Topics in analysis

Tomas Persson

Preface

These are the lecture notes of the course *Topics in analysis*, which I give in Lund during the second half of the spring semester, 2019. The course consists of several somewhat basic topics in analysis that are usually not part of any of the basic courses at LTH, but which, in my opinion, are still interesting, usefull and partially part of what could be called general knowledge. It is my hope that the students will find these topics entertaining, interesting and to some use.

Tomas Persson, Lund, May 22, 2019

Typset in Antykwa Toruńska with LaTeX 2ε

CHAPTER 1

Continued fractions and Diophantine approximation

1. Continued fractions

A *continued fraction* is an expression of the form

$$b_0 + \cfrac{a_1}{b_1 + \cfrac{a_2}{b_2 + \cfrac{a_3}{b_3 + \cfrac{1}{\ddots + \cfrac{a_n}{b_n}}}}}$$

or

$$b_0 + \cfrac{a_1}{b_1 + \cfrac{a_2}{b_2 + \cfrac{a_3}{b_3 + \cfrac{1}{\ddots}}}} = \lim_{n\to\infty} b_0 + \cfrac{a_1}{b_1 + \cfrac{a_2}{b_2 + \cfrac{a_3}{b_3 + \cfrac{1}{\ddots + \cfrac{a_n}{b_n}}}}},$$

provided that the limit exists. We will study the case when $a_k = 1$ for all k and b_k is a natural number for all k. We will use the notation

$$[b_0; b_1, b_2, \ldots, b_n] = b_0 + \cfrac{1}{b_1 + \cfrac{1}{b_2 + \cfrac{1}{b_3 + \cfrac{1}{\ddots + \cfrac{1}{b_n}}}}}$$

and

$$[b_0; b_1, b_2, \ldots] = \lim_{n\to\infty} [b_0; b_1, b_2, \ldots, b_n].$$

For obvious reasons, $[b_0; b_1, b_2, \ldots, b_n]$ is called a finite continued fraction, and $[b_0; b_1, b_2, \ldots]$ is called an infinite continued fraction. The natural numbers $b_1, b_2, \ldots$ are called the digits of the continued fraction.

It is clear that when b_k are natural numbers, then $[b_0; b_1, b_2, \ldots, b_n]$ is a rational number and we write $\frac{A_n}{B_n} = [b_0; b_1, b_2, \ldots, b_n]$. The rational number $\frac{A_n}{B_n}$ is called the *n-th convergent* of the continued fraction $[b_0; b_1, b_2, \ldots]$.

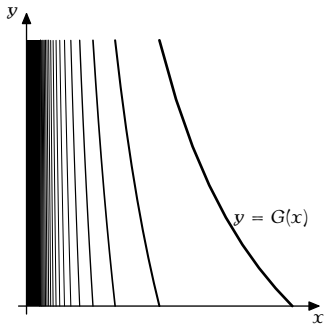
We will show that every real number x can be written as a continued fraction, $x = [b_0, b_1, b_2, \ldots]$. This representation is unique, unless x is rational, in which case x can be written as a continued fraction in two different ways. For instance, we have

$$\frac{3}{7} = \frac{1}{2 + \dfrac{1}{3}} = \frac{1}{2 + \dfrac{1}{2 + \dfrac{1}{1}}}. \tag{1.1}$$

To study continued fractions, we will make use of the Gauß transformation[1] $G\colon (0,1] \to [0,1)$, defined by

$$G(x) = \frac{1}{x} \mod 1 = \frac{1}{x} - \left[\frac{1}{x}\right],$$

where $[\cdot]$ denotes the integer part.



Suppose that $x = [0; b_1, b_2, \ldots, b_n]$, where $n \geq 2$. Then

$$x = \cfrac{1}{b_1 + \cfrac{1}{b_2 + \cfrac{1}{b_3 + \cfrac{1}{\ddots + \cfrac{1}{b_n}}}}} \leq \frac{1}{b_1}$$

and

$$x = \cfrac{1}{b_1 + \cfrac{1}{b_2 + \cfrac{1}{b_3 + \cfrac{1}{\ddots + \cfrac{1}{b_n}}}}} \geq \frac{1}{b_1 + \dfrac{1}{b_2}}.$$

[1] Johann Carl Friedrich Gauß, 1777–1855. German mathematician.

Hence, since b_1 and b_2 are natural numbers, $x \in \left(\frac{1}{b_1+1}, \frac{1}{b_1}\right]$, and $G(x)$ is defined. We have that

$$\frac{1}{x} = b_1 + \cfrac{1}{b_2 + \cfrac{1}{b_3 + \cfrac{1}{\ddots + \cfrac{1}{b_n}}}},$$

from which it is apparent that

$$G(x) = \cfrac{1}{b_2 + \cfrac{1}{b_3 + \cfrac{1}{\ddots + \cfrac{1}{b_n}}}} = [0; b_2, \ldots, b_n].$$

Clearly, $G(\frac{1}{b_n}) = 0$.

Thus, $x = [0; b_1, b_2, \ldots, b_n]$ is a point such that for each k,

$$G^{k-1}(x) \in \left(\frac{1}{b_k + 1}, \frac{1}{b_k}\right], \qquad k = 1, 2, \ldots, n \tag{1.2}$$

$$G^n(x) = 0, \tag{1.3}$$

where G^k denotes the k-fold composition of G with itself. The equations (1.2) and (1.3) define x uniquely, since the restriction of G to any interval $\left(\frac{1}{b_k+1}, \frac{1}{b_k}\right]$ is a one-to-one mapping $\left(\frac{1}{b_k+1}, \frac{1}{b_k}\right] \to [0, 1)$.

The restriction of G to the interval $\left(\frac{1}{b_k+1}, \frac{1}{b_k}\right]$ has an inverse given by

$$S_{b_k} \colon [0, 1) \to \left(\frac{1}{b_k + 1}, \frac{1}{b_k}\right]; \quad S_{b_k} \colon x \mapsto \frac{1}{x + b_k}.$$

The observations (1.2) and (1.3) above can thus be formulated using these mappings, and we conclude that

$$[0; b_1, b_2, \ldots, b_n] = S_{b_1} \circ S_{b_2} \circ \ldots \circ S_{b_n}(0).$$

We shall now use the knowledge aquired to study infinite continued fractions. The functions S_b are strictly decreasing. Hence, $S_{b_1} \circ S_{b_2} \circ \ldots \circ S_{b_n}$ is strictly increasing if n is even and strictly decreasing if n is odd.

Suppose that n is even. Then, since $S_{b_{n+1}}(0) > 0$ and $S_{b_{n+1}} \circ S_{b_{n+2}}(0) > 0$, we have

$$\begin{aligned}[0; b_1, b_2, \ldots, b_n] &= S_{b_1} \circ S_{b_2} \circ \ldots \circ S_{b_n}(0) \\ &< S_{b_1} \circ S_{b_2} \circ \ldots \circ S_{b_n} \circ S_{b_{n+1}}(0) = [0; b_1, b_2, \ldots, b_n, b_{n+1}],\end{aligned}$$

and

$$\begin{aligned}[0; b_1, b_2, \ldots, b_n] &= S_{b_1} \circ S_{b_2} \circ \ldots \circ S_{b_n}(0) \\ &< S_{b_1} \circ S_{b_2} \circ \ldots \circ S_{b_n} \circ S_{b_{n+1}} \circ S_{b_{n+2}}(0) \\ &= [0; b_1, b_2, \ldots, b_n, b_{n+1}, b_{n+2}].\end{aligned}$$

If n is instead odd, then we get the opposite inequalities,

$$[0; b_1, b_2, \ldots, b_n] > [0; b_1, b_2, \ldots, b_n, b_{n+1}],$$
$$[0; b_1, b_2, \ldots, b_n] > [0; b_1, b_2, \ldots, b_n, b_{n+1}, b_{n+2}].$$

This shows that the sequence of convergents $[0; b_1, b_2, \ldots, b_n]$ to the infinite continued fraction $[0; b_1, b_2, \ldots]$ is an oscillating sequence, and that

$$[0; b_1, b_2, \ldots, b_n] < [0; b_1, b_2, \ldots, b_n, b_{n+1}, b_{n+2}], \quad \text{when } n \text{ is even},$$
$$[0; b_1, b_2, \ldots, b_n] > [0; b_1, b_2, \ldots, b_n, b_{n+1}, b_{n+2}], \quad \text{when } n \text{ is odd}.$$

Thus, the convergents $[0; b_1, b_2, \ldots, b_n]$ for even n form a strictly increasing sequence of numbers that are bounded from above by the convergents for odd n, and the convergents for odd n form a strictly decreasing sequene of numbers that are bounded from below by the convergents for even n. Therefore, the limits

$$\lim_{n\to\infty} [0; b_1, b_2, \ldots, b_{2n}] \qquad \text{and} \qquad \lim_{n\to\infty} [0; b_1, b_2, \ldots, b_{2n+1}]$$

exist.

To prove that the limits are equal, we will use that the mappings S_b contract distances. We have that

$$|S_b'(x)| = \frac{1}{(x+b)^2} \leq \frac{1}{b^2}.$$

Hence, $|S_b'(x)|$ is only guaranteed to be less than one if $b > 1$. However, the derivative of $S_b \circ S_c$ is given by

$$(S_b \circ S_c)'(x) = \frac{1}{(bx + bc + 1)^2},$$

which is always at most $\frac{1}{4}$.

Using the information about the derivatives, we can now show that the limits are equal. Consider

$$[0; b_1, b_2, \ldots, b_{2n+1}] \qquad \text{and} \qquad [0; b_1, b_2, \ldots, b_{2n}].$$

We have for some $\xi \in (0, 1)$ that

$$\begin{aligned} 0 \leq [0; b_1, b_2, \ldots, b_{2n+1}] &- [0; b_1, b_2, \ldots, b_{2n}] \\ &= S_{b_1} \circ \ldots \circ S_{b_{2n+1}}(0) - S_{b_1} \circ \ldots \circ S_{b_{2n}}(0) \\ &= (S_{b_1} \circ \ldots \circ S_{b_{2n}})'(\xi)(S_{b_{2n+1}}(0) - 0) \\ &\leq \frac{1}{4^n} |S_{b_{2n+1}}(0) - 0| \leq \frac{1}{4^n}. \end{aligned}$$

This shows that the limits are equal and hence we may define

$$[0; b_1, b_2, \ldots] = \lim_{n\to\infty} \frac{A_n}{B_n} = \lim_{n\to\infty} [0; b_1, b_2, \ldots, b_n].$$

If x is a real number, then we can write x in a unique way as a sum of an integer and a number in $[0, 1)$. Thus, we may write $x = [x] + \{x\}$, where $[x]$ is an integer called the *integer part* of x and $\{x\}$ is a number in $[0, 1)$ which is called the *fractional part* of x. With this notation, we may formulate our finding above as the following theorem.

THEOREM 1.1. *Let x be an irrational number. Then x can be written as an infinite continued fraction expansion*

$$x = b_0 + \cfrac{1}{b_1 + \cfrac{1}{b_2 + \cfrac{1}{b_3 + \cfrac{1}{\ddots}}}},$$

where the continued fraction digits b_k are given by the relations

$$b_0 = [x], \qquad G^k(\{x\}) = \frac{1}{G^{k-1}(\{x\})} - b_k, \qquad k = 1, 2, \ldots$$

Similarly, if x is a rational number, then x can be written as a finite continued fraction

$$x = b_0 + \cfrac{1}{b_1 + \cfrac{1}{b_2 + \cfrac{1}{b_3 + \cfrac{1}{\ddots + \cfrac{1}{b_n}}}}},$$

where the continued fraction digits b_k are given by the relations

$$b_0 = [x], \qquad G^k(\{x\}) = \frac{1}{G^{k-1}(\{x\})} - b_k, \qquad k = 1, 2, \ldots, n$$

and $G^n(\{x\}) = \frac{1}{G^{n-1}(\{x\})} - b_n = 0$.

The numbers A_n and B_n can be calculated recursively, according to the following lemma.

LEMMA 1.2 (The Euler[2]–Wallis[3] relations). *Let*

$$A_{-1} = 1, \qquad A_0 = b_0,$$
$$B_{-1} = 0, \qquad B_0 = 1,$$

and define the sequences A_n and B_n recursively by

$$A_n = b_n A_{n-1} + A_{n-2}, \qquad B_n = b_n B_{n-1} + B_{n-2}.$$

Then, for any $n \geq 0$, the rational number $\frac{A_n}{B_n}$ is the n-th convergent of the continued fraction $[b_0; b_1, b_2, \ldots]$.

PROOF. A Möbius[4] transformation is a transformation of the form

$$z \mapsto \frac{az + b}{cz + d},$$

which we can represent by the matrix $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$.

[2]Leonhard Euler, 1707–1783. Swiss mathematician.

[3]John Wallis, 1616–1703. English mathematician.

[4]August Ferdinand Möbius, 1790–1868. German mathematician.

If T_1 and T_2 are two Möbius transformations represented by the two matrices

$$M_1 = \begin{bmatrix} a_1 & b_1 \\ c_1 & d_1 \end{bmatrix} \qquad \text{and} \qquad M_2 = \begin{bmatrix} a_2 & b_2 \\ c_2 & d_2 \end{bmatrix},$$

then the composition $T_1 \circ T_2$ is represented by the matrix $M_1 M_2$ (Exercise 1.1).

In particular, the composition $S_{b_1} \circ S_{b_2} \circ \ldots \circ S_{b_n}$ is represented by the product

$$\begin{bmatrix} 0 & 1 \\ 1 & b_1 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 1 & b_2 \end{bmatrix} \cdots \begin{bmatrix} 0 & 1 \\ 1 & b_n \end{bmatrix}.$$

We let $A_{-1} = 1$, $A_0 = b_0$, $B_{-1} = 0$, $B_0 = 1$. Since $[0; b_1, b_2, \ldots, b_n] = S_{b_1} \circ S_{b_2} \circ \ldots \circ S_{b_n}(0)$, we have $[b_0; b_1, b_2, \ldots, b_n] = T_{b_0} \circ S_{b_1} \circ S_{b_2} \circ \ldots \circ S_{b_n}(0)$ where $T_{b_0}(z) = z + b_0$. Therefore,

$$\begin{bmatrix} 1 & b_0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 1 & b_1 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 1 & b_2 \end{bmatrix} \cdots \begin{bmatrix} 0 & 1 \\ 1 & b_n \end{bmatrix} = \begin{bmatrix} C_n & A_n \\ D_n & B_n \end{bmatrix}, \tag{1.4}$$

for some C_n and D_n, and with $\frac{A_n}{B_n} = [b_0; b_1, b_2, \ldots, b_n]$.

Clearly,

$$\begin{bmatrix} C_1 & A_1 \\ D_1 & B_1 \end{bmatrix} = \begin{bmatrix} 1 & b_0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 1 & b_1 \end{bmatrix} = \begin{bmatrix} b_0 & 1 + b_0 b_1 \\ 1 & b_1 \end{bmatrix},$$

so that

$$\begin{aligned} A_1 &= 1 + b_0 b_1 = b_1 A_0 + A_{-1}, \\ B_1 &= b_1 = b_1 B_0 + B_{-1}, \\ C_1 &= b_0 = A_0, \\ D_1 &= 1 = B_0. \end{aligned}$$

It is thus clear that A_1 and B_1 satisfy the recursion relation.

We prove by induction that A_n and B_n are given by the recursion relation for all n and that $C_n = A_{n-1}$ and $D_n = B_{n-1}$. Suppose that this is true for all $n < k$. Then

$$\begin{bmatrix} C_k & A_k \\ D_k & B_k \end{bmatrix} = \begin{bmatrix} A_{k-2} & A_{k-1} \\ B_{k-2} & B_{k-1} \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 1 & b_k \end{bmatrix} = \begin{bmatrix} A_{k-1} & b_k A_{k-1} + A_{k-2} \\ B_{k-1} & b_k B_{k-1} + B_{k-2} \end{bmatrix}.$$

Hence, by induction, $A_k = b_k A_{k-1} + A_{k-2}$ and $B_k = b_k B_{k-1} + B_{k-2}$ holds for all k. □

EXERCISE 1.1. Prove the claim about compositions of Möbius transformations in the proof of Lemma 1.2.

EXERCISE 1.2. Use a computer and the Gauß transformation to calculate some of the first digits of the continued fraction of π. Use these digits to calculate some convergents of π. Do the same for other constants such as e and $\sqrt{2}$.

EXERCISE 1.3. Evaluate the infinite continued fraction

$$1 + \cfrac{1}{1 + \cfrac{1}{1 + \cfrac{1}{\ddots}}}.$$

EXERCISE 1.4. Find the continued fraction expansion of $\sqrt{2}$.

EXERCISE 1.5. Which of the continued fractions in (1.1) are obtain from Theorem 1.1?

EXERCISE 1.6. Calculate the first convergents of $[0; 1, 1, \ldots]$ using the Euler–Wallis relations. Do you recognise the recursion and the numbers A_n and B_n?

2. Diophantine approximation

Diophantine approximation is the approximation of real numbers, or other interesting objects such as points in $\mathbb{R}^n$, by rational numbers, or other types of numbers. We shall study some of the basic theory in which we approximate real numbers by rational numbers.

It is clear from the definition of real numbers that any real number x can be approximated by a rational number $\frac{p}{q}$ such that the distance $|x - \frac{p}{q}|$ can be made as small as we please. If $\frac{p}{q}$ is a good approximation of x does not only depend on the mentioned distance, but it also depends on the size of q. It is desireable that the distance $|x - \frac{p}{q}|$ as well as q are small. The question is how small the distance can be made while q is not larger than a specified number? The following classical theorem by Dirichlet[5] gives an answer to this question.

THEOREM 1.3 (Dirichlet). *For any $x \in \mathbb{R}$ and any natural number Q, there exists a natural number $q \leq Q$ and an integer p such that*

$$\left|x - \frac{p}{q}\right| < \frac{1}{Qq}.$$

PROOF. Fix Q. The inequality that we want to prove holds for some p and q can be written as

$$|qx - p| \leq \frac{1}{Q}.$$

Hence, we want to prove that for some natural number $q \leq Q$ holds

$$\min_{p \in \mathbb{Z}} |qx - p| < \frac{1}{Q}$$

or equivalently

$$\min\{qx \mod 1, 1 - (qx \mod 1)\} < \frac{1}{Q}. \tag{1.5}$$

[5]Peter Gustav Lejeune Dirichlet, 1805–1859. German mathematician.

We consider the $Q+1$ numbers $kx \mod 1$ where $k = 0, 1, \ldots, Q$. They are all elements of the interval $[0,1)$. We write this interval as the union of Q subintervals,

$$[0,1) = \Big[0, \frac{1}{Q}\Big) \cup \Big[\frac{1}{Q}, \frac{2}{Q}\Big) \cup \ldots \cup \Big[\frac{Q-1}{Q}, 1\Big).$$

Hence, we have $Q+1$ numbers and Q subintervals. It follows that two of the numbers are elements of the same subinterval. (This is called Dirichlet's pigeon hole principle. If $Q+1$ pigeons are put into Q holes — for instance pigeon holes — then at least one hole must contain at least two pigeons.)

Say we have $0 \le k < l \le Q$ and $kx \mod 1$ and $lx \mod 1$ are in the same subinterval. Then the distance between these numbers is less than $1/Q$.

We either have $(lx \mod 1) \ge (kx \mod 1)$ or $(lx \mod 1) \le (kx \mod 1)$. Assume that $(lx \mod 1) \ge (kx \mod 1)$. Then

$$0 \le (lx \mod 1) - (kx \mod 1) = (l-k)x \mod 1 \le \frac{1}{Q}.$$

Similarly, if $(lx \mod 1) \le (kx \mod 1)$. Then

$$-\frac{1}{Q} \le (lx \mod 1) - (kx \mod 1) \le 0,$$

and hence $1 - \frac{1}{Q} \le (l-k)x \mod 1 \le 1$. This proves that (1.5) holds with $q = l - k \le Q$. □

Corollary 1.4. *If x is irrational, then there are infinitely many natural numbers q for which there is an integer p with*

$$\Big|x - \frac{p}{q}\Big| < \frac{1}{q^2}.$$

Proof. Let Q_1 be a natural number. By Theorem 1.3, there exist p_1 and $q_1 \le Q_1$ such that $\big|x - \frac{p_1}{q_1}\big| < \frac{1}{Q_1 q_1} \le \frac{1}{q_1^2}$.

Take $Q_2 > \big|x - \frac{p_1}{q_1}\big|^{-1}$. Then there exists p_2 and $q_2 \le Q_2$ such that $\big|x - \frac{p_2}{q_2}\big| < \frac{1}{Q_2 q_2} \le \frac{1}{Q_2} < \big|x - \frac{p_1}{q_1}\big|$. Hence $\frac{p_2}{q_2} \ne \frac{p_1}{q_1}$ and $\big|x - \frac{p_2}{q_2}\big| \le \frac{1}{q_2^2}$.

Continuing in this way, we get an infinite sequence $\frac{p_k}{q_k}$ with the desired properties. □

A rational number $\frac{p}{q}$ such that $\big|x - \frac{p}{q}\big| < \frac{1}{q^2}$ is called a *best approximand*. Hence, by Corollary 1.4, every irrational number has infinitely many best approximands. Here is a connection between best approximands and continued fractions.

Theorem 1.5. *Let x be an irrational number, and let $\frac{A_n}{B_n}$ be a convergent. Then $\frac{A_n}{B_n}$ is a best approximand.*

Proof. We prove this using the matrices that appeared in the proof of the Euler–Wallis relations, Lemma 1.2. We have

$$\frac{A_n}{B_n} - \frac{A_{n-1}}{B_{n-1}} = \frac{A_n B_{n-1} - A_{n-1} B_n}{B_n B_{n-1}}.$$

But

$$A_nB_{n-1} - A_{n-1}B_n = -\det\begin{bmatrix} A_{n-1} & A_n \\ B_{n-1} & B_n \end{bmatrix} = \pm 1,$$

since this matrix is a product of matrices of determinant ± 1. Hence

$$\left|\frac{A_n}{B_n} - \frac{A_{n-1}}{B_{n-1}}\right| = \frac{1}{B_nB_{n-1}}. \tag{1.6}$$

Now, x lies inside the interval with end points $\frac{A_n}{B_n}$ and $\frac{A_{n-1}}{B_{n-1}}$ and hence

$$\left|x - \frac{A_{n-1}}{B_{n-1}}\right| < \left|\frac{A_n}{B_n} - \frac{A_{n-1}}{B_{n-1}}\right| = \frac{1}{B_nB_{n-1}} < \frac{1}{B_{n-1}^2}.$$

Thus, $\frac{A_{n-1}}{B_{n-1}}$ is a best approximand. □

There is a partial converse of Theorem 1.5 by Legendre[6]: If p and q are such that $\left|x - \frac{p}{q}\right| < \frac{1}{2q^2}$, then $\frac{p}{q}$ is a convergent of x. We shall not prove this result, but refer the reader to a book by Niven [**10**].

THEOREM 1.6. *Let x be an irrational number and $\frac{A_n}{B_n}$ a convergent. Then*

$$\left|x - \frac{A_n}{B_n}\right| \leq \frac{1}{B_nB_{n+1}}.$$

PROOF. This follows from (1.6), changing $n-1$ to n. □

EXERCISE 1.7. Use Theorem 1.6 to explain why $\frac{A_n}{B_n}$ is a particularly good approximation of $x = [0; b_1, b2, \ldots]$ when b_{n+1} is huge.

EXERCISE 1.8. Prove that if x is a real number, and if $\frac{A_n}{B_n}$ and $\frac{A_{n-1}}{B_{n-1}}$ are two consequtive convergents of x, then either

$$\left|x - \frac{A_{n-1}}{B_{n-1}}\right| < \frac{1}{2B_{n-1}^2} \qquad \text{or} \qquad \left|x - \frac{A_n}{B_n}\right| < \frac{1}{2B_n^2}.$$

EXERCISE 1.9. Use the proof of Theorem 1.5 to give an alternative proof of Corollary 1.4.

EXERCISE 1.10. Find a rational approximation to $\sqrt{2}$ such that the error is less than 10^{-3}.

3. Irrational and transcendental numbers

We shall now relate the theory of Diophantine approximation with transcendental numbers, which we now define.

DEFINITION 1.7 (algebraic number, transcendental number). A real number x is called an *algebraic number of degree d* if d is the smallest natural number such that there are integer $a_0, a_1, \ldots, a_d$, with $a_d \neq 0$, and

$$a_dx^d + a_{d-1}x^{d-1} + \ldots + a_1x + a_0 = 0.$$

[6]Adrien-Marie Legendre, 1752–1833. French mathematician.

A number is called an *algebraic number* if it is an algebraic number of degree d for some d.

A number which is not an algebraic number is called a *transcendental number*.

Note that all rational numbers are algebraic numbers.

THEOREM 1.8 (Liouville[7]). *Let x be an algebraic number of degree $d \geq 2$. Then there is a constant $c > 0$ such that*

$$\left|x - \frac{p}{q}\right| > \frac{c}{q^d}$$

holds for all integers p and natural numbers q.

PROOF. An integer polynomial is a polynomial with integer coefficients. Let P be an integer polynomial of degree d such that $P(x) = 0$, and assume that d is minimal. Then x is a simple root of P, since otherwise P' would be an integer polynomial of degree $d - 1$ such that $P'(x) = 0$, which would violate the minimality of d.

We have $P'(x) \neq 0$, and since P' has finitely many zeroes, there is a number $r > 0$ such that there are no zeroes of P' within a distance $2r$ to x.

Let now $\frac{p}{q}$ be a rational number such that $\left|x - \frac{p}{q}\right| < r$. Since $d \geq 3$, we have $x \neq \frac{p}{q}$. By the mean value theorem, there is a ξ between x and $\frac{p}{q}$ such that

$$P'(\xi)\Big(x - \frac{p}{q}\Big) = P(x) - P\Big(\frac{p}{q}\Big) = -P\Big(\frac{p}{q}\Big).$$

But $P\big(\frac{p}{q}\big)$ is a rational number with denominator at most q^d, and since $P\big(\frac{p}{q}\big) \neq 0$ the modulus of the nominator is at least 1. Hence,

$$|P'(\xi)|\left|x - \frac{p}{q}\right| \geq \frac{1}{q^d}.$$

Since P' is bounded within a distance r to x, there is a constant c such that $|P'(\xi)| < \frac{1}{c}$. Then

$$\left|x - \frac{p}{q}\right| \geq \frac{1}{|P'(\xi)|q^d} > \frac{c}{q^d}.$$

If necessary, we can make c even smaller, so that $\left|x - \frac{p}{q}\right| > \frac{c}{q^d}$ also holds for the at most finitely many rationals $\frac{p}{q}$ for which this inequality is not valid. □

Numbers x such that for every d, there is no constant $c > 0$ such that

$$\left|x - \frac{p}{q}\right| > \frac{c}{q^d}$$

holds for all p and q are called *Liouville numbers*. In other words, Liouville numbers are those irrational numbers x such that for any d, there are infinitely many p and q such that $\left|x - \frac{p}{q}\right| < \frac{1}{q^d}$.

[7] Joseph Liouville, 1809–1882. French mathematician.

Hence, Theorem 1.8 says that every Liouville number is transcendental. Unfortunately, most transcendental numbers are not Liouville numbers. It is therefore not always possible to prove that a number is transcendental by using Theorem 1.8.

Let us now prove that some well-known numbers are irrational or even transcendental. For more results of this type, the reader is refered to Niven's book [**10**].

THEOREM 1.9. *The number e is irrational.*

PROOF. We have $e = \sum_{k=0}^{\infty} \frac{1}{k!} < 3$ and $e > 1 + 1 + \frac{1}{2} = \frac{5}{2}$.

Suppose that $e = \frac{p}{q}$, where p and q are natural numbers. Then, since $\frac{5}{2} < e < 3$, we must have $q > 2$. Furthermore, $q!e$ is a natural number and

$$q!e = \sum_{k=0}^{q} \frac{q!}{k!} + \sum_{k=q+1}^{\infty} \frac{q!}{k!}.$$

The first sum is a natural number, since all its terms are natural numbers. The second sum satisfies

$$0 < \sum_{k=q+1}^{\infty} \frac{q!}{k!} = \frac{1}{q+1} + \frac{1}{(q+1)(q+2)} + \ldots \leq \sum_{k=1}^{\infty} \frac{1}{(q+1)^k} = \frac{1}{q} \leq \frac{1}{3}.$$

Hence we have written the integer $q!e$ as a sum of an integer and a positive number not larger than $\frac{1}{3}$, which is impossible. □

THEOREM 1.10. *If $x \neq 0$ is rational, then e^x is irrational.*

PROOF. Suppose that $x = \frac{p}{q}$ and that e^x is rational. Then $e^p = (e^x)^q$ is also rational. Hence it is sufficient to prove that e^p is irrational when p is a non-zero integer. By considering e^{-p} instead, we may even assume that $p > 0$.

Let $f(x) = \frac{x^n(1-x)^n}{n!} = \frac{1}{n!}\phi(x)\phi(1-x)$, where $\phi(x) = x^n$. We have

$$f^{(k)}(x) = \frac{1}{n!} \sum_{j=0}^{k} \binom{k}{j} \phi^{(j)}(x)(-1)^{k-j}\phi^{(k-j)}(1-x).$$

Clearly $\phi^{(j)}(0) \neq 0$ only if $j = n$ and then $\phi^{(n)}(0) = n!$. Thus, if $k \geq n$

$$f^{(k)}(0) = \frac{1}{n!} \sum_{j=0}^{k} \binom{k}{j} \phi^{(j)}(0)(-1)^{k-j}\phi^{(k-j)}(1) = \binom{k}{n}(-1)^{k-n}\phi^{(k-n)}(1)$$

which is an integer. Otherwise, if $k < n$, $f^{(k)}(0) = 0$. Similary, $f^{(k)}(1) = 0$ if $k < n$ and if $k \geq n$, then $f^{(k)}(1) = \binom{k}{n}(-1)^n\phi^{(k-n)}(0)$ which is an integer. Since f is a polynomial of degree $2n$, we have $f^{(k)} = 0$ for $k > 2n$.

Define

$$g(x) = \sum_{k=0}^{2n} (-1)^k p^{2n-k} f^{(k)}(x).$$

It follows that $g(0)$ and $g(1)$ are integers since $f^{(k)}(0)$ and $f^{(k)}(1)$ are integers. We also have

$$g'(x) = \sum_{k=0}^{2n}(-1)^k p^{2n-k} f^{(k+1)}(x) = \sum_{k=1}^{2n}(-1)^{k-1} p^{2n-k+1} f^{(k)}(x),$$

so that $pg(x) + g'(x) = p^{2n+1} f(x)$. It then follows that

$$\frac{\mathrm{d}}{\mathrm{d}x}(e^{px} g(x)) = e^{px}(pg(x) + g'(x)) = p^{2n+1} e^{px} f(x).$$

Now, if we assume that $e^p = \frac{s}{t}$ is a rational number, then we get a contradiction in the following way. The number

$$I_n = tp^{2n+1} \int_0^1 e^{px} f(x)\,\mathrm{d}x = t\big[e^{px} g(x)\big]_0^1 = t\Big(\frac{s}{t} g(1) - g(0)\Big) = sg(1) - tg(0)$$

is an integer. But by the estimates $0 < e^{px} \le e^p$ and $0 < f(x) < \frac{1}{n!}$ for $0 < x < 1$ we get

$$0 < tp^{2n+1} \int_0^1 e^{px} f(x)\,\mathrm{d}x < \frac{tp^{2n+1} e^p}{n!}.$$

This shows that I_n is an integer, $I_n > 0$ and $I_n \to 0$ as $n \to \infty$, which is impossible. □

THEOREM 1.11. *The number π^2 is irrational.*

PROOF. Assume that $\pi^2 = \frac{p}{q}$, where p and q are natural numbers. As in the proof of Theorem 1.10, we use the function $f(x) = \frac{x^n(1-x)^n}{n!}$. We consider the integral

$$I_n = \int_0^1 q^n \pi^{2n+1} f(x) \sin(\pi x)\,\mathrm{d}x = \int_0^1 \pi p^n f(x) \sin(\pi x)\,\mathrm{d}x.$$

To get a contradiction, we will show that I_n is always an integer and that $0 < I_n < 1$ holds if n is large.

Put

$$F(x) = q^n(\pi^{2n} f(x) - \pi^{2n-2} f^{(2)}(x) + \pi^{2n-4} f^{(4)}(x) - \ldots).$$

Then

$$\begin{aligned} F''(x) &= q^n(\pi^{2n} f^{(2)}(x) - \pi^{2n-2} f^{(4)}(x) + \pi^{2n-4} f^{(6)}(x) - \ldots) \\ &= q^n - \pi^2 F(x) + b^n \pi^{2n} f(x). \end{aligned}$$

Since $f^{(k)}(0)$ and $f^{(k)}(1)$ are integers, and $q^n \pi^{2n} = p^n$, we have that $F(0)$ and $F(1)$ are integers as well.

We differentiate and get

$$\begin{aligned} &\frac{\mathrm{d}}{\mathrm{d}x}(F'(x)\sin\pi x - \pi F(x)\cos\pi x) \\ &\quad = F''(x)\sin\pi x + \pi F'(x)\cos\pi x - \pi F'(x)\cos\pi x + \pi^2 F(x)\sin\pi x \\ &\quad = (F''(x) + \pi^2 F(x))\sin\pi x = q^n \pi^{2n+2} f(x). \end{aligned}$$

Hence

$$I_n = \Big[\frac{1}{\pi}(F'(x)\sin\pi x - \pi F(x)\cos\pi x)\Big]_0^1 = F(1) + F(0),$$

which is an integer.

Clearly $I_n > 0$. Since both x and $1 - x$ are numbers in $[0, 1]$ when $x \in [0, 1]$ we have that the maximum of f on $[0, 1]$ is at most $\frac{1}{n!}$. It follows that

$$I_n \leq \pi \frac{p^n}{n!}.$$

But $p^n/n! \to 0$ as $n \to \infty$, so if n is large we have that I_n is an integer and $0 < I_n < 1$ which is impossible. □

THEOREM 1.12. *The number e is transcendental.*

PROOF. To get a contradiction, we assume that e is algebraic of degree d. Then there are integers $a_0, a_1, \ldots, a_d$ with $a_0 \neq 0$ such that

$$a_d e^d + a_{d-1} e^{d-1} + \ldots + a_1 e + a_0 = 0.$$

The method is similar to previous proofs, but this time we define

$$f(x) = \frac{x^{n-1}(x-1)^n(x-2)^n \ldots (x-d)^n}{(n-1)!}.$$

We are going to consider the sum

$$S_n = \sum_{k=0}^{d} a_k e^k \int_0^k e^{-x} f(x)\, \mathrm{d}x$$

and prove that n can be chosen so that S_n is a non-zero integer such that $|S_n| < 1$. No such integers exists, so this is a contradiction.

First, since

$$|f(x)| < \frac{m^{n-1} m^n m^n \ldots m^n}{(n-1)!} = \frac{m^{dn+n-1}}{(n-1)!},$$

we have

$$|S_n| < \sum_{k=0}^{d} |a_k| e^k k \frac{m^{dn+n-1}}{(n-1)!} \leq \left(\sum_{k=0}^{d} |a_k| \right) e^m \frac{(m^{d+1})^n}{(n-1)!}.$$

But $\frac{(m^{d+1})^n}{(n-1)!} \to 0$ as $n \to \infty$, so we have $|S_n| < 1$ if n is large enough.

We will now prove that S_n is a non-zero integer. Put

$$F(x) = f(x) + f'(x) + f^{(2)}(x) + \ldots + f^{(dn+n-1)}(x).$$

Then

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(e^{-x} F(x)\right) = e^{-x}(F'(x) - F(x)) = -e^{-x} f(x).$$

It then follows that

$$a_k \int_0^k e^{-x} f(x)\, \mathrm{d}x = a_k \left[e^{-x} F(x) \right]_0^k = a_k F(0) - a_k e^{-k} F(k),$$

and

$$S_n = \sum_{k=0}^{d} a_k e^k \int_0^k e^{-x} f(x)\, \mathrm{d}x = \sum_{k=0}^{d} a_k e^k F(0) - \sum_{k=0}^{d} a_k F(k).$$

The first sum on the right hand side is zero because of the assumption that e is algebraic. Hence

$$S_n = -\sum_{k=0}^{d} a_k F(k) = -\sum_{k=0}^{d} \sum_{j=0}^{dn+n-1} a_k f^{(j)}(k).$$

Let n be a prime number. We will prove that all terms $a_k f^{(j)}(k)$ are integer multiples of n, except for the term $a_0 f^{(n-1)}(0)$, which is an integer not divisible by n. This means that

$$S_n = nN + M$$

where N and M are integers, and n does not divide M. It follows that S_n is an integer not divisible by n and in particular S_n is a non-zero integer, which finishes the proof.

It remains only to prove the claims about the terms $a_k f^{(j)}(k)$.

We can write f as $f(x) = \frac{x^{n-1}}{(n-1)!} g_0(x)$, where g_0 is a polynomial with integer coefficients. Since all derivatives of $\frac{x^{n-1}}{(n-1)!}$ of order less than $n-1$ are zero at $x = 0$ we have.

$$f^{n-1}(0) = \frac{\mathrm{d}^n}{\mathrm{d}x^n} \frac{x^{n-1}}{(n-1)!}\bigg|_{x=0} \cdot g_0(0) = g_0(0) = (-1)^n (-2)^n \dots (-d)^n.$$

Hence $f^{(n-1)}(0)$ is an integer, and since n is a prime, $f^{(n-1)}(0)$ is not divisible by n if we choose $n > d$.

Finally, to prove that $f^{(j)}(k)$ is an integer divisible by n unless $j = n-1$ and $k = 0$, we write

$$f(x) = \frac{1}{(n-1)!} \phi_k(x) g_k(x)$$

where $\phi_k(x) = (x-k)^n$ and g_k is a polynomial with integer coefficients. The j-th derivative of f at k is of the form

$$f^{(j)}(k) = \frac{1}{(n-1)!} \sum_{l=0}^{j} \binom{j}{l} \phi_k^{(l)}(k) g_k^{(j-l)}(k).$$

The factor $g_k^{(j-l)}(k)$ is always an integer, whereas

$$\phi_k^{(l)}(k) = \begin{cases} 0 & \text{if } l \neq n, \\ n! & \text{if } l = n. \end{cases}$$

Hence $f^{(j)}(k) = 0$ or $f^{(j)}(k) = n g_k^{(j-n)}(k)$, which in both cases is an integer divisible by n. □

We have proved above that π^2 is irrational, and that e is transcendental. In fact, π is also transcendental, see Niven's book [**10**] for a proof.

EXERCISE 1.11. Show that $\sum_{n=0}^{\infty} 10^{-n!}$ is a Liouville number.

EXERCISE 1.12. Prove that there exist irrational numbers a and b such that a^b is rational. Hint: Consider $\sqrt{2}^{\sqrt{2}}$ and $\left(\sqrt{2}^{\sqrt{2}}\right)^{\sqrt{2}}$.

CHAPTER 2

Riemann–Stieltjes integrals

1. Functions of bounded variation

DEFINITION 2.1 (Total variation). The total variation of a function $f\colon [a,b] \to \mathbb{R}$ on the interval $[a,b]$ is the number

$$\operatorname{var}_{[a,b]} f = \sup\left\{ \sum_{k=1}^{n} |f(x_k) - f(x_{k-1})| : a = x_0 < x_1 < \ldots < x_n = b \right\}.$$

DEFINITION 2.2 (Bounded variation). A function $f\colon [a,b] \to \mathbb{R}$ is said to be of bounded variation in $[a,b]$ if $\operatorname{var}_{[a,b]} f < \infty$. The functions of bounded variation on $[a,b]$ is the set

$$BV([a,b]) = \{ f\colon [a,b] \to \mathbb{R} : \operatorname{var}_{[a,b]} f < \infty \}.$$

LEMMA 2.3 (Jordan[1]). *A function f is of bounded variation if and only if there are two increasing functions $g, h\colon [a,b] \to \mathbb{R}$ such that $f = g - h$.*

PROOF. Suppose first that f is of bounded variation. Define $g(x) = \operatorname{var}_{[a,x]} f$ and $h = g - f$. Clearly, g is increasing. If $a \le x < y \le b$, then

$$\begin{aligned} h(y) - h(x) &= g(y) - g(x) - (f(y) - f(x)) \\ &\ge g(y) - g(x) - |f(y) - f(x)| \ge g(y) - g(x) - \operatorname{var}_{[x,y]} f = 0, \end{aligned}$$

which proves that h is increasing. Finally, $f = g - h$ holds by the definition of h.

Now, assume instead that we know that $f = g - h$, where g and h are increasing functions. Then g and h are functions of bounded variation, since we have $\operatorname{var}_{[a,b]} g = |g(b) - g(a)|$ and similarly for h. It now follows that f is of bounded variation since $\operatorname{var}_{[a,b]}(\phi + \psi) \le \operatorname{var}_{[a,b]} \phi + \operatorname{var}_{[a,b]} \psi$ holds whenever $\phi, \psi\colon [a,b] \to \mathbb{R}$ are two functions. □

EXERCISE 2.1. Suppose that $f \in BV([a,b])$ and let E be the set of points in $[a,b]$ in which f is not continuous. Show that E is at most countable.

2. Definition of the Riemann–Stieltjes integral

We will say that $P = \{x_k\}_{k=0}^{n}$ is a *partition* of an interval $[a,b]$ if

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

[1] Marie Ennemond Camille Jordan, 1838–1922. French mathematician.

If P_1 and P_2 are two partitions of the interval $[a,b]$, then P_1 is said to be finer than P_2 if $P_1 \supset P_2$.

If $P = \{x_k\}_{k=0}^n$ is a partition of I, we let $\Delta(P)$ be the number

$$\Delta(P) = \max\{\, x_k - x_{k-1} : k = 1, 2, \dots, n \,\}.$$

DEFINITION 2.4 (Riemann[2]–Stieltjes[3] integrals). Suppose I is a compact and non-empty interval and let $\alpha\colon I \to \mathbb{R}$. A function $f\colon I \to \mathbb{R}$ is said to be Riemann–Stieltjes integrable over I with respect to α if there is a number $s(f,\alpha)$ such that if $\varepsilon > 0$, then there exists a partition P of I such that

$$\left| s(f,\alpha) - \sum_{k=1}^{n} f(\xi_k)(\alpha(x_k) - \alpha(x_{k-1})) \right| < \varepsilon$$

whenever $\{x_k\}_{k=0}^n$ is a partition which is finer than P, and $\xi_k \in [x_{k-1}, x_k]$.

If f is Riemann–Stieltjes integrable over I with respect to α, then $s(f,\alpha)$ is called the Riemann–Stieltjes integral of f over I with respect to α, and we write

$$\int_I f \,\mathrm{d}\alpha = s(f,\alpha).$$

Note that if $\alpha(x) = x$, then we recover the ordinary Riemann integral.

It is possible to consider also countable partitions, that is partitions of the interval which consists of countably many points. One can then define the Riemann–Stieltjes integral as above, and this makes it possible to integrate some functions that would otherwise not be integrable in the above sense. We shall not do so here, and refer the reader to the books by Apostol [**1**] and McLeod [**9**].

3. Properties of the Riemann–Stieltjes integral

According to Lemma 2.3, any function α of bounded variation can be written as a difference of two increasing functions. If $\alpha = \alpha_1 - \alpha_2$ is of bounded variation and α_1 and α_2 are increasing, and the integrals

$$\int_I f \,\mathrm{d}\alpha_1 \qquad \text{and} \qquad \int_I f \,\mathrm{d}\alpha_1$$

exist, then $\int_I f \,\mathrm{d}\alpha$ exists and

$$\int_I f \,\mathrm{d}\alpha = \int_I f \,\mathrm{d}\alpha_1 - \int_I f \,\mathrm{d}\alpha_1.$$

(See Exercise 2.2.)

We prove the following result.

PROPOSITION 2.5. *If I is a compact interval, $f\colon I \to \mathbb{R}$ is continuous and $\alpha\colon I \to \mathbb{R}$ is increasing, then f is Riemann–Stieltjes integrable with respect to α.*

---

[2]Georg Friedrich Bernhard Riemann, 1826–1866. German mathematician.

[3]Thomas Joannes Stieltjes, 1856–1894. Dutch mathematician.

PROOF. Since I is compact and f is continuous, f is uniformly continuous. Let $\varepsilon > 0$. There is then a $\delta > 0$ such that $|f(x) - f(y)| < \varepsilon$ whenever $|x - y| < \delta$. Let $P = \{x_k\}_{k=1}^{\infty}$ be a partition with $\Delta(P) < \delta$ and suppose $I = [a, b]$.

Whatever choice of $\xi_k, \tilde{\xi}_k \in [x_{k-1}, x_k]$ we make, we have $|\xi_k - \tilde{\xi}_k| < \delta$ and therefore

$$\begin{aligned}\left|\sum_{k=1}^{n} f(\xi_k)(\alpha(x_k) - \alpha(x_{k-1}) - \sum_{k=1}^{n} f(\tilde{\xi}_k)(\alpha(x_k) - \alpha(x_{k-1})\right| \\ \leq \sum_{k=1}^{n} |f(\xi_k) - f(\tilde{x}_k)|(\alpha(x_k) - \alpha(x_{k-1})) \\ \leq \varepsilon \sum_{k=1}^{n} |f(\xi_k) - f(\tilde{x}_k)|(\alpha(x_k) - \alpha(x_{k-1})) = \varepsilon(\alpha(b) - \alpha(a)).\end{aligned}$$

This shows that there is a compact interval J of length at most $\varepsilon(\alpha(b) - \alpha(a))$ such that

$$\sum_{k=1}^{n} f(\xi_k)(\alpha(x_k) - \alpha(x_{k-1})) \in J,$$

for all choices of $\xi_k \in [x_{k-1}, x_k]$.

Since α is increasing, any partition Q finer than P has the property that the Riemann–Stieltjes sum to Q lies in J. This is proved as follows. If Q is finer that P, then Q can be obtained by subdivision of some of the intervals $[x_{k-1}, x_k]$ into smaller intervals. Say that $[x_{k-1}, x_k]$ is subdivided by the points

$$x_{k-1} = \tilde{x}_{l-1} < \tilde{x}_l < \ldots < \tilde{x}_{l+m} = x_k$$

which belong to Q. To get the Riemann–Stieltjes sum corresponding to the partition Q, we then replace the term $f(\xi_k)(\alpha(x_k) - \alpha(x_{k-1}))$ in the Riemann–Stieltjes sum for P by

$$T_k := f(\zeta_l)(\alpha(\tilde{x}_l) - \alpha(\tilde{x}_{l-1})) + \ldots + f(\zeta_{l+m})(\alpha(\tilde{x}_{l+m}) - \alpha(\tilde{x}_{l+m-1})).$$

Then, since

$$(\alpha(\tilde{x}_l) - \alpha(\tilde{x}_{l-1})) + \ldots + (\alpha(\tilde{x}_{l+m}) - \alpha(\tilde{x}_{l+m-1})) = \alpha(x_k) - \alpha(x_{k-1}),$$

we have

$$\inf_{\xi_k \in [x_{k-1}, x_k]} f(\xi_k)(\alpha(x_k) - \alpha(x_{k-1})) \leq T_k \leq \sup_{\xi_k \in [x_{k-1}, x_k]} f(\xi_k)(\alpha(x_k) - \alpha(x_{k-1})).$$

Since this is true for all k, the Riemann–Stieltjes sum corresponding to Q will be in J. This proves that f is Riemann–Stieltjes integrable. □

Since any function of bounded variation can be written as a difference between two increasing functions, we now get the following corollary.

COROLLARY 2.6. *If I is a compact interval, $f: I \to \mathbb{R}$ is continuous and $\alpha \in BV(I)$, then f is Riemann–Stieltjes integrable with respect to α.*

LEMMA 2.7. *If α is of bounded variation and f is Riemann–Stieltjes integrable with respect to α, then*

$$\left|\int_I f\,\mathrm{d}\alpha\right| \leq \sup_I |f| \operatorname{var}_I \alpha.$$

PROOF. The lemma follows from

$$\begin{aligned}\left|\sum_{k=1}^n f(\xi_k)(\alpha(x_k)-\alpha(x_{k-1}))\right| &\leq \sum_{k=1}^n |f(\xi_k)||\alpha(x_k)-\alpha(x_{k-1})| \\ &\leq \sup_I f \sum_{k=1}^n |\alpha(x_k)-\alpha(x_{k-1})| \leq \sup_I f \operatorname{var}_I \alpha. \square\end{aligned}$$

THEOREM 2.8. *If I is a compact interval, $f\colon I \to \mathbb{R}$ is Riemann–Stieltjes integrable with respect to α and α is continuously differentiable, then $f\alpha'$ is Riemann integrable and*

$$\int_I f(x)\,\mathrm{d}\alpha(x) = \int_I f(x)\alpha'(x)\,\mathrm{d}x.$$

PROOF. This is proved using the mean value theorem: If $\alpha\colon [a,b] \to \mathbb{R}$ is differentiable, then there exists a $c \in [a,b]$ such that

$$\alpha'(c) = \frac{\alpha(b)-\alpha(a)}{b-a}.$$

Let $\varepsilon > 0$.

Since f is integrable, it is bounded, and we have $|f| \leq M$ for some M. Since α' is continuous and I is compact, α' is uniformly continuous, and there is a δ such that

$$|\alpha'(x)-\alpha'(y)| < \frac{\varepsilon}{2M|I|}$$

when $|x-y| < \delta$.

Put $s = \int_I f\,\mathrm{d}\alpha$. Since f is Riemann–Stieltjes integrable with respect to α, there exists a partition P of I such that

$$\left|s - \sum_{k=1}^n f(\xi_k)(\alpha(x_k)-\alpha(x_{k-1}))\right| < \frac{\varepsilon}{2} \tag{2.1}$$

holds whenever $\{x_k\}_{k=0}^n$ is a partition which is finer than P and $\xi_k \in [x_{k-1},x_k]$. We may assume that $\Delta(P) < \delta$, and we do so.

Let $\{x_k\}_{k=0}^n$ be finer than P. By the mean value theorem, there are numbers $\tilde{\xi}_k \in [x_{k-1},x_k]$, such that

$$(\alpha(x_k)-\alpha(x_{k-1})) = \alpha'(\tilde{\xi}_k)(x_k-x_{k-1}).$$

We now consider the sum

$$\sum_{k=1}^n f(\xi_k)\alpha'(\xi_k)(x_k-x_{k-1})$$

which is an ordinary Riemann sum of the function $f\alpha'$. We then have

$$\begin{aligned}
&\left|\sum_{k=0}^{n} f(\xi_k)\alpha'(\xi_k)(x_k - x_{k-1}) - \sum_{k=1}^{n} f(\xi_k)(\alpha(x_k) - \alpha(x_{k-1}))\right| \\
&\quad = \left|\sum_{k=0}^{n} f(\xi_k)\alpha'(\xi_k)(x_k - x_{k-1}) - \sum_{k=1}^{n} f(\xi_k)\alpha'(\tilde{\xi}_k)(x_k - x_{k-1})\right| \\
&\quad \leq \sum_{k=1}^{n} |f(\xi_k)||\alpha'(\xi_k) - \alpha'(\tilde{\xi}_k)|(x_k - x_{k-1}) \\
&\quad \leq \sum_{k=1}^{n} M\frac{\varepsilon}{2M|I|}(x_k - x_{k-1}) = \frac{\varepsilon}{2}.
\end{aligned}$$

Combining with (2.1), we conclude that

$$\left|s - \sum_{k=0}^{n} f(\xi_k)\alpha'(\xi_k)(x_k - x_{k-1})\right| < \varepsilon.$$

But since $\varepsilon > 0$ is arbitrary, this is just the definition of that $f\alpha'$ is Riemann integrable over I with integral $s = \int_I f\,\mathrm{d}\alpha$. □

EXERCISE 2.2. Prove Corollary 2.6. That is, prove the claim before Proposition 2.5

EXERCISE 2.3. Suppose $I = [0,1]$ and $0 < c < 1$. Let $f(x) = \alpha(x) = 0$ if $x < c$ and $f(x) = \alpha(x) = 1$ if $x > c$. Let also $\alpha(c) = 1$. Is f integrable with respect to α?

EXERCISE 2.4. Suppose that α is increasing and that f and g are Riemann–Stieltjes integrable with respect to α. Show that

$$f \leq g \qquad \Rightarrow \qquad \int_I f\,\mathrm{d}\alpha \leq \int_I g\,\mathrm{d}\alpha.$$

Is this implication true if α is not increasing?

EXERCISE 2.5. Let a and b be integers with $a < b$. Let $\alpha(x) = [x]$, where $[x]$ denotes the function which is the integer part of x, that is, $[x]$ is the largest integer such that $[x] \leq x$. Prove that

$$\sum_{k=a+1}^{b} f(k) = \int_{[a,b]} f\,\mathrm{d}\alpha$$

holds if f is continuous.

EXERCISE 2.6. Let $a < b < c$. Prove that if f is Riemann–Stieltjes integrable with respect to α on $[a,c]$, then f is Riemann–Stieltjes integrable on $[a,b]$ and $[b,c]$ and

$$\int_{[a,b]} f\,\mathrm{d}\alpha + \int_{[b,c]} f\,\mathrm{d}\alpha = \int_{[a,c]} f\,\mathrm{d}\alpha.$$

4. Integration by parts

THEOREM 2.9 (Integration by parts). *Assume that $f\colon [a,b] \to \mathbb{R}$ is Riemann–Stieltjes integrable with respect to α. Then α is Riemann–Stieltjes integrable with respect to f and*

$$\int_{[a,b]} f(x)\,\mathrm{d}\alpha(x) + \int_{[a,b]} \alpha(x)\,\mathrm{d}f(x) = f(b)\alpha(b) - f(a)\alpha(a).$$

PROOF. Let $\varepsilon > 0$ and $s_1 = \int_I f\,\mathrm{d}\alpha$. There is then a partition P such that if $\{x_k\}_{k=0}^{\infty}$ is finer than P, then

$$\left| s_1 - \sum_{k=1}^{n} f(\xi_k)(\alpha(x_k) - \alpha(x_{k-1})) \right| < \varepsilon,$$

holds whenever $\xi_k \in [x_{k-1}, x_k]$. In particular, we can choose $\xi_k = x_k$.

Parallell to the Riemann–Stieltjes sum above we consider

$$\sum_{k=1}^{n} \alpha(\tilde{\xi}_k)(f(x_k) - f(x_{k-1})),$$

which is a Riemann–Stieltjes sum for the integral $\int_I \alpha\,\mathrm{d}f$. (We have not proved that the integral exists, but we can consider the sum anyway.)

Let $\{x_k\}_{k=0}^{n}$ be a partition finer than P. We write

$$\sum_{k=1}^{n} \alpha(\tilde{\xi}_k)(f(x_k) - f(x_{k-1})) = \sum_{k=1}^{n} \alpha(\tilde{\xi}_k) f(x_k) - \sum_{k=1}^{n} \alpha(\tilde{\xi}_k) f(x_{k-1}),$$

and

$$f(b)\alpha(b) - f(a)\alpha(a) = \sum_{k=1}^{n} f(x_k)\alpha(x_k) - \sum_{k=1}^{n} f(x_{k-1})\alpha(x_{k-1}).$$

Combining these equalities, we get

$$\begin{aligned} f(b)\alpha(b) - f(a)\alpha(a) &- \sum_{k=1}^{n} \alpha(\tilde{\xi}_k)(f(x_k) - f(x_{k-1})) \\ &= \sum_{k=1}^{n} f(x_k)(\alpha(x_k) - \alpha(\tilde{\xi}_k)) - \sum_{k=1}^{n} f(x_{k-1})(\alpha(x_{k-1}) - \alpha(\tilde{\xi}_k)) \\ &= \sum_{k=1}^{n} f(x_k)(\alpha(x_k) - \alpha(\tilde{\xi}_k)) + \sum_{k=1}^{n} f(x_{k-1})(\alpha(\tilde{\xi}_k) - \alpha(x_{k-1})). \end{aligned}$$

By letting $\{t_k\}_{k=0}^{m} = \{x_k\}_{k=0}^{n} \cup \{\xi_k\}_{k=1}^{n}$, we get a new partition which is finer than P, and the equality above can be written as

$$\begin{aligned} f(b)\alpha(b) - f(a)\alpha(a) &- \sum_{k=1}^{n} \alpha(\tilde{\xi}_k)(f(x_k) - f(x_{k-1})) \\ &= \sum_{k=1}^{m} f(\hat{\xi}_k)(\alpha(t_k) - \alpha(t_{k-1})), \end{aligned}$$

where $\hat{\xi}_k$ is either x_k or x_{k-1}.

Now, since $\{t_k\}_{k=0}^m$ is finer than P, we have that

$$\left| f(b)\alpha(b) - f(a)\alpha(a) - s_1 - \sum_{k=1}^{n} \alpha(\tilde{\xi}_k)(f(x_k) - f(x_{k-1})) \right| = \left| s_1 - \sum_{k=1}^{m} f(\hat{\xi}_k)(\alpha(t_k) - \alpha(t_{k-1})) \right| < \varepsilon.$$

If we let $s_2 = f(b)\alpha(b) - f(a)\alpha(a) - s_1$, we then have

$$\left| s_2 - \sum_{k=1}^{n} \alpha(\tilde{\xi}_k)(f(x_k) - f(x_{k-1})) \right| < \varepsilon.$$

Since $\varepsilon > 0$, this says that α is Riemann–Stieltjes integrable with respect to f with

$$\int_I \alpha \, \mathrm{d}f = s_2 = f(b)\alpha(b) - f(a)\alpha(a) - s_1 = f(b)\alpha(b) - f(a)\alpha(a) - \int_I f \, \mathrm{d}\alpha. \quad \square$$

COROLLARY 2.10. *If $f \in BV(I)$ and $\alpha\colon I \to \mathbb{R}$ is continuous, then f is Riemann–Stieltjes integrable with respect to α.*

PROOF. This follows by combining Corollary 2.6 and Theorem 2.9. $\square$

5. Riesz' representation theorem

Let I be a compact and non-empty interval. The set of continuous functions from I to $\mathbb{R}$ is denoted by $\mathcal{C}(I)$, and it is a linear space, where addition of continuous functions and multiplication of a continuous function with a real scalar are defined in the natural way. Most often, one considers $\mathcal{C}(I)$ together with the norm defined by

$$\|f\| = \sup_{x \in I} |f(x)|, \qquad f \in \mathcal{C}(I).$$

We will call this norm the *uniform norm*. It will be used in several of the following chapters.

A *linear functional* L on $\mathcal{C}(I)$ is a linear function $L\colon \mathcal{C}(I) \to \mathbb{R}$. It is continuous if there exists a constant C such that

$$|L(f)| \leq C\|f\|, \qquad \text{for all } f \in \mathcal{C}(I). \tag{2.2}$$

The infimum of all C such that (2.2) holds is called the *operator norm* of L and is denoted by $\|L\|$. We thus have

$$\|L\| = \sup_{\|f\| \neq 0} \frac{|L(f)|}{\|f\|}$$

and

$$|L(f)| \leq \|L\|\|f\|.$$

We now prove the following theorem by Frigyes Riesz[4].

[4]Frigyes Riesz, 1880–1956. Hungarian mathematician and brother of the mathematician Marcel Riesz, 1886–1969. Marcel Riesz was professor in Lund 1926–1952.

THEOREM 2.11 (Riesz' representation theorem). *Suppose $L\colon \mathcal{C}(I) \to \mathbb{R}$. Then L is a continuous linear functional on $\mathcal{C}(I)$ if and only if there is a function α of bounded variation such that*

$$L(f) = \int_I f \,\mathrm{d}\alpha, \qquad \text{for all } f \in \mathcal{C}(I). \tag{2.3}$$

It will take a bit of time to prove this theorem, so let us start by explaining the overall idea of the proof. The operator L is only defined for continuous functions. We will first extend the operator to more functions, including functions of the form χ_J, where J is an interval. These, so called indicator functions, are defined by

$$\chi_J(x) = \begin{cases} 1 & \text{if } x \in J, \\ 0 & \text{if } x \notin J. \end{cases}$$

Hence $\chi_J(x) = 1$ if $x \in J$ and $\chi_j(x) = 0$ otherwise. By extending the operator, we mean that we define $L(\chi_J)$ in such a way that L remains linear. This will be done in the next section, using that the function χ_J is a limit of continuous function.

Once the extension is done, we define $\alpha(x) = L(\chi_{[a,x]})$. We then prove that α is of bounded variation and that $L(f) = \int_I f \,\mathrm{d}\alpha$.

5.1. Extending the operator. We say that a function $f\colon I \to \mathbb{R}$ is the pointwise limit of a decreasing sequence of continuous functions if there is a sequence $(f_n)_{n=1}^{\infty}$ of continuous functions such that $f_1 \geq f_2 \geq \dots$ and for any $x \in I$ holds $f(x) = \lim_{n\to\infty} f_n(x)$.

In this section, we will prove that any linear and continuous functional $L\colon \mathcal{C}(I) \to \mathbb{R}$ can be extended to bounded functions $f\colon I \to \mathbb{R}$, with the property that f is the pointwise limit of a decreasing sequence of continuous functions. In fact, we will extend the operator to an even larger set. Let $\mathcal{D}(I)$ be the set of functions which can be writen as a difference of twbounded o functions, each of which is a pointwise limit of a decreasing sequence of continuous functions. Hence the set $\mathcal{D}(I)$ is closed under linear combinations. We will extend the operator to the set $\mathcal{D}(I)$.

To prove Riesz' representation theorem, we will use that every indicator function of a closed intervals is a pointwise limit of a decreasing sequence of continuous functions, along with the extension as formulated in the following lemma.

LEMMA 2.12. *Suppose that $L\colon \mathcal{C}(I) \to \mathbb{R}$ is a continuous and linear functional. Then it is possible to extend the operator L to an operator $L\colon \mathcal{D}(I) \to \mathbb{R}$ such that the extended operator is linear and continuous and with the same norm, that is*

$$\|L\| = \sup_{\substack{f\in\mathcal{C}(I)\\ \|f\|\neq 0}} \frac{|L(f)|}{\|f\|} = \sup_{\substack{f\in\mathcal{D}(I)\\ \|f\|\neq 0}} \frac{|L(f)|}{\|f\|}.$$

Moreover, if $f \in \mathcal{D}(I)$ then, whenever $(f_n)_{n=1}^{\infty}$ is a decreasing sequence of continuous functions which converge pointwise to f, we have

$$L(f) = \lim_{n\to\infty} L(f_n).$$

PROOF. Suppose that $f \in \mathcal{D}(I)$ is the pointwise limit of the sequence f_n and that $f_1 \geq f_2 \geq \ldots$. Let

$$L(f) = \lim_{n\to\infty} L(f_n).$$

We are going to prove that the limit exists and only depends on f, not on which particular sequence f_n we use. (There are many different sequences of which f is the pointwise limit.)

We consider the sequence $(l_n)_{n=2}^{\infty}$, defined by

$$l_n = |L(f_1) - L(f_2)| + |L(f_2) - L(f_3)| + \ldots + |L(f_{n-1}) - L(f_n)|.$$

Clearly, l_n increases with n. Moreover, there are numbers $\sigma_k \in \{-1, 1\}$ such that $\sigma_k(L(f_{k-1}) - L(f_k)) = |L(f_{k-1}) - L(f_k)|$ and

$$\begin{aligned} l_n &= \sigma_2(L(f_1) - L(f_2)) + \sigma_3(L(f_2) - L(f_3)) + \ldots + \sigma_n(L(f_{n-1}) - L(f_n)) \\ &= L(\sigma_2(f_1 - f_2) + \sigma_3(f_2 - f_3) + \ldots + \sigma_n(f_{n-1} - f_n)) \\ &\leq \|L\| \sup_I |\sigma_2(f_1 - f_2) + \sigma_3(f_2 - f_3) + \ldots + \sigma_n(f_{n-1} - f_n)|. \end{aligned}$$

Since the sequence f_n is decreasing we have that $f_1 - f_2 \geq 0, \ldots, f_{n-1} - f_n \geq 0$. In particular we have

$$\sigma_2(f_1 - f_2) \leq f_1 - f_2, \quad \sigma_3(f_2 - f_3) \leq f_2 - f_3, \quad \sigma_4(f_3 - f_4) \leq f_3 - f_4, \quad \ldots$$

We may now write

$$\begin{aligned} l_n &\leq \|L\| \sup_{x\in I}(f_1(x) - f_2(x) + f_2(x) - f_3(x) + \ldots + f_{n-1}(x) - f_n(x)) \\ &= \|L\| \sup_{x\in I}(f_1(x) - f_n(x)) \leq \|L\| \sup_{x\in I}(f_1(x) - f(x)), \end{aligned}$$

and the quantity $\|L\| \sup_{x\in I}(f_1(x) - f(x))$ is bounded since f and f_1 are bounded functions. Hence the sequence (l_n) is bounded.

Since (l_n) is bounded and increasing, it has a limit. It follows that the sequence

$$L(f_n) = L(f_1) - \sum_{k=2}^{n} L(f_{k-1} - f_k))$$

is a Cauchy sequence, since for $m > n$

$$|L(f_m) - L(f_n)| = \left| \sum_{k=n+1}^{m} L(f_k - f_{k-1}) \right| \leq \sum_{k=n+1}^{\infty} |L(f_k - f_{k-1})|,$$

which converges to 0 as $n \to \infty$ since l_n has a limit. Since $L(f_n)$ is a Cauchy sequence, it has a limit, which we agree to call $L(f)$.

Suppose now that $(f_n)_{n=1}^{\infty}$ and $(g_n)_{n=1}^{\infty}$ are two decreasing sequences of continuous functions that both converge pointwise to f. We prove that

$$\lim_{n\to\infty} L(f_n) = \lim_{n\to\infty} L(g_n),$$

which means that $L(f)$ does not depend on the sequence $(f_n)_{n=1}^{\infty}$, and therefore is well-defined.

We will assume that the sequences are strictly decreasing, that is that we have $f_1 > f_2 > \ldots$ and $g_1 > g_2 > \ldots$. If this is not the case, we can simply replace f_n and g_n by $\tilde{f}_n = f_n + \frac{1}{n}$ and $\tilde{g}_n = g_n + \frac{1}{n}$.

We define a new decreasing sequence $(h_n)_{n=1}^{\infty}$ of continuous function in the following way.

Let $h_1 = f_1$. Since $f_1 > f$, and I is compact, there is a number $\varepsilon > 0$ such that $f_1 > f + \varepsilon$. Since g_n converges uniformly to f, there is a number k such that $\|f - g_k\| < \varepsilon$. Then $h_1 = f_1 > g_k > f$. Let $h_2 = g_k$.

Continuing in this manner, we obtain the sequence $(h_k)_{k=1}^{\infty}$. By construction this sequence is decreasing and it converges pointwise to f. Hence, from what we have proved above, $\lim_{k\to\infty} L(h_n)$ exists. Since the sequence $(h_k)_{k=1}^{\infty}$ contains infinitely many of the functions f_n as well as infinitely many of the functions g_n we must have

$$\lim_{n\to\infty} L(f_n) = \lim_{k\to\infty} L(h_k) = \lim_{n\to\infty} L(g_n).$$

This proves that $L(f)$ is well-defined.

Hence we have extended the operator L to functions that are pointwise limits of decreasing sequences of continuous functions. We will use this to extend L to an operator $L\colon \mathcal{D}(I) \to \mathbb{R}$ and we will prove that the extended L is linear, continuous and with unchanged norm.

If both f_1 and f_2 are pointwise limits of sequences of decreasing continuous functions, then so is $f_1 + f_2$. Clearly, in this case we have $L(f_1 + f_2) = L(f_1) + L(f_2)$.

If f is the pointwise limit of a decreasing sequence of continuous functions (f_n), then we have defined $L(f) = \lim L(f_n)$. If instead f is the pointwise limit of an increasing sequence of continuous functions, then $-f$ is the pointwise limit of a decreasing sequence of continuous functions, and we define $L(f) = -L(-f)$. This does not lead to any contradictions, since if f is both the pointwise limit of a decreasing sequence of continuous functions f_n as well as that of an increasing sequence of continueous functions g_n, then $0 = \lim(f_n - g_n)$, and

$$\begin{aligned} 0 = L(0) = \lim_{n\to\infty} L(f_n - g_n) &= \lim_{n\to\infty} \left(L(f_n) + L(-g_n)\right) \\ &= \lim_{n\to\infty} L(f_n) + \lim_{n\to\infty} L(-g_n), \end{aligned}$$

which shows that $\lim L(f_n) = -\lim L(-g_n)$.

First we note that $L(\lambda f) = \lambda L(f)$ holds whenever $f \in \mathcal{D}(I)$ and $\lambda \in \mathbb{R}$, and we will now define L on $\mathcal{D}(I)$.

If $f = f_1 - f_2$, where f_1 and f_2 are pointwise limits of decreasing sequences of continuous functions, then we define $L(f) = L(f_1) - L(f_2)$. This definition is well defined, since if $f = f_1 - f_2 = g_1 - g_2$, then $f_1 + g_2 = g_1 + g_2$ and so

$$L(f_1) + L(g_2) = L(f_1 + g_2) = L(g_1 + f_2) = L(g_1) + L(f_2).$$

Hence, $L(f_1) - L(f_2) = L(g_1) - L(g_2)$, so the definition of $L(f)$ does not depend on which representation of f as $f = f_1 - f_2$ or $f = g_1 - g_2$ that we use. We have therefore defined $L\colon \mathcal{D}(I) \to \mathbb{R}$ and we shall see that L is still linear.

Now, if $f = f_1 - f_2$ and $g = g_1 - g_2$, where f_1, f_2, g_1, g_2 are pointwise limits of decreasing sequences of continuous functions, then $f + g =$

$(f_1 + g_1) - (f_2 + g_2)$ and

$$\begin{aligned} L(f+g) &= L((f_1+g_1)-(f_2+g_2)) = L(f_1+g_1) - L(f_2+g_2) \\ &= L(f_1) + L(g_1) - L(f_2) - L(g_2) = L(f) + L(g). \end{aligned} \tag{2.4}$$

This proves that the extended operator L is linear.

Similary $f - g = (f_1 + g_2) - (f_2 + g_1)$ and

$$\begin{aligned} L(f-g) &= L((f_1+g_2)-(f_2+g_1)) = L(f_1+g_2) - L(f_2+g_1) \\ &= L(f_1) + L(g_2) - L(f_2) - L(g_1) = L(f) - L(g). \end{aligned} \tag{2.5}$$

These equalities will be used in the proof of Riesz' representation theorem.

Finally, we observe that the extended operator L still satisfies

$$|L(f)| \leq \|L\|\|f\| \tag{2.6}$$

where $\|L\|$ is the same number as before, but where $f \in \mathcal{D}(I)$. To see this assume first that f is the pointwise limit of a decreasing sequence of continuous functions f_n. Then $L(f)$ is the limit of $L(f_n)$ and $\|f\|$ is the limit of $\|f_n\|$, so that (2.6) follows by continuity from the inequality $|L(f_n)| \leq \|L\|\|f_n\|$ which holds by definition of $\|L\|$. More generaly, if $f = g - h \in \mathcal{D}(I)$ and g and h are both limits of decreasing sequences of continuous functions, then the same argument shows that (2.6) holds in this case as well. □

5.2. The proof of Riesz' representation theorem. After all the manipulations in the previous section, the reader might be disappointed to learn that it was not really necessary to explicitely extend the operator L as was done above. An alternative approach is to use the axiom of choice to extend the operator, which is done for instance in Shapiro's book [**12**]. Riesz himself, who was not in the possession of the axiom of choice, used the more explicit method that we used above, see the book by Riesz and Szőkefalvi-Nagy [**11**]. Readers of these notes that feel great uncomfort when using the axiom of choice, or are like Riesz not in possession of the mentioned axiom, are probably glad that the author of these notes follows Riesz' approach (or are at least greatly confused by this entire remark).

Proof of Riesz' representation theorem. To prove one direction of the equivalence is easy. Suppose that there is an α of bounded variation such that (2.3) holds. Clearly, L is then linear. By Lemma 2.7, we also have

$$|L(f)| \leq \operatorname{var}_I \alpha \|f\|,$$

so L is continuous.

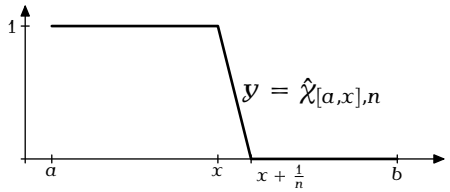
Suppose now that L is continuous and linear. By Lemma 2.12, we can extend L to bounded functions which are pointwise limits of a decreasing sequence of continuous functions.

Let $I = [a, b]$. If J is a subinterval of I, then the indicator function of J is denoted by χ_J and is defined to be 1 on J and 0 elsewhere. That is,

$$\chi_J(x) = \begin{cases} 1 & \text{if } x \in J, \\ 0 & \text{if } x \notin J. \end{cases}$$

We have that $\chi_{[a,x]}\colon I \to \mathbb{R}$ is the pointwise limit of the decreasing sequence $(\hat{\chi}_{[a,x],n})_{n=1}^{\infty}$, where $\hat{\chi}_{[a,x],n}\colon I \to \mathbb{R}$ is defined by

$$\hat{\chi}_{[a,x],n}(t) = \begin{cases} 1 & \text{if } t \le x, \\ 1 - nt & \text{if } x < t < x + \frac{1}{n}, \\ 0 & \text{if } t \ge x + \frac{1}{n}. \end{cases}$$



Hence, by Lemma 2.12, $L(\chi_{[a,x]})$ is defined, and

$$L(\chi_{[a,x]}) = \lim_{n\to\infty} L(\hat{\chi}_{[a,x],n}).$$

We define the function α by $\alpha(a) = 0$ and

$$\alpha(x) = L(\chi_{[a,x]}), \qquad x > a.$$

We shall prove that α is of bounded variation and that (2.3) holds.

To prove that α is of bounded variation, suppose that $a = x_0 < x_1 < \ldots < x_n = b$. Let

$$f = a_1\chi_{[a,x_1]} + a_2\chi_{(x_1,x_2]} + a_3\chi_{(x_2,x_3]} + \ldots + a_n\chi_{(x_{n-1},b]},$$

where

$$a_k = \begin{cases} 1 & \text{if } \alpha(x_k) - \alpha(x_{k-1}) > 0, \\ 0 & \text{if } \alpha(x_k) - \alpha(x_{k-1}) = 0, \\ -1 & \text{if } \alpha(x_k) - \alpha(x_{k-1}) < 0. \end{cases}$$

For $k > 1$, we then have

$$\begin{aligned} \alpha(x_k) - \alpha(x_{k-1}) &= L(\chi_{[a,x_k]}) - L(\chi_{[a,x_{k-1}]}) \\ &= L(\chi_{[a,x_k]} - \chi_{[a,x_{k-1}]}) = L(\chi_{(x_{k-1},x_k]}). \end{aligned}$$

For $k = 0$, since $\alpha(x_0) = \alpha(a) = 0$, we have

$$\alpha(x_1) - \alpha(x_0) = \alpha(x_1) = L(\chi_{[a,x_1]}) = L(\chi_{[x_0,x_1]}).$$

Hence, using (2.4) and (2.5) which says that the extended operator is linear, we have

$$\begin{aligned} L(f) &= a_1L(\chi_{[a,x_1]}) + a_2L(\chi_{(x_1,x_2]}) + a_3L(\chi_{(x_2,x_3]}) + \ldots + a_nL(\chi_{(x_{n-1},b]}) \\ &= a_1(\alpha(x_1) - \alpha(x_0)) + a_2(\alpha(x_2) - \alpha(x_1)) + \ldots + a_n(\alpha(x_n) - \alpha(x_0)) \\ &= |\alpha(x_1) - \alpha(x_0)| + |\alpha(x_2) - \alpha(x_1)| + \ldots + |\alpha(x_n) - \alpha(x_0)|. \end{aligned}$$

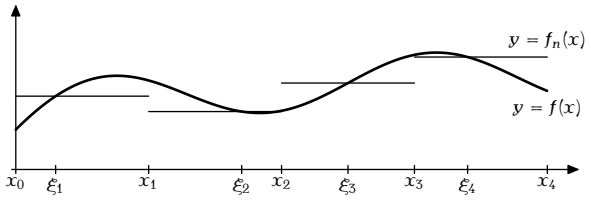
Since $\|f\| = 1$ or $\|f\| = 0$, this shows that

$$\sum_{k=1}^{n} |\alpha(x_k) - \alpha(x_{k-1})| = L(f) \le \|L\|\|f\| \le \|L\|.$$

Hence $\operatorname{var}_I \alpha \le \|L\|$. In particular, α is of bounded variation.

Let now f be a continuous function on $I = [a, b]$. Let $a = x_0 < x_1 < \ldots < x_n = b$ and let $\xi_k \in (x_{k-1}, x_k)$. Define f_n by $f(a) = f(\xi_1)$ and

$$f_n(x) = f(\xi_k), \qquad x \in (x_{k-1}, x_k).$$



Then

$$f_n = f(\xi_1)\chi_{[a,x_1]} + \sum_{k=2}^{n} f(\xi_k)\chi_{(x_{k-1},x_k]}$$

and

$$\begin{aligned} L(f_n) &= f(\xi_1)L(\chi_{[a,x_1]}) + \sum_{k=2}^{n} f(\xi_k)L(\chi_{(x_{k-1},x_k]}) \\ &= \sum_{k=1}^{n} f(\xi_k)(\alpha(x_k) - \alpha(x_{k-1})). \end{aligned}$$

By the definition of the Riemann–Stieltjes integral, we therefore have

$$\lim_{n\to\infty} L(f_n) = \int_I f \,\mathrm{d}\alpha.$$

But since f is continuous, $\|f - f_n\| \to 0$, as $n \to \infty$, and we have by (2.6) that

$$|L(f) - L(f_n)| \leq \|L\|\|f - f_n\| \to 0$$

as $n \to \infty$, which says that $\lim_{n\to\infty} L(f_n) = L(f)$. Hence

$$L(f) = \int_I f \,\mathrm{d}\alpha. \qquad \square$$

EXERCISE 2.7. Let $L\colon \mathcal{C}(I) \to \mathbb{R}$ be a continuous and linear functional given by $L(f) = \int_I f \,\mathrm{d}\alpha$. Prove that $\|L\| = \operatorname{var}_I \alpha$.

EXERCISE 2.8. Let $L\colon \mathcal{C}^r(I) \to \mathbb{R}$ be a continuous and linear function. Show that there are functions $\alpha_0, \alpha_1, \ldots, \alpha_r$ of bounded variation such that for every $f \in \mathcal{C}^r(I)$,

$$L(f) = \int f \,\mathrm{d}\alpha_0 + \int f' \,\mathrm{d}\alpha_1 + \ldots + \int f^{(r)} \,\mathrm{d}\alpha_r.$$

Using the norm

$$\|f\|_{\mathcal{C}^r(I)} = \sup_x |f(x)| + \sup_x |f'(x)| + \ldots + \sup_x |f^{(r)}(x)|$$

find an estimate from above of $\|L\|$ in terms of the variations of the functions $\alpha_0, \alpha_1, \ldots, \alpha_r$.

EXERCISE 2.9. Prove that if $L\colon \mathcal{C}([a,b]) \to \mathbb{R}$ is a continuous linear operator, then there exists a function α of bounded variation and a constant c such that such that

$$L(f) = cf(a) + \int_{[a,b]} \alpha \,\mathrm{d}f.$$

6. Lebesgue–Stieltjes integrals

In this section, we will very briefly discuss how to define a more general version of the Riemann–Stieltjes integral, which allows us to integrate more functions. Suppose that α is of bounded variation on a compact interval I.

According to the remark after Definition 2.4, it is possible get a more general integral by considering more general partitions. Another approach is to define a so called measure from the function α and then define the integral $\int_I f \, d\alpha$, through Lebesgue integration. We will not do any single detail of this, but we mention that this leads to what is called the Lebesgue[5]–Stieltjes integral $\int f \, d\alpha$, which is defined for more general functions f than the Riemann–Stieltjes integral is. Interested readers can for instance find more information and details in the book by Carter and van Brunt [**2**].

However, we have actually made a quite big step towards something as general as the Lebesgue–Stieltjes integral. The integral

$$\int_I f \, d\alpha$$

defines a linear and continuous functional $L\colon \mathcal{C}(I) \to \mathbb{R}$ according to Riesz' representation theorem. Hence according to Lemma 2.12, we may extend L to an operator $L\colon \mathcal{D}(I) \to \mathbb{R}$. We now define the integral $\int_I f \, d\alpha$ by

$$\int_I f \, d\alpha = L(f).$$

The integral $\int_I f \, d\alpha$ is then defined whenever $f \in \mathcal{D}(I)$ and α is of bounded variation, and it is equal to the Lebesgue–Stieltjes integral in this case. However, with the Lebesgue–Stieltjes integral it is possible to integrate even more general functions than those in $\mathcal{D}(I)$.

EXERCISE 2.10. Let $\alpha = 0$ on $[0,1]$ and 1 on $(1,2]$. Put $f = \alpha$ and let $g = 0$ on $[0,1)$ and $g = 1$ on $[1,2]$. Show that

$$\int_{[0,2]} g \, d\alpha = 1$$

as a Riemann–Stieltjes integral. What is the corresponding integral for f?

Calculate the integrals of f and g over the intervals $[0,1]$, $[1,2]$ and $[0,2]$ as Lebesgue–Stieltjes integrals (i.e. calculate them using the extended operator as described above.)

[5]Henri Lebesgue, 1875–1941. French mathematician.

CHAPTER 3

The Euler–Maclaurin summation formula

1. The Euler–Maclaurin summation formula

Let x be a real number. We recall that we can write x in a unique way as $x = [x] + \{x\}$ where $[x]$ is an integer called the integer part of x and $\{x\}$ is a number in $[0, 1)$, called the fractional part of x. The integer $[x]$ is the largest integer, not larger than x, and $\{x\} = x - [x]$.

Let f be a continuous function. Using the function $\alpha(x) = [x]$ we have (Exercise 2.5)

$$\sum_{k=a+1}^{b} f(k) = \int_a^b f \, d\alpha = \int_a^b f \, d[x].$$

The Euler–Maclaurin summation formula gives a very useful relation between the sum $\sum_{k=a}^{b} f(k)$ and the integral $\int_a^b f \, dx$.

THEOREM 3.1 (The Euler–Maclaurin[1] summation formula). *Suppose that f is continuously differentiable on $[a, b]$ and that a and b are integers. Then*

$$\sum_{k=a}^{b} f(k) = \int_a^b f(x) \, dx + \frac{f(a) + f(b)}{2} + \int_a^b f'(x)\Big(\{x\} - \frac{1}{2}\Big) \, dx.$$

PROOF. Integration by parts gives

$$\int_a^b f(x) \, d(x - [x]) = f(b)(b - [b]) - f(a)(a - [a]) - \int_a^b (x - [x]) \, df(x).$$

Hence

$$\int_a^b f(x) \, dx - \sum_{k=a+1}^{b} f(k) = -\int_a^b \{x\} f'(x) \, dx.$$

By adding the quantity

$$\frac{f(b) - f(a)}{2} = \int_a^b f'(x) \frac{1}{2} \, dx$$

to both sides, we obtain

$$\int_a^b f(x) \, dx - \sum_{k=a+1}^{b} f(k) + \frac{f(b) - f(a)}{2} = -\int_a^b f'(x)\Big(\{x\} - \frac{1}{2}\Big) \, dx,$$

from which the theorem follows. □

[1] Colin Maclaurin, 1698–1746. Scottish mathematician.

Note that we don't really need that f is continuously differentiable in the Euler–Maclaurin summation formula. It is only needed that f is differentiable and that the integral $\int_a^b f'(x)\big(\{x\} - \frac{1}{2}\big)\,\mathrm{d}x$ exists.

2. Stirling's formula

Stirling's[2] formula approximates the factorial of natural numbers. We will prove two versions of it, one more precise than the other.

The factorial of a natural number is defined by

$$0! = 1, \qquad n! = n(n-1)!.$$

Hence we have $n! = 1 \cdot 2 \ldots (n-1) \cdot n$.

We start with what is called Wallis' product formula.

LEMMA 3.2 (Wallis' product formula).

$$\frac{\pi}{2} = \lim_{n\to\infty} \prod_{k=1}^{n} \frac{2k}{2k-1}\frac{2k}{2k+1}.$$

PROOF. Let

$$I_n = \int_0^\pi \sin^n x\,\mathrm{d}x.$$

One checks that $I_0 = \pi$ and $I_1 = 2$.

Integration by parts yields the relation $I_n = \frac{n-1}{n} I_{n-2}$. Hence

$$\frac{I_n}{I_{n-2}} = \frac{n-1}{n} \qquad \text{and} \qquad \frac{I_{2n-1}}{I_{2n+1}} = \frac{2n+1}{2n}.$$

Clearly, $I_n \geq I_m$ if $n \leq m$. This implies that

$$1 \leq \frac{I_{2n}}{I_{2n+1}} \leq \frac{I_{2n-1}}{I_{2n+1}} = \frac{2n+1}{2n},$$

so that

$$\lim_{n\to\infty} \frac{I_{2n}}{I_{2n+1}} = 1.$$

But we also have

$$\frac{I_{2n}}{I_{2n+1}} = \frac{\frac{2n-1}{2n} I_{2n-2}}{\frac{2n}{2n+1} I_{2n-1}} = \frac{\frac{2n-1}{2n}\frac{2n-3}{2n-2} I_{2n-4}}{\frac{2n}{2n+1}\frac{2n-2}{2n-1} I_{2n-3}} = \frac{\frac{2n-1}{2n}\frac{2n-3}{2n-2}\cdots\frac{1}{2} I_0}{\frac{2n}{2n+1}\frac{2n-2}{2n-1}\cdots\frac{2}{3} I_1}$$

$$= \frac{\pi}{2}\frac{\frac{2n-1}{2n}\frac{2n-3}{2n-2}\cdots\frac{1}{2}}{\frac{2n}{2n+1}\frac{2n-2}{2n-1}\cdots\frac{2}{3}} = \frac{\pi}{2}\frac{\prod_{k=1}^{n}\frac{2k-1}{2k}}{\prod_{k=1}^{n}\frac{2k}{2k+1}} = \frac{\pi}{2}\prod_{k=1}^{n}\frac{2k-1}{2k}\frac{2k+1}{2k}.$$

Hence

$$\prod_{k=1}^{\infty}\frac{2k-1}{2k}\frac{2k+1}{2k} = \frac{2}{\pi} \qquad \text{and} \qquad \prod_{k=1}^{\infty}\frac{2k}{2k-1}\frac{2k}{2k+1} = \frac{\pi}{2}. \qquad \square$$

LEMMA 3.3. *Let* $a_n = \dfrac{n!}{\sqrt{n}\big(\frac{n}{e}\big)^n}$. *Then* $\lim\limits_{n\to\infty} \dfrac{a_n^2}{a_{2n}} = \sqrt{2\pi}$.

[2] James Stirling, 1692–1770. Scottish mathematician.

PROOF. We first record that

$$2^n n! = 2 \cdot 4 \cdot 6 \cdot \ldots \cdot (2n)$$

and

$$\begin{aligned} 2^{-n}(2n)! &= 3 \cdot 5 \cdot 7 \cdot \ldots \cdot (2n-1) \cdot 2^{-n} 2 \cdot 4 \cdot 6 \cdot \ldots \cdot (2n) \\ &= 3 \cdot 5 \cdot 7 \cdot \ldots \cdot (2n-1) \cdot n!. \end{aligned}$$

Combining these equalities, we get

$$\frac{2^{2n} n! n!}{(2n)!} = \frac{2 \cdot 4 \cdot 6 \cdot \ldots \cdot (2n)}{3 \cdot 5 \cdot 7 \cdot \ldots \cdot (2n-1)}.$$

We therefore have

$$\begin{aligned} \frac{a_n^2}{a_{2n}} &= \frac{n! n!}{(2n)!} \frac{\left(\frac{2}{e}\right)^{2n}}{\left(\frac{1}{e}\right)^{2n}} \sqrt{\frac{2}{n}} = \sqrt{\frac{2}{n}} \frac{2 \cdot 4 \cdot 6 \cdot \ldots \cdot (2n)}{3 \cdot 5 \cdot 7 \cdot \ldots \cdot (2n-1)} \\ &= \sqrt{\frac{2}{n}} \sqrt{\frac{2 \cdot 4 \cdot 6 \cdot \ldots \cdot (2n)}{3 \cdot 5 \cdot 7 \cdot \ldots \cdot (2n-1)} \frac{2 \cdot 4 \cdot 6 \cdot \ldots \cdot (2n)}{3 \cdot 5 \cdot 7 \cdot \ldots \cdot (2n-1)}} \\ &= \sqrt{\frac{2(2n+1)}{n}} \sqrt{\frac{2 \cdot 4 \cdot 6 \cdot \ldots \cdot (2n)}{3 \cdot 5 \cdot 7 \cdot \ldots \cdot (2n-1)} \frac{2 \cdot 4 \cdot 6 \cdot \ldots \cdot (2n)}{3 \cdot 5 \cdot 7 \cdot \ldots \cdot (2n-1)(2n+1)}}. \end{aligned}$$

By Wallis' product formula (Lemma 3.2) we get

$$\lim_{n \to \infty} \frac{a_n^2}{a_{2n}} = 2\sqrt{\frac{\pi}{2}} = \sqrt{2\pi}. \qquad \square$$

We shall now use the Euler–Maclaurin summation formula to obtain Stirling's formula.

THEOREM 3.4 (Stirling's formula). *We have*

$$n! = \sqrt{2\pi n} n^n e^{-n} (1 + \Phi(n)),$$

where Φ *satisfies*

$$\log(1 + \Phi(n)) = -\int_n^\infty \frac{1}{x^2} \frac{\{x\}^2 - \{x\}}{2} \, \mathrm{d}x \tag{3.1}$$

and

$$0 < \Phi(n) < e^{\frac{1}{8n}} - 1. \tag{3.2}$$

PROOF. To turn $n! = 1 \cdot 2 \ldots (n-1) \cdot n$ into something which can be treated with the Euler–Maclaurin summation formula, we apply a logarithm, and obtain

$$\log(n!) = \sum_{k=1}^{n} \log(k).$$

By the Euler–Maclaurin summation formula, we therefore have

$$\begin{aligned} \log(n!) &= \int_1^n \log x \, \mathrm{d}x + \int_1^n \frac{1}{x} \Big(\{x\} - \frac{1}{2} \Big) \, \mathrm{d}x + \frac{\log n}{2} \\ &= n \log n - n + 1 + \frac{1}{2} \log n + \int_1^n \frac{1}{x} \Big(\{x\} - \frac{1}{2} \Big) \, \mathrm{d}x. \end{aligned} \tag{3.3}$$

Let

$$c_n = \int_1^n \frac{1}{x}\Big(\{x\} - \frac{1}{2}\Big)\,dx \qquad \text{and} \qquad f(x) = \int_1^x \Big(\{t\} - \frac{1}{2}\Big)\,dt.$$

Integration by parts shows that

$$c_n = \Big[\frac{1}{x}f(x)\Big]_1^n + \int_1^n \frac{1}{x^2}f(x)\,dx.$$

The function f is bounded, and hence the limit

$$c = \lim_{n\to\infty} c_n = \int_1^\infty \frac{1}{x}\Big(\{x\} - \frac{1}{2}\Big)\,dx = \int_1^\infty \frac{1}{x^2}f(x)\,dx$$

exists.

We have proved in (3.3) that

$$n! = e^{1+c_n}\sqrt{n}\Big(\frac{n}{e}\Big)^n.$$

Hence

$$e^{1+c} = \lim_{n\to\infty} e^{1+c_n} = \lim_{n\to\infty} a_n.$$

Therefore $\lim_{n\to\infty} a_n$ exists and Lemma 3.3 implies that $e^{1+c} = \sqrt{2\pi}$. This shows that $n! \approx \sqrt{2\pi n}\big(\frac{n}{e}\big)^n$ in the sense that

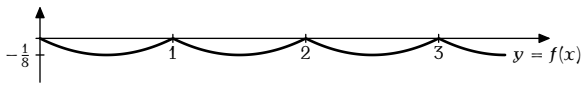
$$\lim_{n\to\infty} \frac{n!}{\sqrt{2\pi n}\big(\frac{n}{e}\big)^n} = 1.$$

Let $1 + \Phi(n) = \frac{n!}{\sqrt{2\pi n}(\frac{n}{e})^n} = \frac{e^{1+c_n}}{e^{1+c}} = e^{c_n - c} = e^{d_n}$, where

$$\begin{aligned} d_n = c_n - c &= \Big[\frac{1}{x}f(x)\Big]_1^n + \int_1^n \frac{1}{x^2}f(x)\,dx - \int_1^\infty \frac{1}{x^2}f(x)\,dx \\ &= \frac{f(n)}{n} - \int_n^\infty \frac{1}{x^2}f(x)\,dx = -\int_n^\infty \frac{1}{x^2}f(x)\,dx, \end{aligned}$$

since $f(n) = 0$ whenever n is an integer.

One easily calculates that $f(x) = \frac{\{x\}^2 - \{x\}}{2}$.



Hence, $-\frac{1}{8} \le f(x) \le 0$ and this implies that

$$0 < d_n < \frac{1}{8}\int_n^\infty \frac{1}{x^2}\,dx = \frac{1}{8n},$$

and

$$0 < \Phi(n) < e^{\frac{1}{8n}} - 1. \qquad \square$$

3. More on the Euler–Maclaurin summation formula

Since the exact value of the integral in (3.1) is not easy to obtain, Stirling's formula is usually only stated with estimates of Φ such as (3.2), rather than an exact but not very useful formula as in (3.1). It is clear however that if we can obtain better estimates on the integral in (3.1), then we can improve the estimate (3.2).

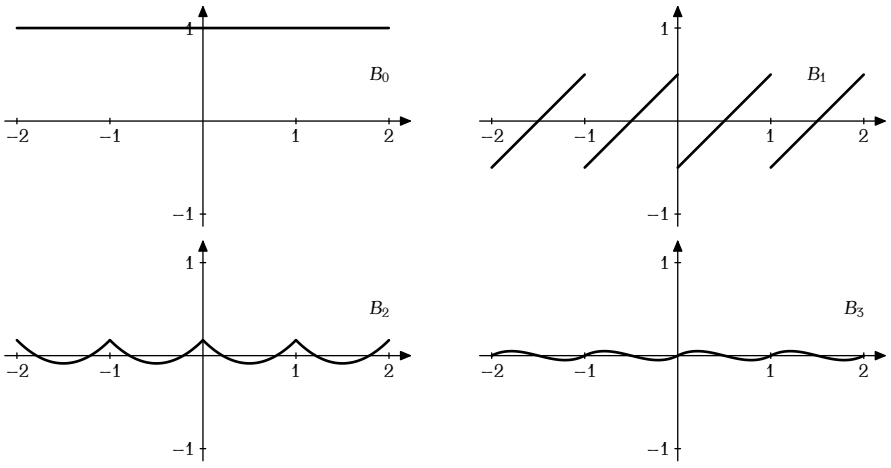
We are going to prove more precise versions of Theorems 3.1 and 3.4. The method is to gain better control over the integral (3.1) and

$$\int_a^b f'(x)\Big(\{x\} - \frac{1}{2}\Big)\,\mathrm{d}x$$

using integration by parts. To do this we need successive primitive functions of $\big(\{x\} - \frac{1}{2}\big)$.

We introduce the so called *Bernoulli*[3] *polynomials*, the first of which are given by

$$B_0(x) = 1, \qquad B_1(x) = x - \frac{1}{2},$$
$$B_2(x) = x^2 - x + \frac{1}{6}, \qquad B_3(x) = x^3 - \frac{3}{2}x^2 + \frac{1}{2}x.$$



They are determined by the properties that

$$B_0(x) = 1, \quad B_n'(x) = nB_{n-1}(x) \tag{3.4}$$

and

$$\int_0^1 B_n(x)\,\mathrm{d}x = 0, \qquad n > 0, \tag{3.5}$$

which imply that

$$\int_0^x B_n(\{x\})\,\mathrm{d}x = \int_0^{\{x\}} B_n(x)\,\mathrm{d}x = \frac{1}{n+1}(B_{n+1}(\{x\}) - B_{n+1}(0)), \tag{3.6}$$

[3]Jacob Bernoulli, 1655–1705. Swiss mathematician.

when $n > 0$. This means that the function $x \mapsto \frac{1}{n+1}B_{n+1}(\{x\})$ is a primitive function of $x \mapsto B_n(\{x\})$. In particular, we have

$$\int_a^b g(x)B_n(\{x\})\,dx = \frac{g(b)B_{n+1}(\{b\}) - g(a)B_{n+1}(\{a\})}{n+1} - \int_a^b g'(x)\frac{B_{n+1}(\{x\})}{n+1}\,dx. \tag{3.7}$$

The numbers $B_n(0)$ are called *Bernoulli numbers*. The first few Bernoulli numbers are

$$B_0(0) = 1, \quad B_1(0) = -\frac{1}{2}, \quad B_2(0) = \frac{1}{6}, \quad \textit{and} \quad B_3(0) = 0.$$

We can now state a more general version of the Euler–Maclaurin summation formula.

THEOREM 3.5 (The Euler–Maclaurin summation formula). *Suppose f is r times continuously differentiable on $[a,b]$ and that a and b are integers. Then*

$$\sum_{k=a}^{b} f(k) = \int_a^b f(x)\,dx + \frac{f(a)+f(b)}{2} + \sum_{k=1}^{[r/2]} \frac{B_{2k}(0)}{(2k)!}(f^{(2k-1)}(b) - f^{(2k-1)}(a)) + (-1)^{r+1}\int_a^b f^{(r)}(x)\frac{B_r(\{x\})}{r!}\,dx.$$

PROOF. We know by Theorem 3.1 that

$$\sum_{k=a}^{b} f(k) = \int_a^b f(x)\,dx + \frac{f(a)+f(b)}{2} + S_1 + R_1,$$

where S_1 and R_1 are given by

$$S_1 = 0, \qquad \text{and} \qquad R_1 = \int_a^b f'(x)\frac{B_1(\{x\})}{1!}\,dx.$$

The proof is of course by induction and integration by parts. Let

$$S_q = \sum_{k=2}^{q}(-1)^k\frac{B_k(0)}{k!}(f^{(k-1)}(b) - f^{(k-1)}(a))$$

and

$$R_q = (-1)^{q+1}\int_a^b f^{(q)}(x)\frac{B_q(\{x\})}{q!}\,dx,$$

which are compatible with the definitions of S_1 and R_1 above.

Suppose that we know that

$$\sum_{k=a}^{b} f(k) = \int_a^b f(x)\,dx + \frac{f(a)+f(b)}{2} + S_q + R_q, \tag{3.8}$$

holds for some $1 \le q < r$. (We know that it holds for $q = 1$.) Integrating by parts (formula (3.7)), we may write

$$\begin{aligned} R_q &= (-1)^{q+1} \int_a^b f^{(q)}(x) \frac{B_q(\{x\})}{q!}\,dx \\ &= \left[(-1)^{q+1} f^{(q)}(x) \frac{B_{q+1}(\{x\})}{(q+1)!}\right]_a^b + (-1)^{q+2} \int_a^b f^{(q+1)}(x) \frac{B_{q+1}(\{x\})}{(q+1)!}\,dx \\ &= (-1)^{q+1} \frac{B_{q+1}(0)}{(2q)!}(f^{(q)}(b) - f^{(q)}(a)) + R_{q+1}. \end{aligned}$$

This shows that

$$S_q + R_q = S_{q+1} + R_{q+1}.$$

Hence

$$\sum_{k=a}^{b} f(k) = \int_a^b f(x)\,dx + \frac{f(a)+f(b)}{2} + S_{q+1} + R_{q+1},$$

and by induction, this shows that (3.8) holds for all $1 \le q \le r$. Now, since $B_k(0) = 0$ whenever k is odd and larger than 1 (Exercise 3.2), we can write S_r as in the theorem. □

EXERCISE 3.1. Prove (3.6) using (3.4) and (3.5).

EXERCISE 3.2. Prove that $B_k(0) = 0$ whenever k is odd and larger than 1.

EXERCISE 3.3. Use the Euler–Maclaurin summation formula to find a good approximation of

$$\sum_{k=1}^{n} k^s.$$

4. An improved version of Stirling's formula

We now apply Theorem 3.5 to get the following improvement of Stirling's formula.

THEOREM 3.6 (Stirling's formula). *We have*

$$n! = \sqrt{2\pi n} n^n e^{-n} \Psi(n),$$

where Ψ *satisfies*

$$\left|\Psi(n) - \sum_{k=1}^{[r/2]} \frac{B_{2k}(0)}{(2k-1)2k} n^{-(2k-1)}\right| \le \frac{(r-2)!}{2^{r-1}} n^{-(r-1)} \tag{3.9}$$

for any natural number r.

PROOF. By Theorem 3.5,

$$\log n! = \int_1^n \log(x)\,dx + \frac{\log n}{2} + S_r + R_r,$$

where

$$S_r = \sum_{k=1}^{[r/2]} \frac{B_{2k}(0)}{(2k)!}\Big(\frac{(2k-2)!}{n^{2k-1}} - (2k-2)!\Big) = \sum_{k=1}^{[r/2]} \frac{B_{2k}(0)}{(2k-1)(2k)}(n^{-(2k-1)} - 1),$$

$$R_r = (-1)^{r+1}\int_1^n (-1)^{r+1}\frac{(r-1)!}{x^r}\frac{B_r(\{x\})}{r!}\,\mathrm{d}x = \int_1^n \frac{B_r(\{x\})}{rx^r}\,\mathrm{d}x.$$

Hence

$$n! = \sqrt{2\pi n}n^n\mathrm{e}^{-n}\Psi(n)$$

with

$$\log\Psi(n) = -\frac{1}{2}\log(2\pi) + 1 + S_r + R_r.$$

Letting

$$c = 1 - \sum_{k=1}^{[r/2]} \frac{B_{2k}(0)}{(2k-1)2k} + \int_1^\infty \frac{B_r(\{x\})}{rx^r}\,\mathrm{d}x,$$

we have

$$\log\Psi(n) = -\frac{1}{2}\log(2\pi) + c + \sum_{k=1}^{[r/2]} \frac{B_{2k}(0)}{(2k-1)2k}n^{-(2k-1)} - \int_n^\infty \frac{B_r(\{x\})}{rx^r}\,\mathrm{d}x.$$

By Theorem 3.4, we know that $\Psi(n) \to 1$ as $n \to \infty$. Hence we must have $-\frac{1}{2}\log(2\pi) + c = 0$ and

$$\log\Psi(n) = \sum_{k=1}^{[r/2]} \frac{B_{2k}(0)}{(2k-1)2k}n^{-(2k-1)} - \int_n^\infty \frac{B_r(\{x\})}{rx^r}\,\mathrm{d}x.$$

By induction, one proves that

$$\max_{x\in[0,1]}|B_r(x)| \le \frac{r!}{2^{r-1}}.$$

Hence

$$\left|\int_n^\infty \frac{B_r(\{x\})}{rx^r}\,\mathrm{d}x\right| \le \frac{(r-1)!}{2^{2-r}}\int_n^\infty \frac{1}{x^r}\,\mathrm{d}x = \frac{(r-2)!}{2^{r-1}}n^{-(r-1)}.$$

This finishes the proof. □

EXERCISE 3.4. Show that Stirling's formula can be written in the form

$$n! = \sqrt{2\pi n}n^n\mathrm{e}^{-n}\left(\frac{1}{12n} - \frac{1}{360n^3} + \frac{1}{1260n^5} - \frac{1}{1680n^7} + \dots\right).$$

5. Applications of Stirling's formula

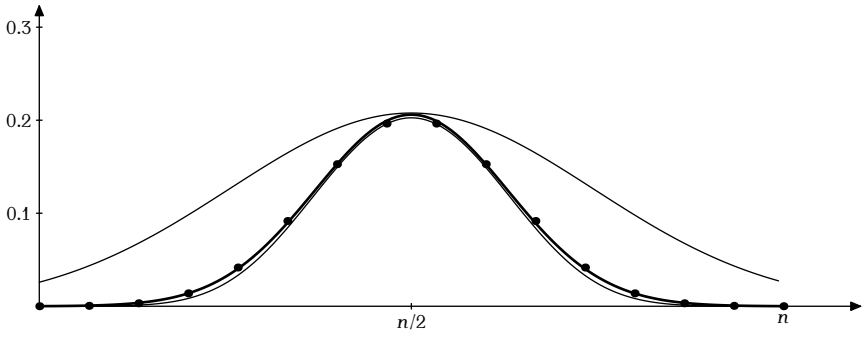
5.1. Binomial coefficients. We use Stirling's formula to estimate the size of the *binomial coefficient* $\binom{n}{k}$ for large n.

THEOREM 3.7. *If $n \geq 10$ then*

$$e^{-\frac{1}{4n}\left(\frac{k-\frac{n}{2}}{\frac{\sqrt{n}}{2}}\right)^4-\frac{1}{4n}} \leq \frac{\binom{n}{k}}{2^n \frac{1}{\sqrt{2\pi}}\frac{2}{\sqrt{n}} e^{-\frac{1}{2}\left(\frac{k-\frac{n}{2}}{\frac{\sqrt{n}}{2}}\right)^2}} \leq e^{\frac{1}{8n}+\frac{2\log n}{n}\left(\frac{k-\frac{n}{2}}{\frac{\sqrt{n}}{2}}\right)^2} \tag{3.10}$$

for all integers k with $0 \leq k \leq n$.

For illustration, the picture below shows the graph of the function $x \mapsto \frac{1}{\sqrt{2\pi}}\frac{2}{\sqrt{n}} e^{-\frac{1}{2}\left(\frac{x-\frac{n}{2}}{\frac{\sqrt{n}}{2}}\right)^2}$ (dark line) and the values of $2^{-n}\binom{n}{k}$ (dots) when $n = 15$. The thin lines are the upper and lower bound on $2^{-n}\binom{n}{k}$ which we obtain from Theorem 3.7.



PROOF. We start by proving that the rightmost inequality of (3.10) holds when $k = 0$ and $k = n$. In this case it is sufficient to prove that

$$1 \leq 2^n \frac{1}{\sqrt{2\pi}}\frac{2}{\sqrt{n}} e^{-\frac{n}{2}+\frac{1}{8n}}. \tag{3.11}$$

Since $-\log n \geq -1 - \frac{1}{2n}$, we have

$$\begin{aligned} 2^n \frac{1}{\sqrt{2\pi}}\frac{2}{\sqrt{n}} e^{-\frac{n}{2}+\frac{1}{8n}} &= \sqrt{\frac{2}{\pi}} e^{-\frac{n}{2}+n\log 2-\frac{1}{2}\log n+\frac{1}{8n}} \\ &\geq \sqrt{\frac{2}{\pi}} e^{-\frac{n}{2}+n\log 2-1-\frac{1}{8}n} \\ &= \sqrt{\frac{2}{\pi}} e^{-1+n(\log 2-\frac{10}{16})}. \end{aligned}$$

But

$$\log 2 = -\log(1-\frac{1}{2}) = \sum_{k=1}^{\infty} \frac{1}{k2^k} > \frac{1}{2}+\frac{1}{8} = \frac{10}{16},$$

so (3.11) holds whenever

$$-1 + n(\log 2 - \frac{9}{16}) > \frac{1}{2}(\log \pi - \log 2) \quad \Leftrightarrow \quad n > \frac{2+\log\pi-\log 2}{2\log 2-\frac{9}{8}} \approx 9.3.$$

We now consider the case when $1 \le k \le n-1$. By Stirling's formula we have

$$e^{-\frac{1}{4n}} \le \frac{\binom{n}{k}}{\frac{1}{\sqrt{2\pi}}\sqrt{\frac{n}{(n-k)k}}\frac{n^n}{k^k(n-k)^{n-k}}} \le e^{\frac{1}{8n}}.$$

Hence, it sufficies to prove that

(3.12)

$$\frac{1}{\sqrt{2\pi}}\sqrt{\frac{n}{(n-k)k}}\frac{n^n}{k^k(n-k)^{n-k}}e^{\frac{1}{8n}} \le 2^n\frac{1}{\sqrt{2\pi}}\frac{2}{\sqrt{n}}e^{-\frac{1}{2}(1-\frac{1}{2n})\left(\frac{k-\frac{n}{2}}{\frac{\sqrt{n}}{2}}\right)^2+\frac{1}{8n}}$$

where $c_n = \frac{2\log n}{n}$, and

(3.13)

$$\frac{1}{\sqrt{2\pi}}\sqrt{\frac{n}{(n-k)k}}\frac{n^n}{k^k(n-k)^{n-k}}e^{-\frac{1}{4n}} \ge 2^n\frac{1}{\sqrt{2\pi}}\frac{2}{\sqrt{n}}e^{-\left(\frac{k-\frac{n}{2}}{\frac{n}{\sqrt{2}}}\right)^2-\frac{1}{4n}-\frac{1}{4n}\left(\frac{k-\frac{n}{2}}{\frac{\sqrt{n}}{2}}\right)^4}$$

holds for all integers k with $0 \le k \le n$ when n is large.

We first prove (3.12). Letting $k = xn$, with $x \in [1/n, 1-1/n]$, we can rewrite (3.12) as

$$-\frac{1}{2}\Big(\log x + \log(1-x)\Big) - \log 2 - 2\log n\Big(x-\frac{1}{2}\Big)^2 \le n\Big(x\log x + (1-x)\log(1-x) + \log 2 - 2\Big(x-\frac{1}{2}\Big)^2\Big).$$

Put

$$\begin{aligned}\phi_n(x) &= -\frac{1}{2}\Big(\log x + \log(1-x)\Big) - \log 2 - 2\log n\Big(x-\frac{1}{2}\Big)^2\\ &= -\frac{1}{2}\log\Big(1-4\Big(x-\frac{1}{2}\Big)^2\Big) - \frac{1}{2}\log n\cdot 4\Big(x-\frac{1}{2}\Big)^2,\\ \psi(x) &= x\log x + (1-x)\log(1-x) + \log 2 - 2\Big(x-\frac{1}{2}\Big)^2,\end{aligned}$$

so that (3.12) can be written as $\phi_n(x) \le n\psi(x)$.

Let us consider the inequality $\phi_n(x) \le 0$. Letting $t = 4\big(x-\frac{1}{2}\big)^2 \in [0, \frac{1}{4}]$ this inequality can be written as

$$-\log\Big(1-t\Big) \le \log n\cdot t.$$

From the fact that $t \mapsto -\log(1-t)$ is a convex function follows that

$$-\log(1-t_0) \le \log n\cdot t_0 \qquad \Rightarrow \qquad -\log(1-t) \le \log n\cdot t \quad \text{for all } t < t_0.$$

Let $t_0 = 1-\frac{8}{n}$. If $1/n \le x \le 1-1/n$, then $0 \le t \le 4\big(\frac{1}{n}-\frac{1}{2}\Big)^2 = 1-\frac{4}{n}+\frac{4}{n^2}$, so that $t \le t_0 = 1-\frac{8}{n}$ when $n \ge 2$. For this choise of t_0 and with $n \ge 8$ we have $-\log(1-t_0) \le \log n\cdot t_0$, since

$$-\log(1-t_0) = \log n - \log 8, \qquad \text{and}$$

$$\log n\cdot t_0 = \log n - 8\frac{\log n}{n}.$$

Hence we have proved that if $n \ge 4$, then $\phi_n(x) \le 0$ for all $1/n \le x \le 1-1/n$.

We now consider ψ. Taylor's[4] formula implies that

$$\psi(x) = \frac{\psi^{(4)}(\xi)}{4!}\Big(x - \frac{1}{4}\Big)^4$$

for some $\xi \in [0,1]$. Since $\psi^{(4)}(x) = 2(x^{-3} + (1-x)^{-3})$, we have $\psi^{(4)}(\xi) \geq 4$ and so $\psi(x) \geq \frac{1}{6}(x - \frac{1}{2})^4$ if $0 \leq x \leq 1$. In particular ψ is non-negative.

We have now proved (3.12) when $n \geq 4$, since (3.12) can be written as $\phi(x) \leq n\psi(x)$ and we have proved that ϕ is non-positive if $n \geq 4$ and ϕ is non-negative.

It now remains to prove (3.13). Similarly to the proof of (3.12), we can write (3.13) as

$$\begin{aligned} &-\frac{1}{2}\Big(\log x + \log(1-x)\Big) - \log 2 \\ &\geq n\Big(x\log x + (1-x)\log(1-x) + \log 2 - 2\Big(x - \frac{1}{2}\Big)^2 - 4\Big(x - \frac{1}{2}\Big)^4\Big), \end{aligned}$$

or equivalently, as

$$\frac{1}{2}\log\Big(1 - 4\Big(x - \frac{1}{2}\Big)^2\Big) \geq n\Big(\psi(x) - 4\Big(x - \frac{1}{2}\Big)^4\Big).$$

But $\phi(x) \geq 0$ and $\psi(x) - 4\Big(x - \frac{1}{2}\Big)^4 \leq 0$, so the inequality is satisfied for any $x \in [0,1]$ and any n. Hence (3.13) holds for all k with $0 \leq k \leq n$. □

5.2. The Central Limit Theorem. Suppose that we generate randomly, for instance by tossing a coin, a sequence $\{x_k\}_{k=1}^{\infty}$ of numbers in $\{0,1\}$, such that for each x_k the probability that x_k is 0 is equal to the probability that x_k is 1, and such that the numbers x_k and x_l are independent.

In probabilistic language, we have a sequence of independent and identically distributed random variables X_k such that

$$\mathrm{P}(X_k = 0) = \mathrm{P}(X_k = 1) = \frac{1}{2},$$

where P denotes probability. The *expected value* or *expectation* of the random variable X_k is then $\mathrm{E}\,X_k = \frac{1}{2}$. (The name "expected value" can be confusing, since the expected value is something you definitely do not expect; mean value is a better name for the same thing.)

Let us form the sum $s_n = x_1 + x_2 + \ldots + x_n$. In probabilistic language, the value of s_n is a random variable S_n and $S_n = X_1 + X_2 + \ldots + X_n$. We are interested in the distribution of S_n, that is the probability that s_n lies in some given interval.

An important and central theorem in the theory of probability is the so called *law of large numbers*. It comes in a weak and a strong version. The weak version states that if $S_n = X_1 + X_2 + \ldots + X_n$ is the sum of the random variables $X_1, X_2, \ldots, X_n$, that are independent and identically distributed, with mean μ, then for any $\varepsilon > 0$ the probability that $|\frac{1}{n}S_n - \mu| \geq \varepsilon$ converges to 0 as $n \to \infty$. Hence one should expect that $\frac{1}{n}S_n$ is close to μ if n is large. For the coin tossing, discussed above,

[4]Brook Taylor, 1685–1731. English mathematician.

this means that if n is large, then the probability that the proportion of 0 or 1 deviates more than ε from $\frac{1}{2}$ is small.

One might be interested in how $\frac{1}{n}S_n$ deviates from μ. Suppose that $X_1, X_2, \ldots$ are independent and identically distributed, with mean μ and *variance* $\sigma^2 := \mathrm{E}((X_k - \mu)^2)$. The *central limit theorem* states that the distribution of $\sqrt{n}(\frac{1}{n}S_n - \mu)$ converges to a normal distribution with mean 0 and variance σ^2. That is

$$\mathrm{P}\Big(\sqrt{n}(\frac{1}{n}S_n - \mu) \in [a,b]\Big) \to \int_a^b \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\frac{x^2}{2\sigma^2}}\,\mathrm{d}x, \qquad n \to \infty.$$

We will prove the central limit theorem in the special case that $\mathrm{P}(X_k = 0) = \mathrm{P}(X_k = 1) = \frac{1}{2}$ as described above. *Warning!* The proof will be based on Theorem 3.7 and the Euler–Maclaurin summation formula. The estimates will be elementary, but somewhat messy. There are several different and better ways to prove the central limit theorem, so the reader is recommended to actually not really read the proof presented here.

It is sufficent to study the distribution function of S_n defined by

$$F_n(a) = \mathrm{P}(S_n \le a).$$

Suppose k is an integer in $[0,n]$. Then the probability that $S_n = k$ is given by

$$\mathrm{P}(S_n = k) = 2^{-n}\binom{n}{k}$$

since there are exactly $\binom{n}{k}$ different ways in which the sum $x_1 + x_2 + \ldots + x_n$ can be k, each of which has probability 2^{-n}. Hence

$$F_n(a) = \sum_{0 \le k \le a} 2^{-n}\binom{n}{k}.$$

To prove a central limit theorem for S_n, we want to relate the probability $\mathrm{P}\big(\sqrt{n}(\frac{1}{n}S_n - \mu) \le a\big)$ to F_n and write

$$\mathrm{P}\Big(\sqrt{n}(\frac{1}{n}S_n - \mu) \le a\Big) = \mathrm{P}(S_n \le n\mu + a\sqrt{n}) = F_n(n\mu + a\sqrt{n}).$$

Using Theorem 3.7 we can estimate $F_n(n\mu + a\sqrt{n})$ which leads to the following theorem.

Corollary 3.8 (The de Moivre[5]–Lagrange[6] central limit theorem). *If $S_n = X_1 + X_2 + \ldots + X_n$ and $(X_k)_{k=1}^{\infty}$ are independent and such that*

$$\mathrm{P}(X_k = 0) = \mathrm{P}(X_k = 1) = \frac{1}{2},$$

then the distribution of $\sqrt{n}(\frac{1}{n}S_n - \frac{1}{2})$ converges to a normal distribution with variance $\frac{1}{4}$ as $n \to \infty$. More precisely, there is a constant c such that

$$\left|\mathrm{P}\Big(\sqrt{n}\Big(\frac{1}{n}S_n - \frac{1}{2}\Big) \le a\Big) - \int_{-\infty}^{a} \sqrt{\frac{2}{\pi}}\mathrm{e}^{-2t^2}\,\mathrm{d}t\right| \le \frac{c}{\sqrt{n}}.$$

[5]Abraham de Moivre, 1667–1754. French mathematician.

[6]Joseph-Louis Lagrange, 1736–1813. Italian mathematician.

PROOF. We will make use of the equality

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\alpha x^2}\,dx = \frac{1}{\sqrt{2a}}, \qquad a > 0. \tag{3.14}$$

We first prove the bound from above. Suppose $a \geq 0$. Let

$$f_n(t) = \frac{1}{\sqrt{2\pi}} \frac{2}{\sqrt{n}} e^{-(\frac{1}{2} - \frac{2\log n}{n})\left(\frac{t-\frac{n}{2}}{\frac{\sqrt{n}}{2}}\right)^2 + \frac{1}{8n}}.$$

By Theorem 3.7, we have

$$\begin{aligned} F_n(\tfrac{n}{2} + a\sqrt{n}) &= \sum_{k=0}^{\frac{n}{2}+a\sqrt{n}} 2^{-n}\binom{n}{k} \leq \sum_{k=0}^{\frac{n}{2}+a\sqrt{n}} f_n(k) \leq \sum_{k=-\infty}^{\frac{n}{2}+a\sqrt{n}} f_n(k) \\ &= \int_{-\infty}^{\frac{n}{2}+a\sqrt{n}} f_n(t)\,d[t] = \int_{-\infty}^{\frac{n}{2}+a\sqrt{n}} f_n(t)\,dt + \Delta_1, \end{aligned}$$

where Δ_1 satisfies

$$\Delta_1 = \frac{f_n(\frac{n}{2} + a\sqrt{n})}{2} + \int_{-\infty}^{\frac{n}{2}+a\sqrt{n}} f_n'(t) B_1(\{t\})\,dt$$

by the Euler–Maclaurin summation formula. By the change of variables $s = \frac{t-\frac{n}{2}}{\sqrt{n}}$ we have

$$F_n(\tfrac{n}{2} + a\sqrt{n}) \leq \int_{-\infty}^{a} \sqrt{\frac{2}{\pi}} e^{-2s^2 + \frac{8\log n}{n}s^2 + \frac{1}{8n}}\,ds + \Delta_1. \tag{3.15}$$

The derivative f_n' is positive on $(-\infty, \frac{n}{2})$ and negative on $(\frac{n}{2}, \infty)$. Since $B_1(\{x\}) = \{x\} - \frac{1}{2}$, we have $|B_1(\{x\})| \leq \frac{1}{2}$ for all x. We may therefore estimate that

$$\begin{aligned} \int_{-\infty}^{a} f_n'(t) B_1(\{t\})\,dt &= \int_{-\infty}^{0} f_n'(t) B_1(\{t\})\,dt + \int_{\frac{n}{2}}^{a} f_n'(t) B_1(\{t\})\,dt \\ &\leq \frac{1}{2}\int_{-\infty}^{\frac{n}{2}} f_n'(t)\,dt + \left|\frac{1}{2}\int_{0}^{a} f_n'(t)\,dt\right| \\ &= \frac{1}{2} f_n\Big(\frac{n}{2}\Big) + \frac{1}{2}\Big|f_n(a) - f_n\Big(\frac{n}{2}\Big)\Big| \leq f_n\Big(\frac{n}{2}\Big) + \frac{1}{2} f_n(a) \\ &\leq \frac{3}{2} f_n\Big(\frac{n}{n}\Big). \end{aligned}$$

Hence

$$\Delta_1 \leq \frac{f_n(a)}{2} + \frac{3}{2} f_n\Big(\frac{n}{n}\Big) \leq 2 f_n\Big(\frac{n}{2}\Big) = \frac{4}{\sqrt{2\pi n}} e^{\frac{1}{8n}}.$$

Finally, by (3.15), we have

$$F_n(\tfrac{n}{2} + a\sqrt{n}) - \int_{-\infty}^{a} \sqrt{\frac{2}{\pi}} e^{-2s^2}\,dt \leq \Delta_1 + A,$$

where

$$\begin{aligned} A &= \int_{-\infty}^{a} e^{-(\frac{1}{2}-\frac{2\log n}{n})s^2+\frac{1}{8n}}\,ds - \int_{-\infty}^{a} e^{-\frac{1}{2}s^2}\,ds \\ &\le \int_{-\infty}^{\infty} e^{-(\frac{1}{2}-\frac{2\log n}{n})s^2+\frac{1}{8n}}\,ds - \int_{-\infty}^{\infty} e^{-\frac{1}{2}s^2}\,ds \\ &= \frac{e^{\frac{1}{8n}}}{\sqrt{1-\frac{4\log n}{n}}} - 1. \end{aligned}$$

We used (3.14) in the last step. We have thus proved that

$$F_n(\tfrac{n}{2}+a\sqrt{n}) - \int_{-\infty}^{a}\sqrt{\frac{2}{\pi}}e^{-2t^2}\,dt \le 2\sqrt{\frac{2}{\pi n}}e^{\frac{1}{8n}} + \frac{e^{\frac{1}{8n}}}{\sqrt{1-\frac{4\log n}{n}}} - 1.$$

We will now prove the bound from below. Let

$$g_n(t) = \frac{1}{\sqrt{2\pi}}\frac{2}{\sqrt{n}}e^{-\frac{1}{2}\left(\frac{t-\frac{n}{2}}{\frac{\sqrt{n}}{2}}\right)^2-\frac{1}{4n}\left(\frac{t-\frac{n}{2}}{\frac{\sqrt{n}}{2}}\right)^4-\frac{1}{4n}}.$$

By Theorem 3.7, we have

$$\begin{aligned} F_n(\tfrac{n}{2}+a\sqrt{n}) &= \sum_{k=0}^{\frac{n}{2}+a\sqrt{n}} 2^{-n}\binom{n}{k} \ge \sum_{k=0}^{\frac{n}{2}+a\sqrt{n}} g_n(k) = \int_{-1}^{\frac{n}{2}+a\sqrt{n}} g_n(t)\,d[t] \\ &= \int_{-1}^{\frac{n}{2}+a\sqrt{n}} g_n(t)\,dt + \Delta_2 \\ &= \int_{-\infty}^{\frac{n}{2}+a\sqrt{n}} g_n(t)\,dt + \Delta_2 - \int_{-\infty}^{-1} g_n(t)\,dt, \end{aligned}$$

where

$$\Delta_2 = \frac{g_n(a)}{2} + \int_{-1}^{a} g_n'(t)B_1(\{t\})\,dt.$$

Again, with the change of variable $s = \frac{t-\frac{n}{2}}{\sqrt{n}}$, we have

$$F_n(\tfrac{n}{2}+a\sqrt{n}) \ge \int_{-\infty}^{a}\sqrt{\frac{2}{\pi}}e^{-2s^2-\frac{4}{n}s^4-\frac{1}{4n}}\,ds + \Delta_2 - \int_{-\infty}^{-\frac{\sqrt{n}}{2}}\sqrt{\frac{2}{\pi}}e^{-2s^2}\,ds. \tag{3.16}$$

In the same way as for Δ_1 we get

$$|\Delta_2| \le 2g_n(0) = \frac{4}{\sqrt{2\pi n}}.$$

Since $2s^2 \ge |s|$ when $s \le -\frac{\sqrt{n}}{2} \le -\frac{1}{2}$ we have

$$\int_{-\infty}^{-\frac{\sqrt{n}}{2}}\sqrt{\frac{2}{\pi}}e^{-2s^2}\,d \le \int_{-\infty}^{-\frac{\sqrt{n}}{2}}\sqrt{\frac{2}{\pi}}e^{-s}\,ds = \sqrt{\frac{2}{\pi}}e^{-\frac{\sqrt{n}}{2}}.$$

Finally, by (3.16) we have

$$F_n(\tfrac{n}{2}+a\sqrt{n}) - \int_{-\infty}^{a}\sqrt{\frac{2}{\pi}}e^{-2s^2}\,ds \ge \Delta_2 + B,$$

where

$$\begin{aligned}
B &= \int_{-\infty}^{a} \sqrt{\frac{2}{\pi}} e^{-2s^2 - \frac{4}{n}s^4 - \frac{1}{4n}} \, ds - \int_{-\infty}^{a} \sqrt{\frac{2}{\pi}} e^{-2s^2} \, ds \\
&\geq \int_{-\infty}^{\infty} \sqrt{\frac{2}{\pi}} e^{-2s^2 - \frac{4}{n}s^4 - \frac{1}{4n}} \, ds - \int_{-\infty}^{\infty} \sqrt{\frac{2}{\pi}} e^{-2s^2} \, ds \\
&\geq e^{-\frac{1}{4n}} \int_{-n^{1/8}}^{n^{1/8}} \sqrt{\frac{2}{\pi}} e^{-2s^2 - \frac{4}{n}s^4} \, ds - 1 \\
&\geq e^{-\frac{1}{4n} - \frac{4}{\sqrt{n}}} \int_{-n^{1/8}}^{n^{1/8}} \sqrt{\frac{2}{\pi}} e^{-2s^2} \, ds - 1 \\
&= e^{-\frac{1}{4n} - \frac{4}{\sqrt{n}}} \left(1 - 2 \int_{n^{1/8}}^{\infty} \sqrt{\frac{2}{\pi}} e^{-2s^2} \, ds \right) - 1 \\
&\geq e^{-\frac{1}{4n} - \frac{4}{\sqrt{n}}} \left(1 - 2 \int_{n^{1/8}}^{\infty} \sqrt{\frac{2}{\pi}} e^{-2s} \, ds \right) - 1 \\
&= e^{-\frac{1}{4n} - \frac{4}{\sqrt{n}}} \left(1 - \sqrt{\frac{2}{\pi}} e^{-2n^{1/4}} \right) - 1.
\end{aligned}$$

Combining the estimates of Δ_2 and B, we have

$$F_n(\tfrac{n}{2} + a\sqrt{n}) - \int_{-\infty}^{a} \sqrt{\frac{2}{\pi}} e^{-2s^2} \, ds \geq -\frac{4}{\sqrt{2\pi n}} + e^{-\frac{1}{4n} - \frac{4}{\sqrt{n}}} \left(1 - \sqrt{\frac{2}{\pi}} e^{-2n^{1/4}} \right) - 1.$$

Using for instance Taylor expansions, it is not difficult to prove that our estimates from above and below, imply that

$$\left| F_n(\tfrac{n}{2} + a\sqrt{n}) - \int_{-\infty}^{a} \sqrt{\frac{2}{\pi}} e^{-2s^2} \, ds \right| \leq \frac{c}{\sqrt{n}}.$$

We refrain from trying to find an explicit such constant, but mention that we may choose $c = 7$ if we require that $n \geq 12$. □

EXERCISE 3.5. Let $0 \leq s \leq n$. Consider the set $\{0,1\}^n$ of sequences of zeroes and ones of length n. Estimate the proportion of such sequences which contains at least sn zeroes.

EXERCISE 3.6. A so called Einstein[7] solid consists of N oscillators, each of which can hold a discrete number of energy units. If the entire solid holds q units of energy, then this energy can be distributed in

$$\Omega = \binom{N + q - 1}{q}$$

different ways. If $q \gg N$, derive the approximation

$$\Omega \approx \left(\frac{eq}{N} \right)^N.$$

[7] Albert Einstein, 1879–1955. German-American physicist

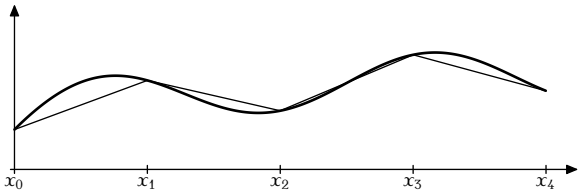
EXERCISE 3.7. The error estimate in Corollary 3.8 is $\frac{c}{\sqrt{n}}$. Can we replace $\frac{1}{\sqrt{n}}$ with something which decays faster, for instance $\frac{1}{n}$? Investigate with a computer experiment.

5.3. Numerical integration. Let $f\colon [a,b] \to \mathbb{R}$ be a continuous function. Suppose that we want to estimate the integral $\int_a^b f(x)\,dx$ for instance using a computer. We will briefly discuss some methods to do this, and the connection to the Euler–Maclaurin summation formula.

The *trapetzoidal rule* is the following. We partition the interval $[a,b]$ into n subintervals of equal length and denote the endpoints of these intervals by the points

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

In each of the intervals $[x_{k-1}, x_k]$, we approximate the graph of f by a line segment through the points $(x_{k-1}, f(x_{k-1}))$ and $(x_k, f(x_k))$. In this way we get an approximation of f which we denote by $\tilde{f}$ and which has a graph consisting of n line segments. This is illustrated in the below picture.



Over the interval $[x_{k-1}, x_k]$, the integral of $\tilde{f}$ is given by

$$\int_{x_{k-1}}^{x_k} \tilde{f}(x)\,dx = \frac{f(x_{k-1}) + f(x_k)}{2}(x_k - x_{k-1}) = \frac{f(x_{k-1}) + f(x_k)}{2}\frac{b-a}{n}.$$

Hence

$$\begin{aligned}\int_a^b \tilde{f}\,dx &= \frac{b-a}{n}\sum_{k=1}^{n}\frac{f(x_{k-1}) + f(x_k)}{2} \\ &= \frac{b-a}{n}\left(\frac{f(x_0)}{2} + f(x_1) + f(x_2) + \ldots + f(x_{n-1}) + \frac{f(x_n)}{2}\right) \\ &= \frac{b-a}{n}\sum_{k=0}^{n} f(x_k) - \frac{b-a}{n}\frac{f(a)+f(b)}{2}.\end{aligned}$$

The trapetzoidal rule is thus the approximation

$$\int_a^b f(x)\,dx \approx \frac{b-a}{n}\sum_{k=0}^{n} f(x_k) - \frac{b-a}{n}\frac{f(a)+f(b)}{2}. \tag{3.17}$$

There are many other and better methods than the trapetzoid rule, but we shall not discuss them here. The book by Davis and Rabinowitz [4] contains more information on the topic.

The approximation (3.17) resembles very much the Euler–Maclaurin summation formula. We let $h = \frac{b-a}{n}$ be the step length, and let $g(t) =$

$f(a+ht)$. By the change of variable $x = a + ht$, we get

$$\int_a^b f(x)\,\mathrm{d}x = h\int_0^n g(t)\,\mathrm{d}t,$$

and

$$\frac{b-a}{n}\sum_{k=0}^{n} f(x_k) - \frac{b-a}{n}\frac{f(a)+f(b)}{2} = h\sum_{k=0}^{n} g(t) - h\frac{g(0)+g(n)}{2}.$$

The trapetzoidal rule, is thus the approximation

$$\int_0^n g(t)\,\mathrm{d}t \approx \sum_{k=0}^{n} g(t) - \frac{g(0)+g(n)}{2}.$$

By Theorem 3.1, we have

$$\int_0^n g(t)\,\mathrm{d}t = \sum_{k=0}^{n} g(t) - \frac{g(0)+g(n)}{2} + \int_0^n g'(t)\Big(\{t\} - \frac{1}{2}\Big)\,\mathrm{d}t,$$

provided that f (and hence also g) is continuously differentiable. So the trapetzoid rule gives an error given by the integral $h\int_0^n g'(t)\big(\{t\} - \frac{1}{2}\big)\,\mathrm{d}t$.

If $|f'| \le C$, then $|g'| \le Ch$ and we have

$$\left|\int_0^n g(t)\,\mathrm{d}t - \left(\sum_{k=0}^{n} g(t) - \frac{g(0)+g(n)}{2}\right)\right| \le nhC\int_0^1 |B_1(x)|\,\mathrm{d}x = \frac{b-a}{4}C.$$

Multiplying by h, we get the following theorem.

THEOREM 3.9. *Suppose that $f\colon [a,b] \to \mathbb{R}$ is differentiable and let*

$$T_n = \frac{b-a}{n}\sum_{k=0}^{n} f(x_k) - \frac{b-a}{n}\frac{f(a)+f(b)}{2}$$

be the approximation of $\int_a^b f(x)\,\mathrm{d}x$ given by the trapetzoidal rule. Then

$$\left|\int_a^b f(x)\,\mathrm{d}x - T_n\right| \le \frac{(b-a)^2}{4n}\|f'\|.$$

If f is r times continuously differentiable, then we can use Theorem 3.5 to get an even better estimate of the integral. In the same way as above one proves the following theorem.

THEOREM 3.10. *Suppose that $f\colon [a,b] \to \mathbb{R}$ is r times differentiable and let*

$$S_n = \frac{b-a}{n}\sum_{k=0}^{n} f(x_k) - \frac{b-a}{n}\frac{f(a)+f(b)}{2} + \frac{b-a}{n}\sum_{k=1}^{[r/2]}\frac{B_{2k}(0)}{(2k)!}(f^{(2k-1)}(b) - f^{(2k-1)}(a)).$$

be an approximation of $\int_a^b f(x)\,\mathrm{d}x$. Then

$$\left|\int_a^b f(x)\,\mathrm{d}x - S_n\right| \le \frac{(b-a)^{r+1}}{n^r r!}\|f^{(r)}\|\int_0^1 |B_r(x)|\,\mathrm{d}x.$$

In particular, if f is periodic, with a period $b-a$, then

$$\left| \int_a^b f(x)\,\mathrm{d}x - \frac{b-a}{n} R_n \right| \leq \frac{(b-a)^{r+1}}{n^r r!} \|f^{(r)}\| \int_0^1 |B_r(x)|\,\mathrm{d}x,$$

where R_n is the Riemann sum $R_n = \sum_{k=0}^{n-1} f(x_k)$.

EXERCISE 3.8. Prove Theorem 3.10.

EXERCISE 3.9. Let $f\colon \mathbb{R} \to \mathbb{R}$ be infinitely many times differentiable as well as periodic, with period 1. Show that if n is such that

$$\lim_{r\to\infty} \frac{\|f^{(r)}\|}{(2\pi n)^r} = 0,$$

then $\int_0^1 f(x)\,\mathrm{d}x = R_n$, where R_n is the Riemann sum $R_n = \frac{1}{n}\sum_{k=0}^{n-1} f\left(\frac{k}{n}\right)$. You may use the following estimate by Lehmer [**7**]:

$$\sup_{x\in[0,1]} |B_n(x)| \leq 2\frac{r!}{(2\pi)^r}.$$

CHAPTER 4

Fourier series

1. The Dirichlet and the Fejér kernels

Let $\mathbb{T} = \mathbb{R}/\mathbb{Z}$, which we will think of as the interval $[0,1)$ with the points 0 and 1 identified. In words, $\mathbb{T}$ is the 1-dimensional torus or circle.

We will study a function $f\colon \mathbb{T} \to \mathbb{R}$, its *Fourier*[1] *coefficients* $(c_k(f))_{k=-\infty}^{\infty}$ and its *Fourier series* $(c_k(f)e^{i2\pi kx})_{k=-\infty}^{\infty}$, where $c_k(f)$ is defined by

$$c_k(f) = \int_0^1 f(x)e^{-i2\pi kx}\,\mathrm{d}x.$$

Hence, the Fourier series is defined if and only if the integrals above are defined.

Suppose that f has a Fourier series. We are interested in knowing when we can expect that

$$f(x) = \sum_{k=-\infty}^{\infty} c_k(f)e^{i2\pi kx} = \lim_{m,n\to\infty} \sum_{k=-m}^{n} c_k(f)e^{i2\pi kx}$$

and if so, we would like to know something about the convergence. It turns out that the regularity, like continuity and differentiability, is important for the behaviour of the Fourier series. Note that since we are identifying the points 0 and 1, continuity of $f\colon \mathbb{T} \to \mathbb{R}$ at 0 means that

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

Since $\mathbb{T} = \mathbb{R}/\mathbb{Z}$, we can also think of a function $f\colon \mathbb{T} \to \mathbb{R}$ as a 1-periodic function $f\colon \mathbb{R} \to \mathbb{R}$. Continuity of $f\colon \mathbb{T} \to \mathbb{R}$ is then the same thing as continuity of f as a 1-periodic function $\mathbb{R} \to \mathbb{R}$.

We will only (except in Exercise 4.6) consider the symmetrically truncated sums

$$S_n(f)(x) = \sum_{k=-n}^{n} c_k(f)e^{i2\pi kx},$$

which are just the truncated sum of the trigonometric Fourier series of f, that is

$$S_n(f)(x) = \sum_{k=0}^{n} (a_k \cos(2\pi kx) + b_k \sin(2\pi kx)),$$

where

$$a_0 = c_0, \qquad a_k = c_k + c_{-k}, \qquad b_k = i(c_k - c_{-k}), \qquad k \geq 1.$$

[1] Joseph Fourier, 1768–1830. French mathematician.

By the definition of c_n, we can write $S_n(f)$ as

$$S_n(f)(x) = \sum_{k=-n}^{n} e^{i2\pi kx} \int_0^1 f(t)e^{-i2\pi kt}\,dt = \int_0^1 \left(\sum_{k=-n}^{n} e^{i2\pi kx} e^{-i2\pi kt} \right) f(t)\,dt$$
$$= \int_0^1 \left(\sum_{k=-n}^{n} e^{i2\pi k(x-t)} \right) f(t)\,dt = \int_0^1 D_n(x-t)f(t)\,dt,$$

where, if x is not an integer,

$$D_n(x) = \sum_{k=-n}^{n} e^{i2\pi kx} = e^{-i2\pi nx} \frac{1 - e^{i2\pi(2n+1)x}}{1 - e^{i2\pi x}}$$
$$= \frac{e^{-i\pi(2n+1)x} - e^{-i\pi(2n+1)x}}{e^{-i\pi x} - e^{i\pi x}} = \frac{\sin(\pi(2n+1)x)}{\sin \pi x}. \tag{4.1}$$

and, if x is an integer,

$$D_n(x) = \sum_{k=-n}^{n} e^{i2\pi kx} = \sum_{k=-n}^{n} 1 = 2n + 1.$$

Since the limit of the expression in (4.1) is $2n + 1$ as x approaches an integer, we will use the expression in (4.1) also when x is an integer, letting it in that case denote the limit.

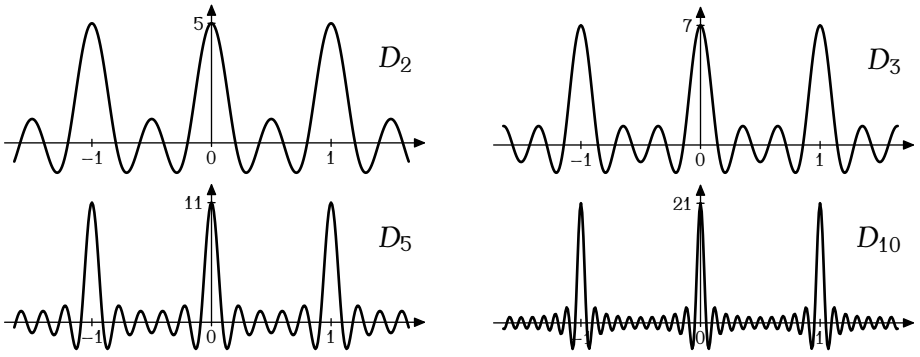
The function D_n is called the n-th Dirichlet kernel. Using convolution, we may write

$$S_n(f)(x) = \int_0^1 D_n(x-t)f(t)\,dt = D_n * f(x).$$

We will study how and when $S_n(f) = D_n * f$ converges to f.

Clearly, the Dirichlet kernel is 1-periodic, and it satisfies

$$\int_0^1 D_n(x)\,dx = 1.$$



From the look of the graphs of D_n, it seems like D_n converges poinwise to 0 except at the integer points. We shall soon prove this and investigate the properties of D_n further.

We will also study $\sigma_n(f)$ which we define by

$$\sigma_n(f) = \frac{1}{n} \sum_{k=0}^{n-1} S_n(f) = \sum_{k=-n}^{n} \left(1 - \frac{|k|}{n}\right) c_k(f) e^{i2\pi kx}.$$

Hence, we can think of $\sigma_n(f)$ as either the average of the partial sums $S_0(f), S_1(f), \dots, S_{n-1}(f)$, or as a truncated sum, similar to $S_n(f)$, but in which the terms are multiplicated with the factor $(1 - \frac{|k|}{n})$ which is 1 when $k = 0$ and decays to 0 at $k = \pm n$.

If we let

$$F_n = \frac{1}{n} \sum_{k=0}^{n-1} D_k,$$

then we may write

$$\sigma_n(f)(x) = \frac{1}{n} \sum_{k=0}^{n-1} S_k(f) = \frac{1}{n} \sum_{k=0}^{n-1} D_k * f = F_n * f.$$

The function F_n is called the n-th Fejér[2] kernel.

Since F_n is the average of Dirichlet kernels, we have

$$\int_0^1 F_n(x)\,\mathrm{d}x = 1. \tag{4.2}$$

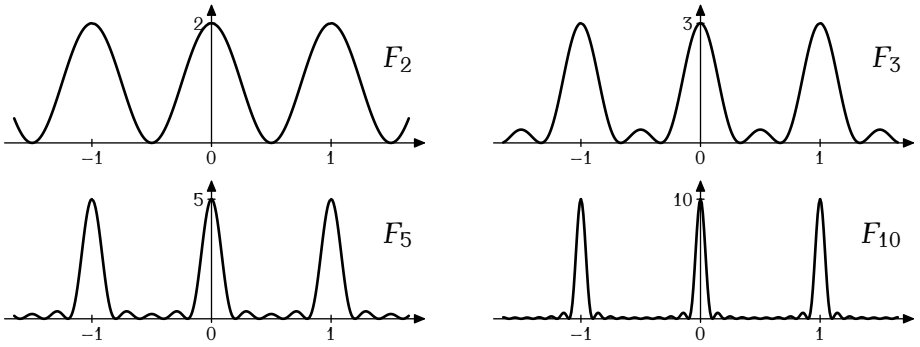
Using (4.1) we can find a simple expression for F_n. We get

$$\begin{aligned} F_n(x) &= \frac{1}{n} \sum_{k=0}^{n-1} \frac{e^{i\pi(2k+1)x} - e^{-i\pi(2k+1)x}}{e^{i\pi x} - e^{-i\pi x}} \\ &= \frac{1}{n} \frac{1}{e^{i\pi x} - e^{-i\pi x}} \left(\sum_{k=0}^{n-1} e^{i\pi(2k+1)x} - \sum_{k=0}^{n-1} e^{-i\pi(2k+1)x} \right). \end{aligned}$$

Summing the two geometric series and simplifying (Exercise 4.1), we find that

$$F_n(x) = \frac{1}{n} \left(\frac{\sin \pi n x}{\sin \pi x} \right)^2. \tag{4.3}$$

From this formula, it is apparent that $F_n(x) \geq 0$ for all x, a property which is important.



[2]Lipót Fejér, 1880–1959. Hungarian mathematician.

EXERCISE 4.1. Prove (4.3).

EXERCISE 4.2. Suppose that x and f are such that $\sum_{-\infty}^{\infty} c_k(f)e^{i2\pi kx} = a$. Prove that $S_n(f)(x) \to a$ as $n \to \infty$.

Suppose that x and f are such that $S_n(f)(x) = D_n * f(x) \to a$ as $n \to \infty$. Prove that $\sigma_n(f)(x) \to a$ as $n \to \infty$.

EXERCISE 4.3. Show that $|\sin \pi x| \geq 2x$ for $0 \leq x \leq \frac{1}{2}$.

EXERCISE 4.4. Prove that if $\frac{1}{\sqrt[4]{n}} \leq x \leq \frac{1}{2}$ then $0 \leq F_n(x) \leq \frac{2}{\sqrt{n}}$. (Use Exercise 4.3.) Conclude that $0 \leq F_n(x) \leq \frac{1}{2\sqrt{n}}$ whenever $\frac{1}{\sqrt[4]{n}} \leq x \leq 1 - \frac{1}{\sqrt[4]{n}}$.

2. Convergence of Fourier series

We formulate the important properties of the Fejér kernel in the following lemma.

LEMMA 4.1. *F_n is non-negative and even, $\int_0^1 F_n(x)\,dx = 1$ and $0 \leq F_n(x) \leq \frac{1}{2\sqrt{n}}$ for $\frac{1}{\sqrt[4]{n}} \leq x \leq 1 - \frac{1}{\sqrt[4]{n}}$.*

PROOF. It is immediately apparent from (4.3) that F_n is non-negative and even. By (4.2) we have $\int_0^1 F_n(x)\,dx = 1$.

Finally, from Exercise 4.4 we know that $0 \leq F_n(x) \leq \frac{1}{\sqrt{n}}$ whenever $\frac{1}{\sqrt[4]{n}} \leq x \leq 1 - \frac{1}{\sqrt[4]{n}}$. □

We can now prove our first result on the convergence of Fourier series.

THEOREM 4.2 (Fejér's theorem). *Suppose that $f\colon \mathbb{T} \to \mathbb{R}$ is bounded, integrable and continuous at x. Then*

$$\lim_{n\to\infty} \sigma_n(f)(x) = f(x).$$

Moreover, if f is continuous on an interval, then $\sigma_n f$ converges uniformly to f on any compact subinterval of the interval on which f is continuous.

PROOF. Since f is integrable, $c_k(f)$ is defined for each k, so the Fourier series of f exists.

Suppose that f is continuous at x. Take $\varepsilon > 0$. Since f is continuous at x there is a $\delta > 0$ such that $|f(t) - f(x)| < \delta$ when $|t - x| < \delta$. Take N so that $1/N \leq \delta^4$.

Suppose that $|f| \leq C$ and let $n \geq N$. We then have that

$$\begin{aligned} |F_n * f(x) - f(x)| &= \left| F_n * f(x) - f(x) \int_0^1 F_n(t)\,dt \right| \\ &= \left| \int_0^1 F_n(t)(f(x-t) - f(x))\,dt \right| \\ &\leq \int_0^1 F_n(x)|f(x-t) - f(x)|\,dt. \end{aligned}$$

We split the last integral into integrals over the intervals $[0, n^{-\frac{1}{4}}]$, $[n^{-\frac{1}{4}}, 1 - n^{-\frac{1}{4}}]$ and $[1 - n^{-\frac{1}{4}}, 1]$ and use the properties of F_n described in Lemma 4.1. In the middle interval we have $0 \leq F_n \leq \frac{1}{2\sqrt{n}}$ and $|f(x-t) - f(x)| \leq 2C$. Hence

$$\int_{n^{\frac{1}{4}}}^{1-n^{\frac{1}{4}}} F_n(x)|f(x-t) - f(x)|\,\mathrm{d}t \leq \int_{n^{\frac{1}{4}}}^{1-n^{\frac{1}{4}}} \frac{1}{2\sqrt{n}} 2C\,\mathrm{d}t \leq \frac{C}{\sqrt{n}}.$$

In the two other intervals, we have $|f(x-t) - f(x)| \leq \varepsilon$. Hence

$$\int_0^{n^{\frac{1}{4}}} F_n(x)|f(x-t) - f(x)|\,\mathrm{d}t + \int_{1-n^{\frac{1}{4}}}^{1} F_n(x)|f(x-t) - f(x)|\,\mathrm{d}t \leq \int_0^{n^{\frac{1}{4}}} F_n(x)\varepsilon\,\mathrm{d}t \leq \varepsilon + \int_{1-n^{\frac{1}{4}}}^{1} F_n(x)\varepsilon\,\mathrm{d}t \leq \varepsilon.$$

If we put these estimates together, we obtain that

$$|\sigma_n(f)(x) - f(x)| \leq \varepsilon + \frac{C}{\sqrt{n}}$$

if $n \geq N$, which proves that $\lim_{n\to\infty} \sigma_n(f)(x) = f(x)$.

Now, if f is continuous on I and $J \subset I$ is compact, then f is uniformly continuous on J. For $\varepsilon > 0$ there then exists a $\delta > 0$ such that for all $x \in J$ we have $|f(x) - f(t)| < \varepsilon$ when $|x - t| < \delta$. The estimates above then shows that

$$|\sigma_n(f)(x) - f(x)| \leq \varepsilon + \frac{C}{\sqrt{n}}$$

whenever $x \in J$ and $n \geq N$. Hence $\sigma_n(f)$ converges uniformly to f on J. □

We will state a more precise version of Fejér's theorem. For this purpose we need what is called modulus of continuity.

Definition 4.3 (Modulus of continuity). Suppose that f is continuous at a point x. Then the *local modulus of continuity at* x of f is the function ω_x defined by

$$\omega_x(\delta) = \max\{\, |f(x) - f(y)| : |x - y| \leq \delta \,\}.$$

If f is a continuous function on a compact interval, then the *modulus of continuity* of f is the function ω defined by

$$\omega(\delta) = \max\{\, |f(x) - f(y)| : |x - y| \leq \delta \,\}, \qquad \delta > 0.$$

Note that if f is continuous at x, then $\omega_x(\delta) \to 0$ as $\delta \to 0$. If f is continuous on a compact interval, then f is uniformly continuous, and $\omega(\delta) \to 0$ as $\delta \to 0$.

Let ω_x be the local modulus of continuity at x of f. If we investigate the proof of Theorem 4.2, we see that we can choose $\varepsilon = \omega_x(n^{-\frac{1}{4}})$. In this way, we obtain the following version of Fejér's theorem.

THEOREM 4.4. *Suppose that $f\colon \mathbb{T} \to \mathbb{R}$ is bounded, integrable, continuous at x and that ω_x is its local modulus of continuity at x. Then*

$$|\sigma_n(f)(x) - f(x)| \leq \omega_x(n^{-\frac{1}{4}}) + \frac{\|f\|}{\sqrt{n}}.$$

In particular, if f in continuous and ω is its modulus of continuity, then

$$\|\sigma_n(f) - f\| \leq \omega(n^{-\frac{1}{4}}) + \frac{\|f\|}{\sqrt{n}}.$$

We shall now investigate what can be said about the convergence of $\sigma_n f$ at points where f is not continuous. Let

$$f_-(x) = \lim_{t \to x^-} f(t), \qquad f_+(x) = \lim_{t \to x^+} f(t),$$

denote the left and right limit of f at x if the limits exist. We give the following theorem.

THEOREM 4.5. *Suppose that f is bounded, integrable and that f has left and right limit at a point x. Then*

$$\lim_{n \to \infty} \sigma_n(f)(x) = \frac{f_-(x) + f_+(x)}{2}.$$

PROOF. Exercise 4.5. □

By Fejér's theorem, we know that if f is continuous, then $\sigma_n(f)(x)$ converges to $f(x)$. For $S_n(f)(x)$ to converges to $f(x)$, it is not enough that f is continuous.

To formulate conditions which guarantee that $S_n(f)(x)$ converges to $f(x)$, we introduce the following notation for the left and right limit of f at a point x

$$f_-(x) = \lim_{t \to x^-} f(t), \qquad f_+(x) = \lim_{t \to x^+} f(t),$$

if the limits exist. Similarly we define the left and right derivative of f at x by

$$f'_-(x) = \lim_{h \to 0^-} \frac{f(x+h) - f_-(x)}{h}, \qquad f'_+(x) = \lim_{h \to 0^+} \frac{f(x+h) - f_+(x)}{h},$$

if the limits exist.

THEOREM 4.6 (Dirichlet). *Suppose that f is bounded, integrable, and that x is such that $f_-(x)$, $f_+(x)$, $f'_-(x)$ and $f'_+(x)$ exists. Then*

$$\lim_{n \to \infty} S_n(f)(x) = \frac{f_-(x) + f_+(x)}{2}.$$

PROOF. Since D_n and f are 1-periodic functions, we have

$$S_n(f)(x) = D_n * f(x) = \int_0^1 D_n(t) f(x-t)\,\mathrm{d}t = \int_{-\frac{1}{2}}^{\frac{1}{2}} D_n(t) f(x-t)\,\mathrm{d}t.$$

Since D_n is even, we have

$$\int_{-\frac{1}{2}}^{0} D_n(t)\,\mathrm{d}t = \int_0^{\frac{1}{2}} D_n(t)\,\mathrm{d}t = \frac{1}{2}.$$

Hence

$$\begin{aligned} S_n(f)(x) - \frac{f_-(x) + f_+(x)}{2} \\ &= S_n(f)(x) - f_-(x) \int_{-\frac{1}{2}}^{0} D_n(t)\,\mathrm{d}t - f_+(x) \int_0^{\frac{1}{2}} D_n(t)\,\mathrm{d}t \\ &= \int_{-\frac{1}{2}}^{0} D_n(t)(f(x-t) - f_+(x))\,\mathrm{d}t + \int_0^{\frac{1}{2}} D_n(t)(f(x-t) - f_-(x))\,\mathrm{d}t. \end{aligned}$$

Consider for instance the first integral on the last line above. By the formula (4.1) for the Dirichlet kernel, we can write

$$\begin{aligned} \int_{-\frac{1}{2}}^{0} D_n(t)(f(x-t) - f_+(x))\,\mathrm{d}t \\ &= \int_{-\frac{1}{2}}^{0} \frac{f(x-t) - f_+(x)}{t} \frac{t}{\sin \pi t} \sin(\pi(2n+1)t)\,\mathrm{d}t. \end{aligned}$$

Because of the assumption, the term $\frac{f(x-t)-f_+(x)}{t}$ has the finite limit $f'_+(x)$ as $t \to 0^-$. Hence this term is bound. Similarly, $\frac{t}{\sin \pi t}$ is bounded and continuous on $(-\frac{1}{2}, 0)$. It follows that

$$\frac{f(x-t) - f_+(x)}{t} \frac{t}{\sin \pi t}$$

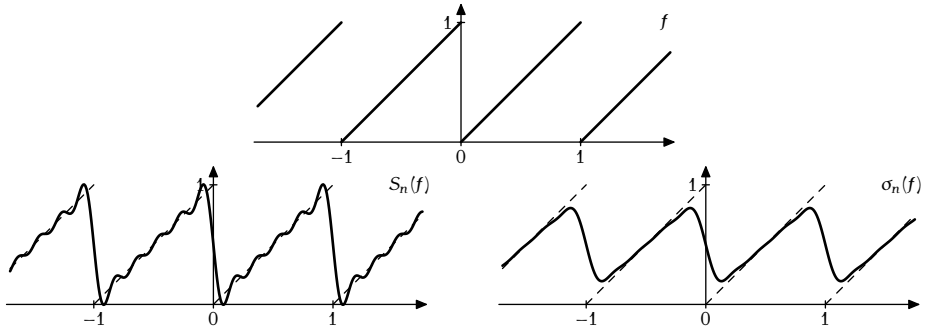
is integrable in the sense of Riemann.

Now, the Riemann–Lebesgue lemma implies that

$$\int_{-\frac{1}{2}}^{0} D_n(t)(f(x-t) - f_+(x))\,\mathrm{d}t \to 0$$

as $n \to 0$. The same argument for the integral over $(0, \frac{1}{2})$ gives the same result. □

In conclusion, less regularity of f is required for $\sigma_n(f)$ to converge to f, than for $S_n(f)$ to converge to f. Let us briefly study what happens when f is piecewise continuous, but not continuous. Let $f(x) = x$ on $\mathbb{T}$. As a function on $\mathbb{R}$, the graph of f is shown below, together with $S_n(f)$ and $\sigma_n(f)$.



EXERCISE 4.5. Prove Theorem 4.5 by modifying the proof of Fejér's theorem.

EXERCISE 4.6. Suppose that f is two times continuously differentiable. Prove that there exists a constant C such that $|c_k(f)| \leq C/k^2$.

Use this to prove that $\sum_{-\infty}^{\infty} c_k(f)e^{i2\pi kx}$ converges. Conclude (Exercise 4.2 and Theorem 4.2) that $f(x) = \sum_{-\infty}^{\infty} c_k(f)e^{i2\pi kx}$.

EXERCISE 4.7. Let f and g be two continuous functions. Show that $f = g$ if and only if f and g have the same Fourier series.

EXERCISE 4.8. Give another proof of the statement in Exercise 3.9, using Fourier series.

3. A Tauberian theorem by Hardy and Landau

Suppose that f is bounded, integrable and continuous on an interval. Then by Theorem 4.2, $\sigma_n(f)$ converges uniformly to f on any compact subinterval of the interval of continuity. We will prove that if we assume a bit more regularity of f, then even $S_n(f)$ will converge uniformly to f on any such compact interval. To do so, we will use a so called Tauberian theorem by Hardy and Landau. Let us first say something about Tauberian theorems.

Consider a sequence $(a_k)_{k=1}^n$ of numbers. We let $s_n = \sum_{k=1}^n a_k$. The series is said to be summable with sum s if $s = \lim_{n\to\infty} s_n$. We may also consider so called *Abel summation*. The series is called Abel summable with Abel sum A if the limit

$$A = \lim_{x\to 1^-} \sum_{k=1}^{\infty} a_k x^k$$

exists.

A theorem by Niels Henrik Abel[3] says that if a sequence is summable, then it is Abel summable and the sum is equal to the Abel sum.

However, if the sequence is Abel summable, then it need not be summable, which happens for instance when $a_k = (-1)^k$. In other words, the converse of Abel's theorem is not true.

Hence, unless we assume special conditions on the numbers a_k, Abel summability does not imply summability.

Similarly, we may consider the so called *Cesàro*[4] *summation* of a sequence. The sequence $(a_k)_{k=1}^{\infty}$ is said to be Cesàro summable with Cesàro sum σ if the limit $\sigma = \lim_{n\to\infty} \sigma_n$ exists, where

$$\sigma_n = \frac{1}{n}\sum_{k=1}^{n} s_k.$$

Then, if the sequence is summable, it is Cesàro summable and the sum is equal to the Cesàro sum.

[3]Niels Henrik Abel, 1802–1829. Norwegian mathematician.

[4]Ernesto Cesàro, 1859–1906. Italian mathematician.

As for Abel summation the converse is not true. If a series is Cesàro summable, then it need not be summable, unless we impose extra conditions on the sequence a_k.

Tauberian theorem are theorems about assumptions on the terms of a sequence, which guarantees that the series is summable if it is summable in the sence of for instance Abel or Cesàro. The first Tauberian theorem was proved by Alfred Tauber.[5] He proved that if $ka_k \to 0$ as $k \to \infty$, then the series is summable if it is Abel summable.

Here we shall study and use the following Tauberian theorem for Cesàro summability.

THEOREM 4.7 (the Hardy[6]–Landau[7] Tauberian theorem [5, 6]). *Assume that $(u_k)_{k=1}^{\infty}$ is a sequence of functions $u_k\colon [a,b] \to \mathbb{R}$ and that there exists a constant A for which $\|u_k\| \leq \frac{A}{k}$ holds for all k.*

Let $s_n(x) = \sum_{k=1}^{n} u_k(x)$ be the partial sums of the series $(u_k)_{k=1}^{\infty}$. If

$$\sigma_n = \frac{1}{n}\sum_{k=1}^{n} s_k$$

converges uniformly to a function f, then s_n converges uniformly to f as $n \to \infty$.

PROOF. Let $\varepsilon > 0$. Since σ_n converges uniformly to f, there exists an N such that $|f(x) - \sigma_n(x)| < \varepsilon$ holds for all $x \in [a,b]$ as long as $n \geq N$.

We take $m > n \geq N$ and write

$$m\sigma_m - n\sigma_n = s_{n+1} + s_{n+2} + \ldots + s_m$$

and

$$\begin{aligned} s_m &= s_m \\ s_m &= s_{m-1} + u_m \\ s_m &= s_{m-2} + u_{m-1} + u_m \\ &\vdots \\ s_m &= s_{n+1} + u_{n+2} + u_{n+3} + \ldots + u_m. \end{aligned}$$

Hence

$$\begin{aligned} (m-n)s_m &= (s_{n+1} + \ldots + s_m) + u_{n+2} + 2u_{n+3} + \ldots + (m-n-1)u_m \\ &= m\sigma_m - n\sigma_n + u_{n+2} + 2u_{n+3} + \ldots + (m-n-1)u_m. \end{aligned}$$

Subtracting $(m-n)\sigma$ from both sides yields

$$(m-n)(s_m-\sigma) = m(\sigma_m-\sigma) - n(\sigma_n-\sigma) + u_{n+2} + 2u_{n+3} + \ldots + (m-n-1)u_m.$$

Using that $|u_k| \leq \frac{A}{k}$ we then get that

$$(m-n)|s_m - \sigma| \leq m|\sigma_m - \sigma| + n|\sigma_n - \sigma| + \sum_{k=n+2}^{m} A\frac{k-n-1}{k}.$$

---

[5] Alfred Tauber, 1866–1942. Hungarian–Austrian mathematician.

[6] Godfrey Harold Hardy, 1877–1947. English mathematician.

[7] Edmund Georg Hermann Landau, 1877–1938. German mathematician.

We estimate the sum by

$$\sum_{k=n+2}^{m} A\frac{k-n-1}{k} \leq \sum_{k=n+2}^{m} A\frac{m-n-1}{n+2} = A\frac{(m-n-1)^2}{n+2},$$

and use that $|\sigma_m - \sigma|$ and $|\sigma_n - \sigma|$ are both less than ε since $m, n \geq N$, to get

$$(m-n)|s_m - \sigma| \leq (m+n)\varepsilon + A\frac{(m-n)^2}{n}.$$

Hence

$$\begin{aligned} |s_m - \sigma| &\leq \varepsilon\frac{m+n}{m-n} + A\frac{(m-n)}{n} \\ &\leq \varepsilon\Big(1 + \frac{2}{\frac{m}{n}-1}\Big) + A\Big(\frac{m}{n} - 1\Big). \end{aligned}$$

If m is large enough, it is possible to choose n so that

$$\sqrt{\varepsilon} \leq \frac{m}{n} - 1 \leq 2\sqrt{\varepsilon}.$$

We then have

$$|s_m - \sigma| \leq \varepsilon + 2\sqrt{\varepsilon} + 2A\sqrt{\varepsilon}.$$

Hence s_m converges uniformly to σ. □

COROLLARY 4.8. *Suppose that f is of bounded variation, and continuous on an interval. Then $S_n(f)$ converges uniformly to f on any compact subinterval of the interval of continuity.*

PROOF. By Theorem 2.8, we have

$$c_n = \int_0^1 f(x)e^{-i2\pi nx}\,dx = \int_0^1 f(x)\,d\left(\frac{e^{-i2\pi nx}}{-i2\pi n}\right).$$

Integration by parts (Theorem 2.9) implies that

$$c_n = \int_0^1 \frac{e^{i2\pi nx}}{-i2\pi n}\,df(x).$$

Hence $|c_n| \leq \frac{1}{2\pi|n|}\operatorname{var}_{[0,1]} f$.

Let I be a compact subinterval of an interval on which f is continuous. By Fejér's theorem, Theorem 4.2, $\sigma_n(f)$ converges uniformly to f on I. Note that $\sigma_n(f)$ is the Cesàro sum of the Fourier series of f. Therefore, Theorem 4.7 implies that $S_n(f)$ converges uniformly to f on I. □

EXERCISE 4.9. Give an example of a series which is Cesàro summable, but not summable.

CHAPTER 5

Polynomial approximation

In this chapter, we are going to study how to approximate a continuous function $f\colon [a,b] \to \mathbb{R}$ by a polynomial p. We shall see that such approximations are possible, and will study various ways in which such approximations can be obtained.

1. Weierstaß' approximation theorem

We start by the following theorem by Karl Weierstraß[1].

THEOREM 5.1 (Weierstraß' approximation theorem). *Let $f\colon [a,b] \to \mathbb{R}$ be a continuous function and let $\varepsilon > 0$. Then there exists a polynomial p such that*

$$|f(x) - p(x)| < \varepsilon$$

for all $x \in [a,b]$.

We will give proofs later.

2. Lagrange polynomials

Suppose that we are given n points

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

such that $x_k \in [a,b]$ for all k and that $x_j \neq x_k$ whenever $j \neq k$. Our goal is to find a polynomial of lowest possible degree such that $p(x_k) = y_k$ for all k. In other words, we are looking for an *interpolating polynomial*.

DEFINITION 5.2. Given n points

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

such that $x_k \in [a,b]$ for all k and that $x_j \neq x_k$ whenever $j \neq k$, the *Lagrange polynomial* to these points is the polynomial

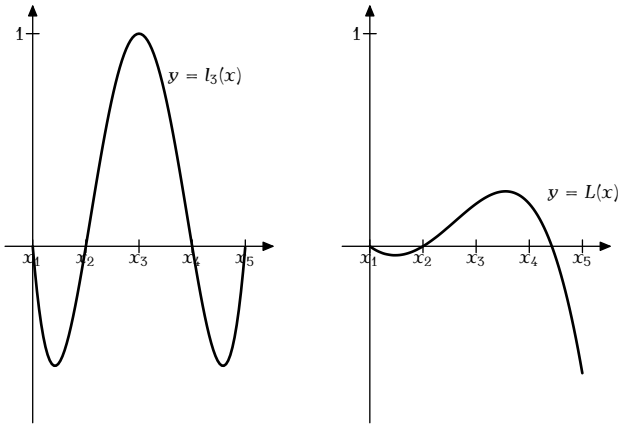
$$L = \sum_{k=1}^{n} y_k l_k, \qquad \text{where} \qquad l_k(x) = \prod_{\substack{1 \leq j \leq k \\ j \neq k}} \frac{x - x_j}{x_k - x_j}.$$

Note that L is a polynomial of degree at most $n-1$.

EXAMPLE 5.3. Consider the points $(x_1, y_1) = (0,0)$, $(x_1, y_1) = (0.25, 0)$, $(x_1, y_1) = (0.5, 0.2)$, $(x_1, y_1) = (0.75, 0.2)$, $(x_1, y_1) = (1, -0.6)$.

The graphs of l_3 and the Lagrange polynomial of these points are shown below.

[1] Karl Theodor Wilhelm Weierstraß, 1815–1897. German mathematician.



PROPOSITION 5.4. *We have $L(x_k) = y_k$ for $k = 1, 2, \ldots, n$.*

PROOF. By the definition of l_k we have

$$l_k(x_j) = \delta_{k,j} := \begin{cases} 0 & \text{if } j \neq k, \\ 1 & \text{if } j = k. \end{cases}$$

(The symbol $\delta_{k,j}$ defined above is called Kronecker's[2] delta.) Hence,

$$L(x_j) = \sum_{k=1}^{n} y_k l_k(x_j) = y_j. \qquad \square$$

Suppose that we have a function $f\colon [a,b] \to \mathbb{R}$ that we want to approximate by a polynomial. We could then choose n points x_k in $[a,b]$, put $y_k = f(x_k)$ and form the Lagrange polynomial L. By Proposition 5.4, we have $L(x_k) = f(x_k)$ for $k = 1, 2, \ldots, n$, but we would like to know how well L approximates f in other points of $[a,b]$.

LEMMA 5.5. *If $g\colon I \to \mathbb{R}$ is n times differentiable and has $n+1$ zeroes in I, then there exists a $\xi \in I$ such that $g^{(n)}(\xi) = 0$.*

PROOF. Let g be $n-1$ times differentiable and suppose that $x_{0,1} < x_{0,2} < \ldots x_{0,n+1}$ are zeroes of g. By Rolle's theorem, on each of the intervals $[x_{0,k}, x_{0,k+1}]$, there exists an $x_{1,k} \in (x_{0,k}, x_{0,k+1})$ such that $g'(\xi_k) = 0$. Hence we have n zeroes of g'.

In the same way, replacing g by g', Rolle's Theorem implies that there exists an $x_{2,k} \in (x_{1,k}, x_{1,k+1})$ such that $g''(x_{2,k}) = 0$, and we have $n-1$ zeroes of g''.

Continuing in this way we eventually end up with a $x_{n,1}$ such that $g^{(n)}(x_{n,1}) = 0$. $\square$

THEOREM 5.6. *Let $f\colon [a,b] \to \mathbb{R}$ be n times diffentiable, $(x_k, y_k) = (x_k, f(x_k))$ for $k = 1, 2, \ldots, n$ with $x_j \neq x_k$ if $j \neq k$, and let L be the corresponding Lagrange polynomial. For every $x \in [a,b]$ there exists a $\xi \in [a,b]$ such that*

$$f(x) - L(x) = \frac{f^{(n)}(\xi)}{n!} \prod_{k=1}^{n} (x - x_k).$$

[2]Leopold Kronecker, 1823–1891. German mathematician.

PROOF. Put $W(x) = \prod_{k=1}^{n}(x - x_k)$. We may assume that x is not equal to any of the x_k, since otherwise, there is nothing to prove. Then $W(x) \neq 0$ and we want to find an expression for

$$C = \frac{f(x) - L(x)}{W(x)}$$

in terms of f only, and not involving the polynomial L. For this purpose, we let $g(t) = f(t) - L(t) - CW(t)$. Then g is zero in each of the points $x, x_1, x_2, \ldots, x_n$. Since g has $n + 1$ zeroes, there is by Lemma 5.5 a point ξ such that $g^{(n)}(\xi) = 0$.

But $L^{(n)} = 0$ since L is of degree at most $n - 1$, and $W^{(n)} = n!$. Hence

$$0 = g^{(n)}(\xi) = f^{(n)}(\xi) - Cn!,$$

which implies that $C = f^{(n)}(\xi)/n!$ and finishes the proof. □

Given the points $x_1, x_2, \ldots, x_n$, we let

$$W(x) = \prod_{k=1}^{n}(x - x_k).$$

An immediate corollary to Theorem 5.6 is then the following.

COROLLARY 5.7. *Let $f\colon [a, b] \to \mathbb{R}$ be n times differentiable and let L be the Lagrange polynomial to the points $(x_k, y_k) = (x_k, f(x_k))$ where the points x_k are distinct. Then*

$$\|f - L\| \leq \frac{\|f^{(n)}\|}{n!}\|W\| \leq \frac{\|f^{(n)}\|}{n!}(b - a)^n.$$

Because of the above corollary, it is of interest to make $\|W\|$ as small as possible for a given n. We prove the following result, which we later on will relate to the so called Chebyshev[3] polynomials.

THEOREM 5.8. *For n points $(x_k)_{k=1}^{n}$ in $[-1, 1]$, the norm $\|W\|$ is minimal when*

$$x_k = \cos\Big(\frac{2k - 1}{2n}\pi\Big),$$

in which case $\|W\| = 2^{1-n}$ and $W = 2^{1-n}T_n$, where T_n is the polynomial of degree n such that

$$T_n(\cos x) = \cos(nx).$$

PROOF. Since

$$\cos^n\theta = \Big(\frac{e^{i\theta} + e^{-i\theta}}{2}\Big)^n = \frac{e^{in\theta} + e^{-in\theta}}{2^n} + \ldots = 2^{1-n}\cos(n\theta) + \ldots,$$

we have

$$\cos(n\theta) = 2^{n-1}\cos^n\theta + \ldots,$$

and we see that there is a polynomial T_n such that $T_n(\cos x) = \cos(nx)$. Apparently, the leading coefficient of T_n is 2^{n-1}.

Let $W = 2^{1-n}T_n$. Then W is of degree n and $W(x) = 0$ when $x = \cos\theta$ with θ such that $\cos(n\theta) = 0$. Hence $W(x) = 0$ if and only if

[3]Pafnuty Lvovich Chebyshev, 1821–1894. Russian mathematician.

x is of the form $x_k = \cos\left(\frac{2k-1}{2n}\pi\right)$ with $k = 1, 2, \ldots, n$. Then W can be written in the form

$$W(x) = \gamma \prod_{k=1}^{n} (x - x_k),$$

for some number γ.

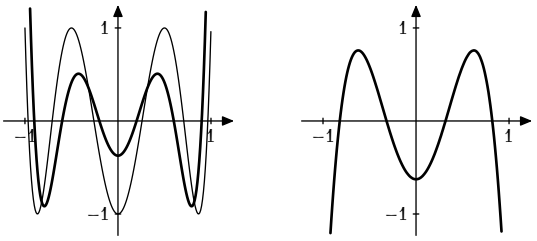
Since the leading coefficient of T_n is 2^{n-1}, the leading coefficient of W is 1 and $\gamma = 1$.

Since $W(\cos\theta) = 2^{1-n}\cos(n\theta)$, we have $\|W\| = 2^{1-n}$.

Suppose now that $V(x) = \prod_{k=1}^{n}(x - \tilde{x}_n)$ for some other choice of interpolation points. Assume that $\|V\| < \|W\|$. We will show that this leads to a contradiction.

We will study $W - V$ and we first note that $W - V$ is of degree at most $n - 1$, since both W and V are of degree n and the terms of degree n cancel.

Below are pictures of the graphs of $2^{n-1}W$ (left, thin) and $2^{n-1}V$ (left, thick) as well as the difference $2^{n-1}(W - V)$ (right).



Because of the relation with $\cos(n\theta)$, we have that W assumes its maximal modulus in $n + 1$ points of $(-1, 1)$, and W assumes different signs in two such neighbouring points. From this and the assumption that $\|V\| < \|W\|$, it follows that $W - V$ changes sign at least once in each of the intervals between the points of maximal modulus of W. Hence $W - V$ has at least n zeroes, but this is impossible since $W - V$ is of degree at most $n - 1$. □

EXERCISE 5.1. Let $f(x) = |x|$. Use a computer to plot the graph of f and the Lagrange polynomial to the n points (x_k, y_k) where $y_k = f(x_k)$ and

$$x_1 = -1, \quad x_2 = -1 + 2\frac{1}{n-1}, \quad \ldots, \quad x_{n-1} = -1 + 2\frac{n-2}{n-1}, \quad x_n = 1.$$

Does it seem like these Lagrange polynomials can be used to obtain a polynomial p such that $|f(x) - p(x)| < \varepsilon$ for all $x \in [-1, 1]$?

EXERCISE 5.2. Use Corollary 5.7 and Theorem 4.2 to prove Theorem 5.1.

EXERCISE 5.3. Let $x_1, x_2, \ldots, x_n$ be distinct points of $[a, b]$. Put

$$p(x) = \sum_{k=1}^{n} (y_k A_k(x) + y'_k B_k(x)),$$

where $A_k(x) = (1 - 2(x - x_k)l_k'(x))l_k(x)^2$ and $B_k(x) = (x - x_k)l_k(x)^2$. Prove that $p(x_k) = y_k$ and $p'(x_k) = y_k'$.

EXERCISE 5.4. Suppose that we want to approximate an n times differentiable function $f\colon [a,b] \to \mathbb{R}$ by a polynomial p of degree $n-1$. Compare the error estimate in Corollary 5.7 by the corresponding error we would have got by letting p be the Taylor polynomial of degree $n-1$ at the point a.

EXERCISE 5.5. Show that the Lagrange polynomial L can be written in the form

$$L(x) = c \begin{vmatrix} 0 & 1 & x & x^2 & \dots & x^{n-1} \\ y_1 & 1 & x_1 & x_1^2 & \dots & x_1^{n-1} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ y_n & 1 & x_n & x_n^2 & \dots & x_n^{n-1} \end{vmatrix},$$

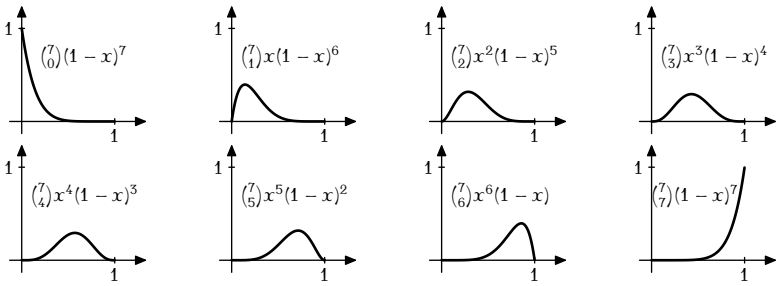
where c is a constant.

3. Bernstein polynomials

To any function $f\colon [0,1] \to \mathbb{R}$, we associate the n-th *Bernstein*[4] *polynomial* $B_n(f)$, defined by

$$B_n(f)(x) = \sum_{k=0}^{n} \binom{n}{k} f\Big(\frac{k}{n}\Big) x^k (1-x)^{n-k}.$$

Each of the polynomials $\binom{n}{k}x^k(1-x)^{n-k}$ have a unique point at which it is maximal, and they tend to get more and more concentrated around their maximum the larger n gets. This is illustrated in the picture below.



Weierstraß' approximation theorem follows immediately from the following theorem.

THEOREM 5.9. *If $f\colon [0,1] \to \mathbb{R}$ is continuous, then $B_n(f) \to f$ uniformly as $n \to \infty$.*

We shall prove this theorem as a special case of the more general theorem below. First we need some terminology. A linear operator $L\colon \mathcal{C}([0,1]) \to \mathcal{C}([0,1])$ is *monotone* if

$$f \le g \qquad \Rightarrow \qquad Lf \le Lg.$$

THEOREM 5.10. *Suppose that for each n, the operator $L_n\colon \mathcal{C}([a,b]) \to \mathcal{C}([a,b])$ is linear and monotone and that*

$$\|L_n f - f\| \to 0 \qquad n \to \infty,$$

[4]Sergei Natanovich Bernstein, 1880–1968. Russian mathematician.

whenever f is given by $f(x) = 1$, $f(x) = x$ or $f(x) = x^2$. Then

$$\|L_n f - f\| \to 0 \qquad n \to \infty,$$

holds for any $f \in \mathcal{C}([a,b])$.

PROOF. Let $\phi_y(x) = (y-x)^2 = y^2 - 2yx + x^2$. Since each L_n is linear, we then have that $L_n\phi_y \to \phi_y$ uniformly as $n \to \infty$, and the convergence is uniform in $y \in [a,b]$.

Since f is continuous on $[a,b]$, it is uniformly continuous on $[a,b]$. Let $\varepsilon > 0$. There is then a $\delta > 0$ such that $|f(x) - f(y)| < \varepsilon$ whenever $|x-y| < \delta$. Put $C = 2\|f\|\delta^{-2}$.

For any $y \in [a,b]$ we have

$$|f(y) - f(x)| \le \varepsilon + C\phi_y(x), \qquad x \in [a,b].$$

This follows immediately from the inequalities

$$\begin{aligned} &|f(y) - f(x)| \le \varepsilon, && \text{if } |x-y| < \delta,\\ &|f(y) - f(x)| \le 2\|f\| = C\delta^2 \le C|x-y|^2 = C\phi_y(x), && \text{if } |x-y| \ge |\delta. \end{aligned}$$

Hence, we have

$$-\varepsilon - C\phi_y \le f(y) - f \le \varepsilon + C\phi_y.$$

By the monotonicity of L_n, it follows that

$$-L_n\varepsilon - CL_n\phi_y \le L_n f(y) - L_n f \le L_n\varepsilon + CL_n\phi_y.$$

In other words, regarding the number $f(y)$ as a constant function, we have $|(L_n f(y))(x) - (L_n f)(x)| \le |(L_n\varepsilon)(x)| + C|(L_n\phi_y)(x)|$ for all $x \in [a,b]$.

But ε and $f(y)$ are constant functions, so $L_n\varepsilon = \varepsilon L_n 1$ and $L_n f(y) = f(y)L_n 1$, and both converge uniformly to ε and $f(y)$. We can therefore take n_0 so large that $\|L_n f(y) - f(y)\| \le \varepsilon$ as well as $\|L_n 1 - 1\| < \varepsilon$ when $n \ge n_0$. Then, for all $x \in [a,b]$, we have

$$\begin{aligned} |f(y) - (L_n f)(x)| &\le \|L_n f(y) - f(y)\| + |(L_n f(y))(x) - (L_n f)(x)|\\ &\le \varepsilon + \varepsilon\|L_n 1\| + C|(L_n\phi_y)(x)|, \end{aligned}$$

when $n \ge n_0$. In particular, $|f(y) - (L_n f)(y)| \le \varepsilon + \varepsilon(1+\varepsilon) + C|(L_n\phi_y)(y)|$. Since $(L_n\phi_y)(y) \to 0$ as $n \to \infty$, uniformly in y, there is an n_1 so that

$$|f(y) - (L_n f)(y)| \le \varepsilon + \varepsilon(1+\varepsilon) + \varepsilon$$

holds for all y and all $n \ge n_1$. This proves that $L_n f$ converges uniformly to f. □

Theorem 5.9 now follows from Theorem 5.10 by using the following lemma.

LEMMA 5.11. *The operators B_n are linear and monotone and*

$$\|B_n(f) - f\| \to 0 \qquad n \to \infty,$$

holds whenever f is given by $f(x) = 1$, $f(x) = x$ or $f(x) = x^2$.

PROOF. It is clear that B_n is linear. The monotonicity of B_n follows since $\binom{n}{k}x^k(1-x)^{n-k} \geq 0$ whenever $x \in [0,1]$.

Let f_0, f_1 and f_2 be defined by $f_j(x) = x^j$. We start by observing that

$$\sum_{k=0}^{n} \binom{n}{k} x^k(1-x)^{n-k} = (x + (1-x))^n = 1. \tag{5.1}$$

Hence, $B_n(f_0) = f_0$ and $B_n(f_0)$ converges to f_0 uniformly.

Next, we use that (Exercise 5.6)

$$\binom{n}{k}\frac{k}{n} = \binom{n-1}{k-1}, \qquad 1 \leq k \leq n.$$

We then have that

$$\begin{aligned} B_n(f_1)(x) &= \sum_{k=0}^{n} \binom{n}{k}\frac{k}{n}x^k(1-x)^{n-k} = \sum_{k=1}^{n} \binom{n-1}{k-1}x^k(1-x)^{n-k} \\ &= x\sum_{k=0}^{n-1} \binom{n-1}{k-1}x^k(1-x)^{n-1-k} = x, \end{aligned}$$

where we used (5.1) in the last step. Hence, $B_n(f_1) = f_1$ and $B_n(f_1)$ converges uniformly to f_1.

Finally, we have

$$\begin{aligned} B_n(f_2)(x) &= \sum_{k=0}^{n} \binom{n}{k}\frac{k^2}{n^2}x^k(1-x)^{n-k} = \sum_{k=1}^{n} \binom{n-1}{k-1}\frac{k}{n}x^k(1-x)^{n-k} \\ &= \frac{n-1}{n}\sum_{k=1}^{n} \binom{n-1}{k-1}\frac{k}{n-1}x^k(1-x)^{n-k} \\ &= \frac{n-1}{n}\sum_{k=1}^{n} \binom{n-1}{k-1}\frac{k-1}{n-1}x^k(1-x)^{n-k} \\ &\qquad + \frac{1}{n}\sum_{k=1}^{n} \binom{n-1}{k-1}x^k(1-x)^{n-k} \\ &= \frac{n-1}{n}x\sum_{k=0}^{n-1} \binom{n-1}{k}\frac{k}{n-1}x^k(1-x)^{n-1-k} \\ &\qquad + \frac{x}{n}\sum_{k=0}^{n-1} \binom{n-1}{k}x^k(1-x)^{n-1-k} \\ &= \frac{n-1}{n}xB_{n-1}(f_1)(x) + \frac{x}{n} = \frac{n-1}{n}x^2 + \frac{x}{n}. \end{aligned}$$

Hence $B_n(f_2)$ converges uniformly to f_2. □

It is possible to refine the proofs of Theorem 5.10 and Lemma 5.11, to get the following result.

THEOREM 5.12. *Suppose that f is Hölder*[5] *continuous on $[0,1]$, i.e. that there are constants $C \geq 0$ and $\alpha \in (0,1]$ such that*

$$|f(x) - f(y)| \leq C|x-y|^{\alpha}$$

holds for all x and y in $[0,1]$. Then

$$\|f - B_n(f)\| \leq \left(1 + \frac{\|f\|}{4} C^{\frac{1}{\alpha}}\right) n^{-\frac{\alpha}{\alpha+1}}.$$

PROOF. Exercise 5.7. □

EXERCISE 5.6. Prove that $\binom{n}{k}\frac{k}{n} = \binom{n-1}{k-1}$.

EXERCISE 5.7. Prove Theorem 5.12. Hint: Prove that $|B_n(\phi_y)(y)| \leq \frac{1}{4n}$ and choose $\varepsilon = n^{-t}$ with $t = \frac{\alpha}{\alpha+1}$.

4. Chebyshev polynomials

We start by defining the Chebyshev polynomials. These have already appeared in the proof of Theorem 5.8.

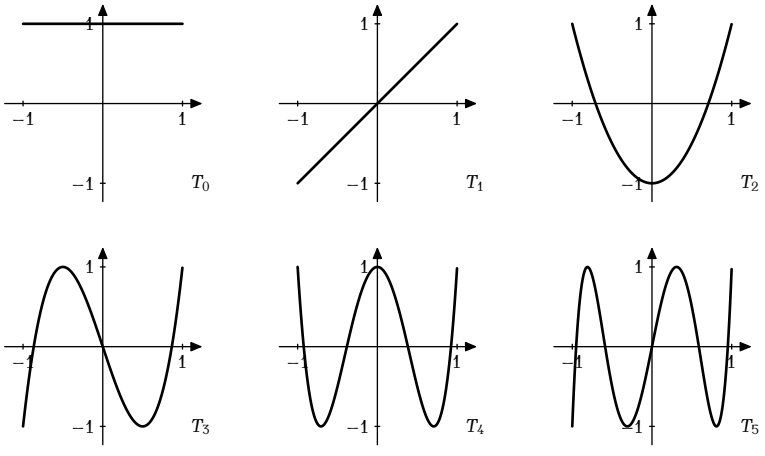
DEFINITION 5.13 (Chebyshev polynomials). The *Chebyshev polynomials* are the polynomials T_n defined by

$$T_n(\cos x) = \cos(nx).$$

As we have seen in Theorem 5.8, T_n is given by

$$T_n(x) = 2^{n-1} \prod_{k=1}^{n} \left(x - \cos\left(\frac{2k-1}{2n}\pi x\right)\right).$$

The graphs of the first few Chebyshev polynomials are shown below.



[5]Otto Ludwig Hölder, 1859–1937. German mathematician.

Consider the scalar product

$$(f,g) = \int_{-1}^{1} f(x)g(x)\frac{1}{\sqrt{1-x^2}}\,\mathrm{d}x.$$

The Chebyshev polynomials are orthogonal with respect to this scalar product, which is easily seen by a change of variables $x = \cos t$:

$$\begin{aligned}(T_n, T_m) &= \int_{-1}^{1} T_n(x)T_m(x)\frac{1}{\sqrt{1-x^2}}\,\mathrm{d}x = \int_0^{\pi} T_n(\cos t)T_m(\cos t)\,\mathrm{d}t \\ &= \int_0^{\pi} \cos(nt)\cos(mt)\,\mathrm{d}t = 0, \qquad \text{if } n \neq m.\end{aligned}$$

In the same way, we see that

$$(T_0, T_0) = \pi, \qquad (T_n, T_n) = \frac{\pi}{2} \quad \text{if } n \geq 1.$$

Given a continuous function f on $[-1,1]$, we define $C_n f$ by

$$C_n(f)(x) = \sum_{k=0}^{n} c_k T_k$$

where

$$c_k = \frac{(T_k, f)}{(T_k, T_k)}.$$

The polynomial $C_n(f)$ is the unique polynomial of degree at most n which minimize the norm

$$\|f - p\|_C^2 = (f-p, f-p),$$

where p is a polynomial of degree at most n. We will see that if f has enough regularity, then $C_n(f)$ can also be used to find a polynomial which approximates f in the uniform norm.

Recall that the modulus of continuity of a function f is the function ω defined by

$$\omega(\delta) = \max\{\, |f(x) - f(y)| : |x-y| \leq \delta \,\}, \qquad \delta > 0.$$

THEOREM 5.14. *Let f be a continuous function on $[-1,1]$ and let ω denote its modulus of continuity. Then there exists a polynomial p of degree at most n such that*

$$\|f - p\| \leq \omega\Big(\frac{2\pi}{\sqrt[4]{n}}\Big) + \frac{\|f\|}{\sqrt{n}}.$$

PROOF. By a change of variable, $x = \cos t$, where $x \in [-1,1]$ and $t \in [0,\pi]$, we define the function g by $g(t) = f(\cos t)$. To approximate f by a polynomial is then equivalent to approximate g by a trigonometric polynomial. We may continue g to an even function on $[-\pi,\pi]$. We then have $g(-\pi) = g(\pi)$.

Let $u = -\pi + 2\pi t$ and consider $h(u) = g(-\pi + 2\pi u)$. Then $h\colon [0,1] \to \mathbb{R}$ and we can extend h to a periodic and continuous function $\mathbb{R} \to \mathbb{R}$. Let $\tilde{\omega}$ be the modulus of continuity of h. Then

$$\tilde{\omega}(\delta) \leq \omega(2\pi\delta)$$

since $\frac{\mathrm{d}u}{\mathrm{d}t} = 2\pi$ and $\big|\frac{\mathrm{d}u}{\mathrm{d}t}\big| \leq 2\pi$.

By Theorem 4.4,

$$\|h - \sigma_n(h)\| \le \tilde{\omega}(1/\sqrt[4]{n}) + \frac{\|h\|}{\sqrt{n}}.$$

Hence there is a polynomial p of degree at most n such that

$$\|f - p\| \le \omega\Big(\frac{2\pi}{\sqrt[4]{n}}\Big) + \frac{\|f\|}{\sqrt{n}}. \qquad \square$$

Using similar arguments to those used above, we will now prove that $C_n(f)$ approximates f well if f has enough regularity.

THEOREM 5.15. *Let f be a two times continuously differentiable function on $[-1,1]$. Then*

$$\|f - C_n(f)\| \le \frac{2\|f''\|}{n-1}.$$

PROOF. We write

$$C_n(f) = \sum_{k=0}^{n} c_k T_k, \qquad c_k = \frac{(T_k, f)}{(T_k, T_k)}.$$

We will estimate the size of c_k. With the change of variables $x = \cos t$, we have

$$\begin{aligned}
c_k(T_k, T_k) &= \int_{-1}^{1} T_k(x) f(x) \frac{1}{\sqrt{1-x^2}}\,\mathrm{d}x = \int_0^{\pi} \cos(kt) f(\cos t)\,\mathrm{d}t \\
&= \left[\frac{\sin(kt)}{k} f(\cos t)\right]_0^{\pi} + \int_0^{\pi} \frac{\sin(kt)}{k} \sin(t) f'(\cos t)\,\mathrm{d}t \\
&= 0 + \int_0^{\pi} \frac{1}{2k}(\cos((k-1)t) - \cos((k+1)t)) f'(\cos t)\,\mathrm{d}t \\
&= \frac{1}{2k}\left[\left(\frac{\sin((k-1)t)}{k-1} - \frac{\sin((k+1)t)}{k+1}\right) f'(\cos t)\right]_0^{\pi} \\
&\quad + \frac{1}{2k}\int_0^{\pi} \left(\frac{\sin((k-1)t)}{k-1} - \frac{\sin((k+1)t)}{k+1}\right) \sin(t) f''(\cos t)\,\mathrm{d}t \\
&= \frac{1}{2k}\int_0^{\pi} \left(\frac{\sin((k-1)t)}{k-1} - \frac{\sin((k+1)t)}{k+1}\right) \sin(t) f''(\cos t)\,\mathrm{d}t
\end{aligned}$$

Hence

$$\begin{aligned}
|c_k|(T_k, T_k) &\le \|f''\| \frac{1}{2k} \int_0^{\pi} \left|\left(\frac{\sin((k-1)t)}{k-1} - \frac{\sin((k+1)t)}{k+1}\right) \sin t\right| \mathrm{d}t \\
&\le \|f''\| \frac{\pi}{2k}\left(\frac{1}{k-1} + \frac{1}{k+1}\right) = \frac{\|f''\|\pi}{(k-1)(k+1)}.
\end{aligned}$$

This estimate implies that $\sum_{k=1}^{\infty} |c_k|$ converges, which in turn implies that $\tilde{f} = \lim_{n\to\infty} C_n(f)$ exists and that $\tilde{f}$ is continuous since T_k is continuous.

We have $f = \lim_{n\to\infty} C_n(f)$ because of Theorem 4.6. This is proved by arguing with the changes of variables as in the proof of Theorem 5.14.

Hence

$$f - C_n(f) = \sum_{k=n+1}^{\infty} c_k T_k.$$

Since $(T_k, T_k) = \frac{\pi}{2}$ for $k \geq 1$, we therefore have

$$\begin{aligned}\|f - C_n(f)\| &\leq \sum_{k=n+1}^{\infty} |c_k| \|T_k\| = \sum_{k=n+1}^{\infty} |c_k| \\ &\leq 2\|f''\| \sum_{k=n+1}^{\infty} \frac{1}{(k-1)(k+1)} \leq 2\|f''\| \int_n^{\infty} \frac{1}{(x-1)(x+1)}\,dx \\ &= 2\|f''\| \frac{1}{2} \log \frac{n+1}{n-1} = \|f''\| \log\Big(1 + \frac{2}{n-1}\Big) \leq \frac{2\|f''\|}{n-1}. \qquad \square\end{aligned}$$

There are many more results in the sprit of Theorems 5.14 and 5.15. Some can be found in the book by Cheney [3], Chapter 4, Section 6.

EXERCISE 5.8. Prove that T_n is even if n is even and that T_n is odd if n is odd.

EXERCISE 5.9. Show that $T_0(x) = 1$ and $T_1(x) = x$ and that

$$T_n(x) = 2x T_{n-1}(x) - T_{n-2}(x).$$

EXERCISE 5.10. Prove that $T_n(T_m(x)) = T_{n+m}(x)$.

CHAPTER 6

Rational approximation

1. Padé approximation

Suppose that $f\colon [-1,1] \to \mathbb{R}$ is $n+1$ times differentiable. By Taylor's formula, there exists for each x a ξ with $|\xi| < x$ such that

$$f(x) = \sum_{k=1}^{n} \frac{f^{(k)}(0)}{k!} x^k + \frac{f^{(n+1)}(\xi)}{(n+1)!} x^{n+1}.$$

Hence, if $f^{(n+1)}$ is bounded, then the *Taylor polynomial* $\sum_{k=1}^{n} \frac{f^{(k)}(0)}{k!} x^k$ is a good approximation of f is $|x|$ is small. More precisely, in case $f^{(n+1)}$ is bounded, then

$$\left| f(x) - \sum_{k=1}^{n} \frac{f^{(k)}(0)}{k!} x^k \right| \leq C|x|^{n+1}$$

holds for some constant C. (We may take $C = \frac{\|f^{(n+1)}\|}{(n+1)!}$.)

We would like to achive something similar for approximations by *rational functions* instead of by polynomials. More precisely, suppose that we desire to approximate f by a rational function

$$g(x) = \frac{\sum_{k=0}^{n} a_k x^k}{\sum_{k=0}^{m} b_k x^k}.$$

Then we say that g is a *Padé*[1] *approximation of order* (n,m) to f if there is a constant $l > n$ and a constant $C > 0$ such that

$$|f(x) - g(x)| \leq C|x|^l$$

holds for all x in some interval $(-r,r)$. The question is if such an approximation is possible, and if so, how do we find the coefficients a_k and b_k? Here is a theorem.

THEOREM 6.1 (Padé approximation). *Assume that f is $m+n+1$ times differentiable on $(-r,r)$ and that $f^{(n+m+1)}$ is bounded on $(-r,r)$. Let*

$$g(x) = \frac{\sum_{k=0}^{n} a_k x^k}{\sum_{k=0}^{m} b_k x^k}$$

where a_k and b_k satisfy the equations

$$b_0 \neq 0, \qquad a_k = 0 \text{ when } k > n \qquad\qquad b_k = 0 \text{ when } k > m$$

and

$$a_k = \sum_{j=0}^{k} b_j \frac{f^{(k-j)}(0)}{(k-j)!}, \qquad k = 1, 2, \ldots, l-1.$$

[1] Henri Padé, 1863–1953. French mathematician.

Then g is a Padé approximation of order (n,m) to f, that is

$$|f(x) - g(x)| \leq C|x|^l$$

holds for some $l > n$ and $C > 0$.

PROOF. Multiplying by $\sum b_k x^k$, the inequality $|f(x) - g(x)| \leq C|x|^l$ can be written as

$$\left| f(x) \sum_{k=0}^{m} b_k x^k - \sum_{k=0}^{n} a_k x^k \right| \leq C|x|^l \left| \sum_{k=0}^{m} b_k x^k \right|.$$

Since we require that $b_0 \neq 0$, the expression $\left|\sum_{k=0}^{m} b_k x^k\right|$ is bounded away from zero on some interval $(-r, r)$. Hence it is sufficient to prove that

$$\left| f(x) \sum_{k=0}^{m} b_k x^k - \sum_{k=0}^{n} a_k x^k \right| \leq C_1|x|^l$$

holds for some constant $C_1 > 0$ and some $l > n$.

By Taylor's formula, we can write

$$f(x) = \sum_{k=0}^{n+m} \frac{f^{(k)}(0)}{k!} x^k + R_{n+m+1}(x),$$

where $|R_{n+m+1}(x)| \leq C_2|x|^{n+m+1}$. We therefore have

$$\begin{aligned}
&\left| f(x) \sum_{k=0}^{m} b_k x^k - \sum_{k=0}^{n} a_k x^k \right| \\
&\leq \left| \left(\sum_{k=0}^{n+m} \frac{f^{(k)}(0)}{k!} x^k \right) \left(\sum_{k=0}^{m} b_k x^k \right) - \sum_{k=0}^{n} a_k x^k \right| + |R_{n+m+1}(x)| \left| \sum_{k=0}^{m} b_k x^k \right| \\
&\leq \left| \left(\sum_{k=0}^{n+m} \frac{f^{(k)}(0)}{k!} x^k \right) \left(\sum_{k=0}^{m} b_k x^k \right) - \sum_{k=0}^{n} a_k x^k \right| + C_3|x|^{n+m+1}.
\end{aligned}$$

Hence it is sufficient to prove that

$$\left| \left(\sum_{k=0}^{n+m} \frac{f^{(k)}(0)}{k!} x^k \right) \left(\sum_{k=0}^{m} b_k x^k \right) - \sum_{k=0}^{n} a_k x^k \right| \leq C_4|x|^l.$$

But this estimate is satisfied for some $C_4 > 0$ if the coefficient of x^k in the polynomial

$$\left(\sum_{k=0}^{n+m} \frac{f^{(k)}(0)}{k!} x^k \right) \left(\sum_{k=0}^{m} b_k x^k \right) - \sum_{k=0}^{n} a_k x^k$$

is zero for all $k < l$.

If we define $a_k = 0$ when $k > n$ and $b_k = 0$ when $k > m$, then the conditions that need to be satisfied are exactly those mentioned in the statement of the theorem. □

Note that the Padé approximation need not be unique. For instance, we may multiply both a_k and b_k by any non-zero constant. In particular,

we may for instance require that $b_0 = 1$. If we do so, then we can write the conditions on a_k and b_k as

$$\begin{bmatrix} 1 & 0 & \dots & 0 & -f(0) & 0 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 & -f'(0) & -f(0) & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & 1 & -\frac{f^{(n)}(0)}{n!} & -\frac{f^{(n-1)}(0)}{(n-1)!} & \dots & \dots & -f(0) \\ 0 & \dots & \dots & 0 & -\frac{f^{(n+1)}(0)}{(n+1)!} & -\frac{f^{(n)}(0)}{n!} & \dots & \dots & -f'(0) \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & \dots & 0 & -\frac{f^{(l-1)}(0)}{(l-1)!} & -\frac{f^{(l-2)}(0)}{(l-2)!} & \dots & \dots & -\frac{f^{(l-n-1)}(0)}{(n-l-1)!} \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \\ 1 \\ b_1 \\ \vdots \\ b_m \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

EXAMPLE 6.2. Let us consider $f(x) = \sin x$. Put $n = m = 2$. We are looking for an approximation of the form

$$g(x) = \frac{a_0 + a_1 x + a_2 x^2}{b_0 + b_1 x + b_2 x^2}.$$

Put $l = 5$. We choose $b_0 = 1$. Evaluating $f^{(k)}(0)$, we see that we need to find a solution to the system

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 & -1 & 0 \\ 0 & 0 & 0 & \frac{1}{3!} & 0 & -1 \\ 0 & 0 & 0 & 0 & \frac{1}{3!} & 0 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ 1 \\ b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}.$$

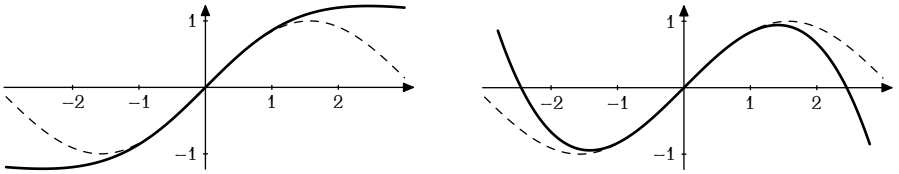
The last two equation implies that $b_1 = 0$ and $b_2 = \frac{1}{6}$. The remaining equations then implies that $a_0 = 0$, $a_1 = 1$ and $a_2 = 0$. Hence

$$g(x) = \frac{x}{1 + \frac{1}{6}x^2} = \frac{6x}{6 + x^2}$$

is a Padé approximation to $\sin x$ and

$$\left| \sin x - \frac{6x}{6 + x^2} \right| \leq C|x|^4.$$

To show that this Padé approximation is rather good, we plot it (to the left) and compare with the approximation by the Taylor polynomial of order 4 (to the right).



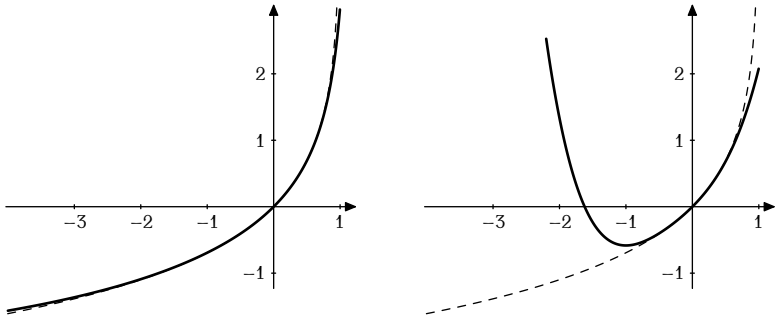
The approximations are about equaly good, which should not be too surprising, since in both cases, the error is of the order $C|x|^5$. (The constant of course, could be very different in the two cases.)

EXAMPLE 6.3. We consider the Padé approximation of order $(2,2)$ to $f(x) = -\log(1-x)$. Computations similar to those above, yield that

$$g(x) = \frac{6x - 3x^2}{6 - 6x + x^2}$$

is the desired Padé approximation with $l = 5$.

It is well known that the Taylor series for f does not converge outside the interval $[-1,1]$. Therefore, the Taylor polynomial cannot be used to approximate f to the left of -1. The Padé approximation however gives a very useful approximation to the left of -1. The pictures below show the graph of g (left) and the graph of the Taylor polynomial of order 4 (right).



EXERCISE 6.1. Find Padé approximations of order $(2,2)$ to the functions $\cos x$ and e^x.

EXERCISE 6.2. Let $f(x) = \sin(x) + \cos(x)$. Let a be the smallest positive root of a. Use Padé approximations to find approximations of a of the form $\sqrt{\frac{p}{q}}$ where $\frac{p}{q}$ is a rational number. Hint: Use Padé approximations of order $(2,m)$.

2. Continued fractions

In the previous section, we studied how to find certain rational approximations of a function, namely the Padé approximations. We will now turn these rational approximations into continued fractions. This can be done by succesive polynomial division.

EXAMPLE 6.4. Consider the Padé approximation

$$g(x) = \frac{6x - 3x^2}{6 - 6x + x^2}$$

of $f(x) = -\log(1-x)$. By successive polynomial division, we obtain

$$\begin{aligned}\frac{6x - 3x^2}{6 - 6x + x^2} &= -3 + \frac{18 - 12x}{6 - 6x + x^2} = -3 + \frac{6}{(6 - 6x + x^2)/(3 - 2x)} \\ &= -3 + \frac{6}{\frac{9}{4} - \frac{1}{2}x + \frac{-\frac{3}{4}}{3 - 2x}} = -3 + \frac{24}{9 - 2x + \frac{3}{2x + 3}}.\end{aligned}$$

Alternatively, we have

$$\frac{6x-3x^2}{6-6x+x^2} = -3 + \frac{18-12x}{6-6x+x^2} = -3 + \frac{1}{(6-6x+x^2)/(18-12x)}$$
$$= -3 + \frac{1}{\frac{3}{8} - \frac{1}{12}x + \frac{-\frac{3}{4}}{18-12x}} = -3 + \frac{1}{\frac{3}{8} - \frac{1}{12}x + \frac{1}{16x-24}}.$$

Hence, in general, we may write

$$\frac{a_0 + a_1x + \ldots + a_nx^n}{b_0 + b_1x + \ldots + b_mx^m} = P_0(x) + \cfrac{c_1}{P_1(x) + \cfrac{c_2}{P_2(x) + \cfrac{c_3}{\ddots + \cfrac{c_k}{P_k(x)}}}},$$

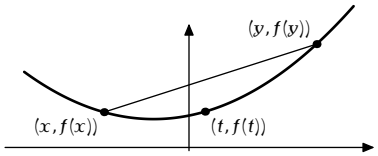
where $P_0, P_1, \ldots, P_k$ are polynomials and $c_1, c_2, \ldots, c_k$ are constants. This has a computational advantage in that it if the constants are suitably chosen, then the evaluation can be done with not more than $\max\{n, m\}$ operations of multiplications or divisions, whereas an evaluation of the power series of order $n + m$ requires $n + m - 1$ such operations. This makes the evaluation of the continued fraction faster than that of a power series, which is important in some applications. See Cheney [3], Chapter 5, for more details.

CHAPTER 7

Inequalities

1. Convexity and Jensen's inequality

DEFINITION 7.1 (Convex function). A function $f\colon [a,b] \to \mathbb{R}$ is *convex* if, whenever $x, y, t \in [a,b]$ are points with $x \le t \le y$, the point $(t, (f))$ does not lie below the straight line through the points $(x, f(x))$ and $(y, f(y))$.



THEOREM 7.2 (Jensen's[1] inequality). *Let α be increasing on $[a,b]$ and ϕ a convex function. If f and $\phi \circ f$ are integrable with respect to α, then*

$$\frac{1}{\alpha(b)-\alpha(a)} \int_a^b \phi \circ f \, d\alpha \ge \phi\left(\frac{1}{\alpha(b)-\alpha(a)} \int_a^b f \, d\alpha \right).$$

PROOF. Since α is increasing, we have by Exercise 2.4 that $\int_a^b g \, d\alpha \le \int_a^b h \, d\alpha$ if $g \le h$, a property that we will soon make use of.

Let t be fixed. Since ϕ is convex, there is a number λ such that the line through the point $(t, \phi(t))$, given by

$$y = \lambda(x-t) + \phi(t),$$

lies below the graph of ϕ. Hence

$$\phi(x) \ge \lambda(x-t) + \phi(t)$$

Let $m = \alpha(b) - \alpha(a)$. We put $x = f(z)$ and $t = \frac{1}{m} \int_a^b f \, d\alpha$ and obtain

$$\phi(f(z)) \ge \lambda\left(f(z) - \frac{1}{m} \int_a^b f \, d\alpha \right) + \phi\left(\int_a^b f \, d\alpha \right).$$

The number λ does not depend on z, so integrating over z with respect to α, we get

$$\begin{aligned} \int_a^b \phi(f(z)) \, d\alpha &\ge \lambda\left(\int_a^b f(z) - \frac{1}{m} \int_a^b f \, d\alpha \right) + \int_a^b \phi\left(\frac{1}{m} \int_a^b f \, d\alpha \right) d\alpha \\ &= \phi\left(\frac{1}{m} \int_a^b f \, d\alpha \right) \int_a^b d\alpha = \phi\left(\frac{1}{m} \int_a^b f \, d\alpha \right) m. \end{aligned}$$

□

[1] Johan Ludwig William Valdemar Jensen, 1859–1925. Danish mathematician.

We will now study a special consequence of Jensen's inequality. Suppose that $x_1, x_2, \ldots, x_n$ are non-negative numbers. We may form the *arithmetic mean*

$$A_n = \frac{1}{n}\sum_{k=1}^{n} x_k, \tag{7.1}$$

as well as the *geometric mean*

$$G_n = \sqrt[n]{x_1 x_2 \ldots x_n}. \tag{7.2}$$

It is sometimes also natural to consider the *harmonic mean*

$$H_n = \frac{n}{\sum_{k=1}^{n} \frac{1}{x_k}}. \tag{7.3}$$

THEOREM 7.3 (The arithmetic–geometric–harmonic mean inequality). *Suppose that $x_1, x_2, \ldots, x_n$ are non-negative numbers. Then*

$$H_n \leq G_n \leq A_n,$$

where A_n, G_n and H_n are given by (7.1)–(7.3).

PROOF. By Jensen's inequality (see Exercise 7.1), we have

$$\log G_n = \frac{1}{n}\log\prod_{k=1}^{n} x_k = \frac{1}{n}\sum_{k=1}^{n}\log(x_k) \leq \log\left(\frac{1}{n}\sum_{k=1}^{n} x_k\right) = \log A_n$$

since log is *concave*. Since log is strictly increasing, this implies the inequality $G_n \leq A_n$.

It remains to prove that $H_n \leq G_n$. By Jensen's inequality we have

$$\log H_n = -\log\left(\frac{1}{n}\sum_{k=1}^{n}\frac{1}{x_k}\right) \leq -\frac{1}{n}\sum_{k=1}^{n}\log\frac{1}{x_k} = \frac{1}{n}\sum_{k=1}^{n}\log x_k = \log G_n,$$

which proves that $H_n \leq G_n$ since log is strictly increasing. □

EXERCISE 7.1. Suppose that ϕ is a convex function and let $(x_k)_{k=1}^n$ be a sequence of n real numbers. Make a particular choice of α and f in Jensen's inequality, and conclude that

$$\frac{1}{n}\sum_{k=1}^{n}\phi(x_k) \geq \phi\left(\frac{1}{n}\sum_{k=1}^{n} x_k\right).$$

EXERCISE 7.2. On his way to work, Dr. Overgaard observes that his bicycle travels at the speed 20 km/h half of the distance and 40 km/h the other half. The distance is 20 km. Prove, without using any artificial aids such as paper and pencil, that it takes less than 40 minutes for Dr. Overgaard to travel from home to work. (Theorem 7.3 is not considered artificial aid.)

2. Some inequalities for integrals

THEOREM 7.4 (Hölder's inequality). *Let α be increasing on I and suppose that $p > 1$ and $q > 1$ are numbers such that $p^{-1} + q^{-1} = 1$. Then*

$$\int_I |fg| \, d\alpha \leq \left(\int_I f^p \, d\alpha \right)^{\frac{1}{p}} \left(\int_I g^q \, d\alpha \right)^{\frac{1}{q}},$$

provided that the integrals exist.

PROOF. We will use Young's[2] inequality for products (there is also a Young's inequality for convolutions): If $p, q > 1$ and $\frac{1}{p} + \frac{1}{q} = 1$, then

$$ab \leq \frac{a^p}{p} + \frac{b^q}{q}, \qquad a, b \geq 0.$$

(See Exercise 7.3.)

We assume that $\int_I |f|^p \, d\alpha \neq 0$ and $\int_I |g|^q \, d\alpha \neq 0$, and put

$$f_1 = \frac{f}{\left(\int_a^b |f|^p \, d\alpha \right)^{\frac{1}{p}}}, \qquad g_1 = \frac{f}{\left(\int_a^b |g|^q \, d\alpha \right)^{\frac{1}{q}}}.$$

Then

$$\int_I |f_1|^p \, d\alpha = \int_I |g_1|^q \, d\alpha = 1.$$

Using Young's inequality, we have

$$|f_1(x) g_1(x)| \leq \frac{|f_1(x)|^p}{p} + \frac{|g_1(x)|}{q}.$$

Integrating (using Exercise 2.4), we get

$$\int_I |f_1 g_1| \, d\alpha \leq \frac{1}{p} + \frac{1}{q} = 1.$$

But

$$\int_I |f_1 g_1| \, d\alpha = \int_I |fg| \frac{1}{\left(\int_I |f|^p \, d\alpha \right)^{\frac{1}{p}} \left(\int_I |g|^q \, d\alpha \right)^{\frac{1}{q}}} \, d\alpha,$$

and Hölder's inequality follows. □

A consequence of Hölder's inequality is the following. Assume that $p, q > 1$ and that $\frac{1}{p} + \frac{1}{q} = 1$. The space $\mathcal{L}^p(I, \alpha)$ is the space of all functions f such that

$$\|f\|_p := \left(\int_I |f|^p \, d\alpha \right)^{\frac{1}{p}} < \infty.$$

Let $g \in \mathcal{L}^q(I, \alpha)$ and consider the linear operator $L_g \colon \mathcal{L}^p(I, \alpha) \to \mathbb{R}$ defined by

$$L_g(f) = \int_I fg \, d\alpha$$

Then Hölder's inequality implies that

$$|L_g(f)| = \int_I |fg| \, d\alpha \leq \|f\|_p \|g\|_q.$$

[2] William Henry Young, 1863–1942. English mathematician.

Hence

$$\|L_g\| := \sup_{f \in \mathcal{L}^p(I,\alpha), \|f\|_p = 1} |L_g(f)| \leq \|g\|_q,$$

and L_g is therefore a continuous linear operator.

There is a converse to this result, which we will not prove however: If $L_g \colon \mathcal{L}^p(I,\alpha) \to \mathbb{R}$ is a continuous linear operator, then there is a $g \in \mathcal{L}^q(I,\alpha)$ such that

$$L_g(f) = \int_I fg \, d\alpha.$$

This is similar to Riesz' representation theorem (Theorem 2.11), and is often also called Riesz' representation theorem.

A special case of Hölder's inequality is the following. (Let $p = q = 2$.)

THEOREM 7.5 (the Cauchy[3]–Bunyakovsky[4]–Schwarz[5] inequality). *Let α be increasing on I. Then*

$$\int_I |fg| \, d\alpha \leq \left(\int_I f^2 \, d\alpha \right)^{\frac{1}{2}} \left(\int_I g^2 \, d\alpha \right)^{\frac{1}{2}},$$

provided that the integrals exist.

The following inequality is much used in probability, perhaps mostly through Corollary 7.7.

THEOREM 7.6 (Markov's[6] inequality). *Let $f \colon I \to [0,\infty)$ be Riemann–Stieltjes integrable with respect to the increasing function α, and let $a > 0$. Then*

$$\int_I \chi_{\{ x : f(x) \geq a \}} \, d\alpha \leq \frac{1}{a} \int_I f \, d\alpha,$$

provided that the integrals exist.

PROOF. The function $\chi_{\{ x : f(x) \geq a \}}$ satisfies

$$\chi_{\{ x : f(x) \geq a \}} \leq f,$$

since if $x \in \{ x : f(x) \geq a \}$ then

$$\chi_{\{ x : f(x) \geq a \}} = a \leq f(x),$$

and if $x \notin \{ x : f(x) \geq a \}$ then

$$\chi_{\{ x : f(x) \geq a \}} = 0 \leq f(x).$$

Since α is increasing (Exercise 2.4), we have

$$\int_I a \chi_{\{ x : f(x) \geq a \}} \, d\alpha \leq \int_I f \, d\alpha. \qquad \square$$

A simple but important corollary of Markov's inequality is the following inequality, which has many applications in probability.

---

[3] Augustin-Louis Cauchy, 1789–1857. French mathematician.

[4] Viktor Bunyakovsky, 1804–1889. Russian mathematician.

[5] Hermann Schwarz, 1843–1921. German mathematician.

[6] Andrey Andreyevich Markov, 1856–1922. Russian mathematician.

COROLLARY 7.7 (Chebyshev's inequality). *Suppose X is a random variable which takes values in a compact interval I. Let $\mu = \mathrm{E}(X)$ be the expectation and $\sigma^2 = \mathrm{E}((X-\mu)^2)$ the variance of X. Then*

$$\mathrm{P}(|X-\mu| \geq a) \leq \frac{\sigma^2}{a^2}.$$

PROOF. The inequality follows immediately once we have translated the language of probability into that of Theorem 7.6.

Let $\alpha(a) = \mathrm{P}(X \leq a)$. Then α is increasing on I. If necessary, we can replace I by a larger interval so that α is zero at the left end-point and one at the right end-point of I. We then have

$$\mu = \int_I x \,\mathrm{d}\alpha(x), \qquad \sigma^2 = \int_I (x-\mu)^2 \,\mathrm{d}\alpha.$$

Let $f(x) = (x-\mu)^2$. Since $|x-\mu| \geq a$ if and only if $(x-\mu)^2 \geq a^2$, we have

$$\mathrm{P}(|X-\mu| \geq a) = \int_I \chi_{\{\,x:|x-\mu|\geq a\,\}} \,\mathrm{d}\alpha = \int_I \chi_{\{\,x:f(x)\geq a^2\,\}} \,\mathrm{d}\alpha.$$

Now, Theorem 7.6 implies that

$$\mathrm{P}(|X-\mu| \geq a) = \int_I \chi_{\{\,x:f(x)\geq a^2\,\}} \,\mathrm{d}\alpha \leq \frac{1}{a^2}\int_I f \,\mathrm{d}\alpha = \frac{\sigma^2}{a^2}. \qquad \square$$

THEOREM 7.8 (Chebyshev's sum inequality). *Let α be increasing on $I = [a,b]$ and suppose that f and g are functions on I that are either both increasing or both decreasing. Then*

$$(\alpha(b)-\alpha(a)) \int_I fg \,\mathrm{d}\alpha \geq \int_I f \,\mathrm{d}\alpha \int_I g \,\mathrm{d}\alpha,$$

provided that the integrals exist.

PROOF. If f is increasing, then

$$f(x) - f(y) \geq 0 \qquad \Leftrightarrow \qquad x - y \geq 0,$$

and if f is decreasing, then

$$f(x) - f(y) \geq 0 \qquad \Leftrightarrow \qquad x - y \leq 0.$$

The corresponding statements are of course true also for g. From these statement we see that we always have

$$(f(x)-f(y))(g(x)-g(y)) \geq 0$$

if either f and g are both increasing or both decreasing, since in that case, the two factors above are always of the same sign.

Since α is increasing we get

$$\int_I \int_I (f(x)-f(y))(g(x)-g(y)) \,\mathrm{d}\alpha(x) \,\mathrm{d}\alpha(y) \geq 0.$$

A rearrangement yields

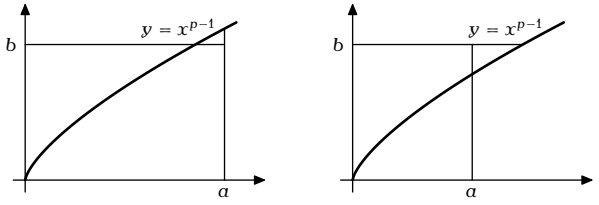
$$\int_I \int_I f(x)g(x) \,\mathrm{d}\alpha(x) \,\mathrm{d}\alpha(y) + \int_I \int_I f(y)g(y) \,\mathrm{d}\alpha(x) \,\mathrm{d}\alpha(y) \geq \int_I \int_I f(x)g(y) \,\mathrm{d}\alpha(x) \,\mathrm{d}\alpha(y) + \int_I \int_I f(y)g(x) \,\mathrm{d}\alpha(x) \,\mathrm{d}\alpha(y).$$

We obtain

$$2(\alpha(b) - \alpha(a)) \int_I fg \, d\alpha \geq 2 \int_I f \, d\alpha \int_I g \, d\alpha,$$

from which the theorem follows. □

EXERCISE 7.3. Prove Young's inequality. Consider for a fixed $a > 0$ the function $f(x) = \frac{a^p}{p} + \frac{x^q}{q} - ax$ or consider the following picture.



How do the pictures show that $ab \leq \int_0^a x^{p-1} \, dx + \int_0^b y^{\frac{1}{p-1}} \, dy$?

EXERCISE 7.4. Let $a_1, a_2, \ldots, a_n$ and $b_1, b_2, \ldots, b_n$ be real numbers. Make a particular choice of f, g and α in Theorem 7.5 and conclude that

$$\sum_{k=1}^{n} |a_k b_k| \leq \left(\sum_{k=1}^{n} a_k^2 \right) \left(\sum_{k=1}^{n} b_k^2 \right).$$

Prove Cauchy's inequality

$$\left| \sum_{k=1}^{n} \overline{\alpha_k} \beta_k \right|^2 \leq \left(\sum_{k=1}^{n} |\alpha_k|^2 \right) \left(\sum_{k=1}^{n} |\beta_k|^2 \right),$$

where α_k and β_k are complex numbers.

EXERCISE 7.5. Suppose you make n independent tosses with a coin such that the probability of heads and tails are equal. Show that the probability that the outcome deviates from equal number of heads and tails by k or more, is not larger than

$$\frac{n}{4k^2}.$$

In particular, if you toss a coin 100 times, then with probability at least $\frac{3}{4}$, you will have between 41 and 59 heads.

Bibliography

[1] T. Apostol, *Mathematical Analysis*, Second edition, Addison-Wesley Publishing Co., Reading, 1974.

[2] M. Carter, B. van Brunt, *The Lebesgue–Stieltjes integral. A practical introduction*, Undergraduate Texts in Mathematics, Springer-Verlag, New York, 2000, ISBN: 0-387-95012-5.

[3] E. W. Cheney, *Introduction to Approximation Theory*, McGraw-Hill Book Co., New York, 1966.

[4] P. J. Davis, P. Rabinowitz, *Methods of numerical integration*, Academic press, New York, 1975.

[5] G. H. Hardy, *Theorems relating to the summability and convergence of slowly oscillating series*, Proceedings of the London Mathematical Society (2) 8, 1910, 301–320.

[6] E. Landau, *Über die Bedeutung einiger neuen Grenzwertsätze der Herren Hardy und Axer*, Prace Matematyczno-Fizyczne 21, 1910, 97–177.

[7] D. H. Lehmer, *On the maxima and minima of Bernoulli polynomials*, American Mathematical Monthly 47, (1940), 533–538.

[8] J. Malmquist, V. Stenström, S. Danielson, *Matematisk analys*, Natur och Kultur, Stockholm, 1953.

[9] R. M. McLeod, *The generalized Riemann integral*, Carus Mathematical Monographs, no. 20, Mathematical Association of America, Washington, 1980.

[10] Niven, *Irrational numbers*, The Carus Mathematical Monographs, No. 11, The Mathematical Association of America, New York, 1956.

[11] F. Riesz, B. Szőkefalvi-Nagy, *Functional analysis*, Frederick Ungar Publishing Co., New York, 1955.

[12] J. H. Shapiro, *Volterra adventures*, Student Mathematical Library 85, American Mathematical Society, Providence, 2018, ISBN: 978-1-4704-4116-6.

Index

Contents