Contents

Notation

Sets

$\mathbb{N}$	the set $\{1, 2, 3, \ldots\}$ of all natural numbers
$\mathbb{Z}$	the set $\{\ldots, -1, 0, 1, 2, 3, \ldots\}$ of all integers
$\mathbb{Q}$	the set of all rational numbers
$\mathbb{R}$	the set of all real numbers
$\mathbb{C}$	the set of all complex numbers
$n!$	the product $n(n-1)\cdots 1$ pronounced as n factorial
$\{n\}$	the set $\{1, 2, \ldots, n\}$
$\emptyset$	the empty set
$a \in A$	a is an element of A
$a \notin A$	a is no element of A
$\lvert A\rvert$	the number of elements of the set A, the *cardinality* of A
$A \subseteq B$	A is a subset of B
$A \supseteq B$	the set A contains the set B, i.e. $B \subseteq A$
$\{x \in A : \cdots\}$	the set of all $x \in A$ for which $\cdots$
$\{x : \cdots\}$	the set of all x for which $\cdots$
$A \cup B$	the union of the sets A and B
$A \cap B$	the intersection of the sets A and B
$\bigcup_{i\in I} A_i$	the elements in *at least* one set A_i
$\bigcap_{i\in I} A_i$	the elements which are in *all* sets A_i
$A \setminus B$	the set of elements in A which are not in B
$\mathcal{P}(A)$	the *power set of* A is the family of all subsets of A
$[A]^k$	the collection of all subsets of A with cardinality k
$A \times B$	the Cartesian product of A and B: $A \times B = \{(a, b) \mid a \in A \text{ and } b \in B\}$
A^n	the n-fold Cartesian product of A : $A^n = A \times A \cdots \times A = \{(a_1, \ldots, a_n) \mid a_1, \ldots, a_n \in A\}$
$\sum_{x\in S} f(x)$	the sum of all $x \in S$ of the $f(x)$

Complex numbers

We noteren complexe getallen als $z = a + bi$ met $a, b \in \mathbb{R}$.

$\operatorname{Re} z$	the real part of z (is equal to a)
$\operatorname{Im} z$	the imaginary part of z (is equal to b)
$\lvert z \rvert$	the modulus of z (is equal to $\sqrt{a^2 + b^2}$)
$\operatorname{Arg} z$	the principal value of the argument of z
$\overline{z}$	the complex conjugate of z (is equal to $a - bi$)

Counting

$\binom{n}{k}$	the binomial coefficient $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ has to be pronounced as 'n choose k'
$\left\lvert {n \atop k} \right\rvert$	the number of surjections of $\{1, 2, \ldots, n\}$ to $\{1, 2, \ldots, k\}$

Probability and Statistics

A^c	the *complement* of A is the set $\{\omega \in \Omega : \omega \notin A\}$
$P(A)$	the probability of event A
$P(A \mid B)$	de *conditional probability* A given B (is equal to $\frac{P(A \cap B)}{P(B)}$)

1

Graphs

Author: D. Gijswijt

Introduction

Graph theory is an important subfield of *Discrete mathematics*. Apart from being a mathematical discipline in its own right, it is known for its many applications in a wide range of disciplines including computer science, chemistry, physics, social and economic sciences and, of course, in pure mathematics. The word *graph* was introduced in 1878 by the English mathematician J. J. Sylvester as a shorthand for 'graphic representation'. He used these graphical representations to depict the structure of organic molecules.

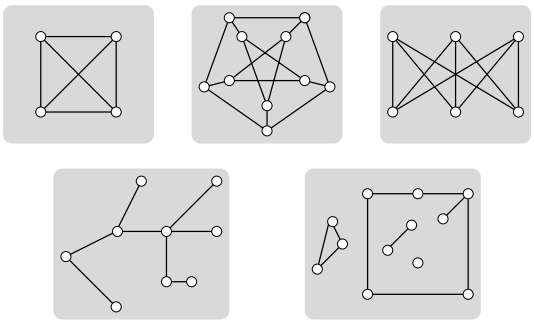


Figure 1.1: Five different graphs

1.1 Graphs

Simply put, a graph consists of a number of objects and connections between these objects. Examples are:

- road networks: pairs of cities are connected by roads,
- the World Wide Web: pairs of webpages are connected through hyperlinks,
- social networks: two people are connected if they are friends,

- Rubik's Cube: two cube configurations are connected if they differ by a singe move.

The objects of a graph are called *nodes* (or *vertices*) and the connections are called *edges*. The only information contained in a graph is the set of objects and which pairs of objects form an edge[1]. All other information is 'forgotten'. In this way, a road network and a social network might be represented by the same abstract graph even though the original meaning is completely different in the two cases. This abstraction allows one to focus on the underlying logical structure without being distracted by irrelevant details.

Graphically, a graph is represented by drawing dots or small circles for the nodes, and drawing lines or curves between pairs of nodes to denote the edges. Figure 1.1 shows five examples of graphs.

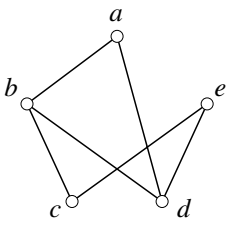
After this informal introduction, we are ready to give a precise, mathematical definition of graphs.

Definition 1.1. A graph G is a pair (V, E), where V is a finite set and E is a set of unordered pairs from V. The elements of V are called the *nodes* of G and the elements of E are the *edges* of G.

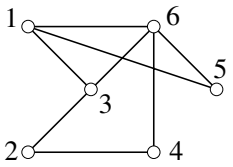
Example 1.2. Let $G = (V, E)$ be the graph where

$$\begin{aligned} V &= \{a, b, c, d, e\} \\ E &= \{\{a, b\}, \{a, d\}, \{b, c\}, \{b, d\}, \{c, e\}, \{d, e\}\}\,. \end{aligned}$$

This graph is graphically represented as follows.



Exercise 1.1 Consider the graph (V, E) with $V = \{1, 2, 3, 4, 5, 6\}$ represented by the following drawing.



Write down the set of edges E.

Exercise 1.2 There are exactly 8 graphs with node set $\{1, 2, 3\}$:

Determine the number of graphs with node set $\{1, 2, 3, 4, 5\}$.

[1] In some applications some additional data is preserved/added, for example the lengths of the roads in the network or capacities of cables in an electrical grid.

Variations on the notion of graph

In some situations, it is convenient to use a slightly more general notion of graph. In *multi graphs* we allow multiple edges between a given pair of nodes and we may also allow 'loops' (edges connecting a node to itself). If we want to consider directed edges, we can use *directed graphs* (or *digraphs* for short). The edges are called *arcs* and are specified by two nodes: the head and the tail. In Figure 1.2 you can see some examples.

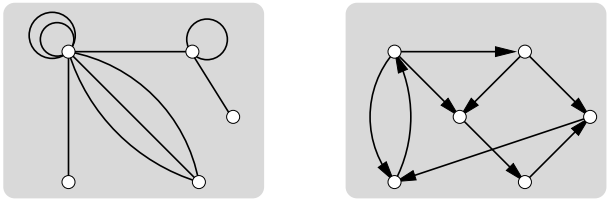


Figure 1.2: A multi graph and a digraph.

In this chapter, we will only consider graphs that are *simple* (no loops and multiple edges) and *undirected*, unless mentioned otherwise.

Degrees

If $e = \{u, v\}$ is an edge of graph G, we say that e is *incident* to u and v. Conversely, we also say that u and v are incident to e. The nodes u and v are called *neighbours*.

The *degree* $d(v)$ of a node v is the number of edges incident to v. Since our graphs are simple, $d(v)$ is also equal to the number of neighbours of v. When all vertices of G have the same degree, the graph is called *regular*. When all nodes have degree k, the graph is said to be *k-regular*.

In Figure 1.3 two regular graphs are drawn. The graph on the left is C_5, the *cyclic graph on* 5 *nodes*. The graph on the right is K_5, the *complete graph on* 5 *nodes*. The word 'complete' refers to the fact that every pair of nodes is an edge.

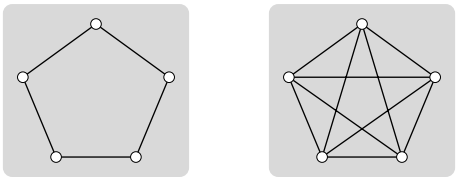


Figure 1.3: The cyclic graph C_5 is 2-regular and the complete graph K_5 is 4-regular.

Exercise 1.3 Denote by C_n the cyclic graph on n nodes ($n \geq 3$) and by K_n the complete graph on n nodes ($n \geq 1$). Determine the number of edges of C_n and K_n (as a function of n).

Our first theoretical result links the degrees of the nodes to the total number of edges. This simple but powerful statement is called the *handshaking lemma*.

Lemma 1.3 (Handshaking lemma)**.** *Let $G = (V, E)$ be a graph. Let $m := |E|$ be the number of edges of G. We have the following relation.*

$$\sum_{v \in V} d(v) = 2m.$$

Proof. Every edge is incident to exactly two nodes. Therefore, every edge contributes 2 to the sum of the degrees of the nodes, which implies that this sum equals $2m$. □

Exercise 1.4 The graph G has 14 nodes and 25 edges. Every node has degree 3 or degree 5. How many nodes have degree 3?

Exercise 1.5 Let $f(n)$ be the number of 2-regular graphs with node set $\{1, 2, \ldots, n\}$. Convince yourself of the fact that $f(1) = f(2) = 0$, $f(3) = 1$, and $f(4) = 3$. Determine $f(5)$.

Exercise 1.6 Draw a 3-regular graph on 17 nodes or show that such a graph does not exist.

Exercise 1.7 Show that every graph has an *even* number of nodes of *odd* degree.

Exercise 1.8 Let G be a graph on $n \geq 2$ nodes. Prove that G has two nodes of the same degree.

Paths and walks

In this section, we consider ways of walking along the edges of a graph.

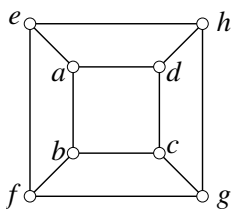
Definition 1.4. Let a and b be two nodes of a graph G. A *walk* from a to b is a sequence $W = (v_0, v_1, \ldots, v_k)$ of nodes of G, such that $a = v_0$, $b = v_k$, and $\{v_0, v_1\}, \{v_1, v_2\}, \ldots, \{v_{k-1}, v_k\}$ are edges of G.

We say that W *traverses* the nodes $v_0, v_1, \ldots v_k$ and the edges $\{v_0, v_1\}, \{v_1, v_2\}, \ldots, \{v_{k-1}, v_k\}$. The number k is called the *length* of the walk. When $a = b$, we say that W is a *closed* walk.

We want to stress that the nodes in the walk need not be distinct and that edges can be tranversed more than once. The case where alle nodes in the walk are distinct deserves a special name.

Definition 1.5. Let $W = (v_0, v_1, \ldots, v_k)$ be a walk. We say that W is a *path* if no two of the nodes $v_0, v_1, \ldots, v_k$ are the same. The walk W is called a *cycle* if $v_1, v_2, \ldots, v_k$ are distinct, $v_0 = v_k$, and $k \geq 3$.

Example 1.6. Consider the following graph (a 'cube').



Consider the following five walks in this graph.

(i) (c),

(ii) (e, a, d, c, g, h, e),

(iii) (h, d, c, b, a, e, f),

(iv) $(b, c, b, f, b, c, d, a, b)$,

(v) (d, h, d).

Walks (i) and (iii) are paths and walk (ii) is a cycle. Walks (i), (iv), and (v) are closed walks, but not a cycle.

Exercise 1.9 Let G be the complete graph with node set $\{1, 2, 3, 4, 5\}$.

(a) How many walks in G start in node 1 and have length 10?

(b) How many paths does G have?

Exercise 1.10 Let a and b be two nodes of a graph G. Prove that:

$$\text{There is a walk in } G \text{ from } a \text{ to } b \quad \Longleftrightarrow \quad \text{there is a path in } G \text{ from } a \text{ to } b.$$

The implication '$\Leftarrow$' is clear, but the implication '$\Rightarrow$' is less obvious.

Definition 1.7. A graph is said to be *connected* if there is a path from every node to every other node.

Exercise 1.11 Let a, b, and c be nodes of a graph $G = (V, E)$. Check the following statements.

(i) There is a path from a to a.

(ii) If there is a path from a to b, then there is a path from b to a.

(iii) If there are paths from a to b and from b to c, then there is a path from a to c.

This exercise shows that the relation "there is a path from a to b" is an *equivalence relation*[2] on the nodes of a graph. It implies that the node set of the graph is partitioned into a number of subsets $V_1, V_2, \ldots, V_k$ (equivalence classes) such that for any two nodes a and b we have:

$$\text{There is a path from } a \text{ to } b \quad \Longleftrightarrow \quad a \text{ and } b \text{ are in the same subset } V_i.$$

In other words, the graph G decomposes into a number of *connected components* $G_1, \ldots, G_k$, where

$$\begin{aligned} G_i &= (V_i, E_i), \\ E_i &= \{\{u, v\} \in E : u, v \in V_i\}. \end{aligned}$$

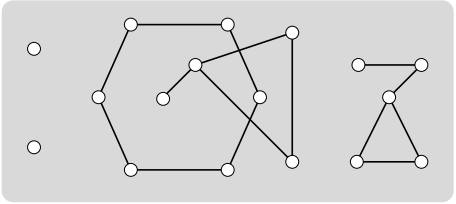
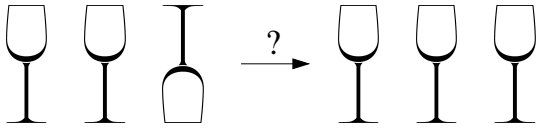


Figure 1.4: A graph with five connected components.

In Figure 1.4, a graph having five connected components is drawn.

Exercise 1.12 Three wine glasses are placed on a table. One of the glasses is upside-down. You are allowed to make the following move: choose two of the glasses and turn them. The goal is to get all three glasses to be the right-side up[3]. Draw the graph in which the nodes are the eight possible orientations of the three glasses (upside-down/correct) and two nodes form an edge if the two situations differ by a single move. How many connected components does this graph have? Can the goal be achieved?



Definition 1.8. Given nodes a and b of a graph, define their *distance* $d(a, b)$ to be the length of a shortest path from a to b. If a and b are in different connected components, define the distance to be $d(a, b) := +\infty$.

Exercise 1.13 In the graph of Figure 1.5, the nodes are Hollywood actors. Two actors are connected by an edge if they have co-starred in a Hollywood movie[4].

a) Determine the distance in this graph between Jennifer Aniston and Johnny Depp.

b) The *diameter* of a graph is the maximum distance between two nodes of the graph. What is the diameter of this graph?

1.2 Eulerian graphs

We start by defining a special type of closed walk.

[2]The notion of equivalence relation is one that is ubiquitous in mathematics and you will encounter it for example in the course 'Mathematical structures'.

[3]This was used in the popular TV-show *Mindf*ck* in episode 2 of 2017.

[4]The distance in the *full co-stardom graph* from a given actor to Kevin Bacon is called their *Bacon number*. Similarly, a mathematician's *Erdős number* is the distance to the legendary Paul Erdős in the graph where two mathematicians are connected by an edge if the have co-authored a paper. The Bacon-number of Paul Erdős is 5, while the Erdős number of Kevin Bacon is $+\infty$.

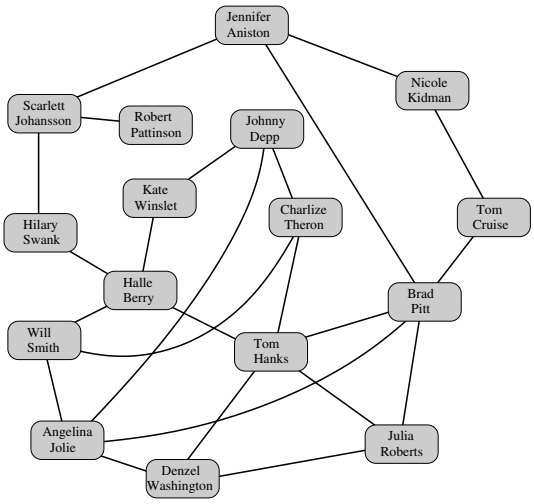


Figure 1.5: Part of the *co-stardom graph*

Definition 1.9. An *Eulerian tour* in a graph G is a closed walk in which every edge is traversed exactly once.

The name refers to a solution found by Leonhard Euler to a problem concerning a route through his town[5], the well-known problem of *the seven bridges of Königsberg*. See Exercise 18 and Figure 1.7).

Not every graph has an Eulerian tour. If it does, the graph is said to be *Eulerian*. If $(v_0, v_1, \ldots, v_m = v_0)$ is one Eulerian tour, then one immediately has $2m$ Eulerian tours. This is because you can simply start the tour in any position: $(v_i, v_{i+1}, \ldots, v_m = v_0, v_1, \ldots, v_{i-1}, v_i)$, or take the tour in the opposite direction. We will consider these $2m$ Eulerian tours to be essentially the same. In general, a graph can have multiple, essentially different, Eulerian tours.

Exercise 1.14 Show that K_4 is not Eulerian. Show that K_5 has two essentially different Eulerian tours.

One may weaken the requirements in Definition 1.9 to demand that the walk traverses every edge exactly once, but not require the walk to be closed. In that situation, the walk is called an *Eulerian trail*. So an Eulerian tour is precisely the same as a closed Eulerian trail. Een graaf that has an Eulerian trail, but no Eulerian tour is said to be *semi-Eulerian*.

Exercise 1.15 Draw a semi-Eulerian graph with 5 nodes and 5 edges.

The existence of an Eulerian tour implies that one can draw the given graph in one penstroke: without lifting your pen from the paper and without drawing any edges more than once.

[5]Königsberg is currently known as Kaliningrad.

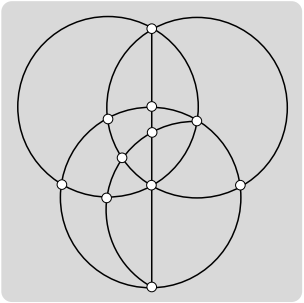


Figure 1.6: A semi-Eulerian graph

Exercise 1.16 Try to draw the graph in Figure 1.6 in a single stroke.

A graph that is Eulerian must clearly be connected (except for nodes of degree 0) since the tour traverses all nodes that are incident to at least one edge. Now, let G be a connected graph and suppose that G is Eulerian. If we consider following an Eulerian tour W in G, then it becomes clear that every node v of G must have *even* degree. Indeed, when traversing the $d(v)$ edges incident to v, half of the time we are moving towards v and half of the time we are moving away from v.

Have only even degree nodes is therefore a *necessary conditions* for a graph to be Eulerian: if the condition does not hold, then the graph cannot be Eulerian. Surprisingly, the condition is also *sufficient*: a connected graph with only even degree nodes is Eulerian.

Theorem 1.10. *A connected graph G is Eulerian if and only if all nodes have even degree and G is connected (except for nodes of degree 0).*

Proof. We have already argued necessity of the condition. We will now show that it is sufficient.

Let $G = (V, E)$ be a connected graph in which every node has *even* degree. Our task is to show that G has an Eulerian tour. Consider all possible walks in G that traverse every edge at most once. Such walks exist: consider any walk of length 0 starting at some node. Also, the length of such a walk can never exceed the total number of edges in the graph since no edge is traversed more than once. This implies that we can take among all these walks one of maximum length, say $W = (v_0, v_1, \ldots, v_k)$. We will prove that W is an Eulerian tour.

Claim 1: The walk W is closed. Indeed, suppose for contradiction that W were not closed (i.e. $v_0 \neq v_k$). Then W traverses an odd number of all edges incident to v_k, since each time W arrives at v_k through some edge, it immediately leaves through another edge, except for the last time it arrives at v_k. Since v_k has even degree by assumption, at least one edge incident to v_k is not traversed by W. But this implies that we can extent W by traversing this edge, contradicting our assumption that W had maximum length.

Claim 2: For every node v_i on the walk W it must be the case that W traverses all edges incident to v_i. Indeed, suppose for contradiction that some edge $e = \{v_i, x\}$ is not traversed. Since W is a closed walk (Claim 1), we can start our walk in v_i instead of v_0 and add one more step to the walk by traversing edge e: $(v_i, v_{i+1}, \ldots, v_k = v_0, v_1, \ldots, v_i, x)$. This again contradicts the fact we choose W to have maximum length.

Claim 3: All nodes of G are traversed by W. Indeed, let u be a node. Since G is connected by assumption, there is a path from v_0 to u, say $P = (v_0 = u_0, u_1, \ldots, u_t = u)$. Observe that Claim 2 implies that if a node is traversed by W, also its neighbours are traversed by W. Since $u_0 = v_0$ is traversed by W, also its neighbour u_1 is traversed by W. But then also u_2 is traversed by W, etc. We conclude that also $u = u_t$ is traversed by W.

From Claim 3 and Claim 2 it follows that W traverses all edges, and hence is an Eulerian tour. □

Exercise 1.17 For which values of n is K_n Eulerian?

In the more general situation of Euler trails we obtain the theorem below. This theorem follows easily from Theorem 1.10, or by modifying its proof.

Theorem 1.11. *A graph G has an Euler trail if and only if G has either zero or two nodes of odd degree.*

Exercise 1.18 In Figure 1.7 you can see a sketch of Königsberg. The two islands in the river are connected to each other and to the two shores by a total of seven bridges. The question is whether there exists a walk throug the city in which every bridge is traversed exactly once.

Apply Theorem 1.11 to this problem.

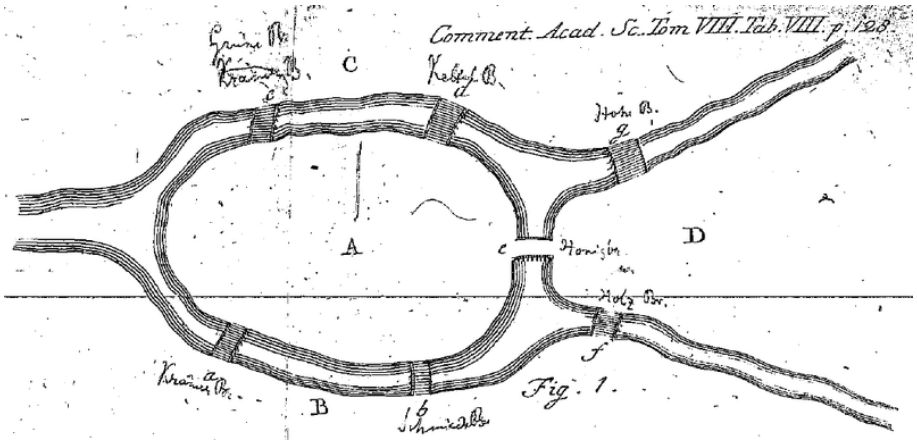


Figure 1.7: The seven bridges of Königsberg. From Euler's paper *Solutio problematis ad geometriam situs pertinentis* (The solution of a problem relating to the geometry of position), 1741.

Applications

As a first application of Eulerian tours we consider the *Chinese postman problem*[6]. A postman needs to traverse all streets in a certain area of town in order to deliver his mail.

[6]The preposition 'Chinese' refers to the nationality of the mathematician M.K. Kwan who studied the problem, not to that of the postmen

Naturally, he wants to take the most efficient route, and he has to start and end at the postoffice. Preferably, he would traverse each street exactly once before returning to his starting point. This is possible precisely when the street plan is an Eulerian graph.

If some nodes have odd degree, one may replace some edges by a pair of *parallel edges* to obtain a (multi)graph in which all nodes have even degree (why is this always possible?) These doubled edges will correspond to the streets traversed twice by the postman. In the Chinese postman problem we want to minimise the number of added parallel edges (or even the total length of the corresponding streets). There are efficient methods to solve this problem based on *matching theory*. We will not pursue this beautiful theory in this short introduction.

A modern variant of the Chinese postman problem in the Netherlands is that of applying salt or brine to the roads in winter. The winter service vehicles need to traverse all relevant roads but want to minimise the number of roads traversed twice. It will be clear that the same graph theoretic methods apply here.

A classical puzzle derived form the game of dominoes in this context is the following. The goal is to make a chain of dominoes in such a way that adjacent dominoes touch in equal numbers of pips. In Figure 1.8, you can see a small example. The puzzle is to make such a chain using *all* 28 dominoes.

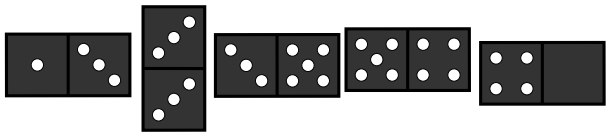


Figure 1.8: A valid chain of dominoes

Exercise 1.19 Solve the domino puzzle. Is it true that both ends of the chain always have the same number of pips? Suppose we had dominoes in which the number of pips ran from 0 to 7. Could we make a chain using all 36 dominoes?

1.3 Hamiltonian graphs

After introducing Eulerian tours as closed walks traversing all edges, it is natural to consider closed walks traversing all nodes.

Definition 1.12. A Hamiltonian cycle is a cycle traversing all nodes of the graph.

A graph having a Hamiltonian cycle is said to be *Hamiltonian*. The naming refers to a puzzle that was marketed in 1859 and was invented by the Irish mathematician and physisist W.R. Hamilton. The puzzle entailed a 'journey around the world', where the world was represented by a dodecahedron and one traveled along the edges from vertex to vertex. Each of the twenty vertices needed to be visited exactly once and one had to return to the starting point to complete a tour. Figure 1.9 shows that this is indeed possible. It should be clear from the picture that the proposed route can be seen as a Hamiltonian cycle in the dodecahedral graph: the graph whose nodes and edges correspond to the vertices and edges of the dodecahedron.

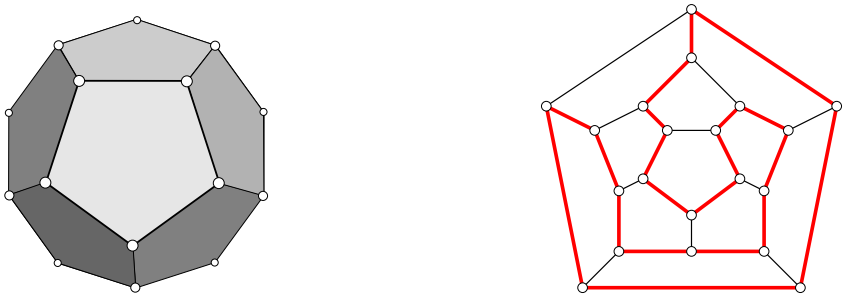


Figure 1.9: A dodecahedron and a Hamiltonian cycle in the corresponding graph.

Exercise 1.20 Find out which of the graphs in Figure 1.10 are Hamiltonian.

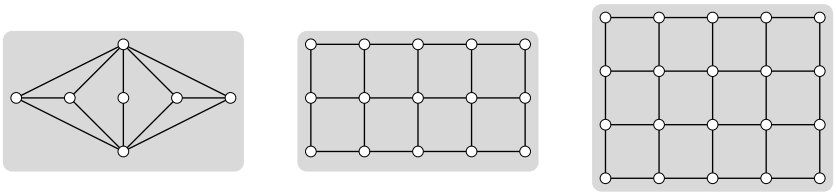


Figure 1.10

Exercise 1.21 Show that K_n is Hamiltonian for every $n \geq 3$.

Just as in the definition of Eulerian tour, one may remove the condition in Definition 1.12 that the walk needs to be closed. A *Hamiltonian path* is a path that traverses all nodes of the graph.

In analogy to the situation for Eulerian graphs, one might expect a simple criterion for determining whether a graph is Hamiltonian: a necessary and sufficient condition. Surprisingly, this is not the case. More precisely: no such criterion has been discovered at present. It is widely conjectured that there cannot be an efficient algorithm to determine whether a given graph is Hamiltonian.

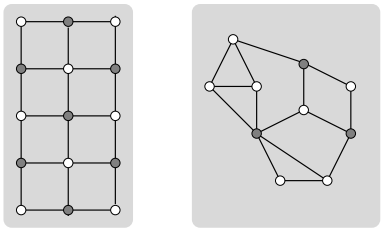
However, there do exist a number of necessary conditions for Hamiltonicity, as well as a number of sufficient conditions. An example of a necessary condition is given in the next theorem.

Theorem 1.13. *Let G be a graph on n nodes and suppose that G is Hamiltonian. Let k be an integer such that $1 \leq k \leq n$. If we remove k nodes from G, then the remaining graph has no more than k connected components.*

Before proving this theorem, we will give two examples.

Example 1.14. In the figure below, two graphs are shown. In the graph on the left, 7 nodes are selected (in grey). When these nodes are removed from the graph (including the edges incident to these nodes), we obtain 8 connected components (each consisting of a single node). In the graph on the right, you can remove the 3 grey nodes to obtain

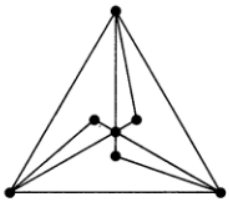
a graph with 4 connected components. Therefore, the theorem implies that these two graphs do not have a Hamiltonian cycle.



Proof. Let C be a Hamiltonian cycle in the graph G with n nodes. Let H be the graph that results when we remove from G all edges that are not traversed by C. So H looks like (is isomorphic to) C_n. Now we remove k nodes from H. Clearly, this results in at most k components (each shaped like a path). Since G has the same nodes as H but has additional edges, removing the same k nodes from G also yields at most k connected components. □

This theorem is useful for showing that a given graph is *not* Hamiltonian as it gives a necessary condition for being Hamiltonian. If the condition is not met, we can conlude that the graph is not Hamiltonian. The theorem cannot be used to show that a given graph *is* Hamiltonian. Indeed, even if the condition holds for every k and every set of k nodes that we remove, this does not imply that the given graph is Hamiltonian.

Exercise 1.22 Consider the graph below. Show that this graph is not Hamiltonian. Also show that when removing k nodes from this graph ($1 \leq k \leq 7$) the number of connected components of the resulting graph is no more than k.



Exercise 1.23 Show that the graph in Figure 1.11 is not Hamiltonian.

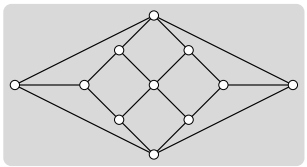


Figure 1.11

We will now give a sufficient condition for Hamiltonicity.

Theorem 1.15 (Dirac)**.** *Let G be a graph on n nodes ($n \geq 3$) in which all nodes have degree at least $n/2$ (rounded up). Then G is Hamiltonian.*

Proof. We give a proof by contradiction. Suppose that the theorem is not true. Then for some n there must be a non-Hamiltonian graph G on n nodes such that $d(v) \geq n/2$ for all $v \in V$. Add to G a maximum number of additional edges (between pairs of nodes that do not yet for an edge). We do this under the additional restriction that we do not create a Hamiltonian cycle. Call the resulting graph G'.

Since G' has no Hamiltonian cycle, we know that $G' \neq K_n$. Hence there exist two nodes v and w that do not form an edge in G'. If we were to add the edge $\{v, w\}$ to G', we would create a Hamiltonian cycle (since otherwise we would have already added this edge in the construction of G'). This implies that there is a Hamiltonian path in G' from v to w, say $(v = v_1, v_2, \ldots, v_n = w)$. We will use this path to show that G' does have a Hamiltonian cycle, contradicting the definition of G'.

First, define two subsets of the nodes of G' as follows. Let

$$\begin{aligned} I &:= \{i : \ v_i \text{ is a neighbour of } v_1\}, \\ J &:= \{i : \ v_{i-1} \text{ is a neighbour of } v_n\}. \end{aligned}$$

You may check that $I \subseteq \{2, \ldots, n-1\}$ and $J \subseteq \{3, \ldots, n\}$ since v_1 and v_n are not neighbours. Since v_1 and v_n both have degree at least $n/2$, we find that $|I| + |J| \geq n$. Since I and J are both subsets of $\{2, \ldots, n\}$, a set with fewer than n elements, we conclude that I and J must have at least one element in common, say $i \in I \cap J$.

But this leads to the following Hamiltonian cycle: $(v_1, \ldots, v_{i-1}, v_n, v_{n-1}, \ldots, v_i, v_1)$. See Figure 1.12.

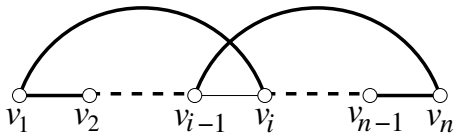


Figure 1.12

□

Exercise 1.24 Find an example of a graph that shows that the condition in Theorem 1.15 is *not a necessary condition* for being Hamiltonian.

An application

In Figure 1.13 a disc is drawn containing 4 tracks. The disk is partitioned into 16 equal sectors. Each part of a sector is covered by either an electrically conductive material or by an insulator.

Let's denote a conducting sector part by a 1 and a non-conducting part by a 0. This way, the 16 sectors can be represented by the binary representations of the numbers 0 to 15. Using sensors, the sectors (binary codes) can be read, thus determining the rotational position of the disk (up to a certain number of degrees).

The precision increases when we divide the disc into more sectors. In general, one uses 2^n sectors, representing the rotational position of the disc by binairy words of length n. This is an example of *analog-digital-conversion.*

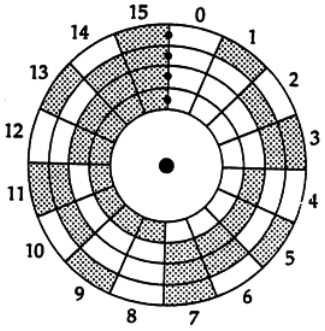


Figure 1.13: A disk with 16 sectors that are labelled by a binary code.

Apart from the precision determined by the length n of the codes, there can be errors in the decoding process, see Figure 1.13. The errors occur when a sensor is at the edge between two sectors. There, it may read either the bit corresponding to one sector, or the bit corresponding to the adjacent sector. In some situations, for instance in between sectors 7 and 8, this problem may occur for any of the four bits independently, leading to any of 2^4 possible outcomes (and hence no information on the rotational position). The best situation occurs when two adjacent sectors differ in only one bit, as is the case between sectors 2 (0010) and 3 (0011).

The obvious question is if there exists a binary encoding such that the codes for any two adjacent sectors differ in only one bit. For 16 sectors, binary words of length 4, this is indeed possible as demonstrated by Figure 1.14.

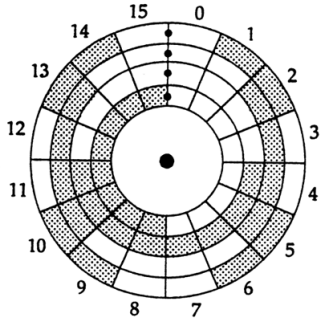


Figure 1.14: The 16 sectors are labelled according to a Gray code of length 4.

Such a labeling is called a *Gray code*, after its inventor Frank Gray at Bell labs. More precisely, a sequence consisting of all 2^n binary words of length n (each word occuring once) is called a *Gray code of length n* if any two consecutive words differ in only 1 bit, and also the first and last word differ in only 1 bit. The sequence

$$(000, 001, 011, 010, 110, 111, 101, 100)$$

is an example of a Gray code of length 3.

Exercise 1.25 Gray codes correspond to Hamiltonian cycles in certain graphs. Which graphs and what is the connection? Come up with a construction for Gray codes of length $n+1$ starting from a Gray code of lengt n.

1.4 Additional exercises

Exercise 1.26 A connected graph containing no cycle is called a *tree*.

a) Show that a tree with more than one node has at least two nodes of degree 1 (the *leaves* of the tree).

b) Draw some trees. What is the precise relation between the number of edges and the number of nodes in a tree?

Exercise 1.27 How many Hamiltonian cycles does K_n have? How many 'essentially different' Hamiltonian cycles? We call two Hamiltonian cycles 'essentially different' if they do not traverse the same set of edges.

Exercise 1.28 A graph is called *planar* if it can be draw in the plane without any two edges intersecting. In Figure 1.9, you can see that the dodecahedral graph is planar. Now consider the graphs corresponding to the four other Platonic solids (tetrahedron, cube, octahedron, icosaedron), see the figure below.

Draw the corresponding four graphs in the plane (without crossing edges) and show that they are all Hamiltonian.

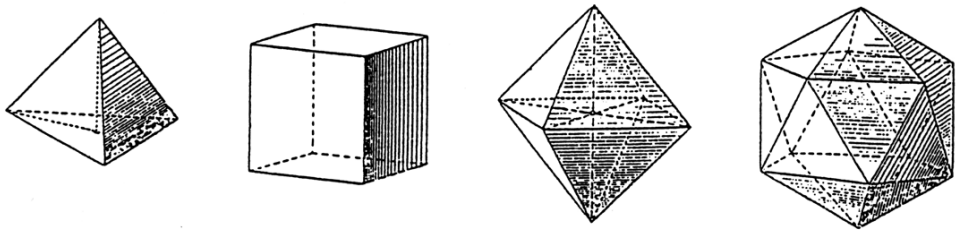


Figure 1.15: Four of the five Platonic solids.

Exercise 1.29 Only one of the Platonic solids has a corresponding graph that is Eulerian. Which one is it?

Literature

[1] A. Schrijver Dictaat *Grafen: Kleuren en Routeren*, zie:
`http://homepages.cwi.nl/~lex/files/graphs1_3.pdf`

[2] R.J. Wilson *Introduction to Graph Theory*, Longman, London, 1996.

2

Complex Numbers

Author: A.T. Hensbergen

Introduction

The German mathematican Leopold Kronecker once remarked: *"Die ganzen Zahlen hat der liebe Gott gemacht, alles andere ist Menschenwerk"*[1].
Throughout your life/math career you must have come across the number systems $\mathbb{N} \subseteq \mathbb{Z} \subseteq \mathbb{Q} \subseteq \mathbb{R} \subseteq$??.
What's the story behind this chain of sets, and does the real number system really mean the end?

2.1 The number systems

$\mathbb{N}$, the natural numbers

The natural numbers:

$$(0,)1,\ 2,\ 3,\ \ldots$$

were probably introduced to you by your parents – without giving them their official name. I put the zero between parentheses since different mathematicians (e.g. teachers ;-) think differently about including or not including zero. Anyway, I will let the natural numbers start at 0. At first these numbers were mainly meant for *counting*. To make counting more efficient *adding* numbers comes up, and the next basic operation, *multiplication*, is in fact iterated addition.
If one would build up the natural numbers in an axiomatic way addition and multiplication could be defined in a recursive way. I will not do so here, but let me mention one way to define the product of two natural numbers: First $a \cdot 0 \underset{\text{def}}{=} 0$ and further $a \cdot (n+1) \underset{\text{def}}{=} a \cdot n + a$.
So for instance $5 \times 4 = 5 \times (3+1) = 5 \times 3 + 5 = 5 \times (2+1) + 5 = \ldots = 5+5+5+5$.

[1] Weber, H. (1893), "Leopold Kronecke", Mathematische Annalen, Springer Berlin, Heidelberg, 43: 1 – 25.

The following properties may then be deduced:

$$
\begin{aligned}
&a+b=b+a \quad \text{(symmetry)}\\
&a+(b+c)=(a+b)+c \quad \text{(associativity)}\\
&a+0=a\\
&a\cdot b=b\cdot a\\
&a\cdot(b\cdot c)=(a\cdot b)\cdot c\\
&a\cdot(b+c)=a\cdot b+a\cdot c \quad \text{(distributivity)}\\
&a\cdot 1=a
\end{aligned}
$$

et cetera

The inverse operations, *subtraction* and *division* are not always possible: $7-5=2$, but $4-9$ is not possible in the realm of the natural numbers. In other words: the equation $9+x=4$ has *no* solution in $\mathbb{N}$.
Likewise for the inverse operation of multiplication. E.g. consider $20/5$ and $13/7$.

The first extension: $\mathbb{Z}$, the integers

To make sure that every equation $a+x=b$ *does* have a solution one can introduce new symbols -1, -2, -3, ... with the 'convention' that, for instance, -3 is a solution for the equation $8+x=5$.
Formally for each equation $a+x=b$ with $b<a$ the solution/symbol/number $-(a-b)$ is introduced. As a result we get the set:

$$\mathbb{Z}=\{\ldots,-3,\,-2,\,-1,\,0,\,1,\,2,\,\ldots\}$$

Addition and multiplication have to be defined for any two numbers in $\mathbb{Z}$ in such a way that the properties of section **1.1** remain valid. By now you should know how this is done.

The second extension: $\mathbb{Q}$, the rational numbers

To make sure the equation $a\cdot x=b$ does have a solution for every $a(\neq 0)$ and $b\in\mathbb{Z}$, one can introduce new symbols $\frac{p}{q}$, $p\in\mathbb{Z}, q=1,2,3,\ldots$, with the 'convention' that $\frac{p}{q}$ is the solution for the equation $q\cdot x=p$. The set of all these 'numbers' is called the set of *rational numbers*:

$$\mathbb{Q}=\left\{\frac{p}{q}\;\middle|\;p\in\mathbb{Z},\ q=1,2,3,\ldots\right\}$$

This is just one way to do it; one might also take $q\in\mathbb{Z}, q\neq 0$ and $p\in\mathbb{Z}$.
To see this, just note that, for instance,the equations $(-2)x=5$ and $2x=-5$ are equivalent.

Some more terminology: for the rational number $\frac{p}{q}$ the number p is called the *numerator*, and the number q is called the *denominator*.

From the observation that if $x=\frac{p}{q}$, $y=\frac{r}{s}$, i.e. $q\cdot x=p$ and $s\cdot y=r$, it follows that $(qs)\cdot(xy)=(q\cdot x)\cdot(s\cdot y)=p\cdot r=pr$, so that $xy=\frac{pr}{qs}$. Hence it makes sense to define the product of $\frac{p}{q}$ and $\frac{r}{s}$ as $\frac{pr}{qs}$.

In a slightly more involved way the following 'consistent' definition of addition may be given support: $\frac{p}{q} + \frac{r}{s} \underset{\text{def}}{=} \frac{ps+qr}{qs}$.
With *consistent* we mean that for the addition and multiplication of rational numbers the properties of section **1.1** still hold.

Note that equality of numbers in $\mathbb{Q}$ has a minor complication:
the rational numbers $\frac{1}{3}$ and $\frac{2}{6}$ should be considered as *the same* rational number,
since they both satisfy the equation $6x = 2$.
To take care of this 'nuisance' we may *define* $\frac{p}{q} = \frac{r}{s} \underset{\text{def}}{\iff} ps = qr$.

Lastly, we may identify the integer k with the rational number $\frac{k}{1}$, and then it can be said that the set of integers is contained in the set of rational numbers.

For someone who has never been exposed to rational numbers the last two definitions may look horrible, but once you grasped the idea of fractions your fear may disappear. If you've never seen complex numbers (don't be scared by the name) I hope the same will happen to you: first you think 'what's going on here ...', but soon you will get the hack of them. And maybe even start to love them.

One step beyond: $\mathbb{R}$, the real numbers

Why go further? Well, why not?! The first two extensions from $\mathbb{N}$ to $\mathbb{Q}$ had to do with the algebraic operators $+$ and $\times$. They sort of 'closed' these operators. There is another important relation/operator giving structure to these number systems, namely the inequality operator.
The extension from $\mathbb{Q}$ to $\mathbb{R}$ has to do with this $\leq$-operator: For each two numbers a and b in either $\mathbb{N}$, $\mathbb{Z}$ or $\mathbb{Q}$, either $a \leq b$ or $b \leq a$, and both hold if and only if $a = b$.
Together with the property if $a \leq b$ and $b \leq c$ then $a \leq c$, this implies that all three number systems can be visualized on a *line*. The rational numbers are 'dense' on the line in the sense that for each two rational numbers $a < b$ there is always a rational number c for which $a < c < b$. However, the rational numbers still leave 'holes' on this line. For instance, the cubes of the rational numbers, i.e. a^3, somewhere between 2 and 3 pass the value 10, but it can be shown that there is no rational number b for which b^3 equals 10. In other words: $\sqrt[3]{10}$ is *irrational*. So there's a 'hole' somewhere between 2 and 3, and there are in fact infinitely many holes. Intuitively, $\mathbb{R}$ is the result of 'filling up' all these holes. In the course "Mathematical Structures" you will learn a more rigorous approach.

Enough is enough?

Now what else could we wish for? Well, there is something a little unsatisfactory with quadratic and higher order equations. Some do have solutions in $\mathbb{R}$, some don't. The simplest equation where $\mathbb{R}$ 'falls short' is the equation $x^2 + 1 = 0$. In the eighteenth century mathematicians came up with the number system $\mathbb{C}$, an even larger system than the real number system, in which all n-th order equations do have solutions. The terms 'complex number' and 'imaginary number' do convey the scepticism people had with regards to these 'numbers'. Battles were fought over this (luckily without too much bloodshed). Nowadays complex numbers are an important tool for mathematicians and engineers.

To quote the writer (ex math-student) John Derbyshire: *"I tell you, with complex numbers you can do anything."*[2] The German philosopher/mathematician Gottfried Leibniz went even further and in 1702 wrote: *"Imaginary* (read: complex) *numbers are a fine and wonderful refuge of the divine spirit almost an amphibian between being and non-being."*[3]

2.2 The complex number system

The definition(s)

The amazing thing is that by introducing only one new symbol ('number') i with the defining property $i^2 = -1$, a new number system may be built in which as far as addition and multiplication are concerned we can still do algebra as in all the above number systems, and in which *all* algebraic equations *do* have solutions.

Definition 2.1. The *complex numbers* are defined as the set:

$$\mathbb{C} = \{\, a + bi \mid a, b \in \mathbb{R} \,\}$$

For $z = a + bi$ the number a is called the *real part* of z, and b is called the *imaginary part* of z. In shorthand: $a = \operatorname{Re} z, b = \operatorname{Im} z$.

Note that $\operatorname{Im} z$ is a *real* number.

Definition 2.2. For two complex numbers the *sum* and the *product* are defined as follows:

$$(a + bi) + (c + di) = (a + c) + (b + d)i$$

and

$$(a + bi) \cdot (c + di) = (ac - bd) + (ad + bc)i$$

The logic behind this:
$$\begin{aligned}(a + bi) \cdot (c + di) &= ac + adi + bic + bidi \\ &= ac + bi \cdot di + adi + bci \\ &= ac + bdi^2 + adi + bci \quad (\text{and } i^2 = -1\,) \\ &= (ac - bd) + (ad + bc)i.\end{aligned}$$

As with real numbers $(a + bi) \cdot (c + di)$ is often written as $(a + bi)(c + di)$.

Definition 2.3. Thirdly, two complex numbers are *equal*, that is $a + bi = c + di$ if and only if $a = c$ and $b = d$. (Easier than for rational numbers!)

It can be easily checked that the basic properties of subsection **1.1** (commutativity, symmetry, ...) still hold for complex numbers.
Subtraction of complex numbers is immediate:

$$(a + bi) - (c + di) = (a - c) + (b - d)i.$$

[2] from "Prime Obsession: Bernhard Riemann and the Greatest Unsolved Problems in Mathematics".
[3] F. Klein, Elementary "Mathematics From an Advanced Standpoint (1932), Vol. 1".

For division of a complex number $a + bi$ by a complex number $c + di$ not equal to zero, we use the same trick that helps to eliminate a square root from the denominator of a fraction. Compare:

$$\frac{4+\sqrt{3}}{1+\sqrt{3}} = \frac{4+\sqrt{3}}{1+\sqrt{3}} \cdot \frac{1-\sqrt{3}}{1-\sqrt{3}} = \frac{4-4\sqrt{3}+\sqrt{3}-3}{1-(\sqrt{3})^2} = \frac{1-\sqrt{3}}{-2} = \tfrac{1}{2}\sqrt{3} - \tfrac{1}{2}$$

with

$$\frac{a+bi}{c+di} \cdot \frac{c-di}{c-di} = \frac{(a+bi)(c-di)}{(c+di)(c-di)} = \frac{ac+bd+(bc-ad)i}{c^2+d^2} = \frac{ac+bd}{c^2+d^2} + \frac{bc-ad}{c^2+d^2}\,i$$

Note that $c + di \neq 0$ is equivalent to $c^2 + d^2 \neq 0$, so we don't get zero in the denominators.

Lastly, by *identifying* the real number a and the complex number $a + 0i$ we can consider $\mathbb{R}$ as a subset of $\mathbb{C}$. Which finally explains the title of the introduction.

Exercise 2.1 Compute (i.e. express in the form $a + bi$ $(a, b \in \mathbb{R})$:

a) $(1 + 2i)(3 + 1)(1 - 2i)$ and $(1 + 2i)(1 - 2i)(3 + 1)$;

b) $(1 + i)^{15}$;

c) $(1 - i)^{15}$;

Exercise 2.2 Compute

a) $\dfrac{1}{1+i}$ and $\dfrac{1}{(1+i)^2}$;

b) $\dfrac{2-i}{1+2i}$ and $\dfrac{i-2}{1+2i}$;

c) $\dfrac{10}{4+3i} + \dfrac{5}{3-4i}$.

Exercise 2.3 Check that the commutative law indeed holds for the addition and multiplication of two complex numbers, i.e. $z + w = w + z$ and $z \cdot w = w \cdot z$.

Exercise 2.4 For $z = -\frac{1}{2} + \frac{1}{2}\sqrt{3}$ find z^2, z^3, z^4, z^5 and z^{1000} and z^{1001}. Also find $z + z^2$, $z + z^2 + z^3$, $z + z^2 + z^3 + z^4$, and $z + z^2 + \ldots + z^{1001}$.

The complex plane

So far, the notions and operations for rational and real numbers could be extended to the complex numbers. For the ordering $(\leq)$ this cannot be done in a way that preserves all properties that hold for real (or rational) numbers. Two of these properties are: if for three real numbers we have $x \leq y$ and $a \geq 0$, then $-y \leq -x$ and $ax \leq ay$.

The assumption that the ordering on $\mathbb{R}$ can be extended to $\mathbb{C}$ leads to a contradiction in the following way: Either we would have $i \geq 0$ or $i \leq 0$.
Well, *if* $i \geq 0$, then (since the rules of $\leq$ still apply) also $i \cdot i \geq i \cdot 0$, but

then -1 would be ≥ 0.
On the other hand *if* $i \leq 0$, then $(-i) \geq 0$, so then also $(-i) \cdot (-i) \geq (-i) \cdot 0$, and again we arrive at the absurdity $-1 \geq 0$.

Since both the assumption $i \geq 0$ and the assumption $i \leq 0$ lead to a contradiction, we conclude that comparing i and 0 does not make sense. Another way to put it: the complex numbers do not fit in the real line anymore. But hey, informally it was already mentioned that the real numbers do fill up all of the line. Is that a problem? Well, no, for you perception it may be so, for a while, but basically it's just something new. Something interesting!
Since the image of a line sometimes does help to visualize matters about real numbers we may ask ourselves: is there a *suitable* geometric representation for the complex numbers? In fact, there is, and it is called the *complex plane*. We may identify the complex number $a + bi$ with the point (a,b) in the plane. The x-axis contains the real numbers $a = a + 0i$, and for this reason is called the *real axis*, and the y-axis, containing the purely imaginary numbers $0 + bi$, is called the *imaginary axis*.
The addition of two complex numbers $(a + bi)$ and $(c + di)$ then nicely corresponds to the addition of the two *vectors* starting at (0,0) and ending at (a, b) and (c, d). See Figure 2.1.
What is the geometric interpretation of the product of $(a + bi)$ and $(c + di)$?
To simplify matters we need a few more notions, which will be defined in the next section.

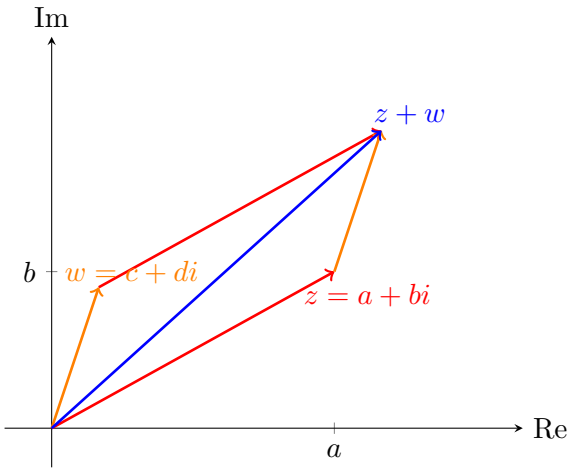


Figure 2.1: Addition in the complex plane.

Exercise 2.5 Find $w = iz$ for $z = 1 + 2i$, $z = -3 + i$, $z = 5 - 2i$, and sketch these (six) points in the complex plane. Which transformation is (or: seems to be) going on?

Exercise 2.6 For $z = 1 + i$, find z^2, z^3, $z^4, \ldots, z^8$. Again: sketch these points in the complex plane.

(Complex) conjugate, modulus and argument

Definition 2.4. The *(complex) conjugate* of $z = a + bi$ is defined as $\bar{z} = a - bi$.

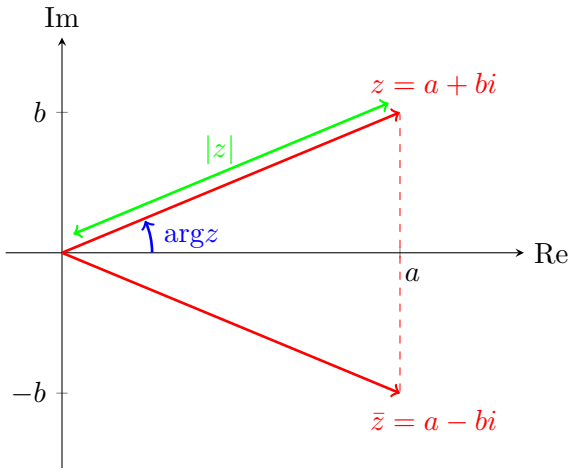


Figure 2.2: Conjugate, Modulus and Argument.

Above the complex conjugate was already used to define the quotient of two complex numbers.

Definition 2.5. The *modulus* of $z = a + bi$ is defined as $|z| = \sqrt{a^2 + b^2}$ so this definition is actually nothing more than applying Pythagorean Theorem to the vector(a, b).

In the complex plane: $|a + bi|$ is the distance from (0,0) to the point (a, b).

For real numbers $a = a + 0i$, the modulus is just the absolute value (which can also be interpreted as the distance from the point a on the line to the point 0). That is, there is no 'clash' of notations.

Definition 2.6. An *argument* of $z = a + bi$, denoted by $\arg z$, is defined as the angle, from the positive x-axis to the line segment connecting $(0, 0)$ to (a, b) (see Figure 2.3). This angle is defined up to a multiple of 2π. The picture tells the story. To make it unique, an interval of length 2π may be chosen in advance, usually either the interval $[-\pi, \pi)$ or the interval $[0, 2\pi)$, and the value of the argument that lies in this interval is called its *principal value*.

Properties of conjugate and modulus

- $\overline{z + w} = \overline{z} + \overline{w}$ and $\overline{z - w} = \overline{z} - \overline{w}$;
- $\overline{zw} = \overline{z} \cdot \overline{w}$ and $\overline{\frac{z}{w}} = \frac{\overline{z}}{\overline{w}}$;
- For $z = a + bi$: $z + \overline{z} = 2a = 2\,\mathrm{Re}\, z \in \mathbb{R}$, and $z\overline{z} = a^2 + b^2 \in \mathbb{R}$;
- $|z| = \sqrt{z\overline{z}}$;
- $|zw| = |z||w|$ and $\left|\frac{z}{w}\right| = \frac{|z|}{|w|}$.

These properties are all easily checked: just take $z = a + bi$ and $w = c + di$, and see what happens ...

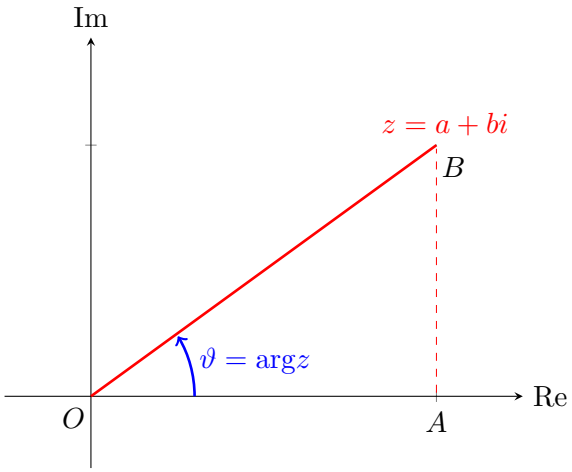


Figure 2.3: Computing $\arg z$

To find the argument ϑ of the complex number $z = a + bi \neq 0$, first note that for $0 \leq \vartheta < \frac{\pi}{2}$

$$\cos\vartheta = \frac{|OA|}{|OB|} = \frac{\operatorname{Re} z}{|z|} \text{ and}$$
$$\sin\vartheta = \frac{|AB|}{|OB|} = \frac{\operatorname{Im} z}{|z|} \text{ (see Figure 2.3),}$$

from which it follows that

$$z = \operatorname{Re} z + i \operatorname{Im} z = |z|\cos\vartheta + i|z|\sin\vartheta = |z|(\cos\vartheta + i\sin\vartheta)$$

We should take care of the signs if z does not lie in the first quadrant.
Check for yourself that the formulas $\cos\vartheta = \dfrac{\operatorname{Re} z}{|z|}$ and $\sin\vartheta = \dfrac{\operatorname{Im} z}{|z|}$ are true everywhere.

Conversely, it is quickly seen that if $z = r(\cos\varphi + i\sin\varphi)$, with $r \in \mathbb{R}_{\geq 0}$, then $r = |z|$ and $\varphi = \arg z \pmod{2\pi}$. If $z = 0$, when there is not really an angle, any value may be taken as $\arg z$.
The next calculation sheds geometric light onto the multiplication of two complex numbers:

$$r_1(\cos\varphi + i\sin\varphi) \cdot r_2(\cos\psi + i\sin\psi) =$$
$$r_1 r_2(\cos\varphi\cos\psi + i\cos\varphi\sin\psi + i\sin\varphi\cos\psi - \sin\varphi\sin\psi) =$$
$$\ldots = r_1 r_2(\cos(\varphi + \psi) + i\sin(\varphi + \psi))$$

From this we may again deduce that $|zw| = |z| \cdot |w|$, but more importantly also the following

Properties of the argument

Up to multiples of 2π the following identities hold (and make life a bit easier, sometimes):

- $\arg(zw) = \arg z + \arg w$;

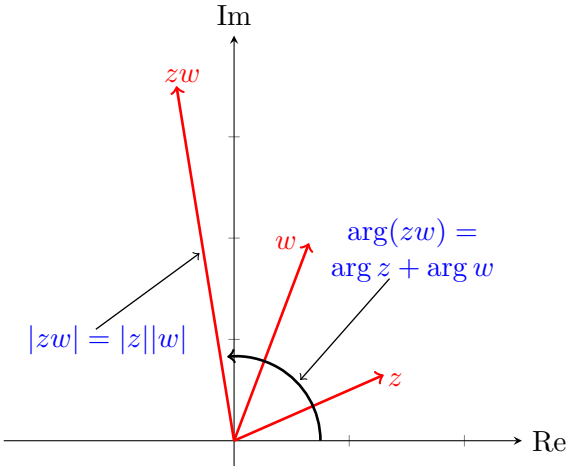


Figure 2.4: Product in the complex plane

- $\arg\left(\frac{z}{w}\right) = \arg z - \arg w$;
- $\arg(z^n) = n\,(\arg z)$.

Check for yourself how the second and the third properties follow from the first!

Geometric interpretation of the product

Above it was shown that each complex number can be written in the form $z = r(\cos\vartheta + i\sin\vartheta)$, namely by taking $r = |z|$ and $\vartheta = \arg z$. This is sometimes called the *polar form*. Multiplying two complex numbers in polar form boils down to *multiplying* their moduli and *adding* their arguments.

Exercise 2.7 For an 'arbitrary' number $z = a + bi$, find $w_1 = \frac{1}{2}(z + \overline{z})$, $w_2 = \frac{1}{2}(z - \overline{z})$ and $w_3 = z + i\overline{z}$, and sketch the corresponding points in the complex plane.

Exercise 2.8 Find the polar form of the complex numbers $-5i$, $-\sqrt{6}+\sqrt{2}i$ and $-3-4i$. (For one of them you may use arcsin (or arccos or arctan) to specify the argument φ.)

Exercise 2.9 By using the fact that $\frac{1}{12} = \frac{1}{3} - \frac{1}{4}$, find $\cos(\frac{1}{12}\pi)$ and $\sin(\frac{1}{12}\pi)$.
Hint: Compute the quotient of two suitable complex numbers in two ways.

Exercise 2.10 Find $w_i = \dfrac{10}{\overline{z_i}}$ for $z_1 = 1+2i$, $z_2 = -3+i$, $z_3 = -4-2i$, and $z_4 = 7-4i$.
And again: sketch. How would you describe the mapping $z \mapsto \dfrac{10}{\overline{z}}$?

(Complex) exponential function

Motivation

As shown in the previous section, the function $g(t) = \cos t + i\sin t$ has the property $g(s)g(t) = g(s+t)$. And it also satisfies $g(0) = 1$. These are the same as for the (real)

functions $f(t) = e^{at}$.
And there's more: it can be shown (and it will be during the course differential equations) that $f(t) = e^{at}$ is the *unique* function with the properties $f(0) = 1$ and $f'(t) = af(t)$.
Now *if* we define differentiation of a function $g : \mathbb{R} \to \mathbb{C}$ in the most obvious way (i.e. just differentiate the real part and the imaginary part), then it is easily seen that the function $g(t) = \cos t + i \sin t$ satisfies $g(0) = \cos 0 + i \sin 0 = 1$ and also $g'(t) = -\sin t + i \cos t \overset{!!}{=} ig(t)$.
After these observations the next definition hopefully does not come as a big shock.

Definition 2.7. First, for real t, e^{it} is *defined* as $e^{it} = \cos t + i \sin t$.
Second, for general complex numbers: $e^z = e^{a+bi} \underset{\text{def}}{=} e^a e^{bi} = e^a(\cos b + i \sin b)$.

Exercise 2.11 Show that for any two complex numbers the following identity holds: $e^z e^w = e^{z+w}$.

Example 2.8. Taking $z = \pi i$ in the definition we get $e^{\pi i} = -1$.
This formula, connecting three (or four, if you want) important numbers is called *Euler's Formula*, and there are quite a few mathematicians that consider it to be the most beautiful formula of all.

Computations using the polar form

First of all note that we can now use the 'abbreviation' $re^{i\vartheta}$ for the complex number $r(\cos \vartheta + i \sin \vartheta)$. A few examples show how certain computations can be simplified.

Example 2.9. Find $\left(\dfrac{i+\sqrt{3}}{2}\right)^{20}$ without doing so many multiplications as follows: First we write $z = \dfrac{i+\sqrt{3}}{2}$ in polar form:
$|z| = 1$, $\cos(\arg z) = \dfrac{\operatorname{Re} z}{|z|} = \frac{1}{2}\sqrt{3}$ and $0 \le \arg z \le \frac{1}{2}\pi$
(since z lies in the first quadrant), so $\arg z = \frac{1}{6}\pi$.
Then $z = 1 \cdot e^{\frac{1}{6}\pi i}$, and it follows that

$$z^{20} = 1^{20} e^{20 \cdot \frac{1}{6}\pi i} = e^{\frac{20}{6}\pi i} = e^{\frac{10}{3}\pi i} = \cos \tfrac{10}{3}\pi + i \sin \tfrac{10}{3}\pi = -\tfrac{1}{2} - \tfrac{1}{2}i\sqrt{3}$$

Example 2.10. Find all solutions of the equation $z^3 = 8i$.
The 'trick' here is to write both terms in polar form and note that $r_1 e^{i\vartheta} = r_2 e^{i\varphi}$, with $r_i \ge 0$, is equivalent to $r_1 = r_2$ *and* $\vartheta - \varphi = 2k\pi$, for some integer k.
Following the hint: first note that $|8i| = 8$ and $(\arg 8i) = \frac{1}{2}\pi$, so that $8i = 8e^{\frac{1}{2}\pi i}$.
So we have to find $z = re^{i\vartheta}$ that satisfies $(re^{i\vartheta})^3 = r^3 e^{3i\vartheta} = 8e^{\frac{1}{2}\pi i}$.
We deduce that $r^3 = 8$ and $3\vartheta = \frac{1}{2}\pi + 2k\pi$, leading to $r = 2$, $\vartheta = \frac{1}{6}\pi + \frac{2}{3}k\pi$.
This gives the solutions $z_k = 2\left(\cos\left(\frac{1}{6}\pi + \frac{2}{3}k\pi\right) + i \sin\left(\frac{1}{6}\pi + \frac{2}{3}k\pi\right)\right)$, $k \in \mathbb{Z}$.
Now it may look like there are infinitely many solutions, but due to the periodicity of the sine and the cosine, it follows that $z_{k+3} = z_k$, so there are actually only three *different* solutions. These three different solutions are found by taking, for instance, the values

$k = 0, 1, 2$, which gives $z_0 = 2\big(\cos\big(\frac{1}{6}\pi\big) + i\sin\big(\frac{1}{6}\pi\big)\big) = \sqrt{3} + i$,
$z_1 = 2\big(\cos\big(\frac{1}{6}\pi + \frac{2}{3}\pi\big) + i\sin\big(\frac{1}{6}\pi + \frac{2}{3}\pi\big)\big) = -1 + i\sqrt{3}$ and
$z_2 = 2\big(\cos\big(\frac{1}{6}\pi + \frac{4}{3}\pi\big) + i\sin\big(\frac{1}{6}\pi + \frac{4}{3}\pi\big)\big) = -1 - i\sqrt{3}$.

Exercise 2.12 Find $\dfrac{(1-i)^{30}}{(2i\sqrt{3}-2)^7}$, by using the polar form.

Exercise 2.13 Find all (five) solutions of the equation $z^5 = -4 - 4i$ and write them in the form $a + bi$. (The answer may contain sines and cosines of 'difficult' angles.)
Sketch all solutions in the complex plane.

Concluding Exercises

Exercise 2.14 An easy starter

a. Check that $\big((3+i)(4-i)\big)(3-i) = (3+i)\big((4-i)(3-i)\big)$.
What would have been the quickest way to this (relatively simple) answer?

b. Compute (i.e. write in the form $a + bi$, $a, b \in \mathbb{R}$):
$(1-2i) + (1-2i)^2 + (1-2i)^4$ and $\dfrac{11+13i}{2+5i}$.

c. Show that $\overline{(a+bi)(c+di)} = (a-bi)(c-di)$. (That is: $\overline{zw} = \overline{z}\,\overline{w}$.)

d. Use **c.** and the fact that $|z| = \sqrt{z\overline{z}}$ to prove that $|zw| = |z||w|$. (Make sure that you only consider the square roots of non-negative (real) numbers!

Exercise 2.15 Algebraic equations (= n-th order equations)
It can be *proved* that any algebraic equation

$$c_n z^n + c_{n-1} z^{n-1} + ... + c_1 z + c_0, c_i \in \mathbb{C}, i = 0, 1, \ldots, n, c_n \neq 0$$

has at least one solution in $\mathbb{C}$.
This exercise sheds some light onto this theorem about complex numbers, also known as the *fundamental theorem of algebra*.

a. Find all solutions of the equation $z^2 = 5 - 12i$.
(Hint: write $z = a + bi$, $a, b \in \mathbb{R}$.)

b. Using the fact that $(1+i)^4 = -4$, find all (four) solutions of the equation $z^4 = -4$.

c. Find all solutions of the equation $z^4 + 2z^2 + 2 = 0$.
(It *is* possible to write the solutions in the form $a + bi$. a and b are a bit awkward, but you are urgently requested to suppress the urge to use your pocket calculator.)

d. Generalization of the *abc*-formula.
Consider the quadratic equation $az^2 + bz + c = 0$ with *real* coefficients a, b, c.
Suppose the discriminant $D = b^2 - 4ac$ is negative.

Show that $z_{1,2} = \dfrac{-b \pm i\sqrt{-D}}{2a}$ are two solutions of the equation.
Note that, since $D < 0$, the square root $\sqrt{-D}$ is 'okay'.
As a 'concrete' example: take the equation $5z^2 + 6z + 5 = 0$.

e. The real equation $x^2 + 6x - 9 = 0$ can be solved by 'completing the square':
$x^2 + 6x - 9 = (x + 3)^2 - 18$, so:
$x^2 + 6x - 9 = 0 \Leftrightarrow (x + 3)^2 - 18 = 0 \Leftrightarrow .... \Leftrightarrow x = -3 \pm \sqrt{18}$
Solve the equation $z^2 + (2 + 2i)z + 2 = 0$, along a similar path.
That is, by rewriting in into the form $(z - u)^2 = v$, with $u, v \in \mathbb{C}$.

Would the *abc*-formula have worked here?

And: What is the connection between the 'methods' of **c.** and **d.**?

f. Show that for algebraic equations with real coefficients the non-real solutions come in complex conjugate pairs, that is:
if $z = a + bi$ is a solution of the equation $c_n z^n + c_{n-1} z^{n-1} + \ldots + c_1 z + c_0 = 0$, with all coefficients $c_i \in \mathbb{R}$, then so is $\overline{z}$.

Check whether this indeed happened in questions **b.**, **c.** and **d.**

Exercise 2.16 Exponential equations

a. Find *all* solutions of the equation $e^z = 1 - i$.
(Hint: put $z = a + bi$, $a, b \in \mathbb{R}$, and write both sides in polar form.)

b. As a generalization: show that the equation $e^z = c$ for any non-zero $c \in \mathbb{C}$ has infinitely many solutions. (And why should we exclude $c = 0$?)

c. For *real* numbers α, show that

$$\cos\alpha = \frac{e^{i\alpha} + e^{-i\alpha}}{2} \text{ and } \sin\alpha = \frac{e^{i\alpha} - e^{-i\alpha}}{2i}.$$

Using this, prove:

$(\cos\alpha)^3 = \frac{1}{4}\cos(3\alpha) + \frac{3}{4}\cos\alpha$ and $(\sin\alpha)^3 = -\frac{1}{4}\sin(3\alpha) + \frac{3}{4}\sin\alpha$.

d. With part **b.** in mind, define

$$\cos z = \frac{e^{iz} + e^{-iz}}{2} \text{ and } \sin z = \frac{e^{iz} - e^{-iz}}{2i}, \text{ for } \textit{all} \text{ complex numbers } z.$$

Find all $z \in \mathbb{C}$ for which $\cos z = 2$.
(Hint: rewrite the equation as a quadratic equation in the variable $w = e^{iz}$.)

Exercise 2.17 Complex numbers and plane geometry.
The modulus and the argument have the geometric interpretations length and angle. Furthermore, note that $|z - w|$ can be likewise interpreted as the *distance* between the points z and w in the complex plane, and that $\arg(z/w)$ can be interpreted as the angle between two line segments. (Make a picture yourself!) In this exercise you can either work analytically (by putting $z = x + iy$) or geometrically.

a. Describe the set of all complex numbers z for which $|z - 3 + 2i| = 4$ and sketch it in the complex plane.

b. Likewise for the set $\{z \,|\, |z - 2| = |z - 4|\}$.

c. The set $\{z \,|\, \operatorname{Im} z = |z - 4i|\}$ may be easier to describe analytically first. Unless you already know a lot about quadratic curves. Which also applies to the next question.

d. How to describe/sketch the set $\{z \,|\, |z + 2i| + |z - 2i| = 6\}$?

e. Which angle is represented by the argument of the quotient $\dfrac{z - z_0}{z - z_1}$?
(Picture?!)

f. Sketch the set of all complex numbers z for which $\arg\left(\dfrac{z - 2i}{z - 6}\right) = \frac{1}{2}\pi$.

g. To conclude quite a hard geometric 'puzzle' which can be solved rather easily using complex numbers: Suppose P and Q are two points in the plane, and let k be a positive number not equal to 1.
Show that the points X for which $|XP| = k \cdot |XQ|$ lie on a *circle*.
Consider the 'concrete' example with the points $P(5, 5)$, $Q(2, 2)$, and the value $k = 2$.

3

Optimization in networks

Author: C. Roos,
translated by L.J.J. van Iersel

3.1 Introduction

Optimization is the part of mathematics that concerns the development and analysis of algorithms for solving problems where some function $f(x)$ needs to be optimized (minimized or maximized) over all elements x in some set X. In the case of minimization, such a problem can be written as

$$\min\{f(x) : x \in X\}.$$

The function f is called the *objective function* and X the *domain* or *feasible region*. An interesting special case is when f is a linear function while $X \subset \mathbb{R}^n$ and can be described by linear (in)equalities. Such a problem is called a *linear programming* problem and can be solved efficiently.

Many practical problems, in many application areas, can be modelled as optimization problems (which are often not just linear programming problems). For example, think about optimizing production processes in factories, finding optimal designs (e.g. of airplanes), finding a most-likely explanation of certain biological data, optimizing traffic, hospitals, airports, energy networks, etc.

During the second world war, optimization was applied for the first time at a large scale, to model and solve logistical and operational problems. As a scientific discipline, the field arose and flourished when the first computers were built. Since then, computers have become faster and faster. However, the size of data sets have exploded even more (eg. DNA data, the internet, the internet of things) and the problems that need to be solved have become far more complex. Therefore, the development of fast algorithms has become even more important.

Optimization in networks is one of the most important areas of optimization, since many practical problems involve networks (*graphs*) and many other problems can be modelled using graphs. In the 1930s and 1940s, pioneering work has been done in graph theory [7] and resource allocation, including the work by the Dutch Tjalling Koopmans who received the Nobel prize in Economics for this work in 1975 together with the Russian economist Kantorovich. The field started flourishing in the 1950s when efficient algorithms for graph problems were found, including the famous algorithms for the

"minimum spanning tree" problem that were found in 1965 (Kruskal), 1957 (Prim) and 1959 (Dijkstra), although, in fact, the Czech mathematician Borøuvka already found an efficient algorithm for "minimum spanning tree" problem in 1926. In the 1950s, also the first efficient algorithms for the shortest path problem were found, including the famous algorithm by Dijkstra (1959). Many of these early algorithms are still being used today.

In this lecture, we will discuss efficient algoritms for several optimization problems in networks.

3.2 Shortest paths

Many people regularly use services like *google maps* to find a shortest route from one location to another. To find your route, google maps needs to solve a variant of the *shortest path problem*, one of the oldest and best studied problems in applied mathematics. This problem can be easily modelled with a network, which is usually called a *graph* in mathematics, see Figure 3.1 for a simple example. The vertices of the graph represent for example street crossings while the streets between crossings are represented by edges (which are undirected) or arcs (which are directed). Each edge or arc is labelled with a positive number representing for example length or travel time (but we will always refer to this label as the "length" of the edge).

In the network in Figure 3.1, there are multiple directed paths from vertex s to vertex t. A shortest directed path is indicated in bold.

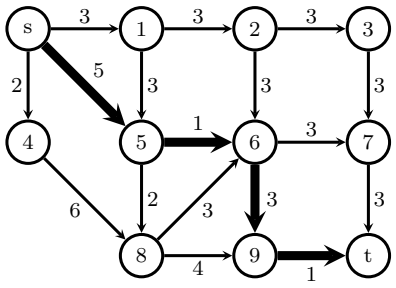


Figure 3.1: Example of a shortest-path problem.

It is clear that in practice the number of vertices and arcs will be much bigger than in this simple example. In 1958, a network with 265 vertices was studied. At the Western Joint Computer Conference in Los Angelos, it was proudly reported that all shortest paths between vertices of this network had been found. Finding these shortes paths took about three hours on an IBM 704 computer (an enormous device). Modern networks contains millions of vertices and results are expected within seconds.

Shortest-path problems have a much larger application area than one would expect at first sight:

- A crucial problem in speach recognition (automatically transforming spoken to written language) is to distinguish similar-sounding words, like *to*, *two* and *too*. Form a directed graph with possible words of some sentence as vertices and an

arc between two vertices that could follow each other in the sentence. The length of each arc is some measure of how likely the two words are to be together in a sentence. Then a best possible interpretation of the sentence can be found by searching for a shortest path from the first to the last word. See Figure 3.2.

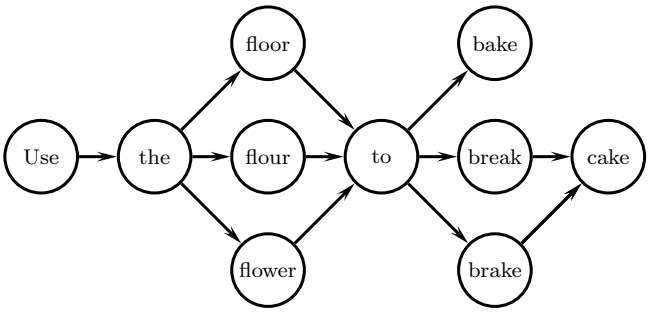


Figure 3.2: Example of a sentence.

- Image segmentation is a different example. Consider the problem of distinguishing two objects on a digital image (for example an MRI scan). This can be done by finding a line between two points containing a minimum number of dark pixels. The pixels are the vertices of the graph and the length of an edge is calculated in such a way that light pixels are connected by short edges and darker pixels by longer edges. A shortest path between the two points forms an optimal way to divide the image. A similar technique is used to find the contour of, for example, the heart in an X-ray image.

In many applications, large numbers of shortest paths need to be found, which makes it even more important to have a fast algorithm. The afore-mentioned Dijkstra algorithm is one of the best known and most used algorithms and will be described below.

Some definitions

A *directed graph* consists of a finite set of *vertices* V and a set of *arcs* A, where each arc is an ordered pair of vertices. If $a = (u, v) \in A$, then u is the *tail* of a and v its *head*. A *(directed) path* in a directed graph is a sequence of vertices $P = (v_1, v_2, \ldots, v_{k+1})$ such that $(v_i, v_{i+1}) \in A$ for all $1 \leq i \leq k$. We assume that each arc has a non-negatie length. The length of arc (u, v) is denoted c_{uv}. The *length* $\ell(P)$ of path P is defined as the some of the lengths of the arcs on the path:

$$\ell(P) := \sum_{i=1}^{k} c_{v_i v_{i+1}}$$

Exercise 3.1 Determine the number of paths from s to t in the directed graph in Figure 3.1.

Dijkstra's algorithm

Dijkstra's alorithm finds a shortest path from s to t, where s and t are two arbitrary vertices in the graph. The algorithm labels each vertex with a number π_v, which is always greater or equal to the length of a shortes path from s to v. The value of π_v at a certain point in the execution of the algorithm is equal to the length of the shortest path from s to v found so far. After finishing the execution of the algorithm, π_t is equal to the length of a shortest path from s to t.

In each iteration, a vertex is *investigated* and the algorithm keeps track of a set Q of vertices that still need to be investigated. Initally, the set Q contains only vertex s, each vertex v has label $\pi_v = \infty$ (the mathematical symbol for infinity) except for vertex s, which gets label 0 (the length of a shortest path from s to s).

A single iteration consists of choosing a vertex u from the set Q and investigating u. We choose a vertex with smallest label π_u, over all vertices in Q. This vertex is removed from Q and investigated as follows.

When investigating vertex u, we update the label π_v, if necessary. If the path from s via u to v is shorter than π_v, then we set π_v to the length of this shorter path. Hence we give v as new label the sum of π_u and the length c_{uv} of the arc from u to v. We also add the vertex v to Q. Figure 3.3 describes the algoriithm in pseudocode.

Initialisation:
$Q := \{s\}$;
$\pi_s := 0$;
$\pi_v := \infty, \forall v \in V \setminus \{s\}$;

while Q is not empty:
    Choose $u \in Q$ such that $\pi_u \leq \pi_v$ for all $v \in Q$;
    Investigate(u)
endwhile

Figure 3.3: Algorithm of Dijkstra.

The procedure Investigate(u) is give in Figure 3.4.

begin
    for all $a = (u, v) \in A$ do the following:
    if $\pi_v > \pi_u + c_{uv}$ then
        $\pi_v := \pi_u + c_{uv}$;
        add v to Q
    end;
    remove u from Q
end

Figure 3.4: Procedure Investigate(u).

As an example we apply the algorithm to the directed graph from Figure 3.1. The

result is summarized in Figure 3.5. Each row of the table in this figure is one iteration, and the table entries contain the values of π_v. The elements of Q are given by a circle around the value of π_v. The element that is chosen to be investigated has a square instead of a circle.

	iteraties											
vertex	1	2	3	4	5	6	7	8	9	10	11	12
s	$\boxed{0}$	0	0	0	0	0	0	0	0	0	0	0
1	∞	③	$\boxed{3}$	3	3	3	3	3	3	3	3	3
2	∞	∞	∞	⑥	$\boxed{6}$	6	6	6	6	6	6	6
3	∞	∞	∞	∞	∞	⑨	⑨	$\boxed{9}$	9	9	9	9
4	∞	$\boxed{2}$	2	2	2	2	2	2	2	2	2	2
5	∞	⑤	⑤	$\boxed{5}$	5	5	5	5	5	5	5	5
6	∞	∞	∞	∞	⑥	$\boxed{6}$	6	6	6	6	6	6
7	∞	∞	∞	∞	∞	∞	⑨	⑨	$\boxed{9}$	9	9	9
8	∞	∞	⑧	⑧	⑦	⑦	$\boxed{7}$	7	7	7	7	7
9	∞	∞	∞	∞	∞	∞	⑨	⑨	⑨	$\boxed{9}$	9	9
t	∞	∞	∞	∞	∞	∞	∞	∞	∞	⑫	$\boxed{10}$	10

Figure 3.5: Execution of Dijkstra's algorithm.

We conclude from the table that the length of a shortest path is equal to 10. We can also use the table to find a shortest path by "backtracking". The label 10 of vertex t first appeared when investigating vertex 9, which had label 9 at that time. This label first appeared when investigating vertex 6, etc. Continuing this way we find out that $P = (s, 5, 6, 9, t)$ is a shortest path from s to t.

Exercise 3.2 Prove that the label π_v of a vertex v is never smaller than the length of a shortest path from s to v.

Exercise 3.3 When a vertex is investigated, it is removed from Q. Prove that it will never be added to Q again.

Exercise 3.4 When a vertex u is investigated, then the label of this vertex becomes *permanent* (i.e. it will not be changed any more). Prove this. Then use this to show that the label of this vertex is then equal to the length of a shortest path from s. naar die knoop.

Exercise 3.5 If the algorithm terminates, then the label π_v of each vertex v is equal to the length of a shortest path from s to v (unless no such path exists, in which case $\pi_v = \infty$). Prove this.

From Exercise 3.5 follows that Dijkstra's algorithm does not only find the lengths of a shortest path from s to t, but the length of a shortest path from s to each other vertex. Therefore, we call it an *one-to-all* shortest path algorithm.

If we only want to know a shortest path from s to t, then we can terminate the algorithm when vertex t is ready to be investigated. A shotest path from s to t has then be found, according to Exercise 4. This variant of Dijkstra's algorithm is a *one-to-one* shortest-path algorithm.

Also note that we can also find a shortest path in an undirected graph using Dijkstra's algorithm. We just replace each undirected edge $\{u, v\}$ by two arcs $(u, v), (v, u)$ both having the same length as $\{u, v\}$ and then apply Dijkstra's algorithm to the obtained directed graph.

The efficiency of an algorithm is determined by the number of elemental steps that the algorithm needs in the worst case, where elemental steps are for example adding/multiplying/dividing/subtracting two numbers, comparing two numbers, adding an element to a set, etc. To describe the efficiency of the algorithm, we need to bound the number of elemental steps that the algorithm needs for a network with n vertices and m arcs. Since we are mostly interested in the behaviour of the algorithm for large values of n and m, we usually simplify the expression using "Big-O" notation. For example, if the expression is $7n^2m + 8n + 5$ then we are mostly interested in the term $7n^2m$, which tells us that when n becomes twice as large, the running time of the algorithm becomes roughly four times as large. This conclusion is independent from the constant 7 in the expression. Therefore, we omit also the 7 and write that the algorithm has running time $O(n^2m)$, where the symbol O indicates that we only describe the order of magnitude of the running time.

Exercise 3.6 Verify that the number of elemental steps in Dijkstra's algorithm is $O(n^2)$.

For now we assume that there exists a path from s to each other vertex. After executing Dijkstra's (one-to-all) algorithm the labels π_v satisfy:

$$\pi_s = 0, \quad \pi_v - \pi_u \leq c_{uv}, \quad \forall (u, v) \in A. \tag{3.1}$$

Exercise 3.7 Prove (3.1).

Now let $P = (s = v_0, v_1, v_2, \ldots, v_{k-1}, t = v_k)$ be an arbitrary path from s to t. Then we may write

$$\ell(P) = \sum_{i=0}^{k-1} c_{v_i v_{i+1}} \geq \sum_{i=0}^{k-1} (\pi_{v_{i+1}} - \pi_{v_i}) = \pi_{v_k} - \pi_{v_0} = \pi_t - \pi_s = \pi_t.$$

In other words, if P is an arbitrary path from s to t, and π satisfies (3.1), then $\ell(P) \geq \pi_t$. Dijkstra's algorithm finds a path P as well as a labelling π satisfying $\ell(P) = \pi_t$. The following theorem is a consequence of this.

Theorem 3.1. *The length of a shortest path from s to t is equal to*

$$\max\{\pi_t : \pi_s = 0, \quad \pi_v - \pi_u \leq c_{uv}, \quad \forall (u, v) \in A\}.$$

This is the duality theorem for the shortest-path problem. The importance of this (and every) duality theorem is the following. Once we have found a shortest path P, we can use the labels π_v to vertify that P is indeed a shortest path. If the labels satisfy (3.1) and $\ell(P) = \pi_t$, then this proves non-algorithmically that P is a shortest path.

Maximum flow

We consider again a directed graph $D = (V, A)$, with vertex set V and arc set A. For each arc (u, v), again a positive number c_{uv} is given, which now describes the *capacity* of the arc. Two special vertices are called s (the *source*) and t (the *sink*), and we are asked to find a maximum flow from s to t. A *flow* is an assignment of a value $x_{uv} \in \mathbb{R}_{\geq 0}$ to each arc (u, v) such that in every vertex other than s and t the flow is preserved (total inflow equals total outflow) and $x_{uv} \leq c_{uv}$ (the flow does not exceed the capacity) on each arc (u, v).

In other words, x needs to satisfy the following *balance equations*:

$$\sum_{(u,v)\in A} x_{uv} = \sum_{(v,w)\in A} x_{vw}, \quad \text{for all } v \in V \setminus \{s, t\} \tag{3.2}$$

and the *capacity limitations*:

$$0 \leq x_{uv} \leq c_{uv}, \quad \text{for all } (u, v) \in A. \tag{3.3}$$

The *value* of the s–t flow x is defined as

$$\text{value}(x) = \sum_{(s,v)\in A} x_{sv} - \sum_{(v,s)\in A} x_{vs}.$$

Exercise 3.8 Verify that the value of an s–t flow x is always equal to the net inflow at t:

$$\text{value}(x) = \sum_{(v,t)\in A} x_{vt} - \sum_{(t,v)\in A} x_{tv}.$$

In Figure 3.6, an example of a max-flow problem is given.

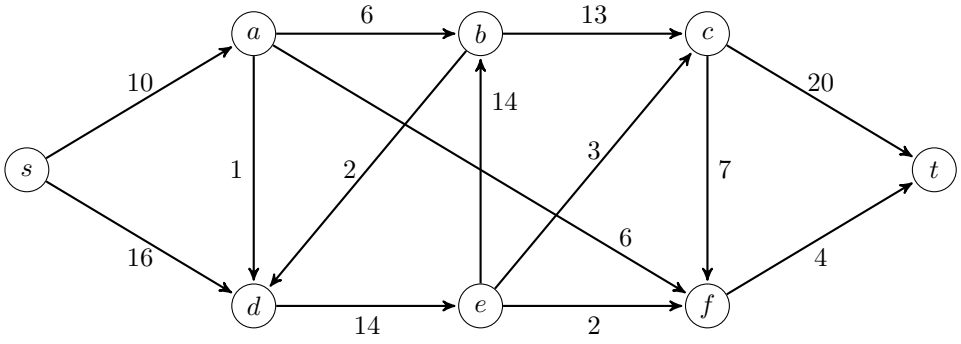


Figure 3.6: Maximum flow problem

Notice that $x_{vw} = 0, (v, w) \in A$ defines a valid flow with value 0. Flows with higher values are easy to find: take an arbitrary path from s to t and send as much flow over the path as possible. For example, over the path (s, a, f, t) in Figure 3.6, we can send a flow of value 4 without violating the capacity constraints on this path.

In addition to that, we can send a flow of 13 over the path (s, d, e, b, c, t). Together, this gives a flow of value 17. This flow is indicated in Figure 3.7.

At every arc (u, c) the flow value and the capacity are given as $x_{uv}|c_{uv}$.

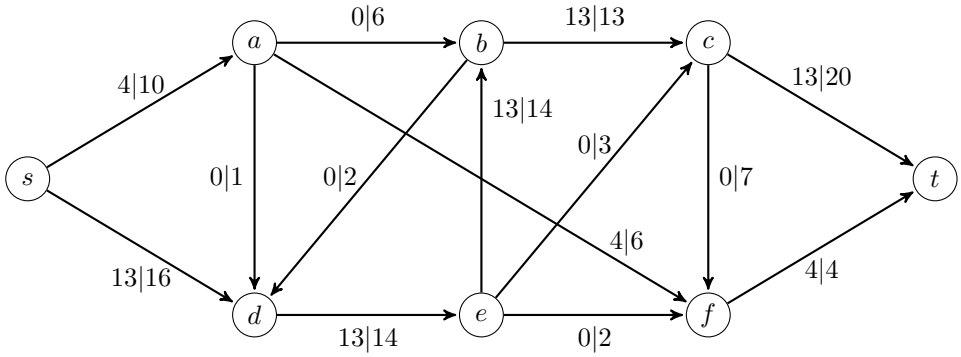


Figure 3.7: A flow with value 17.

The Ford-Fulkerson Algorithm

Given a flow, the Ford-Fulkerson Algorithm (1956) basically searches for a path from s to t over which the flow can be increased or *augmented*. Somehow surprisingly, this algorithm always finds a maximum flow. If there is no path from s to t over which the flow can be augmented, then the flow has maximum possible value, which we will prove below. Importantly, these augmenting paths may also traverse arcs in reverse direction, in which case the flow is decreased, instead of increased, on those arcs.

First, we will describe a systematic method for finding an augmenting path. First, given a certain flow x, we construct an auxiliary directed graph D_x. This directed graph has the same vertices as D, but not the same arcs. Each arc $(u, v) \in A$ also becomes an arc of D_x if it is possible to send more flow over the arc, i.e. if $x_{uv} < c_{uv}$. In addition, for each arc $(u, v) \in A$ over which the flow can be decreased (i.e. if $x_{uv} > 0$), the reverse arc (v, u) is included in D_x. In other words, the set A_x of arcs of D_x is

$$A_x = \{(u, v) \in A : x_{uv} < c_{uv}\} \cup \{(v, u) : (u, v) \in A, x_{uv} > 0\}.$$

If $x_{uv} < c_{uv}$ then the arc (u, v) gets label $c_{uv} - x_{uv}$ in D_x, and if $x_{uv} > 0$ then arc (v, u) gets label x_{uv} in D_x.

We illustrate this using the flow in Figure 3.7. The auxiliary directed graph is depicted in Figure 3.8.

In the directed graph in Figure 3.8, there is a directed path from s to t, indicated in bold. Moreover, we see that over this path we can send at most 1 extra unit of flow. Augmenting the flow by one over this path, we get the new flow given in Figure 3.9.

To find out whether it is possible to further improve the flow, we construct the new auxiliary directed graph depicted in Figure 3.10.

We now find a new augmenting path (s, a, b, e, c, t). This path traverses the arc (e, b) (from the original directed graph, see Figure 3.9) in reverse direction. Therefore, the flow is not increased but decreased on this arc. Therefore, we can augment the flow by $\min(6, 6, 13, 2, 6) = 2$. This way we get the new flow depicted in Figure 3.11.

To determine whether more improvements are possible, we construct the new auxiliary directed graph, see Figure 3.12.

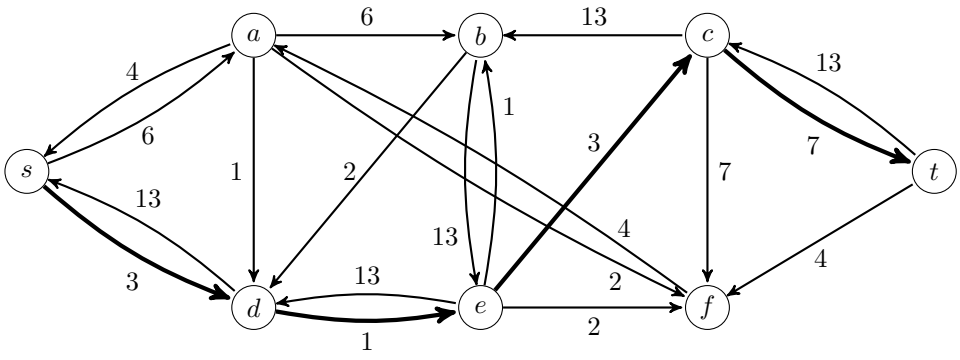


Figure 3.8: Auxiliary directed graph for the flow in Figure 3.7.

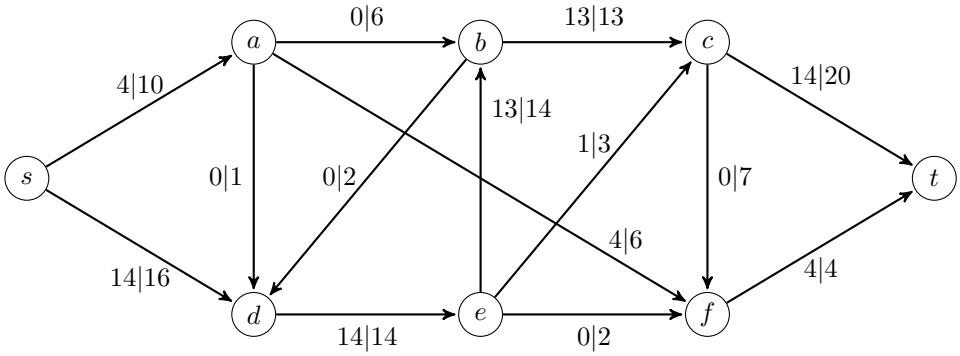


Figure 3.9: A flow with value 8.

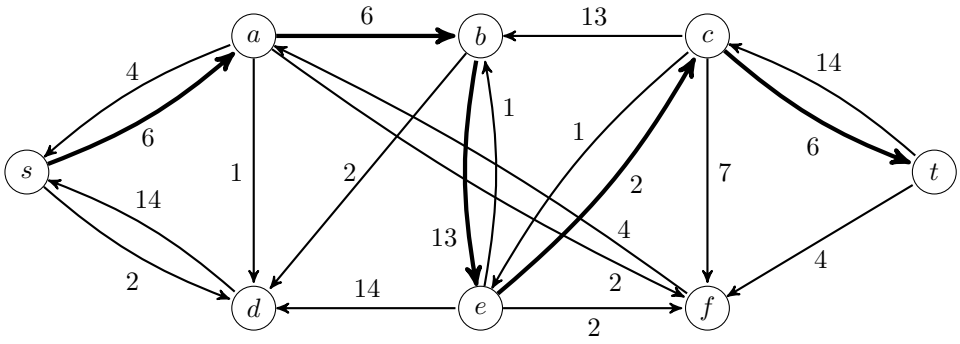


Figure 3.10: Auxiliary directed graph for the flow in Figure 3.9.

At first sight, it seems that no more improvement is possible. To check this systematically, we label all vertices reachable from s with a $*$. We can do this using a simplified version of Dijkstra's algorithm. We first label s with a $*$ and define $Q = \{s\}$. Then we choose a vertex $u \in Q$ and investigate vertex u, i.e. all unlabelled vertices that are

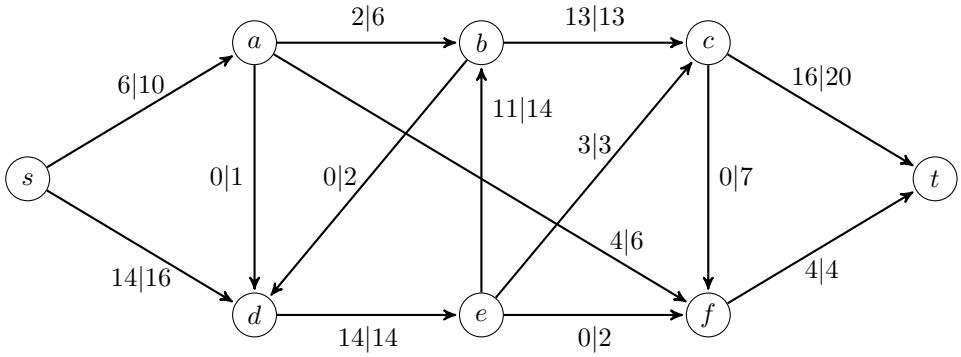


Figure 3.11: A flow with value 20.

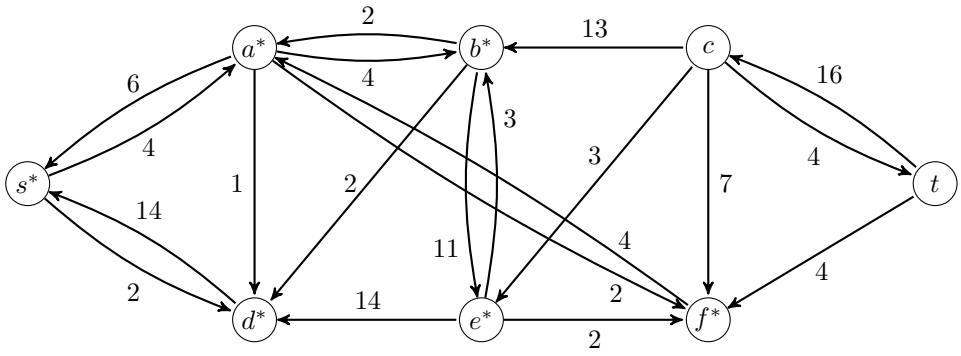


Figure 3.12: Auxiliary directed graph for the flow in Figure 3.11.

reachable from u by traversing a single arc are labelled with a $*$ and added to Q. Then we remove u from Q. We repeat this untill Q is empty. It is clear that after this process all vertices reachable from s are labelled $*$ while all vertices not reachable from s are unlabelled. Hence, there exists an augmenting path if and only if vertex t is labelled.

Since we get the labelling given in Figure 3.12, there is no augmenting path. We will now prove that it follows that the flow in Figure 3.11 has maximum value. We do this as follows.

We define a vertex set S as follows:

$$S = \{v \in V : \ v \text{ is labelled } *\}.$$

It is clear that $s \in S$ and, since there is no augmenting path, $t \notin S$. The set of arcs with tail in S and head not in S is denoted as $\delta^+(S)$, i.e.

$$\delta^+(S) := \{(u, v) \in A : u \in S, v \notin S\}.$$

A sketch of this situation can be found in Figure 3.13.

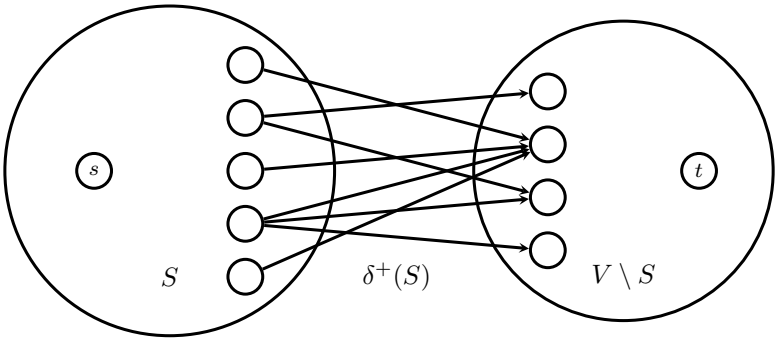


Figure 3.13: The s–t cut $\delta^+(S)$.

If we delete the arcs of $\delta^+(S)$ from A, then there is no more path from s to t. Hence we call $\delta^+(S)$ an *s-t cut* and define the *capacity* $c(\delta^+(S))$ of this cut as follows:

$$c(\delta^+(S)) := \sum_{(u,v)\in\delta^+(S)} c_{uv} = \sum\{c_{uv} : u \in S, v \notin S\}.$$

In the example, $\delta^+(S)$ contains the arcs (b,c), (e,c) and (f,t), and we have $c(\delta^+(S)) = 13+3+4 = 20$. This is exactly the value of the current flow x, which is not a coincidence, as we will show now. In other words, we will show that if x is a flow for which there exists no augmenting path, and S is the set of vertices labelled $*$, then

$$\text{value}(x) = c(\delta^+(S)).$$

For an arbitraty set of vertices U with $s \in U$ and $t \notin U$ is $\delta^+(U)$ the corresponding s-t cut. Let y be an arbitrary flow. From the observation that each path from s to t contains at least one arc from $\delta^+(U)$, it follows that

$$\text{value}(y) \leq c(\delta^+(U)).$$

In particular, $\text{value}(x) \leq c(\delta^+(S))$. Conversely, for each arc $(u,v) \in \delta^+(S)$, vertex u is labelled and v is not. This means that this arc is not contained in the auxiliary directed graph, since otherwise vertex v would also be labelled. From this we conclude that the arc is *saturated*, i.e. $x_{vw} = c_{vw}$.

Now consider the arcs $(u,v) \in \delta^-(S)$, which have their tail $u \notin S$ and their head $v \in S$. These arcs have no flow, since otherwise vertex u would also be labelled. Hence it follows indeed that $\text{value}(x) = c(\delta^+(S))$.

The above proves the max-flow min-cut theorem, the duality theorem for the maximum flow problem.

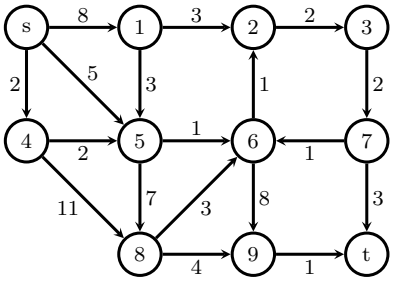
Theorem 3.2 (Max-flow min-cut theorem)**.** *The value of a maximum flow is equal to*

$$\min\{c(\delta^+(U)) : U \subseteq V, s \in U, t \notin U\}.$$

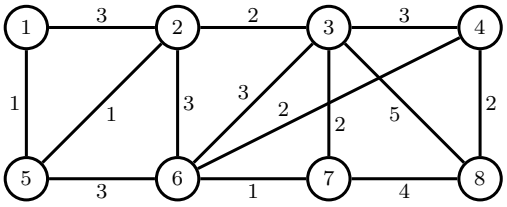
We conclude that the flow x in Figure 3.9 has maximum value, which is proved by the s-t cut formed by the arcs (b,c), (e,c) and (f,t).

3.3 Exercises

Exercise 3.9 Use Dijkstra's algorithm to solve the shortest path problem in the directed graph below.



Exercise 3.10 Consider the graph below.



a) Use Dijkstra's algorithm to determine the distance over a shortest path from vertex 1 to each other vertex.

b) Draw the shortest paths in the graph.

Exercise 3.11 If one or more arcs have negative lengths then it becomes more difficult to find a shortest path, especially when there are circuits with negative total length. In such a case, a shortest path may not exist (using the definition used here, where a path may use a vertex multiple times). Give an example where this is the case.

Can you think of a method for finding negative-length circuits?

Exercise 3.12 Let $\pi : V \to \mathbb{R}$. We call π a *potential* for directed graph $D = (V, A)$ if

$$\pi_v - \pi_u \leq c_{uv}, \quad \forall (u, v) \in A.$$

The Theorem of Gallai says

$$\text{There exists a potential} \quad \Longleftrightarrow \quad D \text{ has no negative-length circuit.}$$

a) Prove the implication '$\Longrightarrow$'.

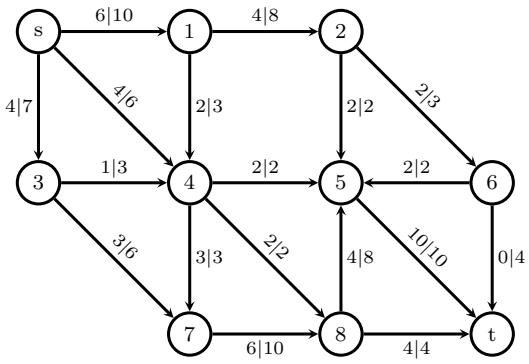
b) Prove the implication '$\Longleftarrow$'.
Hint: take for $\pi(v)$ the length of a shortest path starting at v. Why does such a shortest path exist and why is the resulting π a potential?

Exercise 3.13 Given is the directed graph below.

a) What is the capacity of the cut $\delta^+(U)$ if $U = \{s, 1, 5\}$.

b) Improve the flow to a maximum flow. What is the value of the maximum flow?

c) Prove that your flow has maximum value by finding a cut with capacity equal to the value of the flow. This cut has minimum capacity.



Literature

[1] R. K. Ahuja, T.L. Magnanti, and J.B. Orlin. *Network Flows: Theory, Algorithms, and Applications.* Prentice-Hall, Englewood Cliffs, New Jersey, 1993.

[2] V.K. Balakrishnan and C. Moire. *Network Optimization*, volume 2 of *Chapman and Hall Mathematics Series.* CRC Press, Inc, UK, 1995.

[3] M.O. Ball, T.L. Magnanti, Monma C.L., and G.L. Nemhauser. *Network Models.* Elsevier Science, Amsterdam, 1995.

[4] D.P. Bertsekas. *Network Optimization: Continuous and Discrete Models.* Athena Scientific, P.O. Box 391, Belmont, MA, USA, 1998.

[5] D.Z. Du and P. M. Pardalos. *Network Optimization Problems: Algorithms, Applications And Complexity*, volume 2 of *Series on Applied Mathematics.* World Scientific Publishing Company, Incorporated, 1993.

[6] Fred Glover and Nancy V. Phillips. *Network Models in Optimization and Their Applications in Practice.* John Wiley & Sons, UK, 1992.

[7] D. König. *Theorie der endlichen und unendlichen Graphen.* Reprinted by Chelsea in 1950, New York, USA, 1936.

[8] A. Schrijver. *Combinatorial Optimization. Polyhedra and Efficiency.* Springer, Berlin, 2003. 3 volumes.

4

Differential Equations

Auteur: H.M. Schuttelaars

Introduction

Differential equations are equations that give a relation between the derivatives of a function and the function value itself. For example, a differential equation for the velocity of an object falling to the earth reads

$$\frac{dv}{dt} = 9.8 - \frac{v}{5}, \tag{4.1}$$

where the acceleration of the object dv/dt is related to its velocity v. The velocity v will presumably change with time, so the solution(s) to the differential equation v are function(s) of t. The variable t is called the *independent* variable and v is the *dependent* variable. Equation (4.1) is called an *ordinary* differential equation, because the dependent variable v only depends on one independent variable. If the dependent variable depends on more than one independent variable (for example time t and position x), the dynamic behavior of the dependent variable is usually described by a *partial* differential equation.

When studying differential equations, it is important to know if there are solutions and, if solutions exist, whether there is only one solution or possibly more. In the example above, it is easily seen that the constant (in time) solution $v = 49$ is a solution of the differential equation. It turns out that this is not the only solution, the most general solution reads $v(t) = 49 + c\exp(-t/5)$ with $c \in \mathbb{R}$ an unknown constant. By specifying an initial condition (the velocity at $t = 0$) the unknown constant can be determined.

Equation (4.1) is an example of a differential equation following from a physical conservation law. Many of the principles (or laws) underlying the behavior of the natural world are expressed in terms of (partial) differential equations, other examples are the motion of fluids, the flow of current in electric circuits, the dissipation of heat in solid objects, the propagation of waves, and the increase or decrease of populations. Differential equations that describe physical/biological/economical processes are often called mathematical models. It is noteworthy that simple differential equations can already provide useful models for important processes in physics, biology, economy, engineering, etc.

In the following, we will restrict ourselves to the analysis of *ordinary* differential equations. In section 4.1, we begin with two models leading to equations that are easy to solve. In section 4.2 the geometrical interpretation of first order differential equations

will be discussed. Solution methods for specific classes of differential equations will be given in section 4.3. In section 5, some references are given.

4.1 Some Examples

Population Dynamics

In this section, we will develop a model for the evolution of a population (i.e. a collection of individuals of a particular species) that lives within a well-defined area. The changes in the number of individuals within this population is a result of reproduction, death or migration of individual organisms. Ignoring migration, the population model reads in words:

$$\text{population change} = \text{births} - \text{deaths}. \tag{4.2}$$

This total number of individuals at a specific time t will be denoted by $N(t)$. The change in population size in a small time interval Δt is the difference between $N(t + \Delta t)$ and $N(t)$. The balance equation (4.2) can now be written as

$$\begin{aligned} N(t + \Delta t) - N(t) = {} & \text{number of births during } \Delta t \\ & - \text{number of deaths during } \Delta t, \end{aligned} \tag{4.3}$$

where the number of births and deaths during the time interval Δt can be a function of t itself. A continuous–time version of the balance equation (4.3) is obtained by dividing both sides of the equation by Δt and taking the limit $\Delta t \to 0$. The resulting equation reads

$$\frac{dN}{dt} = B(N) - D(N), \tag{4.4}$$

with $B(N)$ the *population birth rate* and $D(N)$ the *population death rate*, that are explicit functions of the total number of individuals $N(t)$. Since all individuals in the population are assumed to be identical, the population birth rate can be expressed in terms of individual–level birth rate $b(N)$ by

$$b(N) = \frac{B(N)}{N}.$$

Similarly the individual–level death rate is defined as

$$d(N) = \frac{D(N)}{N}.$$

Substituting these expressions in equation (4.4) results in the following general continuous–time population balance equation:

$$N' = b(N)N - d(N)N, \tag{4.5}$$

where N' is a shorthand notation for dN/dt. Equation (4.5) does not specify a complete population dynamic model yet, as it only determines how the size of the population changes in time. We still have to specify the size of the population at some particular moment in time, i.e., we have to specify an *initial condition*. Often this initial condition is prescribed at $t = 0$:

$$N(0) = N_0,$$

with N_0 a given population size at time $t = 0$.

There are many possible choices for the individual–level birth and death rate. The simplest choice is to assume that both rates are constant, and hence independent of the population size N. The resulting population model reads

$$N' = rN, \tag{4.6}$$

with the *population growth rate* r the difference between the constant birth and death rate. The equation is called the exponential growth equation or "Malthus' growth law".

Exercise 4.1 Given the initial population size $N(0) = N_0$, show that

$$N(t) = N_0 \exp(rt)$$

is a solution to equation (4.6). Later we will see that this is the only solution (see Theorem 1). Malthus (1798) investigated the birth and death register of his parish and concluded that the population of his parish doubled every 30 years. Give an estimate for the parameter r (do not forget the unit of this parameter).

The exponential growth of the population size, predicted by this model, is not very realistic because the population would quickly exhaust its natural resources. Hence the model formulation has to be adjusted to account for the fact that the availability of resources is limited. One way to account for this is by assuming that the individual–level birth rate decreases linearly with an increasing population and reaches a value of 0 at some prescribed population size $N = N_{\max}$. Assuming the individual–level death rate is still independent of the population size N, the following expressions are found:

$$b(N) = \beta \left(1 - \frac{N}{N_{\max}}\right), \qquad d(N) = \delta.$$

Inserting these expressions in equation (4.5) results in the logistic growth equation:

$$N' = rN\left(1 - KN\right). \tag{4.7}$$

In section 4.3, we will show how to solve this ordinary differential equation.

Exercise 4.2 Exercise 2: Give explicit expressions for r and K in terms of β, δ and $N_{\max}$, using the definitions for $b(N)$ and $d(N)$

Torricelli's Law

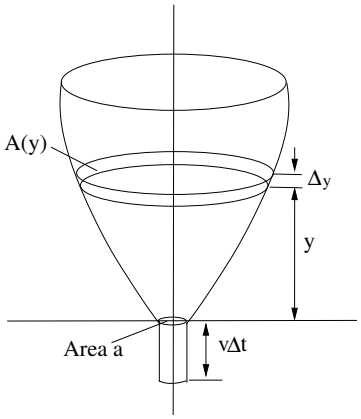


Figure 4.1: Derivation of Torricelli's law.

Water from a water tank is draining through a hole with area a at its bottom. The depth of water in the tank at time t is denoted by $y(t)$ and the volume of water in the tank by $V(t)$. Under ideal conditions the velocity of the stream of water exiting through the hole is

$$v = \sqrt{2gy},$$

with $g = 9.8\,\mathrm{ms}^{-2}$. The velocity v is the velocity a drop of water would acquire in falling freely from the water surface to the hole. This is Torricelli's law of draining. As indicated in Figure 4.1, the amount of water that leaves through the bottom hole during a short time interval Δt amounts to a cylinder with base area a and height $v\Delta t$. Hence the resulting change ΔV in the volume of water in the tank is given by

$$\Delta V = -av\Delta t = -a\sqrt{2gy}\Delta t.$$

If $A(y)$ denotes the horizontal cross-sectional area of the tank at height y above the hole, the change in volume can be written as

$$\Delta V = A(y)\Delta y,$$

resulting in a differential equation for $y(t)$:

$$A(y)y' = -a\sqrt{2gy}. \tag{4.8}$$

Solutions to this equations will be studied in section 4.3.

Exercise 4.3

a) Consider a cylindrical tank, i.e., a tank with a constant cross–section A. Show that equation (4.8) reduces to

$$y' = -k\sqrt{y},$$

with k a constant. Give an explicit expression for k.

b) Next consider a hemispherical tank (i.e., a half of a sphere) with a top radius of 4 m, and a hole at the bottom of 10 cm. Find an explicit expression for $A(y)$ and give an equation for the water depth in this tank.

Springs and Masses

Ordinary differential equations also arise in the study of mechanics. Consider a mass m attached to the end of a spring, as shown in Figure 4.2. The displacement $x(t)$ from its equilibrium position is governed by Newton's law:

$$mx'' = m\frac{d^2x}{dt^2} = F(x,t), \tag{4.9}$$

where $F(x,t)$ represents the forces acting on the mass. In the usual spring and mass model, the net force acting on the mass is considered to be the sum of three terms. The first term is a restoring force kx, which pulls the mass back toward the equilibrium position. The second term describes the damping force (due to friction) cx'. The constants k and c are assumed to be nonnegative. The last term is a time-dependent force $f(t)$ which is independent of the position or velocity of the mass. Given an initial displacement $x(0) = x_0$ and velocity $x'(0) = v_0$, the linear spring model consists of a differential equation with initial conditions

$$mx'' + cx' + kx = f(t), \qquad x(0) = x_0 \quad \text{and} \quad x'(0) = v_0. \tag{4.10}$$

It can be shown that, for any continuous function $f(t)$ on the interval $[0, b)$, there is a unique solution $x(t)$ with two continuous derivatives on $[0, b)$ that satisfies equation (4.10).

Classification of Ordinary Differential Equations

The *order* of a differential equation is the order of the highest order derivative present in the equation. In the examples above, equations (4.1), (4.6), (4.7) and (4.8) are first order equations, and equation (4.10) is a second order ordinary differential equation (remember that a differential equation is called *ordinary* if the dependent variable only depends on *one* independent variable).

Differential equations are also classified as *linear* or *non–linear*. A first order ordinary differential equation is linear if the differential equation can be written as

$$x' = bx + a, \tag{4.11}$$

with the coefficients a and b independent of x. Note that a and b may involve the independent variable t. Similarly, a second order differential equation is linear if it can be written as

$$x'' = cx' + bx + a, \tag{4.12}$$

with the coefficients a, b and c independent of x. In the examples above, equations (4.1), (4.6) and (4.10) are linear, while equations (4.7) and (4.8) are nonlinear.

Finally, it is important to distinguish between *homogeneous* and *inhomogeneous* equations. An equation is called *homogeneous* if there are no terms in the equation that do not depend on the dependent variable. If such a term is present, the equation is called inhomogeneous. Hence equations (4.11) and (4.12) are homogeneous if $a = 0$. In the explicit examples given in the previous sections, all equations are homogeneous, except for the equation describing the motion of the forced mass–spring system.

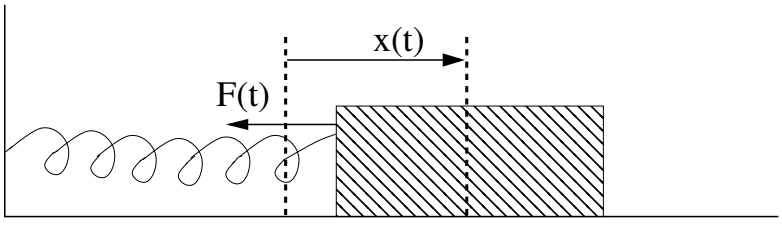


Figure 4.2: The displacement $x(t)$ for a mass–spring system.

Exercise 4.4 It is often useful to write a higher order differential equation as a system of first order differential equations (this will for example be used during the first year modeling course). As an example, rewrite equation (4.10) as a system of (two) coupled first order differential equations by introducing $y(t) = x'(t)$.

4.2 Direction Fields

The behavior of solutions of first order ordinary differential equations can be studied from a geometrical viewpoint. As an example, we consider equation (4.1), which is repeated for convenience:

$$\frac{dv}{dt} = 9.8 - \frac{v}{5}.$$

For a given value of the speed v, the right side of equation (4.1) can be evaluated, resulting in a corresponding value of dv/dt. For instance, if $v = 40$, then $dv/dt = 1.8$. This means that the slope of a solution $v = v(t)$ has the value 1.8 at any point where $v = 40$. We can display this information graphically in the tv–plane by drawing short line segments, or arrows, with slope 1.8 at several points on the line $v = 40$. Similarly, if $v = 50$, then $dv/dt = -0.2$, so we draw line segments with slope -0.2 at several points on the line $v = 50$. By proceeding in the same way with other values of v, the left panel in Figure 4.3 is obtained. This is an example of what is called a *direction* or slope field. Each line

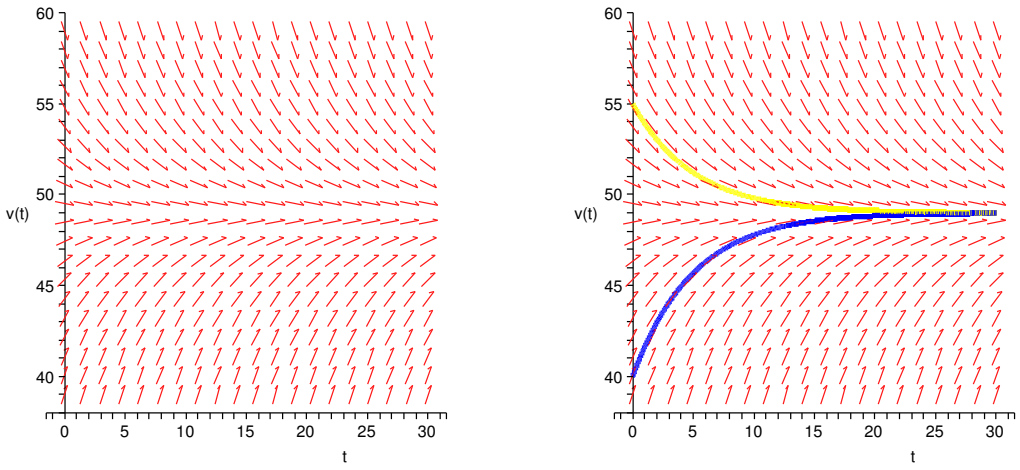


Figure 4.3: Left panel: Direction field for equation (4.1). Right panel: same direction field but now with solution curves for initial conditions $v(0) = 40$ (blue) and $v(0) = 55$ (yellow).

segment is a tangent line to the graph of a solution of equation (4.1). Drawing such a direction field is often very time consuming, but can be done with the aid of computer programs such as MAPLE, MATLAB, MAXIMA, PYTHON, etc.

If an initial condition is chosen, a *solution curve* can be constructed by drawing the line that is tangent to the line segments in the direction field. In Figure 4.3, right panel, the solution curves are plotted for two different initial conditions, $v(0) = 40$ and $v(0) = 55$.

Exercise 4.5 Even though we have not found any explicit solutions, we can nonetheless draw some qualitative conclusions about the behavior of solutions. Indicate the *equilibrium* solution in Figure 4.3, i.e., the solution $v(t)$ that does not change in time. What happens with initial velocities v greater than this equilibrium value? And smaller than this value? Is the equilibrium solution (linearly) stable, i.e, does $v(t)$ evolve back to the equilibrium solution for initial velocities close to the equilibrium velocity?

Exercise 4.6 Construct the direction field for equation (4.7). Are there equilibrium solutions? Are these solutions stable?

The analysis of the above example suggests that, given an initial condition, there is only one *unique* solution. This raises the question of whether this is true of all initial value problems for first order equations. In other words, does every initial value problem have exactly one solution? The answer to this question is given by the following fundamental theorem:

Theorem 4.1. *Let the functions $f(t,u)$ and $\partial f(t,u)/\partial u$ be continuous in some rectangle $\alpha < t < \beta$, $\gamma < u < \delta$ containing the point (t_0,u_0). Then, in some interval $t_0 - h < t < t_0 + h$ contained in $\alpha < t < \beta$, there is a unique solution $u = \phi(t)$ of the initial value problem*

$$u' = f(t,u), \qquad u(t_0) = u_0.$$

In this theorem, we require the functions f and its first *partial derivative* $\partial f/\partial u$ to be *continuous*, concepts that will be defined in detail in later courses. Briefly, continuity of f at (t_0,u_0) means that the value $f(t_0,u_0)$ is defined and that the value $f(t,u)$ is close to $f(t_0,u_0)$ if the point (t,u) is close to (t_0,u_0). The partial derivative $\partial f/\partial u$ denotes the derivative of the expression $f(t,u)$ with respect to the variable u, with t regarded as a constant. For now it suffices that functions f defined by 'formulas' (i.e., polynomials, sine, cosine, etc) satisfy the requirements of Theorem 1, except possibly in some special points.

Exercise 4.7 Consider the initial value problem

$$\frac{du}{dt} = 2\sqrt{u}, \qquad u(0) = 0.$$

Verify that this initial value problem has two solutions for $t > 0$, namely $u_1(t) = t^2$ and $u_2(t) = 0$. Does this result contradict Theorem 1?

4.3 Solution Methods

In this section the solutions to three different types of ordinary differential equations will be discussed. In section 4.3 separable first order differential equations will be introduced, in section 4.3 first order linear differential equations will be discussed, and in section 4.3 the solution method for linear, second order ordinary differential equations with constant coefficients will be presented.

Separable Equations

Consider a differential equation of the form

$$\frac{du}{dt} = g(t)h(u). \tag{4.13}$$

An implicit solution of equation (4.13) reads

$$F(u) = G(t) + C, \tag{4.14}$$

with C an integration constant, and $F(u)$ and $G(t)$ the antiderivatives of $1/h(u)$ and $g(t)$. To verify that $F(u)$ is indeed an implicit solution, differentiate equation (4.14) with respect to t. For the left hand side we have to use the chain rule,

$$\frac{dF(u(t))}{dt} = \frac{dF(u)}{du}\frac{du}{dt} = \frac{1}{h(u)}\frac{du}{dt},$$

and for the right hand side it immediately follows that

$$\frac{dG(t)}{dt} = g(t).$$

Combining these results, it immediately follows that equation (4.14) is the solution of equation (4.13).

Equation (4.13) is said to be *separable* because — upon formal multiplication of both sides by dt and by $1/h(u)$ — it takes the symbolic form

$$\frac{1}{h(u)}du = g(t)dt \tag{4.15}$$

in which the variables t and u (and their respective differentials dt and du) are separated on opposite sides of the equation. The process of rewriting equation (4.13) into equation (4.15) is called *separating the variables*.

Now a solution to equation (4.13) is obtained by integrating each side in equation (4.15) with respect to its "own" variable:

$$\int \frac{1}{h(u)}\, du = \int g(t)\, dt, \tag{4.16}$$

which immediately gives solution (4.14).

Example 4.2. (taken from [1]): Find the solution of the initial value problem

$$\frac{du}{dt} = \frac{u\cos(t)}{1+2u^2}, \qquad u(0) = 1. \tag{4.17}$$

Observe that $u = 0$ is a solution of this differential equation. To find other solutions, assume that $u \neq 0$ and write the differential equation in the form

$$\frac{1+2u^2}{u}du = cos(t)\, dt. \tag{4.18}$$

Integrating the left side with respect to u and the right side with respect to t, we obtain

$$\ln|u| + u^2 = \sin(t) + C. \tag{4.19}$$

To satisfy the initial condition we substitute $t = 0$ and $u = 1$ in equation 1(4.19); this gives $C = 1$. Hence the solution of the initial value problem (4.17) is given implicitly by

$$\ln|u| + u^2 = \sin(t) + 1. \tag{4.20}$$

Exercise 4.8 The implicit solution (4.20) is not readily solved for u as a function of t. To get some insight in the behavior of the solution show that

- no solution crosses the t-axis.
- for the initial value problem discussed above, the absolute value bars can be dropped in the solution (4.20).

Exercise 4.9 Solve the following problems, using separation of variables:

- $du/dt = -6ut$, with $u(0) = 7$.
- $du/dt = -6ut$, with $u(0) = -4$.
- Plot the population size $N(t)$ by solving equation (4.7), with $r = 0.06$, $K = 1/150$ and $N(0) = 20$. What does the solution curve look like for $N(0) = 300$?
- Consider Exercise 3b. At time $t = 0$, the water tank is full of water. How long will it take for all the water to drain from the tank?

Linear First Order Differential Equations

A differential equation of the form

$$\frac{du}{dt} + f(t)u = g(t) \tag{4.21}$$

is called a linear first order differential equation. If the function $g(t) \equiv 0$, the resulting homogeneous equation is a separable equation that can be solved using the method explained in section 4.3. The general solution to the homogeneous differential equation reads

$$u(t) = C\exp(-F(t)), \tag{4.22}$$

with $F(t)$ the antiderivative of $f(t)$ and C an arbitrary constant.

Exercise 4.10 Show that expression (4.22) is a solution to differential equation (4.21) with $g(t) \equiv 0$.

To solve equation (4.21) for arbitrary $g(t)$, a method called *variation of parameters* will be employed. In this approach, the constant C in equation (4.22) is assumed to be a function of t. A so–called *particular* solution $u_p(t)$, defined by

$$u_p(t) = C(t) \exp[-F(t)],$$

can now be found by substituting this expression in equation (4.21), resulting in

$$\frac{dC}{dt} = g(t) \exp[F(t)]. \tag{4.23}$$

From this, a particular solution $u_p(t)$ is obtained:

$$u_p(t) = \exp[-F(t)] \int g(t) \exp[F(t)]\, dt \tag{4.24}$$

Exercise 4.11 Derive expressions (4.23) and (4.24). Is there only one possible particular solution?

Since the problem is linear, the sum of a particular solution and an arbitrary constant times the homogeneous solution is still a solution of the equation. Hence, the general solution can be written as

$$u(t) = u_{\text{hom}}(t) + u_p(t) = B \exp[-F(t)] + \exp[-F(t)] \int_0^t g(t') \exp[F(t')]\, dt',$$

with the lower integration boundary $t' = 0$ and the upper boundary as $t' = t$. The unknown coefficient B follows from the initial condition.

Example 4.3. Find the solution of the initial value problem

$$\frac{du}{dt} - \cos(t)u = \exp[\sin(t)], \qquad u(\pi) = 0. \tag{4.25}$$

First, we use separation of variables to get the homogeneous solution. This solution reads

$$u_{\text{hom}}(t) = C \exp\left[\int \cos(t)\, dt\right] = C \exp[\sin(t)].$$

Next, we assume that C varies with t. Substituting the particular solution

$$u_p(t) = C(t) \exp[\sin(t)]$$

in equation (4.25) results in the following equation for C:

$$C'(t) = 1.$$

Solving this equation, the general solution can be found. This solution reads

$$u(t) = B \exp[\sin(t)] + t \exp[\sin(t)].$$

Using that at $t = \pi$ the solution $u(0) = 0$, it follows that $B = -\pi$.

Exercise 4.12 Solve the following problems:

- $du/dt = u + 2t\exp(2t)t$, with $u(0) = 1$.
- $t\,du/dt + 2u = t^2 - t + 1$, with $u(1) = \frac{1}{2}$, $\qquad t > 0$.
- Solve equation (4.1) with initial condition $v(0) = 70$.

Linear Second Order Differential Equations with Constant Coefficients

A linear second order differential equation is of the form

$$a\frac{d^2u}{dt^2} + b\frac{du}{dt} + cu = g(t), \tag{4.26}$$

with a, b and c constant coefficients, i.e., these coefficients do not depend on the independent variable t or the dependent variable u. If $g(t) \equiv 0$, the equation is homogeneous, otherwise Eqnequation (4.26) is inhomogeneous. Since equation (4.26) is linear, the general solution is again the sum of a particular and homogeneous solution (cf. Section 4.3). In the following, we will focus on obtaining the homogeneous solution, particular solutions can again be obtained by variation of parameters, or (if the forcing term $g(t)$ has a nice enough form) by guessing a trial solution.

To construct solutions to equation (4.26) with $a \neq 0$ and $g(t) \equiv 0$, we need to find *two* independent solutions. The key to finding these solutions is to guess a plausible form of these solutions. A possible solution would be that u' and u'' are constant multiples of u, which implies that $u' \sim u$ and $u'' \sim u$. We have already seen that such functions are exponentials, so we try solutions of the form $u(t) = \exp(rt)$. Substituting this in equation (4.26) results in

$$\big(ar^2 + br + c\big)\exp(rt) = 0,$$

which must be valid for all times t. We therefore have to require that

$$ar^2 + br + c = 0. \tag{4.27}$$

Equation (4.27) is called the characteristic equation, with solutions

$$r_1, r_2 = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}.$$

Now we can distinguish three cases:

1. The characteristic equation has two real and distinct roots, resulting in the following general solution:
$$u(t) = c_1\exp(r_1t) + c_2\exp(r_2t).$$

2. The characteristic equation has two real but equal roots. The general solution reads
$$u(t) = c_1\exp(r_1t) + c_2\,t\exp(r_1t).$$

3. The characteristic equation has complex conjugate roots. After rewriting, the following expression for $u(t)$ can be found:

$$u(t) = exp(pt)\left[c_1 \cos(qt) + c_2 \sin(qt)\right],$$

with p the real part of the roots and q the imaginary part.

Again the unknown coefficients c_1 and c_2 follow from the initial conditions.

Exercise 4.13 Solve the following problems:

- (example of case 1)

$$3d^2u/dt^2 + 7du/dt + 2u = 0. \qquad u(0) = 1 \quad \text{and} \quad u'(0) = 1$$

.

- (example of case 2)

$$4d^2u/dt^2 + 12du/dt + 9u = 0. \qquad u(0) = 4 \quad \text{and} \quad u'(0) = -3.$$

Show for this example explicitly that $t\exp(r_1 t)$ is indeed a solution satisfying equation (4.26).

- (example of case 3)

$$d^2u/dt^2 + 4u = 0. \qquad u(0) = 1 \quad \text{and} \quad u'(0) = -0.$$

Show for this example explicitly that $t\exp(r_1 t)$ is indeed a solution satisfying equation (4.26).Clearly indicate how you get the solution, given in the list above under item number 3, with p and q real, from the general solution

$$u(t) = c_1 \exp(r_1 t) + c_2 \exp(r_2 t),$$

with r_1 and r_2 the complex solutions of the characteristic equation. Verify the expressions for p and q used in item 3 above.

Literature

There are many books on Differential Equations. Below, there are some useful references:

[1] Boyce, W.E. and DiPrima, R.C. *Elementary Differential Equations and Boundary Value Problems*, John Wiley and Sons, Inc, New York.

[2] Van Horssen, W.T. *Differentiaalvergelijkingen*, Epsilon 27, Epsilon Uitgaven, Utrecht.

[3] Duistermaat, J.J. and Eckhaus, W. *Analyse van Gewone Differentiaalvergelijkingen.*, Epsilon 33, Epsilon Uitgaven, Utrecht.

5

Counting

Author: K.P. Hart

Introduction

Everybody knows what it is to count. How many students in the lecture room? Let's count: one, two, three, ..., fifty-one, ...

What is happening here mathematically is the construction of a bijection[1] between two sets: the set that we are counting, and a standard set, one of the form $\{1, 2, 3, \ldots, n\}$. This is normally not a very exciting task and, if the set is rather large, there is a non-zero chance of making errors.

If the set comes with a bit of structure then you can count it more efficiently by, cleverly, dividing it into subsets, count each of those subsets and add the results. In a lecture room it is usually better to count the number of students in each row and then add those numbers.

In this lecture we will see some methods and formulas that will help us count a variety of objects in a systematic way. We will derive formulas for, for example

1. the number of maps from $\{1, 2, \ldots, k\}$ to $\{1, 2, \ldots, n\}$
2. the number of injective maps from $\{1, 2, \ldots, k\}$ to $\{1, 2, \ldots, n\}$
3. the number of surjective from $\{1, 2, \ldots, k\}$ to $\{1, 2, \ldots, n\}$
4. the number of bijective maps from $\{1, 2, \ldots, k\}$ to $\{1, 2, \ldots, n\}$
5. the number of subsets of $\{1, 2, \ldots, n\}$ that have k elements

Many practical/mathematical problems can be reduced to one of the problems given above or solved by the methods that we will develop

1. How many ways are there to fill in a lottery form?
2. How many different hands can you meet in a card game?
3. In how many ways can we divide a mentor group into two or three groups for the homework?

[1] See the course TW1010, [2]*Definition 2.3.6

4. At the drawing-of-names for Saint Nicholas: what is the probabilty that nobody draws their own name?

Notation

Before we start: we need some notation to keep our formulations short and sweet.

We will denote the set $\{1, 2, \ldots, n\}$ by $\mathbf{n}$.

The number of elements of a set X is written as $|X|$. The formal definition is given by: $|X| = n$ means there is a bijection $f : X \to \mathbf{n}$.

That this is a sound definition follows from the following exercise.

Exercise 5.1 Prove: if m and n are natural numbers and there is a bijection $f : \mathbf{m} \to \mathbf{n}$ then $m = n$. *Hint*: induction on m.

If X is a set then $\mathcal{P}(X)$ denotes the family of *all* subsets of X and we use $[X]^k$ denotes the family of subsets that have exactly k elements.

Exercise 5.2 Write down all elements of $[\mathbf{5}]^0$, $[\mathbf{5}]^1$, and $[\mathbf{5}]^2$, respectively.

Also, quite often it is easier to count a set, X say, by making a bijection between X and some set, Y, that we counted before.

5.1 Boxes and balls

In this section we describe a general model for counting things. That model is "boxes and balls". It turns out that very many problems can be (re)formulated as a problem about dividing balls over boxes.

For example: how many solutions does the equation $k + l + m = 10$ have if we allow only natural numbers in our solutions? Some people think that 0 is a natural number and some people think it is not so we actually have two questions: one where we allow zeroes and one where we do not. To formulate this into terms of balls and boxes take three boxes, labeled k, l, and m. Also take ten balls that all look the same and divide them over the boxes. Every distribution gives a solution and every solution determines a distribution. For the people who don't like 0 as a natural number we add the requirement that there should be at least one ball in each box.

The problem of dividing the students in a mentor group into three subgroups looks like this problem but not quite: the students are all different! To translate this problems we use numbered balls and divide these over the three boxes. To ensure some balance between the groups you can demand that each box gets a certain minimum number of balls.

We will start with n boxes, these will be numbered. Next we will take k balls and divide these over the boxes. We consider the following situations/demands:

1. the balls are indistinguishable (same size, colour, ...),

2. the balls are distinguishable, with numbers say.

and

1. at most one ball in each box,

2. arbitrary distributions,

3. at least one ball in each box.

These six possibilities cover lots of situations, as we shall see later.

5.2 At most one ball in each Box

This is the easiest case and it gives rise to some formulas that will be of use in the other situations as well.

To begin, if $k > n$ then the number of good distributions is zero, so we shall assume that $k \leq n$.

Distinguishable balls, factorials

To illustrate that many situations can be translated to "boxes and balls" we show how dealing cards at bridge can be seen as putting balls into boxes.

To simulate dealing we put 52 boxes on the floor and on each box we paste one card from our deck of 52 cards. You are given 13 balls (numbered).

Instead of the dealer handing you a three-of-hearts as your first card he tells you to put ball number 1 into the box with the three-of-hearts on it, and so on. Since you can not get the same card twice you can put at most one ball into any box. This is not a very useful way of dealing cards but it illustrates how one may translate one situation into an other.

Whether you want to think of getting 13 out of 52 cards or of putting 13 balls into 52 boxes, counting the number of hands/distributions proceeds as follows.

You have 52 possibilities for the first card/ball, 51 for the second, ..., and 40 for the thirteenth. That means that there are $52 \times 51 \times \cdots \times 40$ possible sequences of cards that you can see during the dealing process or that there are $52 \times 51 \times \cdots \times 40$ ways of distributing thirteen distinguishable balls over 52 boxes. That number is fairly large:

$$52 \times 51 \times \cdots \times 40 = 3954242643911239680000$$

(that is 22 digits).

In general there are

$$n \times (n-1) \times \cdots \times (n-k+1) \qquad (*)$$

ways to distribute k distinguishable balls over n boxes in such a way that every box contains at most one ball.

Exercise 5.3 Verify formula $(*)$, why does the product end at $n-k+1$?

Notice that every distribution codes an injective map[2] from $\mathbf{k}$ to $\mathbf{n}$ — let $f(i) = j$ mean that ball i goes into box j — and vice versa. So we have counted the number of injective maps from $\mathbf{k}$ to $\mathbf{n}$ as well.

In the special case that $k = n$ we have the number of *bijective* maps from $\mathbf{n}$ to itself (the *permutations* of $\mathbf{n}$); that number is

$$n \times (n-1) \times \cdots \times 1$$

[2]See [2]*Definition 2.3.5

That product is abbreviated as $n!$ (pronounced: n-factorial). Note that the product in $(*)$ kan be written as

$$\frac{n!}{(n-k)!}$$

These formulas can be used to calculate probabilities.

Exercise 5.4 Consider bridge hands. What is the probability of a hand with only red cards (♡ and ♢)? What is the probabilty of getting all cards in one suit (♡, or ♢, or ♣, or ♠)?

Exercise 5.5 What is the probability of having all six correct in a lotto drawing?

Exercise 5.6 At a (traditional) speed date event there are ten men and ten women. The organisers decide there will be one-minute sessions where each man is coupled with a woman and vice versa, *and* that all possible couplings of the ten men and ten women should occur at least once. How long will that event last?

Indistinguishable balls, subsets, binomial coefficients

When playing bridge you are not really interested in the order in which you get your cards; you will arrange them in some useful order yourself. It is the *set* of cards in your hand that is important and we have just seen that that set can be ordered in 13! different ways. The number of *hands* in bridge therefore is not equal to $\frac{52!}{39!}$ but to

$$\frac{52!}{13! \times 39!} = 635013559600$$

(still twelve digits).

The translation to "balls and boxes" comes down to erasing the numbers on the balls, thus making them indistinguishable.

The argument about bridge hands shows what the relation is between the two ways of distributing k balls over n boxes with at most one ball per box: divide the number for distinguishable balls by $k!$ to obtain the number for indistinguishable balls.

Exercise 5.7 What is the number of ways to fill out a lotto card?

The general formula for the number of ways to distribute k balls over n boxes with at most one ball per box is the product $(*)$, divided by $k!$:

$$\frac{n!}{k!(n-k)!} \qquad (**)$$

this last quotient occurs in very many formulas and it has gotten its own abbreviation

$$\binom{n}{k}$$

This is usually pronounced as 'n-choose-k' or 'n-above-k' and it is called a *binomial coefficient*.

An important observation is this: because the balls are indistinguishable a distribution of balls corresponds to a choice of k out of the n boxes. This means that $\binom{n}{k}$ is exactly the number of subsets of $\mathbf{n}$ that have k elements; in formula

$$\binom{n}{k} = \left|[\mathbf{n}]^k\right|$$

5.3 Binomial Coefficients

Because the binomial coefficients occur in many places, in particular in the remaining four cases that we still have to consider, we take some time to study them more closely. We begin by calculating a few easy values of $\binom{n}{k}$.

Exercise 5.8 Verify that $\binom{n}{0} = \binom{n}{n} = 1$ and $\binom{n}{1} = \binom{n}{n-1} = n$. Do this in two ways: by working with the formula and by actually counting the representing sets: $[\mathbf{n}]^0$ and $[\mathbf{n}]^n$, and $[\mathbf{n}]^1$ and $[\mathbf{n}]^{n-1}$.

Theorem 5.1. *If $0 \leq k \leq n$ then*

$$\binom{n}{k} = \binom{n}{n-k}$$

Exercise 5.9 Prove this theorem, in two ways: by working with the formula and by making a bijection between the families $[\mathbf{n}]^k$ and $[\mathbf{n}]^{n-k}$.

An important equality is the following

Theorem 5.2. *If $n \geq k \geq 1$ then*

$$\binom{n+1}{k} = \binom{n}{k-1} + \binom{n}{k}$$

Exercise 5.10 Prove this theorem in two ways

(a) By adding the fractions

$$\frac{n!}{(k-1)!(n-k+1)!} \quad \text{and} \quad \frac{n!}{k!(n-k)!}$$

(b) By dividing $[\mathbf{n+1}]^k$ into two subfamilies: the sets that contain $n+1$ and the ones that do not contain $n+1$.

This theorem enables us to put the binomial coefficients in a nice table, known as *Pascal's Triangle*, see Figure 5.1. Every number in the table is the sum of the two numbers right above it.

Exercise 5.11 Divide $\mathcal{P}(\mathbf{n})$ into two families: the subsets with an even number of elements, and the subsets with an odd number of elements. Make a bijection between these two families. *Hint*: Define for $A \subseteq \mathbf{n}$ a set A' as follows: $A' = A \setminus \{1\}$ if $1 \in A$ and $A' = A \cup \{1\}$ if $1 \notin A$.

Exercise 5.12 Prove, in two ways: $\binom{n+1}{3} = \sum_{k=0}^{n} \binom{k}{2}$.

We can use Exercise 8 and Theorem 5.2 to prove formula $(**)$ in an alternative way. For the moment we write the quotient in $(**)$ as $B(n,k)$ and we write $\left|[\mathbf{n}]^k\right|$ as $C(n,k)$. Then you can interpret the exercise and the proof of the theorem as proofs of

- $B(n,0) = B(n,n) = 1$ and $B(n+1,k) = B(n,k-1) + B(n,k)$, as well as

$$\begin{array}{ccccccccccc}
 & & & & \binom{0}{0} & & & & & & \\
 & & & \binom{1}{0} & & \binom{1}{1} & & & & & \\
 & & \binom{2}{0} & & \binom{2}{1} & & \binom{2}{2} & & & & \\
 & & & & \cdots & & & & & & \\
\binom{n}{0} & & \binom{n}{1} & & \binom{n}{k-1} & \binom{n}{k} & & & \binom{n}{n} & & \\
\binom{n+1}{0} & \binom{n+1}{1} & & \binom{n+1}{2} & & \binom{n+1}{k} & & \binom{n+1}{n} & & \binom{n+1}{n+1} &
\end{array}$$

Figure 5.1: Pascal's Triangle

- $C(n,0) = C(n,n) = 1$ and $C(n+1,k) = C(n,k-1) + C(n,k)$.

Exercise 5.13 Prove, using the above, that $B(n,k) = C(n,k)$ for all n and k. *Hint*: Induction on n.

The binomial coefficients also occur in a formula for the powers of $a+b$.

Theorem 5.3. *For every n we have*

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^{n-k} b^k$$

You can prove this theorem in a various ways. You can expand $(a+b)^n$ completely, without simplifying; you will then see that you will get, for every k, exactly as many products with k factors b as there are subsets of $\mathbf{n}$ that have k elements. Let us look at $(a+b)^2$ and $(a+b)^3$:

$$(a+b)(a+b) = a(a+b) + b(a+b) = aa + ab + ba + bb$$

and

$$(a+b)^3 = a(a+b)^2 + b(a+b)^2 = aaa + aab + aba + abb + baa + bab + bba + bbb$$

In these special cases you can see by inspection that every subset of **2** and **3** is filled with bs exactly once and that after simplification we get the formula in the theorem.

We base our proof on Exercise 8 and Theorem 5.2; we write

$$(a+b)^n = \sum_{k=0}^{n} N(n,k) a^{n-k} b^k$$

and we prove that the numbers $N(n,k)$ have exactly the same properties as $B(n,k)$ and $C(n,k)$.

Start by writing $(a+b)^{n+1}$ as $(a+b)(a+b)^n$:

$$(a+b) \sum_{k=0}^{n} N(n,k) a^{n-k} b^k$$

and remove the parentheses to get

$$\sum_{k=0}^{n} N(n,k)a^{n+1-k}b^k + \sum_{k=0}^{n} N(n,k)a^{n-k}b^{k+1}.$$

Now we rewrite the second sum:

$$\begin{aligned}\sum_{k=0}^{n} N(n,k)a^{n-k}b^{k+1} &= N(n,0)a^n b + N(n,1)a^{n-1}b^2 + \cdots + N(n,n)a^0 b^{n+1} \\ &= \sum_{k=1}^{n+1} N(n,k-1)a^{n+1-k}b^k\end{aligned}$$

Now we can combine both sums into one:

$$N(n,0)a^{n+1} + \sum_{k=1}^{n}\big(N(n,k) + N(n,k-1)\big)a^{n+1-k}b^k + N(n,n)b^{n+1}$$

From this we read off the following:

- $N(n+1,0) = N(n,0)$;
- $N(n+1,k) = N(n,k) + N(n,k-1)$; en
- $N(n+1,n+1) = N(n,n)$.

Now observe that $N(1,0) = N(1,1) = 1$ and, hence by induction it follows that $N(n,0) = N(n,n) = 1$ for all n. Now we use the second formula to prove that $N(n,k) = \binom{n}{k}$ for all n and k.

Using this theorem we can create all kinds of formulas, by making clever choices of a and b.

Exercise 5.14 Show that $\sum_{k=0}^{n}\binom{n}{k} = 2^n$. (NB we will see this later by counting subsets, see Exercise 28.)

Exercise 5.15 Show that $\sum_{k=0}^{n}(-1)^k\binom{n}{k} = 0$. (NB this we already knew, see Exercise 11.)

Exercise 5.16 Show that $\sum_{k=0}^{n}\binom{n}{k}^2 = \binom{2n}{n}$.

Exercise 5.17 Show that $\binom{n+m}{k} = \sum_{i=0}^{k}\binom{n}{i}\binom{m}{k-i}$.

Exercise 5.18 Prove: if $l < k \leq \frac{n}{2}$ then $\binom{n}{l} < \binom{n}{k}$.

Exercise 5.19 Prove that $\binom{2n}{n} > \frac{4^n}{2n+1}$. *Hint*: What is the arithmetic mean of $\binom{2n}{0}$, $\binom{2n}{1}$, …, $\binom{2n}{n}$, …, $\binom{2n}{2n}$?

5.4 Indistinguishable balls, arbitrary, at least one

We return to our balls and boxes and count the four remaing kinds of distributions. It turns out that the two cases for indistinguishable balls have very similar answers.

Arbitrary

First we do arbitrary distributions and we reduce it to the problem of counting subsets. We do that as follows: put the boxes in a row, and glue (or staple) the boxes together. Then when we put balls in (some of) the boxes we actually create a sequence of two kinds of objects: balls and 'walls' (pairs of sides glued together'. Figure 5.2 shows the

Figure 5.2: Nine balls in five boxes

situation with nine balls and five boxes: the four lines show the sides *between* the boxes. The dots represent the balls: one in box 1, two in box 2, zero in box 3, and three in each of boxes 4 and 5.

So, if we have nine balls and five boxes then every distribution creates a sequence of thirteen symbols and every such sequence creates a distribution. As soon as soon as we have placed the four lines in four of the thirteen spots the sequence, and hence the distribution of the balls over the boxes, is fixed. This means that we have $\binom{13}{4} = \binom{13}{9}$ possible distributions of nine indistinguishable balls over five boxes.

The general formula is now easy to see: we must put k dots and $n-1$ lines in a sequence. So: choose $n-1$ positions out of $k+n-1$ and put the lines there, that makes

$$\binom{k+n-1}{n-1} \quad \text{or} \quad \binom{k+n-1}{k}$$

possible distributions.

Exercise 5.20 How many solutions does $x+y+z=10$ have (natural numbers, zero included)?

Exercise 5.21 You have nineteen Euros. In how many ways can you divide this money among five persons, if everybody should get a whole number of Euros?

At least one in each box

This situation appears to be new but it almost the same as the 'arbitrary' one. We need at least n balls of course, so we assume $k \geq n$. All we do is first put one ball in each box and then distribute the remaining $n-k$ balls arbitrarily. So the answer is

$$\binom{k-n+n-1}{n-1} = \binom{k-1}{n-1}$$

possibilities.

Exercise 5.22 Find an other way of seeing this. *Hint*: How many lines are allowed between two balls?

Exercise 5.23 How many solutions does $x+y+z=10$ have (natural numbers, zero excluded)?

Exercise 5.24 You have nineteen Euros. In how many ways can you divide this money among five persons, if everybody should get a whole and positive number of Euros?

5.5 Distinguishable balls, arbitray, maps, powers

We now look at distinguishable balls, numbered 1 through k, say.

As on page 61 we can code our distributions by maps from $\mathbf{k}$ to $\mathbf{n}$. Define $f(i) = j$ if ball i goes into box j. So we count the maps from $\mathbf{k}$ to $\mathbf{n}$ as well.

Definition 2.3.1 in Lay's book [2] states that a map/function from $\mathbf{k}$ to $\mathbf{n}$ is a subset f of the product $\mathbf{k} \times \mathbf{n}$ with the property that for every $i \in \mathbf{k}$ there is exactly one $j \in \mathbf{n}$ such that $(i, j) \in f$, and that j is written $f(i)$.

Counting these subsets is not difficult; you can choose for every $i \in \mathbf{k}$ the value $f(i)$ independently from the other values. In that way you make n^k maps.

Exercise 5.25 How many differents strings of one hundred beads can you create if you have sufficiently many beads of four colours?

The method also works if you have a different number of possible values for $f(i)$.

Exercise 5.26 Take a sequence of natural numbers n_1, n_2, ..., n_k. Show that the number of maps from $\mathbf{k}$ to $\mathbb{N}$ subject to $f(i) \in \mathbf{n}_i$ for all i is equal to $n_1 \times n_2 \times \cdots \times n_k$.

Exercise 5.27 How many possible (Dutch) licence plates (digit - three letters - two digits) are possible?

Exercise 5.28 How many subsets does The set $\mathbf{k}$ have? *Hint*: Every subset determins a map from $\mathbf{k}$ to $\mathbf{2}$. (We already knew this, see Exercise 14.)

This kind of counting problems occurs often in Probabilty theory. Flipping coins and throwing dice will produce functions with values in $\mathbf{2}$ and $\mathbf{6}$ respectively.

Exercise 5.29 Flip a coin twenty times; what is the number of outcomes? What is the number when you throw a die twenty times?

Exercise 5.30 You have nineteen Euro coins, one for each member of the Eurozone. In how many ways can you divide these coins among five persons?

5.6 Distinguishable balls, at least one per box, surjections

We now come to the hardest of our six counting problems: distribute k distinguishable balls over n boxes such that every box gets at least one ball.

As above this amounts to counting certain maps from $\mathbf{k}$ to $\mathbf{n}$, in this case the *surjective* maps ("every box gets a ball" is saying that the map that codes the distribution is surjective: for every j there is an i such that $f(i) = j$).

We use the symbol $\left|{k \atop n}\right|$ to denote the number of surjections from $\mathbf{k}$ to $\mathbf{n}$. In some cases we can write down what $\left|{k \atop n}\right|$ without much effort. Of course $\left|{k \atop n}\right| = 0$ if $n = 0$ of $k < n$.

Exercise 5.31 Verify that $\left|{k \atop 1}\right| = 1$ en $\left|{k \atop k}\right| = k!$

Exercise 5.32 Show that $\left|{k \atop 2}\right| = 2^k - 2$. *Hint*: How many non-surjective maps are there?

Exercise 5.33 Determine $\left|{k \atop 3}\right|$ and $\left|{k \atop 4}\right|$ for a few values of k.

We are going to determine $\left|{k \atop n}\right|$ by using the hint in Exercise 32: we count the maps that are *not* surjective and subtract that number from n^k, the number of all maps.

We divide the non-surjective maps into groups: for every $j \in \mathbf{n}$ let $E(j) = \{f : j \notin f[\mathbf{k}]\}$. All we have to do is count the union $\bigcup_{j=1}^{k} E(j)$.

For example, if $n = 2$, as in Exercise 32 the groups $E(1)$ and $E(2)$ each have just one element: the constant map with value 2 and 1 respectively. We see that $E(1) \cup E(2)$ has exactly two elements and so, indeed, $\left|{k \atop 2}\right| = 2^k - 2$.

We turn to the case $n = 3$. We have three sets $E(1)$, $E(2)$ and $E(3)$ and we want to count their union. We do this a bit naïvely: add the individual numbers of elements:

$$\big|E(1)\big| + \big|E(2)\big| + \big|E(3)\big|$$

This number is too large because we have counted the elements of the intersections $E(1) \cap E(2)$, $E(1) \cap E(3)$ and $E(2) \cap E(3)$ twice. So we subtract those numbers:

$$\big|E(1)\big| + \big|E(2)\big| + \big|E(3)\big| - \big|E(1) \cap E(2)\big| - \big|E(1) \cap E(3)\big| - \big|E(2) \cap E(3)\big|$$

However, now we must look at the elements of $E(1) \cap E(2) \cap E(3)$; those were counted three times in the first sum and three times in the second sum, in total they were counted zero times. To get the right total we must add their number to the sum:

$$\begin{aligned}\big|E(1)\big| + \big|E(2)\big| + \big|E(3)\big| \\ - \big|E(1) \cap E(2)\big| - \big|E(1) \cap E(3)\big| - \big|E(2) \cap E(3)\big| \\ + \big|E(1) \cap E(2) \cap E(3)\big|\end{aligned}$$

Draw a picture (Venn diagram) of three sets to see how this adding and subtracting works.

Now we calculate the total sum. To begin: all the $E(j)$ have the same number of elements, to wit 2^k. Indeed: $E(j)$ consists of all maps from $\mathbf{k}$ to the two-point set $\mathbf{3} \setminus \{j\}$. Likewise all intersections $E(i) \cap E(j)$ have the same size: each has $1^k = 1$ element, and the total intersection $E(1) \cap E(2) \cap E(3)$ is empty. We find that

$$\big|E(1) \cup E(2) \cup E(3)\big| = 3 \cdot 2^k - 3 \cdot 1^k + 0^k$$

We found this equality in a quite straightforward manner and it is also straightforward to verify that we did not miss an element of the union nor that we counted one more than once.

In the general situation we have to a bit more careful and really check that our formula is correct. To show how that works we do the general verification for the special case $n = 3$. We show that every element of $E(1) \cup E(2) \cup E(3)$ is really counted exactly once in the right-hand side. Take a non-surjective $f : \mathbf{k} \to \mathbf{3}$ and look at the complement of $f[\mathbf{k}]$. If the complement consists of one point, 2 say, then f is counted only in $\big|E(2)\big|$, it contributes zero to the other terms. If it consists of two points, say 1 and 2, then f contributes 1 to $\big|E(1)\big|$ and $\big|E(2)\big|$ and to $-\big|E(1) \cap E(2)\big|$. So the contribution to the total sum is $1 + 1 - 1 = 1$.

Now we subtract from 3^k and obtain

$$\left|{k \atop 3}\right| = 3^k - 3 \cdot 2^k + 3 \cdot 1^k - 0^k$$

We have left in the factors 1^k en 0^k to foreshadow the structure of the general formula.

If you look closely you will see that we can also write

$$\left|{k \atop 3}\right| = \binom{3}{0} \cdot 3^k - \binom{3}{1} \cdot 2^k + \binom{3}{2} \cdot 1^k - \binom{3}{3} \cdot 0^k$$

or, in one summation:

$$\left|{k \atop 3}\right| = \sum_{i=0}^{3} (-1)^i \binom{3}{i} (3-i)^k$$

In this form the formula holds in general.

Theorem 5.4. *If $0 < n \leq k$ then*

$$\left|{k \atop n}\right| = \sum_{i=0}^{n} (-1)^i \binom{n}{i} (n-i)^k$$

Proof. We count the number of non-surjective maps, in other words we determine

$$\big|E(1) \cup E(2) \cup \cdots \cup E(k)\big|$$

We start with $\big|E(1)\big| + \big|E(2)\big| + \cdots + \big|E(k)\big|$; each of the $E(j)$ has $(n-1)^k$ elements; so the sum is equal to $n(n-1)^k$.

Because of counting double we must subtract $\big|E(j_1) \cap E(j_2)\big|$, for every pair $\{j_1, j_2\}$. Every intersection has $(n-2)^k$ elements and we have $\binom{n}{2}$ pairs so after subtracting we have $\binom{n}{1}(n-1)^k - \binom{n}{2}(n-2)^k$.

As above we have counted the members of the intersections $E(j_1) \cap E(j_2) \cap E(j_3)$ zero times (three times positive, three times negative), so we add their numbers to our total; there are $\binom{n}{3}$ intersections and these have $(n-3)^k$ elements each.

We have reached $\binom{n}{1}(n-1)^k - \binom{n}{2}(n-2)^k + \binom{n}{3}(n-3)^k$.

At stage i we deal with maps that avoid at least i values; these belong to intersections of the form $E(j_1) \cap \cdots \cap E(j_i)$. There are $\binom{n}{i}$ such intersections and each has $(n-i)^k$ elements. When i is odd we add, when i is even we subtract; the contribution at this stage is

$$(-1)^{i-1} \binom{n}{i} (n-i)^k$$

At the end we obtain the following formula for the number of non-surjective maps

$$\sum_{i=1}^{n} (-1)^{i-1} \binom{n}{i} (n-i)^k \qquad (\dagger)$$

and hence the expression

$$n^k - \sum_{i=1}^{n} (-1)^{i-1} \binom{n}{i} (n-i)^k = \sum_{i=0}^{n} (-1)^i \binom{n}{i} (n-i)^k$$

for the number of surjective maps.

Although, at least intuitively, it seems clear that formula (†) does indeed give the number of non-surjective maps we verify that this is indeed the case by showing that every non-surjective map f contributes exactly 1 to the sum.

Suppose that the complement, J, of $f[\mathbf{k}]$ has exactly j elements. Then f contributes only to the first j terms of (†). Its contribution to the ith term is $(-1)^{i-1}\binom{j}{i}$: it gives 1 for every subset of J that has i elements.

The total contribution of f is therefore

$$\sum_{i=1}^{j}(-1)^{i-1}\binom{j}{i}$$

But now apply Exercise 15:

$$\sum_{i=0}^{j}(-1)^{i}\binom{j}{i}=0$$

and hence

$$1=\sum_{i=1}^{j}(-1)^{i-1}\binom{j}{i}$$

The total contribution of f to (†) is indeed equal to 1. □

Stirling numbers

A surjective map $f:\mathbf{k}\to\mathbf{n}$ determines a partition of $\mathbf{k}$ into n nonempty subsets: for $i\in\mathbf{n}$ we let $A_i=\{j : f(j)=i\}$. This *partition* does not change if we permute the numbers in $\mathbf{n}$.

So, every partition corresponds to $n!$ surjective maps. We conclude that the number of ways to partition $\mathbf{k}$ into n nonempty subsets is equal to

$$\frac{1}{n!}\left|{k \atop n}\right|$$

That number is denoted $\left\{{k \atop n}\right\}$; the nnumbers $\left\{{k \atop n}\right\}$ are called *Stirling numbers of the second kind*.

Exercise 5.34 In how many ways can we divide a mentor group of ten students into three nonempty groups?

Exercise 5.35 In how many ways can we divide a mentor group of ten students into three groups in a reasonably balanced way? We take 'reasonably balanced' to mean a three-three-four division.

Exercise 5.36 You have nineteen Euro coins, one for each member of the Eurozone. In how many ways can you divide these coins among five persons, if everybody should get at least one coin?

5.7 The Inclusion-Exclusion Principle

The proof of Theorem 5.4 contains an important principle that we will discuss now.

Take a n sets A_1, A_2, ..., A_n. We want to determine $|A_1 \cup A_2 \cup \cdots \cup A_n|$ and we do that using the method from the above proof.

For this we need to introduce a few abbreviations:

1. for a subset I of $\mathbf{n}$ we write $A(I) = \bigcap_{i\in A} A_i$;
2. for $k \leq n$ we write $T(k) = \sum\{|A(I)| : I \in [\mathbf{n}]^k\}$

So, for example: $T(1) = \sum_{i=1}^{n}|A_i|$ and $T(2) = \sum_{1\leq i<j\leq n}|A_i \cap A_j|$.

Theorem 5.5. *We have the following equality*

$$|A_1 \cup A_2 \cup \cdots \cup A_n| = \sum_{k=1}^{n}(-1)^{k-1}T(k) \qquad (\ddagger)$$

This theorem is called the *Principle of Inclusion-Exclusion* or *Inclusion-Exclusion Principle*.

It is often convenient to define $A(\emptyset) = A_1 \cup A_2 \cup \cdots \cup A_n$ and then we can rewrite formula (‡) as follows:

$$\sum_{k=0}^{n}(-1)^{k}T(k) = 0 \qquad (\dagger\dagger)$$

The proof is just as in the proof of Theorem 5.4.

Proof.[Proof of Theorem 5.5] Every element of the union contributes 1 to the number of elements of the union. Every element of the union also contributes 1 to the right-hand side of (‡).

To prove that last sentence we assume that x belongs to exactly l of the sets A_i, say A_{i_1}, ..., A_{i_l}. We determine how much x contributes to every $T(k)$.

The contribution to $T(1)$ is l: we get 1 for every i_j. The contribution to $T(2)$ is $\binom{l}{2}$: we get 1 for every pair $\{i, j\}$ for which $x \in A_i$ and $x \in A_j$. Likewise the contribution to $T(3)$ is equal to $\binom{l}{3}$ and, in general, the contribution to $T(k)$ is equal to $\binom{k}{l}$; for $k > l$ the contribution is of course 0.

So, the contribution of x to the sum $\sum_{k=1}^{n}(-1)^{k-1}T(k)$ is equal to $\sum_{k=1}^{l}(-1)^{k-1}\binom{l}{k}$. We apply Exercise 15: $\sum_{k=0}^{l}(-1)^{k}\binom{l}{k} = 0$; bring all terms, except the first to the right-hand side, we get

$$1 = \sum_{k=1}^{l}(-1)^{k-1}\binom{l}{k}$$

So the element x is counted exactly once on the right-hand side of (‡). □

A typical application of the Inclusion-Exclusion Principle is the following

Exercise 5.37 How many numbers from **10000** are divisible by 3, by 5, but not by 7, and also not by 11?

Occasionally you can use 5.5 to check data.

Exercise 5.38 There are 35 math students who have made a choice of elective courses: 18 students want to take Logic (L), 23 want to do Numerical Mathematics (N), 21 choose Set Theory (V), and 17 take Real Analysis (R). In addition we know that 9 students want to do L and N, 7 L and V, 6 choose L and R, 12 take N and V, 9 follow N and R, and 12 opt for V and R. There are also students that choose three courses: 4 do LNV, 3 go for LNR, 5 take LVR, and 7 pick NVR; there are three students who want to take all courses. Is the administration in order?

Saint Nicholas

A well-known application of the Inclusion-Exclusion Principle is the following: count the number of permutations of $\mathbf{n}$ without fixed points.

Exercise 5.39 Solve this problem.

Hint: count the permutations that do have a fixed point, take $A(i) = \{f : f(i) = i\}$ for $i \in \mathbf{n}$.

Exercise 5.40 As an application of this: calculate the probability that at Saint Nicholas when a group of ten friends draws lots with each other's names on them nobody draws themselves.

Speed-dating

Consider a speed-date session with a men and b women. At the end everybody writes the name of the person from the other group that they like best. If two people choose each other then we have a match.

The question is how many possibilities there are without matches.

Exercise 5.41 Investigate this for a few small values of a and b: $a = b = 2$, $a = b = 3$, $a = 2$ and $b = 3$, $a = 3$ and $b = 4$, ...

Exercise 5.42 Find a formula for the number of possibilities without matches for arbitrary a and b.

Exercise 5.43 In case $a = b$ determine the probability that no matches will occur; what is the limit of this probability as a goes to infinity?

Exercise 5.44 Adapt your answer to the previous exercise to give a formula for the number of possibilities with k matches ($k \leq \min\{a, b\}$ of course).

Euler's totient function

Euler's totient function, also called Euler's φ-function, is defined as

$$\varphi(n) = \{i \in \mathbf{n} : \gcd(i, n) = 1\}$$

Here $\gcd(i, n)$ denote the greatest common divisor of i and n.

For some numbers the value of φ is easy to find:

Exercise 5.45 Verify that $\varphi(p) = p - 1$ if p is a prime number.

Exercise 5.46 Let $n \in \mathbb{N}$ be arbitrary and let $p_1, \ldots, p_m$ be its prime divisors. Prove

$$\varphi(n) = n \times \prod_{i=1}^{m} \left(1 - \frac{1}{p_i}\right)$$

You will encounter the function φ again in the course *Algebra 1* and you may like to explore the function a bit more.

Exercise 5.47 Let p be a prime number and $i \geq 1$. Calculate $\varphi(p^i)$ directly.

Exercise 5.48 Let m and n be two natural numbers such that $\gcd(m, n) = 1$. Prove that $\varphi(m \cdot n) = \varphi(m) \cdot \varphi(n)$.

5.8 More problems

Exercise 5.49 We still have n boxes and k balls. Write $k = k_1 + k_2 + \cdots + k_n$, with each k_i a nonnegative integer.

(a) In how many ways can we distribute k indistinguishable balls over the boxes in such a way that box i receives k_i balls?

(b) In how many ways can we do the same with distinguishable balls?

Exercise 5.50 There are twelve people and two round tables that each sit six people. Investigate how many different table settings you can have. Note: this will depend on your definition of 'different'; try to find as many possibilities as you can.

The following is problem 4 from the International Mathematical Olympiad of 2011 held in Amsterdam.

Exercise 5.51 Let $n > 0$ be an integer. We are given a balance and n weights of weight $2^0, 2^1, \ldots, 2^{n-1}$. We are to place each of the n weights on the balance, one after another, in such a way that the right pan is never heavier than the left pan. At each step we choose one of the weights that has not yet been placed on the balance, and place it on either the left pan or the right pan, until all of the weights have been placed. Determine the number of ways in which this can be done.

Exercise 5.52 Let n be a natural number. What is the probability that $\gcd(i, j) = 1$ when i and j are chosen at random from $\mathbf{n}$?

Exercise 5.53 At a (traditional) party there are ten married couples. How many dances can one have if men dance with women but nobody ever dances with their partner?

Exercise 5.54 At a (non-traditional) party there are ten married couples. How many dances can one have if everyone can dance with everyone *except* with their partner?

5.9 Other ways of counting

In the course *Algebra 1* you will encounter a result known variously as Burnside's Lemma, the Orbit-counting Theorem, and the Cauchy-Frobenius lemma — in the lecture notes it is called the 'Banenformule' [5]*5.7 on page 60.

This formula helps when counting the objects with symmetries, for example the number of 'different' ways in which you can colour the faces of a cube with one, two, three, ..., six colours.

Yet another way of counting is via *generating functions*; it provides a systematic way of solving problems like counting solutions to equations. See [1] for more.

Literature

There are tons of books on Discrete Mathematics. We mention a few.

[1] Graham, Ronald L., Knuth, Donald E. en Patashnik, Oren *Concrete mathematics (A foundation for computer science), 2^{nd} edition.* Addison-Wesley Publishing Company, Reading, MA.

[2] Lay, Steven R. *Analysis. With an Introduction to Proof.* Education, Inc., Boston.

[3] Lemmens, P.W.H. en Springer, T.A. *Hoofdstukken uit de Combinatoriek.* Epsilon 25. Epsilon Uitgaven, Utrecht. In Dutch

[4] Lovász, L., Pelikán, J. en Vesztergombi, K. *Discrete mathematics. Elementary and beyond.* Undergraduate Texts in Mathematics, Springer-Verlag, New York.

[5] Stevenhagen, P. *Algebra 1.* Collegedictaat Universiteit Leiden.

6

Probability and Statistics

Author: H.P. Lopuhaä

Introduction

Is everything on this planet determined by randomness? This question is open to philosophical debate. What is certain is that every day thousands and thousands of engineers, scientists, business persons, manufacturers, and others are using tools from probability and statistics.

The theory and practice of probability and statistics were developed during the last century and are still actively being refined and extended. In this chapter we will introduce some basic notions and computational rules from probability theory and discuss an often used statistical method of estimation.[1]

6.1 Events, probabilities and Bayes' rule

In early 2001 the European Commission introduced massive testing of cattle to determine infection with the transmissible form of *Bovine Spongiform Encephalopathy* (BSE) or "mad cow disease." As no test is 100% accurate, most tests have the problem of false positives and false negatives. A *false positive* means that according to the test the cow is infected, but actuality it is not. A *false negative* means an infected cow is not detected by the test.

Now suppose that the selected test has a small probability, say 1%, to produce a false positive or false negative, i.e.,

1. there is a 1% chance that a healthy cow is infected according to the test;

2. there is a 1% chance that an infected cow is healthy according to the test.

The crucial question is of course, suppose my cow is infected according to the test, then what is the probability that it really has BSE? To answer this question, we need some formal concepts and computational rules from probability theory. These will be introduced in the next section.

[1]These topics are based on Chapters 2, 3 and 21 from F.M. Dekking, C. Kraaikamp, H.P. Lopuhaä and L. Meester, *A Modern Introduction to Probability and Statistics - Understanding Why and How*, Springer 2005.

Events: notation and operations

The world around us is full of phenomena we perceive as random or unpredictable. We aim to model these phenomena as *outcomes* of some experiment, where you should think of *experiment* in a very general sense. The set of all possible outcomes is denoted by Ω, called the *sample space*.

For example, one of the most basic experiments is the tossing of a coin. Assuming that we will never see the coin land on its rim, there are two possible outcomes: heads and tails. We therefore take as the sample space associated with this experiment the set $\Omega = \{H, T\}$. In another experiment we ask the next person we meet on the street in which month her birthday falls. An obvious choice for the sample space is

$$\Omega = \{\text{Jan}, \text{Feb}, \text{Mar}, \text{Apr}, \text{May}, \text{Jun}, \text{Jul}, \text{Aug}, \text{Sep}, \text{Oct}, \text{Nov}, \text{Dec}\}\,.$$

Subsets of the sample space are called *events*. We say that an event A *occurs* if the outcome of the experiment is an element of the set A. For example, in the birthday experiment we can ask for the outcomes that correspond to a long month, i.e., a month with 31 days. This is the event

$$L = \{\text{Jan}, \text{Mar}, \text{May}, \text{Jul}, \text{Aug}, \text{Oct}, \text{Dec}\}\,.$$

When a person indicates to be born in the month of May, then we say that the event L occurs.

Events may be combined according to the usual set operations. For example, if R is the event that corresponds to the months that have the letter r in their (full) name, so

$$R = \{\text{Jan}, \text{Feb}, \text{Mar}, \text{Apr}, \text{Sep}, \text{Oct}, \text{Nov}, \text{Dec}\}\,,$$

then the long months that contain the letter r are

$$L \cap R = \{\text{Jan}, \text{Mar}, \text{Oct}, \text{Dec}\}\,.$$

The set $L \cap R$ is called the *intersection* of L and R and occurs if both L and R occur. Similarly, we have the *union* $A \cup B$ of two sets A and B, which occurs if at least one of the events A and B occurs. Another common operation is taking complements. The event $A^c = \{\omega \in \Omega : \omega \notin A\}$ is called the *complement* of A; it occurs if and only if A does *not* occur. The complement of Ω is denoted $\emptyset$, the empty set, which represents the *impossible event*. Figure 6.1 illustrates these three set operations by means of Venn diagrams. See also Section 2.1 in the book of the course TW1010 Mathematical Structures.

We call events A and B *disjoint* or *mutually exclusive* if A and B have no outcomes in common; in set terminology: $A \cap B = \emptyset$. For example, the event L "the birthday falls in a long month" and the event $\{\text{F}eb\}$ are disjoint. Finally, we say that event A *implies* event B if the outcomes of A also lie in B. In set notation: $A \subset B$; see Figure 6.2.

Exercise 6.1 We toss a coin three times. For this experiment we choose the sample space

$$\Omega = \{HHH, THH, HTH, HHT, TTH, THT, HTT, TTT\}$$

where T stands for tails and H for heads.

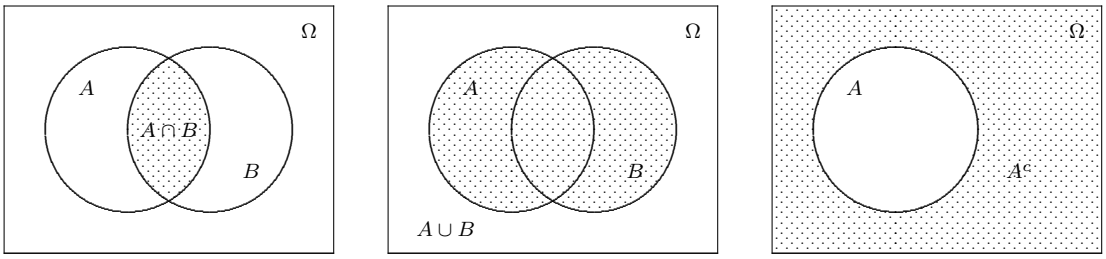


Figure 6.1: Set theoretical operations.

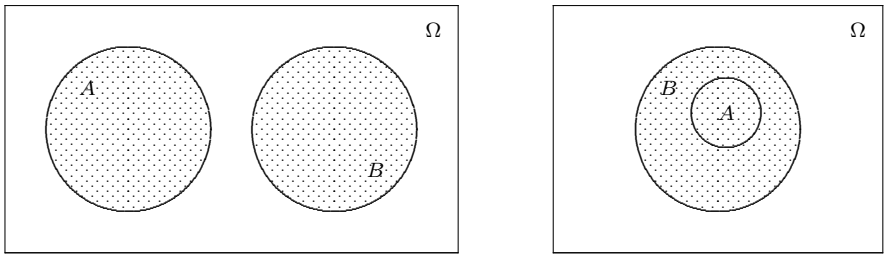


Figure 6.2: Minimal and maximal intersection.

(a) Write down the set of outcomes corresponding to each of the following events:

A : "we throw tails exactly two times."
B : "we throw tails at least two times."
C : "tails did not appear *before* a head appeared."
D : "the first throw results in tails."

(b) Write down the set of outcomes corresponding to each of the following events: A^c, $A \cup (C \cap D)$, and $A \cap D^c$.

Probabilities and computational rules

We want to express how likely it is that an event occurs. To do this we will assign a probability to each event.

Definition 6.1. The probability of an event A is denoted by a number $P(A)$ in $[0, 1]$. We assume that

1. $P(\Omega) = 1$;
2. for disjoint events A and B, it holds that $P(A \cup B) = P(A) + P(B)$.

The number $P(A)$ is called the probability that A occurs.

Note that some useful properties are immediate from this definition. To start with

$$\text{if } A \subset B, \text{ then } P(A) \leq P(B) \tag{6.1}$$

Indeed, we can write B as $B = A \cup (B \cap A^c)$, where A and $B \cap A^c$ are disjoint (make a picture). Therefore, from Definition 6.1 it follows that $P(B) = P(A) + P(B \cap A^c) \geq P(A)$. Another useful rule is

$$P(A^c) = 1 - P(A). \tag{6.2}$$

Exercise 6.2 Provide a proof of (6.2) by means of Definition 6.1.

Consider the events L, "born in a long month," and R, "born in a month with the letter r." If we suppose for convenience that all months are equally likely, their probabilities are easy to compute: since $L = \{\text{Jan, Mar, May, Jul, Aug, Oct, Dec}\}$ and $R = \{\text{Jan, Feb, Mar, Apr, Sep, Oct, Nov, Dec}\}$, one finds

$$P(L) = \frac{7}{12} \quad \text{and} \quad P(R) = \frac{8}{12}.$$

Now suppose that it is *known* about the person we meet in the street that he was born in a "long month," and we wonder whether he was born in a "month with the letter r." The given information excludes five outcomes of our sample space: it cannot be February, April, June, September, or November. Seven possible outcomes are left, of which only four—those in $R \cap L = \{\text{Jan, Mar, Oct, Dec}\}$—are favorable, so we reassess the probability as $4/7$. We call this the *conditional probability of R given L*, and we write:

$$P(R \mid L) = \frac{4}{7}.$$

This is not the same as $P(R \cap L)$, which is $1/3$. Also note that $P(R \mid L)$ is the ratio of $P(R \cap L)$ and $P(L)$.

Exercise 6.3 Let $N = R^c$ be the event "born in a month without r." Compute the conditional probability $P(N \mid L)$.

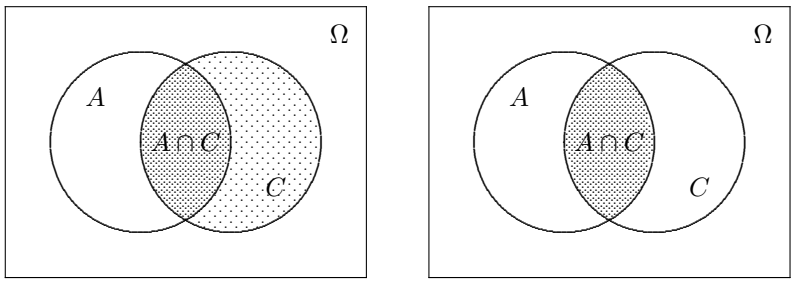
In view of the example above, in general we define the conditional probability of an event A given that an event C occurs, as follows.

Definition 6.2. The conditional probability of A given C is given by:

$$P(A \mid C) = \frac{P(A \cap C)}{P(C)},$$

provided $P(C) > 0$.

Note that $P(A \cap C)$, the probability that A and C occur simultaneously, is different from $P(A \mid C)$, the probability that A occurs, knowing that C occurs. Figuur 6.3 illustrates the difference between both probabilities. Whereas $P(A \cap C)$ can be represented as the proportion of Ω that is contained in $A \cap C$, the conditional probability $P(A \mid C)$ is the proportion of C that is contained in $A \cap C$.



$$P(A \mid C) = \frac{P(A \cap C)}{P(C)} \qquad P(A \cap C)$$

Figure 6.3: Conditional probability and probability of intersection.

Exercise 6.4 Let $P(C) > 0$. Prove that $P(A \mid C) + P(A^c \mid C) = 1$.

The conditional probability $P(A \mid C)$ can be computed by means of its definition. However, in many applications one is well aware of the value of $P(A \mid C)$ and Definition 6.2 is used to calculate other probabilities using the rule

$$P(A \cap C) = P(A \mid C) \times P(C). \tag{6.3}$$

This is called the *multiplication rule*. This allows the computation of $P(A \cap C)$ to be divided into two parts, computation of $P(C)$ and of $P(A \mid C)$, which is often easier than direct calculation of $P(A \cap C)$. One can also use (6.3) in the following way

$$P(A \cap C) = P(C \mid A) \times P(A).$$

Both equalities are correct, but usually only one of $P(A \mid C)$ and $P(C \mid A)$ is easy to determine and the other not.

For example, consider a pack of 52 cards from which we randomly draw two cards and define the events

$$S_1 = \text{first card is } \spadesuit,$$
$$S_2 = \text{second card is } \spadesuit.$$

What is $P(S_1 \cap S_2)$? By means of (6.3) we find

$$P(S_1 \cap S_2) = P(S_2 \mid S_1) \times P(S_1) = \frac{12}{51} \times \frac{13}{52} = \frac{1}{17}.$$

In this example, it also holds that

$$P(S_1 \cap S_2) = P(S_1 \mid S_2) \times P(S_2),$$

but this formula is useless, because $P(S_1 \mid S_2)$ is not so straightforward to compute in contrast with $P(S_2 \mid S_1)$.

Law of total probability and Bayes' rule

We return to the BSE example from the beginning of this chapter. Imagine we test a cow. Let B denote the event "the cow has BSE" and T the event "the test comes up positive" (this is test jargon for: according to the test we should believe the cow is infected with BSE). One can "test the test" by analyzing samples from cows that are known to be infected or known to be healthy and so determine the effectiveness of the test. The European Commission had this done for four tests in 1999 and for several more later.[2] The results for what the report calls Test A may be summarized as follows: an infected cow has a 70% chance of testing positive, and a healthy cow just 10%; in formulas:

$$P(T \mid B) = 0.70,$$
$$P(T \mid B^c) = 0.10.$$

Law of total probability. Which percentage of cows will test positive with Test A, what is $P(T)$? The tested cow is either infected or healthy: event T occurs in combination with B or with B^c (there are no other options). In other words,

$$T = (T \cap B) \cup (T \cap B^c),$$

so that

$$P(T) = P(T \cap B) + P(T \cap B^c),$$

because $T \cap B$ en $T \cap B^c$ are disjoint. Then we apply (6.3) in such a way that the given conditional probabilities can be used:

$$P(T \cap B) = P(T \mid B) \times P(B),$$
$$P(T \cap B^c) = P(T \mid B^c) \times P(B^c),$$

which leads to the following rule

$$P(T) = P(T \mid B) \times P(B) + P(T \mid B^c) \times P(B^c). \tag{6.4}$$

This rule we call the *law of total probability*: computation of a probability by conditioning on various disjoint events, which together contain all possible outcomes.

Exercise 6.5 Consider a deck of 52 cards from which we randomly draw two cards and define the events

$$S_1 = \text{first card is } \spadesuit,$$
$$S_2 = \text{second card is } \spadesuit.$$

Compute $P(S_2)$ by application of the law of total probability.

In fact, the rule is more general and (6.4) is only a special case with two disjoint events B and B^c. Figure 6.4 is an illustration of the rule, when conditioned on five disjoint sets $C_1, C_2, \ldots, C_5$. The set A is the disjoint union of the sets $(A \cap C_1), (A \cap$

[2] See J. Moynagh, H. Schimmel, en G.N. Kramer. The evaluation of tests for the diagnosis of transmissible spongiform encephalopathy in bovines. Technical report, European Commission, Directorate General XXIV, Brussels, 1999.

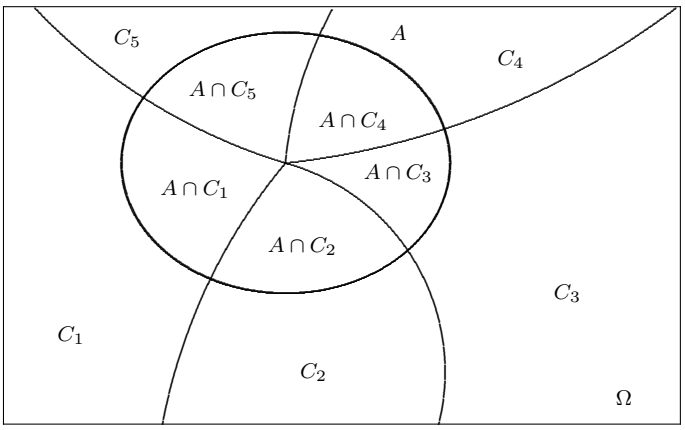
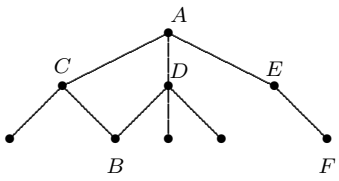


Figure 6.4: Law of total probability with 5 disjoint sets.

$C_2), \ldots, (A \cap C_5)$, so that $P(A) = P(A \cap C_1) + (A \cap C_2) + \cdots + P(A \cap C_5)$. Thereafter, we can apply the multiplication rule (6.3): $P(A \cap C_i) = P(A \mid C_i) \times P(C_i)$, for each $i = 1, 2, \ldots, 5$.

Exercise 6.6 Someone wants to walk from A to B (see the map). To do so, he first randomly selects one of the paths to C, D, or E. Next, he selects randomly one of the possible paths at that moment (so if he first selected the path to E, he can either select the path to A or the path to F), etc. What is the probability that he will reach B after two selections?



We continue with the BSE example. When we process and apply the results of our test to (6.4), we have

$$P(T) = 0.70 \times P(B) + 0.10 \times P(B^c).$$

In short, if we have information about $P(B)$, and thus also about $P(B^c) = 1 - P(B)$ (see Exercise 2), then we can compute $P(T)$. Suppose[3] for the convenience that 2% of the cows are actually infected, i.e., $P(B) = 0.02$, then

$$P(T) = 0.70 \times 0.02 + 0.10 \times (1 - 0.02) = 0.112.$$

We see that, despite the 70% chance of detecting infected cows, and with 10% risk of indicating healthy cows as infected, no less than 11.2% (a factor 5.5 too high) of all cows

[3]This assumption is only done to make the calculations insightful. The actual value is unknown and varies from country to country. The BSE risk for the Netherlands in 2003 was estimated at $P(B) \approx 0.000013$.

test positive. This is of course due to the fact that in our example only 2% of the cows are infected. This causes $P(T \mid B) = 0.70$ to work only in 2% of the cases, and the influence of the false positive $P(T \mid B^c) = 0.10$ is much bigger.

Exercise 6.7 In 2003, the BSE risk for the Netherlands was estimated at $P(B) = 0.000013$. Compute $P(T)$ for this case.

Bayes' rule. The much more important question from the beginning of this section is: if a cow tests positive on BSE, then what is the probability that the cow is actually infected? In formula: what is $P(B \mid T)$? We have information about $P(T \mid B)$, a conditional probability, but the wrong one. In fact, we want to change the roles of T and B.

This is possible as follows. Starting with the definition of conditional probability we get

$$P(B \mid T) = \frac{P(B \cap T)}{P(T)} = \frac{P(T \mid B) \times P(B)}{P(T)}.$$

Then we apply (6.4) to the denominator, so that

$$P(B \mid T) = \frac{P(T \mid B) \times P(B)}{P(T \mid B) \times P(B) + P(T \mid B^c) \times P(B^c)}. \tag{6.5}$$

This computational rule is called *Bayes' rule*, named after the English clergyman Thomas Bayes who derived this in the 18th century. When we apply this rule to our example, with $P(B) = 0.02$, we get

$$P(B \mid T) = \frac{0.70 \times 0.02}{0.70 \times 0.02 + 0.10 \times (1 - 0.02)} = 0.125,$$

and with a similar computation: $P(B \mid T^c) = 0.0068$. These conditional probabilities show that T is a bad test. Indeed, a perfect test would have $P(B \mid T) = 1$ and $P(B \mid T^c) = 0$. In Exercise 8 we redo the computations with a more realistic value for $P(B)$.

Exercise 6.8 In 2003, the BSE risk for the Netherlands was estimated at $P(B) = 0.000013$. Compute $P(B \mid T)$ and $P(B \mid T^c)$ for this case.

Exercise 6.9 We return to the example in the introduction of this chapter, i.e., assume we have a test for which $P(T \mid B^c) = 0.01$ en $P(T^c \mid B) = 0.01$. Furthermore, suppose it is known that 1 in 1000 cows is infected. Compute the probability that a cow (actually) is infected, given that it is infected according to the test.

Independence

Consider three probabilities from the previous section:

$$\begin{aligned} P(B) &= 0.02 \\ P(B \mid T) &= 0.125 \\ P(B \mid T^c) &= 0.0068. \end{aligned}$$

If we know nothing about a cow, we would say that there is a 2% chance it is infected. However, if we know it tested positive, we can say there is a 12.5% chance the cow is

infected. On the other hand, if it tested negative, there is only a 0.68% chance. We see that the two events are related in some way: the probability of B *depends* on whether T occurs or not.

Imagine the opposite: the test is useless. Whether the cow is infected is unrelated to the outcome of the test, and knowing the outcome of the test does not change our probability of B: $P(B \mid T) = P(B)$. In this case we would call B *independent* of T.

Definition 6.3. An event A is called independent of B if

$$P(A \mid B) = P(A). \tag{6.6}$$

The above definition of independence initially appears to be one-sided. You would think that if A is independent of B, then B is also independent of A. That this is indeed the case, can be seen from the following exercise.

Exercise 6.10 Suppose A is independent of B, so that $P(A \mid B) = P(A)$. Show that B is independent A, i.e., $P(B \mid A) = P(B)$.

In short, if A is independent of B, then this also applies the other way round. In this case, we call A and B independent events. Note that from the multiplication rule (6.3) it immediately follows that if A and B are independent events, then

$$P(A \cap B) = P(A \mid B) \times P(B) = P(A) \times P(B). \tag{6.7}$$

Although intuitively, the definition (6.6) of independence is more natural, the above rule (6.7) is used more often, when calculating the probability of independent events. Just yet, we have seen that (6.6) implies (6.7). The next exercise shows that this also hold the other way around, so that (6.6) and (6.7) are equivalent.

Exercise 6.11 Suppose that $P(A \cap B) = P(A) \times P(B)$. Show that A and B are independent, i.e., $P(A \mid B) = P(A)$ and $P(B \mid A) = P(B)$.

6.2 Estimating unknown parameters

During World War II, the Allied forces started to analyze markings and serial numbers obtained from captured German equipment. An extensive description about how one went about and what sort of obstacles one had to overcome is described in an article by Richard Ruggles and Henry Brodie.[4] A start was made by analyzing serial numbers on car tires. For each manufacturer, the serial number started with a two letter code. These were suspected to correspond to the month and year of manufacturing. This turned out to be correct. Table 6.1 shows a number examples of month codes of some manufacturers. For example, the Dunlop code was Dunlop Arbeit spelled backwards. After breaking the year code as well, one was able to recode the serial numbers as numbers running from 1 to some unknown largest number N, and the observed (recoded) serial numbers could be viewed as a part of this. The goal was now to obtain, for each month for each manufacturer, an estimate of N on the basis of the observed serial numbers.

After breaking the month and year codes, the estimation problem reduces to the following statistical problem: given is a vase of balls numbered from 1 to N; estimate N

[4] Ruggles, R. and Brodie, H. (1947) An empirical approach to economic intelligence in World War II, *Journal of the American Statistical Association*, **42**, p. 72-91.

Table 6.1: Coding of German manufacturers of tires.

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Dunlop	T	I	E	B	R	A	P	O	L	N	U	D
Fulda	F	U	L	D	A	M	U	N	S	T	E	R
Phoenix	F	O	N	I	X	H	A	M	B	U	R	G
Sempirit	A	B	C	D	E	F	G	H	I	J	K	L

on the basis of the numbers $x_1, x_2, \ldots, x_n$, which are randomly drawn from the vase without replacement. One idea to get an estimate is the following. The average of the numbers $1, 2, \ldots, N$ in the vase is

$$\frac{1}{N} \times (1 + 2 + \cdots + N) = \frac{1}{N} \times \frac{1}{2} N(N+1) = \frac{N+1}{2}.$$

This should be approximately equal to the average $\bar{x}$ of the numbers drawn:

$$\bar{x} \approx \frac{N+1}{2},$$

so that

$$2\bar{x} - 1 \approx N.$$

Therefore, take $2\bar{x} - 1$ as an estimate for N.

Another idea is that the largest observed number $m = \max(x_1, x_2, \ldots, x_n)$ must give an indication about N. Since m will generally be less than N, this estimate can be improved immediately by multiplying m with a factor greater than 1. Assuming that the n numbers drwan are more or less equally distributed in the interval $[0, N]$, the maximum is approximately

$$m \approx \frac{n}{n+1} \times N$$

so that

$$\frac{n+1}{n} m \approx N.$$

Therefore, take $(n+1)m/n$ as an estimate N. This gives us two possible methods to estimate N. Which of the two estimators is better now?

An important part of Statistics is devoted to estimating an unknown quantity based on observations, such as the monthly number of tires produced per manufacturer. Most often there are several ad hoc methods to be found and the question arises which one is better. In this section we introduce a universal method of maximum likelihood, which provides the best estimates in a particular sentence.

The maximum likelihood principle

Consider the following (minor) problem. Given are two dice: one with 5 white sides and 1 red, the other with 1 white side and 5 red. Someone chooses one of both dice and uses it three times to do the same experiment: throwing just as long as red comes up. We do

not know which dice has been chosen to throw, but we only know the required numbers in each of the three experiments:

$$7, 4 \text{ and } 10.$$

The question is to determine which dice has been used to throw on the basis of these data. It is everyones gut feeling to choose the first dice. Usually the explanation is that the dice has more white faces and that 18 times white is thrown and 3 times red. Probably, the problem is too easy to pose. Too bad, because despite its simplicity, the above example illustrates an important principle in the statistics: the principle of *maximum likelihood*. One chooses the dice, for which the required numbers are 7, 4 and 10 are most likely.

Indeed, when we assume that the results of the throws are independent, then the probability for having to throw 7 times in the first experiment for a first time red for the *first* dice equal to

$$\begin{aligned} P(\text{first time R in throw 7}) &= \underbrace{P(W) \times P(W) \times \cdots \times P(W)}_{6 \text{ times}} \times P(R) \\ &= \frac{5}{6} \times \frac{5}{6} \times \cdots \times \frac{5}{6} \times \frac{1}{6} = \left(\frac{5}{6}\right)^6 \frac{1}{6}. \end{aligned}$$

Similar calculations also apply to the other two experiments, so that because of the independence of the experiments, the probability of having 7, 4, and 10 is equal to

$$\left(\frac{5}{6}\right)^6 \frac{1}{6} \times \left(\frac{5}{6}\right)^3 \frac{1}{6} \times \left(\frac{5}{6}\right)^9 \frac{1}{6} = \frac{5^{18}}{6^{21}} = 0.0001738937.$$

In the same way one finds for the other dice that the probability of having 7, 4, and 10, is equal to

$$\left(\frac{1}{6}\right)^6 \frac{5}{6} \times \left(\frac{1}{6}\right)^3 \frac{5}{6} \times \left(\frac{1}{6}\right)^9 \frac{5}{6} = \frac{5^3}{6^{21}} = 5.7 \times 10^{-15}.$$

Both probabilities are very small, but the probability of the observed outcomes 7, 4 and 10 is many times bigger (5^{15} times as big!) for the first dice than for the second dice. Hence, choosing for the first dice is the same as choosing the dice for which the observed data is most likely. This principle, which is called the *principle of maximum* likelihood, will also be used when we have to choose a value as an estimate for an unknown parameter based on observations.

Definition 6.4. Suppose that we have to estimate an unknown parameter based on observations, shortly called data. According to the principle of maximum likelihood, we take as an estimate of the unknown parameter, the value for which the probability of the data is the largest.

The likelihood: the probability of the data

Let's make the "dice problem" somewhat more complicated. We replace the throwing of dice by turning a "wheel of fortune". The area in which the pointer can end up when the wheel comes to a halt, consists of an unknown proportion p colored red and the rest is white. We do the same experiment three times: turn the wheel until the pointer ends in red. Suppose the required number of attempts to a first time red are again 7, 4

and 10. The question is to estimate the parameter p (the proportion red) based on these observations. We will apply the principle of maximum likelihood. The next step is then to calculate the probability that in the successive experiments with the wheel of fortune, one must turn 7, 4, and 10 times, respectively, for a first time red. In the same way as in the previous paragraph the probability of the data (the observed outcomes 7, 4 and 10) is equal to

$$(1-p)^6 p \times (1-p)^3 p \times (1-p)^9 p.$$

In this way, the probability of the data becomes a function of the parameter $p \in [0,1]$. This function is called the *likelihood function*, or shortly *likelihood*, and is usually denoted by the capital letter L. According to the principle of maximum likelihood, we now take as an estimate for p, the value which maximizes the likelihood $L(p)$.

First note that in the example above we can simplify things:

$$L(p) = (1-p)^{18} p^3, \quad \text{for } p \in [0,1].$$

The maximum of $L(p)$, for $p \in [0,1]$, can be found by setting the derivative equal to zero. The derivative of $L(p)$ is

$$\begin{aligned} L'(p) &= -18(1-p)^{17}p^3 + 3(1-p)^{18}p^2 \\ &= (1-p)^{17}p^2\left[-18p + 3(1-p)\right] \\ &= (1-p)^{17}p^2\left[-21p + 3\right]. \end{aligned}$$

Therefore $L'(p) = 0$ if and only if $p = 0$, $p = 1$, or $p = 1/7$. One can check that this means that $L(p)$ has a unique maximum at $p = 1/7$. We conclude that $p = 1/7$ is the maximum likelihood estimate for p.

Exercise 6.12 A retailer in computer chips is offered two parties on the black market each consisting of 10 000 chips. It is known that one party has 50% defective chips and the other party only 10%. The retailer is willing to buy the party with 10% defective chips, but he does not know which party it is. He is given the opportunity to choose 10 chips from one of the parties to test. After choosing, it appears that the first three chips are defective and the rest is good. When the retailer uses the principle of maximum likelihood, which party should he buy, the party tested or the other one?

Exercise 6.13 Consider the "wheel of fortune" problem. Suppose one observes $x_1, x_2, \ldots, x_n$ as outcomes, i.e., during the i-th experiment one had to turn x_i times before a first time red, for $i = 1, 2, \ldots, n$.

1. Deduce that the likelihood is given by $L(p) = (1-p)^{x_1+x_2+\cdots+x_n-n}p^n$.

2. Show that the maximum likelihood estimate is given by

$$\frac{n}{x_1 + x_2 + \cdots + x_n}.$$

A totally different, and much simpler, argument is that we have seen 3 times red on 21 throws, so that the probability p of red is equal to $p = 3/21 = 1/7$. This last argument illustrates another estimation principle: the *method of moments*. In this

specific example, this method entails that a probability $P(A)$ is estimated by counting the relative frequency of the occurrence of the event A in a series of repetitions of the same experiment. This method seems very attractive because of its simplicity and seems to yield the same as the more complicated principle of maximum likelihood. However, the fact that in this example the two principles yield the same estimate is a coincidence. In the remainder of this section we will first consider an example to which the method of moments cannot be applied, but where the principle of maximum likelihood offers a solution. Then we end this section with an example that will show that the maximum likelihood estimator is better than the moment estimator.

Remark 6.5. The moments method has a long history. Karl Pearson (1857-1936) is often referred to as the one who emphasized the importance of this method. In the beginning of the 20th century there has been a stir about the properties of the method of moments compared to other methods, especially the controversy between Pearson and Ronald Aylmer Fisher (1890-1962). In his first article,[5] Fisher rediscovered the method of maximum likelihood, already known by Lambert (1760) and Bernoulli (1777). Fisher's work in the field of statistics drew Pearson's attention. As editor of the journal *Biometrika*, Pearson published an article by Fisher in 1915 about the probability distribution of the correlation coefficient. In a follow-up article in *Biometrika* in 1916, Pearson criticized the former article by Fisher without prior notice. Pearson did not understand the maximum likelihood method used by Fisher, and wrongly burned it to the ground. Fisher was grieved by Pearson's self-empowered performance and his lack of understanding, which ultimately led to their violent confrontation. Nevertheless, in 1919 Pearson asked Fisher to accept a position at University College in London, where he fulfilled the chair of Eugenetics. Fisher refused.

Maximum likelihood in case of incomplete data

A situation that is very similar to the "wheel of fortune" problem is discussed in a paper by Weinberg and Gladen[6]. They examined the number of menstrual cycles until pregnancy, measured from the time they had decided to conceive. During the study, data was collected from 100 smoking and 486 non-smoking women. These are summarized in Table 6.2. Note that for 7 smoking and 12 non-smoking women the data is incomplete;

Table 6.2: Observed number of cycles until pregnancy.

Number of cycles	1	2	3	4	5	6	7	8	9	10	11	12	>12
Smokers	29	16	17	4	3	9	4	5	1	1	1	3	7
Non-smokers	198	107	55	38	18	22	7	9	5	3	6	6	12

Source: C.R. Weinberg and B.C. Gladen. The beta-geometric distribution applied to comparative fecundability studies. *Biometrics*, 42(3):547–560, 1986.

we only have information that more than 12 cycles were needed for pregnancy. The

[5]Fisher, R.A. (1912) *Messeng. Math.*, **41**, p. 155-160.

[6]C.R. Weinberg and B.C. Gladen (1986) The beta-geometric distribution applied to comparative fecundability studies. *Biometrics*, **42**(3):547–560.

question is now: what are the probabilities of pregnancy for smoking and non-smoking women, and are these very different from each other?

Suppose for the moment that the probability of pregnancy is the same during each cycle, say p, with $0 < p \leq 1$, and that the outcomes during the different cycles are independent of each other. Then, as in the "wheel of fortune" problem, it holds that

$$P(\text{pregnancy in } k\text{-th cycle}) = (1-p)^{k-1}p, \quad \text{for } k = 1, 2, \ldots$$

These probabilities form the so-called *Geometric* distribution with parameter $p \in [0, 1]$. The fact that this is indeed a proper probability distribution, is established in the next exercise.

Exercise 6.14 Show that the probabilities $(1-p)^{k-1}p$, for $k = 1, 2, \ldots$ sum up to 1.

Because $P(\text{pregnancy in first cycle}) = p$, a simple estimate for p is:

$$\frac{\text{number of women with pregnancy in the first cycle}}{\text{total number of women}}.$$

This gives estimates $p = 29/100 = 0.29$ for the smoking women and $p = 198/486 = 0.41$ for non-smoking women. Intuitively, it must be clear that this cannot be the best method. After all, in this way much of the data in Table 6.2 is not used.

Of course we want a method that uses *all* information. However, for 7 smoking and 12 non-smoking women, we only know that the number of attempts has been more than 12, but we do not know exactly how many attempts have been made until pregnancy. When we simply ignore the information in the last column of Table 6.2, then we can compute estimates, analogously to the "wheel of fortune" problem, using the formula from Exercise 13.

Exercise 6.15 Consider the data of Table 6.2. Suppose we ignore the last column and compute the estimates for the probability p on pregnancy using only the first 12 columns. Show that both the maximum likelihood estimate and the method of moments estimate are $p = 0.281$ for smoking women and $p = 0.3688$ for non-smoking women.

Note, however, that this method overestimates the parameter p. After all, the formula from Exercise 13 is $1/\bar{x}$. Suppose that you would know the exact number of attempts of the women in the last column in Table 6.2. In that case, the average number of attempts would be higher, which means that the estimate $1/\bar{x}$ would be lower. In short, by ignoring the information from the last column, we get too high estimates for p.

The nice thing is that the principle of maximum likelihood offers a way out here. Despite the limited information in the last column of Table 6.2, we can still compute the probability of the data. Namely,

$$\begin{aligned} P(\text{pregnancy after the 12-th cycle}) &= P(\text{no pregnancy in cycles 1 to 12}) \\ &= (1-p)^{12}. \end{aligned}$$

Furthermore, we see from Table 6.2 that 29 smoking women are successful during the

first cycle. This occurs with probability p^{29}. In this way, for the smoking women, we find

Event	Probability
29 times pregnancy in cycle 1	p^{29}
16 times pregnancy in cycle 2	$\{(1-p)p\}^{16}$
17 times pregnancy in cycle 3	$\{(1-p)^2p\}^{17}$
$\vdots$	$\vdots$
3 times pregnancy in cycle 12	$\{(1-p)^{11}p\}^{3}$
7 times pregnancy after cycle 12	$\{(1-p)^{12}\}^{7}$

This means that the probability of the data (the likelihood) for the smoking women is given by

$$\begin{aligned} L(p) &= C \times p^{29} \times \{(1-p)p\}^{16} \times \{(1-p)^2p\}^{17} \times \cdots \times \{(1-p)^{11}p\}^{3} \times \{(1-p)^{12}\}^{7} \\ &= C \times p^{93} \times (1-p)^{322}. \end{aligned}$$

Here, C is the number of possibilities of having 29 times a 1, 16 times a 2, ..., 3 times a 12, and 7 times some number larger than 12, for 100 smoking women. We make no effort to calculate this number because it does not depend on p.[7] According to the principle of maximum likelihood the estimate for p is the value that maximizes $L(p)$. One can check that

$$L'(p) = C \times p^{92}(1-p)^{321}\,(93 - 415p)$$

so that $L(p)$ has a unique maximum (check this!) at $p = 93/415 = 0.224$. We say that for the smoking women, the maximum likelihood estimate is given by $p = 0.224$. Note that this estimate is quite smaller than the estimate 0.29, based on only the first column of Table 6.2, and estimate 0.281 from Exercise 15.

Exercise 6.16 Deduce that for the non-smoking women, the likelihood is given by

$$L(p) = \text{constant} \times p^{474} \times (1-p)^{955}.$$

Compute the maximum likelihood estimate for p.

Example of a non-differentiable likelihood

To determine where the likelihood L attains its maximum, so far we have only seen examples where we could differentiate L. This is not always possible. For example, suppose we have a vase of balls numbered $1, 2, \ldots, N$, where N is unknown. We randomly draw (without replacement) out of the vase five balls with numbers:

$$40,\ 28,\ 7,\ 44 \text{ and } 18.$$

Compute the maximum likelihood estimate for N, the number of balls in the vase, based on this data. The first step is to calculate the probability of the data, as a function of the unknown parameter N. The largest number drawn is 44. This means that for

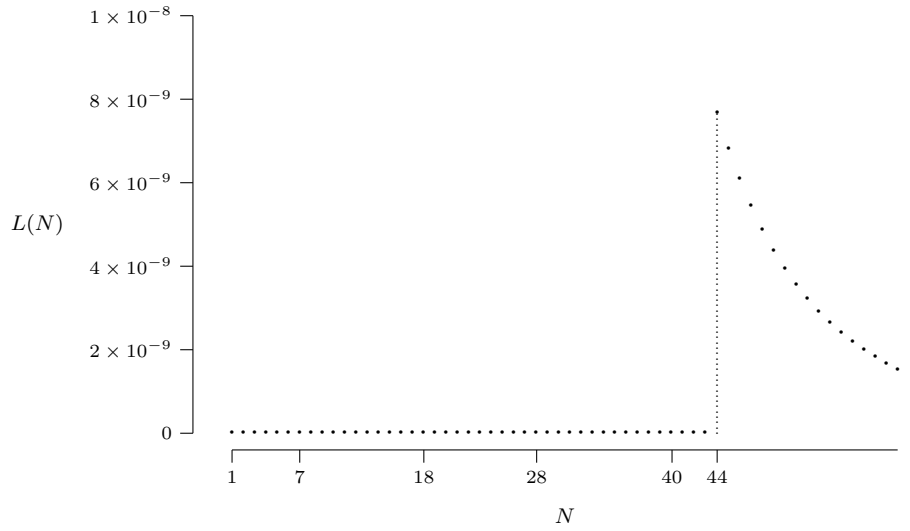


Figure 6.5: Likelihood $L(N)$.

$N = 1, 2, \ldots, 43$, the probability of the above data is equal to zero! For $N = 44, 45, \ldots$, the probability of the data is equal to

$$\frac{1}{N(N-1)(N-2)(N-3)(N-4)}.$$

Hence, the likelihood is given by

$$L(N) = \begin{cases} 0 & \text{, for } N = 1, 2, \ldots, 43; \\ \dfrac{1}{N(N-1)(N-2)(N-3)(N-4)} & \text{, for } N = 44, 45, \ldots. \end{cases}$$

Figure 6.5 contains a graph of the likelihood. The maximum likelihood estimate for N is given by that value of N for which the probability $L(N)$ is maximal. From the description of $L(N)$, see also Figure 6.5, it immediately follows that this is at $N = 44$. Therefore, the largest number drawn is the maximum likelihood estimate for the number of balls in the vase.

Estimating the German war production

The previous example shows that the principle of maximum likelihood does not always provide a good estimate directly. After all, the largest number drawn will generally be smaller than the largest number N in the vase. That means that we sometimes need to adjust the maximum likelihood estimate in order to turn it into a good estimate. However, without going into details, there are theorems in mathematical statistics stating that the (adjusted) maximum likelihood estimates are the best, in a particular sentence. We will illustrate this by means of the example from the beginning of this chapter about estimating German war production.

[7] $C = 311657028822819441451842682167854800096263625208359116504431153487280760832000000000$.

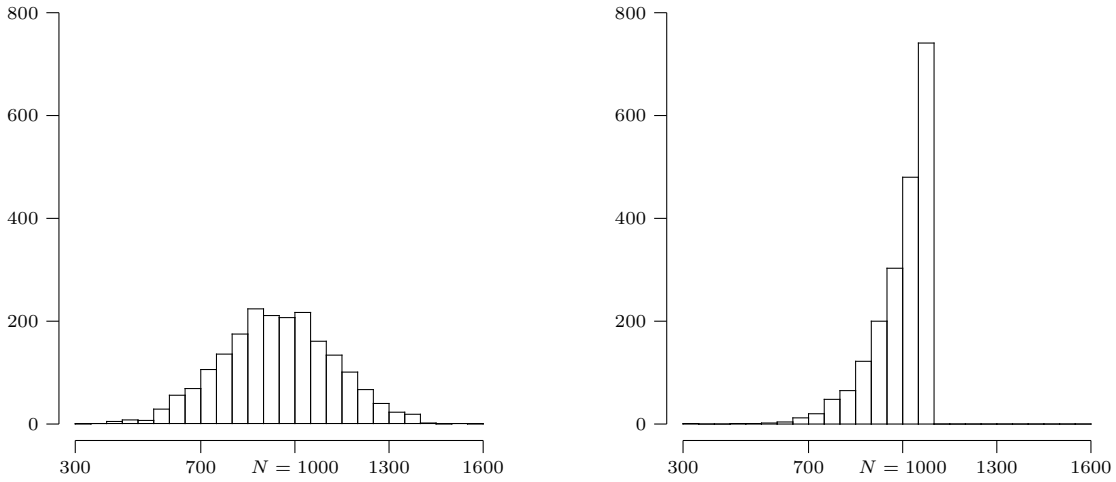


Figure 6.6: Histograms of 2000 values of s_1 (left panel) and s_2 (right panel).

Accuracy of two estimators: a computer simulation. After the problem had been reduced to estimation of the number of balls in a vase numbered $1, 2, \ldots, N$, via some ad hoc reasoning we arrived at two estimators

$$s_1 = 2\bar{x} - 1$$
$$s_2 = \frac{n+1}{n} m.$$

Looking back, we can now see that s_2 is the adjusted maximum likelihood estimate. Furthermore, we note that s_1 is in fact the result of the method of moments. Without going further into mathematical details, we show the performance of both estimators by means of a computer simulation.

Suppose the vase contains $N = 1000$ balls and that we draw $n = 10$ balls from the vase without replacement. We can imitate this on a computer. Thereafter we can calculate estimates s_1 and s_2 based on the numbers drawn. We can easily repeat this procedure on the computer, say 2000 times. Because the 10 numbers are drawn randomly, with any repetition of the procedure slightly different numbers will be drawn and the value of the estimates s_1 and s_2 will vary. For a reasonable method of estimation, we would like the estimates to vary around the parameter of interest, in our case $N = 1000$. Finally, the method that varies the least can then be regarded as the best.

Figure 6.6 shows the results of this computer simulation, summarized in two histograms. We see that for both methods, the estimates vary around $N = 1000$. In fact, for this simulation, the average of the two thousand s_1 values is 998.15. Although the histogram of the s_2 values is skewed, the average of the two thousand s_2 values is equal to 1001.52. In short, both estimation methods are on average "on target". We can also prove this mathematically, by means of the notion of *unbiasedness*, but we will do that in the second-year course *Introduction to Statistics*.

More interesting is that the values of s_2 vary much less around $N = 1000$ than the values of s_1. We can also formalize this mathematically by using the notion of *variance*,

but once more we will do this in a follow-up course. It appears that the amount of variation of estimate s_1 is four times larger than that of s_2. One can show that in general, the amount of variation of estimate s_1 is a factor $(n+2)/3$ times larger than that of s_2. In short, with larger data sets, the difference in accuracy between s_1 and s_2 only increases. The estimate s_2, based on the maximum likelihood estimate is to be preferred above s_1.

Results in the Second World War. During the Second World War estimation method s_2 was used. After the end of the war, it appeared how accurate the estimates were, especially compared with the estimates delivered by the secret service during the war. A good example are the estimates for the average monthly production of tires in 1943. The data are summarized in Table 6.3. The maximum likelihood estimates hardly differ from the actual figures. The estimates of the secret service for total monthly production, likely to be influenced by the German propaganda machine, turned out to be a factor five too high. Other examples concern the production of trucks in 1942, see Table 6.4, and the average monthly production of tanks, see Table 6.5.

Table 6.3: Average monthly production of tires in 1943.

Type of tire	estimate	true production	secret service
Truck and car	147 000	159 000	
Airplane	28 500	26 400	
Total	175 500	186 100	900 000 – 1 200 000

Table 6.4: Production of trucks in 1942.

Type of truck	estimate	true production	secret service
Light truck	16 500	14 436	
Medium truck	62 300	53 439	
Heavy truck	18 500	11 952	
Total	97 300	79 827	200 000

Table 6.5: Average monthly production of tanks in 1940-1942.

Date	estimate	true production	secret service
June 1940	169	122	1000
June 1941	244	271	1550
August 1942	327	342	1550

6.3 Exercises

Exercises for Section 1

Exercise 6.17 Suppose A and B are two events with $P(A) = 0.3$, $P(B) = 0.4$, and $P(A \cap B) = 0.2$. What is $P(A^c \cap B)$?

Exercise 6.18 Give a formal proof of the following rule

$$P(A \cup B) = P(A) + P(B) - P(A \cap B),$$

by making use of Definition 6.1.

Exercise 6.19 Generally $P(A \mid C) + P(A \mid C^c) = 1$ is not true. Come up with a counter example.

Exercise 6.20 A student participates in a multiple-choice exam. Suppose that for each question, he either knows the answer, or randomly choses from the four different options. When he knows the answer, the probability of a correct answer is 1 and equal to 1/4, if he gambles. To pass the exam you must answer 60% of the questions correctly. Suppose the student has "learned for a 6", i.e., the probability that he knows the answer to a question is 0.6. What is the probability that the student actually *knows* the answer, given that he answered the question correctly?

Exercises for Section 2

Exercise 6.21 One conducts a series of 100 experiments that can result in "success" and "failure". The outcomes of the different experiments are independent of each other and in each experiment the probability of "success" is equal to some unknown number $p \in [0, 1]$. One finds 63 "successes".

(i) Give the formula for the likelihood $L(p)$.

(ii) Compute the maximum likelihood estimate for p.

Exercise 6.22 One conducts a series of n experiments that can result in "success" and "failure". The outcomes of the different experiment are independent of each other and in each experiment the probability of "success" is equal to some unknown number $p \in [0, 1]$. One finds k "successes".

(i) Give the formula for the likelihood $L(p)$.

(ii) Compute the maximum likelihood estimate for p.

Exercise 6.23 During a survey in the London subway, at a particular subway station one counted how many women were present in 100 rows each consisting of 10 people. In this way, a data set was formed with numbers $x_1, x_2, \ldots, x_{100}$, where x_i is the observed number of women in the i-th row, for $i = 1, 2, \ldots, 100$. We assume that the rows are independent of each other, as well as the gender of the persons in the different positions in a row. Suppose that for each row and for each position, the probability of a woman is equal to p.

(i) What is the probability of having x_i women in the i-th row?

(ii) Give the formula of the likelihood $L(p)$, the probability of the data $x_1, x_2, \ldots, x_{100}$.

(iii) Give the formula of the maximum likelihood estimate for p.

The data is summarized in the table below.

Number of women	0	1	2	3	4	5	6	7	8	9	10
Number of rows	1	3	4	23	25	19	18	5	1	1	0

Source: R.A. Jinkinson and M. Slater. Critical discussion of a graphical method for identifying discrete distributions. *The Statistician*, 30:239–248, 1981; Table 1 on page 240.

(iv) Compute the maximum likelihood estimate on the basis of the table above.

(v) What is the method of moments estimate?

Exercise 6.24 Events that occur with a very small probability, but which can take place at any position in a particular area, in such a way that the average number of events per unit of area is more or less constant, say $\lambda > 0$, are often modeled by a *Poisson* probability distribution:

$$P(k \text{ events}) = \frac{\lambda^k}{k!} \mathrm{e}^{-\lambda}, \quad \text{for } k = 0, 1, 2, \ldots$$

A nice example concerns bombs hitting London during World War II. An area of 36 km^2 in South London was subdivided into 576 squares with sides of 1/4 kilometer. In each of the 576 square subdivisions, the number of hits was counted of V2 missiles, fired by the German army on London. In this way, a dataset $x_1, x_2, \ldots, x_{567}$ was collected, where x_i is the number of hits in the i-th square. The data are summarized in the table below, which lists the number of squares with 0 hits, 1 hit, 2 hits, and so on.

Aantal hits	0	1	2	3	4	5	6	7
Aantal vierkanten	229	211	93	35	7	0	0	1

Source: R.D. Clarke. An application of the Poisson distribution. *Journal of the Institute of Actuaries*, 72:48, 1946; Table 1 on page 481. © Faculty and Institute of Actuaries.

On the basis of this information we want to compute the maximum likelihood estimate for the parameter λ, which represents the number of hits in a square with sides of 1/4 kilometer.

(i) Deduce that the likelihood of the data is given by

$$L(\lambda) = C \times \frac{1}{(2!)^{93} \times (3!)^{35} \times (4!)^7 \times (7!)} \lambda^{537} \mathrm{e}^{-576\lambda}$$

where C is the number of possibilities for having 229 times a 0, 211 times a 1,..., 1 time a 7.

(ii) Compute the maximum likelihood estimate for λ.

Exercise 6.25 A vase contains balls numbered from $n, n+1, \ldots$, where n is unknown as well as the total number of balls in the vase. We randomly draw 5 balls out of the vase without replacement:

$$40 \quad 28 \quad 7 \quad 44 \quad 18$$

(i) Sketch the graph of the likelihood $L(n)$.

(ii) What is the maximum likelihood estimate for n?

7

Hints and answers

Graphs

Solution 1.1
The edge set is given by $E = \{\{1,3\},\{1,5\},\{1,6\},\{2,3\},\{2,4\},\{3,6\},\{4,6\},\{5,6\}\}$.

Solution 1.2
There are 2^{10} graphs with node set $\{1,2,3,4,5\}$.

Solution 1.3
The graph C_n has n edges. The graph K_n has $\binom{n}{2} = \frac{n(n-1)}{2}$ edges.

Solution 1.4
There are 10 nodes of degree 3.

Solution 1.5
We have $f(5) = 12$.

Solution 1.6
Such a graph does not exist since it would have $\frac{3 \cdot 17}{2} = 25.5$ edges by the Handshaking lemma.

Solution 1.8
Hint. What are the possible degrees of a node in a graph on n nodes?

Solution 1.9
(a) 4^{10}

(b) $5 + 20 + 60 + 120 + 120 = 325$

Solution 1.12
The graph has two connected components.

Solution 1.13

a) 3

b) 5

Solution 1.17

Precisely when n is odd.

Solution 1.18

There is no such walk.

Solution 1.19

Hint. Make a complete graph with node set $\{0, 1, \ldots, 6\}$. The 21 edges represent the dominoes with unequal numbers of pips. This is an Eulerian graph.

Solution 1.20

No, No, Yes.

Solution 1.23

Hint. Remove 5 nodes to obtain a graph with 6 connected components.

Solution 1.24

An example is the cyclic graph on five nodes C_5.

Solution 1.25

Hint. Make a graph whose nodes are the binary words of length n, and two words form an edge if they differ in exactly one bit. For $n = 2$ this is a 'square', for $n = 3$ this is a 'cube', for $n = 4$ a 'hypercube' and so on.

Solution 1.26

a) *Hint.* Consider a path of maximum length. Why do the first and last node have degree 1 in the graph?

b) A tree with n nodes has $n - 1$ edges.

Solution 1.27

$n!$, $\frac{(n-1)!}{2}$.

Solution 1.29

The octahedral graph is Eulerian.

Complex Numbers

Solution 2.1

a) $15 + 5i$

b) $128 - 128i$

c) $128 + 128i$

Solution 2.2

a) $\frac{1}{2} + \frac{1}{2}i$ and $-\frac{1}{2}i$

b) $-i$ and i

c) $\frac{1}{5} - \frac{2}{5}i$

Solution 2.4

$z^2 = -\frac{1}{2} - \frac{1}{2}i\sqrt{3}$, $z^3 = 1$, $z^4 = -\frac{1}{2} + \frac{1}{2}i\sqrt{3}$, etc.
$z^{1000} = -\frac{1}{2} + \frac{1}{2}i\sqrt{3}$, $z^{1001} = -\frac{1}{2} - \frac{1}{2}i\sqrt{3}$

Solution 2.5

$-2 + i$, $1 - 3i$, $2 + 5i$, rotation over angle $\frac{1}{2}\pi$.

Solution 2.6

$2i$, $-2 + 2i$, -4, $-4 - 4i, \cdots, 16$.

Solution 2.7

$w_1 = a$, $w_2 = b$, $w_3 = (a + b) + (a + b)i$

Solution 2.8

$-5i = 5(\cos\frac{3}{2}\pi + i\sin\frac{3}{2}\pi)$;
$-\sqrt{6} + \sqrt{2}i = 2\sqrt{2}(\cos\frac{5}{6}\pi + i\sin\frac{5}{6}\pi)$;
$-3 - 4i = 5(\cos\psi + i\sin\psi), \psi = -\arctan\frac{4}{3}$.

Solution 2.9

$\cos(\frac{1}{12}\pi) = \frac{1}{4}(\sqrt{6} + \sqrt{2})$ and $\sin(\frac{1}{12}\pi) = \frac{1}{4}(\sqrt{6} - \sqrt{2})$

Solution 2.10

$w_1 = 2 + 4i$, $w_2 = -3 + i$, $w_3 = -2 - i$, $w_4 = \frac{8}{5} - \frac{6}{5}i = \frac{2}{5}(3 - 4i)$

Solution 2.12

$\sqrt{3} - i$

Solution 2.13

$z_k = \sqrt{2}\left(\cos(\frac{1}{4} + \frac{2}{5}k\pi) + i\sin(\frac{1}{4} + \frac{2}{5}k\pi)\right), k = 1, 2, 3, 4, 5$;
Note that $z_5 = 1 + i$ and that all solutions lie on a circle of radius $\sqrt{2}$ around the origin.

Optimization in networks

Solution 3.1
Determine, for each vertex v, the number of paths from s to v. Consider the vertices v in a handy order, considering t last. The solution is 16.

Solution 3.2
Verify that this proporty holds in the beginning and is preserved in each iteration.

Solution 3.3
Here you need to use that the lengths of the arcs are non-negative.

Solution 3.4
Use Exercise 3.2 for one inequality and consider the vertices on a shortest path from s to u for the other inequality.

Solution 3.8
Add up the balance equations for all $v \in V \setminus \{s, t\}$.

Solution 3.10
a) The lengths of the shortest paths are 0, 2, 4, 6, 1, 4, 5 and 8.

Solution 3.13
a) 34

b) 17

Differentialequations

Counting

Solution 5.2
$[\mathbf{5}]^0 = \{\emptyset\}$, $[\mathbf{5}]^1 = \{\{1\}, \{2\}, \{3\}, \{4\}, \{5\}\}$ en
$[\mathbf{5}]^2 = \{\{1,2\}, \{1,3\}, \{1,4\}, \{1,5\}, \{2,3\}, \{2,4\}, \{2,5\}, \{3,4\}, \{3,5\}, \{4,5\}\}$.

Solution 5.4
The probability of getting a hand with only red cards is $\frac{26!39!}{13!52!}$.
The probability of getting all cards in one suit is $\frac{13!39!}{52!}$.

Solution 5.29
2^{20} and 6^{20} respectively.

Solution 5.32
There are 2^n maps from **n** to **2**. Which 2 are non-surjective?

Solution 5.34
$\frac{3^{10}-3\cdot 2^{10}+3}{6} = 9330$

Solution 5.35
2100

Probability and Statistics

Solution 6.1

a) $A = \{MMK, MKM, KMM, \}$, $B = \{MMK, MKM, KMM, MMM\}$,
$C = \{KKK, KMK, KKM, KMM\}$ en $D = \{MKK, MMK, MKM, MMM\}$.

b) $A^c = \{KKK, MKK, KMK, KKM, MMM\}$,
$A \cup (C \cap D) = \{MMK, MKM, KMM\}$ and $A \cap D^c = \{KMM\}$.

Solution 6.3
$\frac{3}{7}$

Solution 6.5
$\frac{1}{4}$

Solution 6.7
0.1000078

Solution 6.9
0.09016

Solution 6.17
0.2

Solution 6.20
$\frac{6}{7}$

Solution 6.21

a) $L(p) = \binom{100}{63} p^{63}(1-p)^{37}$

b) $p = 0.63$

Solution 6.24

b) $\lambda = \frac{537}{576}$