David J. C. MacKay

Information Theory, Inference, and Learning Algorithms

CAMBRIDGE

Information Theory, Inference, and Learning Algorithms

David J.C. MacKay

CAMBRIDGE
UNIVERSITY PRESS

PUBLISHED BY THE PRESS SYNDICATE OF THE UNIVERSITY OF CAMBRIDGE
The Pitt Building, Trumpington Sreet, Cambridge, United Kingdom

CAMBRIDGE UNIVERSITY PRESS
The Edinburgh Building, Cambridge CB2 2RU, UK
40 West 20th Sreet, NewYork, NY 10011-4211, USA
477 Williamstown Road, Port Melbourne, VIC 3207, Australia
Ruiz de Alarcón 13, 28014 Madrid, Spain
Dock House, The Waterfront, Cape Town 8001, South Africa

http://www.cambridge.org

First published 2003
Reprinted with corrections 2004 (twice)

Printed in the United Kingdom at the University Press, Cambridge

Typeset in Computer Modern

A record for the book is available from the British Library

ISBN 0 521 64298 1

Contents

Preface

This book is aimed at senior undergraduates and graduate students in Engineering, Science, Mathematics, and Computing. It expects familiarity with calculus, probability theory, and linear algebra as taught in a first- or second-year undergraduate course on mathematics for scientists and engineers.

Conventional courses on information theory cover not only the beautiful *theoretical* ideas of Shannon, but also *practical* solutions to communication problems. This book goes further, bringing in Bayesian data modelling, Monte Carlo methods, variational methods, clustering algorithms, and neural networks.

Why unify information theory and machine learning? Because they are two sides of the same coin. In the 1960s, a single field, cybernetics, was populated by information theorists, computer scientists, and neuroscientists, all studying common problems. Information theory and machine learning still belong together. Brains are the ultimate compression and communication systems. And the state-of-the-art algorithms for both data compression and error-correcting codes use the same tools as machine learning.

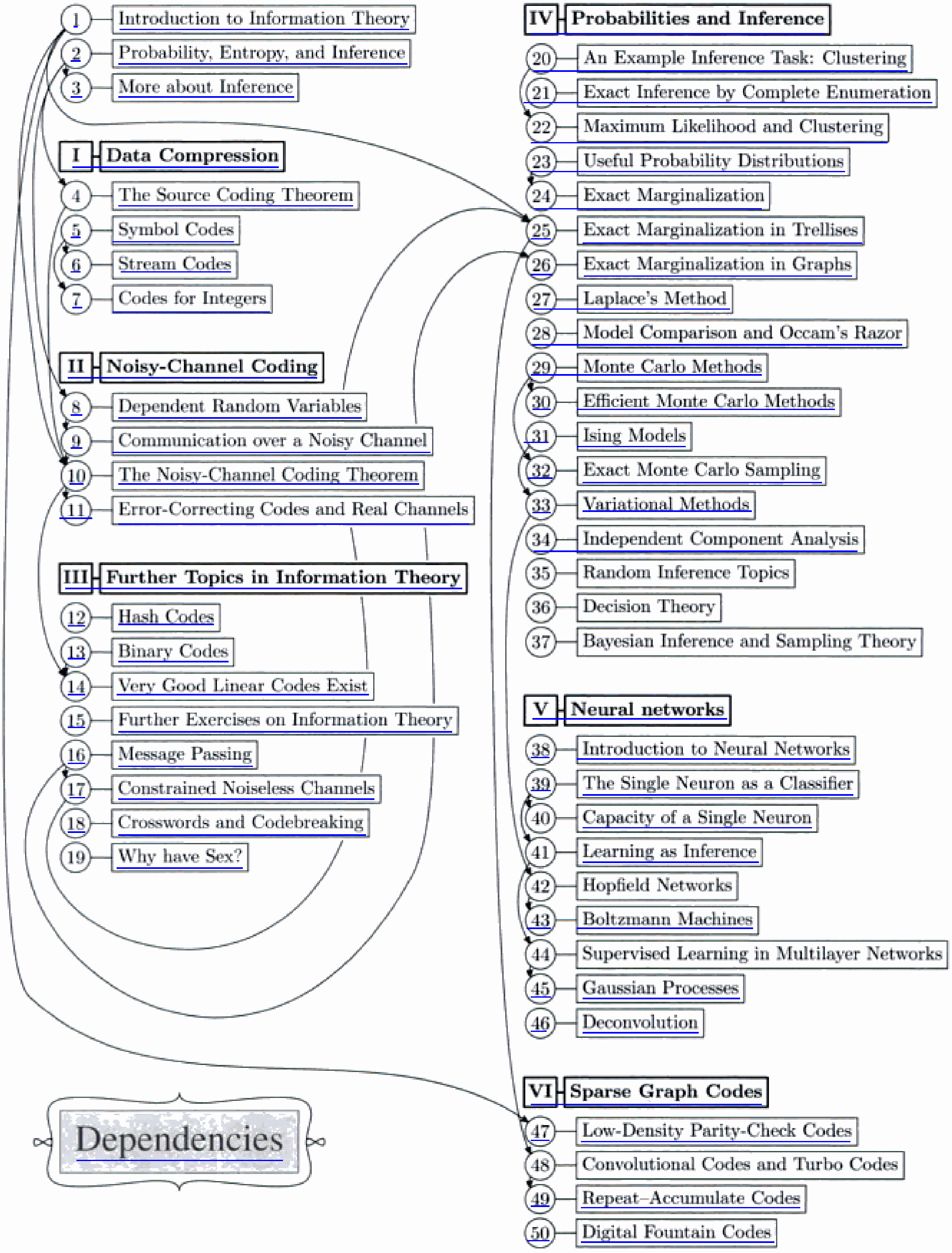
How to use this book

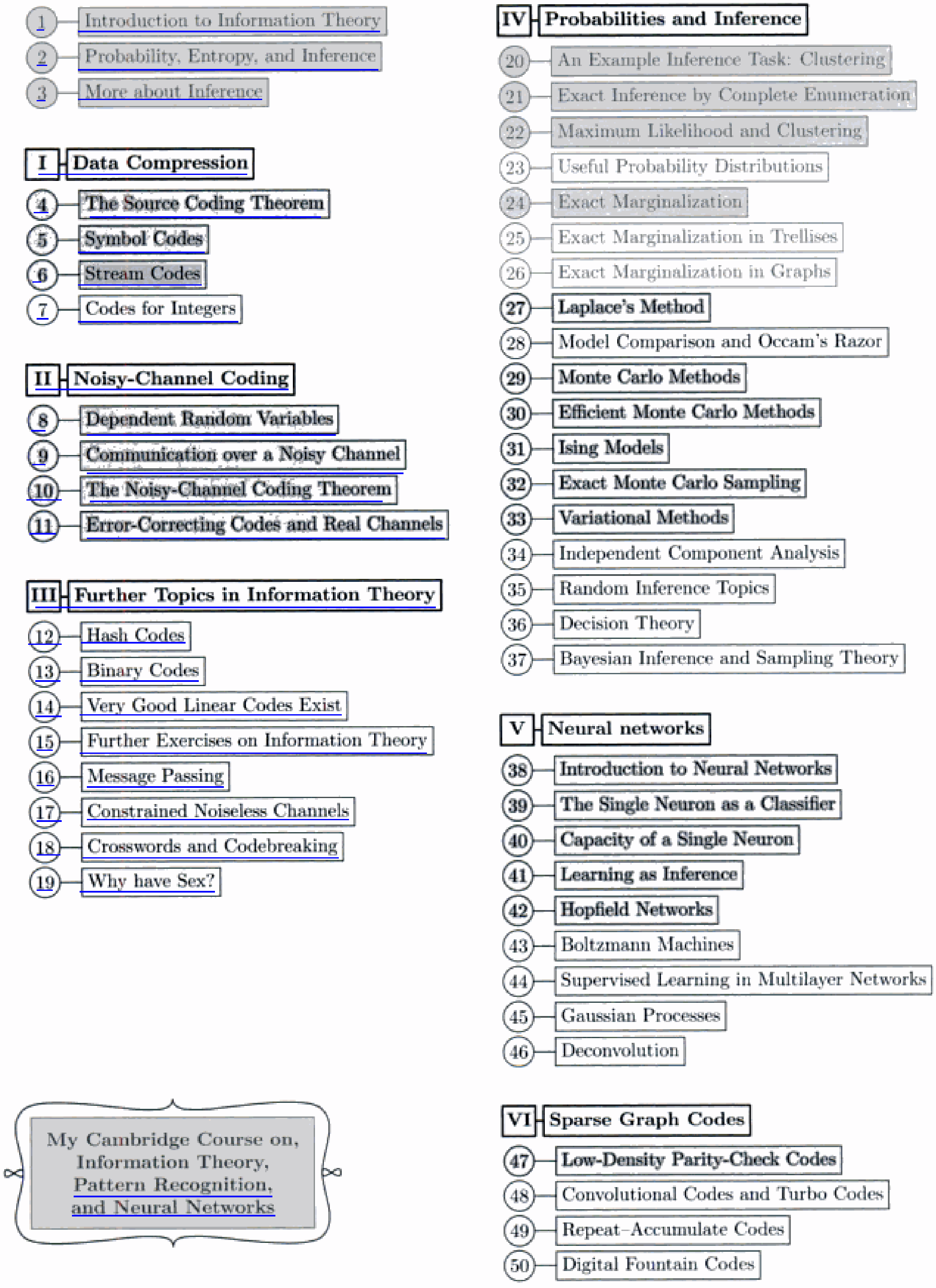
The essential dependencies between chapters are indicated in the figure on the next page. An arrow from one chapter to another indicates that the second chapter requires some of the first.

Within Parts I, II, IV, and V of this book, chapters on advanced or optional topics are towards the end. All chapters of Part III are optional on a first reading, except perhaps for Chapter 16 (Message Passing).

The same system sometimes applies within a chapter: the final sections often deal with advanced topics that can be skipped on a first reading. For example in two key chapters – Chapter 4 (The Source Coding Theorem) and Chapter 10 (The Noisy-Channel Coding Theorem) – the first-time reader should detour at section 4.5 and section 10.4 respectively.

Pages vii–x show a few ways to use this book. First, I give the roadmap for a course that I teach in Cambridge: 'Information theory, pattern recognition, and neural networks'. The book is also intended as a textbook for traditional courses in information theory. The second roadmap shows the chapters for an introductory information theory course and the third for a course aimed at an understanding of state-of-the-art error-correcting codes. The fourth roadmap shows how to use the text in a conventional course on machine learning.


1 Introduction to Information Theory
2 Probability, Entropy, and Inference
3 More about Inference
I Data Compression
4 The Source Coding Theorem
5 Symbol Codes
6 Stream Codes
7 Codes for Integers
II Noisy-Channel Coding
8 Dependent Random Variables
9 Communication over a Noisy Channel
10 The Noisy-Channel Coding Theorem
11 Error-Correcting Codes and Real Channels
III Further Topics in Information Theory
12 Hash Codes
13 Binary Codes
14 Very Good Linear Codes Exist
15 Further Exercises on Information Theory
16 Message Passing
17 Constrained Noiseless Channels
18 Crosswords and Codebreaking
19 Why have Sex?
Dependencies
IV Probabilities and Inference
20 An Example Inference Task: Clustering
21 Exact Inference by Complete Enumeration
22 Maximum Likelihood and Clustering
23 Useful Probability Distributions
24 Exact Marginalization
25 Exact Marginalization in Trellises
26 Exact Marginalization in Graphs
27 Laplace's Method
28 Model Comparison and Occam's Razor
29 Monte Carlo Methods
30 Efficient Monte Carlo Methods
31 Ising Models
32 Exact Monte Carlo Sampling
33 Variational Methods
34 Independent Component Analysis
35 Random Inference Topics
36 Decision Theory
37 Bayesian Inference and Sampling Theory
V Neural networks
38 Introduction to Neural Networks
39 The Single Neuron as a Classifier
40 Capacity of a Single Neuron
41 Learning as Inference
42 Hopfield Networks
43 Boltzmann Machines
44 Supervised Learning in Multilayer Networks
45 Gaussian Processes
46 Deconvolution
VI Sparse Graph Codes
47 Low-Density Parity-Check Codes
48 Convolutional Codes and Turbo Codes
49 Repeat–Accumulate Codes
50 Digital Fountain Codes


1 Introduction to Information Theory
2 Probability, Entropy, and Inference
3 More about Inference
I Data Compression
4 The Source Coding Theorem
5 Symbol Codes
6 Stream Codes
7 Codes for Integers
II Noisy-Channel Coding
8 Dependent Random Variables
9 Communication over a Noisy Channel
10 The Noisy-Channel Coding Theorem
11 Error-Correcting Codes and Real Channels
III Further Topics in Information Theory
12 Hash Codes
13 Binary Codes
14 Very Good Linear Codes Exist
15 Further Exercises on Information Theory
16 Message Passing
17 Constrained Noiseless Channels
18 Crosswords and Codebreaking
19 Why have Sex?
My Cambridge Course on, Information Theory, Pattern Recognition, and Neural Networks
IV Probabilities and Inference
20 An Example Inference Task: Clustering
21 Exact Inference by Complete Enumeration
22 Maximum Likelihood and Clustering
23 Useful Probability Distributions
24 Exact Marginalization
25 Exact Marginalization in Trellises
26 Exact Marginalization in Graphs
27 Laplace's Method
28 Model Comparison and Occam's Razor
29 Monte Carlo Methods
30 Efficient Monte Carlo Methods
31 Ising Models
32 Exact Monte Carlo Sampling
33 Variational Methods
34 Independent Component Analysis
35 Random Inference Topics
36 Decision Theory
37 Bayesian Inference and Sampling Theory
V Neural networks
38 Introduction to Neural Networks
39 The Single Neuron as a Classifier
40 Capacity of a Single Neuron
41 Learning as Inference
42 Hopfield Networks
43 Boltzmann Machines
44 Supervised Learning in Multilayer Networks
45 Gaussian Processes
46 Deconvolution
VI Sparse Graph Codes
47 Low-Density Parity-Check Codes
48 Convolutional Codes and Turbo Codes
49 Repeat–Accumulate Codes
50 Digital Fountain Codes

Short Course on Information Theory

About the exercises

You can understand a subject only by creating it for yourself. The exercises play an essential role in this book. For guidance, each has a rating (similar to that used by Knuth (1968)) from 1 to 5 to indicate its difficulty.

In addition, exercises that are especially recommended are marked by a marginal encouraging rat. Some exercises that require the use of a computer are marked with a *C*.

Answers to many exercises are provided. Use them wisely. Where a solution is provided, this is indicated by including its page number alongside the difficulty rating.

Solutions to many of the other exercises will be supplied to instructors using this book in their teaching; please email `solutions@cambridge.org`.

Summary of codes for exercises

🐀	Especially recommended	[*1*]	Simple (one minute)
		[*2*]	Medium (quarter hour)
▷	Recommended	[*3*]	Moderately hard
C	Parts require a computer	[*4*]	Hard
[p. 42]	Solution provided on page 42	[*5*]	Research project

Internet resources

The website

`http://www.inference.phy.cam.ac.uk/mackay/itila`

contains several resources:

1. *Software.* Teaching software that I use in lectures, interactive software, and research software, written in `perl`, `octave`, `tcl`, `C`, and `gnuplot`. Also some animations.

2. *Corrections to the book.* Thank you in advance for emailing these!

3. *This book.* The book is provided in `postscript`, `pdf`, and `djvu` formats for on-screen viewing. The same copyright restrictions apply as to a normal book.

About this edition

This is the third printing of the first edition. In the second printing, the design of the book was altered slightly. Page-numbering generally remains unchanged, except in chapters 1, 6, and 28, where a few paragraphs, figures, and equations have moved around. All equation, section, and exercise numbers are unchanged. In the third printing, chapter 8 has been renamed 'Dependent Random Variables', instead of 'Correlated', which was sloppy.

Acknowledgments

I am most grateful to the organizations who have supported me while this book gestated: the Royal Society and Darwin College who gave me a fantastic research fellowship in the early years; the University of Cambridge; the Keck Centre at the University of California in San Francisco, where I spent a productive sabbatical; and the Gatsby Charitable Foundation, whose support gave me the freedom to break out of the Escher staircase that book-writing had become.

My work has depended on the generosity of free software authors. I wrote the book in LaTeX 2_{ε}. Three cheers for Donald Knuth and Leslie Lamport! Our computers run the GNU/Linux operating system. I use `emacs`, `perl`, and `gnuplot` every day. Thank you Richard Stallman, thank you Linus Torvalds, thank you everyone.

Many readers, too numerous to name here, have given feedback on the book, and to them all I extend my sincere acknowledgments. I especially wish to thank all the students and colleagues at Cambridge University who have attended my lectures on information theory and machine learning over the last nine years.

The members of the Inference research group have given immense support, and I thank them all for their generosity and patience over the last ten years: Mark Gibbs, Michelle Povinelli, Simon Wilson, Coryn Bailer-Jones, Matthew Davey, Katriona Macphee, James Miskin, David Ward, Edward Ratzer, Seb Wills, John Barry, John Winn, Phil Cowans, Hanna Wallach, Matthew Garrett, and especially Sanjoy Mahajan. Thank you too to Graeme Mitchison, Mike Cates, and Davin Yap.

Finally I would like to express my debt to my personal heroes, the mentors from whom I have learned so much: Yaser Abu-Mostafa, Andrew Blake, John Bridle, Peter Cheeseman, Steve Gull, Geoff Hinton, John Hopfield, Steve Luttrell, Robert MacKay, Bob McEliece, Radford Neal, Roger Sewell, and John Skilling.

Dedication

This book is dedicated to the campaign against the arms trade.

`www.caat.org.uk`

Peace cannot be kept by force.
It can only be achieved through understanding.
– *Albert Einstein*

About Chapter 1

In the first chapter, you will need to be familiar with the binomial distribution. And to solve the exercises in the text – which I urge you to do – you will need to know *Stirling's approximation* for the factorial function, $x! \simeq x^x\, e^{-x}$, and be able to apply it to $\binom{N}{r} = \frac{N!}{(N-r)!\,r!}$. These topics are reviewed below.

Unfamiliar notation? See Appendix A, p.598.

The binomial distribution

Example 1.1. A bent coin has probability f of coming up heads. The coin is tossed N times. What is the probability distribution of the number of heads, r? What are the mean and variance of r?

Solution. The number of heads has a binomial distribution.

$$P(r \mid f, N) = \binom{N}{r} f^r (1-f)^{N-r}. \tag{1.1}$$

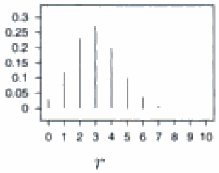


Figure 1.1. The binomial distribution $P(r \mid f = 0.3, N = 10)$.

The mean, $\mathcal{E}[r]$, and variance, $\mathrm{var}[r]$, of this distribution are defined by

$$\mathcal{E}[r] \equiv \sum_{r=0}^{N} P(r \mid f, N)\, r \tag{1.2}$$

$$\begin{aligned} \mathrm{var}[r] &\equiv \mathcal{E}\left[(r - \mathcal{E}[r])^2\right] & (1.3)\\ &= \mathcal{E}[r^2] - (\mathcal{E}[r])^2 = \sum_{r=0}^{N} P(r \mid f, N) r^2 - (\mathcal{E}[r])^2 . & (1.4) \end{aligned}$$

Rather than evaluating the sums over r in (1.2) and (1.4) directly, it is easiest to obtain the mean and variance by noting that r is the sum of N *independent* random variables, namely, the number of heads in the first toss (which is either zero or one), the number of heads in the second toss, and so forth. In general,

$$\begin{array}{rcll} \mathcal{E}[x+y] & = & \mathcal{E}[x] + \mathcal{E}[y] & \text{for any random variables } x \text{ and } y; \\ \mathrm{var}[x+y] & = & \mathrm{var}[x] + \mathrm{var}[y] & \text{if } x \text{ and } y \text{ are independent.} \end{array} \tag{1.5}$$

So the mean of r is the sum of the means of those random variables, and the variance of r is the sum of their variances. The mean number of heads in a single toss is $f \times 1 + (1-f) \times 0 = f$, and the variance of the number of heads in a single toss is

$$\left[f \times 1^2 + (1-f) \times 0^2\right] - f^2 = f - f^2 = f(1-f), \tag{1.6}$$

so the mean and variance of r are:

$$\mathcal{E}[r] = Nf \qquad \text{and} \qquad \mathrm{var}[r] = Nf(1-f). \qquad \Box \tag{1.7}$$

Approximating $x!$ and $\binom{N}{r}$

Let's derive Stirling's approximation by an unconventional route. We start from the Poisson distribution with mean λ,

$$P(r \,|\, \lambda) = e^{-\lambda} \frac{\lambda^r}{r!} \qquad r \in \{0, 1, 2, \ldots\}. \tag{1.8}$$

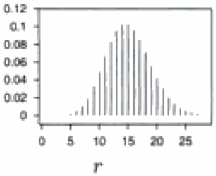


Figure 1.2. The Poisson distribution $P(r \,|\, \lambda = 15)$.

For large λ, this distribution is well approximated – at least in the vicinity of $r \simeq \lambda$ – by a Gaussian distribution with mean λ and variance λ:

$$e^{-\lambda} \frac{\lambda^r}{r!} \simeq \frac{1}{\sqrt{2\pi\lambda}} \, e^{-\frac{(r-\lambda)^2}{2\lambda}}. \tag{1.9}$$

Let's plug $r = \lambda$ into this formula.

$$\begin{aligned} e^{-\lambda} \frac{\lambda^\lambda}{\lambda!} &\simeq \frac{1}{\sqrt{2\pi\lambda}} & (1.10)\\ \Rightarrow \lambda! &\simeq \lambda^\lambda \, e^{-\lambda} \sqrt{2\pi\lambda}. & (1.11) \end{aligned}$$

This is Stirling's approximation for the factorial function.

$$x! \simeq x^x \, e^{-x} \sqrt{2\pi x} \quad \Leftrightarrow \quad \ln x! \simeq x \ln x - x + \tfrac{1}{2} \ln 2\pi x. \tag{1.12}$$

We have derived not only the leading order behaviour, $x! \simeq x^x \, e^{-x}$, but also, at no cost, the next-order correction term $\sqrt{2\pi x}$. We now apply Stirling's approximation to $\ln \binom{N}{r}$:

$$\ln \binom{N}{r} \equiv \ln \frac{N!}{(N-r)! \, r!} \simeq (N-r) \ln \frac{N}{N-r} + r \ln \frac{N}{r}. \tag{1.13}$$

Since all the terms in this equation are logarithms, this result can be rewritten in any base. We will denote natural logarithms ($\log_e$) by 'ln', and logarithms to base 2 ($\log_2$) by 'log'.

Recall that $\log_2 x = \dfrac{\log_e x}{\log_e 2}$.

Note that $\dfrac{\partial \log_2 x}{\partial x} = \dfrac{1}{\log_e 2} \dfrac{1}{x}$.

If we introduce the *binary entropy function*,

$$H_2(x) \equiv x \log \frac{1}{x} + (1-x) \log \frac{1}{(1-x)}, \tag{1.14}$$

then we can rewrite the approximation (1.13) as

$$\log \binom{N}{r} \simeq N H_2(r/N), \tag{1.15}$$

or, equivalently,

$$\binom{N}{r} \simeq 2^{N H_2(r/N)}. \tag{1.16}$$

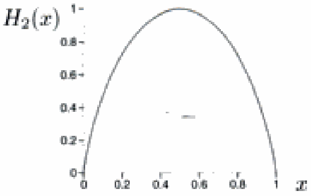


Figure 1.3. The binary entropy function.

If we need a more accurate approximation, we can include terms of the next order from Stirling's approximation (1.12):

$$\log \binom{N}{r} \simeq N H_2(r/N) - \tfrac{1}{2} \log \left[2\pi N \, \frac{N-r}{N} \, \frac{r}{N} \right]. \tag{1.17}$$

1

Introduction to Information Theory

> The fundamental problem of is that of reproducing at one point either exactly or approximately a message selected at another point.
>
> *(Claude Shannon, 1948)*

In the first half of this book we study how to measure information content; we learn how to compress data; and we learn how to communicate perfectly over imperfect communication channels.

We start by getting a feeling for this last problem.

▶ 1.1 How can we achieve perfect communication over an imperfect, noisy communication channel?

Some examples of noisy communication channels are:

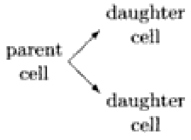
- an analogue telephone line, over which two modems communicate digital information;

modem
phone line
modem

- the radio communication link from Galileo, the Jupiter-orbiting spacecraft, to earth;

Galileo
radio waves
Earth

- reproducing cells, in which the daughter cells' DNA contains information from the parent cells;



- a disk drive.

computer memory
disk drive
computer memory

The last example shows that communication doesn't have to involve information going from one *place* to another. When we write a file on a disk drive, we'll read it off in the same location – but at a later *time*.

These channels are noisy. A telephone line suffers from cross-talk with other lines; the hardware in the line distorts and adds noise to the transmitted signal. The deep space network that listens to Galileo's puny transmitter receives background radiation from terrestrial and cosmic sources. DNA is subject to mutations and damage. A disk drive, which writes a binary digit (a one or zero, also known as a *bit*) by aligning a patch of magnetic material in one of two orientations, may later fail to read out the stored binary digit: the patch of material might spontaneously flip magnetization, or a glitch of background noise might cause the reading circuit to report the wrong value for the binary digit, or the writing head might not induce the magnetization in the first place because of interference from neighbouring bits.

In all these cases, if we transmit data, e.g., a string of bits, over the channel, there is some probability that the received message will not be identical to the

Received sequence $\mathbf{r}$	Likelihood ratio $\frac{P(\mathbf{r}\,|\,s=1)}{P(\mathbf{r}\,|\,s=0)}$	Decoded sequence $\hat{\mathbf{s}}$
000	γ^{-3}	0
001	γ^{-1}	0
010	γ^{-1}	0
100	γ^{-1}	0
101	γ^{1}	1
110	γ^{1}	1
011	γ^{1}	1
111	γ^{3}	1

Algorithm 1.9. Majority-vote decoding algorithm for R_3. Also shown are the likelihood ratios (1.23), assuming the channel is a binary symmetric channel; $\gamma \equiv (1-f)/f$.

At the risk of explaining the obvious, let's prove this result. The optimal decoding decision (optimal in the sense of having the smallest probability of being wrong) is to find which value of $\mathbf{s}$ is most probable, given $\mathbf{r}$. Consider the decoding of a single bit s, which was encoded as $\mathbf{t}(s)$ and gave rise to three received bits $\mathbf{r} = r_1r_2r_3$. By Bayes' theorem, the *posterior probability* of s is

$$P(s\,|\,r_1r_2r_3) = \frac{P(r_1r_2r_3\,|\,s)P(s)}{P(r_1r_2r_3)}. \tag{1.18}$$

We can spell out the posterior probability of the two alternatives thus:

$$P(s=1\,|\,r_1r_2r_3) = \frac{P(r_1r_2r_3\,|\,s=1)P(s=1)}{P(r_1r_2r_3)}; \tag{1.19}$$

$$P(s=0\,|\,r_1r_2r_3) = \frac{P(r_1r_2r_3\,|\,s=0)P(s=0)}{P(r_1r_2r_3)}. \tag{1.20}$$

This posterior probability is determined by two factors: the *prior probability* $P(s)$, and the data-dependent term $P(r_1r_2r_3\,|\,s)$, which is called the *likelihood* of s. The normalizing constant $P(r_1r_2r_3)$ needn't be computed when finding the optimal decoding decision, which is to guess $\hat{s}=0$ if $P(s=0\,|\,\mathbf{r}) > P(s=1\,|\,\mathbf{r})$, and $\hat{s}=1$ otherwise.

To find $P(s=0\,|\,\mathbf{r})$ and $P(s=1\,|\,\mathbf{r})$, we must make an assumption about the prior probabilities of the two hypotheses $s=0$ and $s=1$, and we must make an assumption about the probability of $\mathbf{r}$ given s. We assume that the prior probabilities are equal: $P(s=0) = P(s=1) = 0.5$; then maximizing the posterior probability $P(s\,|\,\mathbf{r})$ is equivalent to maximizing the likelihood $P(\mathbf{r}\,|\,s)$. And we assume that the channel is a binary symmetric channel with noise level $f < 0.5$, so that the likelihood is

$$P(\mathbf{r}\,|\,s) = P(\mathbf{r}\,|\,\mathbf{t}(s)) = \prod_{n=1}^{N} P(r_n\,|\,t_n(s)), \tag{1.21}$$

where $N = 3$ is the number of transmitted bits in the block we are considering, and

$$P(r_n\,|\,t_n) = \begin{cases} (1-f) & \text{if} \quad r_n = t_n \\ f & \text{if} \quad r_n \neq t_n. \end{cases} \tag{1.22}$$

Thus the likelihood ratio for the two hypotheses is

$$\frac{P(\mathbf{r}\,|\,s=1)}{P(\mathbf{r}\,|\,s=0)} = \prod_{n=1}^{N} \frac{P(r_n\,|\,t_n(1))}{P(r_n\,|\,t_n(0))}; \tag{1.23}$$

each factor $\frac{P(r_n|t_n(1))}{P(r_n|t_n(0))}$ equals $\frac{(1-f)}{f}$ if $r_n = 1$ and $\frac{f}{(1-f)}$ if $r_n = 0$. The ratio $\gamma \equiv \frac{(1-f)}{f}$ is greater than 1, since $f < 0.5$, so the winning hypothesis is the one with the most 'votes', each vote counting for a factor of γ in the likelihood ratio.

Thus the majority-vote decoder shown in algorithm 1.9 is the optimal decoder if we assume that the channel is a binary symmetric channel and that the two possible source messages `0` and `1` have equal prior probability.

We now apply the majority vote decoder to the received vector of figure 1.8. The first three received bits are all `0`, so we decode this triplet as a `0`. In the second triplet of figure 1.8, there are two `0`s and one `1`, so we decode this triplet as a `0` – which in this case corrects the error. Not all errors are corrected, however. If we are unlucky and two errors fall in a single block, as in the fifth triplet of figure 1.8, then the decoding rule gets the wrong answer, as shown in figure 1.10.

$$\begin{array}{rccccccc}
\mathbf{s} & \mathtt{0} & \mathtt{0} & \mathtt{1} & \mathtt{0} & \mathtt{1} & \mathtt{1} & \mathtt{0}\\
\mathbf{t} & \overbrace{\mathtt{000}} & \overbrace{\mathtt{000}} & \overbrace{\mathtt{111}} & \overbrace{\mathtt{000}} & \overbrace{\mathtt{111}} & \overbrace{\mathtt{111}} & \overbrace{\mathtt{000}}\\
\mathbf{n} & \mathtt{000} & \mathtt{001} & \mathtt{000} & \mathtt{000} & \mathtt{101} & \mathtt{000} & \mathtt{000}\\
\hline
\mathbf{r} & \underbrace{\mathtt{000}} & \underbrace{\mathtt{001}} & \underbrace{\mathtt{111}} & \underbrace{\mathtt{000}} & \underbrace{\mathtt{010}} & \underbrace{\mathtt{111}} & \underbrace{\mathtt{000}}\\
\hat{\mathbf{s}} & \mathtt{0} & \mathtt{0} & \mathtt{1} & \mathtt{0} & \mathtt{0} & \mathtt{1} & \mathtt{0}\\
\text{corrected errors} & & \star & & & & & \\
\text{undetected errors} & & & & & \star & &
\end{array}$$

Figure 1.10. Decoding the received vector from figure 1.8.

Exercise 1.2.[2, p.16] Show that the error probability is reduced by the use of R_3 by computing the error probability of this code for a binary symmetric channel with noise level f.

The exercise's rating, e.g.'[2]', indicates its difficulty: '1' exercises are the easiest. Exercises that are accompanied by a marginal rat are especially recommended. If a solution or partial solution is provided, the page is indicated after the difficulty rating; for example, this exercise's solution is on page 16.

The error probability is dominated by the probability that two bits in a block of three are flipped, which scales as f^2. In the case of the binary symmetric channel with $f = 0.1$, the R_3 code has a probability of error, after decoding, of $p_b \simeq 0.03$ per bit. Figure 1.11 shows the result of transmitting a binary image over a binary symmetric channel using the repetition code.

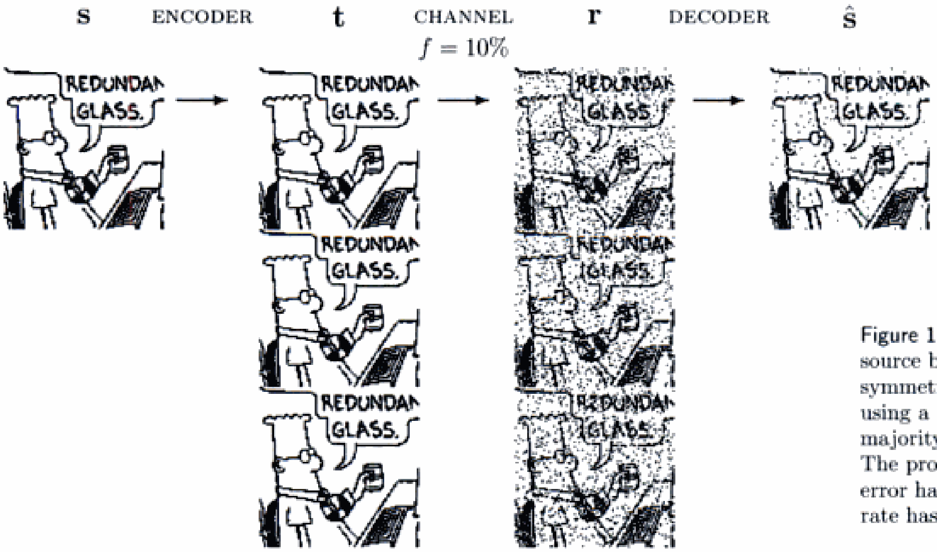


Figure 1.11. Transmitting 10 000 source bits over a binary symmetric channel with $f = 10\%$ using a repetition code and the majority vote decoding algorithm. The probability of decoded bit error has fallen to about 3%; the rate has fallen to 1/3.

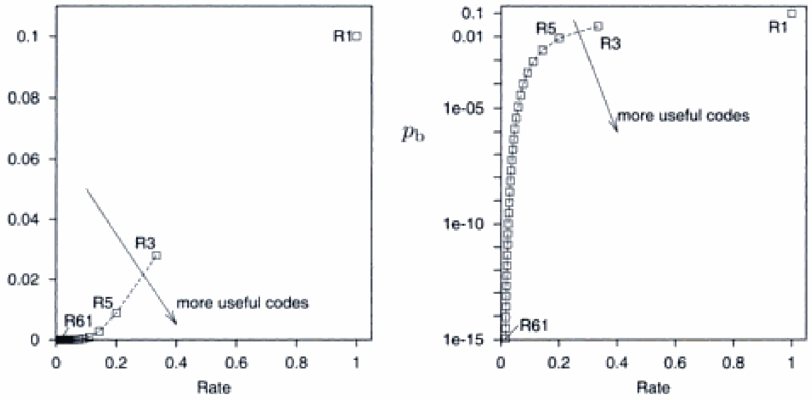


Figure 1.12. Error probability p_b versus rate for repetition codes over a binary symmetric channel with $f = 0.1$. The right-hand figure shows p_b on a logarithmic scale. We would like the rate to be large and p_b to be small.

The repetition code R_3 has therefore reduced the probability of error, as desired. Yet we have lost something: our *rate* of information transfer has fallen by a factor of three. So if we use a repetition code to communicate data over a telephone line, it will reduce the error frequency, but it will also reduce our communication rate. We will have to pay three times as much for each phone call. Similarly, we would need three of the original noisy gigabyte disk drives in order to create a one-gigabyte disk drive with $p_b = 0.03$.

Can we push the error probability lower, to the values required for a sellable disk drive – 10^{-15}? We could achieve lower error probabilities by using repetition codes with more repetitions.

Exercise 1.3.[*3*, p.16] (a) Show that the probability of error of R_N, the repetition code with N repetitions, is

$$p_b = \sum_{n=(N+1)/2}^{N} \binom{N}{n} f^n (1-f)^{N-n}, \tag{1.24}$$

for odd N.

(b) Assuming $f = 0.1$, which of the terms in this sum is the biggest? How much bigger is it than the second-biggest term?

(c) Use Stirling's approximation (p.2) to approximate the $\binom{N}{n}$ in the largest term, and find, approximately, the probability of error of the repetition code with N repetitions.

(d) Assuming $f = 0.1$, find how many repetitions are required to get the probability of error down to 10^{-15}. [Answer: about 60.]

So to build a *single* gigabyte disk drive with the required reliability from noisy gigabyte drives with $f = 0.1$, we would need *sixty* of the noisy disk drives. The tradeoff between error probability and rate for repetition codes is shown in figure 1.12.

Block codes – the (7, 4) Hamming code

We would like to communicate with tiny probability of error *and* at a substantial rate. Can we improve on repetition codes? What if we add redundancy to *blocks* of data instead of encoding one bit at a time? We now study a simple *block code*.

A *block code* is a rule for converting a sequence of source bits $\mathbf{s}$, of length K, say, into a transmitted sequence $\mathbf{t}$ of length N bits. To add redundancy, we make N greater than K. In a *linear* block code, the extra $N - K$ bits are linear functions of the original K bits; these extra bits are called *parity-check bits*. An example of a linear block code is the $(7, 4)$ *Hamming code*, which transmits $N = 7$ bits for every $K = 4$ source bits.

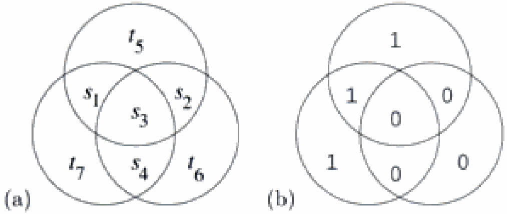


Figure 1.13. Pictorial representation of encoding for the $(7, 4)$ Hamming code.

The encoding operation for the code is shown pictorially in figure 1.13. We arrange the seven transmitted bits in three intersecting circles. The first four transmitted bits, $t_1t_2t_3t_4$, are set equal to the four source bits, $s_1s_2s_3s_4$. The parity-check bits $t_5t_6t_7$ are set so that the *parity* within each circle is even: the first parity-check bit is the parity of the first three source bits (that is, it is 0 if the sum of those bits is even, and 1 if the sum is odd); the second is the parity of the last three; and the third parity bit is the parity of source bits one, three and four.

As an example, figure 1.13b shows the transmitted codeword for the case $\mathbf{s} = 1000$. Table 1.14 shows the codewords generated by each of the $2^4 =$ sixteen settings of the four source bits. These codewords have the special property that any pair differ from each other in at least three bits.

s	t	s	t	s	t	s	t
0000	0000000	0100	0100110	1000	1000101	1100	1100011
0001	0001011	0101	0101101	1001	1001110	1101	1101000
0010	0010111	0110	0110001	1010	1010010	1110	1110100
0011	0011100	0111	0111010	1011	1011001	1111	1111111

Table 1.14. The sixteen codewords $\{\mathbf{t}\}$ of the $(7, 4)$ Hamming code. Any pair of codewords differ from each other in at least three bits.

Because the Hamming code is a linear code, it can be written compactly in terms of matrices as follows. The transmitted codeword $\mathbf{t}$ is obtained from the source sequence $\mathbf{s}$ by a linear operation,

$$\mathbf{t} = \mathbf{G}^{\mathsf{T}}\mathbf{s}, \tag{1.25}$$

where $\mathbf{G}$ is the *generator matrix* of the code,

$$\mathbf{G}^{\mathsf{T}} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \end{bmatrix}, \tag{1.26}$$

and the encoding operation (1.25) uses modulo-2 arithmetic ($1 + 1 = 0$, $0 + 1 = 1$, etc.).

In the encoding operation (1.25) I have assumed that $\mathbf{s}$ and $\mathbf{t}$ are column vectors. If instead they are row vectors, then this equation is replaced by

$$\mathbf{t} = \mathbf{s}\mathbf{G}, \tag{1.27}$$

where

$$\mathbf{G} = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 \end{bmatrix}. \tag{1.28}$$

I find it easier to relate to the right-multiplication (1.25) than the left-multiplication (1.27). Many coding theory texts use the left-multiplying conventions (1.27–1.28), however.

The rows of the generator matrix (1.28) can be viewed as defining four basis vectors lying in a seven-dimensional binary space. The sixteen codewords are obtained by making all possible linear combinations of these vectors.

Decoding the (7, 4) *Hamming code*

When we invent a more complex encoder $\mathbf{s} \rightarrow \mathbf{t}$, the task of decoding the received vector $\mathbf{r}$ becomes less straightforward. Remember that *any* of the bits may have been flipped, including the parity bits.

If we assume that the channel is a binary symmetric channel and that all source vectors are equiprobable, then the optimal decoder identifies the source vector $\mathbf{s}$ whose encoding $\mathbf{t}(\mathbf{s})$ differs from the received vector $\mathbf{r}$ in the fewest bits. [Refer to the likelihood function (1.23) to see why this is so.] We could solve the decoding problem by measuring how far $\mathbf{r}$ is from each of the sixteen codewords in table 1.14, then picking the closest. Is there a more efficient way of finding the most probable source vector?

Syndrome decoding for the Hamming code

For the (7, 4) Hamming code there is a pictorial solution to the decoding problem, based on the encoding picture, figure 1.13.

As a first example, let's assume the transmission was $\mathbf{t} = 1000101$ and the noise flips the second bit, so the received vector is $\mathbf{r} = 1000101 \oplus 0100000 = 1100101$. We write the received vector into the three circles as shown in figure 1.15a, and look at each of the three circles to see whether its parity is even. The circles whose parity is *not* even are shown by dashed lines in figure 1.15b. The decoding task is to find the smallest set of flipped bits that can account for these violations of the parity rules. [The pattern of violations of the parity checks is called the *syndrome*, and can be written as a binary vector – for example, in figure 1.15b, the syndrome is $\mathbf{z} = (1, 1, 0)$, because the first two circles are 'unhappy' (parity 1) and the third circle is 'happy' (parity 0).]

To solve the decoding task, we ask the question: can we find a unique bit that lies *inside* all the 'unhappy' circles and *outside* all the 'happy' circles? If so, the flipping of that bit would account for the observed syndrome. In the case shown in figure 1.15b, the bit r_2 lies inside the two unhappy circles and outside the happy circle; no other single bit has this property, so r_2 is the only single bit capable of explaining the syndrome.

Let's work through a couple more examples. Figure 1.15c shows what happens if one of the parity bits, t_5, is flipped by the noise. Just one of the checks is violated. Only r_5 lies inside this unhappy circle and outside the other two happy circles, so r_5 is identified as the only single bit capable of explaining the syndrome.

If the central bit r_3 is received flipped, figure 1.15d shows that all three checks are violated; only r_3 lies inside all three circles, so r_3 is identified as the suspect bit.

There is a *decoding error* if the four decoded bits $\hat{s}_1, \hat{s}_2, \hat{s}_3, \hat{s}_4$ do not all match the source bits s_1, s_2, s_3, s_4. The *probability of block error* $p_{\rm B}$ is the probability that one or more of the decoded bits in one block fail to match the corresponding source bits,

$$p_{\rm B} = P(\hat{\mathbf{s}} \neq \mathbf{s}). \tag{1.33}$$

The *probability of bit error* $p_{\rm b}$ is the average probability that a decoded bit fails to match the corresponding source bit,

$$p_{\rm b} = \frac{1}{K} \sum_{k=1}^{K} P(\hat{s}_k \neq s_k). \tag{1.34}$$

In the case of the Hamming code, a decoding error will occur whenever the noise has flipped more than one bit in a block of seven. The probability of block error is thus the probability that two or more bits are flipped in a block. This probability scales as $O(f^2)$, as did the probability of error for the repetition code R_3. But notice that the Hamming code communicates at a greater rate, $R = 4/7$.

Figure 1.17 shows a binary image transmitted over a binary symmetric channel using the $(7, 4)$ Hamming code. About 7% of the decoded bits are in error. Notice that the errors are correlated: often two or three successive decoded bits are flipped.

Exercise 1.5.[1] This exercise and the next three refer to the $(7, 4)$ Hamming code. Decode the received strings:

(a) $\mathbf{r} = 1101011$

(b) $\mathbf{r} = 0110110$

(c) $\mathbf{r} = 0100111$

(d) $\mathbf{r} = 1111111$.

Exercise 1.6.[2, p.17] (a) Calculate the probability of block error $p_{\rm B}$ of the $(7, 4)$ Hamming code as a function of the noise level f and show that to leading order it goes as $21f^2$.

(b) [3] Show that to leading order the probability of bit error $p_{\rm b}$ goes as $9f^2$.

Exercise 1.7.[2, p.19] Find some noise vectors that give the all-zero syndrome (that is, noise vectors that leave all the parity checks unviolated). How many such noise vectors are there?

▷ Exercise 1.8.[2] I asserted above that a block decoding error will result whenever two or more bits are flipped in a single block. Show that this is indeed so. [In principle, there might be error patterns that, after decoding, led only to the corruption of the parity bits, with no source bits incorrectly decoded.]

Summary of codes' performances

Figure 1.18 shows the performance of repetition codes and the Hamming code. It also shows the performance of a family of linear block codes that are generalizations of Hamming codes, called BCH codes.

This figure shows that we can, using linear block codes, achieve better performance than repetition codes; but the asymptotic situation still looks grim.

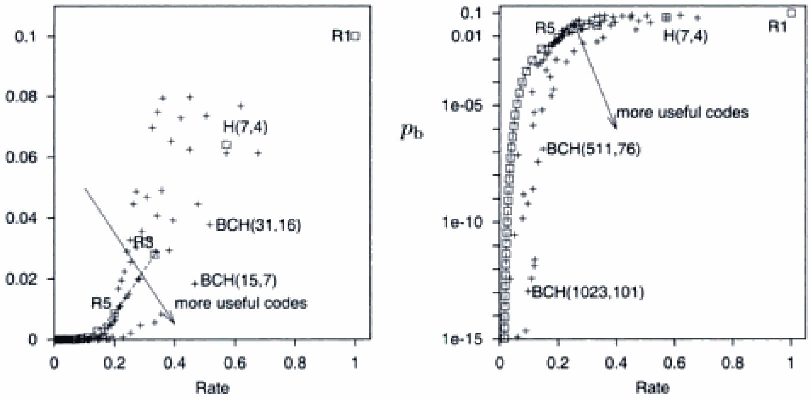


Figure 1.18. Error probability p_b versus rate R for repetition codes, the $(7,4)$ Hamming code and BCH codes with blocklengths up to 1023 over a binary symmetric channel with $f = 0.1$. The righthand figure shows p_b on a logarithmic scale.

Exercise 1.9.$^{[4, \text{ p.19}]}$ Design an error-correcting code and a decoding algorithm for it, estimate its probability of error, and add it to figure 1.18. [Don't worry if you find it difficult to make a code better than the Hamming code, or if you find it difficult to find a good decoder for your code; that's the point of this exercise.]

Exercise 1.10.$^{[3, \text{ p.20}]}$ A $(7,4)$ Hamming code can correct any *one* error; might there be a $(14,8)$ code that can correct any two errors?

Optional extra: Does the answer to this question depend on whether the code is linear or nonlinear?

Exercise 1.11.$^{[4, \text{ p.21}]}$ Design an error-correcting code, other than a repetition code, that can correct any *two* errors in a block of size N.

▶ 1.3 What performance can the best codes achieve?

There seems to be a trade-off between the decoded bit-error probability p_b (which we would like to reduce) and the rate R (which we would like to keep large). How can this trade-off be characterized? What points in the (R, p_b) plane are achievable? This question was addressed by Claude Shannon in his pioneering paper of 1948, in which he both created the field of information theory and solved most of its fundamental problems.

At that time there was a widespread belief that the boundary between achievable and nonachievable points in the (R, p_b) plane was a curve passing through the origin $(R, p_b) = (0, 0)$; if this were so, then, in order to achieve a vanishingly small error probability p_b, one would have to reduce the rate correspondingly close to zero. 'No pain, no gain.'

However, Shannon proved the remarkable result that the boundary between achievable and nonachievable points meets the R axis at a *non-zero* value $R = C$, as shown in figure 1.19. For any channel, there exist codes that make it possible to communicate with *arbitrarily small* probability of error p_b at non-zero rates. The first half of this book (Parts I–III) will be devoted to understanding this remarkable result, which is called the *noisy-channel coding theorem*. *

Example: $f = 0.1$

The maximum rate at which communication is possible with arbitrarily small p_b is called the *capacity* of the channel. The formula for the capacity of a

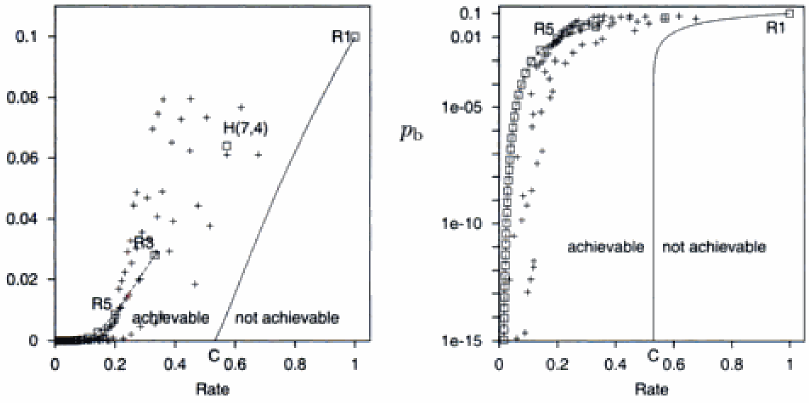


Figure 1.19. Shannon's noisy-channel coding theorem. The solid curve shows the Shannon limit on achievable values of (R, p_b) for the binary symmetric channel with $f = 0.1$. Rates up to $R = C$ are achievable with arbitrarily small p_b. The points show the performance of some textbook codes, as in figure 1.18.

The equation defining the Shannon limit (the solid curve) is $R = C/(1 - H_2(p_b))$, where C and H_2 are defined in equation (1.35).

binary symmetric channel with noise level f is

$$C(f) = 1 - H_2(f) = 1 - \left[f \log_2 \frac{1}{f} + (1-f) \log_2 \frac{1}{1-f} \right]; \qquad (1.35)$$

the channel we were discussing earlier with noise level $f = 0.1$ has capacity $C \simeq 0.53$. Let us consider what this means in terms of noisy disk drives. The repetition code R_3 could communicate over this channel with $p_b = 0.03$ at a rate $R = 1/3$. Thus we know how to build a single gigabyte disk drive with $p_b = 0.03$ from three noisy gigabyte disk drives. We also know how to make a single gigabyte disk drive with $p_b \simeq 10^{-15}$ from sixty noisy one-gigabyte drives (exercise 1.3, p.8). And now Shannon passes by, notices us juggling with disk drives and codes and says:

> 'What performance are you trying to achieve? 10^{-15}? You don't need *sixty* disk drives – you can get that performance with just *two* disk drives (since 1/2 is less than 0.53). And if you want $p_b = 10^{-18}$ or 10^{-24} or anything, you can get there with two disk drives too!'

[Strictly, the above statements might not be quite right, since, as we shall see, Shannon proved his noisy-channel coding theorem by studying sequences of block codes with ever-increasing blocklengths, and the required blocklength might be bigger than a gigabyte (the size of our disk drive), in which case, Shannon might say 'well, you can't do it with those *tiny* disk drives, but if you had two noisy *terabyte* drives, you could make a single high-quality terabyte drive from them'.]

▶ 1.4 Summary

The (7, 4) Hamming Code

By including three parity-check bits in a block of 7 bits it is possible to detect and correct any single bit error in each block.

Shannon's noisy-channel coding theorem

Information can be communicated over a noisy channel at a non-zero rate with arbitrarily small error probability.

Information theory addresses both the *limitations* and the *possibilities* of communication. The noisy-channel coding theorem, which we will prove in Chapter 10, asserts both that reliable communication at any rate beyond the capacity is impossible, and that reliable communication at all rates up to capacity is possible.

The next few chapters lay the foundations for this result by discussing *how to measure information content* and the intimately related topic of *data compression*.

▶ 1.5 Further exercises

▷ Exercise 1.12.[2, p.21] Consider the repetition code R_9. One way of viewing this code is as a *concatenation* of R_3 with R_3. We first encode the source stream with R_3, then encode the resulting output with R_3. We could call this code 'R_3^2'. This idea motivates an alternative decoding algorithm, in which we decode the bits three at a time using the decoder for R_3; then decode the decoded bits from that first decoder using the decoder for R_3.

Evaluate the probability of error for this decoder and compare it with the probability of error for the optimal decoder for R_9.

Do the concatenated encoder and decoder for R_3^2 have advantages over those for R_9?

▶ 1.6 Solutions

Solution to exercise 1.2 (p.7). An error is made by R_3 if two or more bits are flipped in a block of three. So the error probability of R_3 is a sum of two terms: the probability that all three bits are flipped, f^3; and the probability that exactly two bits are flipped, $3f^2(1-f)$. [If these expressions are not obvious, see example 1.1 (p.1): the expressions are $P(r=3 \mid f, N=3)$ and $P(r=2 \mid f, N=3)$.]

$$p_b = p_B = 3f^2(1-f) + f^3 = 3f^2 - 2f^3. \tag{1.36}$$

This probability is dominated for small f by the term $3f^2$.

See exercise 2.38 (p.39) for further discussion of this problem.

Solution to exercise 1.3 (p.8). The probability of error for the repetition code R_N is dominated by the probability that $\lceil N/2 \rceil$ bits are flipped, which goes (for odd N) as

$$\binom{N}{\lceil N/2 \rceil} f^{(N+1)/2}(1-f)^{(N-1)/2}. \tag{1.37}$$

Notation: $\lceil N/2 \rceil$ denotes the smallest integer greater than or equal to $N/2$.

The term $\binom{N}{K}$ can be approximated using the binary entropy function:

$$\frac{1}{N+1} 2^{NH_2(K/N)} \leq \binom{N}{K} \leq 2^{NH_2(K/N)} \Rightarrow \binom{N}{K} \simeq 2^{NH_2(K/N)}, \tag{1.38}$$

where this approximation introduces an error of order $\sqrt{N}$ – as shown in equation (1.17). So

$$p_b = p_B \simeq 2^N (f(1-f))^{N/2} = (4f(1-f))^{N/2}. \tag{1.39}$$

Setting this equal to the required value of 10^{-15} we find $N \simeq 2 \frac{\log 10^{-15}}{\log 4f(1-f)} = 68$. This answer is a little out because the approximation we used overestimated $\binom{N}{K}$ and we did not distinguish between $\lceil N/2 \rceil$ and $N/2$.

A slightly more careful answer (short of explicit computation) goes as follows. Taking the approximation for $\binom{N}{K}$ to the next order, we find:

$$\binom{N}{N/2} \simeq 2^N \frac{1}{\sqrt{2\pi N/4}}. \tag{1.40}$$

This approximation can be proved from an accurate version of Stirling's approximation (1.12), or by considering the binomial distribution with $p = 1/2$ and noting

$$1 = \sum_K \binom{N}{K} 2^{-N} \simeq 2^{-N} \binom{N}{N/2} \sum_{r=-N/2}^{N/2} e^{-r^2/2\sigma^2} \simeq 2^{-N} \binom{N}{N/2} \sqrt{2\pi}\sigma, \tag{1.41}$$

where $\sigma = \sqrt{N/4}$, from which equation (1.40) follows. The distinction between $\lceil N/2 \rceil$ and $N/2$ is not important in this term since $\binom{N}{K}$ has a maximum at $K = N/2$.

Then the probability of error (for odd N) is to leading order

$$\begin{aligned} p_{\rm b} &\simeq \binom{N}{(N+1)/2} f^{(N+1)/2}(1-f)^{(N-1)/2} && (1.42)\\ &\simeq 2^N \frac{1}{\sqrt{\pi N/2}} f[f(1-f)]^{(N-1)/2} \simeq \frac{1}{\sqrt{\pi N/8}} f[4f(1-f)]^{(N-1)/2} && (1.43) \end{aligned}$$

The equation $p_{\rm b} = 10^{-15}$ can be written

$$(N-1)/2 \simeq \frac{\log 10^{-15} + \log \frac{\sqrt{\pi N/8}}{f}}{\log 4f(1-f)} \tag{1.44}$$

which may be solved for N iteratively, the first iteration starting from $\hat{N}_1 = 68$:

$$(\hat{N}_2 - 1)/2 \simeq \frac{-15 + 1.7}{-0.44} = 29.9 \;\;\Rightarrow\;\; \hat{N}_2 \simeq 60.9 \tag{1.45}$$

This answer is found to be stable, so $N \simeq 61$ is the blocklength at which $p_{\rm b} \simeq 10^{-15}$.

Solution to exercise 1.6 (p.13).

(a) The probability of block error of the Hamming code is a sum of six terms – the probabilities that 2, 3, 4, 5, 6, or 7 errors occur in one block.

$$p_{\rm B} = \sum_{r=2}^{7} \binom{7}{r} f^r (1-f)^{7-r}. \tag{1.46}$$

To leading order, this goes as

$$p_{\rm B} \simeq \binom{7}{2} f^2 = 21 f^2. \tag{1.47}$$

(b) The probability of bit error of the Hamming code is smaller than the probability of block error because a block error rarely corrupts all bits in the decoded block. The leading-order behaviour is found by considering the outcome in the most probable case where the noise vector has weight two. The decoder will erroneously flip a *third* bit, so that the modified received vector (of length 7) differs in three bits from the transmitted vector. That means, if we average over all seven bits, the probability that a randomly chosen bit is flipped is 3/7 times the block error probability, to leading order. Now, what we really care about is the probability that

– and there are edges only between nodes in different classes. The graph and the code's parity-check matrix (1.30) are simply related to each other: each parity-check node corresponds to a row of **H** and each bit node corresponds to a column of **H**; for every 1 in **H**, there is an edge between the corresponding pair of nodes.

Having noticed this connection between linear codes and graphs, one way to invent linear codes is simply to think of a bipartite graph. For example, a pretty bipartite graph can be obtained from a dodecahedron by calling the vertices of the dodecahedron the parity-check nodes, and putting a transmitted bit on each edge in the dodecahedron. This construction defines a parity-check matrix in which every column has weight 2 and every row has weight 3. [The weight of a binary vector is the number of 1s it contains.]

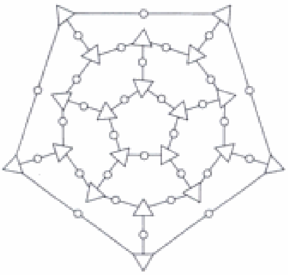


Figure 1.21. The graph defining the (30, 11) dodecahedron code. The circles are the 30 transmitted bits and the triangles are the 20 parity checks. One parity check is redundant.

This code has $N = 30$ bits, and it appears to have $M_{\text{apparent}} = 20$ parity-check constraints. Actually, there are only $M = 19$ *independent* constraints; the 20th constraint is redundant (that is, if 19 constraints are satisfied, then the 20th is automatically satisfied); so the number of source bits is $K = N - M = 11$. The code is a (30, 11) code.

It is hard to find a decoding algorithm for this code, but we can estimate its probability of error by finding its lowest weight codewords. If we flip all the bits surrounding one face of the original dodecahedron, then all the parity checks will be satisfied; so the code has 12 codewords of weight 5, one for each face. Since the lowest-weight codewords have weight 5, we say that the code has distance $d = 5$; the (7, 4) Hamming code had distance 3 and could correct all single bit-flip errors. A code with distance 5 can correct all double bit-flip errors, but there are some triple bit-flip errors that it cannot correct. So the error probability of this code, assuming a binary symmetric channel, will be dominated, at least for low noise levels f, by a term of order f^3, perhaps something like

$$12\binom{5}{3}f^3(1-f)^{27}. \tag{1.55}$$

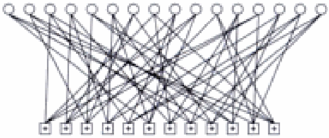


Figure 1.22. Graph of a rate-1/4 low-density parity-check code (Gallager code) with blocklength $N = 16$, and $M = 12$ parity-check constraints. Each white circle represents a transmitted bit. Each bit participates in $j = 3$ constraints, represented by ⊞ squares. The edges between nodes were placed at random. (See Chapter 47 for more.)

Of course, there is no obligation to make codes whose graphs can be represented on a plane, as this one can; the best linear codes, which have simple graphical descriptions, have graphs that are more tangled, as illustrated by the tiny (16, 4) code of figure 1.22.

Furthermore, there is no reason for sticking to linear codes; indeed some nonlinear codes – codes whose codewords cannot be defined by a linear equation like $\mathbf{Ht} = \mathbf{0}$ – have very good properties. But the encoding and decoding of a nonlinear code are even trickier tasks.

Solution to exercise 1.10 (p.14). First let's assume we are making a linear code and decoding it with syndrome decoding. If there are N transmitted bits, then the number of possible error patterns of weight up to two is

$$\binom{N}{2}+\binom{N}{1}+\binom{N}{0}. \tag{1.56}$$

For $N = 14$, that's $91 + 14 + 1 = 106$ patterns. Now, every distinguishable error pattern must give rise to a distinct syndrome; and the syndrome is a list of M bits, so the maximum possible number of syndromes is 2^M. For a (14, 8) code, $M = 6$, so there are at most $2^6 = 64$ syndromes. The number of possible error patterns of weight up to two, 106, is bigger than the number of syndromes, 64, so we can immediately rule out the possibility that there is a (14, 8) code that is 2-error-correcting.

The same counting argument works fine for nonlinear codes too. When the decoder receives $\mathbf{r} = \mathbf{t} + \mathbf{n}$, his aim is to deduce both $\mathbf{t}$ and $\mathbf{n}$ from $\mathbf{r}$. If it is the case that the sender can select any transmission $\mathbf{t}$ from a code of size $S_{\mathbf{t}}$, and the channel can select any noise vector from a set of size $S_{\mathbf{n}}$, and those two selections can be recovered from the received bit string $\mathbf{r}$, which is one of at most 2^N possible strings, then it must be the case that

$$S_{\mathbf{t}} S_{\mathbf{n}} \leq 2^N. \tag{1.57}$$

So, for a (N, K) two-error-correcting code, whether linear or nonlinear,

$$2^K \left[\binom{N}{2} + \binom{N}{1} + \binom{N}{0} \right] \leq 2^N. \tag{1.58}$$

Solution to exercise 1.11 (p.14). There are various strategies for making codes that can correct multiple errors, and I strongly recommend you think out one or two of them for yourself.

If your approach uses a linear code, e.g., one with a collection of M parity checks, it is helpful to bear in mind the counting argument given in the previous exercise, in order to anticipate how many parity checks, M, you might need.

Examples of codes that can correct any two errors are the $(30, 11)$ dodecahedron code in the previous solution, and the $(15, 6)$ pentagonful code to be introduced on p.221. Further simple ideas for making codes that can correct multiple errors from codes that can correct only one error are discussed in section 13.7.

Solution to exercise 1.12 (p.16). The probability of error of R_3^2 is, to leading order,

$$p_{\mathrm{b}}(\mathrm{R}_3^2) \simeq 3 \left[p_{\mathrm{b}}(\mathrm{R}_3)\right]^2 = 3(3f^2)^2 + \cdots = 27f^4 + \cdots, \tag{1.59}$$

whereas the probability of error of R_9 is dominated by the probability of five flips,

$$p_{\mathrm{b}}(\mathrm{R}_9) \simeq \binom{9}{5} f^5 (1-f)^4 \simeq 126 f^5 + \cdots. \tag{1.60}$$

The R_3^2 decoding procedure is therefore suboptimal, since there are noise vectors of weight four that cause it to make a decoding error.

It has the advantage, however, of requiring smaller computational resources: only memorization of three bits, and counting up to three, rather than counting up to nine.

This simple code illustrates an important concept. Concatenated codes are widely used in practice because concatenation allows large codes to be implemented using simple encoding and decoding hardware. Some of the best known practical codes are concatenated codes.

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

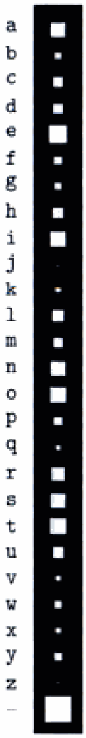


Figure 2.1. Probability distribution over the 27 outcomes for a randomly selected letter in an English language document (estimated from *The Frequently Asked Questions Manual for Linux*). The picture shows the probabilities by the areas of white squares.

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.

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 is 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.

$$P(a=1\,|\,b=1) = \frac{P(b=1\,|\,a=1)P(a=1)}{P(b=1\,|\,a=1)P(a=1) + P(b=1\,|\,a=0)P(a=0)} \tag{2.16}$$

$$= \frac{0.95 \times 0.01}{0.95 \times 0.01 + 0.05 \times 0.99} \tag{2.17}$$

$$= 0.16. \tag{2.18}$$

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

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 is 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 $B(x \mid y)$.

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 $x \mid y$ and the degree of belief in the proposition y. There is a function g such that

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

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

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

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:

$$\begin{aligned} P(u \,|\, n_B, N) &= \frac{P(u, n_B \,|\, N)}{P(n_B \,|\, N)} && (2.21) \\ &= \frac{P(n_B \,|\, u, N) P(u)}{P(n_B \,|\, N)}. && (2.22) \end{aligned}$$

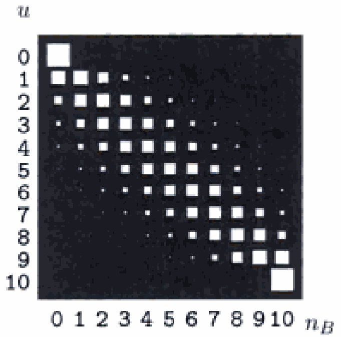


Figure 2.5. Joint probability of u and n_B for Bill and Fred's urn problem, after $N = 10$ draws.

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:

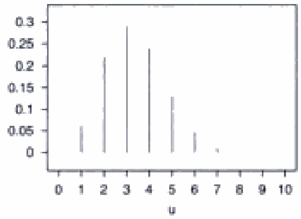
$$P(n_B \,|\, N) = \sum_u P(u, n_B \,|\, N) = \sum_u P(u)P(n_B \,|\, u, N). \tag{2.24}$$

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 $P(u \,|\, n_B, N)$ is called the *posterior probability* of u given n_B. The normalizing constant $P(n_B \,|\, N)$ 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: $P(n_B \,|\, N)$ 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:

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

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,

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

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

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

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

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

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.

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:

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

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:

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

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:

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

Entropy is additive for independent random variables:

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

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

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

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

$$H(X) = {}^1\!/\!_2 \log 2 + {}^1\!/\!_4 \log 4 + {}^1\!/\!_4 \log 4 = 1.5; \qquad (2.41)$$

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 $\mathcal{A} = \{\mathtt{0}, \mathtt{1}, \ldots, \mathtt{9}, \mathtt{a}, \mathtt{b}, \ldots, \mathtt{z}\}$; with probability $1/3$, x is a numeral $(\mathtt{0}, \ldots, \mathtt{9})$; with probability $1/3$, x is a vowel $(\mathtt{a}, \mathtt{e}, \mathtt{i}, \mathtt{o}, \mathtt{u})$; 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 **heist**.

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.

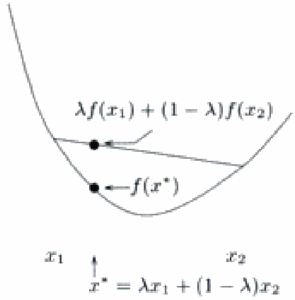

$\lambda f(x_1) + (1-\lambda) f(x_2)$
$f(x^*)$
x_1
x_2
$x^* = \lambda x_1 + (1-\lambda) x_2$

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 $0 \leq \lambda \leq 1$,

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

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 $x > 0$.

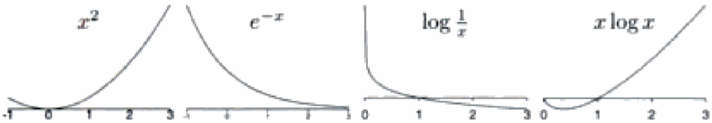


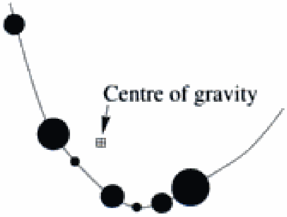
Figure 2.11. Convex $\smile$ functions.

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 $100\,\mathrm{m}^2$. □

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.$^{[3,\ \text{p.41}]}$ (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?

This exercise is intended to help you think about the central-limit theorem, which says that if independent random variables $x_1, x_2, \ldots, x_N$ have means μ_n and finite variances σ_n^2, then, in the limit of large N, the sum $\sum_n x_n$ has a distribution that tends to a normal (Gaussian) distribution with mean $\sum_n \mu_n$ and variance $\sum_n \sigma_n^2$.

Inference problems

Exercise 2.17.$^{[2,\ \text{p.41}]}$ 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.$^{[2,\ \text{p.42}]}$ Let x and y be dependent 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

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

▷ Exercise 2.19.$^{[2,\ \text{p.42}]}$ 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

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

(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 that all three bits are flipped to the probability that exactly two bits are flipped.

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

$$P(\text{error}) = \sum_{\mathbf{r}} P(\mathbf{r}) P(\text{error} \,|\, \mathbf{r}). \qquad (2.55)$$

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

▷ 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 (cf. 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:

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

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

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

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 $\text{var}[n_B]$ 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.

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, $\lambda = \frac{p_2}{\sum_{i=2}^{I} p_i}$.

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

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

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

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

$$p = \frac{1}{1+2^{-a}}. \tag{2.68}$$

Solution to exercise 2.18 (p.36).

$$P(x\,|\,y) = \frac{P(y\,|\,x)P(x)}{P(y)} \tag{2.69}$$

$$\Rightarrow \frac{P(x=1\,|\,y)}{P(x=0\,|\,y)} = \frac{P(y\,|\,x=1)}{P(y\,|\,x=0)}\frac{P(x=1)}{P(x=0)} \tag{2.70}$$

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

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

$$P(x, d_1, d_2) = P(x)P(d_1\,|\,x)P(d_2\,|\,x). \tag{2.72}$$

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

$$\frac{P(x=1\,|\,\{d_i\})}{P(x=0\,|\,\{d_i\})} = \frac{P(\{d_i\}\,|\,x=1)}{P(\{d_i\}\,|\,x=0)}\frac{P(x=1)}{P(x=0)} \tag{2.73}$$

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

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

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

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

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

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. □

(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?

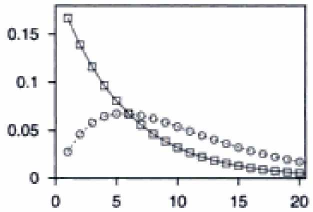


Figure 2.13. The probability distribution of the number of rolls r_1 from one 6 to the next (falling solid line),

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

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:

$$P(\mathbf{r} = 000) = {}^1\!/\!_2(1-f)^3 + {}^1\!/\!_2 f^3, \tag{2.108}$$

$$P(\mathbf{r} = 001) = {}^1\!/\!_2 f(1-f)^2 + {}^1\!/\!_2 f^2(1-f) = {}^1\!/\!_2 f(1-f). \tag{2.109}$$

The posterior probabilities are represented by

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

and

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

The probabilities of error in these representative cases are thus

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

and

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

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

$$\begin{aligned} P(\text{error}) &= \sum_{\mathbf{r}} P(\mathbf{r}) P(\text{error} \mid \mathbf{r}) \\ &= 2[{}^1\!/\!_2(1-f)^3 + {}^1\!/\!_2 f^3]\frac{f^3}{(1-f)^3 + f^3} + 6[{}^1\!/\!_2 f(1-f)]f. \end{aligned}$$

The first two terms are for the cases $\mathbf{r} = 000$ and 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.

About Chapter 3

If you are eager to get on to information theory, data compression, and noisy channels, you can skip to Chapter 4. Data compression and data modelling are intimately connected, however, so you'll probably want to come back to this chapter by the time you get to Chapter 6. Before reading Chapter 3, it might be good to look at the following exercises.

▷ Exercise 3.1.[2, p.59] A die is selected at random from two twenty-faced dice on which the symbols 1–10 are written with nonuniform frequency as follows.

Symbol	1	2	3	4	5	6	7	8	9	10
Number of faces of die A	6	4	3	2	1	1	1	1	1	0
Number of faces of die B	3	3	2	2	2	2	2	2	1	1

The randomly chosen die is rolled 7 times, with the following outcomes:

$$5, 3, 9, 3, 8, 4, 7.$$

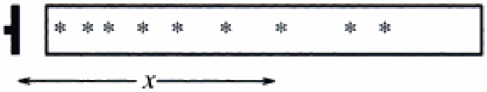
What is the probability that the die is die A?

▷ Exercise 3.2.[2, p.59] Assume that there is a third twenty-faced die, die C, on which the symbols 1–20 are written once each. As above, one of the three dice is selected at random and rolled 7 times, giving the outcomes: 3, 5, 4, 8, 3, 9, 7.

What is the probability that the die is (a) die A, (b) die B, (c) die C?

Exercise 3.3.[3, p.48] Inferring a decay constant

Unstable particles are emitted from a source and decay at a distance x, a real number that has an exponential probability distribution with characteristic length λ. Decay events can only be observed if they occur in a window extending from $x = 1\,\text{cm}$ to $x = 20\,\text{cm}$. N decays are observed at locations $\{x_1, \ldots, x_N\}$. What is λ?



▷ Exercise 3.4.[3, p.55] Forensic evidence

Two people have left traces of their own blood at the scene of a crime. A suspect, Oliver, is tested and found to have type 'O' blood. The blood groups of the two traces are found to be of type 'O' (a common type in the local population, having frequency 60%) and of type 'AB' (a rare type, with frequency 1%). Do these data (type 'O' and 'AB' blood were found at scene) give evidence in favour of the proposition that Oliver was one of the two people present at the crime?

3

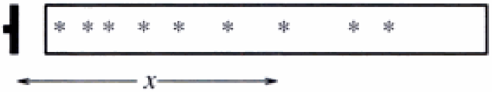
More about Inference

It is not a controversial statement that Bayes' theorem provides the correct language for describing the inference of a message communicated over a noisy channel, as we used it in Chapter 1 (p.6). But strangely, when it comes to other inference problems, the use of Bayes' theorem is not so widespread.

▶ 3.1 A first inference problem

When I was an undergraduate in Cambridge, I was privileged to receive supervisions from Steve Gull. Sitting at his desk in a dishevelled office in St. John's College, I asked him how one ought to answer an old Tripos question (exercise 3.3):

> Unstable particles are emitted from a source and decay at a distance x, a real number that has an exponential probability distribution with characteristic length λ. Decay events can only be observed if they occur in a window extending from $x = 1\,\text{cm}$ to $x = 20\,\text{cm}$. N decays are observed at locations $\{x_1, \ldots, x_N\}$. What is λ?



I had scratched my head over this for some time. My education had provided me with a couple of approaches to solving such inference problems: constructing 'estimators' of the unknown parameters; or 'fitting' the model to the data, or to a processed version of the data.

Since the mean of an unconstrained exponential distribution is λ, it seemed reasonable to examine the sample mean $\bar{x} = \sum_n x_n/N$ and see if an estimator $\hat{\lambda}$ could be obtained from it. It was evident that the estimator $\hat{\lambda} = \bar{x} - 1$ would be appropriate for $\lambda \ll 20\,\text{cm}$, but not for cases where the truncation of the distribution at the right-hand side is significant; with a little ingenuity and the introduction of ad hoc bins, promising estimators for $\lambda \gg 20$ cm could be constructed. But there was no obvious estimator that would work under all conditions.

Nor could I find a satisfactory approach based on fitting the density $P(x \mid \lambda)$ to a histogram derived from the data. I was stuck.

What is the general solution to this problem and others like it? Is it always necessary, when confronted by a new inference problem, to grope in the dark for appropriate 'estimators' and worry about finding the 'best' estimator (whatever that means)?

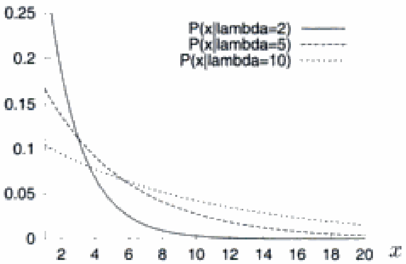


Figure 3.1. The probability density $P(x \mid \lambda)$ as a function of x.

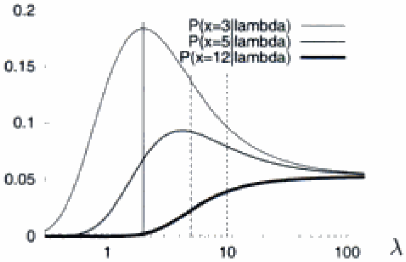


Figure 3.2. The probability density $P(x \mid \lambda)$ as a function of λ, for three different values of x. When plotted this way round, the function is known as the *likelihood* of λ. The marks indicate the three values of λ, $\lambda = 2, 5, 10$, that were used in the preceding figure.

Steve wrote down the probability of one data point, given λ:

$$P(x \mid \lambda) = \begin{cases} \frac{1}{\lambda} e^{-x/\lambda}/Z(\lambda) & 1 < x < 20 \\ 0 & \text{otherwise} \end{cases} \tag{3.1}$$

where

$$Z(\lambda) = \int_1^{20} \mathrm{d}x \, \tfrac{1}{\lambda} e^{-x/\lambda} = \left(e^{-1/\lambda} - e^{-20/\lambda} \right). \tag{3.2}$$

This seemed obvious enough. Then he wrote *Bayes' theorem*:

$$P(\lambda \mid \{x_1, \ldots, x_N\}) = \frac{P(\{x\} \mid \lambda) P(\lambda)}{P(\{x\})} \tag{3.3}$$

$$\propto \frac{1}{(\lambda Z(\lambda))^N} \exp\left(-\textstyle\sum_1^N x_n/\lambda \right) P(\lambda). \tag{3.4}$$

Suddenly, the straightforward distribution $P(\{x_1, \ldots, x_N\} \mid \lambda)$, defining the probability of the data given the hypothesis λ, was being turned on its head so as to define the probability of a hypothesis given the data. A simple figure showed the probability of a single data point $P(x \mid \lambda)$ as a familiar function of x, for different values of λ (figure 3.1). Each curve was an innocent exponential, normalized to have area 1. Plotting the same function as a function of λ for a fixed value of x, something remarkable happens: a peak emerges (figure 3.2). To help understand these two points of view of the one function, figure 3.3 shows a surface plot of $P(x \mid \lambda)$ as a function of x and λ.

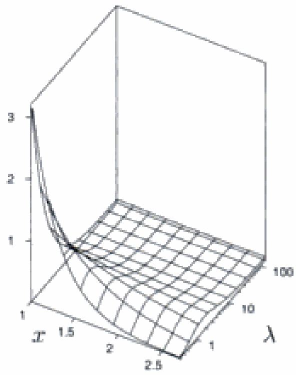


Figure 3.3. The probability density $P(x \mid \lambda)$ as a function of x and λ. Figures 3.1 and 3.2 are vertical sections through this surface.

For a dataset consisting of several points, e.g., the six points $\{x\}_{n=1}^N = \{1.5, 2, 3, 4, 5, 12\}$, the likelihood function $P(\{x\} \mid \lambda)$ is the product of the N functions of λ, $P(x_n \mid \lambda)$ (figure 3.4).

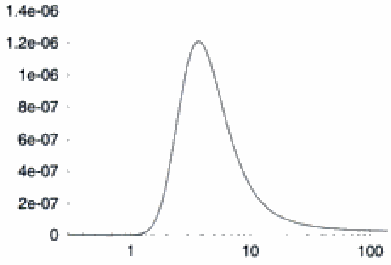


Figure 3.4. The likelihood function in the case of a six-point dataset, $P(\{x\} = \{1.5, 2, 3, 4, 5, 12\} \mid \lambda)$, as a function of λ.

Steve summarized Bayes' theorem as embodying the fact that

> what you know about λ after the data arrive is what you knew before $[P(\lambda)]$, and what the data told you $[P(\{x\} \mid \lambda)]$.

Probabilities are used here to quantify degrees of belief. To nip possible confusion in the bud, it must be emphasized that the hypothesis λ that correctly describes the situation is *not* a *stochastic* variable, and the fact that the Bayesian uses a probability distribution P does *not* mean that he thinks of the world as stochastically changing its nature between the states described by the different hypotheses. He uses the notation of probabilities to represent his *beliefs* about the mutually exclusive micro-hypotheses (here, values of λ), of which only one is actually true. That probabilities can denote degrees of belief, given assumptions, seemed reasonable to me.

The posterior probability distribution (3.4) represents the unique and complete solution to the problem. There is no need to invent 'estimators'; nor do we need to invent criteria for comparing alternative estimators with each other. Whereas orthodox statisticians offer twenty ways of solving a problem, and another twenty different criteria for deciding which of these solutions is the best, Bayesian statistics only offers one answer to a well-posed problem.

If you have any difficulty understanding this chapter I recommend ensuring you are happy with exercises 3.1 and 3.2 (p.47) then noting their similarity to exercise 3.3.

Assumptions in inference

Our inference is conditional on our assumptions [for example, the prior $P(\lambda)$]. Critics view such priors as a difficulty because they are 'subjective', but I don't see how it could be otherwise. How can one perform inference without making assumptions? I believe that it is of great value that Bayesian methods force one to make these tacit assumptions explicit.

First, once assumptions are made, the inferences are objective and unique, reproducible with complete agreement by anyone who has the same information and makes the same assumptions. For example, given the assumptions listed above, $\mathcal{H}$, and the data D, everyone will agree about the posterior probability of the decay length λ:

$$P(\lambda \mid D, \mathcal{H}) = \frac{P(D \mid \lambda, \mathcal{H}) P(\lambda \mid \mathcal{H})}{P(D \mid \mathcal{H})}. \tag{3.5}$$

Second, when the assumptions are explicit, they are easier to criticize, and easier to modify – indeed, we can quantify the sensitivity of our inferences to the details of the assumptions. For example, we can note from the likelihood curves in figure 3.2 that in the case of a single data point at $x = 5$, the likelihood function is less strongly peaked than in the case $x = 3$; the details of the prior $P(\lambda)$ become increasingly important as the sample mean $\bar{x}$ gets closer to the middle of the window, 10.5. In the case $x = 12$, the likelihood function doesn't have a peak at all – such data merely rule out small values of λ, and don't give any information about the relative probabilities of large values of λ. So in this case, the details of the prior at the small–λ end of things are not important, but at the large–λ end, the prior is important.

Third, when we are not sure which of various alternative assumptions is the most appropriate for a problem, we can treat this question as another inference task. Thus, given data D, we can compare alternative assumptions $\mathcal{H}$ using Bayes' theorem:

$$P(\mathcal{H} \mid D, I) = \frac{P(D \mid \mathcal{H}, I) P(\mathcal{H} \mid I)}{P(D \mid I)}, \tag{3.6}$$

Model comparison as inference

In order to perform model comparison, we write down Bayes' theorem again, but this time with a different argument on the left-hand side. We wish to know how probable $\mathcal{H}_1$ is given the data. By Bayes' theorem,

$$P(\mathcal{H}_1 \,|\, \mathbf{s}, F) = \frac{P(\mathbf{s} \,|\, F, \mathcal{H}_1) P(\mathcal{H}_1)}{P(\mathbf{s} \,|\, F)}. \tag{3.17}$$

Similarly, the posterior probability of $\mathcal{H}_0$ is

$$P(\mathcal{H}_0 \,|\, \mathbf{s}, F) = \frac{P(\mathbf{s} \,|\, F, \mathcal{H}_0) P(\mathcal{H}_0)}{P(\mathbf{s} \,|\, F)}. \tag{3.18}$$

The normalizing constant in both cases is $P(\mathbf{s} \,|\, F)$, which is the total probability of getting the observed data. If $\mathcal{H}_1$ and $\mathcal{H}_0$ are the only models under consideration, this probability is given by the sum rule:

$$P(\mathbf{s} \,|\, F) = P(\mathbf{s} \,|\, F, \mathcal{H}_1) P(\mathcal{H}_1) + P(\mathbf{s} \,|\, F, \mathcal{H}_0) P(\mathcal{H}_0). \tag{3.19}$$

To evaluate the posterior probabilities of the hypotheses we need to assign values to the prior probabilities $P(\mathcal{H}_1)$ and $P(\mathcal{H}_0)$; in this case, we might set these to 1/2 each. And we need to evaluate the data-dependent terms $P(\mathbf{s} \,|\, F, \mathcal{H}_1)$ and $P(\mathbf{s} \,|\, F, \mathcal{H}_0)$. We can give names to these quantities. The quantity $P(\mathbf{s} \,|\, F, \mathcal{H}_1)$ is a measure of how much the data favour $\mathcal{H}_1$, and we call it the *evidence* for model $\mathcal{H}_1$. We already encountered this quantity in equation (3.10) where it appeared as the normalizing constant of the first inference we made – the inference of $p_{\mathtt{a}}$ given the data.

How model comparison works: The evidence for a model is usually the normalizing constant of an earlier Bayesian inference.

We evaluated the normalizing constant for model $\mathcal{H}_1$ in (3.12). The evidence for model $\mathcal{H}_0$ is very simple because this model has no parameters to infer. Defining p_0 to be 1/6, we have

$$P(\mathbf{s} \,|\, F, \mathcal{H}_0) = p_0^{F_{\mathtt{a}}} (1 - p_0)^{F_{\mathtt{b}}}. \tag{3.20}$$

Thus the posterior probability ratio of model $\mathcal{H}_1$ to model $\mathcal{H}_0$ is

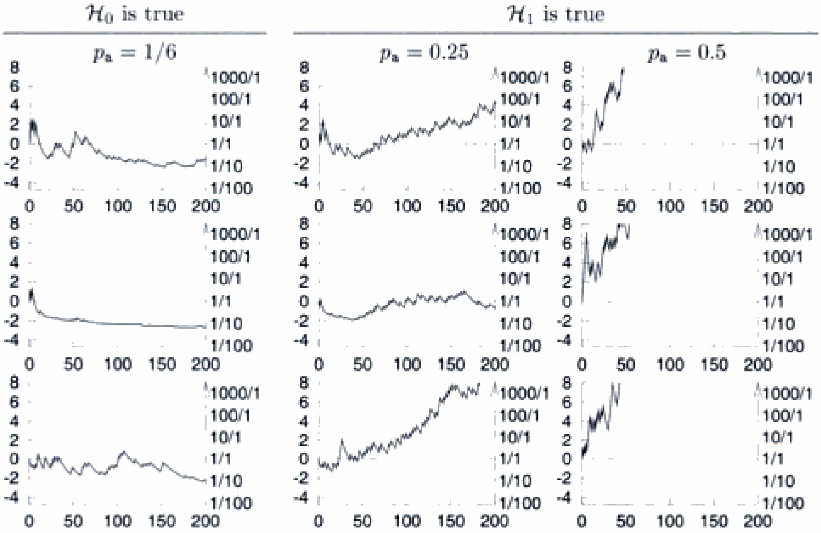
$$\begin{aligned} \frac{P(\mathcal{H}_1 \,|\, \mathbf{s}, F)}{P(\mathcal{H}_0 \,|\, \mathbf{s}, F)} &= \frac{P(\mathbf{s} \,|\, F, \mathcal{H}_1) P(\mathcal{H}_1)}{P(\mathbf{s} \,|\, F, \mathcal{H}_0) P(\mathcal{H}_0)} & (3.21) \\ &= \left. \frac{F_{\mathtt{a}}! F_{\mathtt{b}}!}{(F_{\mathtt{a}} + F_{\mathtt{b}} + 1)!} \right/ p_0^{F_{\mathtt{a}}} (1 - p_0)^{F_{\mathtt{b}}}. & (3.22) \end{aligned}$$

Some values of this posterior probability ratio are illustrated in table 3.5. The first five lines illustrate that some outcomes favour one model, and some favour the other. No outcome is completely incompatible with either model. With small amounts of data (six tosses, say) it is typically not the case that one of the two models is overwhelmingly more probable than the other. But with more data, the evidence against $\mathcal{H}_0$ given by any data set with the ratio $F_{\mathtt{a}}\!:\!F_{\mathtt{b}}$ differing from 1:5 mounts up. You can't predict in advance how much data are needed to be pretty sure which theory is true. It depends what p_0 is.

The simpler model, $\mathcal{H}_0$, since it has no adjustable parameters, is able to lose out by the biggest margin. The odds may be hundreds to one against it. The more complex model can never lose out by a large margin; there's no data set that is actually *unlikely* given model $\mathcal{H}_1$.

F	Data (F_a, F_b)	$\frac{P(\mathcal{H}_1 \mid \mathbf{s}, F)}{P(\mathcal{H}_0 \mid \mathbf{s}, F)}$	
6	$(5, 1)$	222.2	
6	$(3, 3)$	2.67	
6	$(2, 4)$	0.71	$= 1/1.4$
6	$(1, 5)$	0.356	$= 1/2.8$
6	$(0, 6)$	0.427	$= 1/2.3$
20	$(10, 10)$	96.5	
20	$(3, 17)$	0.2	$= 1/5$
20	$(0, 20)$	1.83	

Table 3.5. Outcome of model comparison between models $\mathcal{H}_1$ and $\mathcal{H}_0$ for the 'bent coin'. Model $\mathcal{H}_0$ states that $p_a = 1/6$, $p_b = 5/6$.


$\mathcal{H}_0$ is true

$\mathcal{H}_1$ is true

$p_a = 1/6$

$p_a = 0.25$

$p_a = 0.5$

Figure 3.6. Typical behaviour of the evidence in favour of $\mathcal{H}_1$ as bent coin tosses accumulate under three different conditions. Horizontal axis is the number of tosses, F. The vertical axis on the left is $\ln \frac{P(\mathbf{s} \mid F, \mathcal{H}_1)}{P(\mathbf{s} \mid F, \mathcal{H}_0)}$; the right-hand vertical axis shows the values of $\frac{P(\mathbf{s} \mid F, \mathcal{H}_1)}{P(\mathbf{s} \mid F, \mathcal{H}_0)}$.
(See also figure 3.8, p.60.)

▷ Exercise 3.6.[2] Show that after F tosses have taken place, the biggest value that the log evidence ratio

$$\log \frac{P(\mathbf{s} \mid F, \mathcal{H}_1)}{P(\mathbf{s} \mid F, \mathcal{H}_0)} \tag{3.23}$$

can have scales *linearly* with F if $\mathcal{H}_1$ is more probable, but the log evidence in favour of $\mathcal{H}_0$ can grow at most as $\log F$.

▷ Exercise 3.7.[3, p.60] Putting your sampling theory hat on, assuming F_a has not yet been measured, compute a plausible range that the log evidence ratio might lie in, as a function of F and the true value of p_a, and sketch it as a function of F for $p_a = p_0 = 1/6$, $p_a = 0.25$, and $p_a = 1/2$. [Hint: sketch the log evidence as a function of the random variable F_a and work out the mean and standard deviation of F_a.]

Typical behaviour of the evidence

Figure 3.6 shows the log evidence ratio as a function of the number of tosses, F, in a number of simulated experiments. In the left-hand experiments, $\mathcal{H}_0$ was true. In the right-hand ones, $\mathcal{H}_1$ was true, and the value of p_a was either 0.25 or 0.5.

We will discuss model comparison more in a later chapter.

▶ 3.4 An example of legal evidence

The following example illustrates that there is more to Bayesian inference than the priors.

> Two people have left traces of their own blood at the scene of a crime. A suspect, Oliver, is tested and found to have type 'O' blood. The blood groups of the two traces are found to be of type 'O' (a common type in the local population, having frequency 60%) and of type 'AB' (a rare type, with frequency 1%). Do these data (type 'O' and 'AB' blood were found at scene) give evidence in favour of the proposition that Oliver was one of the two people present at the crime?

A careless lawyer might claim that the fact that the suspect's blood type was found at the scene is positive evidence for the theory that he was present. But this is not so.

Denote the proposition 'the suspect and one unknown person were present' by S. The alternative, $\bar{S}$, states 'two unknown people from the population were present'. The prior in this problem is the prior probability ratio between the propositions S and $\bar{S}$. This quantity is important to the final verdict and would be based on all other available information in the case. Our task here is just to evaluate the contribution made by the data D, that is, the likelihood ratio, $P(D \mid S, \mathcal{H})/P(D \mid \bar{S}, \mathcal{H})$. In my view, a jury's task should generally be to multiply together carefully evaluated likelihood ratios from each independent piece of admissible evidence with an equally carefully reasoned prior probability. [This view is shared by many statisticians but learned British appeal judges recently disagreed and actually overturned the verdict of a trial because the jurors *had* been taught to use Bayes' theorem to handle complicated DNA evidence.]

The probability of the data given S is the probability that one unknown person drawn from the population has blood type AB:

$$P(D \mid S, \mathcal{H}) = p_{\mathrm{AB}} \tag{3.24}$$

(since given S, we already know that one trace will be of type O). The probability of the data given $\bar{S}$ is the probability that two unknown people drawn from the population have types O and AB:

$$P(D \mid \bar{S}, \mathcal{H}) = 2\, p_{\mathrm{O}}\, p_{\mathrm{AB}}. \tag{3.25}$$

In these equations $\mathcal{H}$ denotes the assumptions that two people were present and left blood there, and that the probability distribution of the blood groups of unknown people in an explanation is the same as the population frequencies.

Dividing, we obtain the likelihood ratio:

$$\frac{P(D \mid S, \mathcal{H})}{P(D \mid \bar{S}, \mathcal{H})} = \frac{1}{2p_{\mathrm{O}}} = \frac{1}{2 \times 0.6} = 0.83. \tag{3.26}$$

Thus the data in fact provide weak evidence *against* the supposition that Oliver was present.

This result may be found surprising, so let us examine it from various points of view. First consider the case of another suspect, Alberto, who has type AB. Intuitively, the data do provide evidence in favour of the theory S'

that this suspect was present, relative to the null hypothesis $\bar{S}$. And indeed the likelihood ratio in this case is:

$$\frac{P(D \mid S', \mathcal{H})}{P(D \mid \bar{S}, \mathcal{H})} = \frac{1}{2\, p_{\rm AB}} = 50. \tag{3.27}$$

Now let us change the situation slightly; imagine that 99% of people are of blood type O, and the rest are of type AB. Only these two blood types exist in the population. The data at the scene are the same as before. Consider again how these data influence our beliefs about Oliver, a suspect of type O, and Alberto, a suspect of type AB. Intuitively, we still believe that the presence of the rare AB blood provides positive evidence that Alberto was there. But does the fact that type O blood was detected at the scene favour the hypothesis that Oliver was present? If this were the case, that would mean that regardless of who the suspect is, the data make it more probable they were present; everyone in the population would be under greater suspicion, which would be absurd. The data may be *compatible* with any suspect of either blood type being present, but if they provide evidence *for* some theories, they must also provide evidence *against* other theories.

Here is another way of thinking about this: imagine that instead of two people's blood stains there are ten, and that in the entire local population of one hundred, there are ninety type O suspects and ten type AB suspects. Consider a particular type O suspect, Oliver: without any other information, and before the blood test results come in, there is a one in 10 chance that he was at the scene, since we know that 10 out of the 100 suspects were present. We now get the results of blood tests, and find that *nine* of the ten stains are of type AB, and *one* of the stains is of type O. Does this make it more likely that Oliver was there? No, there is now only a one in ninety chance that he was there, since we know that only one person present was of type O.

Maybe the intuition is aided finally by writing down the formulae for the general case where $n_{\rm O}$ blood stains of individuals of type O are found, and $n_{\rm AB}$ of type AB, a total of N individuals in all, and unknown people come from a large population with fractions $p_{\rm O}, p_{\rm AB}$. (There may be other blood types too.) The task is to evaluate the likelihood ratio for the two hypotheses: S, 'the type O suspect (Oliver) and $N-1$ unknown others left N stains'; and $\bar{S}$, 'N unknowns left N stains'. The probability of the data under hypothesis $\bar{S}$ is just the probability of getting $n_{\rm O}, n_{\rm AB}$ individuals of the two types when N individuals are drawn at random from the population:

$$P(n_{\rm O}, n_{\rm AB} \mid \bar{S}) = \frac{N!}{n_{\rm O}!\, n_{\rm AB}!} p_{\rm O}^{n_{\rm O}} p_{\rm AB}^{n_{\rm AB}}. \tag{3.28}$$

In the case of hypothesis S, we need the distribution of the $N-1$ other individuals:

$$P(n_{\rm O}, n_{\rm AB} \mid S) = \frac{(N-1)!}{(n_{\rm O}-1)!\, n_{\rm AB}!} p_{\rm O}^{n_{\rm O}-1} p_{\rm AB}^{n_{\rm AB}}. \tag{3.29}$$

The likelihood ratio is:

$$\frac{P(n_{\rm O}, n_{\rm AB} \mid S)}{P(n_{\rm O}, n_{\rm AB} \mid \bar{S})} = \frac{n_{\rm O}/N}{p_{\rm O}}. \tag{3.30}$$

This is an instructive result. The likelihood ratio, i.e. the contribution of these data to the question of whether Oliver was present, depends simply on a comparison of the frequency of his blood type in the observed data with the background frequency in the population. There is no dependence on the counts of the other types found at the scene, or their frequencies in the population.

If there are more type O stains than the average number expected under hypothesis $\bar{S}$, then the data give evidence in favour of the presence of Oliver. Conversely, if there are fewer type O stains than the expected number under $\bar{S}$, then the data reduce the probability of the hypothesis that he was there. In the special case $n_{\rm O}/N = p_{\rm O}$, the data contribute no evidence either way, regardless of the fact that the data are compatible with the hypothesis S.

▶ 3.5 Exercises

Exercise 3.8.[2, p.60] The three doors, normal rules.

On a game show, a contestant is told the rules as follows:

> There are three doors, labelled 1, 2, 3. A single prize has been hidden behind one of them. You get to select one door. Initially your chosen door will *not* be opened. Instead, the gameshow host will open one of the other two doors, and *he will do so in such a way as not to reveal the prize.* For example, if you first choose door 1, he will then open one of doors 2 and 3, and it is guaranteed that he will choose which one to open so that the prize will not be revealed.
>
> At this point, you will be given a fresh choice of door: you can either stick with your first choice, or you can switch to the other closed door. All the doors will then be opened and you will receive whatever is behind your final choice of door.

Imagine that the contestant chooses door 1 first; then the gameshow host opens door 3, revealing nothing behind the door, as promised. Should the contestant (a) stick with door 1, or (b) switch to door 2, or (c) does it make no difference?

Exercise 3.9.[2, p.61] The three doors, earthquake scenario.

Imagine that the game happens again and just as the gameshow host is about to open one of the doors a violent earthquake rattles the building and one of the three doors flies open. It happens to be door 3, and it happens not to have the prize behind it. The contestant had initially chosen door 1.

Repositioning his toupée, the host suggests, 'OK, since you chose door 1 initially, door 3 is a valid door for me to open, according to the rules of the game; I'll let door 3 stay open. Let's carry on as if nothing happened.'

Should the contestant stick with door 1, or switch to door 2, or does it make no difference? Assume that the prize was placed randomly, that the gameshow host does not know where it is, and that the door flew open because its latch was broken by the earthquake.

[A similar alternative scenario is a gameshow whose *confused host* forgets the rules, and where the prize is, and opens one of the unchosen doors at random. He opens door 3, and the prize is not revealed. Should the contestant choose what's behind door 1 or door 2? Does the optimal decision for the contestant depend on the contestant's beliefs about whether the gameshow host is confused or not?]

▷ Exercise 3.10.[2] Another example in which the emphasis is not on priors. You visit a family whose three children are all at the local school. You don't

(b) $P(p_{\rm a} \mid {\rm s}={\rm bbb}, F=3) \propto (1-p_{\rm a})^3$. The most probable value of $p_{\rm a}$ (i.e., the value that maximizes the posterior probability density) is 0. The mean value of $p_{\rm a}$ is 1/5.

See figure 3.7b.

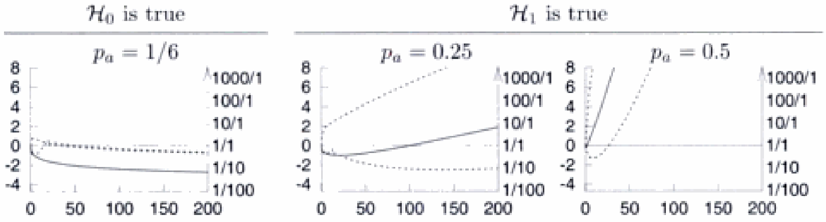


Figure 3.8. Range of plausible values of the log evidence in favour of $\mathcal{H}_1$ as a function of F. The vertical axis on the left is $\log \frac{P({\rm s} \mid F,\mathcal{H}_1)}{P({\rm s} \mid F,\mathcal{H}_0)}$; the right-hand vertical axis shows the values of $\frac{P({\rm s} \mid F,\mathcal{H}_1)}{P({\rm s} \mid F,\mathcal{H}_0)}$.
The solid line shows the log evidence if the random variable F_a takes on its mean value, $F_a = p_a F$. The dotted lines show (approximately) the log evidence if F_a is at its 2.5th or 97.5th percentile.
(See also figure 3.6, p.54.)

Solution to exercise 3.7 (p.54). The curves in figure 3.8 were found by finding the mean and standard deviation of F_a, then setting F_a to the mean ± two standard deviations to get a 95% plausible range for F_a, and computing the three corresponding values of the log evidence ratio.

Solution to exercise 3.8 (p.57). Let $\mathcal{H}_i$ denote the hypothesis that the prize is behind door i. We make the following assumptions: the three hypotheses $\mathcal{H}_1$, $\mathcal{H}_2$ and $\mathcal{H}_3$ are equiprobable *a priori*, i.e.,

$$P(\mathcal{H}_1) = P(\mathcal{H}_2) = P(\mathcal{H}_3) = \frac{1}{3}. \tag{3.36}$$

The datum we receive, after choosing door 1, is one of $D=3$ and $D=2$ (meaning door 3 or 2 is opened, respectively). We assume that these two possible outcomes have the following probabilities. If the prize is behind door 1 then the host has a free choice; in this case we assume that the host selects at random between $D=2$ and $D=3$. Otherwise the choice of the host is forced and the probabilities are 0 and 1.

$$\left|\begin{array}{l|l|l} P(D=2 \mid \mathcal{H}_1) = 1/2 & P(D=2 \mid \mathcal{H}_2) = 0 & P(D=2 \mid \mathcal{H}_3) = 1 \\ P(D=3 \mid \mathcal{H}_1) = 1/2 & P(D=3 \mid \mathcal{H}_2) = 1 & P(D=3 \mid \mathcal{H}_3) = 0 \end{array}\right| \tag{3.37}$$

Now, using Bayes' theorem, we evaluate the posterior probabilities of the hypotheses:

$$P(\mathcal{H}_i \mid D=3) = \frac{P(D=3 \mid \mathcal{H}_i)P(\mathcal{H}_i)}{P(D=3)} \tag{3.38}$$

$$\left| P(\mathcal{H}_1 \mid D=3) = \frac{(1/2)(1/3)}{P(D=3)} \;\middle|\; P(\mathcal{H}_2 \mid D=3) = \frac{(1)(1/3)}{P(D=3)} \;\middle|\; P(\mathcal{H}_3 \mid D=3) = \frac{(0)(1/3)}{P(D=3)} \right| \tag{3.39}$$

The denominator $P(D=3)$ is (1/2) because it is the normalizing constant for this posterior distribution. So

$$\left| P(\mathcal{H}_1 \mid D=3) \;=\; 1/3 \;\middle|\; P(\mathcal{H}_2 \mid D=3) \;=\; 2/3 \;\middle|\; P(\mathcal{H}_3 \mid D=3) \;=\; 0. \right| \tag{3.40}$$

So the contestant should switch to door 2 in order to have the biggest chance of getting the prize.

Many people find this outcome surprising. There are two ways to make it more intuitive. One is to play the game thirty times with a friend and keep track of the frequency with which switching gets the prize. Alternatively, you can perform a thought experiment in which the game is played with a million doors. The rules are now that the contestant chooses one door, then the game show host opens 999,998 doors in such a way as not to reveal the prize, leaving the *contestant's* selected door and *one other door* closed. The contestant may now stick or switch. Imagine the contestant confronted by a million doors, of which doors 1 and 234,598 have not been opened, door 1 having been the contestant's initial guess. Where do you think the prize is?

Solution to exercise 3.9 (p.57). If door 3 is opened by an earthquake, the inference comes out differently – even though visually the scene looks the same. The nature of the data, and the probability of the data, are both now different. The possible data outcomes are, firstly, that any number of the doors might have opened. We could label the eight possible outcomes $\mathbf{d} = (0,0,0), (0,0,1), (0,1,0), (1,0,0), (0,1,1), \ldots, (1,1,1)$. Secondly, it might be that the prize is visible after the earthquake has opened one or more doors. So the data D consists of the value of $\mathbf{d}$, and a statement of whether the prize was revealed. It is hard to say what the probabilities of these outcomes are, since they depend on our beliefs about the reliability of the door latches and the properties of earthquakes, but it is possible to extract the desired posterior probability without naming the values of $P(\mathbf{d} \,|\, \mathcal{H}_i)$ for each $\mathbf{d}$. All that matters are the relative values of the quantities $P(D \,|\, \mathcal{H}_1)$, $P(D \,|\, \mathcal{H}_2)$, $P(D \,|\, \mathcal{H}_3)$, for the value of D that actually occurred. [This is the *likelihood principle*, which we met in section 2.3.] The value of D that actually occurred is '$\mathbf{d} = (0,0,1)$, and no prize visible'. First, it is clear that $P(D \,|\, \mathcal{H}_3) = 0$, since the datum that no prize is visible is incompatible with $\mathcal{H}_3$. Now, assuming that the contestant selected door 1, how does the probability $P(D \,|\, \mathcal{H}_1)$ compare with $P(D \,|\, \mathcal{H}_2)$? Assuming that earthquakes are not sensitive to decisions of game show contestants, these two quantities have to be equal, by symmetry. We don't know how likely it is that door 3 falls off its hinges, but however likely it is, it's just as likely to do so whether the prize is behind door 1 or door 2. So, if $P(D \,|\, \mathcal{H}_1)$ and $P(D \,|\, \mathcal{H}_2)$ are equal, we obtain:

$$
\left|\; P(\mathcal{H}_1|D) = \frac{P(D|\mathcal{H}_1)(1/3)}{P(D)} = 1/2 \;\middle|\; P(\mathcal{H}_2|D) = \frac{P(D|\mathcal{H}_2)(1/3)}{P(D)} = 1/2 \;\middle|\; P(\mathcal{H}_3|D) = \frac{P(D|\mathcal{H}_3)(1/3)}{P(D)} = 0. \;\right| \tag{3.41}
$$

The two possible hypotheses are now equally likely.

If we assume that the host knows where the prize is and might be acting deceptively, then the answer might be further modified, because we have to view the host's words as part of the data.

Confused? It's well worth making sure you understand these two gameshow problems. Don't worry, I slipped up on the second problem, the first time I met it.

There is a general rule which helps immensely when you have a confusing probability problem:

> Always write down the probability of everything.
> *(Steve Gull)*

From this joint probability, any desired inference can be mechanically obtained (figure 3.9).

Which doors opened by earthquake	Where the prize is door 1	door 2	door 3
none	$\frac{p_{\rm none}}{3}$	$\frac{p_{\rm none}}{3}$	$\frac{p_{\rm none}}{3}$
1			
2			
3	$\frac{p_3}{3}$	$\frac{p_3}{3}$	$\frac{p_3}{3}$
1,2			
1,3			
2,3			
1,2,3	$\frac{p_{1,2,3}}{3}$	$\frac{p_{1,2,3}}{3}$	$\frac{p_{1,2,3}}{3}$

Figure 3.9. The probability of everything, for the second three-door problem, assuming an earthquake has just occurred. Here, p_3 is the probability that door 3 alone is opened by an earthquake.

Solution to exercise 3.11 (p.58). The statistic quoted by the lawyer indicates the probability that a randomly selected wife-beater will also murder his wife. The probability that the husband was the murderer, *given that the wife has been murdered*, is a completely different quantity.

To deduce the latter, we need to make further assumptions about the probability that the wife is murdered by someone else. If she lives in a neighbourhood with frequent random murders, then this probability is large and the posterior probability that the husband did it (in the absence of other evidence) may not be very large. But in more peaceful regions, it may well be that the most likely person to have murdered you, if you are found murdered, is one of your closest relatives.

Let's work out some illustrative numbers with the help of the statistics on page 58. Let $m=1$ denote the proposition that a woman has been murdered; $h=1$, the proposition that the husband did it; and $b=1$, the proposition that he beat her in the year preceding the murder. The statement 'someone else did it' is denoted by $h=0$. We need to define $P(h \mid m=1)$, $P(b \mid h=1, m=1)$, and $P(b=1 \mid h=0, m=1)$ in order to compute the posterior probability $P(h=1 \mid b=1, m=1)$. From the statistics, we can read out $P(h=1 \mid m=1) = 0.28$. And if two million women out of 100 million are beaten, then $P(b=1 \mid h=0, m=1) = 0.02$. Finally, we need a value for $P(b \mid h=1, m=1)$: if a man murders his wife, how likely is it that this is the first time he laid a finger on her? I expect it's pretty unlikely; so maybe $P(b=1 \mid h=1, m=1)$ is 0.9 or larger.

By Bayes' theorem, then,

$$P(h=1 \mid b=1, m=1) = \frac{.9 \times .28}{.9 \times .28 + .02 \times .72} \simeq 95\%. \tag{3.42}$$

One way to make obvious the sliminess of the lawyer on p.58 is to construct arguments, with the same logical structure as his, that are clearly wrong. For example, the lawyer could say 'Not only was Mrs. S murdered, she was murdered between 4.02pm and 4.03pm. *Statistically*, only one in a *million* wife-beaters actually goes on to murder his wife between 4.02pm and 4.03pm. So the wife-beating is not strong evidence at all. In fact, given the wife-beating evidence alone, it's extremely unlikely that he would murder his wife in this way – only a 1/1,000,000 chance.'

Solution to exercise 3.13 (p.58). There are two hypotheses. $\mathcal{H}_0$: your number is 740511; $\mathcal{H}_1$: it is another number. The data, D, are 'when I dialed 740511, I got a busy signal'. What is the probability of D, given each hypothesis? If your number is 740511, then we expect a busy signal with certainty:

$$P(D \mid \mathcal{H}_0) = 1.$$

On the other hand, if $\mathcal{H}_1$ is true, then the probability that the number dialled returns a busy signal is smaller than 1, since various other outcomes were also possible (a ringing tone, or a number-unobtainable signal, for example). The value of this probability $P(D \mid \mathcal{H}_1)$ will depend on the probability α that a random phone number similar to your own phone number would be a valid phone number, and on the probability β that you get a busy signal when you dial a valid phone number.

I estimate from the size of my phone book that Cambridge has about 75 000 valid phone numbers, all of length six digits. The probability that a random six-digit number is valid is therefore about $75\,000/10^6 = 0.075$. If we exclude numbers beginning with 0, 1, and 9 from the random choice, the probability α is about $75\,000/700\,000 \simeq 0.1$. If we assume that telephone numbers are clustered then a misremembered number might be more likely to be valid than a randomly chosen number; so the probability, α, that our guessed number would be valid, assuming $\mathcal{H}_1$ is true, might be bigger than

0.1. Anyway, α must be somewhere between 0.1 and 1. We can carry forward this uncertainty in the probability and see how much it matters at the end.

The probability β that you get a busy signal when you dial a valid phone number is equal to the fraction of phones you think are in use or off-the-hook when you make your tentative call. This fraction varies from town to town and with the time of day. In Cambridge, during the day, I would guess that about 1% of phones are in use. At 4am, maybe 0.1%, or fewer.

The probability $P(D \mid \mathcal{H}_1)$ is the product of α and β, that is, about $0.1 \times 0.01 = 10^{-3}$. According to our estimates, there's about a one-in-a-thousand chance of getting a busy signal when you dial a random number; or one-in-a-hundred, if valid numbers are strongly clustered; or one-in-10^4, if you dial in the wee hours.

How do the data affect your beliefs about your phone number? The posterior probability ratio is the likelihood ratio times the prior probability ratio:

$$\frac{P(\mathcal{H}_0 \mid D)}{P(\mathcal{H}_1 \mid D)} = \frac{P(D \mid \mathcal{H}_0)}{P(D \mid \mathcal{H}_1)} \frac{P(\mathcal{H}_0)}{P(\mathcal{H}_1)}. \tag{3.43}$$

The likelihood ratio is about 100-to-1 or 1000-to-1, so the posterior probability ratio is swung by a factor of 100 or 1000 in favour of $\mathcal{H}_0$. If the prior probability of $\mathcal{H}_0$ was 0.5 then the posterior probability is

$$P(\mathcal{H}_0 \mid D) = \frac{1}{1 + \frac{P(\mathcal{H}_1 \mid D)}{P(\mathcal{H}_0 \mid D)}} \simeq 0.99 \text{ or } 0.999. \tag{3.44}$$

Solution to exercise 3.15 (p.59). We compare the models $\mathcal{H}_0$ – the coin is fair – and $\mathcal{H}_1$ – the coin is biased, with the prior on its bias set to the uniform distribution $P(p|\mathcal{H}_1) = 1$. [The use of a uniform prior seems reasonable to me, since I know that some coins, such as American pennies, have severe biases when spun on edge; so the situations $p = 0.01$ or $p = 0.1$ or $p = 0.95$ would not surprise me.]

> When I mention $\mathcal{H}_0$ – the coin is fair – a pedant would say, 'how absurd to even consider that the coin is fair – any coin is surely biased to some extent'. And of course I would agree. So will pedants kindly understand $\mathcal{H}_0$ as meaning 'the coin is fair to within one part in a thousand, i.e., $p \in 0.5 \pm 0.001$'.

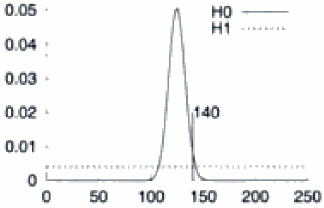


Figure 3.10. The probability distribution of the number of heads given the two hypotheses, that the coin is fair, and that it is biased, with the prior distribution of the bias being uniform. The outcome ($D = 140$ heads) gives weak evidence in favour of $\mathcal{H}_0$, the hypothesis that the coin is fair.

The likelihood ratio is:

$$\frac{P(D|\mathcal{H}_1)}{P(D|\mathcal{H}_0)} = \frac{\frac{140!\,110!}{251!}}{1/2^{250}} = 0.48. \tag{3.45}$$

Thus the data give scarcely any evidence either way; in fact they give weak evidence (two to one) in favour of $\mathcal{H}_0$!

'No, no', objects the believer in bias, 'your silly uniform prior doesn't represent *my* prior beliefs about the bias of biased coins – I was *expecting* only a small bias'. To be as generous as possible to the $\mathcal{H}_1$, let's see how well it could fare if the prior were presciently set. Let us allow a prior of the form

$$P(p|\mathcal{H}_1, \alpha) = \frac{1}{Z(\alpha)} p^{\alpha-1}(1-p)^{\alpha-1}, \quad \text{where } Z(\alpha) = \Gamma(\alpha)^2/\Gamma(2\alpha) \tag{3.46}$$

(a Beta distribution, with the original uniform prior reproduced by setting $\alpha = 1$). By tweaking α, the likelihood ratio for $\mathcal{H}_1$ over $\mathcal{H}_0$,

$$\frac{P(D|\mathcal{H}_1, \alpha)}{P(D|\mathcal{H}_0)} = \frac{\Gamma(140+\alpha)\,\Gamma(110+\alpha)\,\Gamma(2\alpha)2^{250}}{\Gamma(250+2\alpha)\,\Gamma(\alpha)^2}, \tag{3.47}$$

can be increased a little. It is shown for several values of α in figure 3.11. Even the most favourable choice of α ($\alpha \simeq 50$) can yield a likelihood ratio of only two to one in favour of $\mathcal{H}_1$.

In conclusion, the data are not 'very suspicious'. They can be construed as giving at most two-to-one evidence in favour of one or other of the two hypotheses.

Are these wimpy likelihood ratios the fault of over-restrictive priors? Is there any way of producing a 'very suspicious' conclusion? The prior that is best-matched to the data, in terms of likelihood, is the prior that sets p to $f \equiv 140/250$ with probability one. Let's call this model $\mathcal{H}_*$. The likelihood ratio is $P(D|\mathcal{H}_*)/P(D|\mathcal{H}_0) = 2^{250} f^{140}(1-f)^{110} = 6.1$. So the strongest evidence that these data can possibly muster against the hypothesis that there is no bias is six-to-one.

α	$\frac{P(D\|\mathcal{H}_1,\alpha)}{P(D\|\mathcal{H}_0)}$
.37	.25
1.0	.48
2.7	.82
7.4	1.3
20	1.8
55	1.9
148	1.7
403	1.3
1096	1.1

Figure 3.11. Likelihood ratio for various choices of the prior distribution's hyperparameter α.

While we are noticing the absurdly misleading answers that 'sampling theory' statistics produces, such as the p-value of 7% in the exercise we just solved, let's stick the boot in. If we make a tiny change to the data set, increasing the number of heads in 250 tosses from 140 to 141, we find that the p-value goes below the mystical value of 0.05 (the p-value is 0.0497). The sampling theory statistician would happily squeak 'the probability of getting a result as extreme as 141 heads is smaller than 0.05 – we thus reject the null hypothesis at a significance level of 5%'. The correct answer is shown for several values of α in figure 3.12. The values worth highlighting from this table are, first, the likelihood ratio when $\mathcal{H}_1$ uses the standard uniform prior, which is 1:0.61 in favour of the *null hypothesis* $\mathcal{H}_0$. Second, the most favourable choice of α, from the point of view of $\mathcal{H}_1$, can only yield a likelihood ratio of about 2.3:1 in favour of $\mathcal{H}_1$.

Be warned! A p-value of 0.05 is often interpreted as implying that the odds are stacked about twenty-to-one *against* the null hypothesis. But the truth in this case is that the evidence either slightly *favours* the null hypothesis, or disfavours it by at most 2.3 to one, depending on the choice of prior.

The p-values and 'significance levels' of classical statistics should be treated with *extreme caution*. Shun them! Here ends the sermon.

α	$\frac{P(D'\|\mathcal{H}_1,\alpha)}{P(D'\|\mathcal{H}_0)}$
.37	.32
1.0	.61
2.7	1.0
7.4	1.6
20	2.2
55	2.3
148	1.9
403	1.4
1096	1.2

Figure 3.12. Likelihood ratio for various choices of the prior distribution's hyperparameter α, when the data are $D' = 141$ heads in 250 trials.

4

The Source Coding Theorem

4.1 How to measure the information content of a random variable?

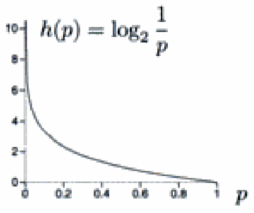
In the next few chapters, we'll be talking about probability distributions and random variables. Most of the time we can get by with sloppy notation, but occasionally, we will need precise notation. Here is the notation that we established in Chapter 2.

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 $\sum_{a_i \in \mathcal{A}_X} P(x\!=\!a_i) = 1$.

How can we measure the information content of an outcome $x = a_i$ from such an ensemble? In this chapter we examine the assertions

1. that the *Shannon information content*,
$$h(x\!=\!a_i) \equiv \log_2 \frac{1}{p_i}, \tag{4.1}$$
is a sensible measure of the information content of the outcome $x = a_i$, and

2. that the *entropy* of the ensemble,
$$H(X) = \sum_i p_i \log_2 \frac{1}{p_i}, \tag{4.2}$$
is a sensible measure of the ensemble's average information content.



p	$h(p)$	$H_2(p)$
0.001	10.0	0.011
0.01	6.6	0.081
0.1	3.3	0.47
0.2	2.3	0.72
0.5	1.0	1.0

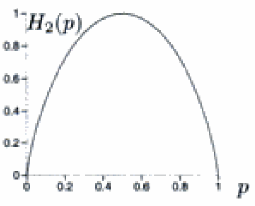


Figure 4.1. The Shannon information content $h(p) = \log_2 \frac{1}{p}$ and the binary entropy function $H_2(p) = H(p, 1\!-\!p) = p \log_2 \frac{1}{p} + (1-p) \log_2 \frac{1}{(1-p)}$ as a function of p.

Figure 4.1 shows the Shannon information content of an outcome with probability p, as a function of p. The less probable an outcome is, the greater its Shannon information content. Figure 4.1 also shows the binary entropy function,

$$H_2(p) = H(p, 1\!-\!p) = p \log_2 \frac{1}{p} + (1-p) \log_2 \frac{1}{(1-p)}, \tag{4.3}$$

which is the entropy of the ensemble X whose alphabet and probability distribution are $\mathcal{A}_X = \{a, b\}, \mathcal{P}_X = \{p, (1-p)\}$.

Information content of independent random variables

Why should $\log 1/p_i$ have anything to do with the information content? Why not some other function of p_i? We'll explore this question in detail shortly, but first, notice a nice property of this particular function $h(x) = \log 1/p(x)$.

Imagine learning the value of two *independent* random variables, x and y. The definition of independence is that the probability distribution is separable into a *product*:

$$P(x,y) = P(x)P(y). \tag{4.4}$$

Intuitively, we might want any measure of the 'amount of information gained' to have the property of *additivity* – that is, for independent random variables x and y, the information gained when we learn x and y should equal the sum of the information gained if x alone were learned and the information gained if y alone were learned.

The Shannon information content of the outcome x, y is

$$h(x,y) = \log \frac{1}{P(x,y)} = \log \frac{1}{P(x)P(y)} = \log \frac{1}{P(x)} + \log \frac{1}{P(y)} \tag{4.5}$$

so it does indeed satisfy

$$h(x,y) = h(x) + h(y), \text{ if } x \text{ and } y \text{ are independent.} \tag{4.6}$$

Exercise 4.2.[*1*, p.86] Show that, if x and y are independent, the entropy of the outcome x, y satisfies

$$H(X,Y) = H(X) + H(Y). \tag{4.7}$$

In words, entropy is additive for independent variables.

We now explore these ideas with some examples; then, in section 4.4 and in Chapters 5 and 6, we prove that the Shannon information content and the entropy are related to the number of bits needed to describe the outcome of an experiment.

The weighing problem: designing informative experiments

Have you solved the weighing problem (exercise 4.1, p.66) yet? Are you sure? Notice that in three uses of the balance – which reads either 'left heavier', 'right heavier', or 'balanced' – the number of conceivable outcomes is $3^3 = 27$, whereas the number of possible states of the world is 24: the odd ball could be any of twelve balls, and it could be heavy or light. So in principle, the problem might be solvable in three weighings – but not in two, since $3^2 < 24$.

If you know how you can determine the odd weight *and* whether it is heavy or light in *three* weighings, then you may read on. If you haven't found a strategy that always gets there in three weighings, I encourage you to think about exercise 4.1 some more.

Why is your strategy optimal? What is it about your series of weighings that allows useful information to be gained as quickly as possible? The answer is that at each step of an optimal procedure, the three outcomes ('left heavier', 'right heavier', and 'balance') are *as close as possible to equiprobable.* An optimal solution is shown in figure 4.2.

Suboptimal strategies, such as weighing balls 1–6 against 7–12 on the first step, do not achieve all outcomes with equal probability: these two sets of balls can never balance, so the only possible outcomes are 'left heavy' and 'right heavy'. Such a binary outcome rules out only half of the possible hypotheses,

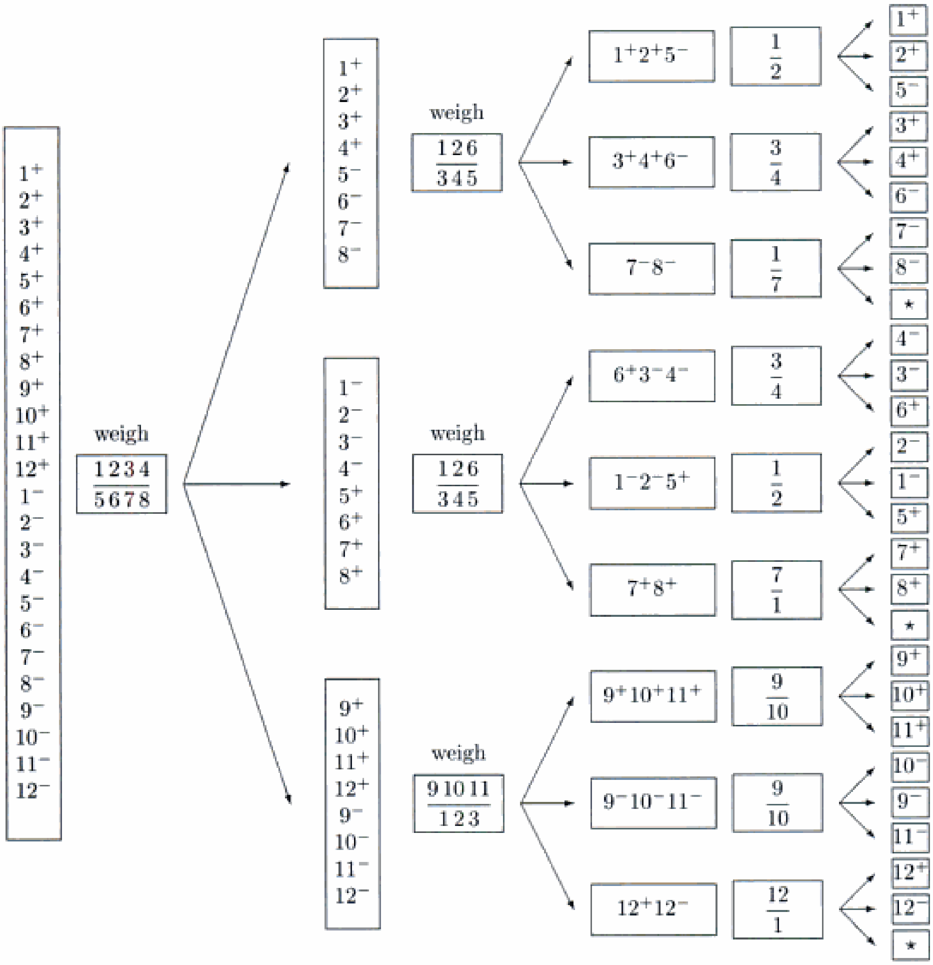


Figure 4.2. An optimal solution to the weighing problem. At each step there are two boxes: the left box shows which hypotheses are still possible; the right box shows the balls involved in the next weighing. The 24 hypotheses are written $1^+, \ldots, 12^-$, with, e.g., 1^+ denoting that 1 is the odd ball and it is heavy. Weighings are written by listing the names of the balls on the two pans, separated by a line; for example, in the first weighing, balls 1, 2, 3, and 4 are put on the left-hand side and 5, 6, 7, and 8 on the right. In each triplet of arrows the upper arrow leads to the situation when the left side is heavier, the middle arrow to the situation when the right side is heavier, and the lower arrow to the situation when the outcome is balanced. The three points labelled ⋆ correspond to impossible outcomes.

so a strategy that uses such outcomes must sometimes take longer to find the right answer.

The insight that the outcomes should be as near as possible to equiprobable makes it easier to search for an optimal strategy. The first weighing must divide the 24 possible hypotheses into three groups of eight. Then the second weighing must be chosen so that there is a 3:3:2 split of the hypotheses.

Thus we might conclude:

> the outcome of a random experiment is guaranteed to be most informative if the probability distribution over outcomes is uniform.

This conclusion agrees with the property of the entropy that you proved when you solved exercise 2.25 (p.37): the entropy of an ensemble X is biggest if all the outcomes have equal probability $p_i = 1/|\mathcal{A}_X|$.

Guessing games

In the game of twenty questions, one player thinks of an object, and the other player attempts to guess what the object is by asking questions that have yes/no answers, for example, 'is it alive?', or 'is it human?' The aim is to identify the object with as few questions as possible. What is the best strategy for playing this game? For simplicity, imagine that we are playing the rather dull version of twenty questions called 'sixty-three'.

Example 4.3. The game 'sixty-three'. What's the smallest number of yes/no questions needed to identify an integer x between 0 and 63?

Intuitively, the best questions successively divide the 64 possibilities into equal sized sets. Six questions suffice. One reasonable strategy asks the following questions:

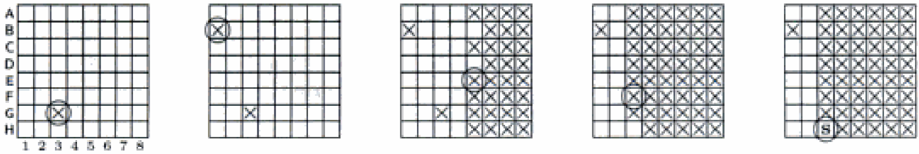
1: is $x \geq 32$?
2: is $x \bmod 32 \geq 16$?
3: is $x \bmod 16 \geq 8$?
4: is $x \bmod 8 \geq 4$?
5: is $x \bmod 4 \geq 2$?
6: is $x \bmod 2 = 1$?

[The notation $x \bmod 32$, pronounced 'x modulo 32', denotes the remainder when x is divided by 32; for example, $35 \bmod 32 = 3$ and $32 \bmod 32 = 0$.]

The answers to these questions, if translated from {yes, no} to {1, 0}, give the binary expansion of x, for example 35 ⇒ 100011. □

What are the Shannon information contents of the outcomes in this example? If we assume that all values of x are equally likely, then the answers to the questions are independent and each has Shannon information content $\log_2(1/0.5) = 1$ bit; the total Shannon information gained is always six bits. Furthermore, the number x that we learn from these questions is a six-bit binary number. Our questioning strategy defines a way of encoding the random variable x as a binary file.

So far, the Shannon information content makes sense: it measures the length of a binary file that encodes x. However, we have not yet studied ensembles where the outcomes have unequal probabilities. Does the Shannon information content make sense there too?



move #	1	2	32	48	49
question	G3	B1	E5	F3	H3
outcome	$x = \mathtt{n}$	$x = \mathtt{n}$	$x = \mathtt{n}$	$x = \mathtt{n}$	$x = \mathtt{y}$
$P(x)$	$\frac{63}{64}$	$\frac{62}{63}$	$\frac{32}{33}$	$\frac{16}{17}$	$\frac{1}{16}$
$h(x)$	0.0227	0.0230	0.0443	0.0874	4.0
Total info.	0.0227	0.0458	1.0	2.0	6.0

Figure 4.3. A game of `submarine`. The submarine is hit on the 49th attempt.

The game of submarine: how many bits can one bit convey?

In the game of battleships, each player hides a fleet of ships in a sea represented by a square grid. On each turn, one player attempts to hit the other's ships by firing at one square in the opponent's sea. The response to a selected square such as 'G3' is either 'miss', 'hit', or 'hit and destroyed'.

In a boring version of battleships called **`submarine`**, each player hides just one submarine in one square of an eight-by-eight grid. Figure 4.3 shows a few pictures of this game in progress: the circle represents the square that is being fired at, and the ×s show squares in which the outcome was a miss, $x = \mathtt{n}$; the submarine is hit (outcome $x = \mathtt{y}$ shown by the symbol $\mathtt{s}$) on the 49th attempt.

Each shot made by a player defines an ensemble. The two possible outcomes are $\{\mathtt{y}, \mathtt{n}\}$, corresponding to a hit and a miss, and their probabilities depend on the state of the board. At the beginning, $P(\mathtt{y}) = 1/64$ and $P(\mathtt{n}) = 63/64$. At the second shot, if the first shot missed, $P(\mathtt{y}) = 1/63$ and $P(\mathtt{n}) = 62/63$. At the third shot, if the first two shots missed, $P(\mathtt{y}) = 1/62$ and $P(\mathtt{n}) = 61/62$.

The Shannon information gained from an outcome x is $h(x) = \log(1/P(x))$. If we are lucky, and hit the submarine on the first shot, then

$$h(x) = h_{(1)}(\mathtt{y}) = \log_2 64 = 6\,\text{bits}. \tag{4.8}$$

Now, it might seem a little strange that one binary outcome can convey six bits. But we have learnt the hiding place, which could have been any of 64 squares; so we have, by one lucky binary question, indeed learnt six bits.

What if the first shot misses? The Shannon information that we gain from this outcome is

$$h(x) = h_{(1)}(\mathtt{n}) = \log_2 \frac{64}{63} = 0.0227\,\text{bits}. \tag{4.9}$$

Does this make sense? It is not so obvious. Let's keep going. If our second shot also misses, the Shannon information content of the second outcome is

$$h_{(2)}(\mathtt{n}) = \log_2 \frac{63}{62} = 0.0230\,\text{bits}. \tag{4.10}$$

If we miss thirty-two times (firing at a new square each time), the total Shannon information gained is

$$\begin{aligned} &\log_2 \frac{64}{63} + \log_2 \frac{63}{62} + \cdots + \log_2 \frac{33}{32} \\ &= 0.0227 + 0.0230 + \cdots + 0.0430 \;=\; 1.0\,\text{bits}. \end{aligned} \tag{4.11}$$

length of each name would be $\log_2 |\mathcal{A}_X|$ bits, if $|\mathcal{A}_X|$ happened to be a power of 2. We thus make the following definition.

The raw bit content of X is

$$H_0(X) = \log_2 |\mathcal{A}_X|. \tag{4.15}$$

$H_0(X)$ is a lower bound for the number of binary questions that are always guaranteed to identify an outcome from the ensemble X. It is an additive quantity: the raw bit content of an ordered pair x, y, having $|\mathcal{A}_X||\mathcal{A}_Y|$ possible outcomes, satisfies

$$H_0(X, Y) = H_0(X) + H_0(Y). \tag{4.16}$$

This measure of information content does not include any probabilistic element, and the encoding rule it corresponds to does not 'compress' the source data, it simply maps each outcome to a constant-length binary string.

Exercise 4.5.[2, p.86] Could there be a compressor that maps an outcome x to a binary code $c(x)$, and a decompressor that maps c back to x, such that *every possible outcome* is compressed into a binary code of length *shorter* than $H_0(X)$ bits?

Even though a simple counting argument shows that it is impossible to make a reversible compression program that reduces the size of *all* files, amateur compression enthusiasts frequently announce that they have invented a program that can do this – indeed that they can further compress compressed files by putting them through their compressor several times. Stranger yet, patents have been granted to these modern-day alchemists. See the `comp.compression` frequently asked questions for further reading.[1]

There are only two ways in which a 'compressor' can actually compress files:

1. A *lossy* compressor compresses some files, but maps some files to the *same* encoding. We'll assume that the user requires perfect recovery of the source file, so the occurrence of one of these confusable files leads to a failure (though in applications such as image compression, lossy compression is viewed as satisfactory). We'll denote by δ the probability that the source string is one of the confusable files, so a lossy compressor has a probability δ of failure. If δ can be made very small then a lossy compressor may be practically useful.

2. A *lossless* compressor maps all files to different encodings; if it shortens some files, it necessarily *makes others longer*. We try to design the compressor so that the probability that a file is lengthened is very small, and the probability that it is shortened is large.

In this chapter we discuss a simple lossy compressor. In subsequent chapters we discuss lossless compression methods.

▶ 4.3 Information content defined in terms of lossy compression

Whichever type of compressor we construct, we need somehow to take into account the *probabilities* of the different outcomes. Imagine comparing the information contents of two text files – one in which all 128 ASCII characters

[1] `http://sunsite.org.uk/public/usenet/news-faqs/comp.compression/`

are used with equal probability, and one in which the characters are used with their frequencies in English text. Can we define a measure of information content that distinguishes between these two files? Intuitively, the latter file contains less information per character because it is more predictable.

One simple way to use our knowledge that some symbols have a smaller probability is to imagine recoding the observations into a smaller alphabet – thus losing the ability to encode some of the more improbable symbols – and then measuring the raw bit content of the new alphabet. For example, we might take a risk when compressing English text, guessing that the most infrequent characters won't occur, and make a reduced ASCII code that omits the characters { !, @, #, %, ^, *, ~, <, >, /, \, _, {, }, [,], | }, thereby reducing the size of the alphabet by seventeen. The larger the risk we are willing to take, the smaller our final alphabet becomes.

We introduce a parameter δ that describes the risk we are taking when using this compression method: δ is the probability that there will be no name for an outcome x.

Example 4.6. Let

$$\begin{aligned} \mathcal{A}_X &= \{\, \texttt{a}, \texttt{b}, \texttt{c}, \texttt{d}, \texttt{e}, \texttt{f}, \texttt{g}, \texttt{h} \,\}, \\ \text{and} \quad \mathcal{P}_X &= \{\, \tfrac{1}{4}, \tfrac{1}{4}, \tfrac{1}{4}, \tfrac{3}{16}, \tfrac{1}{64}, \tfrac{1}{64}, \tfrac{1}{64}, \tfrac{1}{64} \,\}. \end{aligned} \tag{4.17}$$

The raw bit content of this ensemble is 3 bits, corresponding to 8 binary names. But notice that $P(x \in \{\texttt{a},\texttt{b},\texttt{c},\texttt{d}\}) = 15/16$. So if we are willing to run a risk of $\delta = 1/16$ of not having a name for x, then we can get by with four names – half as many names as are needed if every $x \in \mathcal{A}_X$ has a name.

Table 4.5 shows binary names that could be given to the different outcomes in the cases $\delta = 0$ and $\delta = 1/16$. When $\delta = 0$ we need 3 bits to encode the outcome; when $\delta = 1/16$ we need only 2 bits.

$\delta = 0$		$\delta = 1/16$	
x	$c(x)$	x	$c(x)$
a	000	a	00
b	001	b	01
c	010	c	10
d	011	d	11
e	100	e	−
f	101	f	−
g	110	g	−
h	111	h	−

Table 4.5. Binary names for the outcomes, for two failure probabilities δ.

Let us now formalize this idea. To make a compression strategy with risk δ, we make the smallest possible subset S_δ such that the probability that x is not in S_δ is less than or equal to δ, i.e., $P(x \notin S_\delta) \leq \delta$. For each value of δ we can then define a new measure of information content – the log of the size of this smallest subset S_δ. [In ensembles in which several elements have the same probability, there may be several smallest subsets that contain different elements, but all that matters is their sizes (which are equal), so we will not dwell on this ambiguity.]

The smallest δ-sufficient subset S_δ is the smallest subset of $\mathcal{A}_X$ satisfying

$$P(x \in S_\delta) \geq 1 - \delta. \tag{4.18}$$

The subset S_δ can be constructed by ranking the elements of $\mathcal{A}_X$ in order of decreasing probability and adding successive elements starting from the most probable elements until the total probability is $\geq (1-\delta)$.

We can make a data compression code by assigning a binary name to each element of the smallest sufficient subset. This compression scheme motivates the following measure of information content:

The essential bit content of X is:

$$H_\delta(X) = \log_2 |S_\delta|. \tag{4.19}$$

Note that $H_0(X)$ is the special case of $H_\delta(X)$ with $\delta = 0$ (if $P(x) > 0$ for all $x \in \mathcal{A}_X$). [Caution: do not confuse $H_0(X)$ and $H_\delta(X)$ with the function $H_2(p)$ displayed in figure 4.1.]

Figure 4.6 shows $H_\delta(X)$ for the ensemble of example 4.6 as a function of δ.

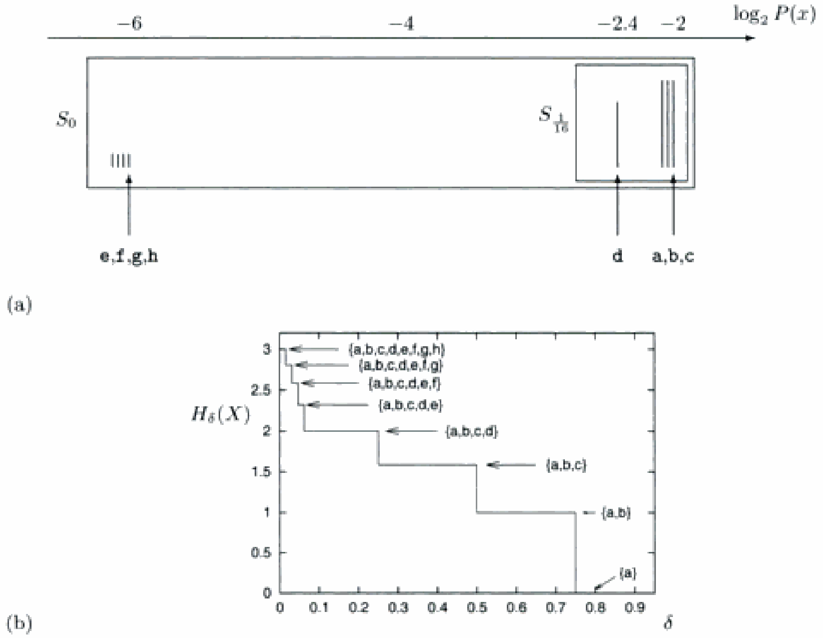


Figure 4.6. (a) The outcomes of X (from example 4.6 (p.75)), ranked by their probability. (b) The essential bit content $H_\delta(X)$. The labels on the graph show the smallest sufficient set as a function of δ. Note $H_0(X) = 3$ bits and $H_{1/16}(X) = 2$ bits.

Extended ensembles

Is this compression method any more useful if we compress *blocks* of symbols from a source?

We now turn to examples where the outcome $\mathbf{x} = (x_1, x_2, \ldots, x_N)$ is a string of N independent identically distributed random variables from a single ensemble X. We will denote by X^N the ensemble $(X_1, X_2, \ldots, X_N)$. Remember that entropy is additive for independent variables (exercise 4.2 (p.68)), so $H(X^N) = NH(X)$.

Example 4.7. Consider a string of N flips of a bent coin, $\mathbf{x} = (x_1, x_2, \ldots, x_N)$, where $x_n \in \{\mathtt{0}, \mathtt{1}\}$, with probabilities $p_0 = 0.9$, $p_1 = 0.1$. The most probable strings $\mathbf{x}$ are those with most $\mathtt{0}$s. If $r(\mathbf{x})$ is the number of $\mathtt{1}$s in $\mathbf{x}$ then

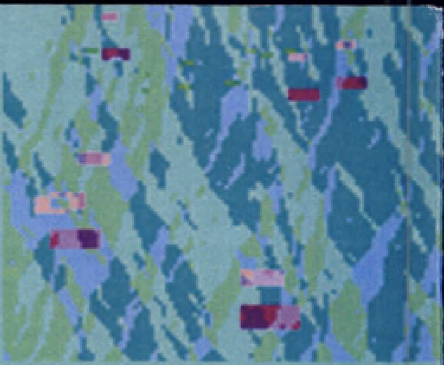
$$P(\mathbf{x}) = p_0^{N-r(\mathbf{x})} p_1^{r(\mathbf{x})}. \tag{4.20}$$

To evaluate $H_\delta(X^N)$ we must find the smallest sufficient subset S_δ. This subset will contain all $\mathbf{x}$ with $r(\mathbf{x}) = 0, 1, 2, \ldots,$ up to some $r_{\max}(\delta) - 1$, and some of the $\mathbf{x}$ with $r(\mathbf{x}) = r_{\max}(\delta)$. Figures 4.7 and 4.8 show graphs of $H_\delta(X^N)$ against δ for the cases $N = 4$ and $N = 10$. The steps are the values of δ at which $|S_\delta|$ changes by 1, and the cusps where the slope of the staircase changes are the points where $r_{\max}$ changes by 1.

Exercise 4.8.[2, p.86] What are the mathematical shapes of the curves between the cusps?

For the examples shown in figures 4.6–4.8, $H_\delta(X^N)$ depends strongly on the value of δ, so it might not seem a fundamental or useful definition of information content. But we will consider what happens as N, the number of independent variables in X^N, increases. We will find the remarkable result that $H_\delta(X^N)$ becomes almost independent of δ – and for all δ it is very close to $NH(X)$, where $H(X)$ is the entropy of one of the random variables.

Figure 4.9 illustrates this asymptotic tendency for the binary ensemble of example 4.7. As N increases, $\frac{1}{N}H_\delta(X^N)$ becomes an increasingly flat function,



Information theory and inference, often taught separately, are here united in one entertaining textbook. These topics lie at the heart of many exciting areas of contemporary science and engineering – communication, signal processing, data mining, machine learning, pattern recognition, computational neuroscience, bioinformatics, and cryptography.

This textbook introduces theory in tandem with applications. Information theory is taught alongside practical communication systems, such as arithmetic coding for data compression and sparse-graph codes for error-correction. A toolbox of inference techniques, including message-passing algorithms, Monte Carlo methods, and variational approximations, are developed alongside applications of these tools to clustering, convolutional codes, independent component analysis, and neural networks.

The final part of the book describes the state of the art in error-correcting codes, including low-density parity-check codes, turbo codes, and digital fountain codes – the twenty-first century standards for satellite communications, disk drives, and data broadcast.

Richly illustrated, filled with worked examples and over 400 exercises, some with detailed solutions, David MacKay's groundbreaking book is ideal for self-learning and for undergraduate or graduate courses. Interludes on crosswords, evolution, and sex provide entertainment along the way.

In sum, this is a textbook on information, communication, and coding for a new generation of students, and an unparalleled entry point into these subjects for professionals in areas as diverse as computational biology, financial engineering, and machine learning.

'This is an extraordinary and important book, generous with insight and rich with detail in statistics, information theory, and probabilistic modeling across a wide swathe of standard, creatively original, and delightfully quirky topics. David MacKay is an uncompromisingly lucid thinker, from whom students, faculty and practitioners all can learn.'

Peter Dayan and Zoubin Ghahramani,
Gatsby Computational Neuroscience Unit, University College, London

'An utterly original book that shows the connections between such disparate fields as information theory and coding, inference, and statistical physics.'

Dave Forney,
Massachusetts Institute of Technology

'An instant classic, covering everything from Shannon's fundamental theorems to the postmodern theory of LDPC codes. You'll want two copies of this astonishing book, one for the office and one for the fireside at home.'

Bob McEliece,
California Institute of Technology

CAMBRIDGE
UNIVERSITY PRESS
www.cambridge.org

ISBN 0-521-64298-1

9 780521 642989