COMP11120 Mathematical Techniques for Computer Science
Part 3b

Andrea Schalk
A.Schalk@manchester.ac.uk

Course webpage
studentnet.cs.manchester.ac.uk/ugt/COMP11120

November 24, 2015

4.3.3 Bayesian updating

In AI it is customary to model the uncertainty regarding a specific situation by keeping probabilities for each of the possible scenarios. As more information becomes available, for example through carrying out controlled experiments, those probabilities are updated to better reflect what is now known about the given situation. This is a way of implementing machine learning. It is also frequently used in spam detection software.

In this section we look at how probabilities should be updated.

Example 4.31. Assume you are given a bag with three socks in it. You are told that every sock in the bag are either red or black. You are asked to guess how many red socks are in the bag. There are four cases:

$$\{0, 1, 2, 3\}.$$

We model this situation by assigning probabilities to the four. At the beginning we know nothing, and so it makes sense to assign the same probability to every one of these. Our first attempt at modelling the situation is to set the following probabilities.

	0	1	2	3
P	1/4	1/4	1/4	1/4

Assume somebody reaches into the bag and draws a red sock which they hold up before returning it to the bag. No we have learned something we didn't know before: there is at least one red sock in the bag. This surely means that we should set $P0$ to 0, but is this all we can do?

The idea is that we should update *all* our probabilities based on this information. The probability $P(i)$ that we have i red socks in the bag should become

$$P(i \mid R),$$

that is, it should be the probability that there are i red socks *given that the drawn sock was red*. Bayes's Theorem helps us to calculate this number since it tells us that

$$P(i \mid R) = \frac{P(R \mid i) \cdot P(i)}{P(R)}.$$

Let us consider the various probabilities that occur in this expression:

- $P(R \mid i)$. This is the probability that a red sock is drawn, given that he total number of sock is red. This is known, and it is equal to $i/3$.
- $P(i)$. We don't know how many red socks there are in the bag, but we are developing an estimated guess for the probability, and that is what we are going to use.
- $P(R)$. This is the the probability that the first sock drawn is red, *independent from how many red socks there are.* It is not clear at first sight whether we can calculate that. The trick is to use the law of total probability, as described below.

We should pause for a moment to think about what the underlying probability space is here:[16]

[16] We cannot use the law of total probability otherwise.

In the table above we have assigned probabilities to the potentially possible numbers of red socks in the bag. But by drawing a sock from the bag we have expanded the possible outcomes:

These now have to be considered as combinations:: They consist of the number of red socks in the bag, plus the outcome of drawing a sock from the bag. We can think of these as being encoded by

- a number from 0 to 3 (the number of red socks in the bag) and
- a colour, R or B, denoting the outcome of the draw.

In other words, for the moment we should think of the sample space as

$$\{0R, 0B, 1R, 1B, 2R, 2B, 3R, 3B\}.$$

Note that our original outcome i now becomes a shortcut for the event

$$\{iR, iB\}.$$

If we draw further socks from the bag then each current outcome iC will become an event

$$\{iCR, iCB\}.$$

Returning to the probability that a red sock is drawn, $P(R)$, we can now see that this is the probability of the event

$$\{0R, 1R, 2R, 3R\}.$$

Since we can split this event into the disjoint union of

$$\{0R\} \cup \{1R\} \cup \{2R\} \cup \{3R\},$$

the law of total probability tells us that

$$\begin{aligned} P(R) &= P(R \mid 0) \cdot P(0) + P(R \mid 1) \cdot P(1) + P(R \mid 2) \cdot P(2) + P(R \mid 3) \cdot P3 \\ &= 0 \cdot 1/4 + 1/3 \cdot 1/4 + 2/3 \cdot 1/4 + 3/3 \cdot 1/4 \\ &= 1/2. \end{aligned}$$

This should be no surprise: At the moment all the events 0 to 3 are considered to be equally likely, which gives us a symmetry that makes drawing a red and drawing a black sock equally likely, based on what we know so far.

The updated probabilities are as follows:

	0	1	2	3
P	0	1/6	$2/6 = 1/3$	$3/6 = 1/2$

Note that the probability that there is just one red sock has gone down, and that the sock are all read has gone up the most.

Assume another sock is drawn, and it is another red sock. This extends the sample space in that events are now of the form

$$iRR, iRB, iBR, iBB.$$

But the way most implementations of the algorithm work is not to look at it from that point of view. Instead of keeping track of the colour of the socks drawn so far the assumption is that everything we know about what happened so far is encoded in the probabilities that describe what we know about the current situation.

This has the advantage that what we have to do now looks very similar to what we did on the previous round of updates, and it means that one can write code that performs Bayesian updating which works for every round.

So again we are seeking to update $P(i)$ by setting it to

$$P(i \mid R) = \frac{P(R \mid i) \cdot P(i)}{P(R)},$$

where now the $P(i)$ are those calculated in the previous iteration. Note that the value of $P(R)$ has changed. It is now

$$\begin{aligned} P(R) &= P(R \mid 0) \cdot P(0) + P(R \mid 1) \cdot P(1) + P(R \mid 2) \cdot P(2) + P(R \mid 3) \cdot P3 \\ &= 0 \cdot 0 + 1/3 \cdot 1/6 + 2/3 \cdot 1/3 + 3/3 \cdot 1/2 \\ &= 7/9. \end{aligned}$$

The updated probabilities are

	0	1	2	3
P	0	1/14	4/14	9/14.

If instead the second drawn sock had been black then we would have to update $P(i)$ to

$$P(i \mid B) = \frac{P(B \mid i) \cdot P(i)}{P(B)},$$

where

$$P(B \mid i) = (3 - i)/3,$$

and based on the probabilities after the first update we have

$$\begin{aligned} P(B) &= P(B \mid 0) \cdot P(0) + P(B \mid 1) \cdot P(1) + P(B \mid 2) \cdot P(2) + P(B \mid 3) \cdot P3 \\ &= 3/3 \cdot 0 + 2/3 \cdot 1/6 + 1/3 \cdot 1/3 + 0 \cdot 1/2 \\ &= 2/9. \end{aligned}$$

leading to updated probabilities of

	0	1	2	3
P	0	1/2	1/2	0.

Note that in this case, the probabilities for both cases that have been ruled out, 0 and 3, have been set to 0. Based on what we have seen in this situation, that is a red sock being drawn followed by a black one, it seems reasonable to have the probabilities for the remaining options to be equal

We can see that Bayesian updating is a way of adjusting our model of the current situation by updating the probabilities we use to judge how likely we are to be in any of the given scenarios.

The preceding example is comparatively simple, but there are two issues worth looking at in the context of this example. The first of these is already hinted at in the example: What is the underlying probability space in a case like this?

The sample space changes with the number of socks drawn—one might think of it as evolving over time. At the stage when n socks have been drawn from the bag the outcomes are best described in the form of strings

$$iX_1X_2\cdots X_n,$$

where $i \in \{0,1,2,3\}$ and $X_i \in \{R,B\}$ for $1 \le i \le n$. In other words, each outcome consists of the number of red socks, and the result of the sock draws conducted.

As we move from one sample space to the next each outcome

$$iX_1X_2\cdots X_n$$

splits into two new outcomes,

$$iX_1X_2\cdots X_nR \qquad \text{and} \qquad iX_1X_2\cdots X_nB.$$

Note that what is happening here is that the number of red socks in the bag is *fixed* for the entirety of the experiment, and so the actual probability distribution for the first probability space (before the first sock is drawn) is one which

- assigns 1 to the actual number of socks and
- 0 to all the other potential numbers of red socks under consideration.

If, for example, the number of red socks in the bag is 1 then the actual probability distribution is

	0	1	2	3
P	0	1	0	0.

Under those circumstances, the actual probabilities for the probability space based on the set of outcomes

$$\{0R, 0B, 1R, 1B, 2R, 2B, 3R, 3B\}.$$

is

	$0R$	$0B$	$1R$	$1B$	$2R$	$2B$	$3R$	$3B$
P	0	0	1/3	2/3	0	0	0	0.

What Bayesian updating is trying to do is to *approximate* this actual probability distribution for the original set of outcomes in a number of steps.

Note that since we do not know what the actual distribution does, it is at first sight surprising that with what little information we have, we can write a procedure that will succeed in approximating the correct distribution. For example, the probabilities used in the Bayesian updating applied in the example are quite different from the actual ones given above.

The underlying probability space is one where the set of events is the powerset of the sample space, but many events have the probability 0. We don't know what the distribution is, and so we cannot describe that space and use that description in our procedure.

Note that we are very careful about which events play a role in our calculation. These are of two kinds:

- The first kind consists of events whose probability we are trying to estimate. These are the outcomes from the original sample space which expand into events whose number of elements doubles each time we draw a sock.

- The second kind consists of events whose approximated probability is calculated by forming a 'weighted average' over all the events of the first kind. In other words, we are using all the data from our current approximation to give an approximated probability for those events. In the example, this is the probability of drawing a red/black sock. The aim here is to ensure that we do not introduce any additional uncertainty or bias into our calculations.

The reason Bayesian updating is so useful is that it allows us to approximate the unknown probability distribution by conducting experiments (or observing events), with very little information being required for the purpose. At each stage we treat the present approximating distribution as if it were the actual distribution, and we are relying on the idea that over time, the available information will tell us enough to ensure that our approximation gets better.

Note that it will not necessarily get better on every step—whenever a comparatively unlikely event (according to the actual distribution) occurs, our approximation is going to get worse on the next step! But there is the 'law of large numbers' which can be thought of as saying that if we keep repeating the same experiment (drawing a sock from the bag) often enough, then almost certainly we will see red socks appearing in the correct proportion.

It is worth pointing out that the idea in Bayesian updating relies on us being able to perform the same experiment more than once—if we don't put the drawn sock back into the bag the idea does not work.[17]

An interesting question is also what we can do if the number of socks in the bag is unknown. It is possible instead to consider the possible ratios between red and black socks. The most general case would require us to cope with infinite sums (since there are infinitely many possible ratios), and that is beyond the scope of this unit. Note also that there is no way of starting with a probability distribution on all possible ratios that assigns to each ratio the same probability in the way we did here, see Proposition 4.20.

If the number of possible ratios is restricted, however, then one may employ the same idea as in the example above, see Exercise 69.

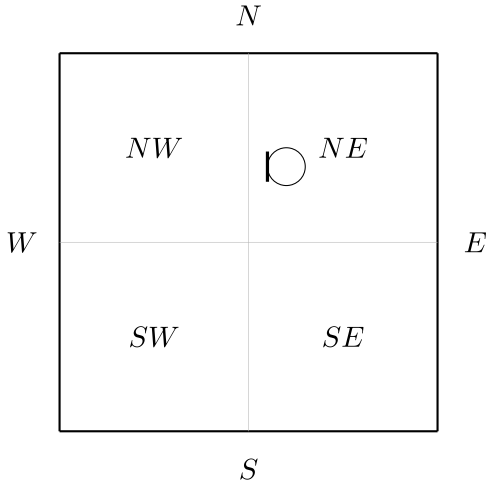
A fun example for Bayesian updating, created by my colleague Gavin Brown, can be found here: `http://www.cs.man.ac.uk/~gbrown/BTTF/`.

The following example is more complicated than the first, but still considerably simpler than its fully grown version which appears in COMP14112.

Example 4.32. We look at a toy version of an example that appears in COMP14112 next semester. Assume you have a robot. The robot is trying to work out where it is. and which way it is facing, and it may have partial, or inaccurate, information about its initial position.

[17] Of course, if the drawn sock is not returned then after three draws how many red socks were in the bag originally.

The robot can be in one of four quarters of the room, NE, NW, SE, SW and it can be facing in one of four directions N, E, S and W. Assume the room is quadratic in shape and it measures $l \times l$ metres. The robot can be anywhere in that room, and the direction in which it is facing is given by the line.



The location and orientation of the robot can then be described by a string of length 3, made up from the symbols $\{N, E, S, W\}$: The first two of the symbols give the quadrant in which the robot is, and the third its orientation. In the picture you can see a robot in state NEW

The first symbol has to be N or S, and the second symbol has to be E or W, so altogether there are

$$2 \cdot 2 \cdot 4 = 16$$

possible states the robot could be in:

$$\begin{array}{cccc} NEN & NEE & NES & NEW \\ NWN & NWE & NWS & NWW \\ SEN & SEE & SES & SEW \\ SWN & SWE & SWS & SWW \end{array}$$

The robot is trying to generate some information about the state it is in. It starts by assigning a probability to each of the 16 possibilities. This could either be a distribution that assigns the same probability to each possible outcome, or it could use partial information it has.

The robot is using a probability space where the outcomes are as in the table above. Since this is a finite set we can calculate the probability for each potential event, that is each subset of the sample space, by having a probability for each of the sixteen cases.

But there are additional pieces of information involved, and the above sample space does not tell the whole story.

The robot has a sensor, which faces in the same direction as it does. It can use that to detect how far it is to the nearest wall. Depending on whether the measured distance is smaller than $l/2$ or larger than $l/2$ the robot can then deduce whether it is in the quarter adjacent to that wall or not. However, the sensor is inaccurate, and will report a wrong distance $1/4$ of the time.

This means there is an event of taking a sensor reading, and the outcome can be that the nearest wall in the direction the robot is facing can be less than $l/2$ or more than $l/2$. So actually, we should think of the sample space as being given by strings of length four, where the last symbol tells us whether the wall is close (C) or far (F). The robot, however, is only interested in the events consisting of the outcome where the last symbol has been ignored.

So where in the table above we wrote, for example, NEN, the underlying event is really $\{NENC, NENF\}$. We call these events 'status events' because they tell us the potential status of the robot.

Querying the sensor is another event, which we can think of as getting the reading C, or getting the reading F, where the former is given by the set

$$\{NENC, NEEC, NESC, NEWC, NWNC, NWEC, NWSC, NWWC,\\ SENC, SEEC, SESC, SEWC, SWNC, SWEC, SWSC, SWWC\}.$$

It is convenient to abbreviate that event with C.

Based on what we've said above it should be clear that we know something about the conditional probabilities for sensor readings.

If the position of the robot is NEN then the nearest wall is close. The probability that the sensor reading will be C is therefore $3/4$ (because the sensor is correct 75% of the time), and $1/4$ that the reading will be F.

This means that we know that $P(C \mid NEN) = 3/4$, and similarly we can determine the conditional probabilities for C and F given the various other status events.

How should the robot update information about its status? It should apply Bayesian updating.

When the robot performs a sensor reading it should update the probability for all status events to reflect the result. If the sensor reading returns F then the probability that the robot is in, for example, square NEN should reduce, since if everything works properly the sensor should return C in that situation. The new value for the probability of NEN should be

the probability of NEN given the outcome F.

In other words, we would like to set $P(NEN)$ to

$$P(NEN \mid F).$$

To calculate that probability we can use Bayes's Theorem which tells us that

$$P(NEN \mid F) = \frac{P(F \mid NEN) \cdot P(NEN)}{PF}.$$

We know that $P(F \mid NEN)$ is $1/4$, and we know the current probability for NEN. Hence it only remains to calculate PF. For this remember that F is a shortcut for all events of the form $???F$, that is, the last symbol is F. We have a pairwise disjoint collection of events with the property that F is a subset of their union, since

$$\begin{aligned}F &= \{NENF\} \cup \{NEEF\} \cup \{NESF\} \cup \{NEWF\}\\ &\quad \cup \{NWNF\} \cup \{NWEF\} \cup \{NWSF\} \cup \{NWWF\}\\ &\quad \cup \{SENF\} \cup \{SEEF\} \cup \{SESF\} \cup \{SEWF\}\\ &\quad \cup \{SWNF\} \cup \{SWEF\} \cup \{SWSF\} \cup \{SWWF\}\\ &= \bigcup_{X\in\{N,S\},Y\in\{E,W\},Z\in\{N,E,S,W\}} \{XYZF\}.\end{aligned}$$

Hence we may use the law of total probability to deduce that

$$PF = \sum_{X\in\{N,S\},Y\in\{E,W\},Z\in\{N,E,S,W\}} P(F \mid XYZF) \cdot P(XYZF).$$

This means we now can calculate the updated probability for NEN.

In general, given a status event L (for location), the robot should update the probability for L to account for the outcome of querying the sensor, so if the outcome is C, it should set

$$P(L) \qquad \text{to} \qquad P(L \mid C).$$

More generally, if we use D (for distance) for an element of the set $\{C, F\}$ then the robot should set

$$P(L) \qquad \text{to} \qquad P(L \mid D),$$

after it has observed the event D. How do we calculate this? We are given

- the probabilities $P(L)$,
- the probabilities $P(D \mid L)$ for $D \in \{C, F\}$.

Bayes's Theorem allows us to calculate the desired probability. It tells us that for each status event L we have

$$P(L \mid D) = \frac{P(D \mid L) \cdot PL}{PD}.$$

Looking at the probabilities that appear on the right hand side of this equality, we know $P(D \mid L)$ from the basic setup (information about the robot's sensor), and we have a value for PL since that is what the robot is keeping track of. What about PD?

Remember that this is a shortcut for all events of the form $???D$, so repeating what we have done above for the case where D is equal to F we can see that we have a pairwise disjoint collection of events with the property that D is a subset of their union, since

$$\begin{aligned}D &= \{NEND\} \cup \{NEED\} \cup \{NESD\} \cup \{NEWD\}\\ &\quad \cup \{NWND\} \cup \{NWED\} \cup \{NWSD\} \cup \{NWWD\}\\ &\quad \cup \{SEND\} \cup \{SEED\} \cup \{SESD\} \cup \{SEWD\}\\ &\quad \cup \{SWND\} \cup \{SWED\} \cup \{SWSD\} \cup \{SWWD\}\\ &= \bigcup_{X\in\{N,S\},Y\in\{E,W\},Z\in\{N,E,S,W\}} \{XYZD\}.\end{aligned}$$

Hence we may use the law of total probability to deduce that

$$PD = \sum_{X\in\{N,S\},Y\in\{E,W\},Z\in\{N,E,S,W\}} P(D \mid XYZD)\cdot P(XYZD).$$

The expressions we have found here get quite unwieldy. In the version of this scenario that you see in COMP14112, the notation changes to make this easier. We show how to adapt that notation to our toy example, and give these equalities using that notation.

Instead of writing XYZ to describe the potential location and orientation of the robot, let's call the events in question

$$L_{i,j,k},$$

where

- $i \in \{0,1\}$, where 0 stands for N and 1 for S,
- $j \in \{0,1\}$, where 0 stands for E and 1 for W, and
- $k \in \{0,1,2,3\}$, where 0 stands for N, 1 for E, 2 for S and 3 for W.

Our encoding means that $L_{0,0,3}$ is equivalent to the status event NES. We can then write the update rule for the probabilities as follows: After a sensor reading resulting in D (where D is still in $\{C,F\}$), the probability

$$PL_{i,j,k}$$

should be set to

$$\begin{aligned}P(L_{i,j,k} \mid D) &= \frac{P(D \mid L_{i,j,k})\cdot PL_{i,j,k}}{\sum_{i'\in\{0,1\},j'\in\{0,1\},k'\in\{0,1,2,3\}} P(D \mid L_{i',j',k'})\cdot PL_{i',j',k'}}\\ &= \frac{P(D \mid L_{i,j,k})\cdot PL_{i,j,k}}{\sum_{i',j',k'} P(D \mid L_{i',j',k'})\cdot PL_{i',j',k'}},\end{aligned}$$

where the last line is a short-cut for the case when it is understood what values the variables i', j' and k' are allowed to take.

You may want to return to these notes when you reach this material on COMP14112 to help with working out how to update the probabilities for the more complicated scenario you will study there.

In general, Bayesian updating is performed in the situation where

- There are a number of possibilities that may apply, say Q_1, Q_2,... Q_n which are events in some probability space such that they are disjoint, and their union is the whole sample space. It is assumed that there are estimates $P(Q_i)$ for all $1 \le i \le n$.
- There is a way of collecting information about the situation, in such a way that there are a number of outcomes s_1, s_2, ..., s_m, and so that each event Q_i can be thought of as

$$Q_i = \{Q_i s_1, Q_i s_2, \ldots, Q_i s_m\}.$$

- When the outcome s_k is observed then for each Q_i its probability is updated to
$$P(Q_i \mid s_k) = \frac{P(s_k \mid Q_i) \cdot P(Q_i)}{P(s_k)},$$
where it is assumed that $P(s_k \mid Q_i)$ is known for all combinations, and where the calculation of $P(s_k)$ is performed as
$$P(s_k) = \sum_{i=1}^{n} P(s_k \mid Q_i) \cdot P(Q_i),$$
giving an overall update of $P(Q_i)$ to
$$\frac{P(s_k \mid Q_i) \cdot P(Q_i)}{\sum_{i=1}^{n} P(s_k \mid Q_i) \cdot P(Q_i)}.$$

Exercise 68. Imagine your friend claims to have a coin that will give the value of heads with a probability of either 2/3 or 3/4. You would like to work out which it is by performing Bayesian updating. Describe the first approximate probability distribution that assigns the same value to all outcomes. Now choose one of the fallowing (or both if you like). In each case we use a random device to simulate the result of throwing a coin that behaves like one of the two cases.

(a) Pick a die. Every time you want to toss our fictitious coin, roll the die. If the die shows from 1 to 4 points, assume the coin showed H, else assume it showed T. Carry out Bayesian updating in this way until you have a probability of at least 80% for one of the two given cases.

(b) Take two coins. Every time you would toss the our fictitious coin, toss both your coins. If at least one of them shows H, assume the result was H, else assume it was T. Carry out Bayesian updating in this way until you have a probability of at least 80% for one of the two given cases.

Exercise 69. Consider Example 4.31. Instead of knowing the total number of socks, all you know is that the ratio of red to black socks is an element of the following set:
$$\{1/4, 1/3, 1/2, 2/3\}.$$
What is the Bayesian update rule for this situation? Assume a black sock is drawn, followed by a red one. Starting from a probability distribution that assigns the value of 1/4 to each ratio, give the updated probabilities for each of the given ratios after each draw.

Optional Exercise 15. Assume you are asked to perform Bayesian updating in a case where there are only two possible options, and where information is gained by performing an experiment which also has two possible outcomes.

The resulting case can be described using three parameters:

- The probability p that we have assigned to the first case'
- the probability q that tells us how likely Outcome 1 is if we are in Case 1 and

- the probability r that tells us how likely Outcome 1 is if we are in Case 2.

Write down the rule for a Bayesian update in this situation. Can you say anything about subsequent calculations?

4.3.4 Independence of events

When we throw two dice, one after the other, or when we throw a coin repeatedly, we are used to a convenient way of calculating the corresponding probabilities for the outcomes.

Example 4.33. Assume we record the outcome of a coin toss with H for head and T for tails. We assume the coin is fair and so the probability for each is 1/2. If we toss the coin twice then the possible outcomes are HH, HT, TH and TT and the probability for each is 1/4. We may calculate the probability HT, that is the first coin toss $C1$ coming up H, and the second, $C2$, T as follows.

$$P((C1 = H) \cap (C2 = T)) = P(C1 = H) \cdot P(C2 = T) = 1/2 \cdot 1/2 = 1/4$$

But it is not safe to assume that for general events A and B we have that the probability of $A \cap B$ can be calculated by multiplying the probabilities of A and B, see Example 4.34.

Definition 23. Given a probability space $(S, \mathcal{E}, P)$ we say that two events A and B are **independent** if and only if

$$P(A \cap B) = PA \cdot PB.$$

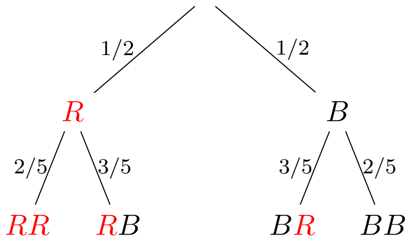
What we mean by 'independent' here is that neither event has an effect on the other. We assume that when we throw a coin multiple times then the outcome of one toss has no effect on the outcome of the next, and similar for dice. We look at this issue again in Section 29 when we have random processes which are more easily described. In particular we talk about independence for processes with a continuous probability distribution.

Example 4.34. Let us look at a situation where this is not the case. We return to the example of pulling socks from a drawer, Example 4.13. Assume you are picking a sock from a drawer without looking. If there are three black and three red socks at the start, and, say, you pick a red sock on the first draw, then the probability of finding a red sock on the second draw is changed.

The probability of drawing a red sock on the first attempt is

$$P(D_1 = R) = 1/2,$$

but what about the probability of drawing a red sock on the second attempt? Again it is best if we look at the tree that shows us how the draw progresses.



We can see that the probability of drawing a red sock on the second attempt is

$$P(D_2 = R) = \frac{1}{2} \cdot \frac{2}{5} + \frac{1}{2} \cdot \frac{3}{5} = \frac{2+3}{10} = \frac{1}{2}.$$

But we can also see from the tree that

$$P((D_1 = R) \cap (D_2 R)) = \frac{1}{2} \cdot \frac{2}{5} = \frac{1}{5},$$

which is not equal to

$$P(D_1 = R) \cdot P(D_2 = R) = \frac{1}{2}\frac{1}{2} = \frac{1}{4},$$

so (very much expectedly) the two events are not independent.

Example 4.35. A more serious example is as follows. SIDS, or ‘Sudden Infant Death Syndrome’ refers to what is also known as ‘cot death’—young children die for no reason that can be ascertained. In 1999 an ‘expert witness’ told the court that the approximate probability of a child of an affluent family dying that way is one in 8500. Since two children in the same family had died this way, the expert argued, the probability was one in 73 million that this would occur, and a jury convicted a young woman called Sally Clark of the murder of her two son, based largely on this assessment.

The conviction was originally upheld on appeal, but overturned on a second appeal a few years later. While Clark was released after three years in prison she later suffered from depression and died from alcohol poisoning a few years after that.

What was wrong with the expert’s opinion? The number of 1 in 73 million came from multiplying 8500 with itself (although 72 million would have been more accurate), that is, arguing that if the probability of one child dying in this way is

$$\frac{1}{8500},$$

then the probability of two children dying in this way is

$$\frac{1}{8500} \cdot \frac{1}{8500}.$$

But we may only multiply the two probabilities if the two events are independent, that is, if the death of a second child cannot possibly be related to the death of the first one. This explicitly assumes that there is no genetic or environmental component to SIDS, or that there may not be other circumstances which makes a second death in the same family more likely. Since then data have been studied that show that the assumption of the independence of two occurrences appears to be wrong.

While there were other issues with the original conviction it is shocking that such evidence could be given by a medical expert without anybody realizing there was a fallacy involved. I hope that this example illustrates why it is important to be clear of the assumptions one makes, and to check whether these can be justified.

Exercise 70. Assume that you have a probability space with two events A and B such that A and B are disjoint, that is $A \cap B = \emptyset$. What can you say about PA, PB and $P(A \cap B)$ under the circumstances? What can you say if you are told that A and B are independent?

Give a sufficient and necessary condition that two disjoint events are independent.

4.4 Random variables

Often when we study situations involving probabilities we want to carry out further calculations. For example, in complexity theory (see COMP11212 and COMP26120) we are frequently looking for the 'average case'—that is, we would like to know what happens 'on average'. By this one typically means taking all the possible cases, weighing each by its relative frequency (not all cases my be equally frequent), and forming the average over all those. For examples of what is meant by an 'average case' for two search algorithms see Examples 4.64 to 4.67.

But in order to carry out these operations we have to be in a situation where we can *calculate* with the values that occur. If we look at some of the examples studied then we can see that some of them naturally lend themselves to calculating averages (it is possible, for example, to ask for the average number of eyes shown when throwing two dice), and some don't (there's no average colour of a sock drawn from one of our bags of socks).

This is why people often design design questionnaires by giving their respondents a scale to choose from. The university does this as well: When you will be asked to fill in course unit questionnaires for all your units, then part of what you are asked to do is to assign numbers. 'On a scale of 1 to 5, how interesting did you find this unit.' This allows the university to form averages. But what does it mean that the average interest level of COMP11120 was 3.65 (value from 2014/15)?[18] Certainly every time you assign numbers so that you may form averages, you should think about what those numbers are supposed to mean, and whether people who are asked to give you numbers are likely to understand the same as you, and as each other, by those numbers.

Nonetheless, forming averages can be a very useful action to perform, and that is why there is a name given to functions that turn the outcomes from some probability spaces into numbers. We see below that this does not merely allow us to calculate averages but also to describe particular events without knowing anything about the events or outcomes from the underlying probability space.

[18] On these questionnaires they try to make the numbers slightly more meaningful by assigning 5 to 'agree' and 1 to 'disagree', but when does one move from one grade to another? Is it really meaningful to average those out?

4.4.1 Random variables defined

Random variables are functions that translate the elements of a sample space, that is the possible outcomes from a random experiment, to real numbers. But this translation has to happen in such a way that we know what the probabilities for the resulting numbers are, and that requires a technical definition. In order to formulate that we have to define another concept.

Definition 24. Let $(S, \mathcal{E}, P)$ be a probability space. The function

$$f \colon S \longrightarrow \mathbb{R}$$

is *measurable* if and only if for all elements r of $\mathbb{R}$ the sets

- $\{s \in S \mid fs \leq r\}$ and
- $\{s \in S \mid r \leq fs\}$

are events, that is, elements of $\mathcal{E}$.

Note that in the case where $\mathcal{E} = \mathcal{P}S$, as is often the case for applications, every function from S to $\mathbb{R}$ is measurable.

Definition 25. A **random variable** is a measurable function from the set of outcomes of a probability space to the set of real numbers $\mathbb{R}$.

If you have a probabilistic game where you might win, lose or draw, you could, for example assign a value of -1 to losing, 0 to a draw and 1 to a win. Or you could give 3 for a win, 1 for a draw, and 0 for a loss. Nothing stops you from using a completely different assignment—it is when you carry out calculations with those numbers that you should think about what they mean. It is customary in game theory to use any items (money or points) won or lost to encode the outcome of a game in a number.

If you have a probability space $(S, \mathcal{E}, P)$ such that the set of outcomes S is a subset of $\mathbb{R}$ then you have a probability variable, provided you can calculate the probabilities of all sets of the form

$$S \cap [r, \infty] \qquad \text{and} \qquad S \cap (-\infty, r],$$

where $r \in \mathbb{R}$. If we know the individual heights for a group of people, then picking a random person from that group, and looking at their height as the outcome, constitutes a random variable. The underlying function from Definition 25 is given by the embedding of S into $\mathbb{R}$, see Example 4.42.

Example 4.36. Consider Example 4.17 where we have given several probability spaces one might use to describe throwing two dice. If you pick as the space the one with outcomes

$$\{(i, j) \mid i, j \in \{1, 2, 3, 4, 5, 6\}\},$$

then the function which maps the the pair (i, j) from that set to the sum of eyes shown

$$i + j,$$

(viewed as an element of $\mathbb{R}$) is a random variable[19]

$$\begin{aligned} X\colon \{1,2,3,4,5,6\} \times \{1,2,3,4,5,6\} &\longrightarrow \mathbb{R} \\ (i,j) &\longmapsto i+j. \end{aligned}$$

Whenever we have a random variable we get an induced probability distribution. We can see that the values taken by X are exactly the elements of

$$\{2,3,4,5,6,7,8,9,10,11,12\}.$$

In order to calculate the probability that X takes the value 4 we have to calculate

$$\begin{aligned} &P(\{(i,j) \in \{1,2,3,4,5,6\} \times \{1,2,3,4,5,6\} \mid X(i,j) = 4\}) \\ &\quad = P(\{(1,3),(2,2),(3,1)\}) \\ &\quad = P(\{(1,3)\}) + P(\{(2,2)\}) + P(\{(3,1)\}) \\ &\quad = \frac{1}{36} + \frac{1}{36} + \frac{1}{36} \\ &\quad = \frac{3}{36} = \frac{1}{12}. \end{aligned}$$

This is usually written in the shortcut notation of

$$P(X = 4).$$

But note that since we have translated our outcomes into real numbers we may also ask, for example, what the following probabilities are:

$$\begin{aligned} &P(X \leq 4) \\ &P(X \leq -4) \\ &P(X \leq 5.5) \\ &P(X \geq 10) \end{aligned}$$

The events described here do not look as if they have anything to do with the original experiment of drawing two dice, but since we have translated the outcome from that experiment into real numbers we may construct such events.

These probabilities can be calculated as follows:

- $P(X \leq 4)$. This can be calculated by splitting it into the possible outcomes satisfying that property.

  $$\begin{aligned} P(X \leq 4) &= P((X=2) \cup (X=3) \cup (X=4)) \\ &= P(X=2) + P(X=3) + P(X=4) \\ &= \frac{1}{36} + \frac{2}{36} + \frac{3}{36} = \frac{1}{6}. \end{aligned}$$

- $P(X \leq -4)$. Clearly there are no possible outcomes which satisfy this condition, so this probability is 0.

[19] Random variables are typically named using capital letters from the end of the alphabet.

- $P(X \leq 5.5)$. This works similar to the first calculation.

$$\begin{aligned} P(X \leq 5.5) &= P((X=2) \cup (X=3) \cup (X=4) \cup (X=5)) \\ &= P(X=2) + P(X=3) + P(X=4) + P(X=5) \\ &= \frac{1}{36} + \frac{2}{36} + \frac{3}{36} + \frac{4}{36} = \frac{10}{36} = \frac{5}{18}. \end{aligned}$$

- $P(X \geq 10)$. This is similar to the previous example.

$$P(X \geq 10) = P(X=10) + P(X=11) + P(X=12) = \frac{3+2+1}{36} = \frac{1}{6}.$$

Below we describe how this works for arbitrary random variables.

In general given a random variable X on a sample space $(S, \mathcal{E}, P)$, and real numbers r and r', we define

- $P(r \leq X \leq r') = P\{s \in S \mid r \leq X(s) \leq r'\}$
- $P(r \leq X) = P\{s \in S \mid r \leq X(s)\}$, $P(r < X) = P\{s \in S \mid r < X(s)\}$,
- $P(X \leq r') = P\{s \in S \mid X(s) \leq r'\}$, $P(X < r) = P\{s \in S \mid X(s) < r\}$..

In other words, if we are given an interval in $\mathbb{R}$ then in order to determine its probability we ask for the probability of the event given by all those elements of the original sample space which are mapped into that interval. Note that the sets that appear on the right hand side of the equal sign appear in the definition of measurability. This ensures that in the original probability space we have a probability for the set in question.

Definition 26. A random variable X is **discrete** if and only if its range is a countable[20] subset of $\mathbb{R}$. A random variable which is not discrete is **continuous**.

While there is a mathematical theory that allows the discrete case to be treated at the same time as the continuous one, covering the mathematics that allows this is beyond the scope of this course unit. In what follows the discrete case is frequently treated separately. In the text, and in some of the results given, some guidance is given on how the discrete case may be seen as a special case of the continuous one.

Exercise 71. Let X be a discrete random variable with range

$$\{r_i \in \mathbb{R} \mid i \in \mathbb{N}\}.$$

Show that

$$\sum_{i \in \mathbb{N}} P(X = r_i) = 1.$$

Exercise 72. Consider the event where a fair coin is thrown four times. Consider the random variable X which records the number of heads thrown. Calculate the following probabilities.

[20] What this means formally is discussed in Section 5.2. Every finite set is countable, and you may think of countable sets as sets which are 'as most as large as $\mathbb{N}$'.

(a) $P(X = 2)$,

(b) $P(X \leq 3)$,

(c) $P(X \leq \pi)$,

(d) $P(X \geq 3)$,

(e) $P(X \geq 10)$,

(f) $P(X < -1)$.

(g) $P((X = 1) \cup (X = 3))$.

(h) $P(X \text{ is even})$. Note that this only makes sense because we know the range of X is a subset of $\mathbb{N}$.

4.4.2 A technical discussion

What follows is a fairly technical discussion regarding why we can define probabilities in the way outlined above. The material from this subsection is not examinable, and you should feel free to skip it when reading the notes.

Optional Exercise 16. Show that if $(S, \mathcal{E}, P)$ is a probability space, and $f\colon S \longrightarrow \mathbb{R}$ is a measurable function, then given r and r' in $\mathbb{R}$ we have that

$$P(r \leq X \leq r')$$

is an event; in other words, the probability that appears above is always defined.

Proposition 4.37. *Let $(S, \mathcal{E}, P)$ be a probability space, and let $f\colon S \longrightarrow \mathbb{R}$ be a function. If f is measurable then for every B in the Borel σ-algebra $\mathcal{E}_B$ on $\mathbb{R}$ we have that*

$$\{s \in S \mid fs \in B\}$$

is an event, that is an element of $\mathcal{E}$.

Proposition 4.38. *Let $(S, \mathcal{E}, P)$ be a probability space, and let be $f\colon S \longrightarrow \mathbb{R}$ a measurable function. For $i \in \mathbb{N}$ let I_i be an interval in $\mathbb{R}$ such that the the I_i are pairwise disjoint. If we define, for $i \in \mathbb{N}$,*

$$E_i = \{s \in S \mid fs \in I_i\},$$

then the E_i are pairwise disjoint and so

$$\begin{aligned} P(\bigcup_{i \in \mathbb{N}} I_i) &= P(\bigcup_{i \in \mathbb{N}} E_i) \\ &= \sum_{i \in \mathbb{N}} PE_i \\ &= \sum_{i \in \mathbb{N}} PI_i. \end{aligned}$$

As a consequence we get the following result:

Theorem 4.39. *Let $(S, \mathcal{E}, P)$ be a probability space, and $f\colon S \longrightarrow \mathbb{R}$ a measurable function. Then a probability space is given by $\mathbb{R}$, the Borel σ-algebra $\mathcal{E}_B$, and the probability distribution*

$$\begin{aligned} \mathcal{E}_B &\xrightarrow{\qquad\qquad} [0,1] \\ E &\longmapsto P(\{s \in S \mid fs \in E\}). \end{aligned}$$

Alternatively we may restrict ourselves to the range of the underlying measurable function to define a probability space—in this way we remove those parts of $\mathbb{R}$ which are assigned a probability of 0.

Theorem 4.40. *Let $(S, \mathcal{E}, P)$ be a probability space, and $f\colon S \longrightarrow \mathbb{R}$ a measurable function. Then a probability space is given by the range T of f, the σ-algebra*

$$\{E \cap T \mid E \in \mathcal{E}_B\},$$

and the probability distribution

$$\begin{aligned} \{E \cap T \mid E \in \mathcal{E}_B\} &\xrightarrow{\qquad\qquad} [0,1] \\ B &\longmapsto P(\{s \in S \mid fs \in B\}). \end{aligned}$$

4.4.3 Calculating probabilities for random variables

Above we define probabilities for random variables. You can think of them as translating the original outcomes into numbers in such a way that we can look at the probabilities of subsets of $\mathbb{R}$ instead of events from the original space. Because we do have a probability distribution (see Theorem 4.40) all the usual results (see Sections 4.2.5 and 4.3.2) hold.

One of the advantages of considering random variables is that it allows us to compute probabilities with very little information, in particular without knowing too much about the original probability space.

Example 4.41. Assume that X is a random variable and that we know that

- $P(X = 1) = 1/2$,
- $P(X = 2) = 1/4$, and
- $P(X \geq 2) = 1/2$.

This is enough to allow us to calculate, for example

- $P(X = 0) = 0$,
- $P(X \leq .5) = 0$,
- $P(X > 2) = 1/4$.

We first of all note that the probabilities of $(X = 1)$ and $(X \geq 2)$ add up to 1, which means that the probability of any set which is a subset of

$$\mathbb{R} \setminus (\{1\} \cup [2, \infty))$$

must be 0. This explains the first two results.

Secondly we can see that $(X \geq 2) = (X = 2) \cup (X > 2)$ and that the two sets we form the union over are disjoint, which means we have

$$1/2 = P(X \geq 2) = P(X = 2) + P(X > 2) = 1/4 + P(X > 2),$$

from which we may deduce the last result.

Note in particular that we do not know whether X is a discrete or a continuous random variable! The fact that it has non-zero probability for being equal to 1 and 2 might suggest it is the former, but it could still be the case that the behaviour is continuous for values beyond 2.

See Example 4.54 for a way of picturing some of this information.

Exercise 73. Assume you have a probability space with outcomes

$$\{s_1, s_2, s_3, s_4, s_5\},$$

and that the following hold:

- The outcomes s_1 and s_2 are equally likely.
- The outcomes s_3, s_4 and s_5 are equally likely.
- The outcomes of the first kind are three times as likely as the outcomes of the second kind.

A random variable is given by the function X defined by

$$\begin{array}{rcl} X\colon \{s_1, s_2, s_3, s_4, s_5\} & \longrightarrow & \mathbb{R} \\ x & \longmapsto & \begin{cases} 1 & x = s_1 \text{ or } x = s_2 \\ 3 & x = s_3 \\ 5 & \text{else.} \end{cases} \end{array}$$

Compute the following:

(a) $P(X \leq 1.5)$,

(b) $P(X \geq 3)$,

(c) $P(2.5 \leq X \leq 3.2)$,

(d) $P(X \geq 6)$.

Example 4.42. Assume that I've been given the measures in height of a group of people, where the measures have been carried out with great precision. We can think of this as a random variable, where the random experiment is to pick a person (randomly) from the group, and the probability distribution is given by the measured heights. One might want to treat this like a continuous random variable if a lot of people are involved.

But maybe for my purposes I only care about how many people I have in much loser categories. Assume that I'm only interested in the following categories:

- People who are at most than 140 cm tall or
- people who are from 140 to 160 cm tall or

- people who are from 160 to 180 cm tall and
- people who are taller than 180 cm.[21]

I would like to count how many people out of the group belong to each category to construct a probability space which allows me to work out the probabilities that a randomly chosen person from that group falls into a particular category.

I do this by counting for each category how many people fall into it, and dividing by the total number of people in the group. I can turn this into a random variable for example by mapping

- people of the first category to 140,
- people of the second category to 150,
- people from the third category to 170 and
- people from the final category to 180.

But note that if I use this random variable to compute an average height for all people I get a very misleading result!

We can see from the preceding example that it can be useful to take a given random variable and use a function on its possible values (here mapping actual heights to representative of some height categories) to get a different (but related) random variable that better expresses whatever we are concerned with.

Example 4.43. In the robot lab exercise in COMP14112 (compare Example 4.32) the orientation of the robot is viewed as an angle from 0 to 360 degrees. The orientation is a continuously varying entity, but for the purpose of that exercise this is split into 100 parts of equal size, creating a discretely valued random variable, which makes it easier to carry out calculations (Bayesian updating in that case).

Most probability spaces with a continuous probability distribution have outcomes that are given by real numbers, and then we immediately have a random variable to compute with. There are, however, other cases. The following example is based on a probability space where the sample set is a subset of $\mathbb{R} \times \mathbb{R}$. We can think of this as a 'two-dimensional' random experiment. In order to get a random variable we have to translate these outcomes to elements of $\mathbb{R}$, and the example suggests how this might work in a particular situation.

Example 4.44. Assume that we have a probability space with underlying sample set $\mathbb{R} \times \mathbb{R}$, and a set of events based on the Borel σ-algebra, where all sets of the form

$$[r, r'] \times [s, s'],$$

[21] Clearly one has to think about what should happen on the borderline—let's assume here this belongs to the lower height category.

for $r, r', s, s' \in \mathbb{R}$ are events. Further assume we have a probability density function given by

$$\begin{aligned} \mathbb{R} \times \mathbb{R} &\longrightarrow \mathbb{R}^+ \\ (x, y) &\longmapsto \frac{e^{-1/2(x^2+y^2)}}{2\pi}, \end{aligned}$$

which might describe the probability of finding an animal at a particular point in a two-dimensional space centred on the animal's den.

For some purposes we might be interested only in the distance of the animal's location from the den, in which case we would like to apply the function

$$\begin{aligned} \mathbb{R} \times \mathbb{R} &\longrightarrow \mathbb{R} \\ (x, y) &\longmapsto \sqrt{x^2 + y^2} \end{aligned}$$

to obtain a random variable which gives us that distance. We have taken two-dimensional data and turned it into a random variable, which requires the restriction to just one dimension.

Note in particular that if a random variable X has a finite range. then we can treat it in much the same way as we did a probability space with a finite sample set where every set of the form $\{s\}$, for $s \in S$, is an event.

Example 4.45. In Example 4.42 the random variable X had four possible values, namely

$$\{140, 150, 170, 180\}.$$

Assume that the probabilities that a randomly chosen person from the monitor group fits into each category is given by the following table:

	140	150	170	180
P	1/2	1/4	1/8	1/8

Then in order to calculate the probability for a particular interval all I need to know is which of those four values appear in it. In order to calculate

$$P(X < 160)$$

I just have to see that

$$(X < 160) = (X = 140) \cup (X = 150),$$

and so

$$P(X < 160) = P(X = 140) + P(X = 150) = 1/2 + 1/4 = 3/4.$$

In general, if a random variable X has a finite range, say

$$\{r_1, r_2, \dots, r_n\} \text{ in } \mathbb{R}$$

then given an interval I we have that

$$P(X \in I) = P(I \cap \{r_1, r_2, \dots, r_n\}) = \sum_{i \in \{1,2,\dots,n\}, r_i \in I} P(X = r_i).$$

In other words we add up all the probabilities for those elements r_i of the range of X which are elements of I.

Note that if we have a random variable X, and combine it with a measurable function from $\mathbb{R}$ to $\mathbb{R}$ we get another random variable, see Examples 4.42 and 4.43.

Example 4.46. Assume that I am in the situation of Example 4.42, but now I am only interested whether somebody is below 160 cm or above.

Then I can take my previous random variable, which produced the possible values

$$\{140, 150, 170, 180\},$$

and combine it with the function

$$\begin{aligned} \mathbb{R} &\longrightarrow \mathbb{R} \\ x &\longmapsto \begin{cases} 150 & x \leq 160 \\ 170 & \text{else.} \end{cases} \end{aligned}$$

to get a new random variable whose only values are

$$\{150, 170\}.$$

For a slightly more interesting example consider the following.

Example 4.47. Assume we have a random variable X that has a range of values

$$\{-n, -(n-1), \ldots, -2, -1, 0, 1, 2, \ldots, n-1, n\}.$$

Maybe for some purposes we are not interested in the values as such, but only in how far distant they are from the mid-point, 0. This might be because we are only interested in the difference between some value and 0, but not whether that difference is positive or negative (compare also Definition 31).

By composing the random variable with the absolute function

$$\begin{aligned} |\cdot| \colon \mathbb{R} &\longrightarrow \mathbb{R} \\ x &\longmapsto |x|, \end{aligned}$$

we obtain a new random variable Y which takes its values in the set

$$\{0, 1, \ldots, n\}.$$

To calculate probabilities for Y we have to know that

$$P(Y = i) = \begin{cases} P(X = i) + P(X = -i) & 0 \leq i \leq n \\ 0 & \text{else.} \end{cases}$$

In general we may state this idea as follows.

Proposition 4.48. *If X is a random variable and $f \colon \mathbb{R} \longrightarrow \mathbb{R}$ is a measurable function then*

$$f \circ X$$

is a random variable.

Exercise 74. Recall the unfair die from Exercise 51. Take as a random variable X the number of eyes shown. Calculate the following.

(a) $P(X \leq 3)$,

(b) $P(X \geq 5)$,

(c) $P(4 \leq X < 6)$,

(d) $P(X \leq \pi)$,

(e) $P(X \geq 7)$.

Now assume that the random variable Y is given by the sum of the eyes shown by two such dice. Calculate the following.

(f) $P(Y \leq 4.5)$,

(g) $P(Y \geq 11.5)$.

Finally assume that we have the random variable Y and we compose it with the following function:

$$\begin{aligned} f\colon \mathbb{R} &\longrightarrow \mathbb{R} \\ x &\longmapsto (x-7)^2. \end{aligned}$$

Calculate the following.

(h) $P(f \circ Y \geq 6)$,

(i) $P(f \circ Y \leq .5)$.

4.4.4 Probability mass functions and cumulative distributions

As we have seen above when we have a random variable we typically do not need to worry about all real numbers but only those that appear in the range of the random variable. We can give a graphical presentation of how the probabilities is spread over that range. It is the equivalent to a probability density function for the case where we have discrete values.

Definition 27. Let X be a random variable with a countable range, say

$$\{r_i \mid i \in \mathbb{N}\}.$$

The **probability mass function (pmf) for** X is given by

$$\begin{aligned} \{r_1, r_2, \ldots, r_n\} &\longrightarrow [0,1] \\ r_i &\longmapsto P(X = r_i). \end{aligned}$$

It is appropriate to think of a probability mass function as the discrete version of a probability density function.

Example 4.49. For the random variable that consists of assigning the total number of eyes to the throw of two dice, see Example 4.36, the pmf is given by

2	3	4	5	6	7	8	9	10	11	12
$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

Example 4.50. If we throw a coin three times and use the random variable that arises from assigning to each output the number of heads that appear then we get the following pmf:

0	1	2	3
1/8	3/8	3/8	1/8.

The following is a version of Proposition 4.18 for random variables with finite range.

Corollary 4.51. *Let X be a random variable with finite range, say T, and pmf p. Then there is a unique probability space $(T, \mathcal{P}T, P)$ with the property that for all elements $t \in T$ we have*

$$P\{t\} = pt.$$

For this space we may calculate for all subsets E of T that

$$PE = \sum_{t \in E} pt.$$

Proof. This is an application of Proposition 4.18. □

What this means is that if we have a probability mass function then we have a uniquely determined probability space, and so for a random variable with finite range all we need to understand the situation is the pmf. For this reason some people call a probability mass function a probability distribution.

In Section 4.2.4 the idea of a cumulative probability distribution is introduced. At this point we are ready to define that concept generally.

Definition 28. Given a random variable X the **cumulative distribution function (cdf) for** X is the function

$$\mathbb{R} \longrightarrow [0, 1]$$

which assigns, for $t \in \mathbb{R}$,

$$t \longmapsto P(X \leq t.)$$

We are using here the fact that the real numbers are ordered and so it makes sense to ask for the probability that the random variable is at most some given number. In particular we can meaningfully draw the graph of this function and visualize the probability distribution in a way that we cannot do with any other form.

Example 4.52. An example for the continuous case is given in Section 4.2.4 in the form of Example 4.19.

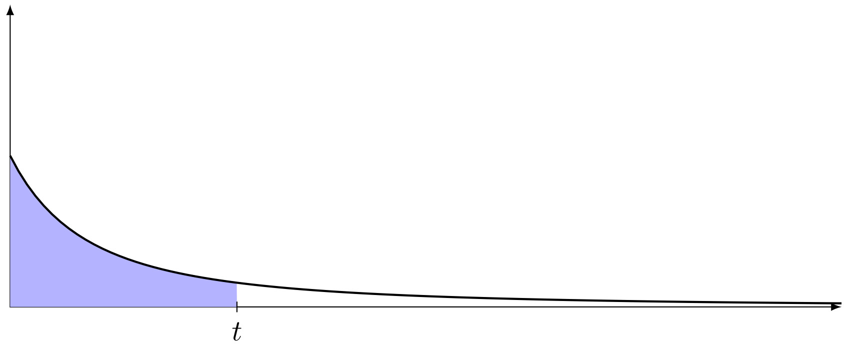
Recall that in the case of a continuous random variable X with range contained in an interval $I \subseteq \mathbb{R}$ the probability distribution is given in the form of a probability density function, say

$$g\colon I \longrightarrow \mathbb{R}^+.$$

There are two cases.

- If the interval is of the form $(-\infty, r')$, $(-\infty, r']$ or $\mathbb{R}$ then the cdf for X is given by
$$P(X \leq t) = \begin{cases} \displaystyle\int_{-\infty}^{t} gxdx & t \leq r' \\ 1 & \text{else.} \end{cases}$$
- If the interval is of the form (r, r'), $(r, r']$, (r, ∞) $[r, \infty)$ then the cdf for X is given by
$$P(X \leq t) = \begin{cases} 0 & t \leq r' \\ \displaystyle\int_{r}^{t} gxdx & r \leq t \leq r' \\ 1 & \text{else.} \end{cases}$$

  In the case below I is $[0, \infty)$, and we calculate the probability that X is below t.

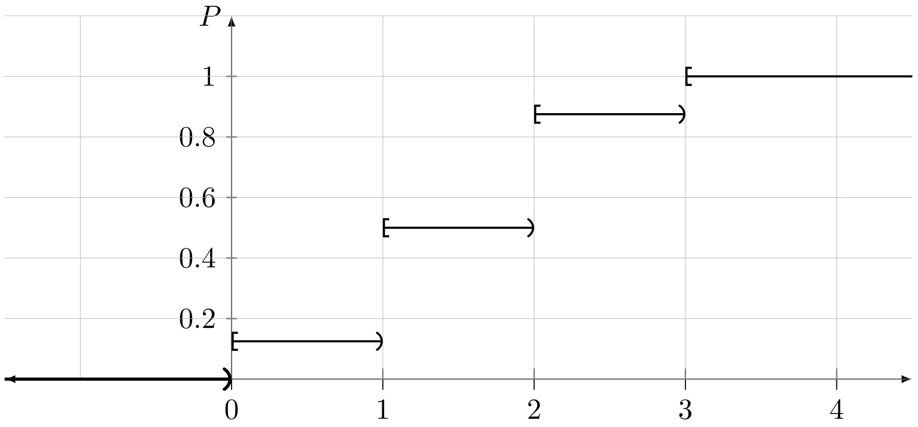


When we have a random variable which can take a finite number of values we have to draw a non-continuous function, and you may find this a bit odd at first. Look at the following example to see how that works.

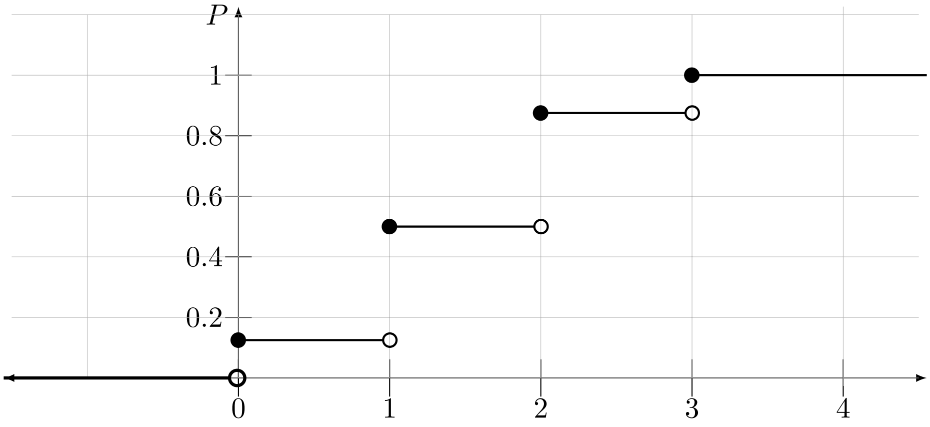
Example 4.53. If we look at the situation from Example 4.50 where the pmf is described in the table

0	1	2	3
1/8	3/8	3/8	1/8.

then the corresponding cdf can be drawn as follows.



Note that when drawing discontinuous functions like the above we have to specify what the value at, say 2, is, the lower or the upper of the two lines. The convention used in the picture above is to use the interval notation, so that [and] mean that the point at the end of the line belongs, and (as well as) mean that it doesn't. An alternative way of drawing the same function is to use the following convention:



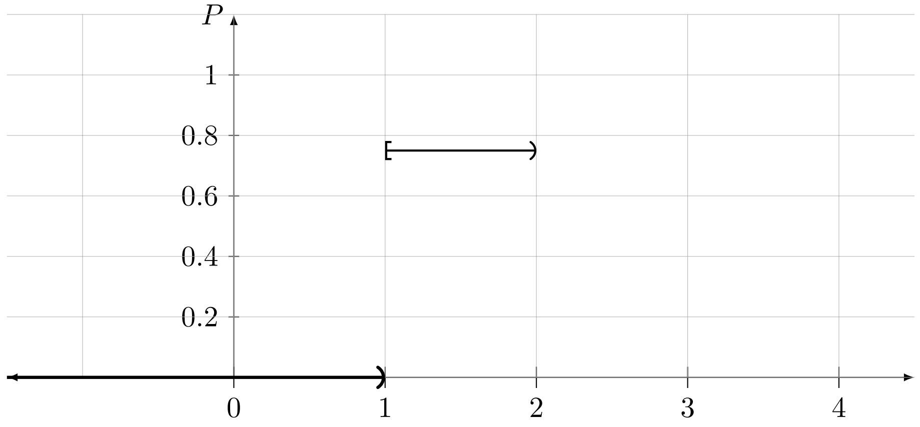
In the picture above the filled circle indicates that the endpoint of the line is included, and the unfilled circle that it is excluded.

In both pictures we can see that the functions jumps to a higher accumulated probability as the next possible value of the random variable is reached. The probability of fewer than 0 heads is 0, the probability of getting at least 0, but fewer than 1 heads is 1/8, and so on.

Example 4.54. We return to Example 4.41, where the information given is that X is a random variable and the following is known about its probability distribution is the following:

- $P(X = 1) = 1/2$,
- $P(X = 2) = 1/4$, and
- $P(X \geq 2) = 1/2$.

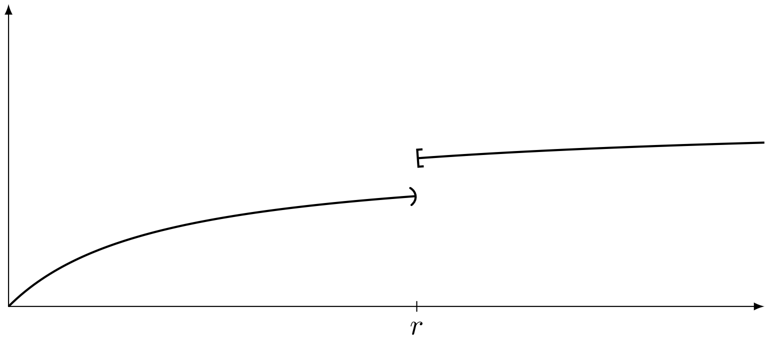
This is sufficient to be able to draw some of the cdf, but there is uncertainty:



We know that the probability is 0 until the value 1 is reached, and that it rises to .5 at that point, and rising further to .75 from 2. What we don't know is when it takes on the value 1 or which values it take between .75 and 1..

Note that the *derivative* of a cumulative distribution function is the corresponding probability density function (which in the discrete case is the corresponding probability mass function). We have no time to discuss here exactly how the derivative is formed in the discrete case.

However there is something that is easy to see. Assume you have a random variable whose cdf F makes a jump as in the following picture.



Then it has to be the case that the probability of the random variable at the point where the jump is concerned is the difference of the two values, that is

$$P(X = r) = F(r) - \lim_{n\to\infty} F(r - \frac{1}{n}).$$

Proposition 4.55. *Let X be a random variable. If P is its cumulative distribution function then its derivative is the corresponding probability density (mass density) function.*

Exercise 75. For Exercise 51 consider the random variable given by the number of eyes the die shows. Give its pmf, and draw a graph for its cdf.

Then do the same with the random variable $f \circ Y$ from the same exercise.

Exercise 76. Assume that teams are regularly playing in a 'best out of 7' series against each other, compare Exercise 50. f one assumes that performance is determined by a random process one may declare a random variable to give the number of matches Team A wins in a given series.

(a) Describe the function that underlies this random variable by writing down a mathematical function that carries out the required assignment. *Hint: You need to have a description of the set of outcomes to do this.*

(b) For the case where team A is equally matched by Team B, give the pmf and draw a graph for its cdf.

(c) Now assume that it is known that A wins the first match. We can now look at the random variable X conditional on this event. Describe the pmf and cdf for the resulting random variable.

4.4.5 Conditional probabilities for random variables

Recall that there is no example for conditional probabilities in the continuous case in Section 4.3 above. The reason for this is that describing the probability density function in the general case requires mathematical techniques beyond this course unit.

However, this is feasible once we restrict ourselves to random variables, where we know that the outcomes are elements of $\mathbb{R}$. The definition does not change.

The conditional probability density function

Recall that given two events A and B, where $PB \neq 0$, the conditional probability of A given B is defined as

$$P(A \mid B) = \frac{P(A \cap B)}{PB}.$$

If X is a random variable with probability distribution function F then we may define the conditional distribution of X given the event B as

$$P(X \leq r \mid B) = \frac{P((X \leq r) \cap B)}{PB}.$$

There is a conditional probability density function, which is once again the derivative of the distribution. The probability that the conditionally distributed random variable falls into a given interval is then the integral over that derivative over the given integral.

Example 4.56. If X is a discrete random variable with pmf p then given an event B with $PB \neq 0$ we can calculate the pmf q of the random variable

$$(X \mid B) \qquad\qquad X \qquad \text{given} \qquad B,$$

by setting, for r in the range of X,

$$qr = \begin{cases} \frac{P(X=r)}{PB} & r \in B \\ 0 & \text{else.} \end{cases}$$

In other words, if we know that B happens, and r is a possible result of X not in B, then it has the probability 0, and otherwise the probability is adjusted by dividing through PB as expected.

Example 4.57. For example, let X be a random variable with range $\mathbb{R}$ and probability distribution f, let r be in $\mathbb{R}$, and assume that B is the event

$$B = (X \leq r).$$

We might then wonder how to calculate, for $s \in \mathbb{R}$,

$$P(X \leq s \mid B).$$

If we know the probability density function g for the resulting random variable we can calculate this probability as

$$\int_{-\infty}^{s} gxdx.$$

That probability density function is given by

$$\begin{aligned} g\colon \mathbb{R} &\longrightarrow \mathbb{R}^+ \\ x &\longmapsto \begin{cases} \frac{fx}{\int_{-\infty}^{r} fx dx} & x \leq r \\ 0 & \text{else} \end{cases} \end{aligned}$$

If we assume that the value of X is below r then the probability that the value is above should indeed by 0. Also note that we are dividing by

$$PB = P(X \leq r) = \int_{-\infty}^{r} fx dx,$$

as expected from the definition of conditional probability.

Note that it is also possible to perform Bayesian updating for random variables: In the case of a discrete random variable, the update procedure is just as described in Section 4.3.3.

If the random variable is continuous then instead of updating the pmf by adjusting all the individual values we have to update the probability density function. Spelling out the resulting definition of the new probability density function goes beyond this course unit.

One random variable depending on another

The material in this subsection is not examinable. You may want to return to it if you ever have to cope with a situation where one random variable depends on another.

Recall Example 4.27, where we were wondering about how to describe the probability density function for the location a fox whose behaviour is influenced by the location of a lynx (if the latter is close enough).

What we have there is one random process, describing the movements of the fox, conditional on another random process, namely the movement of the lynx.

We can only do this in the situation where we have a *jiont distribution*, that is, a probability distribution, or a density function/pmf, that describes the combined probability.

It is then the case that if f is the joint density function for random variables X and Y, we can derive density functions for X and Y, namely

- The probability density function for X is[22]

$$\int_{\infty}^{\infty} f(x, y) dy,$$

- while that for Y is

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

[22] You can calculate with these integrals by treating the other variable as if it were a parameter, that is, you integrate the first expression for y and treat x as if it was a number. You swap the treatment of the two variables for the second expression.

In this situation we can look at the case of the density function g for X given $(Y = s)$, for some $s \in \mathbb{R}$. We get

$$gx = \frac{f(x,s)}{\int_{-\infty}^{\infty} f(x,s)dy}.$$

If instead we are interested in the probability distribution for X given

$$(s \leq Y \leq s'),$$

we have

$$P(X = r \mid s \leq Y \leq s') = \frac{\int_{-\infty}^{r} \left(\int_{s}^{s'} f(x,y)dy \right) dx}{\int_{s}^{s'} \left(\int_{-\infty}^{\infty} f(x,y)dx \right) dy}.$$

If X and Y are discrete random variables then we cam look at their joint pmf. This is a function that, given

- a value r from the range of X and
- a value s from the range of Y,

returns the probability

$$P(X = r \text{ and } Y = s).$$

Independent random variables

When we have two random variables which are independent from each other it becomes easier to calculate with both.

Definition 29. Two random variables X and Y are *independent* if and only if it is the case that for all elements of the Borel σ algebra E and E' we have that

$$P(X \in E \text{ and } Y \in E') = P(X \in E) \cdot P(Y \in E').$$

In particular this means that

- if X is a random variable with density function f and
- Y is a random variable with density function g then

the joint density function for X and Y is given by

$$(x,y) \longmapsto fx \cdot gy.$$

This is important when we wish to look at situation where we have several random variables, for example the failure of a number of pieces of equipment, where we assume that the failure of one is independent from the failure of the others. Note that this assumption is only justified if we can exclude factors that would affect more than one piece of equipment, such as a power surge at some location.

Exercise 77. Assume that you are tasked by your boss with making sure that you have enough servers that the probability of no server being currently online is at most 1% for the entire year. Because you are able to place your servers at separate locations you are allowed to assume that one server failing will have no effect on the other servers.

(a) Assume that the chance of one of your servers failing in a given year is .05. How many servers do you need to comply with your boss's demand? How much safety do you get out of an extra server?

(b) Assume that the probability of one of your servers failing has the probability density function[23]

$$x \longmapsto 2\frac{x^2}{365^3} \ ,$$

where we need to consider it from $x = 0$ to $x = 365$ to cover the year. How many servers do you have to buy and install to ensure that the probability of one of them having failed is below the threshold you were given?

4.4.6 Expected value and standard deviation

One of the motivations for introducing the notion of random variables is the ability to form averages.

Expected value

Example 4.58. Returning to the example of the number of heads when tossing a coin three times, Example 4.50, you may wonder what the average number of heads might be. This case is so simple that you can probably guess the answer, but in more complicated situations you will want to carry out analogous calculations. If we weigh each possible outcome by its probability then this number is given by

$$0 \cdot \frac{1}{8} + 1 \cdot \frac{3}{8} + 2 \cdot \frac{3}{8} + 3 \cdot \frac{1}{8} = \frac{0+3+6+3}{8} = \frac{12}{8} = \frac{3}{2},$$

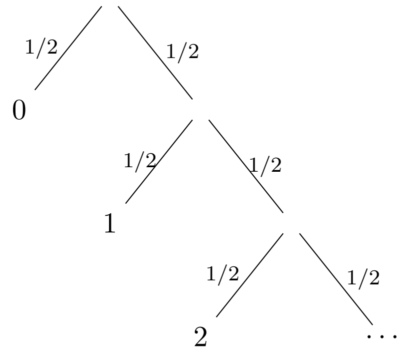
so on average the number of heads is 1.5, which in this simple case you may have been able to guess. Note that if you wanted to bet on the outcome of this experiment then it does not make sense to bet the expected value since it cannot occur.

We look at more interesting examples. Note that solving the following two examples requires knowledge beyond this course unit—it is included here to give you an idea of how powerful the idea is.

Example 4.59. Assume that we have strings which are generated in a random way, in that after each key stroke, with a probability of 1/2, another symbol is added to the string. We would like to calculate the average length of the strings so created. Before we can do this we have to specify when the random decision starts: Are all strings non-empty, or is there a chance that no symbol is ever added? We go for the latter case, but the calculation for the former is very similar.

As is often the case when picturing a step-wise process we can draw a tree that describes the situation. At each stage there is the random decision whether another symbol should be added or not. We give the length of each generated string.

[23] I'm not claiming this is a realistic density function, but hopefully it's not too bad to calculate with.



What this means is that we have a random variable, namely the length of the generated string, and we can see that its probability mass function has the first few values given by the following table.

0	1	2	3	4
1/2	1/4	1/8	1/16	1/32

In other words the pmf is given by the function

$$\begin{aligned} \mathbb{N} &\longrightarrow \mathbb{R} \\ n &\longmapsto \frac{1}{2^{n+1}}. \end{aligned}$$

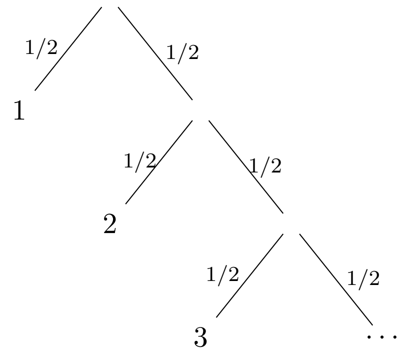
What is the average string length? The idea is that we should give each possible length the probability that it occurs. This means that we should calculate

$$0 \cdot \frac{1}{2} + 1 \cdot \frac{1}{4} + 2 \cdot \frac{1}{8} + \cdots = \sum_{n \in \mathbb{N}} n \cdot \frac{1}{2^{n+1}}.$$

With a bit more mathematics than we can teach on this unit it may be calculated[24] that this required number is 1. So certainly when producing strings in this way we don't have to worry about there being a lot of long ones! But note that we have described a process for producing potentially infinite strings (with a probability of 0), and the power of the methods we use here is such that we can still calculate the average.

We say more about how to cope with situations where we have to compute an infinite sum in Section 4.4.6.

Example 4.60. Assume we are tossing a coin until we get heads for the first time, and then we stop (compare Exercise 53). We wonder what the average number of coin tosses is. Again it makes sense to draw a tree.



[24] In mathematical parlance, we have defined a series whose limit is 1.

This is quite similar to the previous example! The pmf for this random variable is given by

$$\begin{aligned} \mathbb{N} &\longrightarrow \mathbb{R} \\ n &\longmapsto \frac{1}{2^n}. \end{aligned}$$

The expected value is

$$\sum_{n\in\mathbb{N}} n \cdot \frac{1}{2^n} = 2.$$

Again calculating such expected values is not part of this unit, but it gives you one motivation why mathematicians care about what happens if infinitely many numbers are added up.

We say more about how to cope with situations where we have to compute an infinite sum in Section 4.4.6.

What is it that we have calculated in these examples?

Definition 30. Let X be a random variable with probability density function p. Then the **expected value of** X, $E(X)$, is given by

$$E(X) = \int_{-\infty}^{\infty} x \cdot p(x)dx.$$

Note that if X is a discrete random variable with range

$$\{r_i \mid i \in \mathbb{N}\},$$

then its expected value is

$$E(X) = \sum_{i\in\mathbb{N}} r_i \cdot P(X = r_i).$$

This means that if X is a discrete random variable with finite range

$$\{r_1, r_2, \dots, r_n\},$$

then its expected value is

$$\begin{aligned} E(X) &= r_1P(X = r_1) + r_2P(X = r_2) + \cdots + r_nP(X = r_n) \\ &= \sum_{i=1}^{n} r_iP(X = r_i). \end{aligned}$$

Note that in the discrete case, the expected value need not be in the range of X. In Example 4.58 the expected value is 1.5 heads in 3 tosses of a coin, which clearly is not a valid result of tossing a coin three times.

Further note that even if the expected value is a possible outcome it need not in itself be particularly likely.

Example 4.61. Assume we are playing a game with a deck consisting of four aces and the kings of spaces and hearts,

$$\{A\clubsuit, A\spadesuit, A\heartsuit, A\diamondsuit, K\spadesuit, K\heartsuit\}.$$

We each draw a card from the pack. If one of us has an ace and the other a king, the holder of the ace gets two units from the other player. If we both have an ace, then if one of us has a black ace $A♣$ or $A♠$ then he gets 3 units from the other player. If we have aces of the same colour neither of us gets anything. If both of us have a king then the holder of the black king gets one unit from the other player.

We look at the random variable formed by the number of units gained or lost by one of the players (since the rules are symmetric it does not matter which player we pick).

It is clear that the expected pay-off for both players has to be 0 in this situation because the game is completely symmetric, and wins for one player are paid for by the other.[25] So if one player were to expect a gain the other player would have to expect a loss to make up for that game, but the rules are exactly the same for both.

We look at the range of this random variable and its pmf.

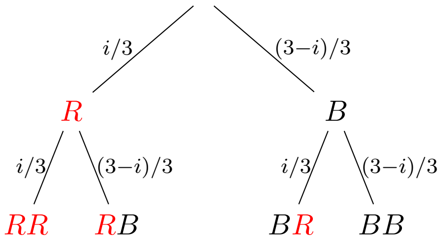
-3	-2	-1	0	1	2	3
$2/15$	$4/15$	$1/30$	$2/15$	$1/30$	$4/15$	$2/15$

We note that the expected value 0 does not occur with a particularly high probability.

Also note that the expected value does not have to be halfway between the extremes of the possible outcomes. This is illustrated (among other things) in the following example, where we calculate the average of an average to show that it is possible to have several layers of random variables, which still allow us to calculate an overall expected value.

Example 4.62. For a more down to earth example let us revisit Example 4.31. There we are faced with 4 possibilities regarding which situation we are in (given by the number of red socks in the bag). This gives us an opportunity to look at an expected value for different probability distributions.

Here is a tree that describes the drawing of two socks (with replacement) from a bag that contains i red socks from a total of 3 socks.



We have here a random variable which maps the outcomes from this tree to the number of red socks drawn. Hence it maps RR to 2, the outcomes RB and BR to 1 and the outcome BB to 0. The pmf of this random variable is

2	1	0
$\frac{i^2}{9}$	$2\frac{i(3-i)}{9}$	$\frac{(3-i)^2}{9}$

[25] This is a *zero-sum game* in the parlance of game theory.

Hence the expected value for the number of socks is

$$2\frac{i^2}{9} + 2\frac{i(3-i)}{9} = \frac{2i^2 + 6i - 2i^2}{9} = \frac{6i}{9}.$$

So the expected value in each case is

i	0	1	2	3
$E(X)$	0	2/3	4/3	2

Note how the expected value varies with the underlying situation, and note that in none of the cases we get as the expected value the halfway point between the two extremes 0 and 2.

We can use these expected values to calculate an *overall expected value* based on our current estimate for the true probability distribution.

At the beginning, the probability of the possible outcomes,

$$\{0, 1, 2, 3\}$$

is equal, 1/4 for each. If we draw two socks (returning the sock to the bag after each draw) then we would expect to draw one red and one black sock on average.

After the first update the pmf is

0	1	2	3
0	1/6	1/3	1/2

If we want the expected value based on our current knowledge, which is given by the current distribution then we should form an average where each of the previously calculated expected values is weighted by the probability that we think it's the correct one, giving an overall expected value of

$$0 \cdot 0 + \frac{2}{3} \cdot \frac{1}{6} + \frac{4}{3} \cdot \frac{1}{3} + 2 \cdot \frac{1}{2} = \frac{2 + 8 + 18}{18} = \frac{28}{18} = 1.\overline{5}.$$

Optional Exercise 17. For the expected value given at the end of the previous example, what is the underlying random variable? Give its range and its pmf.

Exercise 78. You are invited to play the following game: There are three cards:

- One is black on both sides,
- one is red on both sides,
- one is black on one side and red on the other.

You and another person pay one pound each into a kitty. The three cards are put into a bag and mixed together. Without looking into a bag you draw a card. You pull it out of the bag in a way that only the upper side can be seen, and you place it on the table. The card is red on the side you can see.

The other player bets that the card has the same colour on the hidden side as is showing. You're unsure whether you should bet on it having a different colour on the other side. The other player points out that it can't be the card that is black on both sides, so you have a 50-50 chance.

The winner of the bet is to get the two pounds put into the kitty at the start. Should you accept this as a fair game, or should you ask for your pound back?

Using conditioning to calculate expected values

Recall Example 4.60 where we determined the expected number of coin tosses until we get heads for the first time. If we use the definition of the expected value then we have to calculate with an infinite sum to find that number.

We can use conditional probabilities to help with this situation, see Section 4.4.5 for a description of how that works for random variables.

Example 4.63. We are interested in the random variable X which gives the number of tosses of the coin until we see heads for the first time, and we would like to calculate its expected value EX.

We can see that the expected value of X will have to have the property that

$$EX = \frac{1}{2} \cdot 1 + \frac{1}{2}(1 + E(X \mid \text{first toss } T)),$$

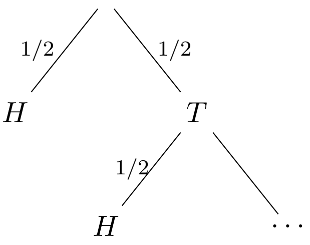
that is, with the probability of 1/2 we will have a first toss of heads and stop, and with the probability of 1/2 we will have a first toss of tails (adding one to the number of tosses), plus however many tosses come out of the conditional probability of the number of tosses given that the first toss is tails.

Now we have reduced the problem to calculating

$$E(X \mid \text{first toss } T),$$

but is that any easier?

If we look at the tree for this situation



we can see that below the node labelled T in the picture, we have another copy *of the same tree.* In other words, the tree branches to

- H, where it ends or
- T, below which another copy of the whole tree appears.[26]

This means that the expected value of the number of tosses for the whole tree has to be the same as that for the tree below T. This means that

$$E(X \mid \text{first toss } T) = EX.$$

By using that for the original equation we find that

$$EX = \frac{1}{2} \cdot 1 + \frac{1}{2}(1 + EX) = 1 + \frac{1}{2}EX.$$

We can now solve this equation for EX to obtain

$$EX = 2.$$

[26] This can only work with infinite structures.

In general we can often avoid having to calculate with infinite sums by using similar techniques. Assume we have a random experiment which has a particular result s with property p, and another result s' with property $1-p$ and that previous experiments have no effect on subsequent ones. We are interested in the expected value of the random variable X of how many times we have to repeat the experiment to get the second outcome we can see that we have

$$\begin{aligned} EX &= (1-p)\cdot 1 + p(1 + E(X \mid \text{first outcome } s)) \\ &= (1-p) + p(1+EX) \\ &= 1 + pEX. \end{aligned}$$

This means that in this situation we get that

$$EX = \frac{1}{1-p}.$$

This idea generalizes to similar experiments with several outcomes. Assume there are n possible outcomes

$$s_1, s_2, \ldots s_n$$

and that

- for $1 \leq i \leq n-1$ outcome s_i occurs with probability p_i and
- outcome s_n occurs with probability $1 - (p_1 + p_2 + \cdots + p_{n-1})$.

Then the expected number of times we have to repeat the experiment to get outcome s_n has to satisfy the equation

$$\begin{aligned} EX &= 1 - (p_1 + p_2 + \cdots + p_{n-1}) + p_1(1 + E(X \mid \text{1st } s_1)) \\ &\quad + \cdots + p_{n-1}(1 + E(X \mid \text{1st } s_n)) \\ &= 1 + (p_1 + p_2 + \cdots + p_{n-1})EX \end{aligned}$$

and so we must have

$$EX = \frac{1}{1 - (p_1 + p_2 + \cdots + p_{n-1})}.$$

Exercise 79. Assume you have a fair coin.

(a) What is the expected number of tosses until you have two heads in a row for the first time?

(b) What is the expected number of tosses until you have heads immediately followed by tails for the first time?

(c) Assume you are invited by one of your friends to play the following game: A coin is tossed unto either

- two heads occur in a row for the first time or
- we have heads immediately followed by tails for the first time.

In the first case you get 6 pounds and in the second case you have to pay the other player 5 pounds. Should you play this game?

Hint: Use the same idea as in Example 4.63.

Averages for algorithms

In COMP11212 and COMP26120 (and COMP36111 if you pick that) you will encounter the notion of the *average complexity* of an algorithm. We do not discuss this notion in detail but we give some examples to indicate which averages are formed here.

Example 4.64. Assume you have an array of integers (for example of student id numbers, pointing to the student file). Assume you are trying to find a particular id number in that array.

A simple minded algorithm for doing this will look at all the possible values in the array until the given number is found. Here's a code snippet.

```
for (int index=1; index < max_index; index++)
    if (array[index]=given_number) ...
```

How many times is the algorithm going to perform look-up for the array on average? In other words, how often will array[index] be invoked?

If the array has 8 entries then the chance that the entry we are looking for is any one of them is 1/8. If we are lucky, and we find the entry on the first attempt[27] at array[1] then we have needed one look-up, whereas if we have to keep checking until we reach array[6] we need 6 look-ups. We have a random variable which takes its values in

$$\{1, 2, 3, 4, 5, 6, 7, 8\},$$

and each of these values occurs with the same probability, namely 1/8. Hence the expected value for this random variable is

$$\begin{aligned}1 \cdot \frac{1}{8} + 2 \cdot \frac{1}{8} + \cdots + 8\frac{1}{8} &= \sum_{i=1}^{8} i\frac{1}{8}\\ &= \frac{1}{8}\sum_{i=1}^{8} i\\ &= \frac{1}{8}\frac{8(8+1)}{2}\\ &= \frac{8+1}{2}.\end{aligned}$$

This means we have to expect 4.5 look-ups on average.

Of course most real-world applications have considerably larger arrays. For this reason it pays to think about the general case.

Example 4.65. We now assume that we have an array with n entries, and we are applying the same search algorithm as in the previous example. The chance that our looked-for number is any one of them is

$$\frac{1}{n}.$$

If the looked-for entry is the first entry of the array then we need one look-up operation, if it is the second entry we need two look-ups, and so on until the

[27] Typically arrays start at index 0 but for our example it makes life less complicated if we start at index 1.

end of the array. So we have a random variable that can take values in the set

$$\{1, 2, \ldots, n\},$$

and for which the probability that any one of them occurs is $1/n$. Hence the expected value for this random variable is

$$\begin{aligned} 1 \cdot \frac{1}{n} + 2 \cdot \frac{1}{n} + \cdots + n\frac{1}{n} &= \sum_{i=1}^{n} i\frac{1}{n} \\ &= \frac{1}{n}\sum_{i=1}^{n} i \\ &= \frac{1}{n}\frac{n(n+1)}{2} \\ &= \frac{n+1}{2}. \end{aligned}$$

In other words we roughly have to look through half the array on average before finding the looked-for entry. You might have been able to work this out without any knowledge of random variables, but this is a particularly simple situation.

People who study algorithms are also interested in the *worst case* which in this example is that we have to perform n look-up operations until we finally find our number.

In the above example we were using an algorithm that is not particularly clever. If the entries appear in the array sorted by their size then we can do much better.

Example 4.66. Assume we are trying to solve the same problem as in the previous example, but this time we have an array whose entries are sorted. In that case we can come up with a faster algorithm effectively by making use of this extra information. Again we assume that we have an array of size 8.

Here's the idea:[28] The first index we try is the one halfway through the array, say the 4th entry. If the entry at that position is the one we were looking for then we are done. If not, then if the entry at that position is below the one we are looking for then we know that the looked-for entry has to be to the right of the current position at a higher index, else to the left at a lower index. Of course we might be really lucky and have found our entry already!

We now apply the same trick again: We find an entry roughly halfway through the appropriate half of the array. If the entry at the current position is below the one we are looking for....

What's the expected number of look-ups required for this algorithm? What we do on each step is to look up one entry, and split the remaining array in two parts whose sizes differ by at most 1.

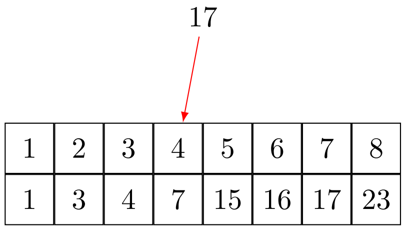
Say our array looks as follows:

1	2	3	4	5	6	7	8
1	3	4	7	15	16	17	23

If we look for the entry 17 we perform the following steps:

[28] I think you will have seen this if you have been at one of our Visit Days.

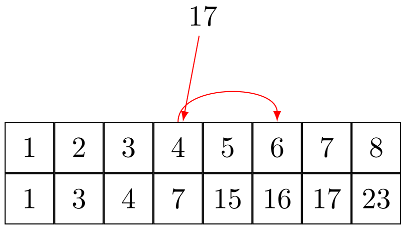
- We look at the entry at index 4, where we find the entry 7. This is smaller then the entry we are looking for.



1	2	3	4	5	6	7	8
1	3	4	7	15	16	17	23

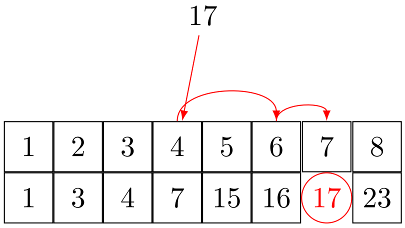
  We know that if our number is in the array it has to be to the right of the index 4.

- On the next step we look halfway through the indices 5, 6, 7 and 8. There are 4 entries, so (roughly) halfway along is at index 6. We find the entry 16, which is again smaller than the one we are looking for.



1	2	3	4	5	6	7	8
1	3	4	7	15	16	17	23

- We now have to look halfway along the indices 7 and 8. There are two entries, so halfway along is at index 7. We have found the number we were looking for,



1	2	3	4	5	6	7	8
1	3	4	7	15	16	17	23

Here is a description of the algorithm when looking for an arbitrary number in this array:

We assume that we cannot be sure that the entry is in the array at all (somebody might have given us an invalid id number). On the first step we look up the entry at index 4. If this doesn't give us the entry we were looking for then this leaves us with

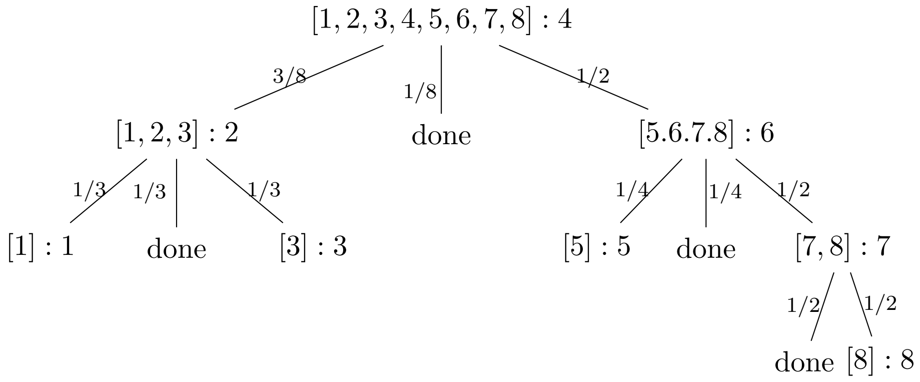
- either our entry is smaller than the one at index 4, so if it is there it must be at indices 1, 2 or 3, in which case
  - we look up the entry at index 2, and if we are not successful then
    - if our entry is below that at index 2 we look up index 1 or
    - if our entry is above that at index 2 we look up index 3,

or

- our entry is greater than the one at index for, so if it is there at all it must be at indices 5, 6, 7 or 8, in which case we
    - look up the entry at index 6, and if that is not the correct one then
        - if our entry is smaller than that at index 6 we look at index 5
        - if our entry is greater than that at index 6 we look at index 7.
            - and if it is not at index 7 we look at index 8,

This information is more usefully collected in a tree. Here the nodes are given labels where

- the first part is a list of indices we still have to look at, then there is a colon and
- the second part is the index we are currently looking at.



We can see that in the worst case we have to look at indices 4, 6, 7 and 8, which makes four look-ups.

We can also calculate the expected value for this situation:

- The probability that we need only one look-up is 1/8;
- we need two look-ups with probability

$$3/8 \cdot 1/3 + 1/2 \cdot 1/4 = 2/8;$$

- we need three look-ups with probability

$$3/8 \cdot (1/3 + 1/3) + 1/2 \cdot (1/4 + 1/2 \cdot 1/2) = 4/8;$$

- we need four look-ups with probability $1/2 \cdot 1/2 \cdot 1/2 = 1/8$.

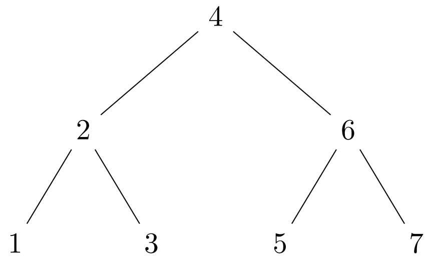
Hence the expected value for the number of look-ups is

$$1 \cdot \frac{1}{8} + 2 \cdot \frac{2}{8} + 3 \cdot \frac{4}{8} + 4 \cdot \frac{1}{8} = \frac{21}{8} = 2.625.$$

Again we want to analyse the general case of this algorithm, which is known as *binary search.*

Example 4.67. From the example above we can see that some cases are easier to analyse than others: If we have indices that exactly fit into a tree then the calculation becomes much easier.

If we look at the example of eight indices we can see that 7 indices would be fit into a tree with three levels of nodes. We can also see that we don't need to have separate nodes labelled 'done'; instead, we can just use the parent node to record that the search is over. In the case where there are seven entries in the array we could calculate the expected value using the following tree, where now we only list the index that is currently looked up for each node:



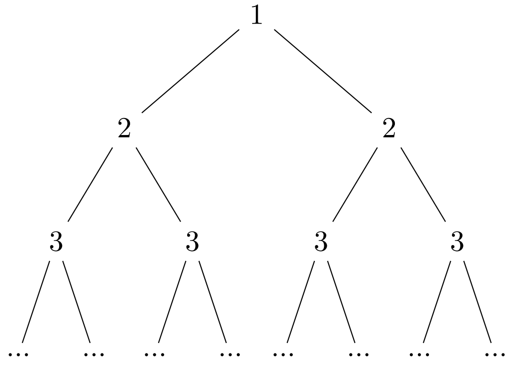
The node on the top level requires one look-up, the two nodes on the second level require two look-ups, and the four nodes on the third level require three look-ups. Each of those nodes will be equally likely to hold our number. Hence we can see that the average number of look-ups is

$$1 \cdot 1 \cdot \frac{1}{7} + 2 \cdot 2 \cdot \frac{1}{7} + 3 \cdot 4 \cdot \frac{1}{7} = \frac{17}{7} \approx 2.43.$$

We can generalize this idea provided that the number of indices is of the form

$$2^0 + 2^1 + \cdots + 2^{k-1} = \sum_{i=0}^{k-1} 2^i = 2^k.$$

We can think of the situation as being given as in the following tree, where on each level we give the number of look-ups required.



We note that on level i (from 0 down to to $k-1$) there are 2^i nodes each requiring $i+1$ look-ups and each occurring with probability $1/(2^k-1)$. Hence the approximated expected value of the number of look-ups is

$$\sum_{i=0}^{k-1} (i+1) \frac{2^i}{2^k-1} = \frac{1}{2^k-1} \sum_{i=0}^{k-1} (i+1) 2^i.$$

To check that we have derived the correct formula we can look at the case for $k = 3$, that is seven entries in the array, and compare the result we get from the formula with the one calculated above. The formula gives approximate 2.43 look-ups which agrees with the result previously calculated.

We give a few (approximate) values of this sum:

k	3	4	5	6	7	8
n	7	15	31	63	127	255
exp no look-ups	2.43	3.27	4.16	5.1	6.06	7.03

As n grows large the sum given above approximates $\log n$— so here were are in a situation where the average case is the same as the worst case!

If we only have values for arrays of sizes of the form

$$2^k - 1,$$

do we have to worry about the other cases? The answer is that the given an array with n entries, we require approximately $\log n$ look-ups, even if n is not of the shape $s^k - 1$. Occasionally it is easier to analyse particular problem sizes, and as long as the values for other values deviate in only a minor way from the function so deduced, this is sufficient for most purposes in computer science. You will learn in COMP11212 that we are typically only interested in the 'rate of growth' of a function describing the number of instructions required for a given problem size, and that all other aspects of the function in question are dropped from consideration.

Often when looking at issues of complexity it is sufficient to have approximate counts, and more generally we only care about how quickly the number of instructions grows as n grows large. We look a little into how one can measure the 'growth' of a function ' in Section 5.1.

You can see from the examples given, however, that a proper analysis can be quite tricky (the cases discussed above are relatively simple ones), and that one often has to make decisions about using approximations. When people claim that an algorithm has, say *an average case quadratic complexity* then this has to be read as an approximate description of its behaviour as the input grows large. The above preceding four examples give you an idea of what is meant by 'average number of steps'. Note that the typical assumption is that every possible configuration is equally likely (that is in our example that the sought-for number is equally likely to occur at any given index in the array), and that these assumptions are not always justified.

Properties of expected values

We know from Proposition 4.48 that we may compose a random variable with a (measurable) function from its range to (a subset of) $\mathbb{R}$ and that gives another random variable. But in general there is no easy formula for the expected value in that situation:

Composing with a function will lead to a different probability density function, and forming the integral over that is not something that can be computed in general. Even if the given random variable is discrete we do not get a simple formula: Assume that f is a measurable function from the range of a random variable X to a subset of $\mathbb{R}$. Then the new random variable has an expected value of

$$E(f \circ X) = \sum_{r \in \text{range}(f \circ X)} r \cdot P(f \circ X = i).$$

In particular there is no easy to to calculate the expected value of $f \circ X$ from that of X.

Exercise 80. Let X be a random variable; consider the following function:

$$\begin{aligned} f \colon \mathbb{R} &\longrightarrow \mathbb{R} \\ x &\longmapsto 1. \end{aligned}$$

Calculate the expected value of the random variable $f \circ X$.

If the function f is a linear function (compare Chapter 0) then we can compute the expected value of $f \circ X$ from that of X. Assume we have a discrete random variable X, with range

$$\{r_i \in \mathbb{R} \mid i \in \mathbb{N}`\}.$$

Let a and b be real numbers. We can compose X with the function

$$\begin{aligned} \mathbb{R} &\longrightarrow \mathbb{R} \\ x &\longmapsto ax + b. \end{aligned}$$

What is the expected value of the resulting random variable? We can calculate

$$\begin{aligned} E(aX + b) &= \sum_{i \in \mathbb{N}} (a \cdot r_i + b) \cdot P(aX + b = a \cdot r_i + b) \\ &= \sum_{i \in \mathbb{N}} (a \cdot r_i + b) \cdot P(X = r_i) \\ &= \sum_{i \in \mathbb{N}} a \cdot r_i \cdot P(X = r_i) + b \cdot P(X = r_i) \\ &= a \sum_{i \in \mathbb{N}} r_i \cdot P(X = r_i) + b \sum_{i \in \mathbb{N}} P(X = r_i) \\ &= a \cdot E(X) + b. \end{aligned}$$

See Exercise 71 for an explanation of the last step.

Proposition 4.68. *Let X be a random variable, and let a and b be real numbers. Then the expected value of the random variable $aX + b$, which is formed by composing X with the function*

$$\begin{aligned} \mathbb{R} &\longrightarrow \mathbb{R} \\ x &\longmapsto ax + b. \end{aligned}$$

has an expected value given by

$$E(aX + b) = aE(X) + b.$$

Proof. This result holds since

$$\begin{aligned} E(aX+b) &= \int_{-\infty}^{\infty} ax+b\cdot p(ax+b)dx \\ &= a\int_{\infty}^{\infty} x\cdot p(x)dx + b\int_{-\infty}^{\infty} p(x)dx \\ &= aE(X)+b. \end{aligned}$$

□

If we have two random variables then we can say something about combining them.

Proposition 4.69. *If X and Y are random variables then*

$$E(X+Y) = EX + EY.$$

If X and Y are independent then we also have

$$E(X\cdot Y) = EX\cdot EY.$$

4.4.7 Variance and standard deviation

The expected value of a random variable allows us to 'concentrate' it's behaviour into just one number. But as Examples 4.58 and 4.61 illustrate, the expected value can be misleading regarding which values are likely to occur. One way of measuring how far a random variable deviates from its expected value is to do the following:

Let X be a random variable.

- Calculate the

  expected value $\quad v \quad$ of X.

- Create a new random variable in two steps:
  - Subtract the expected value from X to form the random variable

    $$X - v.$$

  - To ensure that positive and negative differences from the expected value cannot cancel each other out (and to amplify differences), form the square of the previous random variable to give

    $$(X-v)^2.$$

- Calculate the expected value of the new random variable.

Example 4.70. We return to Example 4.50 of tossing a coin three times, counting the number of heads to get a random variable X. We recall from Example 4.58 that the expected value of X is 1.5.

If we form $X - 1.5$ we get a new random variable with range

$$\{-1.5, -.5, .5, 1.5\}$$

and pmf

-1.5	$-.5$	$.5$	1.5
$1/8$	$3/8$	$3/8$	$1/8.$

If we square the result we have the random variable $(X - 1.5)^2$ with range

$$\{.25, 2.25\}$$

and pmf

$.25$	2.25
$2/8 = 1/4$	$6/8 = 3/4.$

Its expected value is

$$.25 \cdot \frac{1}{4} + 2.25 \cdot \frac{3}{4} = \frac{0.25 + 6.75}{4} = \frac{7}{4} = 1.75.$$

Definition 31. If X be a random variable with expected value v its **variance** is given by

$$E((X - v)^2).$$

As pointed out above, the variance amplifies larger differences from the expected value by squaring the difference, and it returns the square of the expected difference. For some considerations it is preferred not to do the last step, leading to a slightly different way of measuring how far a random variable strays from its expected value.

Definition 32. If X is a random variable then its **standard deviation** is given by the square root of its variance.

The standard deviation gives an idea of what is 'normal' for a given distribution. If we only consider 'normal' those values which are equal to the expected value then this is too narrow for most purposes. If the average height in a given population is 167cm, then we don't consider somebody who measures 168cm far from the norm.

Typically values which are within one standard deviation on either side of the ave considered 'normal'. If the standard deviation is large that means that there are a lot of data points away from the expected value, and we should not have too narrow an idea of what is 'normal'.

Example 4.71. In the example above the standard deviation is $\sqrt{1.75} \approx 1.32$. This means that for the coin example, almost anything is normal. If we increase the number of coin tosses that changes.

4.4.8 Some selected well-studied distributions

In many situations it is hard to determine the probability distribution of a given random variable from the given data. In those cases it is standard to make the assumption that it behaves according to some well known distribution.

Clearly if this assumption is not justified then any calculations based on it are not going to be of much practical use. When you are asked to cope in such a situation you should, at the very least, think about what you know about the given situation and which well-known distribution this suits best.

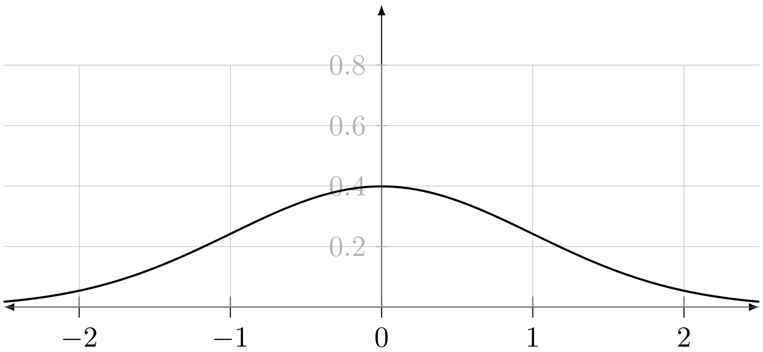
We here give an overview of only a very small number of distributions. There is plenty of material available on this topic, and so there is no need to add to that.

Normal distribution

The normal distribution is used on many occasions. It is a continuous probability distribution—in fact, it is not just one distribution, but a whole family.

In its simplest form the probability density function of the normal distribution is given by

$$\begin{aligned} \mathbb{R} &\longrightarrow \mathbb{R} \\ x &\longmapsto \frac{1}{\sqrt{2\pi}} e^{-x^2/2}. \end{aligned}$$



The expected value of a random variable with this probability density function is 1, and the standard deviation is also 1.

It is possible to create a normal distribution for a given expected value and a given standard deviation. Let v and s be real numbers, where $s > 0$. Then a random variable with probability density function

$$\begin{aligned} \mathbb{R} &\longrightarrow \mathbb{R} \\ x &\longmapsto \frac{1}{s\sqrt{2\pi}} e^{(x-v)^2/2s^2} \end{aligned}$$

has expected value v and standard deviation s.

One of the reasons that this is such a useful distribution is that, under fairly general assumptions, it is the case that the average of a (large) number of random variables which are independent and have independent distributions converges against having a normal distribution. For this reason random variables that are created from a number of independent processes obey a distribution which is close to a normal distribution.

Normal distributions are known to occur in the natural work, for example as the velocities of molecules in an ideal gas. There are many resources available to study phenomena which follow these distributions.

The normal distribution appears in the robot problem in COMP14112 to describe the accuracy of the sensor.

Bernoulli and binomial distributions

We have used Bernoulli distributions already without naming them. Given a random variable with two possible outcomes, say

$$r \qquad \text{and} \qquad r' \qquad \text{in } \mathbb{R},$$

to give a probability distribution of the random variable it is sufficient to determine

$$P(X = r),$$

and all other probabilities are then uniquely determined (compare Corollary 4.51). In particular we know that

$$P(X = r') = 1 - P(X = r),$$

since the probability of all possible outcomes have to add up to 1.

Typically for a Bernoulli distribution we are in that situation, and the assumption is that the only possible values are

$$0 \qquad \text{and} \qquad 1.$$

To make the notation less tedious, assume that

$$P(X = 1) = p.$$

The expected value of this distribution is given by

$$0 \cdot (1 - p) + 1 \cdot p = p,$$

and the variance is

$$\begin{aligned} E((X - p))^2) &= E(X^2 - 2pX + p^2) \\ &= (0^2 - 2p \cdot 0 + p^2)(1 - p) + (1^2 - 2p \cdot 1 + p^2)p \\ &= p^2(1 - p) + (1 - 2p + p^2)p \\ &= p^2 - p^3 + p - 2p^2 + p^3 \\ &= p - p^2 \\ &= p(1 - p). \end{aligned}$$

Example 4.72. Tossing a coin is an experiment that follows a Bernoulli distribution, where one of head or tails is assigned the value 1, and the other the value 0. You can think of this as the random variable that counts the number of heads (or tails) that appear in a single coin toss.

The binomial distributions arise from assuming an experiment with a Bernoulli distribution is carried out repeatedly, such as tossing a coin a number of times, and adding up the results (for example the number of heads that appear).

The Poisson distribution

The Poisson distribution is a discrete distribution that applies to process of a particular kind, namely ones where

- we look at the probability of how many instances of a given event occur within a given time interval or a given space,
- we know the average rate for these events and
- the events occur independently from the time of the last event.

Typical examples are:the following.

- The number of births per hour on a given day.
- The number of mutations in a set region of a chromosome.
- The number of particles emitted by a radioactive source within a given time span.
- The number of sightings of pods of dolphins along a given path followed by an observing plane.
- Failures of machines or components in a given time period.
- The number of calls to a helpline in a given time period.

It is assumed that the expected number of occurrences (on average) of the event is known, so assume we have $v \in \mathbb{R}^+$. A random variable X obeying the Poisson distribution has the pmf

$$P(X = n) = \frac{v^n e^{-n}}{n!}.$$

Its expected value is v, which is also the variance.

Example 4.73. Assume we have motherboards where it is known that on average, .5% are faulty. If we pick a sample of 200 motherboards, what is the probability that three of them are faulty?

From the given data we would expect $.005 \times 200 = 1$ to have one faulty board on average in such a sample. If we assume that this event follows the Poisson distribution then we

$$P(X = 3) = \frac{1^3 e^{-3}}{3!} \approx .14,$$

so the probability is 14%.

4.4.9 Additional exercises

We look at situations here which do not fit into any one section of the preceding notes because they touch on several aspect of random processes.

Exercise 81. Assume you have an array whose entries are natural numbers, and you are given a natural number k that occurs in the array. You want to change the order of the entries in the array in such a way that it satisfies the following two conditions:

- All numbers which are less than k occur to the left of k and[29]
- all numbers which are larger than k occur to the right of k.

This is a part of an important sorting algorithm called *Thickset.* The way this algorithm is implemented is as follows:

- There are two pointers, low and high.

[29] Note that if the element k occurs more than once then we don't say anything about where the other entries k may be placed.

- At the start the low pointer points to the lowest index and the highest pointer points to the highest index.
- You start a loop. This loop runs until the low pointer and the high pointer point at the same entry.
    - Look at the entry the low pointer points to.
        - If the entry is less than or equal to k then increase the low pointer by one, check that it has not reached the index of the high pointer, and repeat.
        - If the entry is greater than k then do the following.
            - Look at the entry the high pointer points to.
            - If the entry is greater than or equal to k then decrease the high pointer, check that it has not reached the low pointer, by one and repeat.
            - If the entry is smaller than k then swap the two entries the low and the high pointer are pointing to.
    - Repeat, looking again at the low pointer.

(a) Carry out this algorithm for the following array and $k = 17$.

1	2	3	4	5	6	7	8
19	2	17	5	1	27	0	31

How many times does the algorithm ask you to swap elements?

Now consider carrying out the algorithm for an arbitrary array of size n.

(b) In the best case, how many times does the algorithm have to swap elements? Justify your answer.

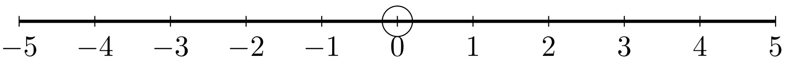
(c) In the worst case, how many times does the algorithm have to swap elements? Justify your answer.

(d) On average, how many times does the algorithm have to swap elements if you may assume that for any given entry

- the probability that it is less than k is $1/2$ and
- the probability that it is greater than k is $1/2$.

(e) Can you say how many times the algorithm has to swap elements on average if you are not allowed to make this assumption?

Exercise 82. Assume you have a line with 11 points points from -5 to 5. There is an ant at point 0.



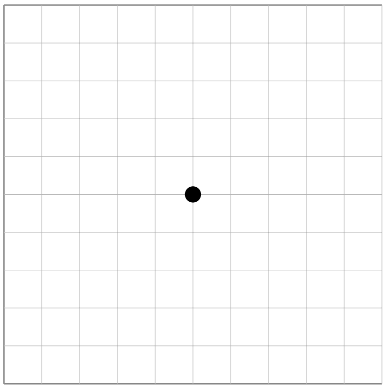
Assume that with probability $1/2$ the ant moves one point to the left, and with probability of $1/2$ it moves one step to the right. If it wants to make a step that causes it to leave the grid it stops.

(a) What is the probability that the ant will have stopped after 10 steps?

(b) What is the expected distance the ant has from its start point after 10 steps?

Exercise 83. This exercise is a generalization to 2 dimensions of the previous one, so you may want to solve that first.

Assume you have an eleven-by-eleven grid, given by the points from $(-5,-5)$ to $(5,5)$ as in the following picture. There is an ant on the grid, initially in position $(0,0)$.



Assume that with the probability of $1/4$ the ant selects a direction from $\{N,E,S,W\}$ and takes one step in that direction. If it wants to move in a direction that would cause it to leavethe grid it stops.

(a) What is the probability that the ant has stopped after ten steps?

(b) What is the average distance the ant will have from the starting point after ten steps?

Assume that the ant is not allowed to change direction by more than 90 degree on each step, and that each of the possible three directions is equally likely.

(c) What is the average distance that the ant will have from the starting point after five steps?

(d) What is the probability that the ant will have have stopped after ten steps?

Exercise 84. Assume you are looking after a cluster containing 50 machines. One of your machines has been infected by a virus that copies itself to one other, randomly chosen, machine from the cluster and infects that.

(a) What is the probability that after eight infection steps, the number of infected computers is 8? (In other words, no computer has been infected twice.)

(b) What is the expected number of infected computers after 50 infection steps?

Chapter 5

Comparing sets and functions

In computer science we are interested in comparing functions to each other because when we decide which algorithm to choose we want to pick the one that shows the better behaviour for the given range of inputs. By 'the better behaviour' we mean an algorithm that performs faster for the given inputs. As you will see in COMP11212 when we do this we only care about comparing these functions regarding how fast they grow, and one of the aims of this section is to introduce that idea.

We also have to be able to compare sets with each other. In Chapter 4 there is frequently a distinction between three cases regarding random processes into

- those with a finite number of outcomes and
- those with a countable number of these and
- those we consider continuous.

This chapter makes these ideas formal. There are other applications for these ideas, and we sketch one here:

- There are countably many Java programs.
- There are uncountably many functions from $\mathbb{N}$ to $\mathbb{N}$.

This mismatch tells us that there are some functions from the natural numbers to the natural numbers which cannot be implemented by a Java program.

5.1 Comparing functions

In Section 4.4.6 there is a discussion of calculating the number of instructions that a program has to carry out on average. It is a first step to analysing the efficiency of an algorithm.

Sometimes we have a choice of programs (or algorithms) to solve a particular problem. For small problem sizes it won't matter too much which one we pick, but as the size of our problem grows (for example, sorting millions of entries in some array as opposed to a few tens) we need to seriously think about what is the best choice. It might be the case that one of the programs takes so long (or requires so many resources in the form or memory) that it is not feasible to use.

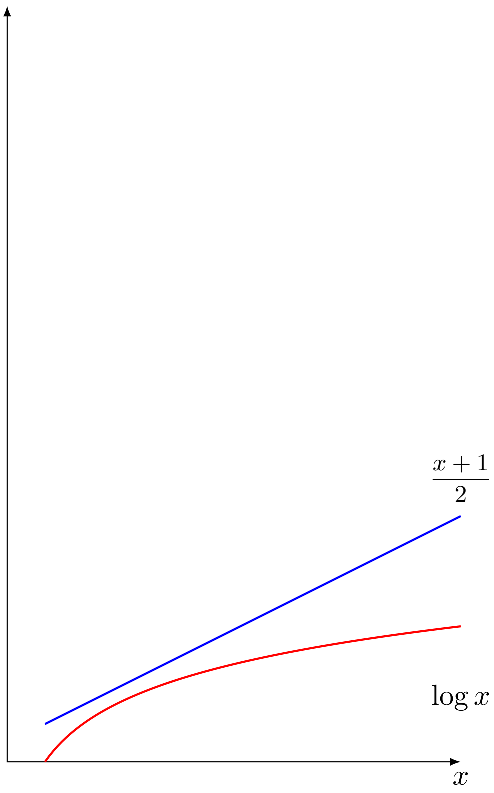
To measure the efficiency of programs it is standard to count the number of some instructions that measures how long the program is taking, depending on the size of the problem. The question then is how to compare such functions. Examples 4.65 and 4.67 give a measure of efficiency of two algorithm, the first one being known as *linear search* while the second is called *binary search*. In that case we counted the number of look-up operations performed to measure the complexity of the two algorithms.

For an array with n entries, the former has an average number of $(n+1)/2$ look-ups to perform, while the latter requires approximately $\log n$ look-ups.

We pictures the corresponding functions by drawing their graph when viewing them as functions

$$\mathbb{R}^+ \longrightarrow \mathbb{R}^+$$

instead of functions from $\mathbb{N}$ to $\mathbb{N}$.



We can see that for every input value binary search requires fewer look-ups than linear search. In this case it looks like an easy choice to make between the two. However, we have to bear in mind that binary search requires the given array to be sorted, and that does require additional computation time and power.

The picture suggests a definition for comparing functions.

Definition 33. Let f and g be two functions from a set S to[1] $\mathbb{N}$. We say

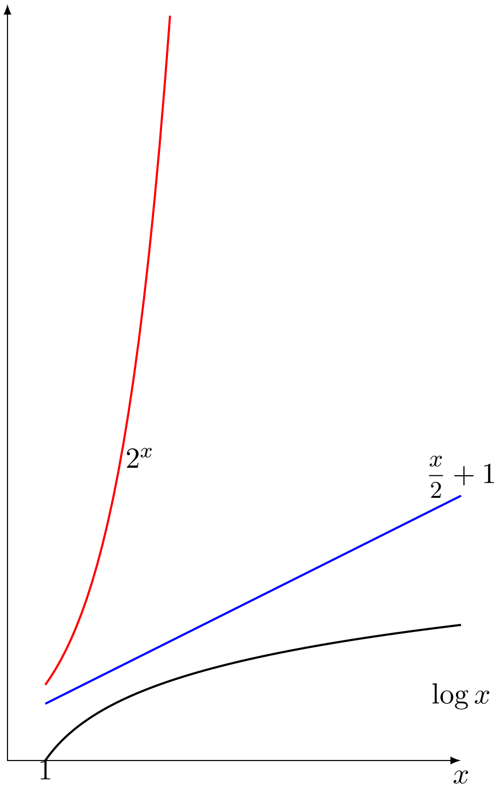
[1] We can use the corresponding definition for functions from some arbitrary set to $\mathbb{Z}$, $\mathbb{Q}$, or $\mathbb{R}$.

that f **dominates** g (or f **is above** g) if and only if

$$\text{for all } s \in S \text{ it is the case that } fs \geq gs.$$

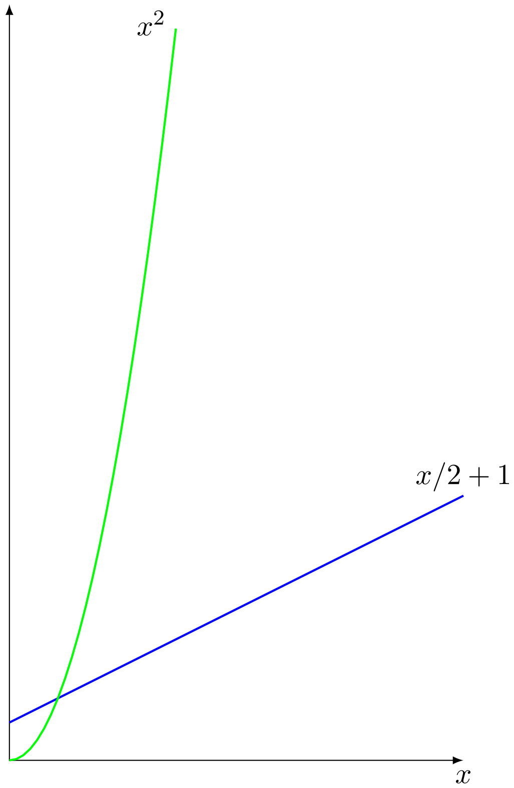
When we draw the graphs of two functions where one dominates the other we can see that the graph of the first is entirely above the graph of the second (but the graphs are allowed to touch).

Here are some examples.



The function $(x \longmapsto 2^x)$ dominates the function $(x \longmapsto x/2+1)$ which in turn dominates the function $(x \longmapsto \log x)$).

But this notion is not sufficient for an important application. If we want to establish whether one program outperforms another then using this idea for, say, the functions giving the number of instructions as a function of the size of the input for each program, may not give a useful result.



Neither function dominates the other. But clearly if the problem size is large (that is, we move to the right in the graph) then the function $(\ x \longmapsto .5x+1\ )$ offers a much preferable solution. This idea is encapsulated by the following definition.[2]

Definition 34. Let f and g be two functions from[3] $\mathbb{N}$ to $\mathbb{N}$. We say that f **eventually dominates** g if and only if

there exists $k \in \mathbb{N}$
such that for all $n \in \mathbb{N}$ with $n \geq k$ we have $fn \geq gn$.

We can think of this as saying that f dominates g if we restrict the source of f to

$$\{n \in \mathbb{N} \mid n \geq k\},$$

or if we only look at the graphs of the two functions to the right of k.

Example 5.1. Consider the two functions

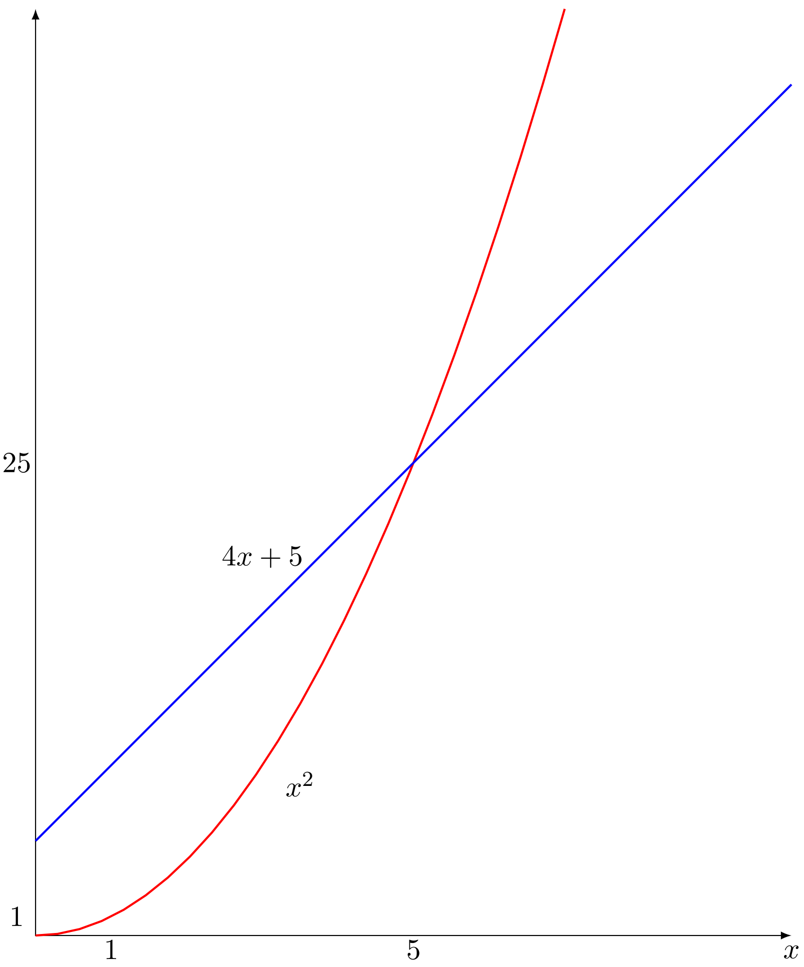
$$\begin{aligned} f\colon \mathbb{N} &\longrightarrow \mathbb{N} \\ n &\longmapsto 4n+5 \end{aligned} \qquad\qquad \begin{aligned} g\colon \mathbb{N} &\longrightarrow \mathbb{N} \\ n &\longmapsto n^2. \end{aligned}$$

Again we use a graph to picture the situation.[4]

[2] You will meet this definition again in COMP11212, and COMP21620.

[3] This works for all functions which have one of $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Q}$ or $\mathbb{R}$ as source/target.

[4] But note that drawing graphs by hand can be time-consuming, and that in order to help with answering the question whether one function is eventually dominated by another a quick imprecise sketch can be sufficient.



Once the two lines have crossed (at $x = 5$) the graph of g stays above that of f. This suggests that we should try to find a proof that g eventually dominates f.

- First of all, we have to give a witness for the 'exists' part of of the statement. The graph helps us to choose $k = 5$, but not that every natural number larger than 5 would also work.[5]
- Now that we have k we have to show that for all $n \in \mathbb{N}$, with $n \geq k$, we have $fn \geq gn$. So let us assume that $n \in \mathbb{N}$, and that $n \geq 5$. Then

$$\begin{aligned} 4n + 5 &\leq 4n + n && 5 \leq n \\ &= 5n \\ &\leq n \cdot n && 5 \leq n \\ &= n^2 \end{aligned}$$

  as required.

  Here's an alternative way of proving the same statement.

  - Again, we have to give a k, but assume this time we have not drawn the graph. We have to guess a k such that for all $n \geq k$

[5] There is no need to find the smallest number that works—any number with the right property will do!

we have

$$4n + 5 \geq n^2.$$

We can see that we require a number k such that multiplying with k is at least as large as multiplying with 4 and adding 5. Say we're a bit unsure, and we are going to try to use $k = 10$ to be on the safe side.

– We have to show that for all $n \in \mathbb{N}$

$$\text{if } n \geq 10 \qquad \text{then} \qquad n^2 \geq 4n + 5.$$

So assume $n \geq 10$. We work out that

$$\begin{aligned} 4n + 5 &\leq 4n + 10 && 5 < 10 \\ &\leq 4n + n && 10 \leq n \\ &= 5n \\ &\leq 10n && 5 < 10. \end{aligned}$$

Note that the shape of the proof has not changed much at all.

- Assume we use $k = 10$ again, but this time we produce a proof where we start by looking at the larger function.
- Let $n \geq 10$. Then

$$\begin{aligned} n^2 &= n \cdot n \\ &\geq n \cdot 10 && n \geq 10 \\ &= 4n + 6n \\ &\geq 4n + 60 && n \geq 10 \\ &\geq 4n + 5 && 60 > 5. \end{aligned}$$

Exercise 85. Determine whether one of the two functions given eventually dominates the other. Give a justification for your answer. You should not use advanced concepts such as limits or derivatives, just basic facts about numbers.

(a) $x \longmapsto 2^x$ and $x \longmapsto 1,000,000x$ as functions from $\mathbb{Z}$ to $\mathbb{Z}$.

(b) sin and cos as functions from $\mathbb{R}$ to $\mathbb{R}$.

(c) $x \longmapsto \log(x + 1)$ and $x \longmapsto x$ as functions from $\mathbb{N}$ to $\mathbb{N}$.

(d) $x \longmapsto x\log(x + 1)$ and $x \longmapsto x^2$ as functions from $\mathbb{R}^+$ to $\mathbb{R}^+$.

5.2 Comparing sets

In the introduction to this chapter we have argued that it is important to be able to compare the sizes of different sets. It turns out that the notions of injective and surjective functions from Section 2.3.3 is useful for this purpose.

In particular, if there is an injective function from a set S to a set T, then for every element of S there is an element of T, and all these elements are different. Hence we know that all elements of S 'fit into' T, and T must be as least as big as S.

Definition 35. Let S and T be sets. We say that the **size of S is smaller than or equal to that of** T if and only if there is an injection from S to T.

You may wonder whether this is a sensible notion of size. Here is one indication that this is so.

Lemma 5.2. *If S and T are sets with finitely many elements then the size of S is less than or equal to the size of T if and only if the number of elements of S is less than or equal to the number of elements of T.*

Proof. Note that in Exercise 34 it is shown that the number of elements in the image of a set S under an injection is the same as the number of elements of S.

We show both implications separately.

- Assume that the size of S is less then or equal to the size of T. Then there is an injection, say f, from S to T. By the above exercise we know that the number of elements of the image $f[S]$ of S under f is the same as the number of elements of S. Since $f[S]$ is a subset of T we know that T has at least as many elements as S.
- Assume that the number of elements of S is less then or equal to the number of elements of T. This means that if name the elements of S as s_1, s_2, ..., s_m, and those of T as t_1, t_2, ..., t_n then $n \geq m$. If we now define the function

  $$\begin{aligned} \{s_1, s_2, \ldots, s_m\} &\longrightarrow \{t_1, t_2, \ldots, t_n\} \\ s_i &\longmapsto t_i \end{aligned}$$

  from S to T it is an injection.

This completes the proof. □

Here is an example with infinite sets.

Example 5.3. The natural numbers $\mathbb{N}$ can be mapped via an injection into the integers $\mathbb{Z}$ by defining

$$n \longmapsto n \; .$$

This is clearly an injection. Hence the size of $\mathbb{N}$ is less than or equal to the size of $\mathbb{Z}$.

If I had asked in the lecture whether the size of $\mathbb{Z}$ is at least that of $\mathbb{N}$ I am sure everybody would have told me that this is true. You may find the following example less intuitive. It shows that once we have sets with infinitely many elements our intuitions about their sizes become suspect.

Example 5.4. What you might find more surprising is that $\mathbb{Z}$ also has a size smaller than or equal to that of $\mathbb{N}$. We give an injection $f\colon \mathbb{Z} \longrightarrow \mathbb{N}$ by setting[6]

$$x \longmapsto \begin{cases} 2n & \text{if } n \in \mathbb{Z}^+ \\ -(2n+1) & \text{else} \end{cases}$$

This is an injection for the following reason. Let m and n in $\mathbb{Z}$. We have to show that $fm = fn$ implies $m = n$. Since the definition of f is by cases we have to distinguish several cases in this proof.

[6] Compare this to the function from the mid-term test in 2015/16,

- m and n both in $\mathbb{Z}^+$. If $2m = fm = fn = 2n$ we may conclude $m = n$.
- m and n both negative. If $-(2m+1) = fm = fn = -(2n+1)$ we may conclude that $2m+1 = 2n+1$ and so $m = n$ as required.
- m in $\mathbb{Z}^+$ and n negative. If $2m = fm = fn = -(2n+1)$ we get $2m = 2n+1$ which can never hold for m, n in $\mathbb{Z}$.
- m negative and n in $\mathbb{Z}^+$. This case is identical to the previous one where n and m have been swapped.

What does it mean that $\mathbb{N}$ is at least as big as $\mathbb{Z}$, and $\mathbb{Z}$ is at least a big as $\mathbb{N}$?

Definition 36. We say that two sets S and T **have the same size** if and only if

- the size of S is less than or equal to the size of T and
- the size of T is less than or equal to the size of S.

The previous two examples show that $\mathbb{N}$ and $\mathbb{Z}$ have the same size.

Exercise 86. Show that the following sets have the same size.

(a) $\mathbb{N}$ and $\mathbb{N} \times \mathbb{N}$;

(b) $\mathbb{N}$ and $\mathbb{N}^k$ where k is a finite number.

(c) $\mathbb{N}$ and $\mathbb{Q}$.

(d) the set of functions from some set S to the two element set $\{0, 1\}$ and the powerset $\mathcal{P}S$ of S.

Exercise 87. Show that if there is a bijection from S to T then S and T have the same size.

Optional Exercise 18. Show that if S and T have the same size then there is a bijection between them. This is known as the *Cantor-Bernstein-Schröder Theorem.*

We may also use this idea to define rigorously the notion of an infinite set. So far we have appealed to the idea that an infinite set is one that does not have a finite number of elements, but what does it mean to have 'infinitely many' elements? Instead we use the following idea, known as *Dedekind infinite.*

Definition 37. A set S is **infinite** if and only if there is an injection from S to a proper subset of S.

Proposition 5.5. *A set S is infinite if and only if there is an injective function from S to itself which is not surjective.*

Proof. We show the statement in two parts.

Assume that the set is infinite. Then there is an injective function

$$f\colon S \longrightarrow S'$$

where S' is a proper subset of S. We can define a function

$$\begin{array}{rcl} g\colon S & \longrightarrow & S \\ s & \longmapsto & fs \end{array}$$

which is obviously also injective, but it is not surjective since we know there is an element of S which is not in S', and so cannot be in the image of g.

Assume that we have an injective function

$$g\colon S \longrightarrow S$$

which is injective but not surjective. Then there is an element s of S which is not in the image of g, that is, there is no $s' \in S$ with $gs' = s$. We define a new function

$$\begin{array}{rcl} f\colon S & \longrightarrow & S \setminus \{s\} \\ s & \longmapsto & gs \end{array}.$$

We note that f is injective since g is, and we note that its image is a proper subset of S. □

Example 5.6. We show that there are infinitely many Java programs. We have to give an injective function from the set of Java programs to itself whose range does not include all Java programs.

We do this as follows: Given a Java' program we map it to the same Java program to which the line

```
System.out.println(”Hello world!”);
```

has been added.

This function is injective: If we have two Java programs that are mapped to the same program then they must be the same program once that new last line has been removed.

The image of this function is a proper subset of the set of all Java programs since there are many programs which do not contain that line and so are not in the image of the function.

Hence we can use the assignment given above to map the set of all Java programs to a proper subset of itself.

Example 5.7. We show that the set of finite subsets $\mathcal{P}_f\mathbb{N}$ of $\mathbb{N}$ is infinite. Again we begin by giving an injective function from that set to itself.

Given a finite non-empty subset

$$\{s_1, s_2, \dots, s_n\}$$

of $\mathbb{N}$ we map it to the set

$$\{s_1, s_2 \dots s_n, (s_1 + s_2 + \cdots s_n)\},$$

and we map the empty set to itself. In other words we have

$$\begin{array}{rcl} \mathcal{P}_f\mathbb{N} & \longrightarrow & \mathcal{P}_f\mathbb{N} \\ \{s_1, s_2, \dots, s_n\} & \longmapsto & \begin{cases} \{s_1, s_2, \dots, s_n, (s_1 + s_2 + \cdots s_n)\} & n > 0 \\ \emptyset & \text{else.} \end{cases} \end{array}$$

In other words we map a given non-empty set to the set where the sum of all the elements of that set has been added as an extra element. We observe that the extra element is always the largest element of the resulting set. Note that if the set we start with has only one element then it is mapped to itself by this function since no extra element is added.

This function is injective. If two sets are mapped to the same set then in particular their greatest elements must be equal, so the original sets must have had elements which add up to the same number. Moreover, all elements (if any) of the set which are below the largest element must also correspond to each other, so the sets must have been equal and our function is injective.

This function is not surjective since the set $\{1, 2\}$ is not in the image of this function. By Proposition 5.5 we know that the given set is infinite.

Exercise 88. Show that the following sets are infinite.

(a) $\mathbb{N}$,

(b) $\mathbb{R}$,

(c) the set of functions from $\mathbb{N}$ to the two element set $\{0, 1\}$ or the powerset $\mathcal{P}\mathbb{N}$ (you choose),

(d) every superset of an infinite set.

(e) Any set which is the target of an injective function whose source is infinite.

Exercise 89. Show that if a set has a finite number of elements then it is not infinite.

Optional Exercise 19. Show that if a set is not infinite then it has a finite number of elements.

Exercise 90. Show that if S is a set with a finite number of elements then so is its powerset $\mathcal{P}S$. Do so by determining the number of elements of $\mathcal{P}S$.

In computer science we particularly care about sets which are at most as large as the natural numbers. This is because given a finite number of symbols there are only countably many strings (and so programs) that can expressed using those symbols.

Definition 38. A set is **countable** if and only if there is an injection from it to the natural numbers. A set is **uncountable** if and only if there is no injection from it to the natural numbers. A set is **countably infinite** if it is both, countable and infinite.

Note that every finite set is countable.

Examples of countably infinite sets are:

- The set of natural numbers $\mathbb{N}$.
- The set of integers $\mathbb{Z}$.
- The set of rational numbers $\mathbb{Q}$.
- The set of finite subsets of $\mathbb{N}$, $\mathcal{P}_f\mathbb{N}$.

- The set of all programs in your favourite programming language.
- The set of all strings over a finite alphabet.

Examples of uncountable sets are:

- The set of real numbers $\mathbb{R}$,
- the set of complex numbers $\mathbb{C}$,
- the set of all subsets of $\mathbb{N}$, $\mathcal{P}\mathbb{N}$,
- the set of all functions from $\mathbb{N}$ to $\mathbb{N}$.

Note that the last example, together with the following exercise, illustrates that there are functions from $\mathbb{N}$ to $\mathbb{N}$ for which we cannot write a computer program!

Optional Exercise 20. Show that every uncountable set is at least as big as any countable set.

Optional Exercise 21. Assume we have a finite set of symbols, say A.

(a) Show that A^k is finite for every $k \in \mathbb{N}$.

(b) Show that

$$\bigcup_{k\in\mathbb{N}} A^k$$

is countable.

(c) Show that there is a bijection between the set of finite strings built with symbols from A and the set $\bigcup_{k\in\mathbb{N}} A^k$,

(d) Conclude that there are countably many strings over the alphabet A.

(e) Put together a set of symbols such that every Java program can be built from those symbols.

(f) Prove that there is an injection from the set of Java programs to the set of strings over this set of symbols.

(g) Conclude that the set of Java programs is countable.

Exercise 91. Show that every subset of a countable set is countable. Conclude that every superset of an uncountable set is uncountable.

Optional Exercise 22. Show that the following sets do not have the same size.

(a) Any set and its powerset;

(b) $\mathbb{N}$ and $\mathbb{R}$. Conclude that $\mathbb{R}$ is not countable.

As a consequence of Exercises 21 and 22 we can see, for example, that there are more real numbers than there are Java programs. This means that if we cannot hope to write a Java program that outputs the digits of a given real number, one at a time, for every real number.

Exercise 92. Show that given a function $f\colon S \longrightarrow T$ the following are equivalent:

(i) f is a surjection and

(ii) the size of S is at least the size of T.

Optional Exercise 23. Show that any two countably infinite sets have the same size.

Glossary

continuous 141

A random variable is continuous if and only if it is not discrete.

countable 186

A set is countable if and only if there is an injective function from it to $\mathbb{N}$.

countably infinite 186

A set is countably infinite if it is both, countable and infinite.

cumulative distribution function (cdf) 149

The cdf of a random variable maps each element r of $\mathbb{R}$ to the probability that the random variable has a value less than or equal to r.

discrete 141

A random variable is discrete if and only if its range is countable.

dominate 178

A function f from a set X to $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Q}$ or $\mathbb{R}$ dominates another g with the same source and target if and only if the graph of f lies entirely above the graph of g (graphs touching is allowed).

eventually dominate 180

A function f from $\mathbb{N}$ to $\mathbb{N}$ eventually dominates another g with the same source and target if and only if there is some number beyond which the graph of f lies above that of g (graphs touching is allowed). The analogous definition works for functions with source and target $\mathbb{Z}$, $\mathbb{Q}$ or $\mathbb{R}$.

expected value 158

The expected value of a random variable can be thought of as the average value it takes. It is given by the integral of the product of a number which the probability that it is the value of the random variable. If the random variable is discrete then this is given by a sum.

independent 136, 155

Two events are independent if and only if the probability of their intersection is the product of their probabilities. Two random variables are independent if and only if for every two events it is the case that the probability that the two variables take values in the product of those events is the product of the probabilities that each random variable takes its value in the corresponding event..

infinite 184

A set is infinite if and only if there is an injection from it to a proper subset.

measurable 139

A function from the sample set of a probability space to the real numbers is measurable if and only if for every interval it is the case that the set of all outcomes mapped to that interval is an event.

probability mass function (pmf) 148

The pmf of a discrete random variable maps each element of the range of that random variable to the probability that it occurs.

random variable 139

A random variable is a measurable function from the set of outcomes of some probability space to the real numbers.

size of a set 183, 184

A set is smaller than another if there exists an injective function from the first to the second. They have the same size if they are both smaller than the other.

standard deviation 171

The standard deviation of a random variable is given by the square root of its variance.

uncountable 186

A set is uncountable if it is not countable.

variance 171

The variance of a random variable with expected value e is given by the expected value of the random variable constructed by squaring the result of subtracting e from the original random variable.

COMP11120, Semester 1

Exercise Sheet 9

For examples classes in Week 11

Core Exercises marked this week

Where the answers are probabilities don't just give a number, give an expression that explains how you got to that number!

Exercise 68.

Exercise 74.

Exercise 75.

Extensional Exercises marked this week

Exercise 76.

Exercise 77.

Remember that

- the **deadline** is the beginning of the examples class, and that you have to be able to promptly answer questions by the TA, referring to your rough work as needed;
- you may only use concepts which are defined in these notes (Chapter 0 establishes concepts for numbers), and for every concept you do use you should find the definition in the notes and work with that;
- you should justify each step in your proofs;
- if you are stuck on an exercise move on to the next one after ten minutes, but write down why you got stuck so that you can explain that to the TA in the examples class. If you couldn't get started then note down all the relevant definitions (use the Glossary to find these),

Exercises you could do this week are those in Section 2.3.3.

COMP11120, Semester 1

Exercise Sheet 10

For examples classes in Week 12

Core Exercises marked this week

Where the answers are probabilities don't just give a number, give an expression that explains how you got to that number!

Exercise 78.

Exercise 84.

Exercise 85.

Extensional Exercises marked this week

Exercise 79.

Exercise 81.

Remember that

- the **deadline** is the beginning of the examples class, and that you have to be able to promptly answer questions by the TA, referring to your rough work as needed;
- you may only use concepts which are defined in these notes (Chapter 0 establishes concepts for numbers), and for every concept you do use you should find the definition in the notes and work with that;
- you should justify each step in your proofs;
- if you are stuck on an exercise move on to the next one after ten minutes, but write down why you got stuck so that you can explain that to the TA in the examples class. If you couldn't get started then note down all the relevant definitions (use the Glossary to find these),

Exercises you could do this week are those in Section 2.3.3.