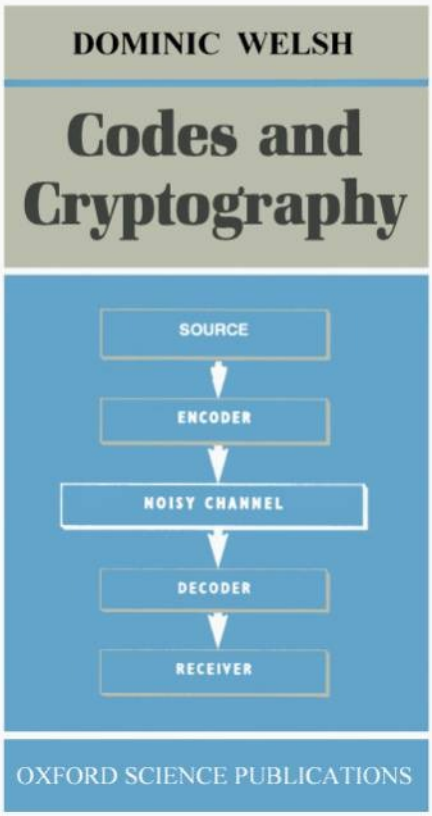
DOMINIC WELSH
Codes and Cryptography
SOURCE
ENCODER
NOISY CHANNEL
DECODER
RECEIVER
OXFORD SCIENCE PUBLICATIONS

Codes and Cryptography

Dominic Welsh
Merton College and the Mathematical Institute,
University of Oxford

CLARENDON PRESS · OXFORD

Oxford University Press, Walton Street, Oxford OX2 6DP

Oxford New York Toronto
Delhi Bombay Calcutta Madras Karachi
Petaling Jaya Singapore Hong Kong Tokyo
Nairobi Dar es Salaam Cape Town
Melbourne Auckland

and associated companies in
Berlin Ibadan

Oxford is a trade mark of Oxford University Press

Published in the United States
by Oxford University Press, New York

First published 1988
Reprinted (with corrections) 1989

British Library Cataloguing in Publication Data

Welsh, Dominic
Codes and cryptography.
1. Cryptography
I. Title
001.54'36 Z103
ISBN 0-19-853288-1
ISBN 0-19-853287-3 Pbk

Library of Congress Cataloging in Publication Data

Welsh, Dominic
Codes and cryptography.
Bibliography: p.
Includes index.
1. Ciphers. 2. Cryptography. I. Title.
Z103.W46 1988 652'.8 87-31354
ISBN 0-19-853288-1
ISBN 0-19-853287-3 (pbk.)

Typeset and Printed by The Universities Press (Belfast) Ltd

Preface

This text is based on a course I have given to undergraduates at Oxford on the mathematics of Communication Theory. Its aim is to introduce the subject in as short a space as possible to students with no previous knowledge of the area.

The foundations of this book are the seminal papers by Claude Shannon on information theory and secrecy systems. The concept of entropy is fundamental to the three main problems of how to encode information (a) economically, (b) reliably, and (c) so as to preserve privacy. However, whereas in coding theory the object is to resurrect a message from an accidentally noisy environment, in cryptography the noise is deliberately superimposed so as to make it difficult for an enemy to recover information contained in the message.

Throughout the text the emphasis is on explaining the main ideas rather than proving results in complete generality. For example, most students seem to find Shannon's noisy coding theorem difficult and I have dealt only with the simplest case of binary alphabets, memoryless channels and sources. In a similar vein Chapters 4 and 9 are crash courses in algebraic coding theory and computational complexity respectively. Both of these are huge areas, rich in exciting problems, and I hope that this treatment will at least enable the reader to make a more informed choice of future options.

The mathematical prerequisites have been kept to a minimum. However, since it is impossible to understand information theory without a working knowledge of very basic probability, this has been taken for granted. Knowledge of elementary modern algebra is also assumed.

The exercises at the end of each section in the book are meant to be elementary and are to be used as a check on the understanding of the preceding principles. The problems at the end of each chapter tend to be harder. Occasionally I have used the device of giving as a problem a result I regard as interesting, together with an original reference; some of these are difficult and to be regarded more as sources of information.

The numbering system used is fairly standard. For example,

Theorem 3.4.1 refers to the first theorem of the fourth section of Chapter 3; when the reference is in the current chapter, this is shortened to Theorem 4.1.

Finally, it gives me great pleasure to thank B. J. Birch, T. F. R. G. Braun, P. J. Cameron, A. Chin, D. A. Cohen, M. J. Collins, P. J. Donnelly, D. R. Heath-Brown, F. C. Piper, J. F. Pratt, D. R. Stirzaker, and I. White for their constructive comments and suggestions about various points arising in the text. I owe a special debt to Keith Edwards, Colin McDiarmid, and Kenneth Regan, who used earlier versions of different parts of this text when giving lectures and classes associated with this course. Their suggestions have been particularly helpful. I should also like to acknowledge the co-operation and technical advice of the staff of Oxford University Press and thank Brenda Willoughby of the Mathematical Institute for her cheerful and accurate typing of what was often hieroglyphic handwriting.

Above all, I am grateful for the help given by my wife Bridget. Her criticisms and suggestions throughout the time of writing have had a very great influence on the final outcome.

D. J. A. W.

Oxford
December 1987

Contents

Notation

A few of the frequently used items which may not be familiar to all readers are listed below.

$\lceil x \rceil$	least integer greater than or equal to x
$\lfloor x \rfloor$	maximum integer not greater than x
$\ln(x)$	$\log_e x$
$\log(x)$	$\log_2 x$ (This is a departure from usual mathematical usage but is a widespread convention in this area.)
$\mathbb{Z}$	integers
$\mathbb{Z}^+$	non-negative integers
$\mathbb{Z}_n$	non-negative integers less than n
Z_n^*	positive integers less than and coprime with n
V_n	set of binary n-tuples $(x_1, \ldots, x_n)$, $x_i = 0$ or 1

Note We often write an ordered n-tuple $(x_1, \ldots, x_n)$ as $x_1x_2 \cdots x_n$ particularly when the x_i are specified numeric entries.

$P(A)$	probability of the event A
$P(A \mid B)$	probability of the event A conditional on the event B
$O(n^k)$	a function f such that for all sufficiently large n, $\lvert f(n) \rvert \leqslant cn^k$ for some constant c

Alphabets and strings

Σ will denote an *alphabet* consisting of a finite set of *symbols* or *letters*. An ordered sequence of symbols from Σ is called a *string* or *word*. If $x = x_1x_2 \cdots x_m$ is such a string its *length* is defined to be m and is denoted by

$$|x| = |x_1x_2 \cdots x_m| = m.$$

If $x = x_1 \cdots x_m$ and $y = y_1 \cdots y_n$ are two strings their *concatenation* is the string $x_1 \cdots x_m y_1 \cdots y_n$.

$\Sigma^{(n)}$	the collection of strings from Σ which have length n
Σ^*	the collection of all finite strings from Σ

1

Entropy = uncertainty = information

1.1 Uncertainty

Consider the following propositions.

[A] A race between two equally matched horses is less uncertain than a race between six evenly matched horses.

[B] The outcome of a spin on a roulette wheel is more uncertain than the throw of a die.

[C] The throw of a fair die is more uncertain than the throw of a biased die in which the probabilities are $\frac{1}{10}$ of getting each of the numbers 1 to 5 and probability $\frac{1}{2}$ of getting a 6.

I suspect (indeed hope) that most readers will agree with each of the propositions [A], [B], and [C]. At the same time, I suspect they will find it difficult to give a formal definition of what they mean by uncertainty. One of Shannon's many achievements was to formalize this abstract idea. Suppose that X and Y are distinct random variables but that

$$P(X=0)=p, \qquad P(X=1)=1-p,$$

while

$$P(Y=100)=p, \qquad P(Y=200)=1-p,$$

where $0<p<1$.

We assert that any definition of uncertainty should give X and Y the same uncertainty. In other words, the uncertainty of X (and Y) should be a function *only* of the probability p. This property of the uncertainty must extend to random variables taking more than two values, and accordingly our first demand of any proposed measure of uncertainty is:

> The uncertainty of a random variable Z, which takes the values a_i with probabilities p_i $(1 \le i \le n)$, is to be a function *only* of the probabilities $p_1, \ldots, p_n$.

Therefore we denote such a function as $H(p_1, \ldots, p_n)$ and demand

of it the following properties which in view of the preceding discussion and examples we regard as minimal requirements.

(A1) $H(p_1, \ldots, p_n)$ is a maximum when $p_1 = p_2 = \ldots = p_n = 1/n$.

(A2) For any permutation π of $(1, 2, \ldots, n)$, we have

$$H(p_1, \ldots, p_n) = H(p_{\pi(1)}, \ldots, p_{\pi(n)}).$$

In other words H must be a symmetric function of the arguments $p_1, \ldots, p_n$.
[Only the probabilities matter, not their order.]

(A3) $H(p_1, \ldots, p_n) \geq 0$ and equals zero only when one of the p_i equals 1.
[Uncertainty is an inherently positive quantity and is zero only when there is no randomness present.]

(A4) $$H(p_1, \ldots, p_n, 0) = H(p_1, \ldots, p_n).$$

[The uncertainty of a six-sided fair die is the same as a seven-sided die that has no chance of showing 7 but is otherwise fair.]

(A5) $$H\left(\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n}\right) \leq H\left(\frac{1}{n+1}, \frac{1}{n+1}, \ldots, \frac{1}{n+1}\right).$$

[A two-horse race is less uncertain than a three-horse race.]

(A6) $H(p_1, \ldots, p_n)$ should be a continuous function of its arguments.
[This is very natural; a small change in the probabilities should not drastically affect the uncertainty.]

(A7) If m and n are positive integers then

$$H\left(\frac{1}{mn}, \frac{1}{mn}, \ldots, \frac{1}{mn}\right) = H\left(\frac{1}{m}, \ldots, \frac{1}{m}\right) + H\left(\frac{1}{n}, \ldots, \frac{1}{n}\right).$$

[This is the linearity condition which essentially says that the uncertainty involved in throwing an m-sided die followed by an n-sided die should be the sum of the individual uncertainties.]

Finally we lay down our last condition. It is not so intuitively obvious and needs a little thought.

(A8) Let $p = p_1 + \ldots + p_m$ and $q = q_1 + \ldots + q_n$ where each p_i and q_j are non-negative; then, if p and q are positive, with $p + q = 1$, we must have

$$H(p_1, \ldots, p_m, q_1, \ldots, q_n) = H(p, q) + pH(p_1/p, \ldots, p_m/p) + qH(q_1/q, \ldots, q_n/q).$$

[Think of a race in which there are m black horses and n grey horses, with p_i the probability the ith black horse wins. The total uncertainty in the outcome is the uncertainty associated with a grey or black winner plus the weighted sum of the uncertainties given that the winner is respectively grey or black.]

From these assumptions we can prove:

Theorem 1 *Let $H(p_1, \ldots, p_n)$ be a function defined for any integer n and for all values of $p_1, \ldots, p_n$ satisfying $p_i \geq 0$ and*

$$\sum_{i=1}^{n} p_i = 1.$$

If H is to satisfy the axioms (A1)–(A8), *then*

$$H(p_1, p_2, \ldots, p_n) = -\lambda \sum_k p_k \log p_k, \tag{1}$$

with λ any positive constant and where the sum is over those k for which $p_k > 0$.

Although Theorem 1 is a beautiful and motivating result its proof is not an integral part of this course and is deferred to Appendix I. It does however add plausibility to the following definitions.

Given a random variable X that takes a finite set of values with probabilities $p_1, p_2, \ldots, p_n$, we define the *uncertainty* or *entropy* of X to be

$$\mathrm{H}(X) = -\sum_k p_k \log_2 p_k \tag{2}$$

where we are taking logarithms to the base 2 (for historical reasons) and where again the sum is over those k with $p_k > 0$.

Note 1 The sum on the right-hand side of (2) is going to occur frequently in what follows and, in order to avoid repeating a caveat about the probabilities p_k being strictly positive, we will henceforth *assume* that, in any sum of this type, the probabilities are nonzero.

Note 2 The conditions (A1)–(A8) are essentially the axioms for entropy proposed by Shannon (1948). They are not minimal and there exists an extensive literature on this subject; see Aczél and Daróczy (1975).

Exercises 1.1

1. Which race has greater uncertainty: a handicap in which there are seven runners, three having probability $\frac{1}{6}$ of winning and four having probability $\frac{1}{8}$ or a selling plate in which there are eight runners with two horses having $\frac{1}{4}$ probability of winning and six horses each having $\frac{1}{12}$ probability?

2. Verify that the entropy function defined by (2) satisfies the conditions (A1)–(A8).

1.2 Entropy and its properties

We have agreed that, for any random variable X taking only a finite number of values with probabilities $p_1, \ldots, p_n$ such that

$$\sum p_i = 1 \quad \text{and} \quad p_i > 0 \quad (1 \leq i \leq n),$$

we define the entropy of X by

$$\mathrm{H}(X) = -\sum_{k=1}^{n} p_k \log p_k.$$

In an exactly analogous way, if $\boldsymbol{X}$ is a random vector which takes only a finite number of values $\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_m$ (say), we define its *entropy* by

(1) $$\mathrm{H}(\boldsymbol{X}) = -\sum_{k=1}^{m} p(\boldsymbol{u}_k) \log p(\boldsymbol{u}_k).$$

For example, when $\boldsymbol{X}$ is a two-dimensional random vector, say $\boldsymbol{X} = (U, V)$ with

$$p_{ij} = P(U = a_i, V = b_j),$$

then we often write

$$\mathrm{H}(\boldsymbol{X}) = \mathrm{H}(U, V) = -\sum_{i,j} p_{ij} \log p_{ij}.$$

More generally, if $X_1, X_2, \ldots, X_m$ is any collection of random variables each taking only a finite set of values, then we may regard $\boldsymbol{X} = (X_1, \ldots, X_m)$ as a random vector taking only a finite set of values and define the *joint entropy* of $X_1, \ldots, X_m$ by

(2) $$\mathrm{H}(X_1, \ldots, X_m) = \mathrm{H}(\boldsymbol{X})$$

$$= -\sum p(x_1, \ldots, x_m) \log p(x_1, \ldots, x_m)$$

where $p(x_1, \ldots, x_m) = P(X_1 = x_1, X_2 = x_2, \ldots, X_m = x_m)$. It is easy to prove that:

(3) $\mathrm{H}(\boldsymbol{X}) = 0$ if and only if $\boldsymbol{X}$ is a constant.

An upper bound for H is given by the following:

Theorem 1 *For any n,*

$$H(p_1, \ldots, p_m) \leq \log_2 n$$

with equality if and only if $p_1 = p_2 = \ldots = p_n = n^{-1}$.

Proof Write

$$\log_2 x = \log_2 e \log_e x.$$

Since the logarithm is a convex function (that is its curve always lies below its tangent) we have

$$\log_e x \leq x - 1$$

with equality if and only if $x = 1$. Hence, if $(q_1, \ldots, q_n)$ is any probability vector, then

$$\log_e(q_k/p_k) \leq (q_k/p_k) - 1,$$

with equality if and only if $q_k = p_k$. Hence,

$$\sum p_i \log_e(q_i/p_i) \leq \sum q_k - \sum p_k = 0,$$

which implies

$$\sum p_i \log q_i \leq \sum p_i \log p_i.$$

Putting $q_i = 1/n$ gives

$$H(p_1, \ldots, p_n) = -\sum p_i \log_2 p_i \leq \log_2 n$$

with equality as stated. □

In the above proof we have proved a very useful inequality, which we name and state as follows.

Key Lemma *If $(p_i : 1 \leq i \leq n)$ is a given probability distribution, then the minimum of*

$$G(q_1, \ldots, q_n) = -\sum p_i \log q_i$$

over all probability distributions $(q_1, \ldots, q_n)$ is achieved when $q_k = p_k$ $(1 \leq k \leq n)$.

Note Here and elsewhere we mean by a probability distribution

$(p_1, \ldots, p_n)$, any set of nonnegative numbers p_i such that

$$\sum p_i = 1.$$

Using this lemma, it is straightforward to prove the following.

Theorem 2 *If X and Y are any two random variables taking only finitely many values, then*

$$\mathrm{H}(X, Y) \leq \mathrm{H}(X) + \mathrm{H}(Y),$$

with equality holding only when X and Y are independent.

Proof Suppose that

$$r_i = P(X = a_i) \quad (1 \leq i \leq m), \qquad s_j = P(Y = b_j) \quad (1 \leq j \leq n),$$
$$t_{ij} = P(X = a_i, Y = b_j) \quad (1 \leq i \leq m, \ 1 \leq j \leq n).$$

Then

$$\mathrm{H}(X) + \mathrm{H}(Y) = -\left(\sum_i r_i \log r_i + \sum_j s_j \log s_j\right)$$
$$= -\left(\sum_i \sum_j t_{ij} \log r_i + \sum_j \sum_i t_{ij} \log s_j\right)$$

because

$$r_i = \sum_j t_{ij}, \qquad s_j = \sum_i t_{ij}.$$

Hence

$$\mathrm{H}(X) + \mathrm{H}(Y) = -\sum_i \sum_j t_{ij} \log (r_i s_j)$$
$$\geq -\sum \sum t_{ij} \log t_{ij} = \mathrm{H}(X, Y),$$

by the preceding Key Lemma. Equality will hold only when

$$r_i s_j = t_{ij} \quad (1 \leq i \leq m, \ 1 \leq j \leq n).$$

But this is exactly the condition that X and Y are independent. □

By a straightforward extension of this method, one can prove:

(4) $$\mathrm{H}(X_1, \ldots, X_n) \leq \mathrm{H}(X_1) + \ldots + \mathrm{H}(X_n),$$

with equality holding only when $X_1, \ldots, X_n$ are mutually independent;

(5) $$H(\boldsymbol{U}, \boldsymbol{V}) \leq H(\boldsymbol{U}) + H(\boldsymbol{V})$$

for any pair of random vectors $(\boldsymbol{U}, \boldsymbol{V})$ with equality holding only when $\boldsymbol{U}$ and $\boldsymbol{V}$ are independent random vectors.

The proofs (which follow in exactly the same way as does that of Theorem 2 from the Key Lemma) are left to the reader.

Exercises 1.2

1. Two fair dice are thrown; X denotes the value obtained by the first, and Y denotes the value obtained by the second. Verify that $H(X, Y) = H(X) + H(Y)$. Show that, if $Z = X + Y$, then
$$H(Z) < H(X, Y).$$
2. Show that, for any random variable X,
$$H(X, X^2) = H(X).$$
3. Show that, for any sequence of random variables $(X_i : 1 \leq i < \infty)$,
$$H(X_1, \ldots, X_n) \leq H(X_1, \ldots, X_{n+1}).$$

1.3 Conditional entropy

Suppose that X is a random variable on a probability space Ω and A is an event in Ω. If X takes a finite set of values $\{a_i : 1 \leq i \leq m\}$, it is natural to define the *conditional entropy* of X given A by

$$H(X \mid A) = -\sum_{k=1}^{m} P(X = a_k \mid A) \log P(X = a_k \mid A).$$

In the same way, if Y is any other random variable taking values b_k $(1 \leq k \leq m)$, we define the *conditional entropy* of X given Y by

$$H(X \mid Y) = \sum_j H(X \mid Y = b_j) P(Y = b_j).$$

We think of $H(X \mid Y)$ as the uncertainty of X given a particular value of Y, averaged over the range of values that Y can take.

Fairly trivial consequences of the definitions are:

(1) $$H(X \mid X) = 0,$$

(2) $$H(X \mid Y) = H(X) \quad \text{if } X \text{ and } Y \text{ are independent.}$$

Example Let X be the value obtained by throwing a fair die. Let Y be another random variable determined by the same experiment with Y equalling 1 if the value thrown is odd and 0 otherwise. Since the die is fair

$$\mathrm{H}(X) = \log 6,$$
$$\mathrm{H}(Y) = \log 2,$$

and

$$\mathrm{H}(X \mid Y) = \log 3. \qquad \square$$

When $\boldsymbol{U}$ and $\boldsymbol{V}$ are random vectors, we naturally extend the definition of conditional entropy by defining

(3) $$\mathrm{H}(\boldsymbol{U} \mid \boldsymbol{V}) = \sum_j \mathrm{H}(\boldsymbol{U} \mid \boldsymbol{V} = \boldsymbol{v}_j) P(\boldsymbol{V} = \boldsymbol{v}_j),$$

where the sum, as usual, is over the (finite) range of values $\boldsymbol{v}_j$ that $\boldsymbol{V}$ has a positive probability of taking.

As a first example of the way in which $\mathrm{H}(\boldsymbol{U} \mid \boldsymbol{V})$ measures the uncertainty about $\boldsymbol{U}$ contained in $\boldsymbol{V}$ we prove:

(4) $\mathrm{H}(\boldsymbol{U} \mid \boldsymbol{V}) = 0$ if and only if $\boldsymbol{U} = g(\boldsymbol{V})$ for some function g.

Proof The right-hand side of (3) is the sum of a finite number of nonnegative quantities. Hence, for it to be zero, we need $\mathrm{H}(\boldsymbol{U} \mid \boldsymbol{V} = \boldsymbol{v}_j)$ to be zero for each j. But again each of these nonnegative quantities is zero only if U is uniquely determined by V. $\square$

Slightly more care gives the following result, which expresses mathematically the idea that our definition of conditional entropy of X given Y correctly measures the remaining uncertainty.

Theorem 1 *For any pair of random variables X and Y that take only a finite set of values,*

$$\mathrm{H}(X, Y) = \mathrm{H}(Y) + \mathrm{H}(X \mid Y).$$

Proof Without loss of generality, we suppose X and Y take integer values and, where necessary, we let $p_{ij} = P(X = i, Y = j)$. Now

$$\begin{aligned}
\mathrm{H}(X, Y) &= -\sum_i \sum_j P(X = i, Y = j) \log P(X = i, Y = j) \\
&= -\sum_i \sum_j P(X = i, Y = j) \log P(X = i \mid Y = j) P(Y = j) \\
&= -\sum \sum p_{ij} \log P(X = i \mid Y = j) - \sum \sum p_{ij} \log P(Y = j)
\end{aligned}$$

$$= -\sum\sum P(X=i \mid Y=j)P(Y=j)\log P(X=i \mid Y=j) + \mathrm{H}(Y)$$

$$= -\sum_j P(Y=j)\sum_i P(X=i \mid Y=j)\log P(X=i \mid Y=j) + \mathrm{H}(Y)$$

$$= \sum_j P(Y=j)\mathrm{H}(X \mid Y=j) + \mathrm{H}(Y)$$

$$= \mathrm{H}(Y) + \mathrm{H}(X \mid Y) \quad \text{as required.}$$ □

The method of proving the above theorem, (essentially definition chasing) carries over to the case of several variables. More precisely, we can prove the following.

Theorem 2 *If $\boldsymbol{U}$ and $\boldsymbol{V}$ random vectors each taking only a finite set of values, then*

$$\mathrm{H}(\boldsymbol{U}, \boldsymbol{V}) = \mathrm{H}(\boldsymbol{V}) + \mathrm{H}(\boldsymbol{U} \mid \mathbf{V}).$$

Proof Follow through the proof of Theorem 1, but instead of X and Y taking integer values i and j, we have $\boldsymbol{U}$ and $\boldsymbol{V}$ taking values $\boldsymbol{u}_i$ and $\boldsymbol{v}_j$, where $\boldsymbol{u}_i$ and $\boldsymbol{v}_j$ are prescribed vectors. □

The following result is an immediate consequence.

Corollary *For any pair of random vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$, $\mathrm{H}(\boldsymbol{X} \mid \boldsymbol{Y}) \leq \mathrm{H}(\boldsymbol{X})$, with equality if and only if $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent.*

Proof

$$\mathrm{H}(\boldsymbol{X} \mid \boldsymbol{Y}) = \mathrm{H}(\boldsymbol{X}, \boldsymbol{Y}) - \mathrm{H}(\boldsymbol{Y})$$

But $\mathrm{H}(\boldsymbol{X}, \boldsymbol{Y}) \leq \mathrm{H}(\boldsymbol{X}) + \mathrm{H}(\boldsymbol{Y})$, with equality if and only if $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent, and the result follows. □

Exercises 1.3

1. Show that, for any random variable X,

$$\mathrm{H}(X^2 \mid X) = 0,$$

but give an example to show that $\mathrm{H}(X \mid X^2)$ is not always zero.

2. The random variable X takes the integer values $1, 2, \ldots, 2N$ with equal probability. The random variable Y is defined by $Y = 0$, if X is even, but $Y = 1$ if X is odd. Show that

$$\mathrm{H}(X \mid Y) = \tfrac{1}{2}\mathrm{H}(X)$$

but that $\mathrm{H}(Y \mid X) = 0$.

1.4 Information

R. V. L. Hartley in 1928 seems to have been the first to attempt to assign a quantitative measure to the concept of information. The rationale behind this attempt can be partly explained as follows.

Suppose E_1 and E_2 are two events in a probability space Ω associated with some experiment and suppose that the function I is to be our measure of information. If E_1 and E_2 have probabilities p_1 and p_2 respectively, then it could be argued that any natural measure of the information content should satisfy

(1) $$I(p_1p_2) = I(p_1) + I(p_2)$$

on the grounds that, for two independent realizations of the experiment, the information that the results of these experiments turned out to be E_1 followed by E_2 should be the sum of the information obtained by carrying out the experiments separately.

Granting that (1) has some validity, and wishing to make our measure of information non-negative and continuous in p, both natural assumptions, we are left with little alternative but to *define* the *information* I of an event E of positive probability by

(2) $$\mathrm{I}(E) = -\log_2 P(E)$$

where we have chosen the base 2 for our logarithms in order to conform with modern conventions. (Hartley originally used logarithms to the base 10.)

Example Suppose we have a source which emits a string of binary digits 0 and 1, each with equal probability and independently for successive digits. Let E be the event that the first n digits are alternately zeros and ones. Then clearly

$$\mathrm{I}(E) = -\log_2(1/2^n) = n$$

and the same applies to any prescribed n-sequences of digits. □

Thus the 'information-theoretic' unit of information, namely the *bit*, corresponds naturally to the use of the word 'bit' to mean a binary digit in present-day computing terminology.

We extend this concept of information to cover random variables and vectors as follows. Suppose $\boldsymbol{U}$ is a random vector taking the values $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m$ with probabilities $p_1, \ldots, p_m$ respectively. Then each of the elementary events $\{\boldsymbol{U} = \boldsymbol{u}_k\}$ $(1 \leq k \leq m)$ has an associated information equal to $-\log_2 p_k$, and we notice that the

entropy of the vector $\boldsymbol{U}$ is given by

$$\mathrm{H}(\boldsymbol{U}) = -\sum p_k \log_2 p_k = \sum p_k \mathrm{I}(\{\boldsymbol{U} = \boldsymbol{u}_k\}),$$

so that $H(\boldsymbol{U})$ has a natural interpretation as the mean value of the information associated with the elementary events determined by $\boldsymbol{U}$.

More generally, if $\boldsymbol{U}$ and $\boldsymbol{V}$ are any two random vectors, we define the *information about* $\boldsymbol{U}$ *conveyed* by $\boldsymbol{V}$ to be the quantity

$$\mathrm{I}(\boldsymbol{U} \mid \boldsymbol{V}) = \mathrm{H}(\boldsymbol{U}) - \mathrm{H}(\boldsymbol{U} \mid \boldsymbol{V}).$$

In other words, $\mathrm{I}(\boldsymbol{U} \mid \boldsymbol{V})$ measures the amount of uncertainty about $\boldsymbol{U}$ that is removed by $\boldsymbol{V}$.

Trivially, we see that

(3) $$\mathrm{I}(\boldsymbol{U} \mid \boldsymbol{U}) = \mathrm{H}(\boldsymbol{U}),$$

(4) $$\mathrm{I}(\boldsymbol{U} \mid \boldsymbol{V}) = 0 \quad \text{if and only if } \boldsymbol{U} \text{ and } \boldsymbol{V} \text{ are independent.}$$

Proof The result follows immediately from the earlier remark that $\mathrm{H}(\boldsymbol{U}) = \mathrm{H}(\boldsymbol{U} \mid \boldsymbol{V})$ only when $\boldsymbol{U}$ and $\boldsymbol{V}$ are independent. □

A somewhat surprising symmetry in I is the following result which, as far as I can see, has no intuitive explanation.

(5)
$$\begin{aligned} \mathrm{I}(\boldsymbol{U} \mid \boldsymbol{V}) &= \mathrm{H}(\boldsymbol{U}) - \mathrm{H}(\boldsymbol{U} \mid \boldsymbol{V}) \\ &= \mathrm{H}(\boldsymbol{U}) - [\mathrm{H}(\boldsymbol{U}, \boldsymbol{V}) - \mathrm{H}(\boldsymbol{V})] \\ &= \mathrm{H}(\boldsymbol{U}) + \mathrm{H}(\boldsymbol{V}) - \mathrm{H}(\boldsymbol{U}, \boldsymbol{V}) \\ &= \mathrm{I}(\boldsymbol{V} \mid \boldsymbol{U}). \end{aligned}$$

Exercises 1.4

1. Which has got the greater information content: a sequence of 10 letters or a sequence of 26 digits from the set $\{0, 1, \ldots, 9\}$? [Assume all sequences are equiprobable.]
2. A fair die is thrown. Show that the information about the value of the die given by the knowledge that it is prime is given by $\log_2 \frac{3}{2}$.

1.5 Conclusion

To sum up, we have shown that, essentially, uncertainty and information are the same quantities, the removal of uncertainty being equated with the giving of information. Both are measured by the mathematical concept of entropy, which is uniquely defined (up to a