Bayesians also use probabilities to describe *inferences*.

2.3 Forward probabilities and inverse probabilities

Probability calculations often fall into one of two categories: *forward probability* and *inverse probability*. Here is an example of a forward probability problem:

Exercise 2.4.[2, p.40] An urn contains K balls, of which B are black and $W = K - B$ are white. Fred draws a ball at random from the urn and replaces it, N times.

(a) What is the probability distribution of the number of times a black ball is drawn, n_B?

(b) What is the expectation of n_B? What is the variance of n_B? What is the standard deviation of n_B? Give numerical answers for the cases $N = 5$ and $N = 400$, when $B = 2$ and $K = 10$.

Forward probability problems involve a *generative model* that describes a process that is assumed to give rise to some data; the task is to compute the probability distribution or expectation of some quantity that depends on the data. Here is another example of a forward probability problem:

Exercise 2.5.[2, p.40] An urn contains K balls, of which B are black and $W = K - B$ are white. We define the fraction $f_B \equiv B/K$. Fred draws N times from the urn, exactly as in exercise 2.4, obtaining n_B blacks, and computes the quantity

$$z = \frac{(n_B - f_B N)^2}{N f_B (1 - f_B)}. \tag{2.19}$$

What is the expectation of z? In the case $N = 5$ and $f_B = 1/5$, what is the probability distribution of z? What is the probability that $z < 1$? [Hint: compare z with the quantities computed in the previous exercise.]

Like forward probability problems, *inverse probability problems* involve a generative model of a process, but instead of computing the probability distribution of some quantity *produced* by the process, we compute the conditional probability of one or more of the *unobserved variables* in the process, *given* the observed variables. This invariably requires the use of Bayes' theorem.

Example 2.6. There are eleven urns labelled by $u \in \{0, 1, 2, \ldots, 10\}$, each containing ten balls. Urn u contains u black balls and $10 - u$ white balls. Fred selects an urn u at random and draws N times with replacement from that urn, obtaining n_B blacks and $N - n_B$ whites. Fred's friend, Bill, looks on. If after $N = 10$ draws $n_B = 3$ blacks have been drawn, what is the probability that the urn Fred is using is urn u, from Bill's point of view? (Bill doesn't know the value of u.)

Solution. The joint probability distribution of the random variables u and n_B can be written

$$P(u, n_B \,|\, N) = P(n_B \,|\, u, N) P(u). \tag{2.20}$$

From the joint probability of u and n_B, we can obtain the conditional distribution of u given n_B:

$$P(u \,|\, n_B, N) = \frac{P(u, n_B \,|\, N)}{P(n_B \,|\, N)} \tag{2.21}$$

$$= \frac{P(n_B \,|\, u, N) P(u)}{P(n_B \,|\, N)}. \tag{2.22}$$

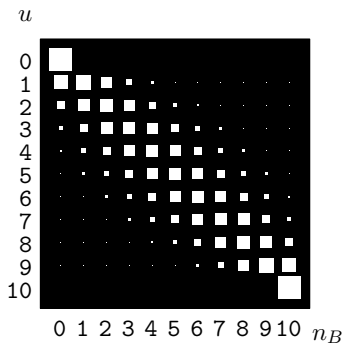


Figure 2.5. Joint probability of u and n_B for Bill and Fred's urn problem, after $N = 10$ draws.

The marginal probability of u is $P(u) = \frac{1}{11}$ for all u. You wrote down the probability of n_B given u and N, $P(n_B \,|\, u, N)$, when you solved exercise 2.4 (p.27). [You *are* doing the highly recommended exercises, aren't you?] If we define $f_u \equiv u/10$ then

$$P(n_B \,|\, u, N) = \binom{N}{n_B} f_u^{n_B} (1 - f_u)^{N - n_B}. \tag{2.23}$$

What about the denominator, $P(n_B \,|\, N)$? This is the marginal probability of n_B, which we can obtain using the sum rule:

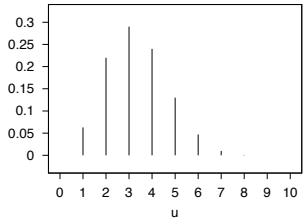
$$P(n_B \,|\, N) = \sum_u P(u, n_B \,|\, N) = \sum_u P(u) P(n_B \,|\, u, N). \tag{2.24}$$

So the conditional probability of u given n_B is

$$P(u \,|\, n_B, N) = \frac{P(u) P(n_B \,|\, u, N)}{P(n_B \,|\, N)} \tag{2.25}$$

$$= \frac{1}{P(n_B \,|\, N)} \frac{1}{11} \binom{N}{n_B} f_u^{n_B} (1 - f_u)^{N - n_B}. \tag{2.26}$$

This conditional distribution can be found by normalizing column 3 of figure 2.5 and is shown in figure 2.6. The normalizing constant, the marginal probability of n_B, is $P(n_B = 3 \,|\, N = 10) = 0.083$. The posterior probability (2.26) is correct for all u, including the end-points $u = 0$ and $u = 10$, where $f_u = 0$ and $f_u = 1$ respectively. The posterior probability that $u = 0$ given $n_B = 3$ is equal to zero, because if Fred were drawing from urn 0 it would be impossible for any black balls to be drawn. The posterior probability that $u = 10$ is also zero, because there are no white balls in that urn. The other hypotheses $u = 1$, $u = 2$, ... $u = 9$ all have non-zero posterior probability. □



u	$P(u \,\|\, n_B = 3, N)$
0	0
1	0.063
2	0.22
3	0.29
4	0.24
5	0.13
6	0.047
7	0.0099
8	0.00086
9	0.0000096
10	0

Figure 2.6. Conditional probability of u given $n_B = 3$ and $N = 10$.

Terminology of inverse probability

In inverse probability problems it is convenient to give names to the probabilities appearing in Bayes' theorem. In equation (2.25), we call the marginal probability $P(u)$ the *prior* probability of u, and $P(n_B \,|\, u, N)$ is called the *likelihood* of u. It is important to note that the terms likelihood and probability are not synonyms. The quantity $P(n_B \,|\, u, N)$ is a function of both n_B and u. For fixed u, $P(n_B \,|\, u, N)$ defines a *probability* over n_B. For fixed n_B, $P(n_B \,|\, u, N)$ defines the *likelihood* of u.

> Never say 'the likelihood of the data'. Always say 'the likelihood of the parameters'. The likelihood function is not a probability distribution.

(If you want to mention the data that a likelihood function is associated with, you may say 'the likelihood of the parameters given the data'.)

The conditional probability $P(u \mid n_B, N)$ is called the *posterior probability* of u given n_B. The normalizing constant $P(n_B \mid N)$ has no u-dependence so its value is not important if we simply wish to evaluate the relative probabilities of the alternative hypotheses u. However, in most data-modelling problems of any complexity, this quantity becomes important, and it is given various names: $P(n_B \mid N)$ is known as the *evidence* or the *marginal likelihood*.

If $\boldsymbol{\theta}$ denotes the unknown parameters, D denotes the data, and $\mathcal{H}$ denotes the overall hypothesis space, the general equation:

$$P(\boldsymbol{\theta} \mid D, \mathcal{H}) = \frac{P(D \mid \boldsymbol{\theta}, \mathcal{H}) P(\boldsymbol{\theta} \mid \mathcal{H})}{P(D \mid \mathcal{H})} \tag{2.27}$$

is written:

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}}. \tag{2.28}$$

Inverse probability and prediction

Example 2.6 (continued). Assuming again that Bill has observed $n_B = 3$ blacks in $N = 10$ draws, let Fred draw another ball from the same urn. What is the probability that the next drawn ball is a black? [You should make use of the posterior probabilities in figure 2.6.]

Solution. By the sum rule,

$$P(\text{ball}_{N+1} \text{ is black} \mid n_B, N) = \sum_u P(\text{ball}_{N+1} \text{ is black} \mid u, n_B, N) P(u \mid n_B, N). \tag{2.29}$$

Since the balls are drawn with replacement from the chosen urn, the probability $P(\text{ball}_{N+1} \text{ is black} \mid u, n_B, N)$ is just $f_u = u/10$, whatever n_B and N are. So

$$P(\text{ball}_{N+1} \text{ is black} \mid n_B, N) = \sum_u f_u P(u \mid n_B, N). \tag{2.30}$$

Using the values of $P(u \mid n_B, N)$ given in figure 2.6 we obtain

$$P(\text{ball}_{N+1} \text{ is black} \mid n_B = 3, N = 10) = 0.333. \qquad \square \tag{2.31}$$

Comment. Notice the difference between this prediction obtained using probability theory, and the widespread practice in statistics of making predictions by first selecting the most plausible hypothesis (which here would be that the urn is urn $u = 3$) and then making the predictions assuming that hypothesis to be true (which would give a probability of 0.3 that the next ball is black). The correct prediction is the one that takes into account the uncertainty by *marginalizing* over the possible values of the hypothesis u. Marginalization here leads to slightly more moderate, less extreme predictions.

Inference as inverse probability

Now consider the following exercise, which has the character of a simple scientific investigation.

Example 2.7. Bill tosses a bent coin N times, obtaining a sequence of heads and tails. We assume that the coin has a probability f_H of coming up heads; we do not know f_H. If n_H heads have occurred in N tosses, what is the probability distribution of f_H? (For example, N might be 10, and n_H might be 3; or, after a lot more tossing, we might have $N = 300$ and $n_H = 29$.) What is the probability that the $N+1$th outcome will be a head, given n_H heads in N tosses?

Unlike example 2.6 (p.27), this problem has a subjective element. Given a restricted definition of probability that says 'probabilities are the frequencies of random variables', this example is different from the eleven-urns example. Whereas the urn u was a random variable, the bias f_H of the coin would not normally be called a random variable. It is just a fixed but unknown parameter that we are interested in. Yet don't the two examples 2.6 and 2.7 seem to have an essential similarity? [Especially when $N = 10$ and $n_H = 3$!]

To solve example 2.7, we have to make an assumption about what the bias of the coin f_H might be. This prior probability distribution over f_H, $P(f_H)$, corresponds to the prior over u in the eleven-urns problem. In that example, the helpful problem definition specified $P(u)$. In real life, we have to make assumptions in order to assign priors; these assumptions will be subjective, and our answers will depend on them. Exactly the same can be said for the other probabilities in our generative model too. We are assuming, for example, that the balls are drawn from an urn independently; but could there not be correlations in the sequence because Fred's ball-drawing action is not perfectly random? Indeed there could be, so the likelihood function that we use depends on assumptions too. In real data modelling problems, priors are subjective *and so are likelihoods*.

Here $P(f)$ denotes a probability density, rather than a probability distribution.

> We are now using $P()$ to denote probability *densities* over continuous variables as well as probabilities over discrete variables and probabilities of logical propositions. The probability that a continuous variable v lies between values a and b (where $b > a$) is defined to be $\int_a^b \mathrm{d}v\, P(v)$. $P(v)\mathrm{d}v$ is dimensionless. The density $P(v)$ is a dimensional quantity, having dimensions inverse to the dimensions of v – in contrast to discrete probabilities, which are dimensionless. Don't be surprised to see probability densities greater than 1. This is normal, and nothing is wrong, as long as $\int_a^b \mathrm{d}v\, P(v) \leq 1$ for any interval (a, b).
>
> Conditional and joint probability densities are defined in just the same way as conditional and joint probabilities.

▷ Exercise 2.8.[2] Assuming a uniform prior on f_H, $P(f_H) = 1$, solve the problem posed in example 2.7 (p.30). Sketch the posterior distribution of f_H and compute the probability that the $N+1$th outcome will be a head, for

(a) $N = 3$ and $n_H = 0$;
(b) $N = 3$ and $n_H = 2$;
(c) $N = 10$ and $n_H = 3$;
(d) $N = 300$ and $n_H = 29$.

You will find the beta integral useful:

$$\int_0^1 \mathrm{d}p_a\, p_a^{F_a}(1-p_a)^{F_b} = \frac{\Gamma(F_a+1)\Gamma(F_b+1)}{\Gamma(F_a+F_b+2)} = \frac{F_a!F_b!}{(F_a+F_b+1)!}. \quad (2.32)$$

You may also find it instructive to look back at example 2.6 (p.27) and equation (2.31).

People sometimes confuse assigning a prior distribution to an unknown parameter such as f_H with making an initial guess of the *value* of the parameter. But the prior over f_H, $P(f_H)$, is not a simple statement like 'initially, I would guess $f_H = 1/2$'. The prior is a probability density over f_H which specifies the prior degree of belief that f_H lies in any interval $(f, f + \delta f)$. It may well be the case that our prior for f_H is symmetric about $1/2$, so that the *mean* of f_H under the prior is $1/2$. In this case, the predictive distribution *for the first toss* x_1 would indeed be

$$P(x_1 = \text{head}) = \int \mathrm{d}f_H \, P(f_H) P(x_1 = \text{head} \,|\, f_H) = \int \mathrm{d}f_H \, P(f_H) f_H = 1/2. \tag{2.33}$$

But the prediction for subsequent tosses will depend on the whole prior distribution, not just its mean.

Data compression and inverse probability

Consider the following task.

Example 2.9. Write a computer program capable of compressing binary files like this one:

```
0000000000000000000010010001000000100000010000000000000000000000000000000001010000000000000110000
1000000000010000100000000010000000000000000000001000000000000000001000000000110000010000000011000100
0000000001001000000000010001000000000000000011000000000000000000000000000010000000000000000100000000
```

The string shown contains $n_1 = 29$ 1s and $n_0 = 271$ 0s.

Intuitively, compression works by taking advantage of the predictability of a file. In this case, the source of the file appears more likely to emit 0s than 1s. A data compression program that compresses this file must, implicitly or explicitly, be addressing the question 'What is the probability that the next character in this file is a 1?'

Do you think this problem is similar in character to example 2.7 (p.30)? I do. One of the themes of this book is that data compression and data modelling are one and the same, and that they should both be addressed, like the urn of example 2.6, using inverse probability. Example 2.9 is solved in Chapter 6.

The likelihood principle

Please solve the following two exercises.

Example 2.10. Urn A contains three balls: one black, and two white; urn B contains three balls: two black, and one white. One of the urns is selected at random and one ball is drawn. The ball is black. What is the probability that the selected urn is urn A?

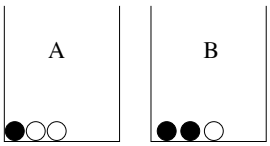


Figure 2.7. Urns for example 2.10.

Example 2.11. Urn A contains five balls: one black, two white, one green and one pink; urn B contains five hundred balls: two hundred black, one hundred white, 50 yellow, 40 cyan, 30 sienna, 25 green, 25 silver, 20 gold, and 10 purple. [One fifth of A's balls are black; two-fifths of B's are black.] One of the urns is selected at random and one ball is drawn. The ball is black. What is the probability that the urn is urn A?

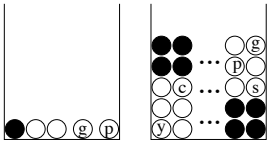


Figure 2.8. Urns for example 2.11.

What do you notice about your solutions? Does each answer depend on the detailed contents of each urn?

The details of the other possible outcomes and their probabilities are irrelevant. All that matters is the probability of the outcome that actually happened (here, that the ball drawn was black) given the different hypotheses. We need only to know the *likelihood*, i.e., how the probability of the data that happened varies with the hypothesis. This simple rule about inference is known as the *likelihood principle*.

> The likelihood principle: given a generative model for data d given parameters $\boldsymbol{\theta}$, $P(d\,|\,\boldsymbol{\theta})$, and having observed a particular outcome d_1, all inferences and predictions should depend only on the function $P(d_1\,|\,\boldsymbol{\theta})$.

In spite of the simplicity of this principle, many classical statistical methods violate it.

2.4 Definition of entropy and related functions

The Shannon information content of an outcome x is defined to be

$$h(x) = \log_2 \frac{1}{P(x)}. \tag{2.34}$$

It is measured in bits. [The word 'bit' is also used to denote a variable whose value is 0 or 1; I hope context will always make clear which of the two meanings is intended.]

In the next few chapters, we will establish that the Shannon information content $h(a_i)$ is indeed a natural measure of the information content of the event $x = a_i$. At that point, we will shorten the name of this quantity to 'the information content'.

The fourth column in table 2.9 shows the Shannon information content of the 27 possible outcomes when a random character is picked from an English document. The outcome $x =$ z has a Shannon information content of 10.4 bits, and $x =$ e has an information content of 3.5 bits.

i	a_i	p_i	$h(p_i)$
1	a	.0575	4.1
2	b	.0128	6.3
3	c	.0263	5.2
4	d	.0285	5.1
5	e	.0913	3.5
6	f	.0173	5.9
7	g	.0133	6.2
8	h	.0313	5.0
9	i	.0599	4.1
10	j	.0006	10.7
11	k	.0084	6.9
12	l	.0335	4.9
13	m	.0235	5.4
14	n	.0596	4.1
15	o	.0689	3.9
16	p	.0192	5.7
17	q	.0008	10.3
18	r	.0508	4.3
19	s	.0567	4.1
20	t	.0706	3.8
21	u	.0334	4.9
22	v	.0069	7.2
23	w	.0119	6.4
24	x	.0073	7.1
25	y	.0164	5.9
26	z	.0007	10.4
27	-	.1928	2.4
		$\sum_i p_i \log_2 \frac{1}{p_i}$	4.1

Table 2.9. Shannon information contents of the outcomes a–z.

The entropy of an ensemble X is defined to be the average Shannon information content of an outcome:

$$H(X) \equiv \sum_{x \in \mathcal{A}_X} P(x) \log \frac{1}{P(x)}, \tag{2.35}$$

with the convention for $P(x) = 0$ that $0 \times \log 1/0 \equiv 0$, since $\lim_{\theta \to 0^+} \theta \log 1/\theta = 0$.

Like the information content, entropy is measured in bits.

When it is convenient, we may also write $H(X)$ as $H(\mathbf{p})$, where $\mathbf{p}$ is the vector $(p_1, p_2, \ldots, p_I)$. Another name for the entropy of X is the uncertainty of X.

Example 2.12. The entropy of a randomly selected letter in an English document is about 4.11 bits, assuming its probability is as given in table 2.9. We obtain this number by averaging $\log 1/p_i$ (shown in the fourth column) under the probability distribution p_i (shown in the third column).

We now note some properties of the entropy function.

- $H(X) \geq 0$ with equality iff $p_i = 1$ for one i. ['iff' means 'if and only if'.]
- Entropy is maximized if $\mathbf{p}$ is uniform:

$$H(X) \leq \log(|\mathcal{A}_X|) \quad \text{with equality iff } p_i = 1/|\mathcal{A}_X| \text{ for all } i. \tag{2.36}$$

Notation: the vertical bars '$|\cdot|$' have two meanings. If $\mathcal{A}_X$ is a set, $|\mathcal{A}_X|$ denotes the number of elements in $\mathcal{A}_X$; if x is a number, then $|x|$ is the absolute value of x.

The *redundancy* measures the fractional difference between $H(X)$ and its maximum possible value, $\log(|\mathcal{A}_X|)$.

The redundancy of X is:

$$1 - \frac{H(X)}{\log |\mathcal{A}_X|}. \tag{2.37}$$

We won't make use of 'redundancy' in this book, so I have not assigned a symbol to it.

The joint entropy of X, Y is:

$$H(X,Y) = \sum_{xy \in \mathcal{A}_X \mathcal{A}_Y} P(x,y) \log \frac{1}{P(x,y)}. \tag{2.38}$$

Entropy is additive for independent random variables:

$$H(X,Y) = H(X) + H(Y) \text{ iff } P(x,y) = P(x)P(y). \tag{2.39}$$

Our definitions for information content so far apply only to discrete probability distributions over finite sets $\mathcal{A}_X$. The definitions can be extended to infinite sets, though the entropy may then be infinite. The case of a probability *density* over a continuous set is addressed in section 11.3. Further important definitions and exercises to do with entropy will come along in section 8.1.

▶ 2.5 Decomposability of the entropy

The entropy function satisfies a recursive property that can be very useful when computing entropies. For convenience, we'll stretch our notation so that we can write $H(X)$ as $H(\mathbf{p})$, where $\mathbf{p}$ is the probability vector associated with the ensemble X.

Let's illustrate the property by an example first. Imagine that a random variable $x \in \{0, 1, 2\}$ is created by first flipping a fair coin to determine whether $x = 0$; then, if x is not 0, flipping a fair coin a second time to determine whether x is 1 or 2. The probability distribution of x is

$$P(x{=}0) = \frac{1}{2}; \quad P(x{=}1) = \frac{1}{4}; \quad P(x{=}2) = \frac{1}{4}. \tag{2.40}$$

What is the entropy of X? We can either compute it by brute force:

$$H(X) = 1/2 \log 2 + 1/4 \log 4 + 1/4 \log 4 = 1.5; \tag{2.41}$$

or we can use the following decomposition, in which the value of x is revealed gradually. Imagine first learning whether $x{=}0$, and then, if x is not 0, learning which non-zero value is the case. The revelation of whether $x{=}0$ or not entails

revealing a binary variable whose probability distribution is $\{1/2, 1/2\}$. This revelation has an entropy $H(1/2, 1/2) = \frac{1}{2}\log 2 + \frac{1}{2}\log 2 = 1$ bit. If x is not 0, we learn the value of the second coin flip. This too is a binary variable whose probability distribution is $\{1/2, 1/2\}$, and whose entropy is 1 bit. We only get to experience the second revelation half the time, however, so the entropy can be written:

$$H(X) = H(1/2, 1/2) + 1/2\, H(1/2, 1/2). \tag{2.42}$$

Generalizing, the observation we are making about the entropy of any probability distribution $\mathbf{p} = \{p_1, p_2, \ldots, p_I\}$ is that

$$H(\mathbf{p}) = H(p_1, 1-p_1) + (1-p_1)H\left(\frac{p_2}{1-p_1}, \frac{p_3}{1-p_1}, \ldots, \frac{p_I}{1-p_1}\right). \tag{2.43}$$

When it's written as a formula, this property looks regrettably ugly; nevertheless it is a simple property and one that you should make use of.

Generalizing further, the entropy has the property for any m that

$$\begin{aligned} H(\mathbf{p}) \;=\; & H\left[(p_1 + p_2 + \cdots + p_m), (p_{m+1} + p_{m+2} + \cdots + p_I)\right] \\ & + (p_1 + \cdots + p_m) H\left(\frac{p_1}{(p_1 + \cdots + p_m)}, \ldots, \frac{p_m}{(p_1 + \cdots + p_m)}\right) \\ & + (p_{m+1} + \cdots + p_I) H\left(\frac{p_{m+1}}{(p_{m+1} + \cdots + p_I)}, \ldots, \frac{p_I}{(p_{m+1} + \cdots + p_I)}\right). \end{aligned} \tag{2.44}$$

Example 2.13. A source produces a character x from the alphabet $\mathcal{A} = \{0, 1, \ldots, 9, a, b, \ldots, z\}$; with probability $1/3$, x is a numeral $(0, \ldots, 9)$; with probability $1/3$, x is a vowel (a, e, i, o, u); and with probability $1/3$ it's one of the 21 consonants. All numerals are equiprobable, and the same goes for vowels and consonants. Estimate the entropy of X.

Solution. $\log 3 + \frac{1}{3}(\log 10 + \log 5 + \log 21) = \log 3 + \frac{1}{3}\log 1050 \simeq \log 30$ bits. □

▶ 2.6 Gibbs' inequality

The 'ei' in L**ei**bler is pronounced the same as in h**ei**st.

The relative entropy *or* Kullback–Leibler divergence between two probability distributions $P(x)$ and $Q(x)$ that are defined over the same alphabet $\mathcal{A}_X$ is

$$D_{\mathrm{KL}}(P||Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)}. \tag{2.45}$$

The relative entropy satisfies *Gibbs' inequality*

$$D_{\mathrm{KL}}(P||Q) \geq 0 \tag{2.46}$$

with equality only if $P = Q$. Note that in general the relative entropy is not symmetric under interchange of the distributions P and Q: in general $D_{\mathrm{KL}}(P||Q) \neq D_{\mathrm{KL}}(Q||P)$, so D_{KL}, although it is sometimes called the 'KL distance', is not strictly a distance. The relative entropy is important in pattern recognition and neural networks, as well as in information theory.

Gibbs' inequality is probably the most important inequality in this book. It, and many other inequalities, can be proved using the concept of convexity.

▶ 2.7 Jensen's inequality for convex functions

The words 'convex $\smile$' and 'concave $\frown$' may be pronounced 'convex-smile' and 'concave-frown'. This terminology has useful redundancy: while one may forget which way up 'convex' and 'concave' are, it is harder to confuse a smile with a frown.

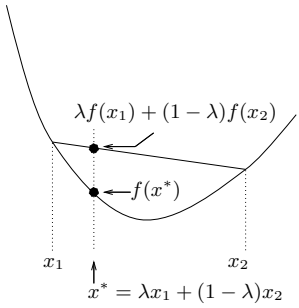


Figure 2.10. Definition of convexity.

Convex $\smile$ functions. A function $f(x)$ is *convex* $\smile$ over (a, b) if every chord of the function lies above the function, as shown in figure 2.10; that is, for all $x_1, x_2 \in (a, b)$ and $0 \leq \lambda \leq 1$,

$$f(\lambda x_1 + (1-\lambda)x_2) \quad \leq \quad \lambda f(x_1) + (1-\lambda) f(x_2). \tag{2.47}$$

A function f is *strictly convex* $\smile$ if, for all $x_1, x_2 \in (a, b)$, the equality holds only for $\lambda = 0$ and $\lambda = 1$.

Similar definitions apply to concave $\frown$ and strictly concave $\frown$ functions.

Some strictly convex $\smile$ functions are

- x^2, e^x and e^{-x} for all x;
- $\log(1/x)$ and $x \log x$ for $x > 0$.

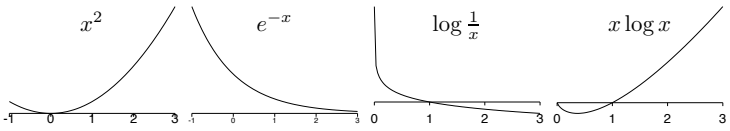


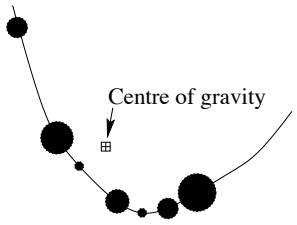
Figure 2.11. Convex $\smile$ functions.

Jensen's inequality. If f is a convex $\smile$ function and x is a random variable then:

$$\mathcal{E}\left[f(x)\right] \geq f(\mathcal{E}[x]), \tag{2.48}$$

where $\mathcal{E}$ denotes expectation. If f is strictly convex $\smile$ and $\mathcal{E}\left[f(x)\right] = f(\mathcal{E}[x])$, then the random variable x is a constant.

Jensen's inequality can also be rewritten for a concave $\frown$ function, with the direction of the inequality reversed.



A physical version of Jensen's inequality runs as follows.

If a collection of masses p_i are placed on a convex $\smile$ curve $f(x)$ at locations $(x_i, f(x_i))$, then the centre of gravity of those masses, which is at $(\mathcal{E}[x], \mathcal{E}\left[f(x)\right])$, lies above the curve.

If this fails to convince you, then feel free to do the following exercise.

Exercise 2.14.[2, p.41] Prove Jensen's inequality.

Example 2.15. Three squares have average area $\bar{A} = 100\,\mathrm{m}^2$. The average of the lengths of their sides is $\bar{l} = 10\,\mathrm{m}$. What can be said about the size of the largest of the three squares? [Use Jensen's inequality.]

Solution. Let x be the length of the side of a square, and let the probability of x be $1/3, 1/3, 1/3$ over the three lengths l_1, l_2, l_3. Then the information that we have is that $\mathcal{E}\left[x\right] = 10$ and $\mathcal{E}\left[f(x)\right] = 100$, where $f(x) = x^2$ is the function mapping lengths to areas. This is a strictly convex $\smile$ function. We notice that the equality $\mathcal{E}\left[f(x)\right] = f(\mathcal{E}[x])$ holds, therefore x is a constant, and the three lengths must all be equal. The area of the largest square is $100\,\mathrm{m}^2$. □