty} |x| \cdot f(x)\,dx$ is convergent. In this case the integral

$$E(X) = \int_{-\infty}^{\infty} x \cdot f(x)\,dx$$

is called the *mean* (or expected value) of X.

Example 17.9 Verify that the function f below defines a density function

$$f(x) = \frac{1}{\pi} \cdot \frac{1}{1+x^2} \qquad -\infty < x < \infty$$

(this is the so-called Cauchy distribution), but it has no mean, since the improper integral

$$\frac{1}{\pi} \int_{-\infty}^{\infty} \frac{x}{1+x^2}\,dx$$

is divergent. See Example 9.5 for the details.

Example 17.10 Consider an interval $[a, b]$ on the real line, and suppose the density function of the random variable X is given by

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{if } a < x < b \\ 0 & \text{elsewhere} \end{cases}$$

Verify that f is really a density function! Then the mean of X is

$$E(X) = \int_a^b \frac{x}{b-a}\,dx = \frac{1}{b-a} \cdot \frac{b^2-a^2}{2} = \frac{a+b}{2}$$

which is the midpoint of the interval $[a, b]$.

17.4 Basic properties of the mean

The mean $E(X^2)$ is called the second moment of the random variable X (if it exists). It can be shown that

$$E(X^2) = \begin{cases} \displaystyle\sum_k x_k^2 p_k & \text{if } X \text{ is discrete} \\ \displaystyle\int_{-\infty}^{\infty} x^2 f(x)\,dx & \text{if } X \text{ is continuous} \end{cases}$$

Below two fundamental properties of the mean are formulated.

Theorem 17.11

1. *If X has a mean, then for any real numbers α and β $E(\alpha X+\beta) = \alpha E(X)+\beta$.*
2. *If $E(X)$, $E(X^2)$ exist, then $E(\alpha X^2+\beta X+\gamma) = \alpha E(X^2)+\beta E(X)+\gamma$.*

Example 17.12 Let λ be a positive number, and assume that the density function of X is given by

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x > 0 \\ 0 & \text{elsewhere.} \end{cases}$$

Based on Example 9.3 this is really a density function, since

$$\int_0^\infty f(x)\,dx = 1\,.$$

On the other hand, Example 9.8 shows that the mean is

$$E(X) = \int_0^\infty x f(x)\,dx = \frac{1}{\lambda}\,.$$

The second moment can be evaluated by integration by parts (see Example 9.9):

$$E(X^2) = \int_0^\infty x^2 f(x)\,dx = \frac{2}{\lambda^2}\,.$$

17.5 Variance and standard deviation

The variance of a random variable is the average squared deviation from the mean.

Definition 17.13 The *variance* of a random variable of X (if it exists) is defined by

$$Var(X) = E((X-E(X))^2)$$

Then the *standard deviation* of X is $D(X)=\sqrt{Var(X)}$.

Sometimes the notation $D^2(X)$ is also used for the variance (for obvious reason).

The variance can be evaluated in the following simplified way:

$$\begin{aligned} Var(X) &= E((X-E(X))^2) = E(X^2 - 2E(X)X + E(X)^2) \\ &= E(X^2) - 2E(X)^2 + E(X)^2 = E(X^2) - E(X)^2 \end{aligned}$$

Basic properties of the variance:

$$Var(\alpha X + \beta) = \alpha^2 Var(X), \qquad D(\alpha X + \beta) = |\alpha| \cdot D(X)$$

Verify these two directly, based on the definition!

Example 17.14 Find the variance and standard deviation of the continuously distributed random variable X in Example 17.12 (where $\lambda > 0$ is a given constant).

$$Var(X) = E(X^2) - E(X)^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2},$$

and in particular

$$D(X) = \frac{1}{\lambda}$$

Example 17.15 Consider now the continuously distributed random variable X examined in Example 17.10. We can calculate the second moment this way:

$$\begin{aligned} E(X^2) &= \int_{-\infty}^{\infty} x^2 f(x)\,dx = \int_a^b \frac{x^2}{b-a}\,dx = \frac{1}{b-a}\left[\frac{x^3}{3}\right]_a^b \\ &= \frac{b^3-a^3}{3(b-a)} = \frac{b^2+ab+a^2}{3} \end{aligned}$$

Therefore, the variance is:

$$Var(X) = E(X^2) - E(X)^2 = \frac{b^2+ab+a^2}{3} - \frac{a^2+2ab+b^2}{4} = \frac{(b-a)^2}{12}$$

moreover, the standard deviation of X is the square root of the variance:

$$D(X) = \frac{b-a}{2\sqrt{3}}.$$

Recitation and Exercises

1. Reading: Textbook-2, Sections 4.1 and 4.2.
2. Homework: Textbook-2, Exercises 4.1, 4.2, 4.4, 4.8, 4.12, 4.13, 4.14, 4.34, 4.37, 4.38, 4.43 and 4.50
3. Review: Calculus, integration, improper integrals and infinite series, and "Probability Exercises"

Chapter 18

Special discrete distributions

This chapter gives a summary of the most widely applied discrete distributions.

18.1 Characteristic distribution

Let $(\Omega, \mathcal{A}, P)$ be a probability space and consider an event $A \in \mathcal{A}$ with $P(A) = p$, and $0 < p < 1$. Then the random variable

$$X = \begin{cases} 1 & \text{if } A \text{ occurs} \\ 0 & \text{if } A \text{ does not occur} \end{cases}$$

possesses the distribution

$$P(X = 0) = 1 - p \quad P(X = 1) = p$$

This is called the *characteristic distribution* associated with the event A.

Theorem 18.1

- *The parameter of the distribution is:* $0 < p < 1$.
- *The mean of this distribution:* $E(X) = p$
- *The variance of this distribution:* $Var(X) = p(1 - p)$.

Proof. We only need to verify the variance. Since the second moment is $E(X^2) = p$, the statement ensues. □

18.2 Binomial distribution

Let $(\Omega, \mathcal{A}, P)$ be a probability space, and consider the Bernoulli experiment, where we carry out n independent experiments in a row, and every time we observe if a given event A occurs. Suppose that $P(A) = p$, $0 < p < 1$ is given. Let X denote how many times A comes out. By the Bernoulli experiment the distribution of X is given by

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \qquad k = 0, 1, 2, \ldots, n$$

This distribution is called the *binomial distribution*.

Theorem 18.2

- *The parameters of the distribution:* $n \in \mathbb{N}$ *and* $0 < p < 1$.
- *The mean of the distribution:* $E(X) = np$
- *The variance of the distribution:* $Var(X) = np(1-p)$.

Proof. In view of Example 17.4 we only need to check the variance First find the second moment.

$$\begin{aligned} E(X^2) &= \sum_{k=1}^{n} k^2 \binom{n}{k} p^k (1-p)^{n-k} = \\ &= \sum_{k=2}^{n} k(k-1) \binom{n}{k} p^k (1-p)^{n-k} + \sum_{k=1}^{n} k \binom{n}{k} p^k (1-p)^{n-k} \\ &= n(n-1)p^2 \sum_{k=2}^{n} \binom{n-2}{k-2} p^{k-2} (1-p)^{n-k} + np = (n^2 - n)p^2 + np\,. \end{aligned}$$

Therefore, the variance is

$$Var(X) = E(X^2) - E(X)^2 = n(n-1)p^2 + np - n^2p^2 = np(1-p)$$

where we observed that the second sum in the second line is precisely the mean.
□

18.3 Hypergeometric distribution

Examine the following sampling without replacement problem. Consider a set of N objects in which m of them are defective. Select a sample of n objects without replacement from the whole set ($n \le m$). Let X denote the number of defective objects in the sample. Then the distribution of X is:

$$P(X = k) = \frac{\binom{m}{k} \cdot \binom{N-m}{n-k}}{\binom{N}{n}} \qquad k = 0, 1, 2, \ldots, n$$

This distribution is called the *hypergeometric distribution.*

Theorem 18.3

- *The parameters of the distribution:* $N, m, n \in \mathbb{N}$.
- *The mean of the distribution:*

$$E(X) = n \cdot \frac{m}{N}$$

- *The variance of the distribution:*

$$Var(X) = \frac{N-n}{N-1} \cdot n \cdot \frac{m}{N} \left(1 - \frac{m}{N}\right) .$$

Proof. By applying the argument of Example 17.3, we again only have to calculate the variance. First find the second moment.

$$\begin{aligned} E(X^2) &= \sum_{k=1}^{n} k^2 \frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}} = \sum_{k=2}^{n} k(k-1) \frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}} + \sum_{k=1}^{n} k \frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}} \\ &= \frac{m(m-1)n(n-1)}{N(N-1)} \sum_{k=2}^{n} \frac{\binom{m-2}{k-2} \cdot \binom{N-m}{n-k+2}}{\binom{N-2}{n-2}} + n\frac{m}{N} \\ &= \frac{m(m-1)n(n-1)}{N(N-1)} + n\frac{m}{N} . \end{aligned}$$

Then we conclude

$$Var(X) = \frac{m(m-1)n(n-1)}{N(N-1)} + n\frac{m}{N} - n^2\frac{m^2}{N^2} = \frac{N-n}{N-1} \cdot n \cdot \frac{m}{N}\left(1 - \frac{m}{N}\right) ,$$

just as we stated. □

18.4 Geometric distribution

Take a probability space $(\Omega, \mathcal{A}, P)$, and consider an event A such that $P(A) = p$, wher $0 < p < 1$ is given. Keep performing the experiment until the event A occurs for the first time. Let X denote the number of trials. The distribution of X is given by:

$$P(X = k) = (1-p)^{k-1}p \qquad k = 1, 2, \ldots$$

This distribution is called the *geometric distribution*.

Theorem 18.4

- *The parameter of the distribution:* $0 < p < 1$.
- *The mean of the distribution:*
$$E(X) = \frac{1}{p}$$
- *The variance of the distribution:*
$$Var(X) = \frac{1-p}{p^2}\,.$$

Proof. The mean of this distribution is easily obtained by following the argument of Example 17.5, so we only need to find the variance. The second moment can be evaluated the following way. Using the second derivative of the power series at $|x| < 1$, we have

$$\sum_{k=2}^{\infty} k(k-1)x^{k-2} = \frac{2}{(1-x)^3}$$

If we employ this identity with $x = 1-p$ we receive

$$\begin{aligned} E(X^2) &= \sum_{k=1}^{\infty} k^2(1-p)^{k-1}p = \sum_{k=2}^{\infty} k(k-1)(1-p)^{k-1}p + \sum_{k=1}^{\infty} k(1-p)^{k-1}p \\ &= p(1-p)\sum_{k=1}^{\infty} k(k-1)(1-p)^{k-2} + \frac{1}{p} = \frac{2p(1-p)}{p^3} + \frac{1}{p}\,. \end{aligned}$$

Thus we get

$$Var(X) = E(X^2) - E(X)^2 = \frac{2p(1-p)}{p^3} + \frac{1}{p} - \frac{1}{p^2} = \frac{1-p}{p^2}$$

and this is what we needed. □

18.5 Poisson distribution

Suppose that X is a random variable, whose range is $\{0\} \cup \mathbb{N}$ and its distribution is defined by

$$P(X = k) = \frac{\lambda^k}{k!}e^{-\lambda} \qquad k = 0, 1, 2, \ldots$$

wher $\lambda > 0$ is a given number.

It is not hard to see that we really defined a distribution. Indeed,

$$\sum_{k=0}^{\infty} \frac{\lambda^k}{k!} e^{-\lambda} = e^{-\lambda} \cdot e^{\lambda} = 1$$

based on the power series of the natural exponential function. This infinite distribution is called the *Poisson distribution*.

Theorem 18.5

- *The parameter of the distribution:* $\lambda > 0$.
- *The mean of the distribution:* $E(X) = \lambda$,
- *The variance of the distribution:* $Var(X) = \lambda$.

Proof. In view of Example 17.6 we only have to calculate the variance. The second moment is obtained as follows.

$$\begin{aligned} E(X^2) &= \sum_{k=1}^{\infty} k^2 \frac{\lambda^k}{k!} e^{-\lambda} = \sum_{k=1}^{\infty} k(k-1) \frac{\lambda^k}{k!} e^{-\lambda} + \sum_{k=1}^{\infty} k \frac{\lambda^k}{k!} e^{-\lambda} \\ &= \lambda^2 \sum_{k=2}^{\infty} \frac{\lambda^{k-2}}{(k-2)!} e^{-\lambda} + \lambda \,. \end{aligned}$$

Hence, the variance is

$$Var(X) = E(X^2) - E(X)^2 = \lambda^2 + \lambda - \lambda^2 = \lambda$$

and that completes the proof. □

Let us remark that the Poisson distribution can be regarded as the "limit distribution" of the binomial distribution as it is explained in the following.

Theorem 18.6 *If $\lambda > 0$ is fixed and $0 < p_n < 1$ is a sequence with $np_n = \lambda$, then*

$$\lim_{n\to\infty} \binom{n}{k} p_n^k (1-p_n)^{n-k} = \frac{\lambda^k}{k!} e^{-\lambda}$$

for every $k = 0, 1, 2, \ldots$.

Proof. Indeed, for each fixed index k we have

$$\begin{aligned} \binom{n}{k} p_n^k (1-p_n)^{n-k} &= \binom{n}{k} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k} \\ &= \frac{n(n-1)\ldots(n-k+1)}{n^k} \cdot \frac{\lambda^k}{k!} \left(1-\frac{\lambda}{n}\right)^n \left(1-\frac{\lambda}{n}\right)^{-k} \end{aligned}$$

Here examine the limits of the four factors separately. It is easy to see that they are 1, $\lambda^k/k!$, $e^{-\lambda}$ and 1 respectively. That proves our theorem. □

Practically, this theorem means that for large values of n and for small values of p the binomial distribution can be approximated by the Poisson distribution, i.e.

$$\binom{n}{k} p_n^k (1-p_n)^{n-k} \approx \frac{\lambda^k}{k!} e^{-\lambda}$$

for every index $0 \leq k \leq n$.

Example 18.7 Let us suppose that in a brand new Suzuki Vitara the probability that the airbag is defective, is 0.002 independently from each other. The factory announces withdrawal if at least 10 malfunctions are reported for the 2000 cars that are manufactured in a month. Find the probability that no withdrawal has to be announced.

let X denote the number of defective cars in a given month. Since this is a Bernoulli-experiment (the probability of malfunction is 0.002 independently from each other), it follows that X has binomial distribution with parameters $n = 2000$ and $p = 0.002$. Therefore the exact value of the probabilty is

$$P(X \leq 9) = \sum_{k=0}^{9} \binom{2000}{k} 0.002^k 0.998^{2000-k}$$

which not easy to handle. Based on our theorem, we can give an approximation of this probability by using the Poisson distribution (we say that "n is sufficiently large and p is sufficiently small"), moreover $\lambda = np = 4$, so

$$\sum_{k=0}^{9} \binom{2000}{k} 0.002^k 0.998^{2000-k} \approx \sum_{k=0}^{9} \frac{4^k}{k!} e^{-4} \approx 0.9919$$

This latter value can be determined by looking up in the Poisson tables that can be found on page 732 in our Textbook.

Recitation and Exercises

1. Reading: Textbook-2, Sections 5.1, 5.2, 5.3 and 5.5
2. Homework: Textbook-2, Exercises 5.5, 5.9, 5.10, 5.15, 5.27, 5.33, 5.47, 5.56, 5.60, 5.66, 5.70 and 5.72
3. Review: Calculus, integration, improper integrals and infinite series, and "Probability Exercises"

Chapter 19

Special continuous distributions

19.1 Uniform distribution

Let $[a, b]$ be a given finite interval. Consider a random variable X with the following density function:

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{if } a < x < b \\ 0 & \text{elsewhere} \end{cases}$$

This random variable X is said to have *uniform distribution* on the interval $[a, b]$. The name comes from the fact that the probability that X is in a subinterval of $[a, b]$ is proportional to the length of the subinterval.

Theorem 19.1

- *The parameters of the distribution: a and b, $a < b$.*
- *The mean of the distribution:*

$$E(X) = \frac{a+b}{2}$$

- *The variance of the distribution:*

$$Var(X) = \frac{(b-a)^2}{12}\,.$$

Proof. These statements are immediate consequences of the results in Examples 17.10 and 17.15. □

Example 19.2 Let X be a uniformly distributed random variable with $E(X) = 5$ and $Var(X) = 3$. Find the probability $P(4 < X < 10)$.

The unknown endpoints of the interval a and b satisfy the following equations:

$$\begin{aligned}\frac{a+b}{2} &= 5\\ \frac{(b-a)^2}{12} &= 3\end{aligned}$$

whose solutions are $a = 2$ and $b = 8$. Therefore,

$$P(4 < X < 10) = P(4 < X < 8) = \frac{2}{3}$$

since the subinterval beyond $[4, 8]$ comes with 0 probability.

19.2 Exponential distribution

Let $\lambda > 0$ be a fixed number. Consider the random variable X with the following density function

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x > 0 \\ 0 & \text{elsewhere} \end{cases}$$

In this case we say that X has *exponential distribution*. nevezzük.

ATTENTION: Verify that f really defines a density function! Sketch the graph of the function!

Theorem 19.3

- *The parameter of the distribution:* $\lambda > 0$.
- *The mean of the distribution:* $E(X) = 1/\lambda$,
- *The variance of the distribution:* $Var(X) = 1/\lambda^2$.

Proof. Our theorem is an immediate consequence of the eqalities in Examples 17.12 and 17.14. □

Example 19.4 Consider an exponentially distributed random variable X with a given parameter $\lambda > 0$. Find the probability $P(X > E(X))$.

Our theorem claims that $E(X) = 1/\lambda$, thus

$$P(X > E(X)) = P\left(X > \frac{1}{\lambda}\right) = \int_{1/\lambda}^{\infty} \lambda e^{-\lambda x}\, dx = \left[-e^{-\lambda x}\right]_{1/\lambda}^{\infty} = \frac{1}{e}.$$

We say that the exponential distribution is memoryless in the following sense. If X is exponentially distributed with a given parameter $\lambda > 0$, and $t, s > 0$ are given positive numbers, then

$$P(X > t + s | X > t) = P(X > s)\,.$$

Indeed, the event $\{X > t + s\}$ implies the event $\{X > t\}$, therefore, the conditional probability on the left-hand side can be written like

$$\begin{aligned} P(X > t + s | X > t) &= \frac{P(X > t + s)}{P(X > t)} = \frac{1 - \int_0^{t+s} \lambda e^{-\lambda x}\, dx}{1 - \int_0^{t} \lambda e^{-\lambda x}\, dx} \\ &= \frac{e^{-\lambda(t+s)}}{e^{-\lambda t}} = e^{-\lambda s} = 1 - \int_0^{s} \lambda e^{-\lambda x}\, dx = P(X > s)\,. \end{aligned}$$

If for instance X denotes the waiting time between two occurances (i.e. two telephone calls, or two customers, etc.), then the lack of memory means that the further waiting time does not depend on how much we have been waiting.

Conversely, it can also be proven that if a continuous distribution is memoryless, then it is necessarily the exponential distribution.

There is an interesting relationship between the Poisson distribution and the exponential distribution. In particular, if the waiting times between successive occurances are independent, exponentially distributed random variables with identical parameter $\lambda > 0$, then the number of occurances in a unit time interval has Poisson distribution with the same parameter. These features will be discussed in later chapters.

19.3 The standard normal distribution

Because of the central role of the standard normal distribution we use a distinguished notation for its density function and cumulative distribution function.

Definition 19.5 We say that the random variable Z has *standard normal distribution*, if its density function is given by

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \qquad -\infty < x < \infty$$

In view of formula (9.2), we see that φ really defines a density function. As an exercise analyze the function φ, and show that it possesses the following properties.

$$\lim_{x\to-\infty} \varphi(x) = \lim_{x\to+\infty} \varphi(x) = 0$$

moreover φ is strictly monotone increasing on the interval $(-\infty, 0)$, strictly monotone decreasing on the interval $(0, \infty)$, and reaches its global maximum at $x = 0$.

By analyzing the second derivative, we can see that φ is convex on the intervals $(-\infty, 1)$ and $(1, +\infty)$, while it is concave on the interval $(-1, 1)$, and consequently has points of inflection at $x = -1$ and $x = 1$ respectively.

EXERCISE: CREATE THE GRAPH OF THE FUNCTION!

Theorem 19.6

- *The parameter of the distribution: no parameter.*
- *The mean of the distribution:* $E(Z) = 0$.
- *The variance of the distribution:* $Var(Z) = 1$.

Proof. Example 9.6 shows that $E(Z) = 0$, and equality (9.3) tells us that $E(Z^2) = 1$. Therefore

$$Var(Z) = E(Z^2) - E(Z)^2 = 1\,.$$

as we stated. □

Let Φ denote the standard normal cumulative distribution function, i.e.

$$\Phi(x) = \int_{-\infty}^{x} \varphi(t)\, dt\,.$$

This function has the properties of cumulative distribution functions, but its interesting feature is that it cannot be expressed explicitly in terms of elementary functions or their finite combinations.

Observe however that φ is an even function, in other words it is symmetric with respect to the y-axis. This implies that $\Phi(0) = 1/2$, and further

$$\Phi(-x) = 1 - \Phi(x) \tag{19.1}$$

for every real number x.

Example 19.7 Because of its central role in Statistics and other applications we can find tables for the values of the Φ function in most probability textbooks and spreadsheet programs like the Microsoft Windows Office Excel application. See the tables on pages 735–736 of our Textbook!

If for example Z is a standard normally distributed random variable, the find the probability

$$P(-2 < Z < 2)$$

Using the table on page 736 of our Textbook, we get

$$\begin{aligned} P(-2 < Z < 2) &= \Phi(2) - \Phi(-2) = \Phi(2) - (1 - \Phi(2)) = 2\Phi(2) - 1 = \\ &= 2 \cdot 0.9772 - 1 = 0.9544 \end{aligned}$$

where we exploited the symmetry property (19.1).

19.4 Normal distribution

Definition 19.8 Let m and σ be given real numbers where $\sigma > 0$. Let Z be a standard normally distributed random variable, then the random variable

$$X = \sigma Z + m$$

is said to have *normal distribution with* (m, σ)*-parameters* (or briefly (m, σ)-normal distribution).

Making use of the properties of the standard normal distribution, and the properties of the mean and the variance (refer to Theorem 17.11) we get the following theorem for (m, σ)-normal distributions.

Theorem 19.9

- *The parameters of the distribution:* $m \in \mathbb{R}$, $\sigma > 0$.
- *The mean of the distribution:* $E(X) = m$,
- *The variance of the distribution:* $Var(X) = \sigma^2$.

How can we find the cumulative distribution function and the density function of this random variable X? Let F denote the cumulative distribution function of X, and take a real number x arbitrarily. Then

$$F(x) = P(X < x) = P(\sigma Z + m < x) = P\left(Z < \frac{x-m}{\sigma}\right) = \Phi\left(\frac{x-m}{\sigma}\right)$$

IMPORTANT! It is vital that $\sigma > 0$, so when we divide by σ the inequality will not change!

We get the density function of X by differentiating F: for every $x \in \mathbb{R}$ we have

$$f(x) = F'(x) = \frac{1}{\sigma}\varphi\left(\frac{x-m}{\sigma}\right) = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-m)^2}{2\sigma^2}}$$

by the Chain-Rule. This function has a global maximum at $x = m$, furthermore it has points of inflection at $x = m - \sigma$ and $x = m + \sigma$ respectively. CREATE A PICTURE!

Example 19.10 For an (m, σ)-normally distributed random variable X the probability of being in an interval can always be expressed in terms of the standard normal cumulative distribution function Φ.

Indeed, if $a < b$ are arbitrarily taken real numbers, then

$$P(a < X < b) = F(b) - F(a) = \Phi\left(\frac{b-m}{\sigma}\right) - \Phi\left(\frac{a-m}{\sigma}\right).$$

For example, for a normally distributed random variable X with parameters $m = 10$ and $\sigma = 2$ we have

$$\begin{aligned} P(7 < X < 13) &= F(13) - F(7) = \Phi(1.5) - \Phi(-1.5) = 2\Phi(1.5) - 1 = \\ &= 2 \cdot 0.9332 - 1 = 0.8664 \end{aligned}$$

where we used the symmetry of Φ, and the tables on page 736 in the Textbook.

Recitation and Exercises

1. Reading: Textbook, Sections 6.1, 6.2, 6.3, 6.4, 6.6
2. Homework: Textbook, Exercises 6.2, 6.3, 6.4, 6.6, 6.7, 6.9, 6.11, 6.15, 6.17, 6.18, 6.45 and 6.46 5.66, 5.70 and 5.72
3. Review: Calculus, integration, improper integrals and infinite series and "Probability Exercises"

Chapter 20

Joint distributions

20.1 Joint cumulative distribution function

Defintion 20.1 Let X and Y be random variables (not necessarily on the same sample space). For any real numbers x and y the function

$$F(x,y) = P(X < x, Y < y)$$

is called the *joint cumulative distribution function* of X and Y.

The following statement comes directly from the definition.

Proposition 20.2 *If F is a joint cumulative distribution function, then*

$$\lim_{x\to-\infty} F(x,y) = \lim_{y\to-\infty} F(x,y) = 0$$

for any fixed real y and x respectively, moreover

$$\lim_{x,y\to+\infty} F(x,y) = 1$$

Similarly to the one dimensional case, we separately discuss discrete and continuous distributions.

20.2 Discrete joint distributions

Definition 20.3 Assume that the range of the variable X is $\{x_1, x_2, \ldots\}$, and the range of the variable Y is $\{y_1, y_2, \ldots\}$. Then the joint distribution of X and Y is given by

$$p_{ik} = P(X = x_i, Y = y_k) \qquad i = 1, 2, \ldots \quad k = 1, 2, \ldots$$

These values can be arranged in a chart:

$y \setminus x$	x_1	x_2	x_3	$\cdots$
y_1	p_{11}	p_{21}	p_{31}	$\cdots$
y_2	p_{12}	p_{22}	p_{32}	$\cdots$
y_3	p_{13}	p_{23}	p_{33}	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

Obviously for all indeces $p_{ik} \geq 0$ and $\sum_i \sum_k p_{ik} = 1$.

Let A be a subset of the plane. By using the joint distribution, how can we evaluate the probability $P((X,Y) \in A)$? Collect all values x_i and y_k for which $(x_i, y_k) \in A$, then

$$P((X,Y) \in A) = \sum_{(x_i,y_k)\in A} p_{ik}$$

Example 20.4 For instance, if we consider the following joint distribution

$y \setminus x$	0	1	2	3
0	0.1	0.08	0.13	0.04
1	0.04	0.2	0.08	0
2	0.03	0	0.05	0.25

then for the subset $A = \{(x,y) \in \mathbb{R}^2 : x + y \geq 3\}$ we have:

$$P(X + Y \geq 3) = 0.04 + 0.08 + 0.05 + 0.25 = 0.42$$

A natural question to ask is that based on the joint distribution, how can we determine the distributions of X and Y alone? As we conclude from the definition

$$p_i = P(X = x_i) = \sum_k p_{ik} = \sum_k P(X = x_i, Y = y_k) \qquad i = 1, 2, \ldots$$

Namely, the probability $p_i = P(X = x_i)$ can be obtained by taking the sum of the elements in the i-th column. Therefore, the sums of columns provide the distribution of X.

In an analogous way,

$$q_k = P(Y = x_k) = \sum_i p_{ik} = \sum_i P(X = x_i, Y = y_k) \qquad k = 1, 2, \ldots$$

which means that the distribution of Y is obtained by taking the sums of rows.

Definition 20.5 The distributions of X and Y are called the *marginal distributions* of the joint distribution.

20.3 Continuous joint distributions

Definition 20.6 We say that X and Y are continuously distributed, if there exists a non-negative integrable function f on the plane suvh that for all real numbers x and y we have

$$F(x,y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(t,s)\, ds\, dt$$

where F is the joint cumulative distribution function of the random variables X and Y. This function f is called the *joint density function* of X and Y.

Clearly, if f is a joint density function, then

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x,y)\, dx\, dy = 1$$

Example 20.7 Let A be a subset of the plane. How can we find the probability $P((X,Y) \in A)$? If f is the joint density function of X and Y, then

$$P((X,Y) \in A) = \iint_A f(x,y)\, dy\, dx$$

For example if we consider the joint density function

$$f(x,y) = \begin{cases} \frac{2}{3}(x+2y) & \text{if } 0 < x < 1,\ 0 < y < 1 \\ 0 & \text{elsewhere} \end{cases} \tag{20.1}$$

then for the set $A = \{(x,y) \in \mathbb{R}^2 : x < 1/2,\ y < 1/2\}$ we have

$$P(X < 1/2,\, Y < 1/2) = \frac{2}{3} \int_0^{1/2} \int_0^{1/2} (x+2y)\, dy\, dx = \frac{1}{8}$$

If the joint density function is given, how can we find the density of X or Y alone? It can be shown that if f_X denotes the density of X, then for every point x

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)\, dy$$

and analogously

$$f_Y(y) = \int_{-\infty}^{\infty} f(x,y)\, dx$$

for every point y.

Definition 20.8 The functions f_X and f_Y are called the *marginal densities* of the joint density function.

Example 20.9 For instance in the case of the joint density in the previous example

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)\,dy = \begin{cases} \int_0^1 \frac{2}{3}(x+2y)\,dy & \text{if } 0 < x < 1 \\ 0 & \text{elsewhere} \end{cases}$$

The marginal density of Y in a similar way

$$f_Y(y) = \begin{cases} \frac{1}{3}(4y+1) & \text{if } 0 < y < 1 \\ 0 & \text{elsewhere} \end{cases}$$

20.4 Independence

Definition 20.10 Let X and Y be random variables with joint cumulative distribution function F. Denote by F_X and F_Y the marginal cumulative distribution functions of X and Y respectively. We say that X and Y are *independent*, if

$$F(x,y) = F_X(x) \cdot F_Y(y)$$

for all real numbers x, y.

In other words we may say that X and Y are independent, if

$$P(X < x, Y < y) = P(X < x) \cdot P(Y < y)$$

for all real numbers x, y. Now we reformulate this definition for the discrete and for the continuous case.

Let X and Y be discrete random variables with joint distribution

$$P(X = x_i, Y = y_k) = p_{ik} \qquad i = 1, 2, \ldots \quad k = 1, 2, \ldots$$

Consider the marginal distributions of X and Y:

$$P(X = x_i) = p_i \quad i = 1, 2, \ldots \qquad P(Y = y_k) = q_k \quad k = 1, 2, \ldots$$

Theorem 20.11 *X and Y are independent if and only if*

$$p_{ik} = p_i \cdot q_k$$

for all indices i and k.

Our theorem states that the random variables are independent if and only if their joint distribution can be expressed as the product of the marginal distributions. For instance in Example 20.4 the variables are not independent, since for the very first element

$$0.17 \cdot 0.35 = p_1 \cdot q_1 \neq p_{11} = 0.1\,,$$

VERIFY!

Now let X and Y be continuously distributed random variables with joint density function f. Denote by f_X and f_Y the marginal densities of X and Y respectively.

Theorem 20.12 *X and Y are independent if and only if*

$$f(x, y) = f_X(x) \cdot f_Y(y)$$

for every real x and y.

Proof. Easily follows from the equality $F(x, y) = F_X(x) \cdot F_Y(y)$. □

Example 20.13 In Example 20.1 the random variables are not independent, since

$$f_X(x) \cdot f_Y(y) \neq f(x, y)\,,$$

i.e. the joint density cannot be expressed as the product of the marginal densities.

However, if the joint density of X and Y is given by

$$f(x, y) = \begin{cases} 4xy & \text{if } 0 < x < 1,\ 0 < y < 1 \\ 0 & \text{elsewhere} \end{cases}$$

then X and Y are independent. Indeed

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y)\, dy = \int_0^1 4xy\, dy = \begin{cases} 2x & \text{if } 0 < x < 1 \\ 0 & \text{elsewhere}\ . \end{cases}$$

and by the symmetry of f the marginal density f_Y has the same form with respect to y. Thus

$$f(x, y) = f_X(x) \cdot f_Y(y)$$

for all real numbers x and y.

20.5 Conditional distributions

Consider the discrete random variables X és Y with joint distribution $P(X = x_i, Y = y_k) = p_{ik}$, where $i = 1, 2, \dots$ and $k = 1, 2, \dots$.

Definition 20.14 Suppose that for a specific index k we have $P(Y = y_k) > 0$. Then by the *conditional distribution* of X under the condition $Y = y_k$ we mean the distribution

$$P(X = x_i | Y = y_k) = \frac{P(X = x_i, Y = y_k)}{P(Y = y_k)}, \quad i = 1, 2, \dots$$

ATTENTION! Verify directly that we really have defined a distribution!

Definition 20.15 By the *conditional expected value* of X under the condition $Y = y_k$ we mean the sum

$$E(X|Y = y_k) = \sum_{i=1} x_i \cdot P(X = x_i | Y = y_k)$$

that may consist of finitely many or infinitely many terms depending on the range of X (this is why we do not indicate the upper bound of the summation).

Example 20.16 Let us examine again the joint distribution in Example 20.4. Then $P(Y = 1) = 0.32$, and the conditional expected value of X under the condition $Y = 1$

$$E(X|Y = 1) = 0 \cdot 0.04 + 1 \cdot 0.2 + 2 \cdot 0.08 + 3 \cdot 0 = 0.36$$

Verify this calculation!

Recitation and Exercises

1. Reading: Textbook-2, Section 3.4
2. Homework: Textbook-2, Exercises 3.39, 3.40, 3.41, 3.42, 3.43, 3.45, 3.47, 3.49, 3.50, 3.51, 3.52 and 3.53
3. Review: Calculus, integration, improper integrals and infinite series, and "Probability Exercises"

Chapter 21

Covariance and correlation

21.1 Mean of a sum

Tétel 21.1 *If the random variables X and Y both have a mean, then so does $X+Y$ and*

$$E(X+Y) = E(X) + E(Y)$$

Proof. We give an outline of the proof in the discrete case, the continuous case is analogous.

$$\begin{aligned} E(X+Y) &= \sum_i \sum_k (x_i + y_k) P(X = x_i, Y = y_k) \\ &= \sum_i x_i \sum_k p_{ik} + \sum_k y_k \sum_i p_{ik} \\ &= \sum_i x_i P(X = x_i) + \sum_k y_k P(Y = y_k) = E(X) + E(Y) \qquad \square \end{aligned}$$

This theorem remains true for a sum with a finite number of terms (use induction!).

Example 21.2 Suppose that on n pieces of cards we wrote the integers $1, \ldots, n$, and then placed them in a hat. We choose m pieces of cards from the hat at random, with replacement. Let X denote the sum of the integers. Find $E(X)$.

The distribution of X in that problem is hard to find. Give it a try!

Denote by $X_1, \ldots, X_m$ the numbers selected. In view of the selection with replacement, each X_k is identically distributed, namely:

$$P(X_k = i) = \frac{1}{n} \qquad i = 1, \ldots, n$$

This means that for every k

$$E(X_k) = \sum_{i=1}^{n} i \cdot \frac{1}{n} = \frac{n(n+1)}{2} \cdot \frac{1}{n} = \frac{n+1}{2}.$$

On the other hand, clearly $X = X_1 + \ldots + X_m$, and therefore

$$E(X) = E(X_1) + \ldots + E(X_m) = m \cdot \frac{n+1}{2}$$

Thus $E(X)$ can be found without even knowing the distribution of X!

21.2 Mean of a product

If the discrete random variables X and Y, then

$$E(XY) = \sum_i \sum_k x_i y_k \cdot p_{ik}$$

where the range of X is $\{x_1, x_2, \ldots\}$, and the range of Y is $\{y_1, y_2, \ldots\}$ respectively, and p_{ik} denotes their joint distribution.

In a conpletely similar way, if X and Y continuously distributed, both have a mean, and their joint density function is f, then

$$E(XY) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy \cdot f(x, y)\, dx\, dy$$

Theorem 21.3 *If X and Y are independent, then*

$$E(XY) = E(X) \cdot E(Y)$$

Proof. We just focus on the continuous case, the discrete case can be treated similarly.

$$\begin{aligned} E(XY) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy \cdot f(x, y)\, dx\, dy = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy \cdot f_X(x) \cdot f_Y(y)\, dx\, dy \\ &= \int_{-\infty}^{\infty} x f_X(x)\, dx \cdot \int_{-\infty}^{\infty} y f_Y(y)\, dy = E(X) \cdot E(Y) \end{aligned}$$

since the independence implies that the joint density is the product of the marginal densities, i.e. $f(x, y) = f_X(x) \cdot f_Y(y)$. $\square$

21.3 Variance of a sum

Theorem 21.4 *Assume that X and Y are independent, and they both have a variance. Then*

$$Var(X+Y) = Var(X) + Var(Y)$$

The statement can be extended to any finite number of terms.

Proof. Exploit our theorem about the mean of the product, then we get:

$$\begin{aligned} Var(X+Y) &= E((X+Y-E(X+Y))^2) \\ &= E((X-E(X))^2) + E((Y-E(Y))^2) \\ &\quad +2E((X-E(X))(Y-E(Y))) \\ &= Var(X) + Var(Y) + 2(E(XY) - E(X)E(Y)) \\ &= Var(X) + Var(Y)\,. \qquad \square \end{aligned}$$

Example 21.5 Why do we think that by repeatedly performing an experiment and taking the average of the results we can expect a more accurate result?

Let us suppose that for determining an unknown quantity m we perform n observations, and the results are the random variables $X_1, \ldots, X_n$. We assume that the variables are independent and identically distributed with

$$E(X_k) = m, \quad D(X_k) = \sigma, \quad k = 1, 2, \ldots, n\,.$$

The assumption that all variables have the same distribution means that the observations (measurments) are carried out in the same circumstances. Then σ is interpreted as the expected error. Take the arithmetic average of our results, i.e. introduce the random variable

$$Y_n = \frac{X_1 + \ldots + X_n}{n}$$

Then clearly $E(Y_n) = m$, moreover, according to our theorem above

$$Var(Y_n) = Var\left(\frac{1}{n}(X_1 + \ldots + X_n)\right) = \frac{1}{n^2} n \cdot \sigma^2 = \frac{\sigma^2}{n}\,.$$

as a consequence of independence. Thus, for the standard deviation of Y_n we obtain:

$$D(Y_n) = \frac{\sigma}{\sqrt{n}}$$

for which $D(Y_n) \to 0$ as $n \to \infty$. Hence, the expected error tends to zero, when n approaches infinity.

21.4 Covariance and correlation

The following concepts are used for measuring the degree of dependence of random variables.

Definition 21.6 The *covariance* of random variables X and Y is defined by

$$Cov(X,Y) = E((X - E(X)) \cdot (Y - E(Y)))$$

and their *correlation coefficient* is given by

$$Corr(X,Y) = \frac{Cov(X,Y)}{D(X) \cdot D(Y)}$$

As it is easy to see

$$Cov(X,Y) = E(XY - E(X)Y - E(Y)X + E(X)E(Y)) = E(XY) - E(X)E(Y)\,,$$

and most of the time, this simpler expression is used to evaluate the covariance.

The covariance is NOT an absolute measurment of the in dependence, since for any $\alpha \neq 0$ we have

$$Cov(\alpha X, Y) = \alpha Cov(X,Y)$$

so it dependends on the dimensions . Just think of the case when X and Y are costs given in Euro, but if we convert them to Forint, then their covariance will change to approximately 340^2 times higher. However, the correlation coefficient is independent of the dimension, since for any real numbers $\alpha \neq 0$ and β we have:

$$Corr(\alpha X + \beta, \alpha Y + \beta) = Corr(X,Y)$$

which means that the correlation is independent of linear transformations. ATTENTION! Verify this equality directly by the definition!

Theorem 21.7

1. $-1 \leq Corr(X,Y) \leq 1$
2. *If X and Y are independent, then* $Cov(X,Y) = 0$

Proof. For proving the first statement, take a real number $t \in \mathbb{R}$ arbitrarily, and consider the random variable

$$W = [X - E(X) + t(Y - E(Y))]^2$$

Since W is nonnegative, so is its mean. This means that

$$E(W) = E((X - E(X))^2) + 2tCov(X,Y) + t^2 E((Y - E(Y))^2) \geq 0$$

for every real number t. This expression is quadratic with respect to t, and therefore it can only nonnegative, if its discriminant is nonpositive, that is:

$$4Cov(X,Y)^2 - 4E((X - E(X))^2)E((Y - E(Y))^2) \leq 0\,.$$

Rearranging the terms, and taking the square root of both sides, we get:

$$|Cov(X,Y)| \leq D(X)D(Y)$$

The second statement is an immediate consequence of Theorem 21.3. □

Example 21.8 ATTENTION! The example below shows that the converse of the second statement of our theorem is not true! Toss a coin twice in a row, and introduce the random variables:

$$X_k = \begin{cases} 0 & \text{if toss } k \text{ is a Head} \\ 1 & \text{if toss } k \text{ is a Tail} \end{cases}$$

($k = 1, 2$). Consider the variables $Y_1 = X_1 + X_2$ and $Y_2 = X_1 - X_2$. Then their joint distribution is:

$Y_2 \setminus Y_1$	0	1	2
-1	0	0.25	0
0	0.25	0	0.25
1	0	0.25	0

By examining the joint distribution, we see that Y_1 and Y_2 are not independent, but we can easily calculate that $Cov(Y_1, Y_2) = 0$

21.5 Theorem of Total Expectation

Consider the discrete random variables X and Y that have a joint distribution $P(X = x_i, Y = y_k) = p_{ik}$, and $P(Y = y_k) > 0$ for all indeces $i = 1, 2, \dots$ and $k = 1, 2 \dots$.

Definition 21.9 Create the conditional expected values of X under the conditions $Y = y_k$ that is:

$$m_k = E(X|Y = y_k) = \sum_{i=1} x_i P(X = x_i | Y = y_k)$$

for every $k = 1, 2 \dots$. This sequence is called the *conditional expectation* of X with respect to the variable Y. Its notation is $E(X|Y)$.

Observe that this way we have defined a random variable, namely

$$E(X|Y) = m_k\,, \quad ha \quad Y = y_k\,, k = 1, 2, \ldots$$

Below we determine the mean of this random variable. This result can be regarded as the generalization of Theorem of Total Probability.

Tétel 21.10 (Theorem of Total Expectation) $E(E(X|Y)) = E(X)$.

Proof. Indeed,

$$\begin{aligned} E(E(X|Y)) &= \sum_{k=1} m_k P(Y = y_k) = \sum_{k=1}\sum_{i=1} x_i P(X = x_i|Y = y_k)P(Y = y_k) \\ &= \sum_{i=1} x_i \sum_{k=1} P(X = x_i, Y = y_k) = \sum_{i=1} x_i P(X = x_i) = E(X) \end{aligned}$$

since, in the second line, we obtain precisely the marginal distribution of X. $\square$

ATTENTION! Why can we interchange the sums in the second line?

Example 21.11 In some situations it is easier to find $E(X)$ by our theorem than by the direct approach. The number of calls received by a call center on a given day has Poisson distribution with a parameter $\lambda > 0$. Every call is a wrong number with a given probability $p > 0$, independently from each other. Find the expected value of the wrong number calls on that day.

Let X denote the number of wrong calls, and Y the total number of calls. It is clear that for any fixed $n \in \mathbb{N}$ under the condition $Y = n$ we face the Bernoulli-experiment. Therefore,

$$P(X = k|Y = n) = \binom{n}{k} p^k (1-p)^{n-k} \quad \text{if} \quad n \geq k$$

while $P(X = k|Y = n) = 0$, if $n < k$. Hence, the conditional expected value is given by

$$m_n = E(X|Y = n) = np\,, \quad n = 1, 2, \ldots$$

Making use of the Theorem of Total Expectation, we obtain

$$E(X) = E(E(X|Y)) = \sum_{n=1}^{\infty} np\frac{\lambda^n}{n!}e^{-\lambda} = \lambda p$$

ATTENTION! Find $E(X)$ directly by using the distribution of X as well!

Recitation and Exercises

1. Reading: Textbook-2, Sections 4.1, 4.2 and 4.3.
2. Homework: Textbook-2, Exercises 4.23, 4.24, 4.52, 4.59, 4.60, 4.64, 4.70, 4.98.
3. Review: "Probability Exercises"

Chapter 22

Sums of random variables

22.1 Sums of discrete variables

Assume that X and Y are independent variables, and both have Poisson-distribution, with parameters $\lambda > 0$ and $\mu > 0$ respectively. Find the distribution of $X + Y$. Then for any fixed integer k

$$\begin{aligned}
P(X+Y=k) &= \sum_{i=0}^{k} P(X=i, Y=k-i) = \sum_{i=0}^{k} P(X=i) \cdot P(Y=k-i) \\
&= \sum_{i=0}^{k} \frac{\lambda^i}{i!} e^{-\lambda} \cdot \frac{\mu^{k-i}}{(k-i)!} e^{-\mu} = \frac{e^{-(\lambda+\mu)}}{k!} \sum_{i=0}^{k} \binom{k}{i} \lambda^i \mu^{k-i} \\
&= \frac{(\lambda+\mu)^k}{k!} e^{-(\lambda+\mu)}
\end{aligned}$$

by the independence. Thus, $X + Y$ has Poisson-distribution with the parameter $\lambda + \mu$.

Using induction, this result can be extended to any finite number of terms.

Tétel 22.1 *Assume that $X_1, \ldots, X_n$ are independent variables, and have Poisson-distribution with parameters $\lambda_1, \ldots, \lambda_n$ respectively. Then the random variable*

$$Y_n = X_1 + \ldots + X_n$$

has Poisson-distribution with parameter $\lambda_1 + \ldots + \lambda_n$.

22.2 Sums of continuous variables

Let X and Y be independent, continuously distributed random variables with density functions f and g respectively. Denote by F and G their cumulative distribution functions. Let H denote the cumulative distribution function of $X+Y$. To find H pick a real number $x \in \mathbb{R}$. Then (sketch a picture!):

$$\begin{aligned} H(x) &= \int\int_{t+s<x} f(s)g(t)\,ds\,ds = \int_{-\infty}^{\infty}\int_{-\infty}^{x-s} f(s)g(t)\,dt\,ds \\ &= \int_{-\infty}^{\infty} f(s)\left(\int_{-\infty}^{x-s} g(t)\,dt\right)\,ds = \int_{-\infty}^{\infty} f(s)G(x-s)\,ds\,. \end{aligned}$$

By taking the derivative of H, we get the density function h of $X+Y$

$$h(x) = \int_{-\infty}^{\infty} f(s)g(x-s)\,ds$$

This formula is called the *convolution integral* of f and g.

ATTENTION! Differentiating the integral is not straightforward! Examine this rule in some simple cases!

Example 22.2 Suppose now that X and Y are independent random variables that are uniformly distributed on the interval $[0,1]$. Then (X,Y) is uniformly distributed on the unit square of the plane. By sketching a picture, show that if h stands for the density function of $X+Y$, then

$$h(x) = \begin{cases} x & \text{if } 0 < x < 1 \\ 2-x & \text{ha } 1 < x < 2 \\ 0 & \text{elsewhere.} \end{cases}$$

Example 22.3 Let X and Y be independent, exponentially distributed random variables, both with parameter $\lambda > 0$. Let h denote the density function of $X+Y$. If f denotes the density function of the exponential distribution with parameter λ, then the convolution integral is:

$$h(x) = \int_{-\infty}^{\infty} f(s)f(x-s)\,ds$$

Behind the integral sign f is zero on the negative part of the real line. Therefore, the integrand is not zero if and only if $s > 0$ and $x - s > 0$, that is $0 < s < x$. Thus,

$$h(x) = \int_0^x \lambda^2 e^{-\lambda s} e^{-\lambda(x-s)}\,ds = \lambda^2 \int_0^x e^{-\lambda x}\,ds = \lambda^2 x e^{-\lambda x}$$

for any given $x > 0$, since the last integrand does not depend on s.

By using induction, we can extend the above result to any finite number of terms.

Theorem 22.4 *Assume that $X_1, \ldots, X_n$ are independent, exponentially distributed random variables with the same parameter $\lambda > 0$. Let h_n denote the* *density function of the random variable*

$$Y_n = X_1 + \ldots + X_n$$

Then

$$h_n(x) = \frac{\lambda^n}{(n-1)!} x^{n-1} e^{-\lambda x}$$

if $x > 0$, *and* $h_n(x) = 0$, *if* $x \leq 0$.

22.3 The Poisson process

In this section we describe a deeper relationship between the exponential and the Poisson distributions.

Consider the random variables $T_1, T_2, \ldots$ which mean waiting times between consecutive "occurances".

We can think of times between successive vehicles on a highway, times between incoming claims received by an insurance company, waiting times between consecutive clients at a customer service desk, time intervals between incoming calls to a call center, etc.

Assume that $T_1, T_2, \ldots$ independent, exponentially distributed random variables with identical parameter $\lambda > 0$. The smaller the value of λ, the longer are the expected waiting times (check the expectation!). The memoryless property of the exponential distribution means that the waiting time is independent on how long we have been waiting before.

Set $S_0 = 0$ denote by

$$S_n = T_1 + \ldots + T_n$$

the total waiting time until the n-th occurance. For a given $t > 0$ the event

$$\{S_n \leq t\}$$

means that the n-th occurance arrives before t. This means that the number of occurances in the time interval $[0, t]$ is at least n.

Denote by $N(t)$ the number of occurances in the time interval $[0, t]$, then the events

$$\{N(t) \geq n\} = \{S_n \leq t\}$$

coincide. For every $t > 0$ we defined a random variable $N(t)$, this correspondence is called the *Poisson process*.

How can we find the distribution of $N(t)$ for a fixed $t > 0$? The event that there are exactly n occurances in the time interval $[0, t]$ is given by

$$\{N(t) = n\} = \{S_n \leq t\} \cap \overline{\{S_{n+1} \leq t\}} = \{S_n \leq t < S_{n+1}\}.$$

Clearly $\{S_{n+1} \leq t\} \subset \{S_n \leq t\}$, and this implies

$$P(N(t) = n) = P(S_n \leq t) - P(S_{n+1} \leq t) \,.$$

Let h_n be the density of S_n, and h_{n+1} be the density of S_{n+1}. Since $T_1, T_2, \ldots$ are independent, exponentially distributed random variables with the same parameter λ, then in view of Theorem 22.4 of the previous section we get

$$h_n(x) = \frac{\lambda^n}{(n-1)!} x^{n-1} e^{-\lambda x} \quad \text{and} \quad h_{n+1}(x) = \frac{\lambda^{n+1}}{n!} x^n e^{-\lambda x}$$

for every $x > 0$. Therefore

$$P(N(t) = n) = P(S_n \leq t) - P(S_{n+1} \leq t) = \int_0^t h_n(x)\,dx - \int_0^t h_{n+1}(x)\,dx \,.$$

Evaluate the first integral on the righ-hand side by integration by parts:

$$\begin{aligned}
\int_0^t h_n(x)\,dx &= \frac{\lambda^n}{(n-1)!} \int_0^t x^{n-1} e^{-\lambda x}\,dx \\
&= \frac{\lambda^n}{(n-1)!} \left[\frac{x^n}{n} e^{-\lambda x} \right]_0^t + \frac{\lambda^n}{(n-1)!} \int_0^t \frac{x^n}{n} \lambda e^{-\lambda x}\,dx \\
&= \frac{(\lambda t)^n}{n!} e^{-\lambda t} + \frac{\lambda^{n+1}}{n!} \int_0^t x^n e^{-\lambda x}\,dx \,.
\end{aligned}$$

We can recognize that in the last integral we pecisely have h_{n+1}. Hence,

$$P(N(t) = n) = \frac{(\lambda t)^n}{n!} e^{-\lambda t}$$

Theorem 22.5 *In the Poisson process the number of occurances in the time interval $[0, t]$ is a Poisson random variable with parameter λt.*

22.4 Sum of standard normal distributions

Let Z_1 and Z_2 be independent, standard normally distributed random variables, and find the distribution of their sum:

$$Y = Z_1 + Z_2$$

Now, the convolution integral is

$$h(x) = \int_{-\infty}^{\infty} \varphi(s)\varphi(x - s)\,ds$$

where h is the density function of Y. Then

$$\begin{aligned} h(x) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-s^2/2}e^{-(x-s)^2/2}\,ds = \frac{1}{2\pi}e^{-\frac{x^2}{2}}\int_{-\infty}^{\infty} e^{xs-s^2}\,ds \\ &= \frac{1}{2\pi}e^{-\frac{x^2}{2}}\int_{-\infty}^{\infty} e^{-(s-x/2)^2}e^{x^2/4}\,ds = \frac{1}{2\pi}e^{-\frac{x^2}{4}}\int_{-\infty}^{\infty} e^{-(s-x/2)^2}\,ds \end{aligned}$$

The last integral is precisely the Gauss integral, whose value is $\sqrt{\pi}$, thus

$$h(x) = \frac{1}{2\sqrt{\pi}}e^{-\frac{x^2}{4}} \qquad -\infty < x < \infty$$

This is exactly the density function of the normal distribution with parameters $m = 0$ and $\sigma = \sqrt{2}$.

Using completely analogous arguments, we can formulate the following result.

Theorem 22.6 *Let $Z_1, \ldots, Z_n$ be independent, standard normally distributed random variables. Then $Y = Z_1 + \ldots + Z_n$ is a normally distributed random variable with parameters $m = 0$ and $\sigma = \sqrt{n}$.*

22.5 Central Limit Theorem

Imagine the following experiment. To determine an unknown quantity m we carry out n independent observations (measurments). To approximate the unknown quantity we use the arithmetic mean (average) of the n outcomes.

Let us denote the outcomes by $X_1, \ldots, X_n$ and assume that they are independent and identically distributed random variables with

$$E(X_k) = m, \qquad D(X_k) = \sigma, \qquad k = 1, 2, \ldots, n$$

(Identical distribution means that the observations are performed in identical circumstances.) For the stantardized average let us introduce the following notation:

$$Y_n = \frac{\frac{1}{n}(X_1 + \ldots + X_n) - m}{\sigma/\sqrt{n}}$$

Then Y_n has a mean of 0 and standard deviation 1.

It was the amazing discovery of the Russian mathematician Alexandr Lyapunov and the mathematics of his time (early 20-th century) that the distribution of this variable Y_n converges to the standard normal distribution.

Tétel 22.7 (Central Limit Theorem) *Under the above conditions let F_n denote the cumulative distribution function of Y_n. Then for every $x \in \mathbb{R}$ we have*

$$\lim_{n\to\infty} F_n(x) = \Phi(x)\,.$$

Example 22.8 On a given day the number of visitors to a local convenience store is 100. Every visitor buys something with probability $p = 0.2$ (independently from each other). Find the probability that on that given day the the number of purchases will be between 15 and 25.

Let X be the number of purchases. Then X is binomially distributed (Bernoulli-experiment!) with parameters $n = 100$ and $p = 0.2$. For each visitor introduce the following notation:

$$X_k = \begin{cases} 0 & \text{if does not buy anything} \\ 1 & \text{if buys something} \end{cases}$$

then $X = X_1 + \ldots + X_{100}$ and the terms are independent random variables. It is easy to see that for each k we have $E(X_k) = 0.2$ and $Var(X_k) = 0.16$, hence $D(X_k) = 0.4$. Therefore,

$$\begin{aligned} P(15 < X < 25) &= P\left(-\frac{5}{4} < \frac{X-20}{4} < \frac{5}{4}\right) \\ &= P\left(-\frac{5}{4} < \frac{\frac{1}{100}(X_1 + \ldots + X_{100}) - 0.2}{0.4/10} < \frac{5}{4}\right) \end{aligned}$$

Making use of the Central Limit Theorem

$$\begin{aligned} P\left(15 < X < 25\right) &\approx \Phi(1.25) - \Phi(-1.25) \\ &= 2\Phi(1.25) - 1 = 0.7888 \end{aligned}$$

by looking up the number in the table for the standard normal distribution, see Textbook-2, page 736 (Appendix A).

Recitation and Exercises

1. Reading: Textbook-2, Sections 6.5 and 6.6.
2. Homework: Textbook-2, Exercises 6.24, 6.26, 6.29, 6.34 and 6.38.
3. Review: "Probability Exercises"

Chapter 23

Law of Large Numbers

23.1 Chebyshev's Theorem

So far we have had to determine probabilities of the form

$$P(a < X < b)$$

This is easy to do if the distribution of the random variable X is known. In particular, in the case of a discrete variable we get

$$P(a < X < b) = \sum_{a<x_k<b} P(X = x_k)$$

while for a continuously distributed variable

$$P(a < X < b) = \int_a^b f(x)\,dx$$

where f is the density function of X. However, there are situations when this procedure cannot be completed. Namely, if

1. either the distribution of X is not known,
2. or the distribution of X is known, but too complicated to use.

In cases like these, we can be satisfied with an appropriate estimate on the given probability. This estimate is provided by Chebyshev's Theorem. Consider a random variable X that has a mean and a variance.

Theorem 23.1 (Chebyshev's Theorem) *The mean of X is $E(X) = m$ and its standard deviation is $D(X) = \sigma$. Then*

$$P(|X - m| < k \cdot \sigma) \geq 1 - \frac{1}{k^2}$$

for any $k > 0$.

Proof. We present the proof for a continuously distributed random variable. In the discrete case the proof can be carried out in a completely analogous way. Let f be the density function of X, then

$$\sigma^2 = \int_{-\infty}^{\infty} (x-m)^2 f(x)\, dx$$

If $k > 0$ is given, then the value of the integral on the right-hand side will not increase if we skip the interval $[m - k\sigma, m + k\sigma]$. In fact:

$$\sigma^2 \geq \int_{-\infty}^{m-k\sigma} (x-m)^2 f(x)\, dx + \int_{m+k\sigma}^{\infty} (x-m)^2 f(x)\, dx \tag{23.1}$$

since the integrand is nonnegative. On the other hand, at every point x of the interval $(-\infty, m - k\sigma]$ we have $(x-m)^2 \geq k^2\sigma^2$, and hence

$$\int_{-\infty}^{m-k\sigma} (x-m)^2 f(x)\, dx \geq \int_{-\infty}^{m-k\sigma} k^2\sigma^2 f(x)\, dx \geq k^2\sigma^2 P(X \leq m - k\sigma)\,.$$

Completely similarly, at every point x of the interval $[m + k\sigma, \infty)$ we get $(x - m)^2 \geq k^2\sigma^2$, and consequently

$$\int_{m+k\sigma}^{\infty} (x-m)^2 f(x)\, dx \geq \int_{m+k\sigma}^{\infty} k^2\sigma^2 f(x)\, dx \geq k^2\sigma^2 P(X \geq m + k\sigma)\,.$$

If we combine the latter two inequalities with the inequality (23.1), then we obtain

$$\sigma^2 \geq k^2\sigma^2 P(X \leq m - k\sigma) + k^2\sigma^2 P(X \geq m + k\sigma)\,.$$

Dividing both sides with the positive expression $k^2\sigma^2$ we get

$$\frac{1}{k^2} \geq P(X \leq m - k\sigma) + P(X \geq m + k\sigma) = P(|X - m| \geq k\sigma)\,.$$

By converting to the complement event, the proof is completed. □

Note that the theorem gives an irrelevant result if $k \leq 1$, so we apply the inequality only for $k > 1$.

Example 23.2 For instance, if the distribution of the random variable X is not known, but its mean $E(X) = 8$ and its standard deviation $D(X) = 2$ are given, then

$$P(2 < X < 14) \geq 1 - \frac{1}{9} \approx 0.8889$$

since in this case $k = 3$.

23.2 Chebyshev's Theorem in equivalent form

Sometimes it is more covenient to use Csebishev's Theorem in the following form:

$$P(|X - E(X)| < \varepsilon) \geq 1 - \frac{Var(X)}{\varepsilon^2}$$

where $\varepsilon > 0$. Indeed, this inequality is equivalent to our theorem by setting $k \cdot D(X) = \varepsilon > 0$, and then

$$\frac{1}{k^2} = \frac{Var(X)}{\varepsilon^2}$$

Let us formulate the theorem in the following equivalent form.

Theorem 23.3 *Consider a random variable X with a mean $E(X) = m$, and standard deviation $D(X) = \sigma$. Then for every fixed $\varepsilon > 0$ we have*

$$P(|X - m| < \varepsilon) \geq 1 - \frac{\sigma^2}{\varepsilon^2} \tag{23.2}$$

Example 23.4 On a given day a call center receives 2000 incoming calls. Every call is a wrong number with probability 0.002 (independently from each other). Find the probability that on that given day there are at most 8 wrong number calls.

Let X denote the number of wrong number calls. Clearly X is binomially distributed (Bernoulli experiment!), with parameters $n = 2000$ and $p = 0.002$. The solution to our problem is:

$$P(X \leq 8) = \sum_{k=0}^{8} \binom{2000}{k} 0.002^k \cdot 0.998^{2000-k}$$

which is not easy to evaluate (although the distribution is known).

However, we can give a reasonable estimate by using Chebyshev's Theorem. Now $m = 4$ and $\sigma^2 = 4 \cdot 0.998 \approx 4$, and therefore

$$P(X \leq 8) = P(|X - 4| < 5) \geq 1 - \frac{4}{25} = 0.84$$

23.3 Poisson approximation

Example 23.5 In a large hospital with 2000 beds, the probability that a patient needs intensive care is 0.002 on any given day (independently from each other). The director wants to establish a new emergency ward so that if a patient needs intensive care, must get a bed with probability of at least 0.99. What should be the size of the emergency ward with minimal cost (smallest number of beds)?

Let N denote the number of beds in the emergency ward, and X be the number of patients who need intensive care on a given day. Then X is clearly binomially distributed (Bernoulli experiment!) with a mean of $m = 4$ and variance $\sigma^2 = 4 \cdot 0.998 \approx 4$. Then the inequality

$$P(X \leq N) = \sum_{k=0}^{N} \binom{2000}{k} 0.002^k 0.998^{2000-k} \geq 0.99$$

has to be solved for the smallest N (which means the lowest cost).

This is the situation when the distribution of X is known, but too complicated to use. Apply Chebyshev's Theorem instead:

$$P(|X-4| < \varepsilon) \geq 1 - \frac{4}{\varepsilon^2} = 0.99$$

The lowest solution is $\varepsilon = 20$ and therefore $N = 23$ is obtained for the optimal smallest number of beds in the new emergency ward.

Chebyshev's Theorem is true for any distribution, so we cannot expect a very sharp estimate. We can get a much more accurate solution if we apply the Poisson approximation. The theorem on how to approximate the binomial distribution by the Poisson distribution is discussed in Section 18.5. In particular, in the present example:

$$\sum_{k=0}^{N} \binom{2000}{k} 0.002^k 0.998^{2000-k} \approx \sum_{k=0}^{N} \frac{4^k}{k!} e^{-4}$$

since "$n = 2000$ is large enough, and $p = 0.002$ sufficiently small", moreover $np = 4$. When we look at the Poisson tables (see Textbook-2, page 732, Appendix A) we can see that the sum on the right-hand side exceeds 0.99 at $N = 9$. Based on this approximation we claim that even an emergency ward of size $N = 9$ fulfills the criteria. (Examining how sharp this approximation is, goes beyond the scope of this book.)

23.4 Law of Large Numbers

We carry out an experiment n times in a row (independently from each other) and each time we observe whether or not a given event A occurs (Bernoulli experiment).

Suppose that the probability of the event A is $P(A) = p$ (where $0 \leq p \leq 1$) and let X_n be the number of experiments in which A occurs. The quotient X_n/n means the relative frequency of the event A.

We want to examine whether the relative frequency converges to the real value of the probability when the number of experiments is increased that is $n \to \infty$?

From theoretical point of view, this question is of fundamental importance. If the answer is affirmative, it justifies our axiomatic approach to probability. Indeed, within the framework of our theory that we have developed from the axioms, we are able to derive a theorem that can directly be verified in reality. In other words, our axioms are set properly, and their consequences reflect real phenomena.

As is well known, X_n is binomially distributed with parameters n and p paraméterekkel. Pick a number $\varepsilon > 0$ and apply Chebyshev's Theorem:

$$P\left(\left|\frac{X_n}{n} - p\right| \geq \varepsilon\right) = P(|X_n - np| \geq n\varepsilon)$$

Since $E(X_n) = np$ and $Var(X_n) = np(1-p)$, we get

$$P(|X_n - np| \geq n\varepsilon) \leq \frac{np(1-p)}{n^2\varepsilon^2}$$

We have $p(1-p) \leq 1/4$ for any real number p, so from here

$$P\left(\left|\frac{X_n}{n} - p\right| \geq \varepsilon\right) \leq \frac{1}{4n\varepsilon^2} \to 0$$

if $n \to \infty$. We formulate this result in the theorem below.

Theorem 23.6 (Bernoulli's Law of Large Numbers)

$$\lim_{n\to\infty} P\left(\left|\frac{X_n}{n} - p\right| < \varepsilon\right) = 1$$

for every $\varepsilon > 0$.

This theorem is sometimes called "Bernoulli's Weak Law of Large Numbers" to distinguish it from more advanced and complicated "Strong Law" results.

Example 23.7 A consulting agency makes a forecast of the support of a political party before the upcoming parliamentary election. They interview potential voters about their preferences. The agency wants to be 90% sure that their prediction should be within the 1% margin (i.e. the difference between the predicted ratio and real ratio is less than 1%). How many people have to be interviewed?

Let $0 < p < 1$ denote the unknown real ratio (the real support of the party), this will be estimated by the relative frequency. Assume that the size of sample (number of interviews) is n (yet to be determined) and X_n is the number of voters who support the party. Then the anticipated support ratio is X_n/n.

This is a Bernoulli experiment, therefore X_n is binomially distributed with $E(X_n) = np$ and $Var(X_n) = np(1-p)$. Then the following inequality holds:

$$P\left(\left|\frac{X_n}{n} - p\right| \leq 0.01\right) \geq 1 - \frac{1}{4n \cdot 10^{-4}}$$

If the agency wants to guarantee this accuracy with at least 90% certainty, then

$$1 - \frac{1}{4n \cdot 10^{-4}} = 0.90$$

from which we have $n = 25\,000$.

In reality, using advanced statistical methods, even a smaller sample might be sufficient. However, in most situations it is hard to guarantee that the set of interviewed voters is homogeneous and representative (in the sense that the sample ratio reflects the ratio for the whole voting society).

Under the conditions of Theorem 23.6 the following stronger statement can also be proven.

Theorem 23.8 *Under the conditions of Theorem 23.6 we have*

$$P\left(\lim_{n\to\infty} \frac{X_n}{n} = p\right) = 1$$

Intuitively, Theorem 23.6 claims that very likely the relative frequency gets close to the probability p as n increases. However, it does not exclude that large differences can occur beyond any arbitrarily large index n. It just says that such large differences are unlikely. Theorem 23.8 tells us however, that such large differences come with probability zero. (The proof is due to Lyapunov and to Kolmogorov in a more general form in the 30's of the last century.)

Recitation and Exercises

1. Reading: Textbook-2, Section 4.4.
2. Homework: Textbook-2, Exercises 4.75, 4.76, 4.77, 4.78 and 4.91.
3. Review: "Probability Exercises"