Solution to exercise 2.14 (p.35). We wish to prove, given the property

$$f(\lambda x_1 + (1-\lambda)x_2) \;\leq\; \lambda f(x_1) + (1-\lambda) f(x_2), \tag{2.60}$$

that, if $\sum p_i = 1$ and $p_i \geq 0$,

$$\sum_{i=1}^{I} p_i f(x_i) \geq f\left(\sum_{i=1}^{I} p_i x_i\right). \tag{2.61}$$

We proceed by recursion, working from the right-hand side. (This proof does not handle cases where some $p_i = 0$; such details are left to the pedantic reader.) At the first line we use the definition of convexity (2.60) with $\lambda = \frac{p_1}{\sum_{i=1}^{I} p_i} = p_1$; at the second line, $\lambda = \frac{p_2}{\sum_{i=2}^{I} p_i}$.

$$\begin{aligned} f\left(\sum_{i=1}^{I} p_i x_i\right) &= f\left(p_1 x_1 + \sum_{i=2}^{I} p_i x_i\right) \\ &\leq\; p_1 f(x_1) + \left[\sum_{i=2}^{I} p_i\right] \left[f\left(\sum_{i=2}^{I} p_i x_i \Big/ \sum_{i=2}^{I} p_i\right)\right] \qquad (2.62) \\ &\leq\; p_1 f(x_1) + \left[\sum_{i=2}^{I} p_i\right] \left[\frac{p_2}{\sum_{i=2}^{I} p_i} f\left(x_2\right) + \frac{\sum_{i=3}^{I} p_i}{\sum_{i=2}^{I} p_i} f\left(\sum_{i=3}^{I} p_i x_i \Big/ \sum_{i=3}^{I} p_i\right)\right], \end{aligned}$$

and so forth. □

Solution to exercise 2.16 (p.36).

(a) For the outcomes $\{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$, the probabilities are $\mathcal{P} = \{\frac{1}{36}, \frac{2}{36}, \frac{3}{36}, \frac{4}{36}, \frac{5}{36}, \frac{6}{36}, \frac{5}{36}, \frac{4}{36}, \frac{3}{36}, \frac{2}{36}, \frac{1}{36}\}$.

(b) The value of one die has mean 3.5 and variance 35/12. So the sum of one hundred has mean 350 and variance $3500/12 \simeq 292$, and by the central-limit theorem the probability distribution is roughly Gaussian (but confined to the integers), with this mean and variance.

(c) In order to obtain a sum that has a uniform distribution we have to start from random variables some of which have a spiky distribution with the probability mass concentrated at the extremes. The unique solution is to have one ordinary die and one with faces 6, 6, 6, 0, 0, 0.

(d) Yes, a uniform distribution can be created in several ways, for example by labelling the rth die with the numbers $\{0, 1, 2, 3, 4, 5\} \times 6^r$.

To think about: does this uniform distribution contradict the central-limit theorem?

Solution to exercise 2.17 (p.36).

$$a = \ln \frac{p}{q} \qquad \Rightarrow \qquad \frac{p}{q} = e^a \tag{2.63}$$

and $q = 1 - p$ gives

$$\frac{p}{1-p} \;=\; e^a \tag{2.64}$$

$$\Rightarrow \qquad p \;=\; \frac{e^a}{e^a + 1} = \frac{1}{1 + \exp(-a)}. \tag{2.65}$$

The hyperbolic tangent is

$$\tanh(a) = \frac{e^a - e^{-a}}{e^a + e^{-a}} \tag{2.66}$$

so

$$f(a) \equiv \frac{1}{1+\exp(-a)} = \frac{1}{2}\left(\frac{1-e^{-a}}{1+e^{-a}}+1\right)$$
$$= \frac{1}{2}\left(\frac{e^{a/2}-e^{-a/2}}{e^{a/2}+e^{-a/2}}+1\right) = \frac{1}{2}(\tanh(a/2)+1). \quad (2.67)$$

In the case $b = \log_2 p/q$, we can repeat steps (2.63–2.65), replacing e by 2, to obtain

$$p = \frac{1}{1+2^{-b}}. \quad (2.68)$$

Solution to exercise 2.18 (p.36).

$$P(x\,|\,y) = \frac{P(y\,|\,x)P(x)}{P(y)} \quad (2.69)$$
$$\Rightarrow \frac{P(x=1\,|\,y)}{P(x=0\,|\,y)} = \frac{P(y\,|\,x=1)}{P(y\,|\,x=0)}\frac{P(x=1)}{P(x=0)} \quad (2.70)$$
$$\Rightarrow \log\frac{P(x=1\,|\,y)}{P(x=0\,|\,y)} = \log\frac{P(y\,|\,x=1)}{P(y\,|\,x=0)} + \log\frac{P(x=1)}{P(x=0)}. \quad (2.71)$$

Solution to exercise 2.19 (p.36). The conditional independence of d_1 and d_2 given x means

$$P(x, d_1, d_2) = P(x)P(d_1\,|\,x)P(d_2\,|\,x). \quad (2.72)$$

This gives a separation of the posterior probability ratio into a series of factors, one for each data point, times the prior probability ratio.

$$\frac{P(x=1\,|\,\{d_i\})}{P(x=0\,|\,\{d_i\})} = \frac{P(\{d_i\}\,|\,x=1)}{P(\{d_i\}\,|\,x=0)}\frac{P(x=1)}{P(x=0)} \quad (2.73)$$
$$= \frac{P(d_1\,|\,x=1)}{P(d_1\,|\,x=0)}\frac{P(d_2\,|\,x=1)}{P(d_2\,|\,x=0)}\frac{P(x=1)}{P(x=0)}. \quad (2.74)$$

Life in high-dimensional spaces

Solution to exercise 2.20 (p.37). The volume of a hypersphere of radius r in N dimensions is in fact

$$V(r,N) = \frac{\pi^{N/2}}{(N/2)!}r^N, \quad (2.75)$$

but you don't need to know this. For this question all that we need is the r-dependence, $V(r,N) \propto r^N$. So the fractional volume in $(r-\epsilon, r)$ is

$$\frac{r^N-(r-\epsilon)^N}{r^N} = 1-\left(1-\frac{\epsilon}{r}\right)^N. \quad (2.76)$$

The fractional volumes in the shells for the required cases are:

N	2	10	1000
$\epsilon/r = 0.01$	0.02	0.096	0.99996
$\epsilon/r = 0.5$	0.75	0.999	$1-2^{-1000}$

Notice that no matter how small ϵ is, for large enough N essentially all the probability mass is in the surface shell of thickness ϵ.

Solution to exercise 2.21 (p.37). $p_a = 0.1$, $p_b = 0.2$, $p_c = 0.7$. $f(a) = 10$, $f(b) = 5$, and $f(c) = 10/7$.

$$\mathcal{E}\left[f(x)\right] = 0.1 \times 10 + 0.2 \times 5 + 0.7 \times 10/7 = 3. \tag{2.77}$$

For each x, $f(x) = 1/P(x)$, so

$$\mathcal{E}\left[1/P(x)\right] = \mathcal{E}\left[f(x)\right] = 3. \tag{2.78}$$

Solution to exercise 2.22 (p.37). For general X,

$$\mathcal{E}\left[1/P(x)\right] = \sum_{x \in \mathcal{A}_X} P(x) 1/P(x) = \sum_{x \in \mathcal{A}_X} 1 = |\mathcal{A}_X|. \tag{2.79}$$

Solution to exercise 2.23 (p.37). $p_a = 0.1$, $p_b = 0.2$, $p_c = 0.7$. $g(a) = 0$, $g(b) = 1$, and $g(c) = 0$.

$$\mathcal{E}\left[g(x)\right] = p_b = 0.2. \tag{2.80}$$

Solution to exercise 2.24 (p.37).

$$P\left(P(x) \in [0.15, 0.5]\right) = p_b = 0.2. \tag{2.81}$$

$$P\left(\left|\log \frac{P(x)}{0.2}\right| > 0.05\right) = p_a + p_c = 0.8. \tag{2.82}$$

Solution to exercise 2.25 (p.37). This type of question can be approached in two ways: either by differentiating the function to be maximized, finding the maximum, and proving it is a global maximum; this strategy is somewhat risky since it is possible for the maximum of a function to be at the boundary of the space, at a place where the derivative is not zero. Alternatively, a carefully chosen inequality can establish the answer. The second method is much neater.

Proof by differentiation (not the recommended method). Since it is slightly easier to differentiate $\ln 1/p$ than $\log_2 1/p$, we temporarily define $H(X)$ to be measured using natural logarithms, thus scaling it down by a factor of $\log_2 e$.

$$H(X) = \sum_i p_i \ln \frac{1}{p_i} \tag{2.83}$$

$$\frac{\partial H(X)}{\partial p_i} = \ln \frac{1}{p_i} - 1 \tag{2.84}$$

we maximize subject to the constraint $\sum_i p_i = 1$ which can be enforced with a Lagrange multiplier:

$$G(\mathbf{p}) \equiv H(X) + \lambda \left(\sum_i p_i - 1 \right) \tag{2.85}$$

$$\frac{\partial G(\mathbf{p})}{\partial p_i} = \ln \frac{1}{p_i} - 1 + \lambda. \tag{2.86}$$

At a maximum,

$$\ln \frac{1}{p_i} - 1 + \lambda = 0 \tag{2.87}$$

$$\Rightarrow \ln \frac{1}{p_i} = 1 - \lambda, \tag{2.88}$$

so all the p_i are equal. That this extremum is indeed a maximum is established by finding the curvature:

$$\frac{\partial^2 G(\mathbf{p})}{\partial p_i \partial p_j} = -\frac{1}{p_i} \delta_{ij}, \tag{2.89}$$

which is negative definite. □

Proof using Jensen's inequality (recommended method). First a reminder of the inequality.

If f is a convex $\smile$ function and x is a random variable then:

$$\mathcal{E}\left[f(x)\right] \geq f\left(\mathcal{E}[x]\right).$$

If f is strictly convex $\smile$ and $\mathcal{E}\left[f(x)\right] = f\left(\mathcal{E}[x]\right)$, then the random variable x is a constant (with probability 1).

The secret of a proof using Jensen's inequality is to choose the right function and the right random variable. We could define

$$f(u) = \log \frac{1}{u} = -\log u \tag{2.90}$$

(which is a convex function) and think of $H(X) = \sum p_i \log \frac{1}{p_i}$ as the mean of $f(u)$ where $u = P(x)$, but this would not get us there – it would give us an inequality in the wrong direction. If instead we define

$$u = 1/P(x) \tag{2.91}$$

then we find:

$$H(X) = -\mathcal{E}\left[f(1/P(x))\right] \leq -f\left(\mathcal{E}[1/P(x)]\right); \tag{2.92}$$

now we know from exercise 2.22 (p.37) that $\mathcal{E}[1/P(x)] = |\mathcal{A}_X|$, so

$$H(X) \leq -f\left(|\mathcal{A}_X|\right) = \log |\mathcal{A}_X|. \tag{2.93}$$

Equality holds only if the random variable $u = 1/P(x)$ is a constant, which means $P(x)$ is a constant for all x. □

Solution to exercise 2.26 (p.37).

$$D_{\mathrm{KL}}(P||Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)}. \tag{2.94}$$

We prove Gibbs' inequality using Jensen's inequality. Let $f(u) = \log 1/u$ and $u = \frac{Q(x)}{P(x)}$. Then

$$\begin{aligned} D_{\mathrm{KL}}(P||Q) &= \mathcal{E}[f(Q(x)/P(x))] & (2.95)\\ &\geq f\left(\sum_x P(x) \frac{Q(x)}{P(x)}\right) = \log\left(\frac{1}{\sum_x Q(x)}\right) = 0, & (2.96) \end{aligned}$$

with equality only if $u = \frac{Q(x)}{P(x)}$ is a constant, that is, if $Q(x) = P(x)$. □

Second solution. In the above proof the expectations were with respect to the probability distribution $P(x)$. A second solution method uses Jensen's inequality with $Q(x)$ instead. We define $f(u) = u \log u$ and let $u = \frac{P(x)}{Q(x)}$. Then

$$\begin{aligned} D_{\mathrm{KL}}(P||Q) &= \sum_x Q(x) \frac{P(x)}{Q(x)} \log \frac{P(x)}{Q(x)} = \sum_x Q(x) f\left(\frac{P(x)}{Q(x)}\right) & (2.97)\\ &\geq f\left(\sum_x Q(x) \frac{P(x)}{Q(x)}\right) = f(1) = 0, & (2.98) \end{aligned}$$

with equality only if $u = \frac{P(x)}{Q(x)}$ is a constant, that is, if $Q(x) = P(x)$. □

Solution to exercise 2.28 (p.38).

$$H(X) = H_2(f) + fH_2(g) + (1-f)H_2(h). \tag{2.99}$$

Solution to exercise 2.29 (p.38). The probability that there are $x-1$ tails and then one head (so we get the first head on the xth toss) is

$$P(x) = (1-f)^{x-1}f. \tag{2.100}$$

If the first toss is a tail, the probability distribution for the future looks just like it did before we made the first toss. Thus we have a recursive expression for the entropy:

$$H(X) = H_2(f) + (1-f)H(X). \tag{2.101}$$

Rearranging,

$$H(X) = H_2(f)/f. \tag{2.102}$$

Solution to exercise 2.34 (p.38). The probability of the number of tails t is

$$P(t) = \left(\frac{1}{2}\right)^t \frac{1}{2} \text{ for } t \geq 0. \tag{2.103}$$

The expected number of heads is 1, by definition of the problem. The expected number of tails is

$$\mathcal{E}[t] = \sum_{t=0}^{\infty} t \left(\frac{1}{2}\right)^t \frac{1}{2}, \tag{2.104}$$

which may be shown to be 1 in a variety of ways. For example, since the situation after one tail is thrown is equivalent to the opening situation, we can write down the recurrence relation

$$\mathcal{E}[t] = \frac{1}{2}(1 + \mathcal{E}[t]) + \frac{1}{2}0 \;\Rightarrow\; \mathcal{E}[t] = 1. \tag{2.105}$$

The probability distribution of the 'estimator' $\hat{f} = 1/(1+t)$, given that $f = 1/2$, is plotted in figure 2.12. The probability of $\hat{f}$ is simply the probability of the corresponding value of t.

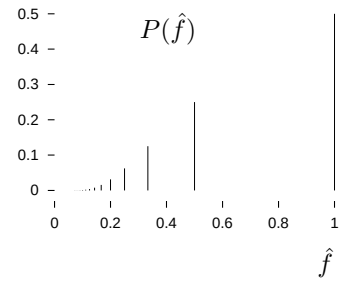

Figure 2.12. The probability distribution of the estimator $\hat{f} = 1/(1+t)$, given that $f = 1/2$.

Solution to exercise 2.35 (p.38).

(a) The mean number of rolls from one six to the next six is six (assuming we start counting rolls after the first of the two sixes). The probability that the next six occurs on the rth roll is the probability of *not* getting a six for $r-1$ rolls multiplied by the probability of then getting a six:

$$P(r_1 = r) = \left(\frac{5}{6}\right)^{r-1} \frac{1}{6}, \text{ for } r \in \{1, 2, 3, \ldots\}. \tag{2.106}$$

This probability distribution of the number of rolls, r, may be called an exponential distribution, since

$$P(r_1 = r) = e^{-\alpha r}/Z, \tag{2.107}$$

where $\alpha = \ln(6/5)$, and Z is a normalizing constant.

(b) The mean number of rolls from the clock until the next six is six.

(c) The mean number of rolls, going back in time, until the most recent six is six.

(d) The mean number of rolls from the six before the clock struck to the six after the clock struck is the sum of the answers to (b) and (c), less one, that is, eleven.

(e) Rather than explaining the difference between (a) and (d), let me give another hint. Imagine that the buses in Poissonville arrive independently at random (a Poisson process), with, on average, one bus every six minutes. Imagine that passengers turn up at bus-stops at a uniform rate, and are scooped up by the bus without delay, so the interval between two buses remains constant. Buses that follow gaps bigger than six minutes become overcrowded. The passengers' representative complains that two-thirds of all passengers found themselves on overcrowded buses. The bus operator claims, 'no, no – only one third of our buses are overcrowded'. Can both these claims be true?

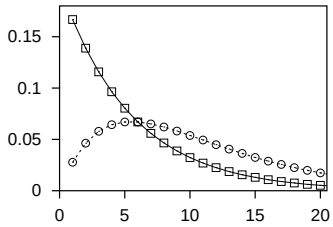

Figure 2.13. The probability distribution of the number of rolls r_1 from one 6 to the next (falling solid line),

$$P(r_1 = r) = \left(\frac{5}{6}\right)^{r-1} \frac{1}{6},$$

and the probability distribution (dashed line) of the number of rolls from the 6 before 1pm to the next 6, $r_{\rm tot}$,

$$P(r_{\rm tot} = r) = r \left(\frac{5}{6}\right)^{r-1} \left(\frac{1}{6}\right)^2.$$

The probability $P(r_1 > 6)$ is about 1/3; the probability $P(r_{\rm tot} > 6)$ is about 2/3. The mean of r_1 is 6, and the mean of $r_{\rm tot}$ is 11.

Solution to exercise 2.38 (p.39).

Binomial distribution method. From the solution to exercise 1.2, $p_B = 3f^2(1-f) + f^3$.

Sum rule method. The marginal probabilities of the eight values of $\mathbf{r}$ are illustrated by:

$$P(\mathbf{r} = \mathtt{000}) = 1/2(1-f)^3 + 1/2 f^3, \tag{2.108}$$

$$P(\mathbf{r} = \mathtt{001}) = 1/2 f(1-f)^2 + 1/2 f^2(1-f) = 1/2 f(1-f). \tag{2.109}$$

The posterior probabilities are represented by

$$P(s = \mathtt{1} \,|\, \mathbf{r} = \mathtt{000}) = \frac{f^3}{(1-f)^3 + f^3} \tag{2.110}$$

and

$$P(s = \mathtt{1} \,|\, \mathbf{r} = \mathtt{001}) = \frac{(1-f)f^2}{f(1-f)^2 + f^2(1-f)} = f. \tag{2.111}$$

The probabilities of error in these representative cases are thus

$$P(\text{error} \,|\, \mathbf{r} = \mathtt{000}) = \frac{f^3}{(1-f)^3 + f^3} \tag{2.112}$$

and

$$P(\text{error} \,|\, \mathbf{r} = \mathtt{001}) = f. \tag{2.113}$$

Notice that while the average probability of error of R_3 is about $3f^2$, the probability (given $\mathbf{r}$) that any *particular* bit is wrong is either about f^3 or f.

The average error probability, using the sum rule, is

$$\begin{aligned} P(\text{error}) &= \sum_{\mathbf{r}} P(\mathbf{r}) P(\text{error} \,|\, \mathbf{r}) \\ &= 2[1/2(1-f)^3 + 1/2 f^3] \frac{f^3}{(1-f)^3 + f^3} + 6[1/2 f(1-f)]f. \end{aligned}$$

The first two terms are for the cases $\mathbf{r} = \mathtt{000}$ and $\mathtt{111}$; the remaining 6 are for the other outcomes, which share the same probability of occurring and identical error probability, f.

So

$$P(\text{error}) = f^3 + 3f^2(1-f).$$

Solution to exercise 2.39 (p.40). The entropy is 9.7 bits per word.