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.

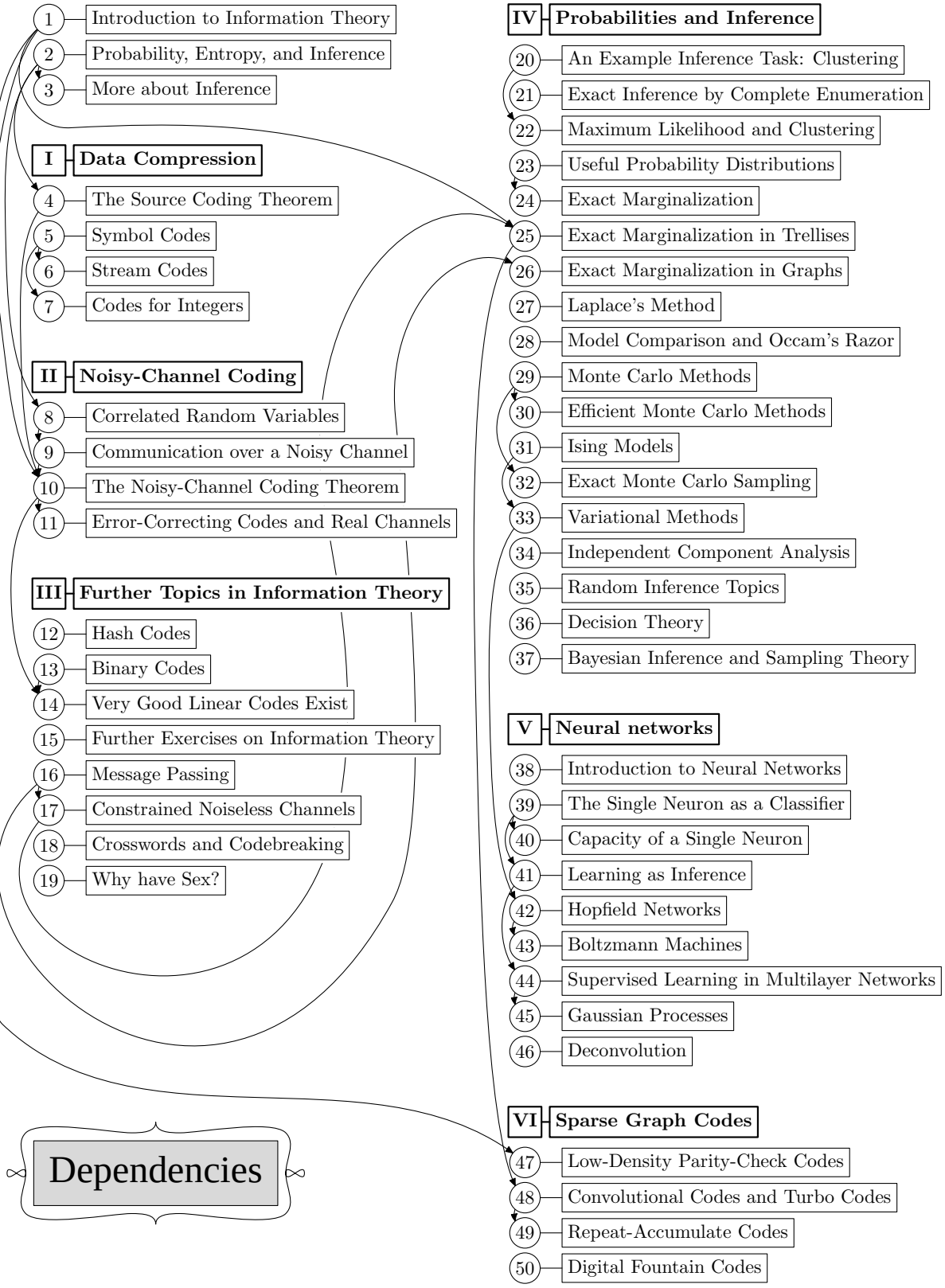
How to use this book

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.

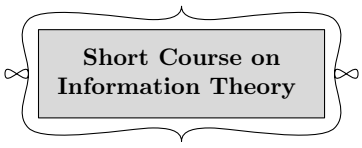
Within Parts I, II, IV, and V of this book, chapters on advanced or optional topics are towards the end. All chapters of Part III are optional on a first reading, except perhaps for Chapter 16 (Message Passing).

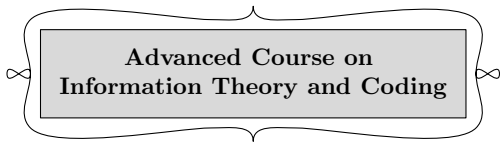
The same system sometimes applies within a chapter: the final sections often deal with advanced topics that can be skipped on a first reading. For example in two key chapters – Chapter 4 (The Source Coding Theorem) and Chapter 10 (The Noisy-Channel Coding Theorem) – the first-time reader should detour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



My Cambridge Course on, Information Theory, Pattern Recognition, and Neural Networks





A Course on Bayesian Inference and Machine Learning

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

In this second printing, a small number of typographical errors were corrected, and 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.

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.600.

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.

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

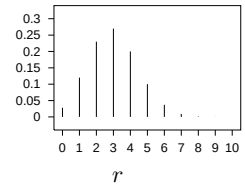


Figure 1.1. The binomial distribution $P(r \mid f = 0.3,\ N = 10)$.

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

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

$$\mathrm{var}[r] \equiv \mathcal{E}\left[(r - \mathcal{E}[r])^2\right] \tag{1.3}$$

$$= \mathcal{E}[r^2] - (\mathcal{E}[r])^2 = \sum_{r=0}^{N} P(r \mid f, N) r^2 - (\mathcal{E}[r])^2\,. \tag{1.4}$$

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:

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

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,

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

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:

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

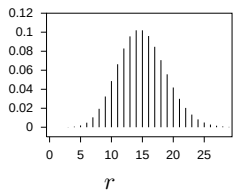


Figure 1.2. The Poisson distribution $P(r \mid \lambda = 15)$.

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

$$e^{-\lambda} \frac{\lambda^\lambda}{\lambda!} \simeq \frac{1}{\sqrt{2\pi\lambda}} \tag{1.10}$$

$$\Rightarrow \lambda! \simeq \lambda^\lambda e^{-\lambda} \sqrt{2\pi\lambda}. \tag{1.11}$$

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,

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

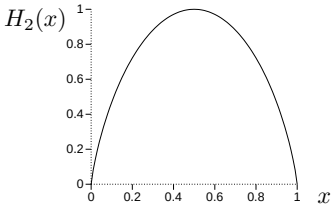


Figure 1.3. The binary entropy function.

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

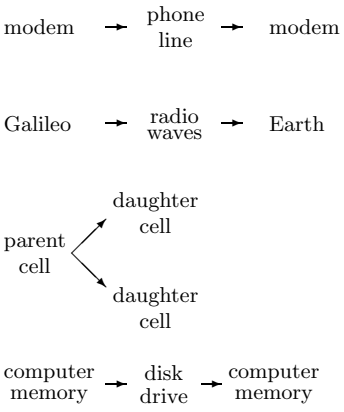
> The fundamental problem of is that of reproducing at one point either exactly or approximately a message selected at another point.
>
> *(Claude Shannon, 1948)*

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;
- the radio communication link from Galileo, the Jupiter-orbiting spacecraft, to earth;
- reproducing cells, in which the daughter cells's DNA contains information from the parent cells;
- a disk drive.

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

transmitted message. We would prefer to have a communication channel for which this probability was zero – or so close to zero that for practical purposes it is indistinguishable from zero.

Let's consider a noisy disk drive that transmits each bit correctly with probability $(1-f)$ and incorrectly with probability f. This model communication channel is known as the *binary symmetric channel* (figure 1.4).

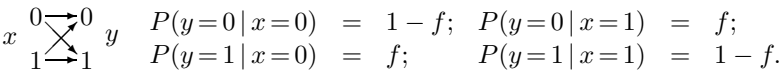


Figure 1.4. The binary symmetric channel. The transmitted symbol is x and the received symbol y. The noise level, the probability that a bit is flipped, is f.

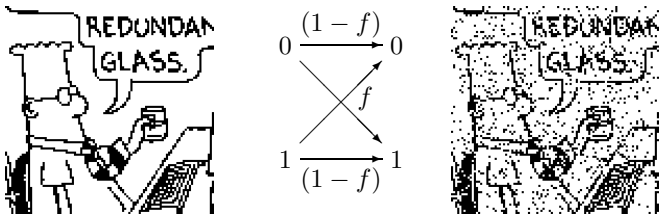


Figure 1.5. A binary data sequence of length 10 000 transmitted over a binary symmetric channel with noise level $f = 0.1$. [Dilbert image Copyright©1997 United Feature Syndicate, Inc., used with permission.]

As an example, let's imagine that $f = 0.1$, that is, ten per cent of the bits are flipped (figure 1.5). A useful disk drive would flip no bits at all in its entire lifetime. If we expect to read and write a gigabyte per day for ten years, we require a bit error probability of the order of 10^{-15}, or smaller. There are two approaches to this goal.

The physical solution

The physical solution is to improve the physical characteristics of the communication channel to reduce its error probability. We could improve our disk drive by

1. using more reliable components in its circuitry;
2. evacuating the air from the disk enclosure so as to eliminate the turbulence that perturbs the reading head from the track;
3. using a larger magnetic patch to represent each bit; or
4. using higher-power signals or cooling the circuitry in order to reduce thermal noise.

These physical modifications typically increase the cost of the communication channel.

The 'system' solution

Information theory and coding theory offer an alternative (and much more exciting) approach: we accept the given noisy channel as it is and add communication *systems* to it so that we can detect and correct the errors introduced by the channel. As shown in figure 1.6, we add an *encoder* before the channel and a *decoder* after it. The encoder encodes the source message $\mathbf{s}$ into a *transmitted* message $\mathbf{t}$, adding *redundancy* to the original message in some way. The channel adds noise to the transmitted message, yielding a received message $\mathbf{r}$. The decoder uses the known redundancy introduced by the encoding system to infer both the original signal $\mathbf{s}$ and the added noise.

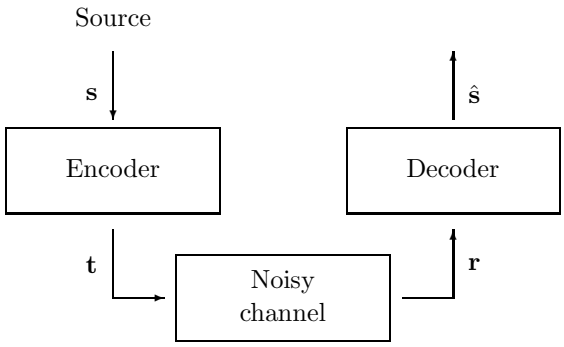


Figure 1.6. The 'system' solution for achieving reliable communication over a noisy channel. The encoding system introduces systematic redundancy into the transmitted vector **t**. The decoding system uses this known redundancy to deduce from the received vector **r** *both* the original source vector *and* the noise introduced by the channel.

Whereas physical solutions give incremental channel improvements only at an ever-increasing cost, system solutions can turn noisy channels into reliable communication channels with the only cost being a *computational* requirement at the encoder and decoder.

Information theory is concerned with the theoretical limitations and potentials of such systems. 'What is the best error-correcting performance we could achieve?'

Coding theory is concerned with the creation of practical encoding and decoding systems.

1.2 Error-correcting codes for the binary symmetric channel

We now consider examples of encoding and decoding systems. What is the simplest way to add useful redundancy to a transmission? [To make the rules of the game clear: we want to be able to detect *and* correct errors; and retransmission is not an option. We get only one chance to encode, transmit, and decode.]

Repetition codes

A straightforward idea is to repeat every bit of the message a prearranged number of times – for example, three times, as shown in table 1.7. We call this *repetition code* 'R_3'.

Source sequence **s**	Transmitted sequence **t**
0	000
1	111

Table 1.7. The repetition code R_3.

Imagine that we transmit the source message

$$\mathbf{s} = 0\ 0\ 1\ 0\ 1\ 1\ 0$$

over a binary symmetric channel with noise level $f = 0.1$ using this repetition code. We can describe the channel as 'adding' a sparse noise vector **n** to the transmitted vector – adding in modulo 2 arithmetic, i.e., the binary algebra in which 1+1=0. A possible noise vector **n** and received vector $\mathbf{r} = \mathbf{t} + \mathbf{n}$ are shown in figure 1.8.

s	0	0	1	0	1	1	0
t	000	000	111	000	111	111	000
n	000	001	000	000	101	000	000
r	000	001	111	000	010	111	000

Figure 1.8. An example transmission using R_3.

How should we decode this received vector? The optimal algorithm looks at the received bits three at a time and takes a majority vote (algorithm 1.9).

Received sequence $\mathbf{r}$	Likelihood ratio $\frac{P(\mathbf{r}\,\vert\,s=1)}{P(\mathbf{r}\,\vert\,s=0)}$	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 $\mathbf{r} = r_1 r_2 r_3$. By Bayes' theorem, the *posterior probability* of s is

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

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

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

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

This posterior probability is determined by two factors: the *prior probability* $P(s)$, and the data-dependent term $P(r_1r_2r_3\,\vert\,s)$, 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 $P(s=0\,\vert\,\mathbf{r}) > P(s=1\,\vert\,\mathbf{r})$, and $\hat{s}=1$ otherwise.

To find $P(s=0\,\vert\,\mathbf{r})$ and $P(s=1\,\vert\,\mathbf{r})$, 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 $P(s\,\vert\,\mathbf{r})$ is equivalent to maximizing the likelihood $P(\mathbf{r}\,\vert\,s)$. And we assume that the channel is a binary symmetric channel with noise level $f < 0.5$, so that the likelihood is

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

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

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

Thus the likelihood ratio for the two hypotheses is

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

each factor $\frac{P(r_n\vert t_n(1))}{P(r_n\vert t_n(0))}$ 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 0s 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.

s	0	0	1	0	1	1	0
t	000	000	111	000	111	111	000
n	000	001	000	000	101	000	000
r	000	001	111	000	010	111	000
$\hat{\mathbf{s}}$	0	0	1	0	0	1	0
corrected errors		⋆					
undetected errors					⋆		

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.

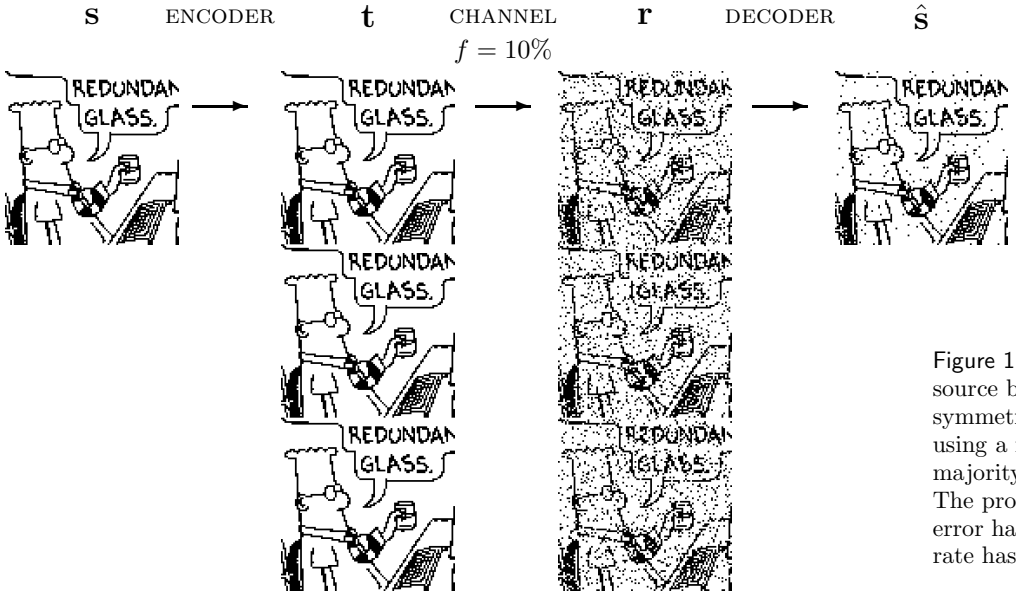


Figure 1.11. Transmitting 10 000 source bits over a binary symmetric channel with $f = 10\%$ using a repetition code and the majority vote decoding algorithm. The probability of decoded bit error has fallen to about 3%; the rate has fallen to 1/3.

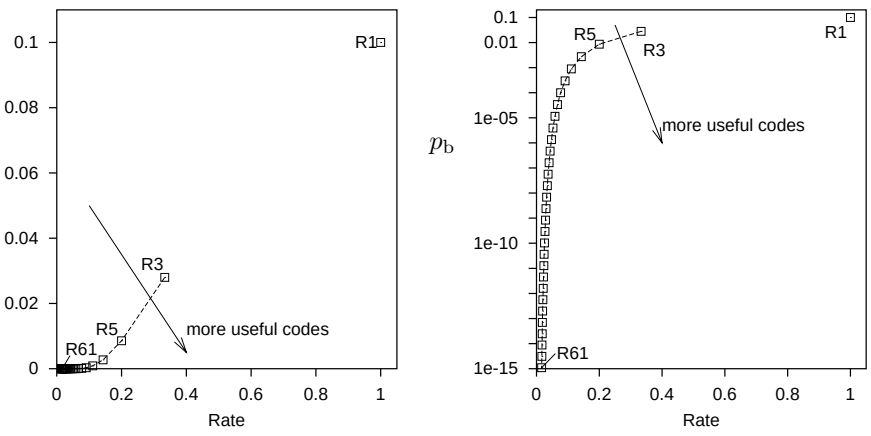


Figure 1.12. Error probability p_b versus rate for repetition codes over a binary symmetric channel with $f = 0.1$. The right-hand figure shows p_b on a logarithmic scale. We would like the rate to be large and p_b to be small.

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.

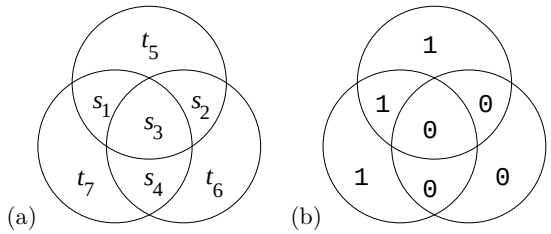


Figure 1.13. Pictorial representation of encoding for the $(7, 4)$ Hamming code.

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 $\mathbf{t} = \mathtt{1000101}$ and the noise flips the second bit, so the received vector is $\mathbf{r} = \mathtt{1000101} \oplus \mathtt{0100000} = \mathtt{1100101}$. 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 $\mathbf{z} = (\mathtt{1}, \mathtt{1}, \mathtt{0})$, because the first two circles are 'unhappy' (parity $\mathtt{1}$) and the third circle is 'happy' (parity $\mathtt{0}$).]

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.

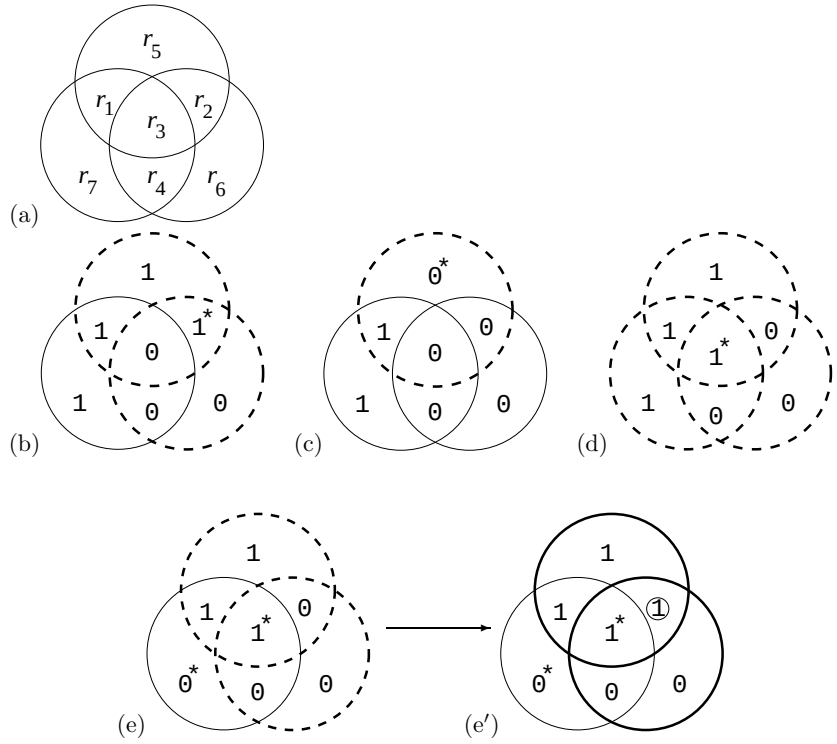


Figure 1.15. Pictorial representation of decoding of the Hamming $(7, 4)$ code. The received vector is written into the diagram as shown in (a). In (b,c,d,e), the received vector is shown, assuming that the transmitted vector was as in figure 1.13b and the bits labelled by $\star$ were flipped. The violated parity checks are highlighted by dashed circles. One of the seven bits is the most probable suspect to account for each 'syndrome', i.e., each pattern of violated and satisfied parity checks.
In examples (b), (c), and (d), the most probable suspect is the one bit that was flipped.
In example (e), two bits have been flipped, s_3 and t_7. The most probable suspect is r_2, marked by a circle in (e′), which shows the output of the decoding algorithm.

Syndrome **z**	000	001	010	011	100	101	110	111
Unflip this bit	*none*	r_7	r_6	r_4	r_5	r_1	r_2	r_3

Algorithm 1.16. Actions taken by the optimal decoder for the $(7, 4)$ Hamming code, assuming a binary symmetric channel with small noise level f. The syndrome vector **z** lists whether each parity check is violated (1) or satisfied (0), going through the checks in the order of the bits r_5, r_6, and r_7.

If you try flipping any one of the seven bits, you'll find that a different syndrome is obtained in each case – seven non-zero syndromes, one for each bit. There is only one other syndrome, the all-zero syndrome. So if the channel is a binary symmetric channel with a small noise level f, the optimal decoder unflips at most one bit, depending on the syndrome, as shown in algorithm 1.16. Each syndrome could have been caused by other noise patterns too, but any other noise pattern that has the same syndrome must be less probable because it involves a larger number of noise events.

What happens if the noise actually flips more than one bit? Figure 1.15e shows the situation when two bits, r_3 and r_7, are received flipped. The syndrome, 110, makes us suspect the single bit r_2; so our optimal decoding algorithm flips this bit, giving a decoded pattern with three errors as shown in figure 1.15e′. If we use the optimal decoding algorithm, any two-bit error pattern will lead to a decoded seven-bit vector that contains three errors.

General view of decoding for linear codes: syndrome decoding

We can also describe the decoding problem for a linear code in terms of matrices. The first four received bits, $r_1r_2r_3r_4$, purport to be the four source bits; and the received bits $r_5r_6r_7$ purport to be the parities of the source bits, as defined by the generator matrix $\mathbf{G}$. We evaluate the three parity-check bits for the received bits, $r_1r_2r_3r_4$, and see whether they match the three received bits, $r_5r_6r_7$. The differences (modulo 2) between these two triplets are called the *syndrome* of the received vector. If the syndrome is zero – if all three parity checks are happy – then the received vector is a codeword, and the most probable decoding is

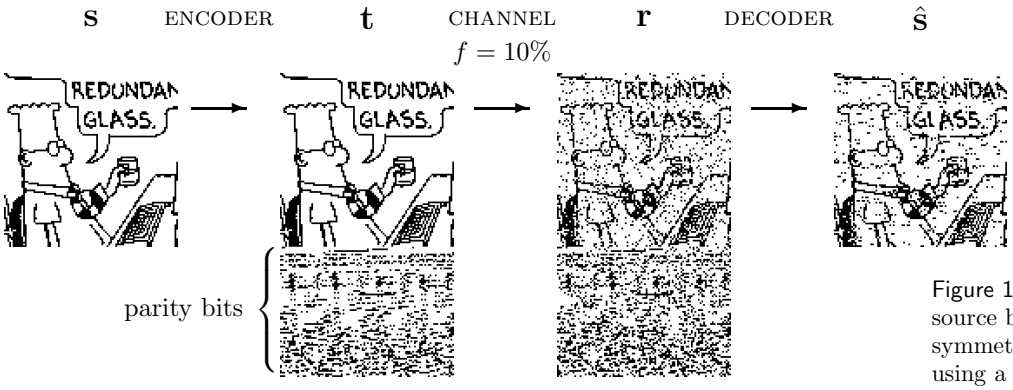


Figure 1.17. Transmitting 10 000 source bits over a binary symmetric channel with $f = 10\%$ using a $(7, 4)$ Hamming code. The probability of decoded bit error is about 7%.

given by reading out its first four bits. If the syndrome is non-zero, then the noise sequence for this block was non-zero, and the syndrome is our pointer to the most probable error pattern.

The computation of the syndrome vector is a linear operation. If we define the 3×4 matrix $\mathbf{P}$ such that the matrix of equation (1.26) is

$$\mathbf{G}^{\mathsf{T}} = \begin{bmatrix} \mathbf{I}_4 \\ \mathbf{P} \end{bmatrix}, \tag{1.29}$$

where $\mathbf{I}_4$ is the 4×4 identity matrix, then the syndrome vector is $\mathbf{z} = \mathbf{H}\mathbf{r}$, where the *parity-check matrix* $\mathbf{H}$ is given by $\mathbf{H} = \begin{bmatrix} -\mathbf{P} & \mathbf{I}_3 \end{bmatrix}$; in modulo 2 arithmetic, $-1 \equiv 1$, so

$$\mathbf{H} = \begin{bmatrix} \mathbf{P} & \mathbf{I}_3 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 1 \end{bmatrix}. \tag{1.30}$$

All the codewords $\mathbf{t} = \mathbf{G}^{\mathsf{T}}\mathbf{s}$ of the code satisfy

$$\mathbf{H}\mathbf{t} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}. \tag{1.31}$$

▷ Exercise 1.4.[1] Prove that this is so by evaluating the 3×4 matrix $\mathbf{H}\mathbf{G}^{\mathsf{T}}$.

Since the received vector $\mathbf{r}$ is given by $\mathbf{r} = \mathbf{G}^{\mathsf{T}}\mathbf{s} + \mathbf{n}$, the syndrome-decoding problem is to find the most probable noise vector $\mathbf{n}$ satisfying the equation

$$\mathbf{H}\mathbf{n} = \mathbf{z}. \tag{1.32}$$

A decoding algorithm that solves this problem is called a *maximum-likelihood decoder*. We will discuss decoding problems like this in later chapters.

Summary of the $(7, 4)$ *Hamming code's properties*

Every possible received vector of length 7 bits is either a codeword, or it's one flip away from a codeword.

Since there are three parity constraints, each of which might or might not be violated, there are $2 \times 2 \times 2 = 8$ distinct syndromes. They can be divided into seven non-zero syndromes – one for each of the one-bit error patterns – and the all-zero syndrome, corresponding to the zero-noise case.

The optimal decoder takes no action if the syndrome is zero, otherwise it uses this mapping of non-zero syndromes onto one-bit error patterns to unflip 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) $\mathbf{r} = \mathtt{1101011}$

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

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

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

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.

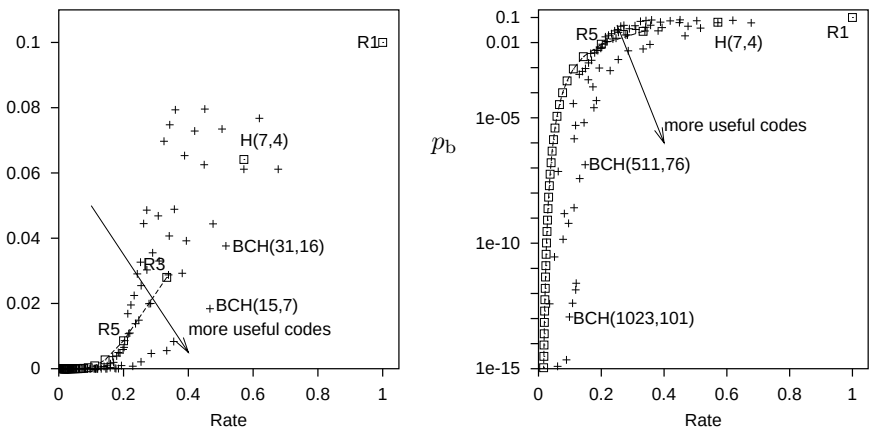


Figure 1.18. Error probability p_b versus rate R for repetition codes, the $(7,4)$ Hamming code and BCH codes with block lengths up to 1023 over a binary symmetric channel with $f = 0.1$. The righthand figure shows p_b on a logarithmic scale.

Exercise 1.9.[4, 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, 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, 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

*

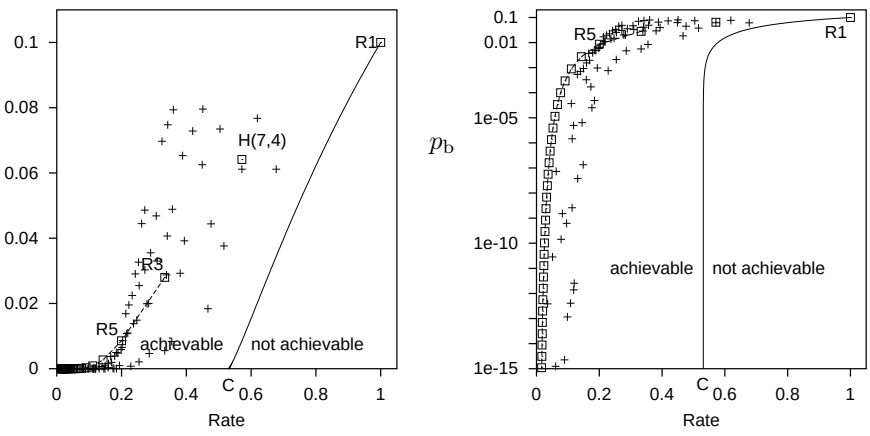


Figure 1.19. Shannon's noisy-channel coding theorem. The solid curve shows the Shannon limit on achievable values of (R, p_b) for the binary symmetric channel with $f = 0.1$. Rates up to $R = C$ are achievable with arbitrarily small p_b. The points show the performance of some textbook codes, as in figure 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 $P(r=2 \mid f, N=3)$.]

$$p_{\rm b} = p_{\rm B} = 3f^2(1-f) + f^3 = 3f^2 - 2f^3. \tag{1.36}$$

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 of $\lceil N/2 \rceil$ bits' being 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

$$p_{\rm b} = p_{\rm B} \simeq 2^N (f(1-f))^{N/2} = (4f(1-f))^{N/2}. \tag{1.39}$$

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 $N/2$.