ENCYCLOPEDIA OF MATHEMATICS AND ITS APPLICATIONS

The Theory of Information and Coding
Second Edition

R. J. McELIECE

California Institute of Technology

CAMBRIDGE
UNIVERSITY PRESS

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

CAMBRIDGE UNIVERSITY PRESS
The Edinburgh Building, Cambridge CB2 2RU, UK
40 West 20th Street, New York, 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

Published by Addison-Wesley 1977, 1979, 1982

First published by Cambridge University Press 1985

Second edition 2002

Printed in the United Kingdom at the University Press, Cambridge

Typeface Times 10/13pt. *System* 3B2 [KT]

A catalogue record for this book is available from the British Library

ISBN 0 521 00095 5 hardback

Contents

Part two: Coding theory

Appendices

References

1
Entropy and mutual information

1.1 Discrete random variables

Suppose X is a discrete random variable, that is, one whose range $R = \{x_1, x_2, \ldots\}$ is finite or countable. Let $p_i = P\{X = x_i\}$. (For probabilistic terminology consult Appendix A.) The *entropy* of X is defined by

$$H(X) = \sum_{i \geqslant 1} p_i \log \frac{1}{p_i}. \tag{1.1}$$

This definition needs elaboration. First, the base of the logarithm is purposely left unspecified. If necessary, however, we shall denote the base-b entropy by $H_b(X)$, and say that the entropy of X is being measured in base-b units. Base-2 units are called *bits* (binary igits), and base-e units are called *nats* (natural digits). Second, if $p_i = 0$, the term $p_i \log p_i^{-1}$ in (1.1) is indeterminate; we define it to be 0, however. (This convention is by no means arbitrary; see Prob. 1.1.) Finally, if R is infinite the sum (1.1) may not converge; in this case we set $H(X) = +\infty$.

Example 1.1 Let X represent the outcome of a single roll of a fair die. Then $R = \{1, 2, 3, 4, 5, 6\}$ and $p_i = \frac{1}{6}$ for each i. Here $H(X) = \log 6 = 2.58$ bits $= 1.79$ nats. □

Example 1.2 Let $R = \{0, 1\}$, and define X by $P\{X = 0\} = p$, $P\{X = 1\} = 1 - p$. Then $H(X) = -p \log p - (1-p)\log(1-p)$, and so $H_2(X)$, as a function of $0 \leqslant p \leqslant 1$, is identical to the binary entropy function $H_2(p)$, which was defined in Eq. (0.13). In what follows, we will frequently represent the function-$p \log p - (1-p)\log(1-p)$, where the bases of the logarithms are unspecified, by $H(p)$, and call it the *entropy function*. Figure 1.1 gives its graph (cf. Fig. 0.4). More generally, if $\mathbf{p} = (p_1, \ldots, p_r)$ is any *probability*

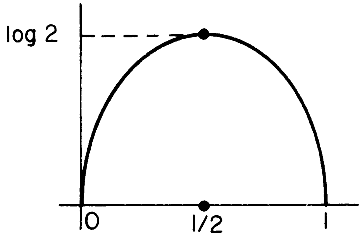


Figure 1.1 The entropy function $H(p)$.

vector, that is, $p_i \geqslant 0$ and $\sum p_i = 1$, we define $H(\mathbf{p}) = H(p_1, p_2, \ldots, p_r) = \sum p_i \log p_i^{-1}$. This notation is not quite consistent, since for $r = 2$ we have $H(p, 1 - p) = H(p)$. (Thus we use the symbol H in three slightly different ways: $H(X)$ is the entropy of the random variable X; $H(p) = -p \log p - (1 - p) \log(1 - p)$ for $0 \leqslant p \leqslant 1$; and $H(p_1, p_2, \ldots, p_r) = \sum p_i \log p_i^{-1}$ if $\mathbf{p}$ is a probability vector.) □

Example 1.3 If the sum $\sum_{n=2}^{\infty}(n \log^2 n)^{-1}$ is denoted by A, and if the random variable X is defined by $P\{X = n\} = (An \log^2 n)^{-1}$ for $n = 2, 3, \ldots,$ then $H(X) = +\infty$. (See Prob. 1.2.) □

It turns out that $H(X)$ can be thought of as a measure of the following things about X:

(a) The amount of "information" provided by an observation of X.
(b) Our "uncertainty" about X.
(c) The "randomness" of X.

In the next few paragraphs we will discuss these properties informally, but the reader should be told immediately that $H(X)$ does in fact measure these things in a deep mathematical sense as well. Indeed there are many possible functions of a random variable X that share the properties to be discussed below, but only $H(X)$ will do for the study of communications problems.

For each $x \in R$ define $I(x) = -\log P\{X = x\}$. Then I is a new random variable, and $H(X)$ is its average. The function $I(x)$ (see Fig. 1.2) can be interpreted as the amount of information provided by the event $\{X = x\}$. According to this interpretation, the less probable an event is, the more information we receive when it occurs. A certain event (one that occurs with probability 1) provides no information, whereas an unlikely event provides a very large amount of information. For example, suppose you visited an oracle

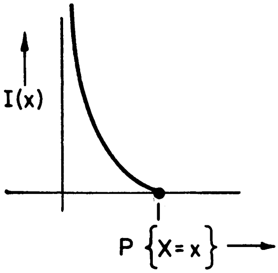


Figure 1.2 The function $I(x)$.

who could answer any "yes or no" question. If you asked, "Will I live to be 125?" and got a "no" answer, you would have gained very little information, since such extreme longevity is exceedingly improbable. Conversely, if you got a "yes," you would have learned much. If now millions of people visited the oracle and asked the same question, most would get a "no," a few would get a "yes," and the average amount of information provided would be $H(p)$, where $p = P\,\{\text{age at death} \geqslant 125\}$. Moreover, just before receiving the oracle's reply you would probably be slightly anxious; this reflects the fact that a small amount of uncertainty exists about the answer. $H(p)$ is equally a measure of this uncertainty.[1] Finally, if a dispassionate census worker were assigned to record the oracle's answers, he would become extremely bored and might begin to suspect the oracle of being a machine that always says "no." This reflects the fact that the random variable X representing the oracle's reply is not very random. Here $H(p)$ measures the randomness of X.

As a less transcendental example, define X by $P\{X = 0\} = P\{X = 1\} = \frac{1}{2}$. Then $I(0) = I(1) = H(X) = \log 2 = 1$ bit, that is, the observation of the "bit" X provides one "bit" of information.

Our first theorem concerns the maximum possible value for $H(X)$ in terms of the size of R.

Theorem 1.1 *Let X assume values in $R = \{x_1, x_2, \ldots, x_r\}$. Then $0 \leqslant H(X) \leqslant \log r$. Furthermore $H(X) = 0$ iff $p_i = 1$ for some i, and $H(X) = \log r$ iff $p_i = 1/r$ for all i.*

Proof Since each p_i is $\leqslant 1$, each term $p_i \log p_i^{-1}$ in (1.1) is $\geqslant 0$, so $H(X) \geqslant 0$. Furthermore $p \log p^{-1} = 0$ iff $p = 0$ or 1, and so $H(X) = 0$ iff each $p_i = 0$ or 1, i.e., one $p_i = 1$ and all the rest are 0.

Now by Jensen's inequality (see Appendix B), since $\log x$ is strictly convex $\cap$,

$$H(X) = \sum_{i=1}^{r} p_i \log \frac{1}{p_i} \leqslant \log \sum_{i=1}^{r} p_i \frac{1}{p_i} = \log r,$$

with equality iff p_i is a constant independent of i, i.e., $p_i = 1/r$ for all i. □

Informally, Theorem 1.1 identifies a uniformly distributed random variable as the most "random" kind of random variable. Formally, it asserts that the maximum value of the function $H(p_1, p_2, \ldots, p_r)$, as $\mathbf{p} = (p_1, \ldots, p_r)$ ranges over the $r - 1$ dimensional simplex $\{p_i \geqslant 0, \sum p_i = 1\}$, is $\log r$ and is achieved uniquely at $\mathbf{p} = (1/r, 1/r, \ldots, 1/r)$.

Our next goal is to define, for a pair of random variables X and Y, a quantity $H(X|Y)$ called the *conditional entropy*[2] of X, given Y. In order to do this neatly, we introduce some streamlined notation. For x in the range of X, y in the range of Y, define:

$$\begin{aligned}
p(x) &= P\{X = x\},\\
p(y) &= P\{Y = y\},\\
p(x, y) &= P\{X = x, Y = y\}, \qquad (1.2)\\
p(x|y) &= P\{X = x|Y = y\} = p(x, y)/p(y),\\
p(y|x) &= P\{Y = y|X = x\} = p(x, y)/p(x).
\end{aligned}$$

(This notation is occasionally ambiguous, and if absolutely necessary appropriate subscripts will be added, for example, $p_X(x)$, $p_{Y|X}(y, x)$. This need will arise, however, only when actual numbers are substituted for the letters x, y; see Example 1.6.) Our definition is:

$$\begin{aligned}
H(X|Y) &= E\left[\log \frac{1}{p(x|y)}\right]\\
&= \sum_{x,y} p(x, y) \log \frac{1}{p(x|y)}. \qquad (1.3)
\end{aligned}$$

(In (1.3) we observe the same conventions as we did for sum (1.1): $0 \log 0^{-1} = 0$; a divergent sum means $H(X|Y) = +\infty$.) Let us pause to motivate the definition via a simple model for a communications channel, called a *discrete memoryless channel* (DMC).

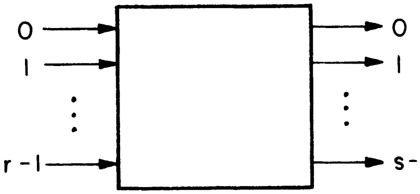


Figure 1.3 A discrete memoryless channel.

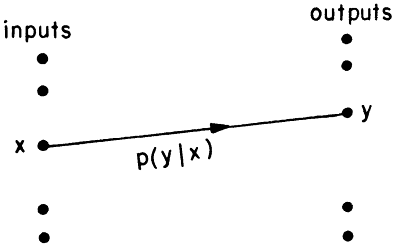


Figure 1.4 Another view of a DMC.

A DMC (Fig. 1.3) is an object that accepts, every unit of time, one of r input symbols, and in response expels one of s output symbols. (This channel is “discrete” because there are only finitely[3] many input and output symbols, “memoryless” because the current output depends only on the current input and not on any of the previous ones.) The precise labeling of the input and output symbols is of no real importance, but it is often convenient to let $\{0, 1, \ldots, r-1\}$ and $\{0, 1, \ldots, s-1\}$ represent the input and output alphabets.

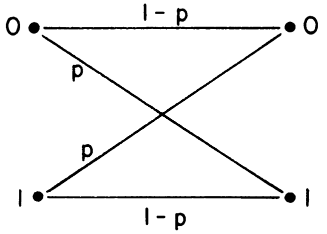
The output is not a definite function of the input, however; rather the channel’s behavior is governed by an $r \times s$ matrix of *transition probabilities* $(p(y|x))$. The number $p(y|x)$ represents the probability that y will be the output, given that x is the input. Clearly the number $p(y|x)$ must satisfy

$$p(y|x) \geqslant 0 \qquad \text{for all}\, x,\ y,$$

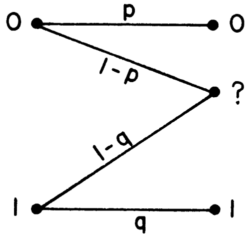
$$\sum_y p(y|x) = 1 \qquad \text{for all}\, x.$$

Sometimes. when r and s are not too big, the DMC is depicted graphically as shown in Fig. 1.4. In such a picture each pair (x, y) with $p(y|x) > 0$ is joined by a line labeled with the number $p(y|x)$.

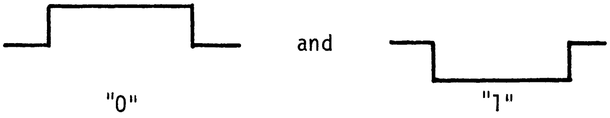
Example 1.4 (the binary symmetric channel, already discussed in the introduction). Here $r = s = 2$, and the graph looks like this:



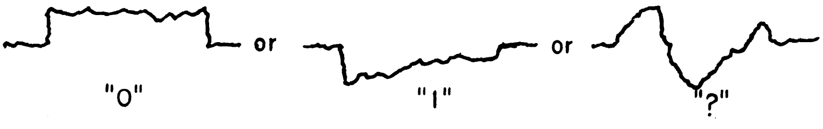
Example 1.5 (The binary erasure channel). Here $r = 2$, $s = 3$. The inputs are labeled "0" and "1," and the outputs are labeled "0," "1," and "?."



Such a channel might arise in practice for example if the inputs to a physical channel were the two squarewaves.



The detector at the output would receive a noisy version of these square waves, $r(t)$:



It might base its decision about whether "0" or "1" was sent on the value of the integral $\int r(t)\,dt = I$. If I is positive, the detector could decide "0" was sent; if negative, "1." However, if $|I|$ is very small, it might be best not to make a "hard decision" about the transmitted bit, but rather to output a special erasure symbol "?." If the channel is relatively quiet, the transitions $0 \to 1$ and $1 \to 0$ would be much less likely than $0 \to ?$ and $1 \to ?$, so the assumptions $P\{Y = 1|X = 0\} = P\{Y = 0|X = 1\} = 0$ might be reasonable. (For more on "hard decisions," see Prob. 4.15.) □

Suppose now that the inputs to a DMC are selected according to a probability distribution $p(x)$ on $\{0, 1, \ldots, r-1\}$, that is, assume the input X to the channel is characterized by

$$P\{X = x\} = p(x), \qquad x \in \{0, 1, \ldots, r-1\}.$$

Having specified X, we can now define a random variable Y which will represent the *output* of the channel. The joint distribution of X and Y is given by

$$\begin{aligned} p(x, y) &= P\{X = x, Y = y\} \\ &= P\{X = x\}P\{Y = y|X = x\} \\ &= p(x)p(y|x), \end{aligned}$$

and the marginal distribution of Y is

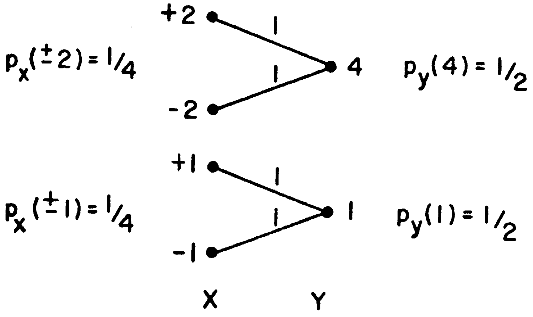
$$\begin{aligned} p(y) &= P\{Y = y\} \\ &= \sum_x P\{Y = y|X = x\}P\{X = x\} \\ &= \sum_x p(y|x)p(x). \end{aligned}$$

Similarly,

$$\begin{aligned} p(x|y) &= p(x, y)/p(y) \\ &= p(y|x)p(x)/\sum_{x'} p(y|x')p(x'). \end{aligned}$$

Hence corresponding to every DMC and input distribution there is a pair of random variables: X, the "input" to, and Y the "output" from, the channel. Conversely, given any pair (X, Y) of discrete random variables , there exist a DMC and input distribution such that X is the input and Y is the output: simply define the channel's transition probabilities by $p(y|x) = P\{Y = y|X = x\}$. In other words, given any ordered pair (X, Y) of random variables, it is possible to think of Y as a "noisy" version of X, that is, as the result of transmitting X through a certain DMC.

Example 1.6 Let X assume the values ± 1, ± 2, each with probability $\frac{1}{4}$, and let $Y = X^2$. The corresponding DMC looks like this:



In this example X and Y are uncorrelated, and yet it is clear that Y provides a considerable amount of "information" about X (see Prob. 1.10). □

Given that we think of Y as a noisy version of X, and that $H(X)$ is a measure of our prior uncertainty about X, how can we measure our uncertainty about X after observing Y? Well, suppose we have observed that $Y = y$. Then, since the numbers $p(x|y) = P\{X = x|Y = y\}$ for fixed y represent the conditional distribution of X, given that $Y = y$, we define the *conditional entropy* of X, given $Y = y$:

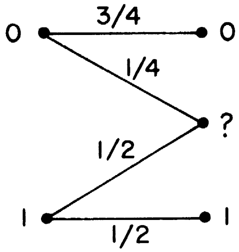
$$H(X|Y = y) = \sum_x p(x|y) \log \frac{1}{p(x|y)}.$$

This quantity is itself a random variable defined on the range of Y; let us define the *conditional entropy* $H(X|Y)$ as its expectation:

$$\begin{aligned} H(X|Y) &= \sum_y p(y) H(X|Y = y) \\ &= \sum_y p(y) \sum_x p(x|y) \log \frac{1}{p(x|y)} \\ &= \sum_{x,y} p(x, y) \log \frac{1}{p(x|y)}, \end{aligned}$$

in agreement with Eq. (1.3). Thus, for a given pair X, Y of random variables, *$H(X|Y)$ represents the amount of uncertainty remaining about X after Y has been observed.*

Example 1.7 Consider the following DMC, which is a particular case of the binary erasure channel of Example 1.5:



Here $p_X(0) = \frac{2}{3}$, $p_X(1) = \frac{1}{3}$. Then a simple calculation yields:

$$H_2(X) = 0.9183 \text{ bits},$$

$$H_2(X|Y = 0) = 0,$$

$$H_2(X|Y = 1) = 0,$$

$$H_2(X|Y = ?) = 1.$$

Thus, if $Y = 0$ or 1, there is no remaining uncertainty about X, but if $Y = ?$, we are more uncertain about X after receiving Y than before! However,

$$H_2(X|Y) = 0.3333 \text{ bits},$$

so that, on the average, at least, an observation of Y reduces our uncertainty about X. □

We now present a technical lemma on $H(X|Y)$ that will be useful later.

Theorem 1.2 *Let X, Y, Z be discrete random variables. Using obvious notation (see Eqs.* (1.2)*), define, for each z, $A(z) = \sum_{x,y} p(y)p(z|x, y)$. Then*

$$H(X|Y) \leqslant H(Z) + E(\log A).$$

Proof

$$\begin{aligned} H(X|Y) &= E\left[\log \frac{1}{p(x|y)}\right] \\ &= \sum_{x,y,z} p(x, y, z) \log \frac{1}{p(x|y)}. \\ &= \sum_{z} p(z) \sum_{x,y} \frac{p(x, y, z)}{p(z)} \log \frac{1}{p(x|y)}. \end{aligned}$$

For fixed z, $p(x, y, z)/p(z) = p(x, y|z)$ is a probability distribution, and so we can apply Jensen's inequality to the inner sum. The result is

$$H(X|Y) \leqslant \sum_z p(z) \log \left[\frac{1}{p(z)} \cdot \sum_{x,y} \frac{p(x, y, z)}{p(x|y)} \right]$$

$$= \sum_z p(z) \log \frac{1}{p(z)} + \sum_z p(z) \log \sum_{x,y} \frac{p(x, y, z)}{p(x|y)}.$$

But $p(x, y, z)/p(x|y) = p(x, y, z)p(y)/p(x, y) = p(y)p(z|x, y)$. □

Corollary ("*Fano's inequality*"). *Let X and Y be random variables, each taking values in the set $\{x_1, x_2, \ldots, x_r\}$. Let $P_e = P\{X \neq Y\}$. Then*

$$H(X|Y) \leqslant H(P_e) + P_e \log(r - 1).$$

Proof In Theorem 1.2 define $Z = 0$ if $X = Y$ and $Z = 1$ if $X \neq Y$. Then $A(0) = 1$ and $A(1) = r - 1$. □

[*Note*: The proof of Theorem 1.2 via our streamlined notation contains some subtle features; see Prob. 1.11.]

Fano's inequality has an interesting heuristic interpretation. Think of $H(X|Y)$ as the amount of information needed to determine X once Y is known. One way to determine X is to first determine whether or not $X = Y$; if $X = Y$, we are done. If, however, $X \neq Y$, there are $r - 1$ remaining possibilities for X. Determining whether or not $X = Y$ is equivalent to determining the random variable Z defined in the proof; since $H(Z) = H(P_e)$, it takes $H(P_e)$ bits to do this. If $X \neq Y$ (this happens with probability P_e), the amount of information needed to find out which of the remaining $r - 1$ values X has is, by Theorem 1.1. at most $\log(r - 1)$.

Example 1.8 We apply Fano's inequality to the channel of Example 1.7. Here $r = 3$, and $P\{X = Y\} = \frac{2}{3}$, $P_e = \frac{1}{3}$. Fano's bound is thus $H(X|Y) \leqslant H\left(\frac{1}{3}\right) + \frac{1}{3}\log 2 = \frac{1}{3}\log 3 + \frac{2}{3}\log\frac{3}{2} + \frac{1}{3}\log 2 = \log 3 - \frac{1}{3}\log 2 = 1.2520$ bits. (For examples where Fano's inequality does better, see Prob.1.11.) □

Now since $H(X)$ represents our uncertainty about X before we know Y, and $H(X|Y)$ represents our uncertainty after, the difference $H(X) - H(X|Y)$ must represent the amount of information provided about X by Y. This important quantity is called the *mutual information* between X and Y, and is denoted by $I(X; Y)$:

$$I(X;\, Y) = H(X) - H(X|Y). \tag{1.4}$$

(In Example 1.7, $I_2(X;\, Y) = 0.9183 - 0.3333 = 0.5850$; thus, informally at least, the observation of a channel output provides 0.5850 bits of information about the input, on the average.) Using the notation of Eq. (1.2), we obtain several important alternative forms for $I(X;\, Y)$:

$$I(X;\, Y) = \sum_{x,y} p(x,\, y) \log \frac{p(x|y)}{p(x)}, \tag{1.5}$$

$$= \sum_{x,y} p(x,\, y) \log \frac{p(x,\, y)}{p(x)p(y)}, \tag{1.6}$$

$$= \sum_{x,y} p(x,\, y) \log \frac{p(y|x)}{p(y)}. \tag{1.7}$$

(The details are left as Prob. 1.14.)

We thus see that $I(X;\, Y)$ is the average, taken over the X, Y sample space, of the random variable[4]

$$I(x;\, y) = \log \frac{p(x|y)}{p(x)} = \log \frac{p(x,\, y)}{p(x)p(y)} = \log \frac{p(y|x)}{p(y)}.$$

Now $I(x;\, y)$ can be either positive or negative (e.g., in Example 1.7 $I(0;\, 0) = \log \frac{3}{2}$ and $I(0;\, ?) = \log \frac{3}{4}$); however, we shall now prove the important fact that $I(X;\, Y)$ cannot be negative. This is surely reasonable, given our heuristics: we do not expect to be misled (on the average) by observing the output of the channel.

Theorem 1.3 *For any discrete random variables X and Y, $I(X;\, Y) \geqslant 0$. Moreover $I(X;\, Y) = 0$ if and only if X and Y are independent.*

Proof We apply Jensen's inequality to Eq. (1.6):

$$\begin{aligned} -I(X;\, Y) &= \sum_{x,y} p(x,\, y) \log \frac{p(x)p(y)}{p(x,\, y)} \\ &\leqslant \log \sum_{x,y} p(x)p(y) \\ &= \log 1 = 0. \end{aligned}$$

Furthermore, in view of the strict convexity $\cap$ of $\log x$, equality holds iff $p(x)p(y) = p(x,\, y)$ for all x, y, that is, iff X and Y are independent. □

(Although we shall not emphasize it, Theorem 1.3 shows that $I(X;\, Y)$ is a good measure of the dependence between X and Y, better for example than the covariance $\mathrm{Cov}(X;\, Y)$. for example, recall Example 1.6. There, as is easily verified, $\mathrm{Cov}(X;\, Y) = 0$ but $I_2(X;\, Y) = 1$ bit.)

Using Eqs. (1.4)–(1.7), it is possible to prove immediately several important facts about mutual information:

$$I(X;\, Y) = I(Y;\, X), \tag{1.8}$$

$$I(X;\, Y) = H(Y) - H(Y|X), \tag{1.9}$$

$$I(X;\, Y) = H(X) + H(Y) - H(X,\, Y), \tag{1.10}$$

where in (1.10) we have defined the *joint entropy* of X and Y by

$$H(X,\, Y) = \sum_{x,y} p(x,\, y) \log \frac{1}{p(x,\, y)}. \tag{1.11}$$

The proofs of these relationships are left as Prob. 1.14. They can be easily remembered by means of the Venn diagram shown in Fig. 1.5. It is a fruitful exercise to give informal interpretations of each of the relations implied by Fig. 1.5. For example, Eq. (1.8) expresses the "mutuality" of mutual information; $H(X,\, Y) = H(X) + H(Y|X)$ becomes "our uncertainty about X and Y is the sum of our uncertainty about X and our uncertainty about Y, once X is known," and so on.

Now if we are given three random variables $X,\, Y,\, Z$, we define the mutual information $I(X,\, Y;\, Z)$ ("the amount of information X and Y provide about Z"), analogously with Eq. (1.7), by

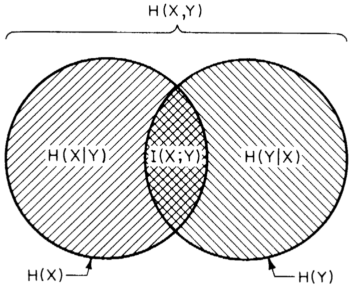


Figure 1.5 A mnemonic Venn diagram for Eqs. (1.4) and (1.8)–(1.10).

$$I(X,\ Y;\ Z) = E\left[\log\frac{p(z|x,\ y)}{p(z)}\right]$$

$$= \sum_{x,y,z} p(x,\ y,\ z)\log\frac{p(z|x,\ y)}{p(z)}.$$

We would not expect X and Y together to provide less information about Z than Y alone does, and indeed this is the case.

Theorem 1.4 *$I(X,\ Y;\ Z) \geqslant I(Y;\ Z)$, with equality iff $p(z|x,\ y) = p(z|y)$ for all $(x,\ y,\ z)$ with $p(x,\ y,\ z) > 0$.*

Proof

$$I(Y;\ Z) - I(X,\ Y;\ Z) = E\left[\log\frac{p(z|y)}{p(z)} - \log\frac{p(z|x,\ y)}{p(z)}\right]$$

$$= E\left[\log\frac{p(z|y)}{p(z|x,\ y)}\right]$$

$$= \sum_{xy,z} p(x,\ y,\ z)\log\frac{p(z|y)}{p(z|x,\ y)}.$$

Applying Jensen's inequality, we have

$$I(Y;\ Z) - I(X,\ Y;\ Z) \leqslant \log\sum_{xy,z} p(x,\ y,\ z)\frac{p(z|y)}{p(z|x,\ y)}$$

$$= \log\sum_{xy,z} p(x,\ y)\cdot p(z|y)$$

$$= \log 1 = 0.$$

The conditions for equality follow from the discussion of Jensen's inequality in Appendix B. □

The condition for equality in Theorem 1.4 is very interesting; it says that the sequence $(X,\ Y,\ Z)$ is a Markov chain, which for our purposes means simply that X, Y, *and* Z can be viewed as shown in Fig. 1.6. Here DMC 1 is characterized by the transition probabilities $p(y|x)$, and DMC 2 by the transition probabilities $p(z|y) = p(z|x,\ y)$. We have already observed that given any pair of random variables $(X,\ Y)$, it is possible to devise a DMC with X as the input and Y as the output. However it is not true that if

(X, Y, Z) is any triple of random variables, there exists a pair of DMC's such that X, Y, Z have the relationship of Fig. 1.6. Indeed, it is clear that a necessary and sufficient condition for this is that (X, Y, Z) forms a Markov chain, that is, $p(z|y) = p(z|x, y)$ (i.e., Z depends on X only through Y).

Now let's assume that (X, Y, Z) is a Markov chain, as in Fig. 1.6. Then by Theorem 1.4, $I(X; Z) \leqslant I(X, Y; Z)$, and since (X, Y, Z) is a Markov chain, $I(X, Y; Z) = I(Y; Z)$. Hence $I(X; Z) \leqslant I(Y; Z)$. Now if (X, Y, Z) is a Markov chain, so is (Z, Y, X) (see Prob. 1.15), and hence $I(X; Z) \leqslant I(X; Y)$. Since this is an extremely important information-theoretic property of Markov chains, we display it as a theorem.

Theorem 1.5 *If (X, Y, Z) is a Markov chain, then*

$$I(X; Z) \leqslant \begin{cases} I(X; Y) \\ I(Y; Z). \end{cases} \qquad \square$$

Referring again to Fig. 1.6, we find that DMC's tend to "leak" information. If the DMC's are deterministic (i.e., if Y is a definite function of X and Z a definite function of Y), we can think of the casade in Fig. 1.6 as a kind of data-processing configuration. Paradoxically, Theorem 1.5 says that data processing can only destroy information! (For an important generalization of this, see Eq. (1.15).)

Example 1.9 Let X_1, X_2, X_3 be independent random variables; then $(X_1, X_1 + X_2, X_1 + X_2 + X_3)$ is a Markov chain, and so $I(X_1; X_1 + X_2 + X_3) \leqslant I(X_1; X_1 + X_2)$ (see Probs. 1.16 and 1.39). □

Example 1.10 In Fig. 1.6 assume that X is described by $P\{X = 0\} = P\{X = 1\} = \frac{1}{2}$, and that both DMC's are binary symmetric channels with error probability p. Then

$$I(X; Y) = 1 - H_2(p) \quad \text{bits,}$$
$$I(X; Z) = 1 - H_2[2p(1 - p)] \quad \text{bits.}$$

These two functions are plotted as follows: (For an extension of this

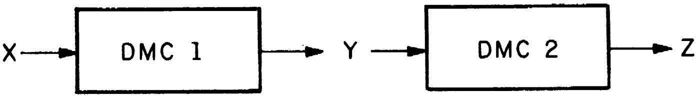
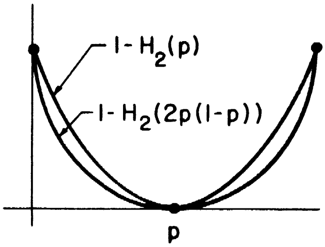


Figure 1.6 An information theorist's view of a Markov chain.



example, see Probs. 1.18 and 1.20.) □

We conclude this section with two results about the convexity of $I(X;\, Y)$ when it is viewed as a function either of the input probabilities $p(x)$ or of the transition probabilities $p(y|x)$.

Theorem 1.6 *$I(X;\, Y)$ is a convex $\cap$ function of the input probabilities $p(x)$.*

Proof We think of the transition probabilities $p(y|x)$ as being fixed, and consider two input random variables X_1 and X_2 with probability distributions $p_1(x)$ and $p_2(x)$. If X's probability distribution is a convex combination $p(x) = \alpha p_1(x) + \beta p_2(x)$, we must show that

$$\alpha I(X_1;\, Y_1) + \beta I(X_2;\, Y_2) \leqslant I(X;\, Y),$$

where Y_1, Y_2 and Y are the channel outputs corresponding to X_1, X_2, and X, respectively. To do this consider the following manipulation, which uses obvious notational shorthand:

$$\begin{aligned}
&\alpha I(X_1;\, Y_1) + \beta I(X_2;\, Y_2) - I(X;\, Y)\\
&\quad = \sum_{x,y} \alpha p_1(x,\, y) \log \frac{p(y|x)}{p_1(y)} + \sum_{x,y} \beta p_2(x,\, y) \log \frac{p(y|x)}{p_2(y)} \qquad \text{(see Eq.(1.7))}\\
&\qquad - \sum_{x,y} [\alpha p_1(x,\, y) + \beta p_2(x,\, y)] \log \frac{p(y|x)}{p(y)}\\
&\quad = \alpha \sum_{x,y} p_1(x,\, y) \log \frac{p(y)}{p_1(y)} + \beta \sum_{x,y} p_2(x,\, y) \log \frac{p(y)}{p_2(y)}. \qquad (1.12)
\end{aligned}$$

We now apply Jensen's inequality to each of the above sums. For example,

$$\sum_{x,y} p_1(x,\, y) \log \frac{p(y)}{p_1(y)} \leqslant \log \sum_{x,y} p_1(x,\, y) \frac{p(y)}{p_1(y)}.$$

But

$$\begin{aligned}\sum_{x,y} p_1(x, y)\frac{p(y)}{p_1(y)} &= \sum_y \frac{p(y)}{p_1(y)} \sum_x p_1(x, y)\\ &= \sum_y \frac{p(y)}{p_1(y)} \cdot p_1(y)\\ &= 1.\end{aligned}$$

Hence the first sum in (1.12) is $\leqslant 0$; similarly, so is the second. □

Corollary *The entropy function $H(p_1, p_2, \ldots, p_r)$ is convex $\cap$.*

Proof Let X be a random variable distributed according to $P\{X = i\} = p_i$. Then $I(X; X) = H(X) = H(p_1, p_2, \ldots, p_r)$. The result now follows from Theorem 1.6. □

Theorem 1.7 *$I(X; Y)$ is convex $\cup$ in the transition probabilities $p(y|x)$.*

Proof Here the input probabilities $p(x)$ are fixed, but we are given two sets of transition probabilities $p_1(y|x)$ and $p_2(y|x)$ and a convex combination $p(y|x) = \alpha p_1(y|x) + \beta p_2(y|x)$. It is required to show that

$$I(X; Y) \leqslant \alpha I(X; Y_1) + \beta I(X; Y_2), \tag{1.13}$$

where Y, Y_1, Y_2 are the channel outputs corresponding to the transition probabilities $p(y|x)$, $p_1(y|x)$, and $p_2(y|x)$. Again using obvious notation, the difference between the left and right sides of (1.13) is (see Eq. (1.5))

$$\begin{aligned}&\sum_{x,y}[\alpha p_1(x, y) + \beta p_2(x, y)] \log \frac{p(x|y)}{p(x)}\\ &\qquad - \sum_{x,y} \alpha p_1(x, y) \log \frac{p_1(x|y)}{p(x)} - \sum_{x,y} \beta p_2(x, y) \log \frac{p_2(x|y)}{p(x)}\\ &\qquad = \alpha \sum_{x,y} p_1(x, y) \log \frac{p(x|y)}{p_1(x|y)} + \beta \sum_{x,y} p_2(x, y) \log \frac{p(x|y)}{p_2(x|y)}.\end{aligned} \tag{1.14}$$

The first sum in (1.14) is, by Jensen's inequality,

$$\leqslant \alpha \log\left[\sum_{x,y} p_1(x,\, y)\frac{p(x|y)}{p_1(x|y)}\right]$$

$$= \alpha \log\left[\sum_{x,y} p(x|y)p_1(y)\right]$$

$$= \alpha \log \sum_{y} p_1(y) = 0.$$

Similarly the second sum is $\leqslant 0$. □

1.2 Discrete random vectors

In Eq. (1.11) we defined the entropy $H(X,\, Y)$ of a pair of random variables, and on p. 28 we defined the mutual information $I(X,\, Y;\, Z)$ between a pair of random variables and a third random variable. In this section we will generalize those definitions and define $H(\mathbf{X})$, $H(\mathbf{X}|\mathbf{Y})$, and $I(\mathbf{X};\, \mathbf{Y})$, where $\mathbf{X}$ and $\mathbf{Y}$ are arbitrary random vectors.

Our point of view is that a random vector $\mathbf{X} = (X_1,\, X_2,\, \ldots,\, X_n)$ is just a finite list of random variables X_i. The distribution of $\mathbf{X}$ (the joint distribution of $X_1,\, X_2,\, \ldots,\, X_n$) is the function $p(x_1,\, x_2,\, \ldots,\, x_n) = P\{X_1 = x_1,\, X_2 = x_2,\, \ldots,\, X_n = x_n\}$, where each x_i is in the range of X_i. A glance at the definitions in Section 1.1 should convince the reader that $H(X)$, $H(X|Y)$, $I(X;\, Y)$ depend only on the distribution functions $p(x)$, $p(y|x)$, etc., and not in any way on the fact that the values assumed by X and Y are real numbers. Hence we can immediately extend these definitions to arbitrary random vectors; for example the entropy of $\mathbf{X} = (X_1,\, \ldots,\, X_n)$ is defined as

$$H(\mathbf{X}) = \sum_{\mathbf{x}} p(\mathbf{x}) \log \frac{1}{p(\mathbf{x})},$$

where the summation is extended over all vectors $\mathbf{x}$ in the range of $\mathbf{X}$. And obviously Theorems 1.1–1.7 remain true.

The generalization of Theorem 1.5 to arbitrary random vectors has a particularly important application, which we now discuss. Consider the model for a communication system shown in Fig. 1.7 (cf. Figs. 0.2 and 5.1). In Fig. 1.7 the random vector $\mathbf{U}$ is a model for k consecutive source outputs; the encoder is a device that takes $\mathbf{U}$ and maps it into an n-tuple $\mathbf{X}$ for transmission over the channel; $\mathbf{Y}$ is the channel's noisy version of $\mathbf{X}$; and the decoder is a device that takes $\mathbf{Y}$ and maps it into a k-tuple $\mathbf{V}$, which is delivered to the destination and is supposed to reproduce $\mathbf{U}$, at least approximately.

The point of all this is that, for any realizable communication system, the sequence $(\mathbf{U}, \mathbf{X}, \mathbf{Y}, \mathbf{V})$ of random vectors forms a Markov chain (see Fig. 1.6). Informally this says that the output of each box in Fig. 1.7 depends only on its input and not on any of the earlier random vectors. Formally it gives many conditions on the various conditional probabilities, for example, $p(\mathbf{y}|\mathbf{x}, \mathbf{u}) = p(\mathbf{y}|\mathbf{x})$, $p(\mathbf{v}|\mathbf{y}, \mathbf{x}) = p(\mathbf{v}|\mathbf{y})$. (There is really no question of proving this part; it is one of the fundamental assumptions we make about a communication system.) Applying Theorem 1.5 to the sub-Markov chain $(\mathbf{U}, \mathbf{X}, \mathbf{V})$, we get $I(\mathbf{U}; \mathbf{V}) \leqslant I(\mathbf{X}; \mathbf{V})$. Similarly $I(\mathbf{X}; \mathbf{V}) \leqslant I(\mathbf{X}; \mathbf{Y})$. Hence for the random variables of Fig. 1.7,

$$I(\mathbf{U}; \mathbf{V}) \leqslant I(\mathbf{X}; \mathbf{Y}). \tag{1.15}$$

This result is called the *data-processing theorem*. Stated bluntly, it says that the information processing (the work done by the encoder and decoder of Fig. 1.7) can only destroy information! It says, for example, that the noisy channel output $\mathbf{Y}$ in Fig. 1.7 contains more information about the source sequence $\mathbf{U}$ than does the decoder's estimate $\mathbf{V}$. (While this is true theoretically, the data processing of the decoder is nevertheless required to render this information usable.)

We now come to a pair of inequalities involving $I(\mathbf{X}; \mathbf{Y})$ and $\sum_{i=1}^{n} I(X_i; Y_i)$, where $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ and $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)$ are a pair of n-dimensional random vectors.

Theorem 1.8 *If the components* $(X_1, X_2, \ldots, X_n)$ *of* $\mathbf{X}$ *are independent, then*

$$I(\mathbf{X}; \mathbf{Y}) \geqslant \sum_{i=1}^{n} I(X_i; Y_i).$$

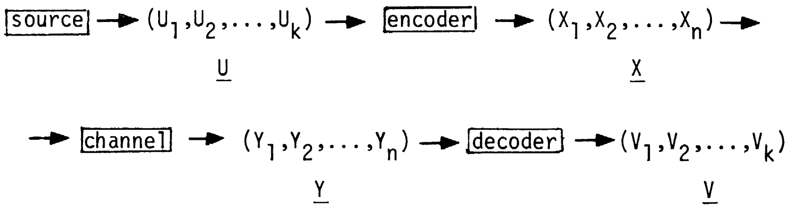

source
$(U_1, U_2, \ldots, U_k)$
$\underline{U}$
encoder
$(X_1, X_2, \ldots, X_n)$
$\underline{X}$
channel
$(Y_1, Y_2, \ldots, Y_n)$
$\underline{Y}$
decoder
$(V_1, V_2, \ldots, V_k)$
$\underline{V}$

Figure 1.7 A general communication system.

Proof Letting E denote expectation on the joint sample space of $\mathbf{X}$ and $\mathbf{Y}$, we have

$$\begin{aligned} I(\mathbf{X};\,\mathbf{Y}) &= E\left[\log\frac{p(\mathbf{x}|\mathbf{y})}{p(\mathbf{x})}\right] \qquad (\textit{see Eq.}\ (1.5)) \\ &= E\left[\log\frac{p(\mathbf{x}|\mathbf{y})}{p(x_1)p(x_2)\ldots p(x_n)}\right], \end{aligned}$$

since $X_1, X_2, \ldots, X_n$ are assumed independent. On the other hand,

$$\begin{aligned} \sum_{i=1}^{n} I(X_i,\,Y_i) &= \sum_{i=1}^{n} E\left[\log\frac{p(x_i|y_i)}{p(x_i)}\right] \\ &= E\left[\log\frac{p(x_1|y_1)\ldots p(x_n|y_n)}{p(x_1)\ldots p(x_n)}\right]. \end{aligned}$$

Hence

$$\begin{aligned} &\sum_{i=1}^{n} I(X_i,\,Y_i) - I(\mathbf{X};\,\mathbf{Y}) \\ &\qquad = E\left[\log\frac{p(x_1|y_1)\ldots p(x_n|y_n)}{p(\mathbf{x}|\mathbf{y})}\right] \\ &\qquad \leqslant \log E\left[\frac{p(x_1|y_1)\ldots p(x_n|y_n)}{p(\mathbf{x}|\mathbf{y})}\right] = 0 \end{aligned}$$

by Jensen's inequality, since this last expectation is

$$\begin{aligned} \sum_{\mathbf{x},\mathbf{y}} p(\mathbf{x},\,\mathbf{y})\{\cdots\} &= \sum_{\mathbf{x},\mathbf{y}} p(x_1|y_1)\ldots p(x_n|y_n)p(\mathbf{y}) \\ &= 1. \end{aligned}$$

□

Example 1.11 Let $X_1, X_2, \ldots, X_n$ be independent identically distributed random variables with common entropy H. Also let π be a permutation of the set $\{1, 2, \ldots, n\}$, and let $Y_i = X_{\pi(i)}$. Then $I(\mathbf{X};\,\mathbf{Y}) = nH$, but $\sum I(X_i;\,Y_i) = kH$, where k is the number of fixed points of π, that is, the number of integers i with $\pi(i) = i$. In particular if π has no fixed points, for example if $\pi(i) \equiv i + 1 \pmod n$, then $\sum I(X_i;\,Y_i) = 0$ (see Prob. 1.23). □

If we think of $(Y_1, Y_2, \ldots, Y_n)$ as the n outputs of a noisy channel when the inputs are $X_1, X_2, \ldots, X_n$, Theorem 1.8 tells us that, if the inputs are independent, $\mathbf{Y}$ provides more information about $\mathbf{X}$ than the total amount of information provided about each X_i by the corresponding Y_i. The next theorem will tell us that if we drop the assumption of independence about the X_i and assume instead that the $(\mathbf{X}, \mathbf{Y})$ channel is memoryless, that is,

$$p(y_1, \ldots, y_n|x_1, \ldots, x_n) = \prod_{i=1}^{n} p(y_i|x_i), \tag{1.16}$$

the situation is quite different!

Theorem 1.9 *If* $\mathbf{X} = (X_1, \ldots, X_n)$ *and* $\mathbf{Y} = (Y_1, \ldots, Y_n)$ *are random vectors and the channel is memoryless, that is, if* (1.16) *holds, then*

$$I(\mathbf{X}; \mathbf{Y}) \leqslant \sum_{i=1}^{n} I(X_i; Y_i).$$

Proof Again letting E denote expectation on the joint sample space of $\mathbf{X}$ and $\mathbf{Y}$, we have

$$\begin{aligned} I(\mathbf{X}; \mathbf{Y}) &= E\left[\log \frac{p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y})}\right] \qquad (\textit{see Eq. } (1.7)) \\ &= E\left[\log \frac{p(y_1|x_1) \ldots p(y_n|x_n)}{p(\mathbf{y})}\right] \end{aligned}$$

by (1.16). On the other hand,

$$\begin{aligned} \sum_{i=1}^{n} I(X_i; Y_i) &= \sum_{i=1}^{n} E\left[\log \frac{p(y_i|x_i)}{p(y_i)}\right] \\ &= E\left[\log \frac{p(y_1|x_1) \ldots p(y_n|x_n)}{p(y_1) \ldots p(y_n)}\right]. \end{aligned}$$

Hence

$$\begin{aligned} &I(\mathbf{X}; \mathbf{Y}) - \sum_{i=1}^{n} I(X_i; Y_i) \\ &= E\left[\log \frac{p(y_1) \ldots p(y_n)}{p(\mathbf{y})}\right] \\ &\leqslant \log E\left[\frac{p(y_1) \ldots p(y_n)}{p(\mathbf{y})}\right] \\ &= 0 \end{aligned}$$