2

Probability, Entropy, and Inference

This chapter, and its sibling, Chapter 8, devote some time to notation. Just as the White Knight distinguished between the song, the name of the song, and what the name of the song was called (Carroll, 1998), we will sometimes need to be careful to distinguish between a random variable, the value of the random variable, and the proposition that asserts that the random variable has a particular value. In any particular chapter, however, I will use the most simple and friendly notation possible, at the risk of upsetting pure-minded readers. For example, if something is 'true with probability 1', I will usually simply say that it is 'true'.

2.1 Probabilities and ensembles

An ensemble X is a triple $(x, \mathcal{A}_X, \mathcal{P}_X)$, where the *outcome* x is the value of a random variable, which takes on one of a set of possible values, $\mathcal{A}_X = \{a_1, a_2, \ldots, a_i, \ldots, a_I\}$, having probabilities $\mathcal{P}_X = \{p_1, p_2, \ldots, p_I\}$, with $P(x = a_i) = p_i$, $p_i \geq 0$ and $\sum_{a_i \in \mathcal{A}_X} P(x = a_i) = 1$.

The name $\mathcal{A}$ is mnemonic for 'alphabet'. One example of an ensemble is a letter that is randomly selected from an English document. This ensemble is shown in figure 2.1. There are twenty-seven possible letters: a–z, and a space character '–'.

i	a_i	p_i
1	a	0.0575
2	b	0.0128
3	c	0.0263
4	d	0.0285
5	e	0.0913
6	f	0.0173
7	g	0.0133
8	h	0.0313
9	i	0.0599
10	j	0.0006
11	k	0.0084
12	l	0.0335
13	m	0.0235
14	n	0.0596
15	o	0.0689
16	p	0.0192
17	q	0.0008
18	r	0.0508
19	s	0.0567
20	t	0.0706
21	u	0.0334
22	v	0.0069
23	w	0.0119
24	x	0.0073
25	y	0.0164
26	z	0.0007
27	–	0.1928

Figure 2.1. Probability distribution over the 27 outcomes for a randomly selected letter in an English language document (estimated from *The Frequently Asked Questions Manual for Linux*). The picture shows the probabilities by the areas of white squares.

Abbreviations. Briefer notation will sometimes be used. For example, $P(x = a_i)$ may be written as $P(a_i)$ or $P(x)$.

Probability of a subset. If T is a subset of $\mathcal{A}_X$ then:

$$P(T) = P(x \in T) = \sum_{a_i \in T} P(x = a_i). \tag{2.1}$$

For example, if we define V to be vowels from figure 2.1, $V = \{\mathtt{a}, \mathtt{e}, \mathtt{i}, \mathtt{o}, \mathtt{u}\}$, then

$$P(V) = 0.06 + 0.09 + 0.06 + 0.07 + 0.03 = 0.31. \tag{2.2}$$

A joint ensemble XY is an ensemble in which each outcome is an ordered pair x, y with $x \in \mathcal{A}_X = \{a_1, \ldots, a_I\}$ and $y \in \mathcal{A}_Y = \{b_1, \ldots, b_J\}$.

We call $P(x, y)$ the joint probability of x and y.

Commas are optional when writing ordered pairs, so $xy \Leftrightarrow x, y$.

N.B. In a joint ensemble XY the two variables are not necessarily independent.

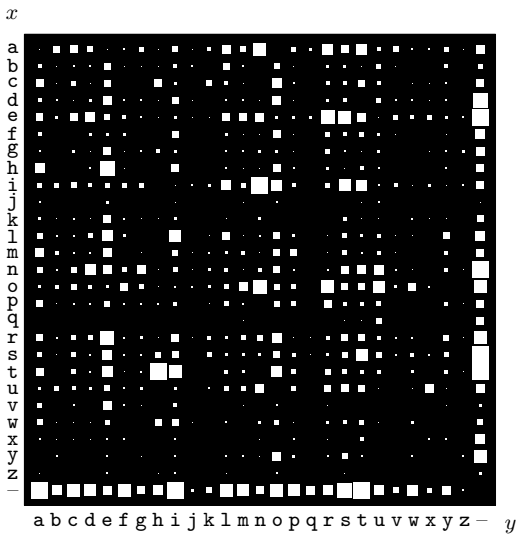


Figure 2.2. The probability distribution over the 27×27 possible bigrams xy in an English language document, *The Frequently Asked Questions Manual for Linux*.

Marginal probability. We can obtain the marginal probability $P(x)$ from the joint probability $P(x, y)$ by summation:

$$P(x = a_i) \equiv \sum_{y \in \mathcal{A}_Y} P(x = a_i, y). \tag{2.3}$$

Similarly, using briefer notation, the marginal probability of y is:

$$P(y) \equiv \sum_{x \in \mathcal{A}_X} P(x, y). \tag{2.4}$$

Conditional probability

$$P(x = a_i \,|\, y = b_j) \equiv \frac{P(x = a_i, y = b_j)}{P(y = b_j)} \quad \text{if} \quad P(y = b_j) \neq 0. \tag{2.5}$$

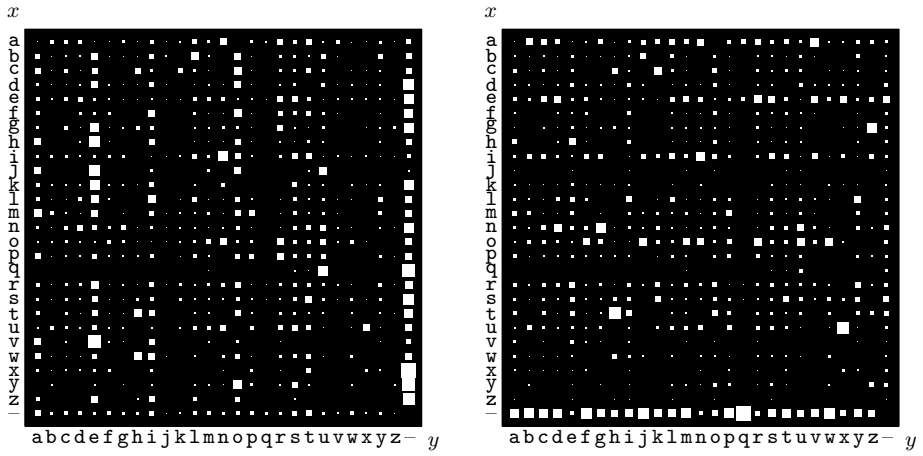
[If $P(y = b_j) = 0$ then $P(x = a_i \,|\, y = b_j)$ is undefined.]

We pronounce $P(x = a_i \,|\, y = b_j)$ 'the probability that x equals a_i, given y equals b_j'.

Example 2.1. An example of a joint ensemble is the ordered pair XY consisting of two successive letters in an English document. The possible outcomes are ordered pairs such as `aa`, `ab`, `ac`, and `zz`; of these, we might expect `ab` and `ac` to be more probable than `aa` and `zz`. An estimate of the joint probability distribution for two neighbouring characters is shown graphically in figure 2.2.

This joint ensemble has the special property that its two marginal distributions, $P(x)$ and $P(y)$, are identical. They are both equal to the monogram distribution shown in figure 2.1.

From this joint ensemble $P(x, y)$ we can obtain conditional distributions, $P(y \,|\, x)$ and $P(x \,|\, y)$, by normalizing the rows and columns, respectively (figure 2.3). The probability $P(y \,|\, x = \mathtt{q})$ is the probability distribution of the second letter given that the first letter is a `q`. As you can see in figure 2.3a, the two most probable values for the second letter y given



(a) $P(y\,|\,x)$ (b) $P(x\,|\,y)$

Figure 2.3. Conditional probability distributions. (a) $P(y\,|\,x)$: Each *row* shows the conditional distribution of the second letter, y, given the first letter, x, in a bigram xy. (b) $P(x\,|\,y)$: Each *column* shows the conditional distribution of the first letter, x, given the second letter, y.

that the first letter x is `q` are `u` and `-`. (The space is common after `q` because the source document makes heavy use of the word `FAQ`.)

The probability $P(x\,|\,y=\texttt{u})$ is the probability distribution of the first letter x given that the second letter y is a `u`. As you can see in figure 2.3b the two most probable values for x given $y=\texttt{u}$ are `n` and `o`.

Rather than writing down the joint probability directly, we often define an ensemble in terms of a collection of conditional probabilities. The following rules of probability theory will be useful. ($\mathcal{H}$ denotes assumptions on which the probabilities are based.)

Product rule – obtained from the definition of conditional probability:

$$P(x, y\,|\,\mathcal{H}) = P(x\,|\,y, \mathcal{H})P(y\,|\,\mathcal{H}) = P(y\,|\,x, \mathcal{H})P(x\,|\,\mathcal{H}). \tag{2.6}$$

This rule is also known as the chain rule.

Sum rule – a rewriting of the marginal probability definition:

$$P(x\,|\,\mathcal{H}) = \sum_y P(x, y\,|\,\mathcal{H}) \tag{2.7}$$

$$= \sum_y P(x\,|\,y, \mathcal{H})P(y\,|\,\mathcal{H}). \tag{2.8}$$

Bayes' theorem – obtained from the product rule:

$$P(y\,|\,x, \mathcal{H}) = \frac{P(x\,|\,y, \mathcal{H})P(y\,|\,\mathcal{H})}{P(x\,|\,\mathcal{H})} \tag{2.9}$$

$$= \frac{P(x\,|\,y, \mathcal{H})P(y\,|\,\mathcal{H})}{\sum_{y'} P(x\,|\,y', \mathcal{H})P(y'\,|\,\mathcal{H})}. \tag{2.10}$$

Independence. Two random variables X and Y are *independent* (sometimes written $X \perp Y$) if and only if

$$P(x, y) = P(x)P(y). \tag{2.11}$$

Exercise 2.2.[1, p.40] Are the random variables X and Y in the joint ensemble of figure 2.2 independent?

I said that we often define an ensemble in terms of a collection of conditional probabilities. The following example illustrates this idea.

Example 2.3. Jo has a test for a nasty disease. We denote Jo's state of health by the variable a and the test result by b.

$$\begin{array}{ll} a=1 & \text{Jo has the disease} \\ a=0 & \text{Jo does not have the disease.} \end{array} \tag{2.12}$$

The result of the test is either 'positive' ($b = 1$) or 'negative' ($b = 0$); the test is 95% reliable: in 95% of cases of people who really have the disease, a positive result is returned, and in 95% of cases of people who do not have the disease, a negative result is obtained. The final piece of background information is that 1% of people of Jo's age and background have the disease.

OK – Jo has the test, and the result was positive. What is the probability that Jo has the disease?

Solution. We write down all the provided probabilities. The test reliability specifies the conditional probability of b given a:

$$\begin{array}{ll} P(b=1\,|\,a=1) = 0.95 & P(b=1\,|\,a=0) = 0.05 \\ P(b=0\,|\,a=1) = 0.05 & P(b=0\,|\,a=0) = 0.95; \end{array} \tag{2.13}$$

and the disease prevalence tells us about the marginal probability of a:

$$P(a=1) = 0.01 \qquad P(a=0) = 0.99. \tag{2.14}$$

From the marginal $P(a)$ and the conditional probability $P(b\,|\,a)$ we can deduce the joint probability $P(a, b) = P(a)P(b\,|\,a)$ and any other probabilities we are interested in. For example, by the sum rule, the marginal probability of $b=1$ – the probability of getting a positive result – is

$$P(b=1) = P(b=1\,|\,a=1)P(a=1) + P(b=1\,|\,a=0)P(a=0). \tag{2.15}$$

Jo has received a positive result $b=1$ and is interested in how plausible it is that she has the disease (i.e., that $a=1$). The man in the street might be duped by the statement 'the test is 95% reliable, so Jo's positive result implies that there is a 95% chance that Jo has the disease', but this is incorrect. The correct solution to an inference problem is found using Bayes' theorem.

$$\begin{aligned} P(a=1\,|\,b=1) &= \frac{P(b=1\,|\,a=1)P(a=1)}{P(b=1\,|\,a=1)P(a=1) + P(b=1\,|\,a=0)P(a=0)} && (2.16) \\ &= \frac{0.95 \times 0.01}{0.95 \times 0.01 + 0.05 \times 0.99} && (2.17) \\ &= 0.16 && (2.18) \end{aligned}$$

So in spite of the positive result, the probability that Jo has the disease is only 16%. □

▶ 2.2 The meaning of probability

Probabilities can be used in two ways.

Probabilities can describe *frequencies of outcomes in random experiments*, but giving noncircular definitions of the terms 'frequency' and 'random' is a challenge – what does it mean to say that the frequency of a tossed coin's

Notation. Let 'the degree of belief in proposition x' be denoted by $B(x)$. The negation of x (NOT-x) is written $\overline{x}$. The degree of belief in a conditional proposition, 'x, assuming proposition y to be true', is represented by $B(x \,|\, y)$.

Axiom 1. Degrees of belief can be ordered; if $B(x)$ is 'greater' than $B(y)$, and $B(y)$ is 'greater' than $B(z)$, then $B(x)$ is 'greater' than $B(z)$.

[Consequence: beliefs can be mapped onto real numbers.]

Axiom 2. The degree of belief in a proposition x and its negation $\overline{x}$ are related. There is a function f such that

$$B(x) = f[B(\overline{x})].$$

Axiom 3. The degree of belief in a conjunction of propositions x, y (x AND y) is related to the degree of belief in the conditional proposition $x \,|\, y$ and the degree of belief in the proposition y. There is a function g such that

$$B(x, y) = g\left[B(x \,|\, y), B(y)\right].$$

Box 2.4. The Cox axioms. If a set of beliefs satisfy these axioms then they can be mapped onto probabilities satisfying $P(\text{FALSE}) = 0$, $P(\text{TRUE}) = 1$, $0 \leq P(x) \leq 1$, and the rules of probability:

$$P(x) = 1 - P(\overline{x}),$$

and

$$P(x, y) = P(x \,|\, y)P(y).$$

coming up heads is $1/2$? If we say that this frequency is the average fraction of heads in long sequences, we have to define 'average'; and it is hard to define 'average' without using a word synonymous to probability! I will not attempt to cut this philosophical knot.

Probabilities can also be used, more generally, to describe *degrees of belief* in propositions that do not involve random variables – for example 'the probability that Mr. S. was the murderer of Mrs. S., given the evidence' (he either was or wasn't, and it's the jury's job to assess how probable it is that he was); 'the probability that Thomas Jefferson had a child by one of his slaves'; 'the probability that Shakespeare's plays were written by Francis Bacon'; or, to pick a modern-day example, 'the probability that a particular signature on a particular cheque is genuine'.

The man in the street is happy to use probabilities in both these ways, but some books on probability restrict probabilities to refer only to frequencies of outcomes in repeatable random experiments.

Nevertheless, degrees of belief *can* be mapped onto probabilities if they satisfy simple consistency rules known as the Cox axioms (Cox, 1946) (figure 2.4). Thus probabilities can be used to describe assumptions, and to describe inferences given those assumptions. The rules of probability ensure that if two people make the same assumptions and receive the same data then they will draw identical conclusions. This more general use of probability to quantify beliefs is known as the *Bayesian* viewpoint. It is also known as the *subjective* interpretation of probability, since the probabilities depend on assumptions. Advocates of a Bayesian approach to data modelling and pattern recognition do not view this subjectivity as a defect, since in their view,

> you cannot do inference without making assumptions.

In this book it will from time to time be taken for granted that a Bayesian approach makes sense, but the reader is warned that this is not yet a globally held view – the field of statistics was dominated for most of the 20th century by non-Bayesian methods in which probabilities are allowed to describe only random variables. The big difference between the two approaches is that

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:

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

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

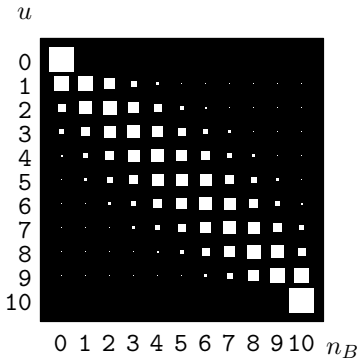


Figure 2.5. Joint probability of u and n_B for Bill and Fred's urn problem, after $N = 10$ draws.

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:

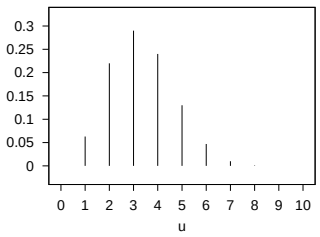
$$P(n_B \,|\, N) = \sum_u P(u, n_B \,|\, N) = \sum_u P(u) P(n_B \,|\, u, N). \tag{2.24}$$

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 $\int_a^b \mathrm{d}v\, P(v) < 1$ 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:

```
0000000000000000000010010001000000100000010000000000000000000000000000000000001010000000000000110000
1000000000100001000000000100000000000000000000001000000000000000001000000000110000010000000110001000
0000000001001000000000010001000000000000000001100000000000000000000000000000100000000000000000100000000
```

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?

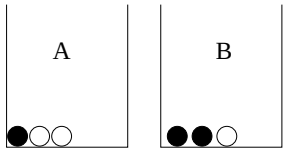


Figure 2.7. Urns for example 2.10.

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?

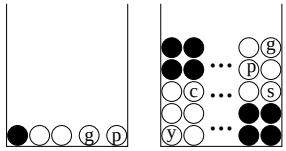


Figure 2.8. Urns for example 2.11.

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 $\boldsymbol{\theta}$, $P(d \mid \boldsymbol{\theta})$, and having observed a particular outcome d_1, all inferences and predictions should depend only on the function $P(d_1 \mid \boldsymbol{\theta})$.

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:

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

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

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

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 Leibler 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 ⌣' and 'concave ⌢' 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.

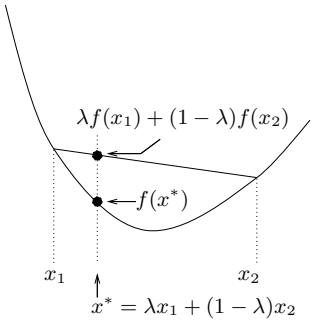


Figure 2.10. Definition of convexity.

Convex ⌣ functions. A function $f(x)$ is *convex* ⌣ 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* ⌣ 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 ⌢ and strictly concave ⌢ functions.

Some strictly convex ⌣ functions are

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

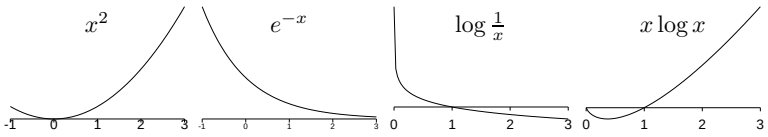


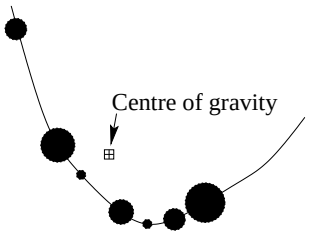
Figure 2.11. Convex ⌣ functions.

Jensen's inequality. If f is a convex ⌣ 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 ⌣ 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 ⌢ 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 ⌣ 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 ⌣ 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 $100\,\mathrm{m}^2$. □

Convexity and concavity also relate to maximization

If $f(\mathbf{x})$ is concave $\frown$ and there exists a point at which

$$\frac{\partial f}{\partial x_k} = 0 \text{ for all } k, \tag{2.49}$$

then $f(\mathbf{x})$ has its maximum value at that point.

The converse does not hold: if a concave $\frown$ $f(\mathbf{x})$ is maximized at some $\mathbf{x}$ it is not necessarily true that the gradient $\nabla f(\mathbf{x})$ is equal to zero there. For example, $f(x) = -|x|$ is maximized at $x = 0$ where its derivative is undefined; and $f(p) = \log(p)$, for a probability $p \in (0, 1)$, is maximized on the boundary of the range, at $p = 1$, where the gradient $\mathrm{d}f(p)/\mathrm{d}p = 1$.

▶ 2.8 Exercises

Sums of random variables

Exercise 2.16.[3, p.41] (a) Two ordinary dice with faces labelled $1, \ldots, 6$ are thrown. What is the probability distribution of the sum of the values? What is the probability distribution of the absolute difference between the values?

(b) One hundred ordinary dice are thrown. What, roughly, is the probability distribution of the sum of the values? Sketch the probability distribution and estimate its mean and standard deviation.

(c) How can two cubical dice be labelled using the numbers $\{0, 1, 2, 3, 4, 5, 6\}$ so that when the two dice are thrown the sum has a uniform probability distribution over the integers 1–12?

(d) Is there any way that one hundred dice could be labelled with integers such that the probability distribution of the sum is uniform?

Inference problems

Exercise 2.17.[2, p.41] If $q = 1 - p$ and $a = \ln p/q$, show that

$$p = \frac{1}{1 + \exp(-a)}. \tag{2.50}$$

Sketch this function and find its relationship to the hyperbolic tangent function $\tanh(u) = \frac{e^u - e^{-u}}{e^u + e^{-u}}$.

It will be useful to be fluent in base-2 logarithms also. If $b = \log_2 p/q$, what is b as a function of p?

▷ Exercise 2.18.[2, p.42] Let x and y be correlated random variables with x a binary variable taking values in $\mathcal{A}_X = \{0, 1\}$. Use Bayes' theorem to show that the log posterior probability ratio for x given y is

$$\log \frac{P(x=1 \mid y)}{P(x=0 \mid y)} = \log \frac{P(y \mid x=1)}{P(y \mid x=0)} + \log \frac{P(x=1)}{P(x=0)}. \tag{2.51}$$

▷ Exercise 2.19.[2, p.42] Let x, d_1 and d_2 be random variables such that d_1 and d_2 are conditionally independent given a binary variable x. Use Bayes' theorem to show that the posterior probability ratio for x given $\{d_i\}$ is

$$\frac{P(x=1 \mid \{d_i\})}{P(x=0 \mid \{d_i\})} = \frac{P(d_1 \mid x=1)}{P(d_1 \mid x=0)} \frac{P(d_2 \mid x=1)}{P(d_2 \mid x=0)} \frac{P(x=1)}{P(x=0)}. \tag{2.52}$$

Life in high-dimensional spaces

Probability distributions and volumes have some unexpected properties in high-dimensional spaces.

Exercise 2.20.[2, p.42] Consider a sphere of radius r in an N-dimensional real space. Show that the fraction of the volume of the sphere that is in the surface shell lying at values of the radius between $r - \epsilon$ and r, where $0 < \epsilon < r$, is:

$$f = 1 - \left(1 - \frac{\epsilon}{r}\right)^N . \tag{2.53}$$

Evaluate f for the cases $N = 2$, $N = 10$ and $N = 1000$, with (a) $\epsilon/r = 0.01$; (b) $\epsilon/r = 0.5$.

Implication: points that are uniformly distributed in a sphere in N dimensions, where N is large, are very likely to be in a thin shell near the surface.

Expectations and entropies

You are probably familiar with the idea of computing the expectation of a function of x,

$$\mathcal{E}\left[f(x)\right] = \langle f(x) \rangle = \sum_x P(x) f(x). \tag{2.54}$$

Maybe you are not so comfortable with computing this expectation in cases where the function $f(x)$ depends on the probability $P(x)$. The next few examples address this concern.

Exercise 2.21.[1, p.43] Let $p_a = 0.1$, $p_b = 0.2$, and $p_c = 0.7$. Let $f(a) = 10$, $f(b) = 5$, and $f(c) = 10/7$. What is $\mathcal{E}\left[f(x)\right]$? What is $\mathcal{E}\left[1/P(x)\right]$?

Exercise 2.22.[2, p.43] For an arbitrary ensemble, what is $\mathcal{E}\left[1/P(x)\right]$?

▷ Exercise 2.23.[1, p.43] Let $p_a = 0.1$, $p_b = 0.2$, and $p_c = 0.7$. Let $g(a) = 0$, $g(b) = 1$, and $g(c) = 0$. What is $\mathcal{E}\left[g(x)\right]$?

▷ Exercise 2.24.[1, p.43] Let $p_a = 0.1$, $p_b = 0.2$, and $p_c = 0.7$. What is the probability that $P(x) \in [0.15, 0.5]$? What is

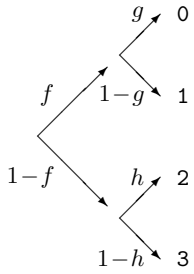
$$P\left(\left|\log \frac{P(x)}{0.2}\right| > 0.05\right)?$$

Exercise 2.25.[3, p.43] Prove the assertion that $H(X) \leq \log(|X|)$ with equality iff $p_i = 1/|X|$ for all i. ($|X|$ denotes the number of elements in the set $\mathcal{A}_X$.) [Hint: use Jensen's inequality (2.48); if your first attempt to use Jensen does not succeed, remember that Jensen involves both a random variable and a function, and you have quite a lot of freedom in choosing these; think about whether your chosen function f should be convex or concave.]

▷ Exercise 2.26.[3, p.44] Prove that the relative entropy (equation (2.45)) satisfies $D_{\mathrm{KL}}(P||Q) \geq 0$ (Gibbs' inequality) with equality only if $P = Q$.

▷ Exercise 2.27.[2] Prove that the entropy is indeed decomposable as described in equations (2.43–2.44).

▷ Exercise 2.28.[2, p.45] A random variable $x \in \{0, 1, 2, 3\}$ is selected by flipping a bent coin with bias f to determine whether the outcome is in $\{0, 1\}$ or $\{2, 3\}$; then either flipping a second bent coin with bias g or a third bent coin with bias h respectively. Write down the probability distribution of x. Use the decomposability of the entropy (2.44) to find the entropy of X. [Notice how compact an expression is obtained if you make use of the binary entropy function $H_2(x)$, compared with writing out the four-term entropy explicitly.] Find the derivative of $H(X)$ with respect to f. [Hint: $\mathrm{d}H_2(x)/\mathrm{d}x = \log((1-x)/x)$.]



▷ Exercise 2.29.[2, p.45] An unbiased coin is flipped until one head is thrown. What is the entropy of the random variable $x \in \{1, 2, 3, \ldots\}$, the number of flips? Repeat the calculation for the case of a biased coin with probability f of coming up heads. [Hint: solve the problem both directly and by using the decomposability of the entropy (2.43).]

▶ 2.9 Further exercises

Forward probability

▷ Exercise 2.30.[1] An urn contains w white balls and b black balls. Two balls are drawn, one after the other, without replacement. Prove that the probability that the first ball is white is equal to the probability that the second is white.

▷ Exercise 2.31.[2] A circular coin of diameter a is thrown onto a square grid whose squares are $b \times b$. ($a < b$) What is the probability that the coin will lie entirely within one square? [Ans: $(1 - a/b)^2$]

▷ Exercise 2.32.[3] Buffon's needle. A needle of length a is thrown onto a plane covered with equally spaced parallel lines with separation b. What is the probability that the needle will cross a line? [Ans, if $a < b$: $2a/\pi b$] [Generalization – Buffon's noodle: on average, a random curve of length A is expected to intersect the lines $2A/\pi b$ times.]

Exercise 2.33.[2] Two points are selected at random on a straight line segment of length 1. What is the probability that a triangle can be constructed out of the three resulting segments?

Exercise 2.34.[2, p.45] An unbiased coin is flipped until one head is thrown. What is the expected number of tails and the expected number of heads?

Fred, who doesn't know that the coin is unbiased, estimates the bias using $\hat{f} \equiv h/(h + t)$, where h and t are the numbers of heads and tails tossed. Compute and sketch the probability distribution of $\hat{f}$.

NB, this is a forward probability problem, a sampling theory problem, not an inference problem. Don't use Bayes' theorem.

Exercise 2.35.[2, p.45] Fred rolls an unbiased six-sided die once per second, noting the occasions when the outcome is a six.

(a) What is the mean number of rolls from one six to the next six?

(b) Between two rolls, the clock strikes one. What is the mean number of rolls until the next six?

(c) Now think back before the clock struck. What is the mean number of rolls, going back in time, until the most recent six?

(d) What is the mean number of rolls from the six before the clock struck to the next six?

(e) Is your answer to (d) different from your answer to (a)? Explain.

Another version of this exercise refers to Fred waiting for a bus at a bus-stop in Poissonville where buses arrive independently at random (a Poisson process), with, on average, one bus every six minutes. What is the average wait for a bus, after Fred arrives at the stop? [6 minutes.] So what is the time between the two buses, the one that Fred just missed, and the one that he catches? [12 minutes.] Explain the apparent paradox. Note the contrast with the situation in Clockville, where the buses are spaced exactly 6 minutes apart. There, as you can confirm, the mean wait at a bus-stop is 3 minutes, and the time between the missed bus and the next one is 6 minutes.

Conditional probability

▷ Exercise 2.36.[2] You meet Fred. Fred tells you he has two brothers, Alf and Bob.

What is the probability that Fred is older than Bob?

Fred tells you that he is older than Alf. Now, what is the probability that Fred is older than Bob? (That is, what is the conditional probability that $F > B$ given that $F > A$?)

▷ Exercise 2.37.[2] The inhabitants of an island tell the truth one third of the time. They lie with probability 2/3.

On an occasion, after one of them made a statement, you ask another 'was that statement true?' and he says 'yes'.

What is the probability that the statement was indeed true?

▷ Exercise 2.38.[2, p.46] Compare two ways of computing the probability of error of the repetition code R_3, assuming a binary symmetric channel (you did this once for exercise 1.2 (p.7)) and confirm that they give the same answer.

Binomial distribution method. Add the probability of all three bits' being flipped to the probability of exactly two bits' being flipped.

Sum rule method. Using the sum rule, compute the marginal probability that $\mathbf{r}$ takes on each of the eight possible values, $P(\mathbf{r})$. $[P(\mathbf{r}) = \sum_s P(s)P(\mathbf{r} \mid s).]$ Then compute the posterior probability of s for each of the eight values of $\mathbf{r}$. [In fact, by symmetry, only two example cases $\mathbf{r} = (000)$ and $\mathbf{r} = (001)$ need be considered.] Notice that some of the inferred bits are better determined than others. From the posterior probability $P(s \mid \mathbf{r})$ you can read out the case-by-case error probability, the probability that the more probable hypothesis is not correct, $P(\text{error} \mid \mathbf{r})$. Find the average error probability using the sum rule,

Equation (1.18) gives the posterior probability of the input s, given the received vector $\mathbf{r}$.

$$P(\text{error}) = \sum_{\mathbf{r}} P(\mathbf{r})P(\text{error} \mid \mathbf{r}). \tag{2.55}$$

▷ Exercise 2.39.[3C, p.46] The frequency p_n of the nth most frequent word in English is roughly approximated by

$$p_n \simeq \begin{cases} \frac{0.1}{n} & \text{for } n \in 1 \ldots 12\,367 \\ 0 & n > 12\,367. \end{cases} \tag{2.56}$$

[This remarkable $1/n$ law is known as Zipf's law, and applies to the word frequencies of many languages (Zipf, 1949).] If we assume that English is generated by picking words at random according to this distribution, what is the entropy of English (per word)? [This calculation can be found in 'Prediction and entropy of printed English', C.E. Shannon, *Bell Syst. Tech. J.* **30**, pp.50–64 (1950), but, inexplicably, the great man made numerical errors in it.]

2.10 Solutions

Solution to exercise 2.2 (p.24). No, they are not independent. If they were then all the conditional distributions $P(y \mid x)$ would be identical functions of y, regardless of x (c.f. figure 2.3).

Solution to exercise 2.4 (p.27). We define the fraction $f_B \equiv B/K$.

(a) The number of black balls has a binomial distribution.

$$P(n_B \mid f_B, N) = \binom{N}{n_B} f_B^{n_B} (1 - f_B)^{N - n_B}. \tag{2.57}$$

(b) The mean and variance of this distribution are:

$$\mathcal{E}[n_B] = N f_B \tag{2.58}$$

$$\mathrm{var}[n_B] = N f_B (1 - f_B). \tag{2.59}$$

These results were derived in example 1.1 (p.1). The standard deviation of n_B is $\sqrt{\mathrm{var}[n_B]} = \sqrt{N f_B (1 - f_B)}$.

When $B/K = 1/5$ and $N = 5$, the expectation and variance of n_B are 1 and 4/5. The standard deviation is 0.89.

When $B/K = 1/5$ and $N = 400$, the expectation and variance of n_B are 80 and 64. The standard deviation is 8.

Solution to exercise 2.5 (p.27). The numerator of the quantity

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

can be recognized as $(n_B - \mathcal{E}[n_B])^2$; the denominator is equal to the variance of n_B (2.59), which is by definition the expectation of the numerator. So the expectation of z is 1. [A random variable like z, which measures the deviation of data from the expected value, is sometimes called χ^2 (chi-squared).]

In the case $N = 5$ and $f_B = 1/5$, $N f_B$ is 1, and $\mathrm{var}[n_B]$ is 4/5. The numerator has five possible values, only one of which is smaller than 1: $(n_B - f_B N)^2 = 0$ has probability $P(n_B = 1) = 0.4096$; so the probability that $z < 1$ is 0.4096.