COMP11120 Mathematical Techniques for Computer Science Part 3

Andrea Schalk
A.Schalk@manchester.ac.uk

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

November 10, 2015

Contents

0 Basics **14**
- 0.1 Numbers . . . 14
- 0.2 Sets . . . 24
- 0.3 Functions . . . 34
- 0.4 Constructions for functions . . . 42

1 Complex Numbers **45**
- 1.1 Basic definitions . . . 45
- 1.2 Operations . . . 46
- 1.3 Properties . . . 52
- 1.4 Applications . . . 52

2 Statements and Proofs **54**
- 2.1 Motivation . . . 54
- 2.2 Precision . . . 56
- 2.3 Important examples . . . 66

3 Formal Logic Systems **89**

4 Probability Theory **90**
- 4.1 Analysing probability questions . . . 90
- 4.2 Axioms for probability . . . 105
- 4.3 Conditional probabilities and independence . . . 118
- 4.4 Random variables . . . 138
- 4.5 Random events and averages in computer science . . . 156

5 More About Statements and Proofs **133**

6 Recursion and Induction **142**
- 6.1 The natural numbers . . . 143
- 6.2 Lists . . . 153
- 6.3 Trees . . . 158
- 6.4 Formal languages . . . 160

7 Relations **166**
- 7.1 General relations . . . 166
- 7.2 Partial functions . . . 170
- 7.3 Equivalence relations . . . 174
- 7.4 Partial orders . . . 200

Exercise Sheets **217**
Exercise Sheet 0 217
Exercise Sheet 1 218
Exercise Sheet 2 219
Exercise Sheet 3 220
Exercise Sheet 7 221
Exercise Sheet 8 222
Exercise Sheet 9 223
Exercise Sheet 10 224

Chapter 4

Probability Theory

Probabilities play a significant role in computer science. Here are some examples:

- One mechanism in machine learning is to have *estimates* for the relative probabilities of something happening, and to adjust those probabilities as the system gets more data.
- If you are running a server of some kind you need to analyse what the average, and the worst case, load on that server might be to ensure that it can satisfy your requirements.[1] Calculating such averages is one of the techniques you learn in probability theory.
- When trying to analyse data you have to make some assumptions in order to calculate anything from the data. We look at the question of what assumptions have what consequences.
- In order to calculate the *average complexity* of a program you have to work out what the average input is like—this is effectively the expected value of a random variable.
- There are sophisticated algorithms that make use of random sampling, such as *Monte Carlo methods*. In order to understand how to employ these you have to understand probability theory.

4.1 Analysing probability questions

Before we look at what is required formally to place questions of probability on a sound mathematical footing we look at some examples of the kinds of issues that we would like to be able to analyse.

In computer science we are often faced with situations where probabilities play a role, and where we have to make the decision about how to model the situation.

Every time we are trying to judge the risk or potential benefits of a given decision we are using probabilistic reasoning, possibly without realizing it. We have to come up with a measure of how big the potential benefit, or the potential disadvantage is, and temper that judgement by the likelihood of it occurring.

[1] You wouldn't the student system to go down if all students are trying to access their exam timetable at the same time.

When somebody buys a lottery ticket, the potential disadvantage is losing their stake money, and the potential advantage is winning something. How many people know exactly what their chances are of doing the latter?

Many games include elements of chance, typically in the form of throwing dice, or dealing cards. When deciding how to play, how many people can realistically assess their chances of being successful?

In machine learning, one technique is to model a situation by assigning probabilities to various potential properties of the studied situation. As more information becomes available, these probabilities are updated (this constitutes 'learning' about the situation in question). How should that occur?

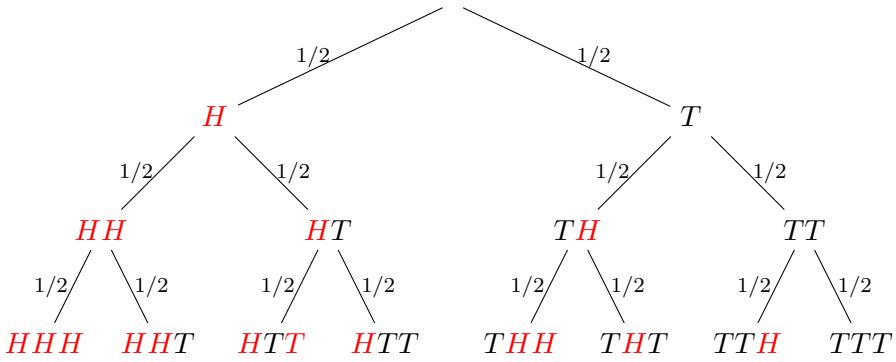
When looking at questions of the complexity of algorithms, one often applied measure is the 'average complexity', by which we mean the complexity of the 'average case' the program will be applied to. How does one form an 'average' in a situation like that?

All these questions are addressed in probability theory, but we have to restrict ourselves here to fairly basic situations to study the general principles. The first few problems we look at are particularly simple-minded.

4.1.1 Simple examples

Most people will have been confronted with issues like the following.

Example 4.1. An example much beloved by those teaching probabilities is that of a coin toss. When a fair coin is thrown we expect it to show 'heads' with the same probability as tails. For the chances to be even, we expect each to occur with the probability of $1/2$. What if we throw a coin more than once? We also expect that the outcome of any previous toss has no influence on the next one. This means we expect it to behave along the following lines.



In order to work out the probability of throwing, say, HTH, we follow down the unique path in the tree that leads us to that result, and we multiply the probabilities we encounter on the way down, so the probability in question is

$$1/8.$$

Note that because each probability that occurs in the tree is $1/2$, the effect will be that each outcome on the same level will have the same probability, which is as expected.

The tree also allows us to work out what the probability is of having one of

$$HHH \qquad \text{and} \qquad TTT$$

all we have to do is to add up the probabilities for each of the outcomes in the set, so the probability in question is

$$1/8 + 1/8 = 1/4.$$

See Section 4.1.4 for more examples where it is useful to draw trees.

Example 4.2. Whenever we throw a die, we expect each face to come up with equal probability, so that the chance of throwing, say, a 3 at any given time is 1/6. It is quite easy to construct more complicated situations here. What if we throw two dice? What are the chances of throwing two 1s? What about throwing the dice such that the eyes shown add up to 7? See Exercise 49 and Example 4.17 for a detailed discussion of this particular question.

There are games where even more dice come into the action (for example Risk and Yahtzee), and while computing all probabilities that occur there while you're playing the game may not be feasible, it might be worth estimating whether you are about to bet on something very unlikely to occur.

Example 4.3. A typical source of examples for probability questions is as a measure of uncertainly of something happening. For example, a company might know that the chance of a randomly chosen motherboard failing within a year is some given probability. This allows both, the producing company and other manufacturers using the part, to make some calculations regarding how many cases of repairs under warranty they are likely to be faced with.

In particular, if you are a manufacturer seeking to buy 100,000 motherboards, then you have to factor in the costs of using a cheaper, less reliable part, compared with a more expensive and more reliable one. If you have a 10$ part which has a 5% chance to be faulty within the given period, you would expect to have around

$$100,000 \cdot .05 = 5000$$

cases. If on the other hand, you have a 12$ part that has a 3% chance of being faulty then you will have to pay 200,000$ more for the parts, and expect to have only

$$100,000 \cdot .03 = 3000$$

cases of failure under warranty. What is the better choice depends on how expensive it is to deal with each case, how many people you expect to make a claim, and whether you worry about the reputation of your company among consumers. Decisions, decisions...

Example 4.4. When you are writing software you may wonder how well your program performs on the 'average' case it will be given.

For a toy example, assume that your program takes in an input string, does some calculations, and returns a number. The number of calculation steps it has to carry out depend on the length of the string. You would like to know how many calculation steps it will have to carry out on average so that you have an idea how long a typical call to that program will take.

Assume we have a string of length n. There is a function which assigns to each $n \in \mathbb{N}$ the number of calculation steps performed for a string of that length. It may not be easy to *calculate* that function, and you will learn more about how one might do that in both, COMP112 and COMP261. For the moment let's assume the function in question is given by the assignment[2]

$$n \longmapsto n^2$$

from $\mathbb{N}$ to $\mathbb{N}$.

So now all we need is the average length of an input string to calculate the average number of calculations carried out. But what is that? This will depend on where the strings come from. Here are some possibilities:

- The strings describe the output of another program.
- The strings are addresses for customers.
- The strings encode DNA sequences.
- The strings describe the potential status of a robot (see Example 4.32).
- The strings are last names of customers.

In each situation the average length will be different. You need to know something about where they come from to even start thinking about an 'average' case.

If we have a probability for each length to occur then we can calculate an average, see Definition 29 for that.

Note that typically the number of instructions that has to be carried out in a typical computer program depends on more than just the size of the input. With many interesting algorithms (for example searching or sorting ones) what exactly has to be done depends on the precise nature of the input. See Examples 4.59 and 4.60 for a discussion of two such situations.

4.1.2 Counting

When modelling situations using probability we often have to count how many possibilities there are, and how many of those have particular properties.

We give some rules here that help with taking care of this.

Selection with return

Assume we are in a situation where there are n options to choose from, and that we may choose the same option as many times as we like. If we choose i many times and we record the choices in the order we made them, then there are

$$n^i$$

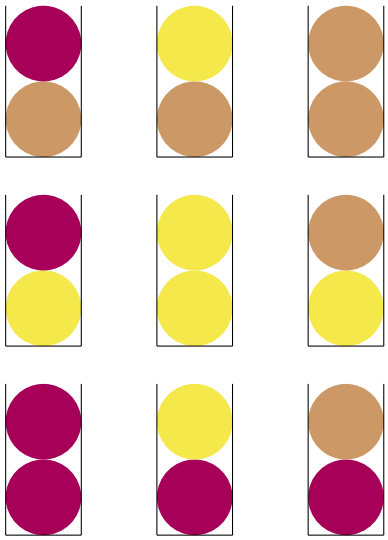
possible different possibilities.

[2] It is not unusual for such a function to have to be defined by cases rather than having one neat expression for it.

Example 4.5. If we toss a coin then on each toss there are two options, heads and tails. If we toss a coin i times then there are 2^i many possible combinations.

Example 4.6. Let's assume we have various flavours of ice cream, and we put scoops into a tall glass so that they sit one above each other. If you may choose 3 scoops of ice cream from a total of n flavours then there are n^3 many combinations, assuming all flavours remain available.

Below we show all the combinations of picking two scoops from three flavours, say lemon, raspberry, and hazelnut.



There are $3^2 = 9$ possible combinations.

The reason this is known as 'selection with return' is that if we think of the choice being made by pulling different coloured balls from an urn (without being able to look into the urn), then one should picture this as drawing a ball, recording its colour before returning it to the urn, drawing a second ball, recording its colour before returning it, and so on.

Selection without return

If we have a choice of n possibilities, and we choose i times in a row, but we may not choose the same item twice, then there are

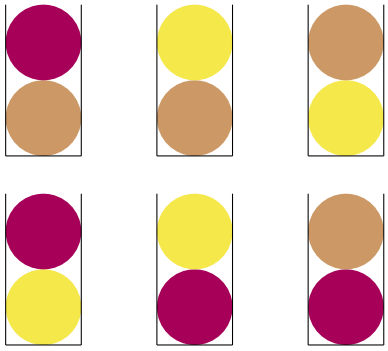
$$n(n-1)\dots(n-i+1) = \frac{n!}{(n-i)!}$$

different combinations, that is listings of choices in the order they were made in.

Example 4.7. If you have to pick three out of fifteen possible runners to finish first, second and third in that order there are $15 \cdot 14 \cdot 13 = 2730$ possibilities

Example 4.8. If you have a program that gives you a design for a webpage, where you have to pick three colours to play specific roles (for example, background, page banner, borders), and there are 10 colours overall, then you have $10 \cdot 9 \cdot 8 = 720$ combinations.

Example 4.9. Returning to the ice cream example, if children are given a tall glass in which they each are allowed two scoops from three flavours, but they may pick every flavour at most once (to make sure popular flavours don't run out) then they have the following choices.



There are now $3 \cdot 2 = 6$ possibilities.

This is known as selection without return because we can think of it as having an urn with n differently coloured balls, from which we choose one ball after the other (without being able to look into the urn) and recording the colours in the order they appear.

What happens if the balls don't each have a unique colour?

Ordering

If we have n different items then there are $n!$ many ways of ordering them, that is, of writing them one after the other. This is the same as choosing without return n times from n possible options. If the items are not all different then the number of visibly different possibilities is smaller.

Example 4.10. If we have a red, a blue, and three black mugs and we are lining them up in a row then the number of possibilities is

$$\frac{5!}{3!} = 20$$

There would be 5! possibilities for lining up 5 different mugs, but in each one of those we wouldn't spot the difference if some of the black mugs were swapped. There are 3! ways of lining up the three black mugs (but if we assume that the mugs are indistinguishable then we cannot tell the difference between the different orderings).

In general, if we have n items and there are n_1 copies of the first design, n_2 copies of the second, and so on, to n_i items of the ith design then there are

$$\frac{(n_1 + n_2 + \cdots + n_i)!}{n_1! \cdot n_2! \cdot \cdots \cdot n_i!}$$

visibly different ways of lining up the items.

Selection without ordering

Sometimes we are confronted with the situation where we have to count how many different selections there are, but where we are not told the order in which this selection arises. A typical example is a lottery draw:

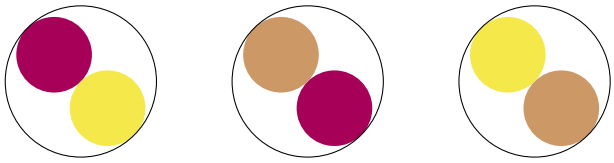
Example 4.11. If we return to Example 4.9, then we can look at the situation where the children are given a shallow bowl rather than a tall glass with scoops of ice cream. Again they are allowed to choose two scoops from three flavours, and again they may pick every flavour at most once.

One way of counting these is to list all the options as we have done above, but that gets cumbersome if the numbers involved are bigger. An alternative way of counting is to count how many selections there are with ordering, and then dividing by the number of different orderings there are for each choice.

We know from Example 4.9 that there are 6 possible combinations when the order is taken into account. For each choice of two flavours there are two ways of ordering them, so we now have

$$\frac{3 \cdot 2}{2} = 3$$

combinations.



In general, when i items are picked from a choice of n different ones, there are

$$\frac{n(n-1)\dots(n-i+1)}{i!} = \frac{n!}{(n-i)!i!}$$

different selections.

Summary

The formulae given above are summarized in the following table. Here n is the number of items available and i is the number of items that are selected. Note that the assumption is that in the unordered case, all items are different.

ordered		unordered
with return	without return	
n^i	$\frac{n!}{(n-i)!}$	$\frac{n!}{(n-i)!i!}$

Note that there is no simple formula for the number of possibilities there are when looking at unordered selections of items some of which may be identical.

Optional Exercise 7. Work out why there is no simple formula as described in the previous sentence by looking at some examples.

Exercise 45. Assume you have 3 red socks and 5 black ones. Answer the following questions

(a) Assume we put all the socks into a bag. Four times we draw a sock from the bag, putting it back each time. How many different draws are there?

(b) Make the same assumption as for the previous part, but now assume we don't put the drawn socks back into the bag. How many draws are there?

(c) Assume we put the socks onto a pile, close our eyes, mix them around, and pick four socks from the pile. How many different combinations do we get?

(d) Can you answer the same questions if you assume we have m red and n black socks? What if we pick k socks (for $k \leq m + n$) many socks on each occasion?

Exercise 46. A researcher in the rain forest has left his laptop unattended and a curious monkey has come to investigate. When the researcher looks up from the plant he is studying he sees the monkey at the keyboard. He makes threatening noises as he runs back. Assume that there's a 50% chance that he will manage to disrupt the monkey before it makes another key stroke, and that he will have reached the laptop before the monkey can manage 6 key strokes. Draw a tree for the situation. What do you think is the average number of key strokes the monkey will manage in this situation?

4.1.3 Combinations

Sometimes we have to combine these ideas to correctly count something.[3]

Example 4.12. If we throw a coin three times then there are 2^3 many possible outcomes. If we want to know how many of those contain at least two heads we have to think about how best to count the number of possibilities.

One possibility is to say that we are interested in

- the situation where there are three heads, of which there is one combination, and
- the situation where there are two heads and one tails. This amounts to the number of different ways of ordering H, H, T and there are

$$\frac{3!}{2!} = 3$$

of those (or there are there positions where the unique T can go and then the two H take up the remaining positions).

But this way of thinking does not scale well. What if we want to know how many outcomes have at least 10 heads when we toss the coin 20 times? Following the above idea we have to add up the number of combinations with 20, 19, 18, and so on, down to 10 occurrences of H. Or we can argue that there are 2^{20} possibilities over all, of these $20!/(10! \cdot 10!)$ contain exactly ten times heads and ten times tails and of the remaining combinations half will have a higher count of heads, and half will have a higher count of tails.

One can calculate this by hand in this particular example

$$\frac{20!}{10! \cdot 10!} = \frac{2 \cdot 19 \cdot 2 \cdot 17 \cdot 2 \cdot 15 \cdot 2 \cdot 13 \cdot 2 \cdot 11}{5!} = \frac{19 \cdot 17 \cdot 2 \cdot 13 \cdot 2 \cdot 11}{1} = 184756.$$

As a consequence our number of possibilities is

$$\frac{2^{20}}{2} + 184756 = 709044.$$

[3] another ex

By thinking about how to count in the right way calculations can be shortened significantly.

Exercise 47. Work out how many outcomes there are in the following cases. Please give an expression that explains the number you have calculated.

(a) Four digit personal identification numbers (PINs). How many times do you have to guess to have a 10% chance of finding the correct PIN?

(b) How many passwords are there using lower case letters? How many times do you have to guess now to have a 10% chance of being correct?

(c) What if upper case letters are included?

(d) How many possible lottery draws are there if six numbers are drawn from 49? How many bets do you have to make to have a 1% chance of having all numbers correct?

(e) Assume you have an array consisting of 10 integers. What is the probability that the array is sorted?

(f) Assume you have an array consisting of 30,0000 id numbers. What is the probability that you randomly pick the one you were looking for? What can you say about the case where the array is sorted?

(g) In an examples class there are 60 students and 6 TAs. Each TA marks 10 students. Assuming the students all have sat down in groups of ten, how many different combinations of TAs and groups are there? What is your chance of having a particular TA this week?

(h) Assume that there are 6 people who want to randomly split into three teams. For this purpose they put two red, two green and two yellow ribbons into a bag, and each person picks one of those out without looking into the bag.

What is the probability that Amy will be on the red team? What is the chance that she will be on the same team as Zenia? How many different ways of splitting the six members into teams are there?

(i) Students from CSSOC are wearing their hoodies. Four of them have a purple, 2 a green, and one a black one. They line up in a queue to leave the room they are in. What is the probability that all the people in the same colour hoodie are next to each other? What is the probability that no two people wearing a purple hoodie are next to each other?

(j) Assume you have a monkey that randomly types letters on a keyboard. How many keystrokes does the monkey have to perform to have a 1% chance to have typed 'hello world' somewhere in the typed text? You may assume that the keyboard only has lower case letters, and that the space bar is the only non-letter key.

(k) Assume you are at a party. Somebody asks each person when their birthday is. How many people have to be at the party for the probability that two of them share a birthday to be larger than 50%?[4]

[4]This is known as the *birthday paradox*, although it is not strictly a paradox, merely a question with a surprising answer. It is why computer scientist have to worry about *collisions* when designing hash tables.

(l) Assume you are composing a phrase of music over two four beat bars. You may use one octave, and any duration from a quaver (an eighth note) to a semibreve (a whole note). How many melodies are there?

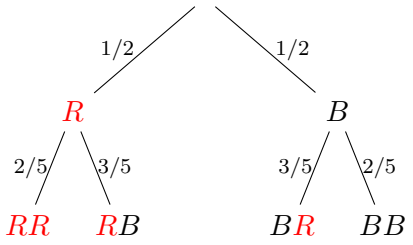
4.1.4 Using trees

Sometimes we can picture what happens in a situation by using trees to provide structure.

Example 4.13 (**Drawing socks**)**.** We may use trees to gain a better understanding of a particular situation. The name 'decision tree' is slightly misleading here since we do not just model decisions that somebody might make but also random moves.

Assume you have a drawer with six individual socks, three red and three black (let's not worry about how you ended up with odd number of socks in both colours). We may answer the question of how many socks we have to pick in order to be sure to get one matched pair—if we pick three socks then there will be at least two which are the same.

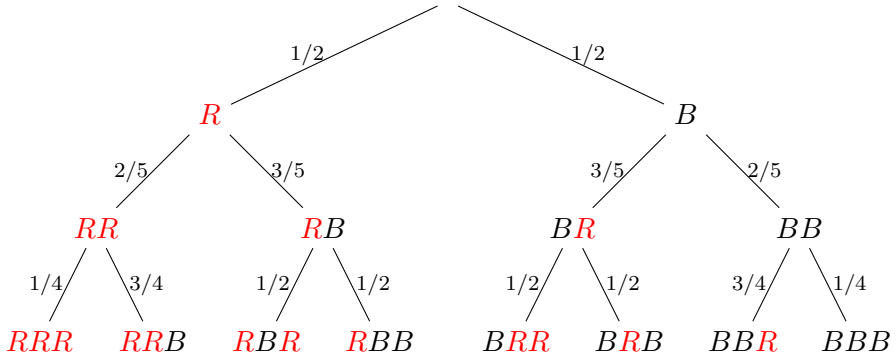
But what if we want to know how many socks we have to pick to have a chance of at least 50% to achieve this? We picture our first two draws as follows.



What is the chance of having two socks of the same colour after two attempts? Of the four possible outcomes two are of the kind we want, namely RR and BB. In order to find out the probability of these two events we *multiply* probabilities as we go down the tree. In other words, the probability of the outcome RR is $2/10$ while that of BB is the same. In order to find the probability of the event $\{RR, BB\}$ we *add* the probabilities of the two outcomes contained, and so we have $4/10 = 2/5$.

Hence in order to guarantee a success rate of at least 50% we have to have (at least) three draws, and in that case we know we will have a 100% success rate.

Let us look at the question of picking at least two black socks. With two draws the chance of succeeding is $2/10 = 1/5$. If we add a third draw we get the following.



The outcomes where we have two black socks are

$$\{RBB, BRB, BBR, BBB\},$$

And if we add up their probabilities we get

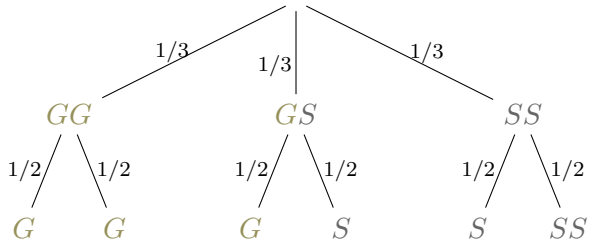
$$3/20 + 3/20 + 6/20 + 1/20 = 13/20 = 65\%,$$

and that gives us a chance of succeeding which is better than 50%.

Example 4.14 (**Gold and Silver**)**.** Assume there are three bags, each with two coins. One has two coins of gold, another two coins of silver and a third one coin of each kind. Somebody randomly picks a bag, and then draws a coin from the bag without looking inside.

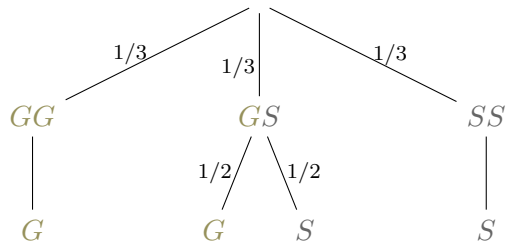
We are shown that the selected coin is gold. What is the chance that the remaining coin from that bag is also gold?

Again we use a tree to understand what is happening.



If we know that a gold coin has been drawn we must be seeing the first, second or third outcome from above. All these are equally likely, with a probability of 1/6 each. Two out of the three have a second coin which is also gold, so the desired probability is 2/3.

Instead of explicitly looking at both coins in the bag, as we did in the tree above, we could have a different event, namely the colour of the drawn coin. If those are our chosen outcomes then the corresponding tree looks like this.



Now we argue that knowing the drawn coin is gold tells us that we have either the first or the second outcome. The former occurs with probability 1/3, the second with probability 1/6 overall, so the former is twice as likely as the latter, again giving a probability of 2/3 that the second coin is also gold.

Example 4.15 (**The Monty Hall problem**)**.** A well-known problem that we may use for illustrative purposes is known as the *Monty Hall problem*.[5]

Imagine you are in a game show. There are three closed doors labelled A, B and C, and you know that behind one of them is a valuable prize (in the original story a car) and behind two of them is something not worth having (in the original story a goat).

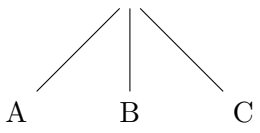
The way the game works is that you pick a door, and then the show master opens one of the remaining doors. You see the booby prize. You are now offered the chance to switch to the other closed doors. Should you switch, or stick with your original choice?

This situation has been endlessly discussed among various groups of people, often because somebody knows the solution and somebody else doesn't want to believe it.

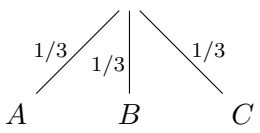
So how does one model a situation like that reliably? Usually when there are steps in a situation it is worth modelling these steps one by one.

What do we know for sure? We know that at the beginning there are three doors, let's assume with two goats and a car. We assume that the probability of the car being behind any one of the doors is the same. From the point of view of the contestant this is like a random event. The production company picks an actual door, and there is no way of telling how they decide which one to hide the main prize behind, but one might hope that they really do pick any door with the probability of 1/3, and that's the assumption the contestant should make. The action of the show master afterwards has to depend on the choice made by the contestant, and we make the additional assumption that if the show master has a choice of opening a door he will open them with equal probability.

We can model the choices step by step using a tree. In the first step we model the fact that the car might be behind any one of the doors.

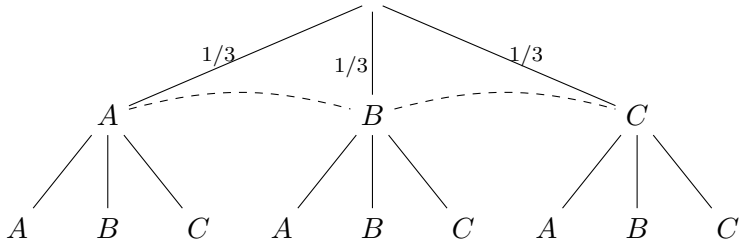


We put probabilities in the tree which indicate that the car can be behind each of them with equal probability.

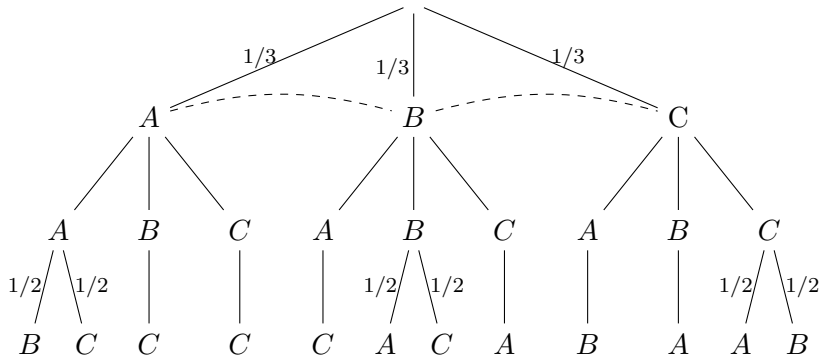


[5]This example also occurs in COMP14112 so we won't look at it in detail in this course unit.

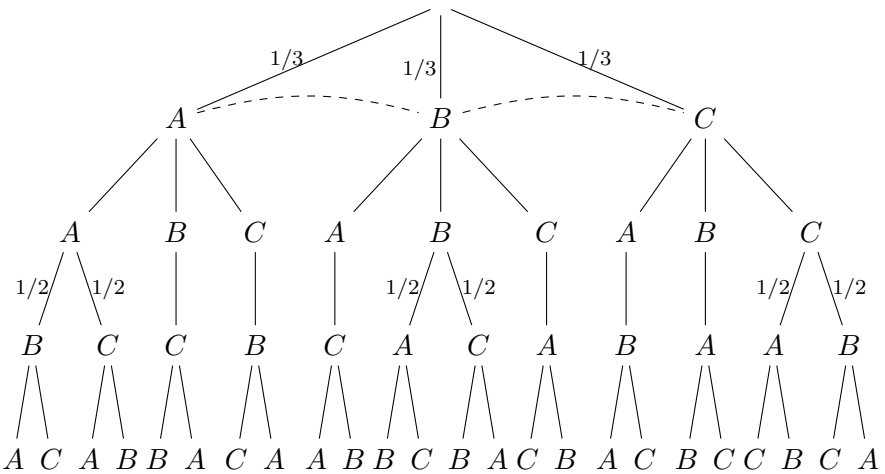
There are three possibilities for the player to choose a door. But note that the player does not know which of the three positions she is in. In game theory one says that the leaves of the tree are in the same *information set*. So from the player's point of view there are three choices (pick door A, B or C), and she cannot make that dependent on where the car is since she does not have that information. This is similar to the situation in many card games where the player has to choose what to play without knowing where all the cards are situated. Only in the course of further play does it become clear what situation the players were in. In the tree we denote this by a dashed line connecting the positions which the player cannot distinguish.



The next step is for the show master to open one of the doors showing the booby prize. In some cases there is only one possible door to open, in others there is a choice between two, and we assume that he picks either one of them with equal probability.

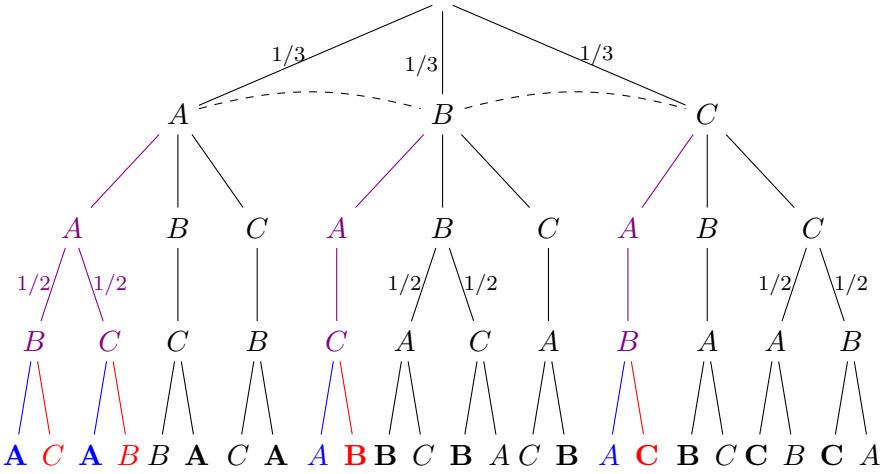


Now the player has to decide whether she wants to switch or not. Again we draw the possible options. We first give the door where the player has not switched, and then the one where she has.



We note that the three principal subtrees are the same up to renaming of nodes. This is because from the Player's point of view, there are only three options (which door to pick), and the first step drawn in the tree (the selection of the door which hides the prize) is completely hidden. The result of that step is not revealed to the player until the end of the game. We look at the question of what happens if the player switches or sticks with her first choice. In purple we highlight the case where the player's first choice is A. The remaining two possibilities give the same result. We also highlight those position where the player wins the main prize by giving it in bold. We now look at two strategies:

- Pick A on the first move and then stick with this choice, given in blue (this is the left choice in the fourth layer of the tree).
- Pick A on the first move and then switch when given the chance to do so, given in red (this is the right choice in the fourth layer of the tree).



If the player picks door A on the first move, and then sticks to that choice, there is a chance of $1/3$ that the original choice was correct, and then no matter which door is opened the main prize is achieved. So a player who does not switch will get the main prize with probability of $1/3$.

A player who picks door A on the first move and then switches, had the correct door with probability $1/3$ and then switches away, but with probability $2/3$ the original door was incorrect, but after the switch the player does gain the main prize, so in that case the player obtains the main prize with probability $2/3$.

For this reason a player who picks a door and then switches has a chance which is twice as high to get the main prize than the player who sticks.

Note that in particular we can see from this example that it may be that what looks like one choice to the player (for example 'pick door A') is effectively a choice taken in a number of different situations the player cannot distinguish between (here the player does not know behind which door the prize is hidden)—in that case the choice will be reflected in several subtrees.

Note that we can tell if a tree properly describes a probabilistic situation:

Proposition 4.16. *A tree with probabilities along some of the edges describes a situation of choices and probabilistic moves if and only if for every node in the tree the probabilities of the edges going down from that node add up to 1.*

Exercise 48. Suppose we have a deck of two cards, $\{Q\spadesuit, A\spadesuit, Q\heartsuit, A\heartsuit\}$. I draw two cards from this pack without being able to see the front of the cards.

You ask me whether one of my cards is an ace, and I answer in the affirmative.

You then tell me to drop one of my cards, making sure I keep an ace in my hand.

You ask me whether I have the ace of spades $A\spadesuit$, and I answer yes.

What is the probability that the card I dropped is also an ace?

Exercise 49. Assume you are throwing two dice, a red and a blue one. Draw a tree that illustrates what happens in this situation.

(a) What is the probability that the sum of the eyes is exactly 4?

(b) What is the probability that the sum of the eyes is at least seven?

(c) What is the probability that there is an even number of eyes visible?

(d) What is the probability that the number on the red die is higher than that on the blue?

4.1.5 Further examples

In the previous sections it was clear from the context which principles you had to apply to find a solution. The point of the following exercises is that you first have to think about what would make sense in the given situation.

Exercise 50. Assume two teams are playing a 'best out of seven' series—that is, the team to win the most out of seven matches is the winner of the series. Once it is clear that one side has won, the remaining matches are no longer played. For example, if one team wins the first four matches the series is over.

(a) Assume that the two teams are equally matched. After what number of matches is the series most likely to end?

(b) How does the answer change if the probability of one team winning is 60%?

Exercise 51. Imagine you have a die that is loaded in that even numbers are twice as likely to occur than odd numbers. Assume that all even numbers are equally likely, as are all odd numbers.

(a) What is the probability of throwing an even number?

(b) What is the probability that the thrown number is at most 4?

(c) With two dice of this kind what is the probability that the combined number of eyes shown is at most 5?

Exercise 52. Assume you have a coin that shows heads half the time and tails the other half, also known as a *fair coin*. Assume the coin is thrown 10 times in a row.

(a) What is the probability that no two successive throws show the same side?

(b) What is the probability that we have exactly half heads and half tails?

(c) What is the chance of having at least five subsequent throws showing the same symbol?

Exercise 53. Assume we toss a fair coin until we see the first heads. We want to record the number of tosses it takes. What is the probability that we require 10 tosses ore more?

4.2 Axioms for probability

In the examples above we have assumed that we know what we mean be 'probability', and that we have some rules for calculating with such numbers.

4.2.1 Overview

This section puts these intuitive ideas onto a firm mathematical footing. It does so in a very general way which you may find difficult to grasp. However by setting this up so generally we give rules that can be applied to *any* situation. Thinking about these rules also encourages you to think about how to model specific situations you are interested in, and to take care with how you do so.

The idea underlying probability theory is that we often find ourselves in a situation where we can work with

- a *sample space* S of all possible outcomes,
- a *set of events* $\mathcal{E}$ (which is a subset of the powerset of S) and
- a *probability distribution* which is given by a function

$$P\colon \mathcal{E} \longrightarrow [0,1],$$

where $[0,1]$ is the interval of real numbers from 0 to 1.

We give precise definitions of what we mean with these notions below, but for the moment let's look at an example.

Example 4.17. In a simple dice game the participants might have two dice which they throw together. If the aim of the game is to score the highest number when adding up the faces of the dice then it makes sense to have the possible outcomes

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

We call the set of possible outcomes the *sample space*.

We could now ask what the probability is of throwing at most 5, which is the event

$$\{2,3,4,5\}.$$

This example is continued below.

Typically when we have a finite set of outcomes S we assume that the set of events $\mathcal{E}$ is the whole powerset $\mathcal{P}S$. When we have an infinite set of outcomes this is not always possible. There is a field in mathematics called *measure theory* which is concerned with which sets of events can be equipped with probability functions, but that goes beyond this course.

Often it is possible to make the sample space finite, and this frequently (but not always) happens for computer science applications. For example in COMP14112 you will be looking at a program that learns the location of a robot in a two-dimensional space. If we think of the location as being given by two coordinates, and the coordinates as real numbers, then there are uncountably many locations in the unit square

$$[0, 1] \times [0, 1].$$

But we cannot measure the location of the robot up to infinite precision (and indeed, we're not interested in the answer to that level of precision), and in the robot exercise a 100×100 grid is imposed on the space, and we are only interested in which one of the squares in the grid the robot inhabits. This means the sample space now has only $100 \cdot 100 = 10,000$ elements.[6]

Similarly if you are interested in the price of a commodity, it typically makes sense to measure the price only up to a limited precision (typically a few post decimal digits), and again this has the effect of making the sample space finite. We concentrate on finite sample spaces for this reason, but we do have a look at infinite notions as well.

Not every function satisfies the requirements of a probability function, and we look at what properties we expect below. In order to formulate what we expect from a probability function P we first have to look at what we expect from the set of events.

4.2.2 Events and probability distributions

The following two definitions are given here for completeness' sake—in the context of this course you will rarely have to think about whether you have a suitable set of events.[7]

Definition 18. Let S be a set. A subset $\mathcal{E}$ of $\mathcal{P}S$ is a *σ-algebra* provided that

- the set S is in $\mathcal{E}$,
- if E is in $\mathcal{E}$ then so is its complement $S \setminus E$ and
- if E_i is in $\mathcal{E}$ for $i \in \mathbb{N}$ their union $\bigcup_{i \in \mathbb{N}} E_i$ is in $\mathcal{E}$.

We note some consequences of this definition. First of all, since S is in $\mathcal{E}$ we may form its complement to get another element of $\mathcal{E}$, and so

$$\emptyset = E \setminus E$$

is in $\mathcal{E}$.

[6] See Example 4.32 for a simplified version of this scenario.

[7] In particular these two definitions are not part of the examinable material.

Further note that the union of a finite number of events must also be an event: If we nave events E_0, E_1, ..., E_n then we can set $E_i = \emptyset$ for $i > n$, and then

$$\bigcup_{i\in\mathbb{N}} E_i = E_0 \cup E_1 \cup \cdots \cup E_n.$$

Note that for every set S the powerset $\mathcal{P}S$ is a σ-algebra. However, it is not always possible to give a probability function that assigns a probability to each subset of S (for example, it is not possible to do this when the sample set is $\mathbb{R}$).

Events which are disjoint play a particular role: If we have two sets of possible outcomes, say E and E', and these sets are disjoint, then we expect that the probability of $E \cup E'$ is the probability of E added to that of E'. But this is not a property of just two sets of outcomes—sometimes we need to apply it to to larger collections of sets. This means we have to worry about what the appropriate generalization of 'disjoint' is.

If we have three sets of outcomes, events E, E' and E'', then in order for

$$P(E \cup E' \cup E'')$$

to be equal to

$$PE + PE' + PE''$$

to hold it must be the case that none of these sets 'overlap', in other words, we need that

$$E \cap E' = \emptyset, \qquad E \cap E'' = \emptyset, \qquad E' \cap E'' = \emptyset.$$

If we want to apply this to even more sets we really need to use a general definition.

Definition 19. Let S be a set. Further assume that we have an arbitrary set I, and that for each element $i \in I$ we have picked a subset S_i of S. We say that the collection of the S_i, where $i \in I$ is **pairwise disjoint** if and only if

$$\text{for } i, j \in \mathbb{N} \text{ we have} \qquad i \neq j \qquad \text{implies} \qquad S_i \cap S_j = \emptyset.$$

This means that the sets we have picked for different elements of I do not overlap.

Exercise 54. Assume that we have two disjoint subsets B_1 and B_2 of S and that we also have a collection E_i, for $i \in \mathbb{N}$ pairwise disjoint subsets of a set S.

(a) Show that for $A \subseteq S$ we have that $A \cap B_1$ and $A \cap B_2$ are disjoint. *If you can do the next part without doing this one you may skip it.*

(b) Show that for $A \subseteq S$ we have that $A \cap E_i$ is a collection of pairwise disjoint sets.

(c) Show that for for $A \subseteq S$ we have that

$$A \cap (B_1 \cup B_2) = (A \cap B_1) \cup (A \cap B_2).$$

If you can do the next part without doing this one you may skip it.

(d) Show that for $A \subseteq S$ we have that

$$A \cap \bigcup_{i \in \mathbb{N}} E_i = \bigcup_{i \in \mathbb{N}} (A \cap E_i).$$

(e) Show that if $A \subseteq B_1 \cup B_2$ then A is the disjoint union of $A \cap B_1$ and $A \cap B_2$. *If you can do the next part without doing this one you may skip it.*

(f) Show that if $A \subseteq \bigcup_{i \in \mathbb{N}} E_i$ then A is the disjoint union of the $A \cap E_i$.

Definition 20. A **probability space** is given by

- a *sample set* S;
- a *set of events* $\mathcal{E} \subseteq \mathcal{P}S$ which is a σ-algebra and
- a *probability distribution*, that is a function

  $$P \colon \mathcal{E} \longrightarrow [0, 1],$$

  with the properties that
  - $PS = 1$ and
  - given E_i, for $i \in \mathbb{N}$, pairwise disjoint[8], then[9]

    $$P(\bigcup_{i \in \mathbb{N}} E_i) = \sum_{i \in \mathbb{N}} P(E_i).$$

These axioms for probability go back to the Russian mathematician *Andrey Kolmogorov* who was trying to determine what the rules are that make probabilities work so well when describing phenomena from the real world. His rules date from 1933. What we have done here is translate them into a more modern setting.

These axioms may seem complicated, but they are quite short, and they have a lot of consequences which you may have learned about when studying probability previously. We look at these in the following section.

Optional Exercise 8. In the definition of a probability distribution we can see an infinite sum. Under which circumstances does it make sense to write something like that? Try to find a probability distribution for the natural numbers, with $P\mathbb{N}$ as the set of events. *Hint: It is sufficient to give probabilities for events of the form* $\{n\}$.

Optional Exercise 9. Can you find a probability space with a sample set being $[0, 1]$? Don't worry too much about defining a σ algebra. Instead assume that every interval of the form $[r, r']$ with $r < r'$ is an events.

Optional Exercise 10. Assume you want to find a probability distribution for the sample space $[0, 1]$ with a σ-algebra which contains all sets of the form $\{r\}$ as events. What can you say about the probabilities of these sets?

[8] Note that some authors write a disjoint union using the addition symbol $+$, and $\sum$ for infinite such unions, but we do not adopt that practice here in case it causes confusion.

[9] Note that below appears a potentially infinite sum, that is, a sum which adds infinitely many numbers. We do not discuss these situations in general in this unit. We say a bit more about how to think of this rule in Definition 23 below.

Example 4.17 (ctd)**.** In the example where two dice are thrown we may use

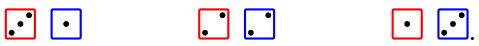
- $S = \{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$ and
- $\mathcal{E} = \mathcal{P}S$

but what is the probability distribution we should use here? Since every subset of S can be written as a disjoint union of sets containing one element each the second condition for probability distributions tells us that it is sufficient to know the probability for each outcome since, for example

$$\begin{aligned} P\{2,3,4,5\} &= P(\{2\} \cup \{3\} \cup \{4\} \cup \{5\}) \\ &= P\{2\} + P\{3\} + P\{4\} + P\{5\}. \end{aligned}$$

This still leaves us with the question of what $P\{2\}$, $P\{3\}$, and so on, should be. If we look at our sample space more closely we find that it in itself can be viewed as a collection of simpler events.

If we look at the outcome 'the sum of the eyes shown by the two dice is 4' then we see that this is a complex event: Assume we have a red die and a blue die, then the following combinations will give the sum of four (giving the red die followed by the blue one):



So we might instead decide that our sample space should look different to make the outcomes as simple as possible to make it easier to determine their probabilities.

IF we record the result of throwing the two dice simultaneously as a pair

$$(i, j),$$

where the first component i tells us the value of the red, and j the value of the blue die. Then our new sample space becomes

$$\{(i, j) \mid 1 \leq i, j \leq 6\}.$$

If we assume that our two dice are both 'fair', that is, every number appears with equal probability then the event of throwing, say, a three with the red die will be $1/6$, as will be the probability for all the other possible outcomes from 1 to 6. The same is true for the blue die. If we now assume that throwing the red die has no effect on the blue die[10] then the probability of each possible outcome

$$(i, j) \qquad \text{is} \qquad \frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36}.$$

The outcomes in our previous sample space are now *events in the new space*, and the probability that the sum thrown is 4, for example, (the old event $\{4\}$) is given by the new event

$$\{(1, 3), (2, 2), (3, 1)\},$$

[10] This property is known as *independence*, see Definition 23.

and its probability is the sum of the probabilities for each singleton, that is

$$\begin{aligned} P\{(1,3),(2,2),(3,1)\} &= P\{(1,3)\} + P\{(2,2)\} + P\{(3,1)\} \\ &= \frac{1}{36} + \frac{1}{36} + \frac{1}{36} \\ &= \frac{3}{36} \\ &= \frac{1}{12}. \end{aligned}$$

For completeness' sake we give a full description of both probability spaces. Because the set of events is the powerset of the sample set it is sufficient to give the probability of each outcome. We begin by describing the second probability space in the somewhat boring table below, where the probability for the outcome (i,j) is the entry in the row labelled i and the column labelled j.

$i\backslash j$	1	2	3	4	5	6
1	1/6	1/6	1/6	1/6	1/6	1/6
2	1/6	1/6	1/6	1/6	1/6	1/6
3	1/6	1/6	1/6	1/6	1/6	1/6
4	1/6	1/6	1/6	1/6	1/6	1/6
5	1/6	1/6	1/6	1/6	1/6	1/6
6	1/6	1/6	1/6	1/6	1/6	1/6

The outcome k from the original space can be thought of as an event in the new space, namely that of

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

and the probability of outcome k in the original space is equal to the probability of the corresponding event in the new space.

Hence the original probability space has a probability distribution determined by the following table:

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

There is a third sample space one could use here: As outcomes use an ordered list $[i,j]$ of numbers, to mean 'the die with the lower number shows i and the die with the higher number shows j'. The whole sample space is then

$$\{[i,j] \mid i,j \in \{1,2,3,4,5,6\},\ i \leq j\},$$

and we give the probabilities for those outcomes below. We give the lower number in the set to determine the row and the higher number for the column.

$i\backslash j$	1	2	3	4	5	6
1	1/6	2/6	2/6	2/6	2/6	2/6
2		1/6	2/6	2/6	2/6	2/6
3			1/6	2/6	2/6	2/6
4				1/6	2/6	2/6
5					1/6	2/6
6						1/6

We learn from this example that there may be more than one suitable sample space, and that by making the possible outcomes as simple as possible we may find their probabilities easier to determine. If the sample space is finite then calculating the probability of any event amounts to adding up the probabilities for the individual outcomes.

So picking a suitable sample space for the problem that one tries to solve is important. It's not unusual to have a number of candidates, but some of them will be easier to describe correctly than others.

4.2.3 Probability spaces for finite sample sets

The above example suggests the idea of the following result. It tells you that if you have a finite sample space then describing a probability space for it can be quite easy.

Proposition 4.18. *Let S be a finite set.*

(i) If for each $s \in S$ we have the probability $ps \in [0,1]$ that s occurs, and the sum of these probabilities is 1, then a probability space is given by

- *the sample space is S,*
- *the set of events is the power set of S, $\mathcal{P}S$,*
- *the probability distribution P is given by*

$$\{s_1, s_2, \ldots, s_n\} \longmapsto ps_1 + ps_2 + \cdots ps_n,$$

where $n \in \mathbb{N}$ and $s_1, s_2, \ldots, s_n \in S$, which means that for for every subset E of S, the probability of E is given by

$$PE = \sum_{s \in E} ps.$$

Moreover this is the only probability space where

- *all sets of the form $\{s\}$ are events and*
- *the probability of the event $\{s\}$ occurring is ps.*

(ii) If $(S, \mathcal{E}, P)$ is a probability space with the property that for $s \in S$, $\{s\} \in \mathcal{E}$ then

- *$\mathcal{E} = \mathcal{P}S$ and*
- *we may read off the probability ps that any given outcome s occurs by considering $P\{s\}$.*

Proof. (i) We have already stated that the powerset of any set is a σ-algebra, so it is sufficient to check that the probability distribution we selected satisfies the required properties.

- We note that the way we have defined the probability distribution, the probability of S is the sum pr the probabilities for the outcomes, and the assumption explicitly stated is that this adds to 1, so $P(S) = 1$.

- If we have pairwise disjoint events E_i for $i \in \mathbb{N}$ then the probability of

$$\bigcup_{i\in\mathbb{N}} E_i$$

is the sum of all the probabilities of elements in this set. But if the E_i are pairwise disjoint then each element of $\bigcup_{i\in\mathbb{N}} E_i$ occurs in exactly one of the E_i, and so

$$\begin{aligned} P(\bigcup_{i\in\mathbb{N}} E_i) &= \sum_{s\in\bigcup_{i\in\mathbb{N}} E_i} Ps && \text{def } P \\ &= \sum_{i\in I}\sum_{s\in E_i} Ps && E_i \text{ pairwise disjoint} \\ &= \sum_{i\in\mathbb{N}} PE_i && \text{def } P.. \end{aligned}$$

(ii) The second statement really has only one property that we need to prove, namely that $\mathcal{P}S$ is the set of events for the given space.

But if S is finite, and all sets of the form $\{s\}$ are events, then for an arbitrary subset S' of S we can list the elements, for example

$$S' = \{s_1, s_2, \ldots s_n\},$$

and by setting

$$E_i = \begin{cases} \{s_i\} & \text{for } 1 \le i \le n \\ \emptyset & \text{else} \end{cases}$$

we have events E_i for $i \in \mathbb{N}$ with the property that

$$S' = \bigcup_{i\in\mathbb{N}} E_i,$$

and since $\mathcal{E}$ is a σ-algebra we know that $S' \in \mathcal{E}$. Hence every subset of S is an event, and so $\mathcal{E} = \mathcal{P}S$. □

Further we learn from the above that if a sample space S is finite then

- the set of events may be $\mathcal{P}S$, the set of all subsets of S and
- in order to know the probability for an arbitrary event it is sufficient to know the probability for each element of the sample space, and we may calculate the probability of the event by adding up the probabilities for all its elements.

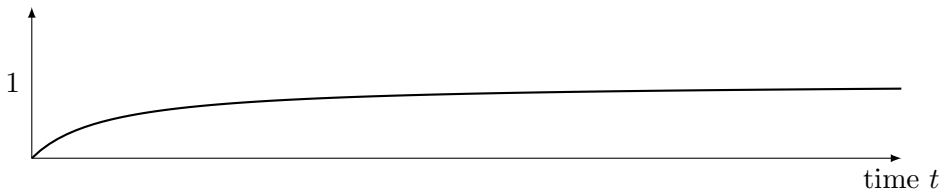
Optional Exercise 11. Extend the statement of Proposition 4.18 to countable sets (these are defined in the next chapter of the notes, but you can find definitions on the web). If you don't want to think about an arbitrary set, think instead of a probability space whose sample set is $\mathbb{N}$.

Exercise 55. Find probability spaces to describe the various situations from Exercises 47 and Exercises 51 to 53. It is fine to describe these in text where you find it difficult to use set and function notation.

4.2.4 Continuous probability distributions

Sometimes it is more appropriate to have a continuous description of a problem. This is often the case when we are plotting events over time. See Definition 26 for a formal definition of what we mean by the discrete versus continuous case here.

Example 4.19. The following curve of a function, say f, might describe[11] the probability that a piece of hardware will have failed by time t.

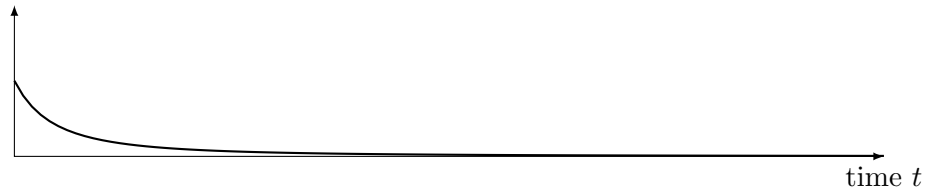


As time progresses the probability of the component having failed approaches 1. But how do we turn this kind of function into a probability space?

We need to identify a set of events, and we need to be able to derive the probability of that event. What we know is how to read off a probability the our device will have failed in the time interval from 0 to t: That probability is given by ft.

This is, in fact, known as a *cumulative probability distribution*: As time progresses the probability becomes higher and higher because the time interval covered becomes bigger and bigger.

In order to give a probability space we need the *probability density function*, which tells us the probability of the device failing at time t. For the function above this is given by the function g plotted below.



The relationship between the two functions is that for all t in $\mathbb{R}^+$ we have

$$ft = \int_{x=0}^{t} gxdx.$$

The reason for this becomes clear in Section 29. It is possible to give a probability space, but the precise description is quite complicated. For completeness' sake we say here that:

There is a σ-algebra $\mathcal{E}_B$ on the set of real numbers $\mathbb{R}$ known as the *Borel σ-algebra* with the property that

- all intervals $[r, r']$, where $r, r' \in \mathbb{R}$, are elements of $\mathcal{E}_B$.

[11] One would that for a real piece of hardware the probability rises more slowly at first!

We may adjust this σ-algebra E_B to subsets of $\mathbb{R}$, for example for our example we may use

$$\mathcal{E}_B^+ = \{E \cap \mathbb{R}^+ \mid E \in \mathcal{E}_B\}.$$

The probability space for the above example is then given by $(\mathbb{R}^+, \mathcal{E}_B^+, P_B)$ where $P_B(g)$ is a probability distribution[12] satisfying

$$P[r, r'] = \int_r^{r'} gxdx.$$

Whenever you are asked to define a continuous probability space you may assume that

- you may use the Borel σ-algebra adjusted as in the above example and
- we can calculate a probability distribution for this σ-algebra from any probability density function.

So it is sufficient for you to give a probability density function in this case.

Definition 21. Let I be a sub-interval of the real numbers. A **probability density function for** I is given by a function

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

with the property that

$$\int_I gxdx = 1,$$

and such that

$$\int_r^{r'} gxdx$$

exists for all $r \leq r'$ in I.

Exercise 56. Describe probability density functions for the fallowing situations:

(a) It is known that the probability of a component having failed rises from 0 to 1 over the time interval from 0 to 1 unit of time at a constant rate.

(b) A bacterium lives for two hours. It is known that its chance of dying in any 10 minute interval during those two hours is the same. What do you think the probability density function should be?

(c) Assume you have an animal which lives in a one dimensional space described by the real line $\mathbb{R}$. Assume that its den is at 0, and that the probability of the animal being at a particular spot r falls at a constant rate and reaches 0 when the animal is one unit away from its den. Give a probability density function for this situation. What does the corresponding cumulative probability distribution look like in this case?

(d) Try to extend the previous part to an animal that lives in a two dimensional space described by the real plane, $\mathbb{R} \times \mathbb{R}$.

[12] One can show that the condition given below uniquely determines $P_B(g)$.

4.2.5 Consequences from Kolmogorov's axioms

The axioms have a number of consequences that are useful to know about.

We look at them one by one here and summarize them in a table at the end of the section.

The empty set

The following properties are consequences of the Kolmogorov axioms.

- The empty set $\emptyset$ is an event since it is the complement of S, and the collection of events must be closed under forming complements, so

  $$\emptyset = S \setminus S$$

  is the reason why this holds.

- We calculate

  $$1 = PS = P(S \cup \emptyset) = PS + P\emptyset = 1 + P\emptyset,$$

  and so

  $$P\emptyset = 0.$$

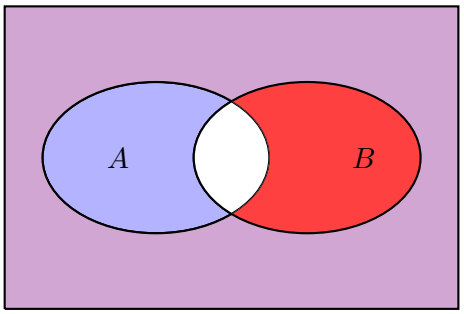
Intersection

If we know that A and B are events, what can we say about $A \cap B$? We note that there is nothing in the axioms that talks about intersections. But it turns out that we can use the axioms to argue that the intersection is an event.

We calculate[13]

$$A \cap B = S \setminus ((S \setminus A) \cup (S \setminus B)).$$

In the following diagram $(S \setminus A) \cup (S \setminus B)$ is the coloured area, and the white part is its complement, that is the desired set.



$(S \setminus A) \cup (S \setminus B)$

Since the complement of an event is an event we know that $S \setminus A$ and $S \setminus B$ are events, and we have seen that the union of a finite number of events is another event.[14]

In general there is no way of calculating the probability of $A \cap B$ from the probabilities of A and B. When the two events are *independent* then this situation changes, see Definition 23.

[13] See Exercise 5.

[14] Note that we can also show that the countable intersection of events is an event by generalizing this idea.

Complement and relative complement

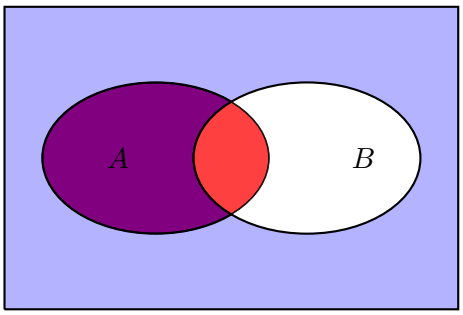
If A is an event then we know that its complement $S \setminus A$ is also an event. We also know that a set and its complement are disjoint sets whose union is S. Hence we know that

$$1 = PS = P(A \cup (S \setminus A)) = PA + P(S \setminus A),$$

and so [15]

$$P(S \setminus A) = 1 - PA.$$

More generally, assume we have events A and B. The picture shows A in red and $S \setminus B$ in blue, with violet giving the overlap. The set whose probability we wish to compute is that overlap.



$A \setminus B$

then

$$A \setminus B = A \cap (S \setminus A)$$

is also an event and we have

$$A = (A \setminus B) \cup (A \cap B),$$

giving us a disjoint union and so (compare Exercise 54

$$PA = P((A \setminus B) \cap (A \cap B)) = P(A \setminus B) + P(A \cap B)$$

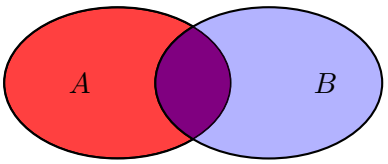
which gives us

$$P(A \setminus B) = PA - P(A \cap B).$$

Union

If we want to calculate the probability of the union of two events then in order to apply Kolmogorov's axiom we must write it as the union of disjoint events.

The Venn diagram for two non-disjoint set looks like this:



$A \cup B$

[15] Some people write this as $P(\neg A) = 1 - PA$ or $P(A^C) = 1 - PA$, but we do not use that notation here.

We can see that if we want to write $A \cup B$ as a disjoint union we have to pick for example the red and violet regions, which make up A, and the blue region, which is $B \setminus A$, and write

$$A \cup B = A \cup (B \setminus A).$$

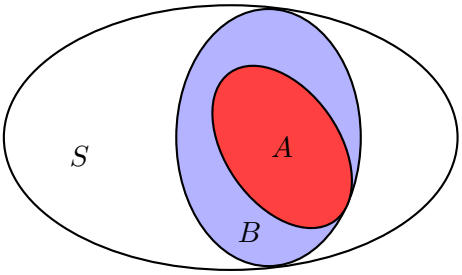
With the result from the previous section that means

$$P(A \cup B) = PA + P(B \setminus A) = PA + PB - P(A \cap B).$$

Note that if A and B do not overlap then $A \cap B = \emptyset$ and we get $P(A \cup B) = PA + PB$ as expected.

Order preservation

Assume we have two events A and B with the property that A is a subset of B.



What can we say about the probabilities of A and B? Certainly we can see that B is the disjoint union of A and $B \setminus A$, and so

$$PB = PA + P(B \setminus A).$$

Since the probability of $B \setminus A$ is greater than or equal to 0 we must have

$$PA \leq PB.$$

Summary

We give all the rules derived above. Let $(S, \mathcal{A}, P)$ be a probability space, and let A and B be events. Then the following rules hold.

$$\begin{aligned} PS &= 1 & P\emptyset &= 0 \\ P(S \setminus A) &= 1 - PA & P(A \setminus B) &= PA - P(A \cap B) \\ P(A \cup B) &= PA + PB - P(A \cap B) \\ A \subseteq B &\text{ implies } PA \leq PB. \end{aligned}$$

It may be worth pointing out that these conditions hold for all probability spaces, in particular they also hold for the case where we are given a probability density function. The first two conditions are trivially true, and the others are standard properties of integrals.

Optional Exercise 12. Convince yourself that the various equalities hold if the probability distribution is given by a probability density function. You may want to draw some pictures for this purpose.

4.2.6 Kolmogorov's axioms revisited

How should we think of the Kolmogorov axioms? The definition of a σ-algebra is something of a formality that ensures that the sets for which we have a probability (namely the *events*) allow us to carry out operations on them.

We may think of the probability distribution as a way of splitting the probability of 1 (which applies to the whole set S) into parts (namely those subsets of S which are events). If S is finite then we only have to know how the probability of 1 is split among the elements of S, and then we can assign a probability to each subset of S by adding up all the probabilities of its elements.

This becomes significantly more complicated if the set is infinite.

Proposition 4.20. *If S is an infinite set then there is no probability distribution which assigns a probability to each element of S which is the same for each element.*

The simplest infinite set we have met is the set of natural numbers $\mathbb{N}$. If we had a probability distribution on $\mathbb{N}$ which assigns a probability to each element then it would have to be the case that the sum of all these probabilities is 1, that is

$$\sum_{i\in\mathbb{N}} P\{i\} = 1$$

and there is no real number r with the property that

$$\sum_{i\in\mathbb{N}} r = 1.$$

If we move from $\mathbb{N}$ as a sample space to $[0,1]$ then there is a way of assigning a uniform probability:

Proposition 4.21. *There is a σ algebra on $[0,1]$ with the property that all intervals $[r,r']$ with $r,r' \in [0,1]$ are events. Further there is a probability distribution on this σ-algebra such that*

$$P[r,r'] = r' - r.$$

This means that every interval of the length r has the same probability, and in that sense this is a uniform distribution.

Proposition 4.22. *There is no probability function P that turns $(\mathbb{R}, \mathcal{P}R, P)$ into a probability space.*

This proposition explains why we cannot have a simpler definition of probability space, where the set of events is always the powerset of the sample space.

4.3 Conditional probabilities and independence

One of the questions that appears frequently in the context of probability theory is that of how information can be used. In other words, can we say something more specific if we already know something about the situation at hand.

4.3.1 Conditional information

For example, if I have to guess the colour of somebody's eyes, but I already know something about the colour of their hair then I can use that information to guide my choice.

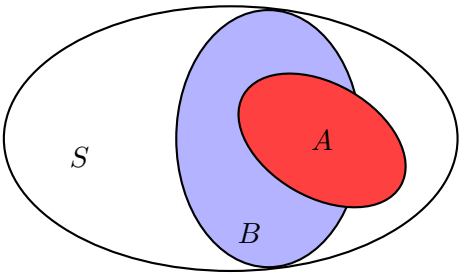
Example 4.23. Let us assume we have a particular part of the population where 56% have dark hair and brown eyes, 14% have dark hair and blue eyes, 3% have fair hair and brown eyes and 27% have fair hair and blue eyes.

If I know a person has been randomly picked from the population, and I have to guess the colour of their eyes, what should I say to have the best chance of being right?

	brown eyes	blue eyes
dark haired	56%	14%
fair haired	3%	27%

We can see from the numbers given that we are better off guessing brown (lacking additional information). But what if we can see that the person in question has fair hair? In that case we are better off guessing blue. What is the appropriate way of expressing these probabilities?

What we are doing here can be pictured by assuming that in the sample space S we have two sets, A and B.



We are interested in the probability of A (say blue eye colour) already knowing that B holds (say fair hair). In the picture above this means the probability that we are in the blue set, provided we already know that we are in the red set.

What we are doing effectively is to change the sample space S to B, and we want to know the probability of $A \cap B$.

Proposition 4.24. *If $(S, \mathcal{E}, P)$ is a probability space and B an event with non-zero probability then a probability space is given by the following data:*

- *sample set B,*
- *set of events*
$$\{B \cap E \mid E \in \mathcal{E}\},$$
- *probability distribution P' defined by*
$$B \cap E \longmapsto \frac{P(B \cap E)}{PB} .$$

We can think of the new space as a restriction of the old space to B, where we have redistributed the probability entirely to the set B, and adjusted all the other probabilities accordingly.

Optional Exercise 13. Define a probability space that is an alternative to the one given in Proposition 4.24. Again assume that you have a probability space $(S, \mathcal{E}, P)$ and a subset B of S with non-zero probability. Use

- sample set S,
- set of events: $\mathcal{E}$,
- a probability density function that assigns to every event of the form $B \cap E$, where $E \in \mathcal{E}$, the same probability as the function given in said proposition.

Optional Exercise 14. Show that the new set of events in Proposition 4.24 is a σ-algebra.

Exercise 57. For the probability distribution P' from Proposition 4.24 carry out the following:

(a) Calculate $P'B$.

(b) For $A \subseteq B$ calculate $P'A$.

(c) Show that P is a probability distribution.

Definition 22. Let $(S, \mathcal{E}, P)$ be a probability space, and let A and B be events, where B has a non-zero probability. We say that the **conditional probability of A given B** is given as

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

It is the probability of the event $A \cap B$ in the probability space based on the restricted sample set B.

Example 4.24 (ctd.). Revisiting the example from above we can see that the probability that a randomly selected person has blue eyes, given that he or she has fair hair, is

$$P(\text{blue eyes} \mid \text{fair hair}) = \frac{P(\text{blue eyes and fair hair})}{P(\text{fair hair})} = \frac{.27}{.3} = .9.$$

In other words, if I am presented with a randomly selected person whose hair I happen to know to be fair then by guessing their eye colour is blue I have a 90% chance of being correct.

On the other hand, if I can see the person has dark hair, then the chance that they have brown eyes is

$$P(\text{brown eyes} \mid \text{dark hair}) = \frac{P(\text{brown eyes and dark hair})}{P(\text{dark hair})} = \frac{.56}{.7} = .8.$$

Hence we can use conditional probabilities to take into account additional information we have been given before making a decision.

Example 4.25. If we revisit Example 4.14 we can see that what we calculated was the probability that we have the bag GG given that we have seen a gold coin. According to the above

$$P(GG \mid G) = \frac{P(GG \cap G)}{PG} = \frac{\frac{1}{3}}{\frac{1}{2}} = \frac{2}{3},$$

just as we concluded above.

Example 4.26. In the Monty Hall problem, Example 4.15, we can think of being shown that there is a booby prize behind one of the doors as adding information. Effectively the show master is asking us: What is the probability that you picked the correct door, knowing that the door I've just shown you is not the correct door? In other words we are interested in the event that

- the price is behind the door the player chose under the condition that
- we were shown the booby prize behind another door.

The probability that the player has chosen the correct door on the first move is $1/3$. The probability that the player chose the incorrect door on the first move is $2/3$, and that is the probability that the prize is hidden behind the door to which the player can switch.

Exercise 58. Assume you have a probability space $(S, \mathcal{E}, P)$ and events A, A' and B. Further assume that $PB \neq 0$.

(a) If you know that $PA \leq PA'$ what can you say about $P(A \mid B)$ and $P(A' \mid B)$?

(b) If you know that $A \cap B = \emptyset$ what can you say about $P(A \mid B)$?

(c) If you know that $A \subseteq B$ what can you say about $P(A \mid B)$?

(d) What is $P(B \mid B)$?

(e) How do PA and $P(A \mid B)$ compare?

In each case justify your answer.

Exercise 59. Assume we know a family with two children.

(a) If we know the family has at least one boy what is the chance that both children are boys?

(b) If we know that the family's firstborn was a boy, what is the probability that both children are boys?

Exercise 60. Go back to Exercise 48 and answer the following questions:

(a) What is the probability that I have at least one ace after the draw?

(b) What is the probability that I have two aces after the draw?

(c) What is the probability that the dropped card is an ace?

(d) What is the probability that I have the ace of spades $A\spadesuit$ given that I dropped a queen?

(e) What is the probability that I have the ace of spades $A\spadesuit$ given that I dropped an ace?

(f) What is the probability that the dropped card was a queen given that I have the ace of spades $A\spadesuit$?

(g) In the original exercise you were asked to calculate the probability that the dropped card was an ace given that I have the ace of spades $A\spadesuit$. Express this using conditional probabilities and recalculate the answer.

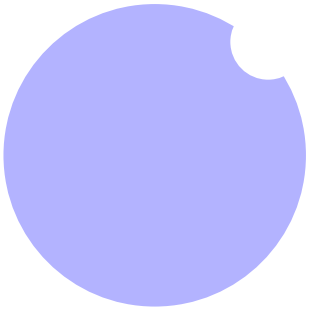
Exercise 61. Assume you have a probability space $(S, \mathcal{E}, P)$, A and B are events, and you know the following:

- $PA\,0$, $PB > 0$, $P(A \cap B) > 0$;
- $P(A \mid B) = P(B \mid A)$ and
- $P(A \cup B) = 1$.

Show that $PA \geq 1/2$. Why is the condition $P(A \cap B) > 0$ needed?

Note that it does make sense to apply the same ideas in the case where we have a probability density function.

Example 4.27. Assume that you have a probability distribution for an animal's location. Further assume that the space in question is centred on the animal's den. Let's assume the animal is a fox, and that we know that its presence is influenced by the presence of another animal, say a lynx. To make the situation simpler let's say that the fox avoids a circle around the lynx.



If we assume the fox avoids the lynx completely then the fox being in the white area of its range, given there is a lynx at the centre of the white circle, has a probability of 0. But that means the 'mass' of probability that resided in the white area has to go somewhere else (since the overall probability that the fox is somewhere in the area has to be equal to 1)!In the above example we haven't got enough information to decide where it goes.

Further if the situation is more interesting, and the lynx only inhibits, but does not prevent, the foxes presence, the analysis is more complicated. We return to this question in Section 4.4 where we restrict how we think of the events that occur, which makes it substantially easier to discuss this.

4.3.2 Equalities for conditional probabilities

From the definition of the conditional probability we may derive some useful equalities.

Recall that the probability of an event conditional on another is defined only if the latter has a probability greater than 0, but the following equality is true even if the probability is 0:

$$P(A \mid B) \cdot PB = P(A \cap B),$$

which is also known as the **multiplication law**.

Note that the expression on the left hand side is symmetric in A and B since $A \cap B = B \cap A$ and so we have

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

If we like we can use this equality to determine $P(B \mid A)$ from $P(A \mid B)$, provided that $PA \neq 0$. The equality

$$P(B \mid A) = \frac{P(A \mid B) \cdot PB}{PA},$$

is known as **Bayes's Theorem**. It allows us to compute the probability of B given A, provided we have the probabilities for A given B, A and B.

Example 4.28. Revisiting Example 4.23 we have calculated the probability that a fair-haired person has blue eyes. What about the probability that a blue-eyed person has fair hair? Using Bayes's law we have

$$\begin{aligned} P(\text{fair hair} \mid \text{blue eyes}) &= \frac{P(\text{blue eyes} \mid \text{fair hair}) \cdot P(\text{fair hair})}{P(\text{blue eyes})} \\ &= \frac{.9 \cdot .3}{.41} \\ &\approx 65.9\%. \end{aligned}$$

On the other hand the probability that a brown-eyed person has dark hair is

$$\begin{aligned} P(\text{dark hair} \mid \text{brown eyes}) &= \frac{P(\text{brown eyes} \mid \text{dark hair}) \cdot P(\text{dark hair})}{P(\text{brown eyes})} \\ &= \frac{.8 \cdot .7}{.59} \\ &\approx 95\%. \end{aligned}$$

There are further equalities based around conditional probabilities that can be useful in practice. Sometimes the sample space can be split into disjoint events, where we know something about those.

In particular, given an even B we know that B and $S \setminus B$ cover the whole sample space S. This means we know that (see Exercise 54)

$$A = (A \cap B) \cup (A \cap (S \setminus B)),$$

and this is a disjoint union. By Kolmogorov's axioms this implies

$$\begin{aligned} PA &= P((A \cap B) \cup (A \cap (S \setminus B))) \\ &= P(A \cap B) + P(A \cap (S \setminus B)), \end{aligned}$$

and if we use the multiplication law twice, and the properties for probability distributions as needed, this is equal to

$$\begin{aligned} &= P(A \mid B) \cdot PB + P(A \mid S \setminus B) \cdot P(S \setminus B) \\ &= P(A \mid B) \cdot PB + P(A \mid S \setminus B) \cdot (1 - PB). \end{aligned}$$

This law is useful, for example, when there is a given property and whether or not that property holds has an influence on whether a second property holds.

Example 4.29. Assume that motherboards from different suppliers have been stored in such a way that it is no longer possible to tell which motherboard came from which supplier.

Further assume that subsequently it has become clear that those from Supplier 1 ($S1$) have a 5% chance of being faulty, while that chance is 10% for ones from Supplier 2 ($S2$). It is known that 70% of supplies in the warehouse came from Supplier 1, and the remainder from Supplier 2. What is the probability that a randomly chosen motherboard is defective?

The law of total probability tells us that

$$\begin{aligned} &P(\text{defect}) \\ &= P(\text{defect} \mid \text{from S1}) \cdot P(\text{from S1}) + P(\text{defect} \mid \text{from S2}) \cdot P(\text{from S"2}) \\ &= .05 \cdot .7 + .1 \cdot .3 \\ &= .065. \end{aligned}$$

Example 4.30. The following is an important case that applies to diagnostic testing in those cases where there is some error (certainly medical tests fall into this category).

Assume a test is being carried out whether some test subject suffers from an undesirable condition. From previous experience it is known that

- if the subject suffers from the condition then with a probability of .99 the test will show this correctly and
- if the subject does not have the condition then with a probability of .95 the test will show this correctly.

We assume that for an arbitrary member of the test population the chance of suffering from the condition is .00001. If a subject tests positive for the condition, what is the probability that they have the condition?

We would like to calculate

$$P(\text{has condition} \mid \text{test positive}).$$

We do not have this data given, but we do have

$$P(\text{test pos} \mid \text{has cond}) \qquad \text{and} \qquad P(\text{has cond}).$$

If we apply Bayes's theorem we get

$$P(\text{has cond} \mid \text{test pos}) = \frac{P(\text{test pos} \mid \text{has cond}) \cdot P(\text{has cond})}{P(\text{test pos})}.$$

We miss

$$P(\text{test pos}),$$

but we may use the law derived above to calculate

$$\begin{aligned}
&P(\text{test pos}) \\
&= P(\text{test pos} \mid \text{has cond}) \cdot P(\text{has cond}) \\
&\quad + P(\text{test pos} \mid \text{doesn't have cond}) \cdot P(\text{doesn't have cond}) \\
&= .99 \cdot .00001 + .05 \cdot .99999 \\
&\approx .05
\end{aligned}$$

So we may calculate the desired probability as

$$\begin{aligned}
P(\text{has condond} \mid \text{test pos}) &= \frac{P(\text{test pos} \mid \text{has condond}) \cdot P(\text{has cond})}{P(\text{test pos})} \\
&\approx \frac{.99 \cdot .00001}{.05} \\
&\approx .0002.
\end{aligned}$$

So if we test something, and in the event it tests positive, there's a .02% chance that the subject is ill, would we think this is a good test?

The issue in this example is the extremely low probability that anybody has the condition at all. If we change the numbers and instead assume that the chance that an arbitrary member of the test population has the condition is .1 instead then we get

$$\begin{aligned}
&P(\text{test pos}) \\
&= P(\text{test pos} \mid \text{has cond}) \cdot P(\text{has cond}) \\
&\quad + P(\text{test pos} \mid \text{doesn't have cond}) \cdot P(\text{doesn't have cond}) \\
&= .99 \cdot .1 + .05 \cdot .9 \\
&= .144
\end{aligned}$$

and

$$\begin{aligned}
P(\text{has cond} \mid \text{test pos}) &= \frac{P(\text{test pos} \mid \text{has cond}) \cdot P(\text{has cond})}{P(\text{test pos})} \\
&= \frac{.99 \cdot .01}{.144} \\
&= .06875.
\end{aligned}$$

So in this case the chance that a subject that tests positive has the condition is almost 69%.

In general when you are given the outcome of a test you should ideally also be given enough data to judge what that information means!

What we have derived is a special case of a more general law. Instead of splitting the sample space into two disjoint sets, B and $S \setminus B$, we split it into more parts. If B_1, B_2,..., B_n is a collection of pairwise disjoint events such that

$$A \subseteq B_1 \cup B_2 \cup \cdots \cup B_n$$

then it is the case that

$$\begin{aligned}
PA &= P(A \mid B_1) \cdot PB_1 + P(A \mid B_2) \cdot PB_2 + \cdots + P(A \mid B_n) \cdot PB_n \\
&= \sum_{i=1}^{n} P(A \mid B_i) \cdot PB_1.
\end{aligned}$$

This is sometimes referred to as the **law of total probability**. The way to think about it is that if we split the event A into disjoint parts of the form

$$A \cap B_i,$$

then the probability of A can be recovered from the probabilities of the parts, and the probabilities of these parts can be calculated using the multiplication law.

This is the law that is used for a procedure known as Bayesian updating which is discussed in the following section. Examples for the application of this rule can be found there.

Exercise 62. Let $(S, \mathcal{E}, P)$ be a probability space, and assume that A, B and C are events. Show that the probability of both A and B holding, given C, is the same as the probability that A holds, given B and C.

Exercise 63. Prove that the law of total probability holds.

Exercise 64. Assume that you have found the following statistical facts about your favourite football team:

- If they score the first goal they win the game with a probability of .65.
- If the other team scores the first goal your team has a probability of .25 of winning the game.
- If your team scores then the probability that the game is a draw is .1.

You have further worked out that in all the matches your team has played, in 55% of all games they have scored, and in half of those they have scored first. What is the probability that your team wins a randomly picked game?

After further analysis you have worked out that they win 45% of all games in which they have scored. What is the probability that a randomly picked game your team is involved in is a draw?

Exercise 65. One of your friend claims she has an unfair coin that shows heads 75% of the time. She gives you a coin, but you can't tell whether it's that one or a fair version.

You toss the coin three times and get HHT. What is the probability that the coin you were given is the unfair one?

Exercise 66. Assume you have an unfair coin that shows heads with probability $p \in (0, 1]$. You toss the coin until heads appears for the first time. Show that the probability that this happens after an even number of tosses is

$$\frac{1-p}{2-p}.$$

This is a tricky exercise. It depends on cleverly choosing events, and using the law of total probability.

Exercise 67. Consider the following situation: Over a channel bits are transmitted. The chance that a bit is incorrectly received is p. From observing previous traffic it is known that the ratio of bits of value 1 to bits of value 0 is 4 to 3.

If the sequence 011 is observed what it the probability that this was transmitted?

COMP11120, Semester 1

Exercise Sheet 7

For examples classes in Week 9

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 47. Do four of the parts, one from (a)–(d), one from (e)–(f). one from (g)–(i) and one from (j) to (l).

Exercise 48.

Exercise 51.

Extensional Exercises marked this week

Exercise 50.

Exercise 54. *This is ahead of the lecture material but only requires calculating with sets. It covers important ideas for material to come.*

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 4.1.

COMP11120, Semester 1

Exercise Sheet 8

For examples classes in Week 10

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 55. Do one from Exercise 47 (a)–(d), one from (e)–(l), and one from Exercises 51 to 53.

Exercise 60.

Exercise 64. Do two of the parts, one from (a)–(c) and one from (d)–(f).

Extensional Exercises marked this week

Exercise 56. Do any two parts.

Exercise 58.

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 Sections 4.2 to Section 4.3.2.