CHAPTER 2

ENTROPY, RELATIVE ENTROPY, AND MUTUAL INFORMATION

In this chapter we introduce most of the basic definitions required for subsequent development of the theory. It is irresistible to play with their relationships and interpretations, taking faith in their later utility. After defining entropy and mutual information, we establish chain rules, the nonnegativity of mutual information, the data-processing inequality, and illustrate these definitions by examining sufficient statistics and Fano's inequality.

The concept of information is too broad to be captured completely by a single definition. However, for any probability distribution, we define a quantity called the *entropy*, which has many properties that agree with the intuitive notion of what a measure of information should be. This notion is extended to define *mutual information*, which is a measure of the amount of information one random variable contains about another. Entropy then becomes the self-information of a random variable. Mutual information is a special case of a more general quantity called *relative entropy*, which is a measure of the distance between two probability distributions. All these quantities are closely related and share a number of simple properties, some of which we derive in this chapter.

In later chapters we show how these quantities arise as natural answers to a number of questions in communication, statistics, complexity, and gambling. That will be the ultimate test of the value of these definitions.

2.1 ENTROPY

We first introduce the concept of *entropy*, which is a measure of the uncertainty of a random variable. Let X be a discrete random variable with alphabet $\mathcal{X}$ and probability mass function $p(x) = \Pr\{X = x\}$, $x \in \mathcal{X}$.

Elements of Information Theory, Second Edition, By Thomas M. Cover and Joy A. Thomas

We denote the probability mass function by $p(x)$ rather than $p_X(x)$, for convenience. Thus, $p(x)$ and $p(y)$ refer to two different random variables and are in fact different probability mass functions, $p_X(x)$ and $p_Y(y)$, respectively.

Definition The *entropy* $H(X)$ of a discrete random variable X is defined by

$$H(X) = -\sum_{x\in\mathcal{X}} p(x)\log p(x). \tag{2.1}$$

We also write $H(p)$ for the above quantity. The log is to the base 2 and entropy is expressed in bits. For example, the entropy of a fair coin toss is 1 bit. We will use the convention that $0\log 0 = 0$, which is easily justified by continuity since $x\log x \to 0$ as $x \to 0$. Adding terms of zero probability does not change the entropy.

If the base of the logarithm is b, we denote the entropy as $H_b(X)$. If the base of the logarithm is e, the entropy is measured in *nats*. Unless otherwise specified, we will take all logarithms to base 2, and hence all the entropies will be measured in bits. Note that entropy is a functional of the distribution of X. It does not depend on the actual values taken by the random variable X, but only on the probabilities.

We denote expectation by E. Thus, if $X \sim p(x)$, the expected value of the random variable $g(X)$ is written

$$E_p g(X) = \sum_{x\in\mathcal{X}} g(x)p(x), \tag{2.2}$$

or more simply as $Eg(X)$ when the probability mass function is understood from the context. We shall take a peculiar interest in the eerily self-referential expectation of $g(X)$ under $p(x)$ when $g(X) = \log\frac{1}{p(X)}$.

Remark The entropy of X can also be interpreted as the expected value of the random variable $\log\frac{1}{p(X)}$, where X is drawn according to probability mass function $p(x)$. Thus,

$$H(X) = E_p \log\frac{1}{p(X)}. \tag{2.3}$$

This definition of entropy is related to the definition of entropy in thermodynamics; some of the connections are explored later. It is possible to derive the definition of entropy axiomatically by defining certain properties that the entropy of a random variable must satisfy. This approach is illustrated in Problem 2.46. We do not use the axiomatic approach to justify the definition of entropy; instead, we show that it arises as the answer to a number of natural questions, such as "What is the average length of the shortest description of the random variable?" First, we derive some immediate consequences of the definition.

Lemma 2.1.1 $H(X) \geq 0$.

Proof: $0 \leq p(x) \leq 1$ implies that $\log \frac{1}{p(x)} \geq 0$. □

Lemma 2.1.2 $H_b(X) = (\log_b a) H_a(X)$.

Proof: $\log_b p = \log_b a \log_a p$. □

The second property of entropy enables us to change the base of the logarithm in the definition. Entropy can be changed from one base to another by multiplying by the appropriate factor.

Example 2.1.1 Let

$$X = \begin{cases} 1 & \text{with probability } p, \\ 0 & \text{with probability } 1-p. \end{cases} \tag{2.4}$$

Then

$$H(X) = -p \log p - (1-p)\log(1-p) \overset{\text{def}}{=\!=} H(p). \tag{2.5}$$

In particular, $H(X) = 1$ bit when $p = \frac{1}{2}$. The graph of the function $H(p)$ is shown in Figure 2.1. The figure illustrates some of the basic properties of entropy: It is a concave function of the distribution and equals 0 when $p = 0$ or 1. This makes sense, because when $p = 0$ or 1, the variable is not random and there is no uncertainty. Similarly, the uncertainty is maximum when $p = \frac{1}{2}$, which also corresponds to the maximum value of the entropy.

Example 2.1.2 Let

$$X = \begin{cases} a & \text{with probability} \frac{1}{2}, \\ b & \text{with probability} \frac{1}{4}, \\ c & \text{with probability} \frac{1}{8}, \\ d & \text{with probability} \frac{1}{8}. \end{cases} \tag{2.6}$$

The entropy of X is

$$H(X) = -\frac{1}{2}\log\frac{1}{2} - \frac{1}{4}\log\frac{1}{4} - \frac{1}{8}\log\frac{1}{8} - \frac{1}{8}\log\frac{1}{8} = \frac{7}{4} \text{ bits}. \tag{2.7}$$

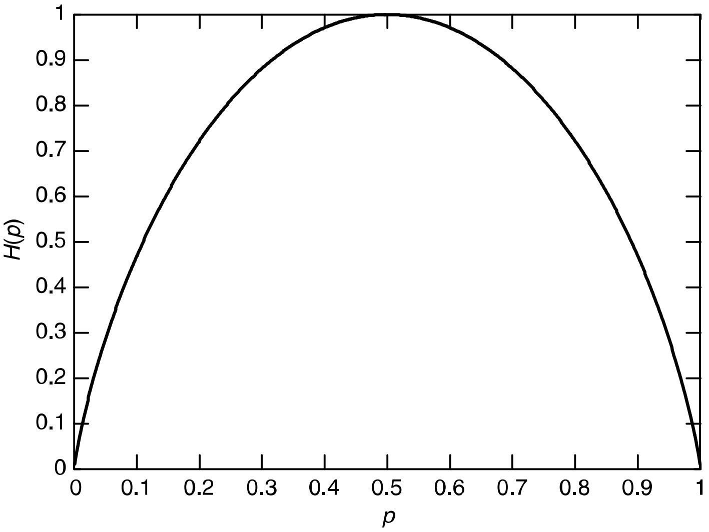

FIGURE 2.1. $H(p)$ vs. p.

Suppose that we wish to determine the value of X with the minimum number of binary questions. An efficient first question is "Is $X = a$?" This splits the probability in half. If the answer to the first question is no, the second question can be "Is $X = b$?" The third question can be "Is $X = c$?" The resulting expected number of binary questions required is 1.75. This turns out to be the minimum expected number of binary questions required to determine the value of X. In Chapter 5 we show that the minimum expected number of binary questions required to determine X lies between $H(X)$ and $H(X) + 1$.

2.2 JOINT ENTROPY AND CONDITIONAL ENTROPY

We defined the entropy of a single random variable in Section 2.1. We now extend the definition to a pair of random variables. There is nothing really new in this definition because (X, Y) can be considered to be a single vector-valued random variable.

Definition The *joint entropy* $H(X, Y)$ of a pair of discrete random variables (X, Y) with a joint distribution $p(x, y)$ is defined as

$$H(X, Y) = -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x, y) \log p(x, y), \tag{2.8}$$

which can also be expressed as

$$H(X, Y) = -E \log p(X, Y). \tag{2.9}$$

We also define the conditional entropy of a random variable given another as the expected value of the entropies of the conditional distributions, averaged over the conditioning random variable.

Definition If $(X, Y) \sim p(x, y)$, the *conditional entropy* $H(Y|X)$ is defined as

$$H(Y|X) = \sum_{x\in\mathcal{X}} p(x) H(Y|X = x) \tag{2.10}$$

$$= -\sum_{x\in\mathcal{X}} p(x) \sum_{y\in\mathcal{Y}} p(y|x) \log p(y|x) \tag{2.11}$$

$$= -\sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p(x, y) \log p(y|x) \tag{2.12}$$

$$= -E \log p(Y|X). \tag{2.13}$$

The naturalness of the definition of joint entropy and conditional entropy is exhibited by the fact that the entropy of a pair of random variables is the entropy of one plus the conditional entropy of the other. This is proved in the following theorem.

Theorem 2.2.1 (*Chain rule*)

$$H(X, Y) = H(X) + H(Y|X). \tag{2.14}$$

Proof

$$H(X, Y) = -\sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p(x, y) \log p(x, y) \tag{2.15}$$

$$= -\sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p(x, y) \log p(x) p(y|x) \tag{2.16}$$

$$= -\sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p(x, y) \log p(x) - \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p(x, y) \log p(y|x) \tag{2.17}$$

$$= -\sum_{x\in\mathcal{X}} p(x) \log p(x) - \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p(x, y) \log p(y|x) \tag{2.18}$$

$$= H(X) + H(Y|X). \tag{2.19}$$

Equivalently, we can write

$$\log p(X, Y) = \log p(X) + \log p(Y|X) \tag{2.20}$$

and take the expectation of both sides of the equation to obtain the theorem. □

Corollary

$$H(X, Y|Z) = H(X|Z) + H(Y|X, Z). \tag{2.21}$$

Proof: The proof follows along the same lines as the theorem. □

Example 2.2.1 Let (X, Y) have the following joint distribution:

$Y \backslash X$	1	2	3	4
1	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{32}$
2	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{32}$	$\frac{1}{32}$
3	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$
4	$\frac{1}{4}$	0	0	0

The marginal distribution of X is $(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8})$ and the marginal distribution of Y is $(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$, and hence $H(X) = \frac{7}{4}$ bits and $H(Y) = 2$ bits. Also,

$$H(X|Y) = \sum_{i=1}^{4} p(Y = i)H(X|Y = i) \tag{2.22}$$

$$= \frac{1}{4}H\left(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8}\right) + \frac{1}{4}H\left(\frac{1}{4}, \frac{1}{2}, \frac{1}{8}, \frac{1}{8}\right)$$

$$+ \frac{1}{4}H\left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right) + \frac{1}{4}H(1, 0, 0, 0) \tag{2.23}$$

$$= \frac{1}{4} \times \frac{7}{4} + \frac{1}{4} \times \frac{7}{4} + \frac{1}{4} \times 2 + \frac{1}{4} \times 0 \tag{2.24}$$

$$= \frac{11}{8} \text{ bits.} \tag{2.25}$$

Similarly, $H(Y|X) = \frac{13}{8}$ bits and $H(X, Y) = \frac{27}{8}$ bits.

Remark Note that $H(Y|X) \neq H(X|Y)$. However, $H(X) - H(X|Y) = H(Y) - H(Y|X)$, a property that we exploit later.

2.3 RELATIVE ENTROPY AND MUTUAL INFORMATION

The entropy of a random variable is a measure of the uncertainty of the random variable; it is a measure of the amount of information required on the average to describe the random variable. In this section we introduce two related concepts: relative entropy and mutual information.

The *relative entropy* is a measure of the distance between two distributions. In statistics, it arises as an expected logarithm of the likelihood ratio. The relative entropy $D(p||q)$ is a measure of the inefficiency of assuming that the distribution is q when the true distribution is p. For example, if we knew the true distribution p of the random variable, we could construct a code with average description length $H(p)$. If, instead, we used the code for a distribution q, we would need $H(p) + D(p||q)$ bits on the average to describe the random variable.

Definition The *relative entropy* or *Kullback–Leibler distance* between two probability mass functions $p(x)$ and $q(x)$ is defined as

$$D(p||q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)} \tag{2.26}$$

$$= E_p \log \frac{p(X)}{q(X)}. \tag{2.27}$$

In the above definition, we use the convention that $0 \log \frac{0}{0} = 0$ and the convention (based on continuity arguments) that $0 \log \frac{0}{q} = 0$ and $p \log \frac{p}{0} = \infty$. Thus, if there is any symbol $x \in \mathcal{X}$ such that $p(x) > 0$ and $q(x) = 0$, then $D(p||q) = \infty$.

We will soon show that relative entropy is always nonnegative and is zero if and only if $p = q$. However, it is not a true distance between distributions since it is not symmetric and does not satisfy the triangle inequality. Nonetheless, it is often useful to think of relative entropy as a "distance" between distributions.

We now introduce mutual information, which is a measure of the amount of information that one random variable contains about another random variable. It is the reduction in the uncertainty of one random variable due to the knowledge of the other.

Definition Consider two random variables X and Y with a joint probability mass function $p(x, y)$ and marginal probability mass functions $p(x)$ and $p(y)$. The *mutual information* $I(X; Y)$ is the relative entropy between the joint distribution and the product distribution $p(x)p(y)$:

$$I(X;Y) = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \tag{2.28}$$

$$= D(p(x,y)||p(x)p(y)) \tag{2.29}$$

$$= E_{p(x,y)}\log\frac{p(X,Y)}{p(X)p(Y)}. \tag{2.30}$$

In Chapter 8 we generalize this definition to continuous random variables, and in (8.54) to general random variables that could be a mixture of discrete and continuous random variables.

Example 2.3.1 Let $\mathcal{X} = \{0, 1\}$ and consider two distributions p and q on $\mathcal{X}$. Let $p(0) = 1 - r$, $p(1) = r$, and let $q(0) = 1 - s$, $q(1) = s$. Then

$$D(p||q) = (1-r)\log\frac{1-r}{1-s} + r\log\frac{r}{s} \tag{2.31}$$

and

$$D(q||p) = (1-s)\log\frac{1-s}{1-r} + s\log\frac{s}{r}. \tag{2.32}$$

If $r = s$, then $D(p||q) = D(q||p) = 0$. If $r = \frac{1}{2}$, $s = \frac{1}{4}$, we can calculate

$$D(p||q) = \frac{1}{2}\log\frac{\frac{1}{2}}{\frac{3}{4}} + \frac{1}{2}\log\frac{\frac{1}{2}}{\frac{1}{4}} = 1 - \frac{1}{2}\log 3 = 0.2075 \text{ bit}, \tag{2.33}$$

whereas

$$D(q||p) = \frac{3}{4}\log\frac{\frac{3}{4}}{\frac{1}{2}} + \frac{1}{4}\log\frac{\frac{1}{4}}{\frac{1}{2}} = \frac{3}{4}\log 3 - 1 = 0.1887 \text{ bit}. \tag{2.34}$$

Note that $D(p||q) \neq D(q||p)$ in general.

2.4 RELATIONSHIP BETWEEN ENTROPY AND MUTUAL INFORMATION

We can rewrite the definition of mutual information $I(X;Y)$ as

$$I(X;Y) = \sum_{x,y} p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \tag{2.35}$$

$$= \sum_{x,y} p(x, y) \log \frac{p(x|y)}{p(x)} \tag{2.36}$$

$$= -\sum_{x,y} p(x, y) \log p(x) + \sum_{x,y} p(x, y) \log p(x|y) \tag{2.37}$$

$$= -\sum_{x} p(x) \log p(x) - \left(-\sum_{x,y} p(x, y) \log p(x|y) \right) \tag{2.38}$$

$$= H(X) - H(X|Y). \tag{2.39}$$

Thus, the mutual information $I(X; Y)$ is the reduction in the uncertainty of X due to the knowledge of Y.

By symmetry, it also follows that

$$I(X; Y) = H(Y) - H(Y|X). \tag{2.40}$$

Thus, X says as much about Y as Y says about X.

Since $H(X, Y) = H(X) + H(Y|X)$, as shown in Section 2.2, we have

$$I(X; Y) = H(X) + H(Y) - H(X, Y). \tag{2.41}$$

Finally, we note that

$$I(X; X) = H(X) - H(X|X) = H(X). \tag{2.42}$$

Thus, the mutual information of a random variable with itself is the entropy of the random variable. This is the reason that entropy is sometimes referred to as *self-information*.

Collecting these results, we have the following theorem.

Theorem 2.4.1 (*Mutual information and entropy*)

$$I(X; Y) = H(X) - H(X|Y) \tag{2.43}$$

$$I(X; Y) = H(Y) - H(Y|X) \tag{2.44}$$

$$I(X; Y) = H(X) + H(Y) - H(X, Y) \tag{2.45}$$

$$I(X; Y) = I(Y; X) \tag{2.46}$$

$$I(X; X) = H(X). \tag{2.47}$$

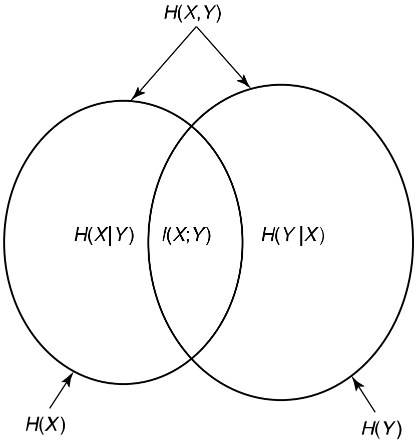

FIGURE 2.2. Relationship between entropy and mutual information.

The relationship between $H(X)$, $H(Y)$, $H(X,Y)$, $H(X|Y)$, $H(Y|X)$, and $I(X;Y)$ is expressed in a Venn diagram (Figure 2.2). Notice that the mutual information $I(X;Y)$ corresponds to the intersection of the information in X with the information in Y.

Example 2.4.1 For the joint distribution of Example 2.2.1, it is easy to calculate the mutual information $I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) = 0.375$ bit.

2.5 CHAIN RULES FOR ENTROPY, RELATIVE ENTROPY, AND MUTUAL INFORMATION

We now show that the entropy of a collection of random variables is the sum of the conditional entropies.

Theorem 2.5.1 (*Chain rule for entropy*) *Let $X_1, X_2, \ldots, X_n$ be drawn according to $p(x_1, x_2, \ldots, x_n)$. Then*

$$H(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} H(X_i|X_{i-1}, \ldots, X_1). \tag{2.48}$$

Proof: By repeated application of the two-variable expansion rule for entropies, we have

$$H(X_1, X_2) = H(X_1) + H(X_2|X_1), \tag{2.49}$$

$$H(X_1, X_2, X_3) = H(X_1) + H(X_2, X_3|X_1) \tag{2.50}$$

$$= H(X_1) + H(X_2|X_1) + H(X_3|X_2, X_1), \tag{2.51}$$

$$\vdots$$

$$H(X_1, X_2, \ldots, X_n) = H(X_1) + H(X_2|X_1) + \cdots + H(X_n|X_{n-1}, \ldots, X_1) \tag{2.52}$$

$$= \sum_{i=1}^{n} H(X_i|X_{i-1}, \ldots, X_1). \quad \square \tag{2.53}$$

Alternative Proof: We write $p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p(x_i|x_{i-1}, \ldots, x_1)$ and evaluate

$$H(X_1, X_2, \ldots, X_n)$$

$$= - \sum_{x_1, x_2, \ldots, x_n} p(x_1, x_2, \ldots, x_n) \log p(x_1, x_2, \ldots, x_n) \tag{2.54}$$

$$= - \sum_{x_1, x_2, \ldots, x_n} p(x_1, x_2, \ldots, x_n) \log \prod_{i=1}^{n} p(x_i|x_{i-1}, \ldots, x_1) \tag{2.55}$$

$$= - \sum_{x_1, x_2, \ldots, x_n} \sum_{i=1}^{n} p(x_1, x_2, \ldots, x_n) \log p(x_i|x_{i-1}, \ldots, x_1) \tag{2.56}$$

$$= - \sum_{i=1}^{n} \sum_{x_1, x_2, \ldots, x_n} p(x_1, x_2, \ldots, x_n) \log p(x_i|x_{i-1}, \ldots, x_1) \tag{2.57}$$

$$= - \sum_{i=1}^{n} \sum_{x_1, x_2, \ldots, x_i} p(x_1, x_2, \ldots, x_i) \log p(x_i|x_{i-1}, \ldots, x_1) \tag{2.58}$$

$$= \sum_{i=1}^{n} H(X_i|X_{i-1}, \ldots, X_1). \quad \square \tag{2.59}$$

We now define the conditional mutual information as the reduction in the uncertainty of X due to knowledge of Y when Z is given.

Definition The *conditional mutual information* of random variables X and Y given Z is defined by

$$I(X; Y|Z) = H(X|Z) - H(X|Y, Z) \tag{2.60}$$

$$= E_{p(x,y,z)} \log \frac{p(X, Y|Z)}{p(X|Z)p(Y|Z)}. \tag{2.61}$$

Mutual information also satisfies a chain rule.

Theorem 2.5.2 (*Chain rule for information*)

$$I(X_1, X_2, \ldots, X_n; Y) = \sum_{i=1}^{n} I(X_i; Y|X_{i-1}, X_{i-2}, \ldots, X_1). \tag{2.62}$$

Proof

$$\begin{aligned}
&I(X_1, X_2, \ldots, X_n; Y) \\
&\quad = H(X_1, X_2, \ldots, X_n) - H(X_1, X_2, \ldots, X_n|Y) && (2.63)\\
&\quad = \sum_{i=1}^{n} H(X_i|X_{i-1}, \ldots, X_1) - \sum_{i=1}^{n} H(X_i|X_{i-1}, \ldots, X_1, Y) \\
&\quad = \sum_{i=1}^{n} I(X_i; Y|X_1, X_2, \ldots, X_{i-1}). \quad \square && (2.64)
\end{aligned}$$

We define a conditional version of the relative entropy.

Definition For joint probability mass functions $p(x, y)$ and $q(x, y)$, the *conditional relative entropy* $D(p(y|x)||q(y|x))$ is the average of the relative entropies between the conditional probability mass functions $p(y|x)$ and $q(y|x)$ averaged over the probability mass function $p(x)$. More precisely,

$$\begin{aligned}
D(p(y|x)||q(y|x)) &= \sum_{x} p(x) \sum_{y} p(y|x) \log \frac{p(y|x)}{q(y|x)} && (2.65)\\
&= E_{p(x,y)} \log \frac{p(Y|X)}{q(Y|X)}. && (2.66)
\end{aligned}$$

The notation for conditional relative entropy is not explicit since it omits mention of the distribution $p(x)$ of the conditioning random variable. However, it is normally understood from the context.

The relative entropy between two joint distributions on a pair of random variables can be expanded as the sum of a relative entropy and a conditional relative entropy. The chain rule for relative entropy is used in Section 4.4 to prove a version of the second law of thermodynamics.

Theorem 2.5.3 (*Chain rule for relative entropy*)

$$D(p(x, y)||q(x, y)) = D(p(x)||q(x)) + D(p(y|x)||q(y|x)). \tag{2.67}$$

Proof

$$D(p(x,y)||q(x,y))$$

$$=\sum_x\sum_y p(x,y)\log\frac{p(x,y)}{q(x,y)} \tag{2.68}$$

$$=\sum_x\sum_y p(x,y)\log\frac{p(x)p(y|x)}{q(x)q(y|x)} \tag{2.69}$$

$$=\sum_x\sum_y p(x,y)\log\frac{p(x)}{q(x)}+\sum_x\sum_y p(x,y)\log\frac{p(y|x)}{q(y|x)} \tag{2.70}$$

$$= D(p(x)||q(x)) + D(p(y|x)||q(y|x)). \quad \square \tag{2.71}$$

2.6 JENSEN'S INEQUALITY AND ITS CONSEQUENCES

In this section we prove some simple properties of the quantities defined earlier. We begin with the properties of convex functions.

Definition A function $f(x)$ is said to be *convex* over an interval (a,b) if for every $x_1, x_2 \in (a,b)$ and $0 \le \lambda \le 1$,

$$f(\lambda x_1 + (1-\lambda)x_2) \le \lambda f(x_1) + (1-\lambda)f(x_2). \tag{2.72}$$

A function f is said to be *strictly convex* if equality holds only if $\lambda = 0$ or $\lambda = 1$.

Definition A function f is *concave* if $-f$ is convex. A function is convex if it always lies below any chord. A function is concave if it always lies above any chord.

Examples of convex functions include x^2, $|x|$, e^x, $x\log x$ (for $x \ge 0$), and so on. Examples of concave functions include $\log x$ and $\sqrt{x}$ for $x \ge 0$. Figure 2.3 shows some examples of convex and concave functions. Note that linear functions $ax+b$ are both convex and concave. Convexity underlies many of the basic properties of information-theoretic quantities such as entropy and mutual information. Before we prove some of these properties, we derive some simple results for convex functions.

Theorem 2.6.1 *If the function f has a second derivative that is nonnegative (positive) over an interval, the function is convex (strictly convex) over that interval.*

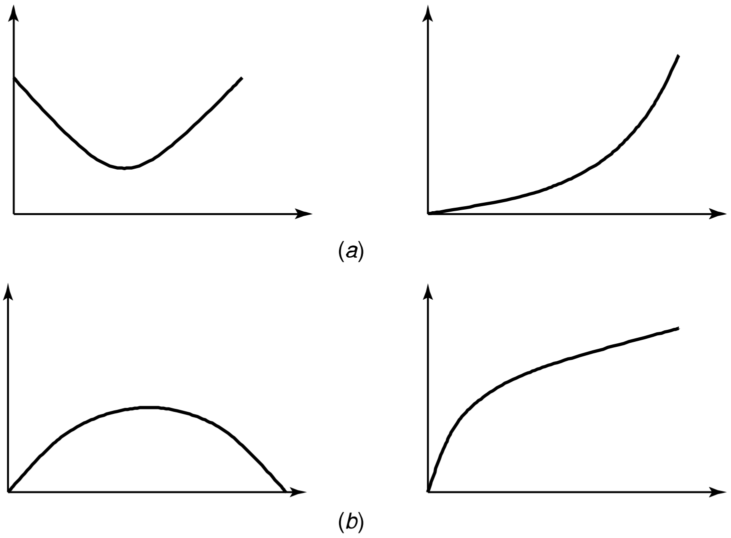

FIGURE 2.3. Examples of (*a*) convex and (*b*) concave functions.

Proof: We use the Taylor series expansion of the function around x_0:

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{f''(x^*)}{2}(x - x_0)^2, \tag{2.73}$$

where x^* lies between x_0 and x. By hypothesis, $f''(x^*) \geq 0$, and thus the last term is nonnegative for all x.

We let $x_0 = \lambda x_1 + (1 - \lambda)x_2$ and take $x = x_1$, to obtain

$$f(x_1) \geq f(x_0) + f'(x_0)((1 - \lambda)(x_1 - x_2)). \tag{2.74}$$

Similarly, taking $x = x_2$, we obtain

$$f(x_2) \geq f(x_0) + f'(x_0)(\lambda(x_2 - x_1)). \tag{2.75}$$

Multiplying (2.74) by λ and (2.75) by $1 - \lambda$ and adding, we obtain (2.72). The proof for strict convexity proceeds along the same lines. □

Theorem 2.6.1 allows us immediately to verify the strict convexity of x^2, e^x, and $x \log x$ for $x \geq 0$, and the strict concavity of $\log x$ and $\sqrt{x}$ for $x \geq 0$.

Let E denote expectation. Thus, $EX = \sum_{x \in \mathcal{X}} p(x)x$ in the discrete case and $EX = \int x f(x)\, dx$ in the continuous case.

The next inequality is one of the most widely used in mathematics and one that underlies many of the basic results in information theory.

Theorem 2.6.2 (*Jensen's inequality*) *If f is a convex function and X is a random variable,*

$$Ef(X) \geq f(EX). \tag{2.76}$$

Moreover, if f is strictly convex, the equality in (2.76) implies that $X = EX$ with probability 1 (i.e., X is a constant).

Proof: We prove this for discrete distributions by induction on the number of mass points. The proof of conditions for equality when f is strictly convex is left to the reader.

For a two-mass-point distribution, the inequality becomes

$$p_1 f(x_1) + p_2 f(x_2) \geq f(p_1 x_1 + p_2 x_2), \tag{2.77}$$

which follows directly from the definition of convex functions. Suppose that the theorem is true for distributions with $k-1$ mass points. Then writing $p'_i = p_i/(1-p_k)$ for $i = 1, 2, \ldots, k-1$, we have

$$\sum_{i=1}^{k} p_i f(x_i) = p_k f(x_k) + (1-p_k) \sum_{i=1}^{k-1} p'_i f(x_i) \tag{2.78}$$

$$\geq p_k f(x_k) + (1-p_k) f\left(\sum_{i=1}^{k-1} p'_i x_i\right) \tag{2.79}$$

$$\geq f\left(p_k x_k + (1-p_k) \sum_{i=1}^{k-1} p'_i x_i\right) \tag{2.80}$$

$$= f\left(\sum_{i=1}^{k} p_i x_i\right), \tag{2.81}$$

where the first inequality follows from the induction hypothesis and the second follows from the definition of convexity.

The proof can be extended to continuous distributions by continuity arguments. □

We now use these results to prove some of the properties of entropy and relative entropy. The following theorem is of fundamental importance.

Theorem 2.6.3 (*Information inequality*) *Let $p(x), q(x), x \in \mathcal{X}$, be two probability mass functions. Then*

$$D(p||q) \geq 0 \tag{2.82}$$

with equality if and only if $p(x) = q(x)$ for all x.

Proof: Let $A = \{x : p(x) > 0\}$ be the support set of $p(x)$. Then

$$\begin{aligned}
-D(p||q) &= -\sum_{x \in A} p(x) \log \frac{p(x)}{q(x)} && (2.83)\\
&= \sum_{x \in A} p(x) \log \frac{q(x)}{p(x)} && (2.84)\\
&\leq \log \sum_{x \in A} p(x) \frac{q(x)}{p(x)} && (2.85)\\
&= \log \sum_{x \in A} q(x) && (2.86)\\
&\leq \log \sum_{x \in \mathcal{X}} q(x) && (2.87)\\
&= \log 1 && (2.88)\\
&= 0, && (2.89)
\end{aligned}$$

where (2.85) follows from Jensen's inequality. Since $\log t$ is a strictly concave function of t, we have equality in (2.85) if and only if $q(x)/p(x)$ is constant everywhere [i.e., $q(x) = cp(x)$ for all x]. Thus, $\sum_{x \in A} q(x) = c \sum_{x \in A} p(x) = c$. We have equality in (2.87) only if $\sum_{x \in A} q(x) = \sum_{x \in \mathcal{X}} q(x) = 1$, which implies that $c = 1$. Hence, we have $D(p||q) = 0$ if and only if $p(x) = q(x)$ for all x. □

Corollary (*Nonnegativity of mutual information*) *For any two random variables, X, Y,*

$$I(X; Y) \geq 0, \tag{2.90}$$

with equality if and only if X and Y are independent.

Proof: $I(X; Y) = D(p(x, y)||p(x)p(y)) \geq 0$, with equality if and only if $p(x, y) = p(x)p(y)$ (i.e., X and Y are independent). □

Corollary

$$D(p(y|x)||q(y|x)) \geq 0, \tag{2.91}$$

with equality if and only if $p(y|x) = q(y|x)$ *for all* y *and* x *such that* $p(x) > 0$.

Corollary

$$I(X;Y|Z) \geq 0, \tag{2.92}$$

with equality if and only if X and Y are conditionally independent given Z.

We now show that the uniform distribution over the range $\mathcal{X}$ is the maximum entropy distribution over this range. It follows that any random variable with this range has an entropy no greater than $\log|\mathcal{X}|$.

Theorem 2.6.4 $H(X) \leq \log|\mathcal{X}|$, *where* $|\mathcal{X}|$ *denotes the number of elements in the range of* X, *with equality if and only* X *has a uniform distribution over* $\mathcal{X}$.

Proof: Let $u(x) = \frac{1}{|\mathcal{X}|}$ be the uniform probability mass function over $\mathcal{X}$, and let $p(x)$ be the probability mass function for X. Then

$$D(p \parallel u) = \sum p(x) \log \frac{p(x)}{u(x)} = \log|\mathcal{X}| - H(X). \tag{2.93}$$

Hence by the nonnegativity of relative entropy,

$$0 \leq D(p \parallel u) = \log|\mathcal{X}| - H(X). \quad \square \tag{2.94}$$

Theorem 2.6.5 (*Conditioning reduces entropy*)(*Information can't hurt*)

$$H(X|Y) \leq H(X) \tag{2.95}$$

with equality if and only if X *and* Y *are independent.*

Proof: $0 \leq I(X;Y) = H(X) - H(X|Y)$. □

Intuitively, the theorem says that knowing another random variable Y can only reduce the uncertainty in X. Note that this is true only on the average. Specifically, $H(X|Y = y)$ may be greater than or less than or equal to $H(X)$, but on the average $H(X|Y) = \sum_y p(y)H(X|Y = y) \leq H(X)$. For example, in a court case, specific new evidence might increase uncertainty, but on the average evidence decreases uncertainty.

Example 2.6.1 Let (X, Y) have the following joint distribution:

$Y \backslash X$	1	2
1	0	$\frac{3}{4}$
2	$\frac{1}{8}$	$\frac{1}{8}$

Then $H(X) = H(\frac{1}{8}, \frac{7}{8}) = 0.544$ bit, $H(X|Y = 1) = 0$ bits, and $H(X|Y = 2) = 1$ bit. We calculate $H(X|Y) = \frac{3}{4}H(X|Y = 1) + \frac{1}{4}$ $H(X|Y = 2) = 0.25$ bit. Thus, the uncertainty in X is increased if $Y = 2$ is observed and decreased if $Y = 1$ is observed, but uncertainty decreases on the average.

Theorem 2.6.6 (*Independence bound on entropy*) *Let $X_1, X_2, \ldots, X_n$ be drawn according to $p(x_1, x_2, \ldots, x_n)$. Then*

$$H(X_1, X_2, \ldots, X_n) \le \sum_{i=1}^{n} H(X_i) \tag{2.96}$$

with equality if and only if the X_i are independent.

Proof: By the chain rule for entropies,

$$H(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} H(X_i|X_{i-1}, \ldots, X_1) \tag{2.97}$$

$$\le \sum_{i=1}^{n} H(X_i), \tag{2.98}$$

where the inequality follows directly from Theorem 2.6.5. We have equality if and only if X_i is independent of $X_{i-1}, \ldots, X_1$ for all i (i.e., if and only if the X_i's are independent). □

2.7 LOG SUM INEQUALITY AND ITS APPLICATIONS

We now prove a simple consequence of the concavity of the logarithm, which will be used to prove some concavity results for the entropy.

Theorem 2.7.1 (*Log sum inequality*) *For nonnegative numbers,* $a_1, a_2, \ldots, a_n$ *and* $b_1, b_2, \ldots, b_n$,

$$\sum_{i=1}^{n} a_i \log \frac{a_i}{b_i} \geq \left(\sum_{i=1}^{n} a_i\right) \log \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i} \tag{2.99}$$

with equality if and only if $\frac{a_i}{b_i} = const.$

We again use the convention that $0 \log 0 = 0$, $a \log \frac{a}{0} = \infty$ if $a > 0$ and $0 \log \frac{0}{0} = 0$. These follow easily from continuity.

Proof: Assume without loss of generality that $a_i > 0$ and $b_i > 0$. The function $f(t) = t \log t$ is strictly convex, since $f''(t) = \frac{1}{t} \log e > 0$ for all positive t. Hence by Jensen's inequality, we have

$$\sum \alpha_i f(t_i) \geq f\left(\sum \alpha_i t_i\right) \tag{2.100}$$

for $\alpha_i \geq 0$, $\sum_i \alpha_i = 1$. Setting $\alpha_i = \frac{b_i}{\sum_{j=1}^{n} b_j}$ and $t_i = \frac{a_i}{b_i}$, we obtain

$$\sum \frac{a_i}{\sum b_j} \log \frac{a_i}{b_i} \geq \sum \frac{a_i}{\sum b_j} \log \sum \frac{a_i}{\sum b_j}, \tag{2.101}$$

which is the log sum inequality. □

We now use the log sum inequality to prove various convexity results. We begin by reproving Theorem 2.6.3, which states that $D(p||q) \geq 0$ with equality if and only if $p(x) = q(x)$. By the log sum inequality,

$$D(p||q) = \sum p(x) \log \frac{p(x)}{q(x)} \tag{2.102}$$

$$\geq \left(\sum p(x)\right) \log \sum p(x) \Big/ \sum q(x) \tag{2.103}$$

$$= 1 \log \frac{1}{1} = 0 \tag{2.104}$$

with equality if and only if $\frac{p(x)}{q(x)} = c$. Since both p and q are probability mass functions, $c = 1$, and hence we have $D(p||q) = 0$ if and only if $p(x) = q(x)$ for all x.

Theorem 2.7.2 (*Convexity of relative entropy*) *$D(p||q)$ is convex in the pair (p, q); that is, if (p_1, q_1) and (p_2, q_2) are two pairs of probability mass functions, then*

$$D(\lambda p_1 + (1-\lambda)p_2||\lambda q_1 + (1-\lambda)q_2) \le \lambda D(p_1||q_1) + (1-\lambda)D(p_2||q_2) \tag{2.105}$$

for all $0 \le \lambda \le 1$.

Proof: We apply the log sum inequality to a term on the left-hand side of (2.105):

$$\begin{aligned}(\lambda p_1(x) + (1-\lambda)p_2(x))\log\frac{\lambda p_1(x) + (1-\lambda)p_2(x)}{\lambda q_1(x) + (1-\lambda)q_2(x)}& \\ \le \lambda p_1(x)\log\frac{\lambda p_1(x)}{\lambda q_1(x)} + (1-\lambda)p_2(x)\log\frac{(1-\lambda)p_2(x)}{(1-\lambda)q_2(x)}.&\end{aligned} \tag{2.106}$$

Summing this over all x, we obtain the desired property. □

Theorem 2.7.3 (*Concavity of entropy*) *$H(p)$ is a concave function of p.*

Proof

$$H(p) = \log|\mathcal{X}| - D(p||u), \tag{2.107}$$

where u is the uniform distribution on $|\mathcal{X}|$ outcomes. The concavity of H then follows directly from the convexity of D. □

Alternative Proof: Let X_1 be a random variable with distribution p_1, taking on values in a set A. Let X_2 be another random variable with distribution p_2 on the same set. Let

$$\theta = \begin{cases} 1 & \text{with probability } \lambda, \\ 2 & \text{with probability } 1-\lambda. \end{cases} \tag{2.108}$$

Let $Z = X_\theta$. Then the distribution of Z is $\lambda p_1 + (1-\lambda)p_2$. Now since conditioning reduces entropy, we have

$$H(Z) \ge H(Z|\theta), \tag{2.109}$$

or equivalently,

$$H(\lambda p_1 + (1-\lambda)p_2) \ge \lambda H(p_1) + (1-\lambda)H(p_2), \tag{2.110}$$

which proves the concavity of the entropy as a function of the distribution. □

One of the consequences of the concavity of entropy is that mixing two gases of equal entropy results in a gas with higher entropy.

Theorem 2.7.4 *Let $(X, Y) \sim p(x, y) = p(x)p(y|x)$. The mutual information $I(X; Y)$ is a concave function of $p(x)$ for fixed $p(y|x)$ and a convex function of $p(y|x)$ for fixed $p(x)$.*

Proof: To prove the first part, we expand the mutual information

$$I(X; Y) = H(Y) - H(Y|X) = H(Y) - \sum_x p(x)H(Y|X = x). \tag{2.111}$$

If $p(y|x)$ is fixed, then $p(y)$ is a linear function of $p(x)$. Hence $H(Y)$, which is a concave function of $p(y)$, is a concave function of $p(x)$. The second term is a linear function of $p(x)$. Hence, the difference is a concave function of $p(x)$.

To prove the second part, we fix $p(x)$ and consider two different conditional distributions $p_1(y|x)$ and $p_2(y|x)$. The corresponding joint distributions are $p_1(x, y) = p(x)p_1(y|x)$ and $p_2(x, y) = p(x)p_2(y|x)$, and their respective marginals are $p(x)$, $p_1(y)$ and $p(x)$, $p_2(y)$. Consider a conditional distribution

$$p_\lambda(y|x) = \lambda p_1(y|x) + (1 - \lambda)p_2(y|x), \tag{2.112}$$

which is a mixture of $p_1(y|x)$ and $p_2(y|x)$ where $0 \le \lambda \le 1$. The corresponding joint distribution is also a mixture of the corresponding joint distributions,

$$p_\lambda(x, y) = \lambda p_1(x, y) + (1 - \lambda)p_2(x, y), \tag{2.113}$$

and the distribution of Y is also a mixture,

$$p_\lambda(y) = \lambda p_1(y) + (1 - \lambda)p_2(y). \tag{2.114}$$

Hence if we let $q_\lambda(x, y) = p(x)p_\lambda(y)$ be the product of the marginal distributions, we have

$$q_\lambda(x, y) = \lambda q_1(x, y) + (1 - \lambda)q_2(x, y). \tag{2.115}$$

Since the mutual information is the relative entropy between the joint distribution and the product of the marginals,

$$I(X; Y) = D(p_\lambda(x, y)||q_\lambda(x, y)), \tag{2.116}$$

and relative entropy $D(p||q)$ is a convex function of (p, q), it follows that the mutual information is a convex function of the conditional distribution.
□

2.8 DATA-PROCESSING INEQUALITY

The data-processing inequality can be used to show that no clever manipulation of the data can improve the inferences that can be made from the data.

Definition Random variables X, Y, Z are said to *form a Markov chain in that order* (denoted by $X \rightarrow Y \rightarrow Z$) if the conditional distribution of Z depends only on Y and is conditionally independent of X. Specifically, X, Y, and Z form a Markov chain $X \rightarrow Y \rightarrow Z$ if the joint probability mass function can be written as

$$p(x, y, z) = p(x)p(y|x)p(z|y). \tag{2.117}$$

Some simple consequences are as follows:

- $X \rightarrow Y \rightarrow Z$ if and only if X and Z are conditionally independent given Y. Markovity implies conditional independence because

  $$p(x, z|y) = \frac{p(x, y, z)}{p(y)} = \frac{p(x, y)p(z|y)}{p(y)} = p(x|y)p(z|y). \tag{2.118}$$

  This is the characterization of Markov chains that can be extended to define Markov fields, which are n-dimensional random processes in which the interior and exterior are independent given the values on the boundary.
- $X \rightarrow Y \rightarrow Z$ implies that $Z \rightarrow Y \rightarrow X$. Thus, the condition is sometimes written $X \leftrightarrow Y \leftrightarrow Z$.
- If $Z = f(Y)$, then $X \rightarrow Y \rightarrow Z$.

We can now prove an important and useful theorem demonstrating that no processing of Y, deterministic or random, can increase the information that Y contains about X.

Theorem 2.8.1 (*Data-processing inequality*) *If $X \rightarrow Y \rightarrow Z$, then $I(X; Y) \geq I(X; Z)$.*

Proof: By the chain rule, we can expand mutual information in two different ways:

$$I(X; Y, Z) = I(X; Z) + I(X; Y|Z) \tag{2.119}$$

$$= I(X; Y) + I(X; Z|Y). \tag{2.120}$$

Since X and Z are conditionally independent given Y, we have $I(X; Z|Y) = 0$. Since $I(X; Y|Z) \geq 0$, we have

$$I(X; Y) \geq I(X; Z). \tag{2.121}$$

We have equality if and only if $I(X; Y|Z) = 0$ (i.e., $X \to Z \to Y$ forms a Markov chain). Similarly, one can prove that $I(Y; Z) \geq I(X; Z)$. □

Corollary *In particular, if $Z = g(Y)$, we have $I(X; Y) \geq I(X; g(Y))$.*

Proof: $X \to Y \to g(Y)$ forms a Markov chain. □

Thus functions of the data Y cannot increase the information about X.

Corollary *If $X \to Y \to Z$, then $I(X; Y|Z) \leq I(X; Y)$.*

Proof: We note in (2.119) and (2.120) that $I(X; Z|Y) = 0$, by Markovity, and $I(X; Z) \geq 0$. Thus,

$$I(X; Y|Z) \leq I(X; Y). \quad \square \tag{2.122}$$

Thus, the dependence of X and Y is decreased (or remains unchanged) by the observation of a "downstream" random variable Z. Note that it is also possible that $I(X; Y|Z) > I(X; Y)$ when X, Y, and Z do not form a Markov chain. For example, let X and Y be independent fair binary random variables, and let $Z = X + Y$. Then $I(X; Y) = 0$, but $I(X; Y|Z) = H(X|Z) - H(X|Y, Z) = H(X|Z) = P(Z = 1)H(X|Z = 1) = \frac{1}{2}$ bit.

2.9 SUFFICIENT STATISTICS

This section is a sidelight showing the power of the data-processing inequality in clarifying an important idea in statistics. Suppose that we have a family of probability mass functions $\{f_\theta(x)\}$ indexed by θ, and let X be a sample from a distribution in this family. Let $T(X)$ be any statistic (function of the sample) like the sample mean or sample variance. Then $\theta \to X \to T(X)$, and by the data-processing inequality, we have

$$I(\theta; T(X)) \leq I(\theta; X) \tag{2.123}$$

for any distribution on θ. However, if equality holds, no information is lost.

A statistic $T(X)$ is called sufficient for θ if it contains all the information in X about θ.

Definition A function $T(X)$ is said to be a *sufficient statistic* relative to the family $\{f_\theta(x)\}$ if X is independent of θ given $T(X)$ for any distribution on θ[i.e., $\theta \to T(X) \to X$ forms a Markov chain].

This is the same as the condition for equality in the data-processing inequality,

$$I(\theta; X) = I(\theta; T(X)) \tag{2.124}$$

for all distributions on θ. Hence sufficient statistics preserve mutual information and conversely.

Here are some examples of sufficient statistics:

1. Let $X_1, X_2, \ldots, X_n$, $X_i \in \{0, 1\}$, be an independent and identically distributed (i.i.d.) sequence of coin tosses of a coin with unknown parameter $\theta = \Pr(X_i = 1)$. Given n, the number of 1's is a sufficient statistic for θ. Here $T(X_1, X_2, \ldots, X_n) = \sum_{i=1}^n X_i$. In fact, we can show that given T, all sequences having that many 1's are equally likely and independent of the parameter θ. Specifically,

$$\Pr\left\{(X_1, X_2, \ldots, X_n) = (x_1, x_2, \ldots, x_n) \middle| \sum_{i=1}^n X_i = k\right\}$$
$$= \begin{cases} \frac{1}{\binom{n}{k}} & \text{if } \sum x_i = k, \\ 0 & \text{otherwise.} \end{cases} \tag{2.125}$$

Thus, $\theta \to \sum X_i \to (X_1, X_2, \ldots, X_n)$ forms a Markov chain, and T is a sufficient statistic for θ.

The next two examples involve probability densities instead of probability mass functions, but the theory still applies. We define entropy and mutual information for continuous random variables in Chapter 8.

2. If X is normally distributed with mean θ and variance 1; that is, if

$$f_\theta(x) = \frac{1}{\sqrt{2\pi}} e^{-(x-\theta)^2/2} = \mathcal{N}(\theta, 1), \tag{2.126}$$

and $X_1, X_2, \ldots, X_n$ are drawn independently according to this distribution, a sufficient statistic for θ is the sample mean $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$. It can be verified that the conditional distribution of $X_1, X_2, \ldots, X_n$, conditioned on $\overline{X}_n$ and n does not depend on θ.

3. If $f_\theta = \text{Uniform}(\theta, \theta + 1)$, a sufficient statistic for θ is

$$\begin{aligned} T&(X_1, X_2, \ldots, X_n) \\ &= (\max\{X_1, X_2, \ldots, X_n\}, \min\{X_1, X_2, \ldots, X_n\}). \qquad (2.127) \end{aligned}$$

   The proof of this is slightly more complicated, but again one can show that the distribution of the data is independent of the parameter given the statistic T.

The minimal sufficient statistic is a sufficient statistic that is a function of all other sufficient statistics.

Definition A statistic $T(X)$ is a *minimal sufficient statistic* relative to $\{f_\theta(x)\}$ if it is a function of every other sufficient statistic U. Interpreting this in terms of the data-processing inequality, this implies that

$$\theta \to T(X) \to U(X) \to X. \qquad (2.128)$$

Hence, a minimal sufficient statistic maximally compresses the information about θ in the sample. Other sufficient statistics may contain additional irrelevant information. For example, for a normal distribution with mean θ, the pair of functions giving the mean of all odd samples and the mean of all even samples is a sufficient statistic, but not a minimal sufficient statistic. In the preceding examples, the sufficient statistics are also minimal.

2.10 FANO'S INEQUALITY

Suppose that we know a random variable Y and we wish to guess the value of a correlated random variable X. Fano's inequality relates the probability of error in guessing the random variable X to its conditional entropy $H(X|Y)$. It will be crucial in proving the converse to Shannon's channel capacity theorem in Chapter 7. From Problem 2.5 we know that the conditional entropy of a random variable X given another random variable Y is zero if and only if X is a function of Y. Hence we can estimate X from Y with zero probability of error if and only if $H(X|Y) = 0$.

Extending this argument, we expect to be able to estimate X with a low probability of error only if the conditional entropy $H(X|Y)$ is small. Fano's inequality quantifies this idea. Suppose that we wish to estimate a random variable X with a distribution $p(x)$. We observe a random variable Y that is related to X by the conditional distribution $p(y|x)$. From Y, we calculate a function $g(Y) = \hat{X}$, where $\hat{X}$ is an estimate of X and takes on values in $\hat{\mathcal{X}}$. We will not restrict the alphabet $\hat{\mathcal{X}}$ to be equal to $\mathcal{X}$, and we will also allow the function $g(Y)$ to be random. We wish to bound the probability that $\hat{X} \neq X$. We observe that $X \to Y \to \hat{X}$ forms a Markov chain. Define the probability of error

$$P_e = \Pr\left\{\hat{X} \neq X\right\}. \tag{2.129}$$

Theorem 2.10.1 (*Fano's Inequality*) *For any estimator $\hat{X}$ such that $X \to Y \to \hat{X}$, with $P_e = \Pr(X \neq \hat{X})$, we have*

$$H(P_e) + P_e \log|\mathcal{X}| \geq H(X|\hat{X}) \geq H(X|Y). \tag{2.130}$$

This inequality can be weakened to

$$1 + P_e \log|\mathcal{X}| \geq H(X|Y) \tag{2.131}$$

or

$$P_e \geq \frac{H(X|Y) - 1}{\log|\mathcal{X}|}. \tag{2.132}$$

Remark Note from (2.130) that $P_e = 0$ implies that $H(X|Y) = 0$, as intuition suggests.

Proof: We first ignore the role of Y and prove the first inequality in (2.130). We will then use the data-processing inequality to prove the more traditional form of Fano's inequality, given by the second inequality in (2.130). Define an error random variable,

$$E = \begin{cases} 1 & \text{if } \hat{X} \neq X, \\ 0 & \text{if } \hat{X} = X. \end{cases} \tag{2.133}$$

Then, using the chain rule for entropies to expand $H(E, X|\hat{X})$ in two different ways, we have

$$H(E, X|\hat{X}) = H(X|\hat{X}) + \underbrace{H(E|X, \hat{X})}_{=0} \tag{2.134}$$

$$= \underbrace{H(E|\hat{X})}_{\leq H(P_e)} + \underbrace{H(X|E, \hat{X})}_{\leq P_e \log|\mathcal{X}|}. \tag{2.135}$$

Since conditioning reduces entropy, $H(E|\hat{X}) \leq H(E) = H(P_e)$. Now since E is a function of X and $\hat{X}$, the conditional entropy $H(E|X, \hat{X})$ is equal to 0. Also, since E is a binary-valued random variable, $H(E) = H(P_e)$. The remaining term, $H(X|E, \hat{X})$, can be bounded as follows:

$$\begin{aligned} H(X|E, \hat{X}) &= \Pr(E = 0)H(X|\hat{X}, E = 0) + \Pr(E = 1)H(X|\hat{X}, E = 1) \\ &\le (1 - P_e)0 + P_e \log |\mathcal{X}|, \end{aligned} \tag{2.136}$$

since given $E = 0$, $X = \hat{X}$, and given $E = 1$, we can upper bound the conditional entropy by the log of the number of possible outcomes. Combining these results, we obtain

$$H(P_e) + P_e \log |\mathcal{X}| \ge H(X|\hat{X}). \tag{2.137}$$

By the data-processing inequality, we have $I(X; \hat{X}) \le I(X; Y)$ since $X \to Y \to \hat{X}$ is a Markov chain, and therefore $H(X|\hat{X}) \ge H(X|Y)$. Thus, we have

$$H(P_e) + P_e \log |\mathcal{X}| \ge H(X|\hat{X}) \ge H(X|Y). \quad \Box \tag{2.138}$$

Corollary *For any two random variables X and Y, let $p = \Pr(X \neq Y)$.*

$$H(p) + p \log |\mathcal{X}| \ge H(X|Y). \tag{2.139}$$

Proof: Let $\hat{X} = Y$ in Fano's inequality. □

For any two random variables X and Y, if the estimator $g(Y)$ takes values in the set $\mathcal{X}$, we can strengthen the inequality slightly by replacing $\log |\mathcal{X}|$ with $\log(|\mathcal{X}| - 1)$.

Corollary *Let $P_e = \Pr(X \neq \hat{X})$, and let $\hat{X} : \mathcal{Y} \to \mathcal{X}$; then*

$$H(P_e) + P_e \log(|\mathcal{X}| - 1) \ge H(X|Y). \tag{2.140}$$

Proof: The proof of the theorem goes through without change, except that

$$H(X|E, \hat{X}) = \Pr(E = 0)H(X|\hat{X}, E = 0) + \Pr(E = 1)H(X|\hat{X}, E = 1) \tag{2.141}$$

$$\le (1 - P_e)0 + P_e \log(|\mathcal{X}| - 1), \tag{2.142}$$

since given $E = 0$, $X = \hat{X}$, and given $E = 1$, the range of possible X outcomes is $|\mathcal{X}| - 1$, we can upper bound the conditional entropy by the $\log(|\mathcal{X}| - 1)$, the logarithm of the number of possible outcomes. Substituting this provides us with the stronger inequality. □

Remark Suppose that there is no knowledge of Y. Thus, X must be guessed without any information. Let $X \in \{1, 2, \ldots, m\}$ and $p_1 \geq p_2 \geq \cdots \geq p_m$. Then the best guess of X is $\hat{X} = 1$ and the resulting probability of error is $P_e = 1 - p_1$. Fano's inequality becomes

$$H(P_e) + P_e \log(m-1) \geq H(X). \tag{2.143}$$

The probability mass function

$$(p_1, p_2, \ldots, p_m) = \left(1 - P_e, \frac{P_e}{m-1}, \ldots, \frac{P_e}{m-1}\right) \tag{2.144}$$

achieves this bound with equality. Thus, Fano's inequality is sharp.

While we are at it, let us introduce a new inequality relating probability of error and entropy. Let X and X' by two independent identically distributed random variables with entropy $H(X)$. The probability at $X = X'$ is given by

$$\Pr(X = X') = \sum_x p^2(x). \tag{2.145}$$

We have the following inequality:

Lemma 2.10.1 *If X and X' are i.i.d. with entropy $H(X)$,*

$$\Pr(X = X') \geq 2^{-H(X)}, \tag{2.146}$$

with equality if and only if X has a uniform distribution.

Proof: Suppose that $X \sim p(x)$. By Jensen's inequality, we have

$$2^{E \log p(X)} \leq E 2^{\log p(X)}, \tag{2.147}$$

which implies that

$$2^{-H(X)} = 2^{\sum p(x) \log p(x)} \leq \sum p(x) 2^{\log p(x)} = \sum p^2(x). \quad \square \tag{2.148}$$

Corollary *Let X, X' be independent with $X \sim p(x)$, $X' \sim r(x)$, $x, x' \in \mathcal{X}$. Then*

$$\Pr(X = X') \geq 2^{-H(p) - D(p||r)}, \tag{2.149}$$

$$\Pr(X = X') \geq 2^{-H(r) - D(r||p)}. \tag{2.150}$$

Proof: We have

$$\begin{aligned} 2^{-H(p)-D(p\|r)} &= 2^{\sum p(x)\log p(x)+\sum p(x)\log\frac{r(x)}{p(x)}} && (2.151)\\ &= 2^{\sum p(x)\log r(x)} && (2.152)\\ &\le \sum p(x)2^{\log r(x)} && (2.153)\\ &= \sum p(x)r(x) && (2.154)\\ &= \Pr(X = X'), && (2.155) \end{aligned}$$

where the inequality follows from Jensen's inequality and the convexity of the function $f(y) = 2^y$. □

The following telegraphic summary omits qualifying conditions.

SUMMARY

Definition The *entropy* $H(X)$ of a discrete random variable X is defined by

$$H(X) = -\sum_{x\in\mathcal{X}} p(x)\log p(x). \qquad (2.156)$$

Properties of *H*

1. $H(X) \ge 0$.
2. $H_b(X) = (\log_b a)H_a(X)$.
3. (Conditioning reduces entropy) For any two random variables, X and Y, we have

   $$H(X|Y) \le H(X) \qquad (2.157)$$

   with equality if and only if X and Y are independent.
4. $H(X_1, X_2, \ldots, X_n) \le \sum_{i=1}^{n} H(X_i)$, with equality if and only if the X_i are independent.
5. $H(X) \le \log|\mathcal{X}|$, with equality if and only if X is distributed uniformly over $\mathcal{X}$.
6. $H(p)$ is concave in p.

Definition The *relative entropy* $D(p \parallel q)$ of the probability mass function p with respect to the probability mass function q is defined by

$$D(p \parallel q) = \sum\nolimits_x p(x) \log \frac{p(x)}{q(x)}. \tag{2.158}$$

Definition The *mutual information* between two random variables X and Y is defined as

$$I(X;Y) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}. \tag{2.159}$$

Alternative expressions

$$H(X) = E_p \log \frac{1}{p(X)}, \tag{2.160}$$

$$H(X,Y) = E_p \log \frac{1}{p(X,Y)}, \tag{2.161}$$

$$H(X|Y) = E_p \log \frac{1}{p(X|Y)}, \tag{2.162}$$

$$I(X;Y) = E_p \log \frac{p(X,Y)}{p(X)p(Y)}, \tag{2.163}$$

$$D(p||q) = E_p \log \frac{p(X)}{q(X)}. \tag{2.164}$$

Properties of D and I

1. $I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) = H(X) + H(Y) - H(X,Y)$.
2. $D(p \parallel q) \geq 0$ with equality if and only if $p(x) = q(x)$, for all $x \in \mathcal{X}$.
3. $I(X;Y) = D(p(x,y)||p(x)p(y)) \geq 0$, with equality if and only if $p(x,y) = p(x)p(y)$ (i.e., X and Y are independent).
4. If $|\mathcal{X}| = m$, and u is the uniform distribution over $\mathcal{X}$, then $D(p \parallel u) = \log m - H(p)$.
5. $D(p||q)$ is convex in the pair (p,q).

Chain rules

Entropy: $H(X_1, X_2, \ldots, X_n) = \sum_{i=1}^n H(X_i|X_{i-1}, \ldots, X_1)$.

Mutual information:
$I(X_1, X_2, \ldots, X_n; Y) = \sum_{i=1}^n I(X_i; Y|X_1, X_2, \ldots, X_{i-1})$.

Relative entropy:
$D(p(x, y)||q(x, y)) = D(p(x)||q(x)) + D(p(y|x)||q(y|x))$.

Jensen's inequality. If f is a convex function, then $Ef(X) \geq f(EX)$.

Log sum inequality. For n positive numbers, $a_1, a_2, \ldots, a_n$ and $b_1, b_2, \ldots, b_n$,

$$\sum_{i=1}^{n} a_i \log \frac{a_i}{b_i} \geq \left(\sum_{i=1}^{n} a_i\right) \log \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i} \tag{2.165}$$

with equality if and only if $\frac{a_i}{b_i}$ = constant.

Data-processing inequality. If $X \to Y \to Z$ forms a Markov chain, $I(X; Y) \geq I(X; Z)$.

Sufficient statistic. $T(X)$ is sufficient relative to $\{f_\theta(x)\}$ if and only if $I(\theta; X) = I(\theta; T(X))$ for all distributions on θ.

Fano's inequality. Let $P_e = \Pr\{\hat{X}(Y) \neq X\}$. Then

$$H(P_e) + P_e \log |\mathcal{X}| \geq H(X|Y). \tag{2.166}$$

Inequality. If X and X' are independent and identically distributed, then

$$\Pr(X = X') \geq 2^{-H(X)}, \tag{2.167}$$

PROBLEMS

2.1 *Coin flips*. A fair coin is flipped until the first head occurs. Let X denote the number of flips required.

(a) Find the entropy $H(X)$ in bits. The following expressions may be useful:

$$\sum_{n=0}^{\infty} r^n = \frac{1}{1-r}, \qquad \sum_{n=0}^{\infty} nr^n = \frac{r}{(1-r)^2}.$$

(b) A random variable X is drawn according to this distribution. Find an "efficient" sequence of yes–no questions of the form, "Is X contained in the set S?" Compare $H(X)$ to the expected number of questions required to determine X.

2.2 *Entropy of functions*. Let X be a random variable taking on a finite number of values. What is the (general) inequality relationship of $H(X)$ and $H(Y)$ if

(a) $Y = 2^X$?

(b) $Y = \cos X$?

2.3 *Minimum entropy*. What is the minimum value of $H(p_1, \ldots, p_n) = H(\mathbf{p})$ as $\mathbf{p}$ ranges over the set of n-dimensional probability vectors? Find all $\mathbf{p}$'s that achieve this minimum.

2.4 *Entropy of functions of a random variable*. Let X be a discrete random variable. Show that the entropy of a function of X is less than or equal to the entropy of X by justifying the following steps:

$$H(X, g(X)) \overset{(a)}{=} H(X) + H(g(X) \mid X) \tag{2.168}$$

$$\overset{(b)}{=} H(X), \tag{2.169}$$

$$H(X, g(X)) \overset{(c)}{=} H(g(X)) + H(X \mid g(X)) \tag{2.170}$$

$$\overset{(d)}{\geq} H(g(X)). \tag{2.171}$$

Thus, $H(g(X)) \leq H(X)$.

2.5 *Zero conditional entropy*. Show that if $H(Y|X) = 0$, then Y is a function of X [i.e., for all x with $p(x) > 0$, there is only one possible value of y with $p(x, y) > 0$].

2.6 *Conditional mutual information vs. unconditional mutual information*. Give examples of joint random variables X, Y, and Z such that

(a) $I(X; Y \mid Z) < I(X; Y)$.

(b) $I(X; Y \mid Z) > I(X; Y)$.

2.7 *Coin weighing*. Suppose that one has n coins, among which there may or may not be one counterfeit coin. If there is a counterfeit coin, it may be either heavier or lighter than the other coins. The coins are to be weighed by a balance.

(a) Find an upper bound on the number of coins n so that k weighings will find the counterfeit coin (if any) and correctly declare it to be heavier or lighter.

(b) (Difficult) What is the coin- weighing strategy for $k = 3$ weighings and 12 coins?

2.8 *Drawing with and without replacement.* An urn contains r red, w white, and b black balls. Which has higher entropy, drawing $k \geq 2$ balls from the urn with replacement or without replacement? Set it up and show why. (There is both a difficult way and a relatively simple way to do this.)

2.9 *Metric.* A function $\rho(x, y)$ is a metric if for all x, y,

- $\rho(x, y) \geq 0$.
- $\rho(x, y) = \rho(y, x)$.
- $\rho(x, y) = 0$ if and only if $x = y$.
- $\rho(x, y) + \rho(y, z) \geq \rho(x, z)$.

(a) Show that $\rho(X, Y) = H(X|Y) + H(Y|X)$ satisfies the first, second, and fourth properties above. If we say that $X = Y$ if there is a one-to-one function mapping from X to Y, the third property is also satisfied, and $\rho(X, Y)$ is a metric.

(b) Verify that $\rho(X, Y)$ can also be expressed as

$$\rho(X, Y) = H(X) + H(Y) - 2I(X; Y) \tag{2.172}$$

$$= H(X, Y) - I(X; Y) \tag{2.173}$$

$$= 2H(X, Y) - H(X) - H(Y). \tag{2.174}$$

2.10 *Entropy of a disjoint mixture.* Let X_1 and X_2 be discrete random variables drawn according to probability mass functions $p_1(\cdot)$ and $p_2(\cdot)$ over the respective alphabets $\mathcal{X}_1 = \{1, 2, \ldots, m\}$ and $\mathcal{X}_2 = \{m + 1, \ldots, n\}$. Let

$$X = \begin{cases} X_1 & \text{with probability } \alpha, \\ X_2 & \text{with probability } 1 - \alpha. \end{cases}$$

(a) Find $H(X)$ in terms of $H(X_1)$, $H(X_2)$, and α.

(b) Maximize over α to show that $2^{H(X)} \leq 2^{H(X_1)} + 2^{H(X_2)}$ and interpret using the notion that $2^{H(X)}$ is the effective alphabet size.

2.11 *Measure of correlation.* Let X_1 and X_2 be identically distributed but not necessarily independent. Let

$$\rho = 1 - \frac{H(X_2 \mid X_1)}{H(X_1)}.$$

(a) Show that $\rho = \frac{I(X_1;X_2)}{H(X_1)}$.

(b) Show that $0 \le \rho \le 1$.

(c) When is $\rho = 0$?

(d) When is $\rho = 1$?

2.12 *Example of joint entropy.* Let $p(x, y)$ be given by

$X \backslash Y$	0	1
0	$\frac{1}{3}$	$\frac{1}{3}$
1	0	$\frac{1}{3}$

Find:

(a) $H(X), H(Y)$.

(b) $H(X \mid Y), H(Y \mid X)$.

(c) $H(X, Y)$.

(d) $H(Y) - H(Y \mid X)$.

(e) $I(X; Y)$.

(f) Draw a Venn diagram for the quantities in parts (a) through (e).

2.13 *Inequality.* Show that $\ln x \ge 1 - \frac{1}{x}$ for $x > 0$.

2.14 *Entropy of a sum.* Let X and Y be random variables that take on values $x_1, x_2, \ldots, x_r$ and $y_1, y_2, \ldots, y_s$, respectively. Let $Z = X + Y$.

(a) Show that $H(Z|X) = H(Y|X)$. Argue that if X, Y are independent, then $H(Y) \le H(Z)$ and $H(X) \le H(Z)$. Thus, the addition of *independent* random variables adds uncertainty.

(b) Give an example of (necessarily dependent) random variables in which $H(X) > H(Z)$ and $H(Y) > H(Z)$.

(c) Under what conditions does $H(Z) = H(X) + H(Y)$?

2.15 *Data processing.* Let $X_1 \to X_2 \to X_3 \to \cdots \to X_n$ form a Markov chain in this order; that is, let

$$p(x_1, x_2, \ldots, x_n) = p(x_1)p(x_2|x_1) \cdots p(x_n|x_{n-1}).$$

Reduce $I(X_1; X_2, \ldots, X_n)$ to its simplest form.

2.16 *Bottleneck.* Suppose that a (nonstationary) Markov chain starts in one of n states, necks down to $k < n$ states, and then fans back to $m > k$ states. Thus, $X_1 \to X_2 \to X_3$, that is,

$p(x_1, x_2, x_3) = p(x_1)p(x_2|x_1)p(x_3|x_2)$, for all $x_1 \in \{1, 2, \ldots, n\}$, $x_2 \in \{1, 2, \ldots, k\}$, $x_3 \in \{1, 2, \ldots, m\}$.

(a) Show that the dependence of X_1 and X_3 is limited by the bottleneck by proving that $I(X_1; X_3) \leq \log k$.

(b) Evaluate $I(X_1; X_3)$ for $k = 1$, and conclude that no dependence can survive such a bottleneck.

2.17 *Pure randomness and bent coins.* Let $X_1, X_2, \ldots, X_n$ denote the outcomes of independent flips of a *bent* coin. Thus, $\Pr\{X_i = 1\} = p$, $\Pr\{X_i = 0\} = 1 - p$, where p is unknown. We wish to obtain a sequence $Z_1, Z_2, \ldots, Z_K$ of *fair* coin flips from $X_1, X_2, \ldots, X_n$. Toward this end, let $f : \mathcal{X}^n \to \{0, 1\}^*$ (where $\{0, 1\}^* = \{\Lambda, 0, 1, 00, 01, \ldots\}$ is the set of all finite-length binary sequences) be a mapping $f(X_1, X_2, \ldots, X_n) = (Z_1, Z_2, \ldots, Z_K)$, where $Z_i \sim$ Bernoulli $(\frac{1}{2})$, and K may depend on $(X_1, \ldots, X_n)$. In order that the sequence $Z_1, Z_2, \ldots$ appear to be fair coin flips, the map f from bent coin flips to fair flips must have the property that all 2^k sequences $(Z_1, Z_2, \ldots, Z_k)$ of a given length k have equal probability (possibly 0), for $k = 1, 2, \ldots$. For example, for $n = 2$, the map $f(01) = 0$, $f(10) = 1$, $f(00) = f(11) = \Lambda$ (the null string) has the property that $\Pr\{Z_1 = 1|K = 1\} = \Pr\{Z_1 = 0|K = 1\} = \frac{1}{2}$. Give reasons for the following inequalities:

$$\begin{aligned} nH(p) &\overset{\text{(a)}}{=} H(X_1, \ldots, X_n) \\ &\overset{\text{(b)}}{\geq} H(Z_1, Z_2, \ldots, Z_K, K) \\ &\overset{\text{(c)}}{=} H(K) + H(Z_1, \ldots, Z_K|K) \\ &\overset{\text{(d)}}{=} H(K) + E(K) \\ &\overset{\text{(e)}}{\geq} EK. \end{aligned}$$

Thus, no more than $nH(p)$ fair coin tosses can be derived from $(X_1, \ldots, X_n)$, on the average. Exhibit a good map f on sequences of length 4.

2.18 *World Series.* The World Series is a seven-game series that terminates as soon as either team wins four games. Let X be the random variable that represents the outcome of a World Series between teams A and B; possible values of X are AAAA, BABABAB, and BBBAAAA. Let Y be the number of games played, which ranges from 4 to 7. Assuming that A and B are equally matched and that the games are independent, calculate $H(X)$, $H(Y)$, $H(Y|X)$, and $H(X|Y)$.

2.19 *Infinite entropy*. This problem shows that the entropy of a discrete random variable can be infinite. Let $A = \sum_{n=2}^{\infty}(n \log^2 n)^{-1}$. [It is easy to show that A is finite by bounding the infinite sum by the integral of $(x \log^2 x)^{-1}$.] Show that the integer-valued random variable X defined by $\Pr(X = n) = (An \log^2 n)^{-1}$ for $n = 2, 3, \ldots$, has $H(X) = +\infty$.

2.20 *Run-length coding*. Let $X_1, X_2, \ldots, X_n$ be (possibly dependent) binary random variables. Suppose that one calculates the run lengths $\mathbf{R} = (R_1, R_2, \ldots)$ of this sequence (in order as they occur). For example, the sequence $\mathbf{X} = 0001100100$ yields run lengths $\mathbf{R} = (3, 2, 2, 1, 2)$. Compare $H(X_1, X_2, \ldots, X_n)$, $H(\mathbf{R})$, and $H(X_n, \mathbf{R})$. Show all equalities and inequalities, and bound all the differences.

2.21 *Markov's inequality for probabilities*. Let $p(x)$ be a probability mass function. Prove, for all $d \geq 0$, that

$$\Pr\{p(X) \leq d\} \log \frac{1}{d} \leq H(X). \tag{2.175}$$

2.22 *Logical order of ideas*. Ideas have been developed in order of need and then generalized if necessary. Reorder the following ideas, strongest first, implications following:

(a) Chain rule for $I(X_1, \ldots, X_n; Y)$, chain rule for $D(p(x_1, \ldots, x_n)||q(x_1, x_2, \ldots, x_n))$, and chain rule for $H(X_1, X_2, \ldots, X_n)$.

(b) $D(f||g) \geq 0$, Jensen's inequality, $I(X; Y) \geq 0$.

2.23 *Conditional mutual information*. Consider a sequence of n binary random variables $X_1, X_2, \ldots, X_n$. Each sequence with an even number of 1's has probability $2^{-(n-1)}$, and each sequence with an odd number of 1's has probability 0. Find the mutual informations

$$I(X_1; X_2), \quad I(X_2; X_3|X_1), \ldots, \quad I(X_{n-1}; X_n|X_1, \ldots, X_{n-2}).$$

2.24 *Average entropy*. Let $H(p) = -p \log_2 p - (1 - p) \log_2(1 - p)$ be the binary entropy function.

(a) Evaluate $H(\frac{1}{4})$ using the fact that $\log_2 3 \approx 1.584$. (*Hint:* You may wish to consider an experiment with four equally likely outcomes, one of which is more interesting than the others.)

(b) Calculate the average entropy $H(p)$ when the probability p is chosen uniformly in the range $0 \le p \le 1$.

(c) (*Optional*) Calculate the average entropy $H(p_1, p_2, p_3)$, where (p_1, p_2, p_3) is a uniformly distributed probability vector. Generalize to dimension n.

2.25 *Venn diagrams*. There isn't really a notion of mutual information common to three random variables. Here is one attempt at a definition: Using Venn diagrams, we can see that the mutual information common to three random variables X, Y, and Z can be defined by

$$I(X; Y; Z) = I(X; Y) - I(X; Y|Z).$$

This quantity is symmetric in X, Y, and Z, despite the preceding asymmetric definition. Unfortunately, $I(X; Y; Z)$ is not necessarily nonnegative. Find X, Y, and Z such that $I(X; Y; Z) < 0$, and prove the following two identities:

(a) $I(X; Y; Z) = H(X, Y, Z) - H(X) - H(Y) - H(Z) + I(X; Y) + I(Y; Z) + I(Z; X)$.

(b) $I(X; Y; Z) = H(X, Y, Z) - H(X, Y) - H(Y, Z) - H(Z, X) + H(X) + H(Y) + H(Z)$.

The first identity can be understood using the Venn diagram analogy for entropy and mutual information. The second identity follows easily from the first.

2.26 *Another proof of nonnegativity of relative entropy*. In view of the fundamental nature of the result $D(p||q) \ge 0$, we will give another proof.

(a) Show that $\ln x \le x - 1$ for $0 < x < \infty$.

(b) Justify the following steps:

$$-D(p||q) = \sum_x p(x) \ln \frac{q(x)}{p(x)} \tag{2.176}$$

$$\le \sum_x p(x) \left(\frac{q(x)}{p(x)} - 1 \right) \tag{2.177}$$

$$\le 0. \tag{2.178}$$

(c) What are the conditions for equality?

2.27 *Grouping rule for entropy*. Let $\mathbf{p} = (p_1, p_2, \ldots, p_m)$ be a probability distribution on m elements (i.e., $p_i \ge 0$ and $\sum_{i=1}^m p_i = 1$).

Define a new distribution $\mathbf{q}$ on $m-1$ elements as $q_1 = p_1, q_2 = p_2, \ldots, q_{m-2} = p_{m-2}$, and $q_{m-1} = p_{m-1} + p_m$ [i.e., the distribution $\mathbf{q}$ is the same as $\mathbf{p}$ on $\{1, 2, \ldots, m-2\}$, and the probability of the last element in $\mathbf{q}$ is the sum of the last two probabilities of $\mathbf{p}$]. Show that

$$H(\mathbf{p}) = H(\mathbf{q}) + (p_{m-1} + p_m) H\left(\frac{p_{m-1}}{p_{m-1}+p_m}, \frac{p_m}{p_{m-1}+p_m}\right). \tag{2.179}$$

2.28 *Mixing increases entropy.* Show that the entropy of the probability distribution, $(p_1, \ldots, p_i, \ldots, p_j, \ldots, p_m)$, is less than the entropy of the distribution $(p_1, \ldots, \frac{p_i+p_j}{2}, \ldots, \frac{p_i+p_j}{2}, \ldots, p_m)$. Show that in general any transfer of probability that makes the distribution more uniform increases the entropy.

2.29 *Inequalities.* Let X, Y, and Z be joint random variables. Prove the following inequalities and find conditions for equality.

(a) $H(X, Y|Z) \geq H(X|Z)$.

(b) $I(X, Y; Z) \geq I(X; Z)$.

(c) $H(X, Y, Z) - H(X, Y) \leq H(X, Z) - H(X)$.

(d) $I(X; Z|Y) \geq I(Z; Y|X) - I(Z; Y) + I(X; Z)$.

2.30 *Maximum entropy.* Find the probability mass function $p(x)$ that maximizes the entropy $H(X)$ of a nonnegative integer-valued random variable X subject to the constraint

$$EX = \sum_{n=0}^{\infty} n p(n) = A$$

for a fixed value $A > 0$. Evaluate this maximum $H(X)$.

2.31 *Conditional entropy.* Under what conditions does $H(X|g(Y)) = H(X|Y)$?

2.32 *Fano.* We are given the following joint distribution on (X, Y):

Y X	a	b	c
1	$\frac{1}{6}$	$\frac{1}{12}$	$\frac{1}{12}$
2	$\frac{1}{12}$	$\frac{1}{6}$	$\frac{1}{12}$
3	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{6}$

Let $\hat{X}(Y)$ be an estimator for X (based on Y) and let $P_e = \Pr\{\hat{X}(Y) \neq X\}$.

(a) Find the minimum probability of error estimator $\hat{X}(Y)$ and the associated P_e.

(b) Evaluate Fano's inequality for this problem and compare.

2.33 *Fano's inequality.* Let $\Pr(X = i) = p_i$, $i = 1, 2, \ldots, m$, and let $p_1 \geq p_2 \geq p_3 \geq \cdots \geq p_m$. The minimal probability of error predictor of X is $\hat{X} = 1$, with resulting probability of error $P_e = 1 - p_1$. Maximize $H(\mathbf{p})$ subject to the constraint $1 - p_1 = P_e$ to find a bound on P_e in terms of H. This is Fano's inequality in the absence of conditioning.

2.34 *Entropy of initial conditions.* Prove that $H(X_0|X_n)$ is nondecreasing with n for any Markov chain.

2.35 *Relative entropy is not symmetric.*
Let the random variable X have three possible outcomes $\{a, b, c\}$. Consider two distributions on this random variable:

Symbol	$p(x)$	$q(x)$
a	$\frac{1}{2}$	$\frac{1}{3}$
b	$\frac{1}{4}$	$\frac{1}{3}$
c	$\frac{1}{4}$	$\frac{1}{3}$

Calculate $H(p)$, $H(q)$, $D(p||q)$, and $D(q||p)$. Verify that in this case, $D(p||q) \neq D(q||p)$.

2.36 *Symmetric relative entropy.* Although, as Problem 2.35 shows, $D(p||q) \neq D(q||p)$ in general, there could be distributions for which equality holds. Give an example of two distributions p and q on a binary alphabet such that $D(p||q) = D(q||p)$ (other than the trivial case $p = q$).

2.37 *Relative entropy.* Let X, Y, Z be three random variables with a joint probability mass function $p(x, y, z)$. The relative entropy between the joint distribution and the product of the marginals is

$$D(p(x, y, z)||p(x)p(y)p(z)) = E\left[\log \frac{p(x, y, z)}{p(x)p(y)p(z)}\right]. \quad (2.180)$$

Expand this in terms of entropies. When is this quantity zero?

2.38 *The value of a question.* Let $X \sim p(x)$, $x = 1, 2, \ldots, m$. We are given a set $S \subseteq \{1, 2, \ldots, m\}$. We ask whether $X \in S$ and receive the answer

$$Y = \begin{cases} 1 & \text{if } X \in S \\ 0 & \text{if } X \notin S. \end{cases}$$

Suppose that $\Pr\{X \in S\} = \alpha$. Find the decrease in uncertainty $H(X) - H(X|Y)$.
Apparently, any set S with a given α is as good as any other.

2.39 *Entropy and pairwise independence.* Let X, Y, Z be three binary Bernoulli($\frac{1}{2}$) random variables that are pairwise independent; that is, $I(X; Y) = I(X; Z) = I(Y; Z) = 0$.

(a) Under this constraint, what is the minimum value for $H(X, Y, Z)$?

(b) Give an example achieving this minimum.

2.40 *Discrete entropies.* Let X and Y be two independent integer-valued random variables. Let X be uniformly distributed over $\{1, 2, \ldots, 8\}$, and let $\Pr\{Y = k\} = 2^{-k}$, $k = 1, 2, 3, \ldots$.

(a) Find $H(X)$.

(b) Find $H(Y)$.

(c) Find $H(X + Y, X - Y)$.

2.41 *Random questions.* One wishes to identify a random object $X \sim p(x)$. A question $Q \sim r(q)$ is asked at random according to $r(q)$. This results in a deterministic answer $A = A(x, q) \in \{a_1, a_2, \ldots\}$. Suppose that X and Q are independent. Then $I(X; Q, A)$ is the uncertainty in X removed by the question–answer (Q, A).

(a) Show that $I(X; Q, A) = H(A|Q)$. Interpret.

(b) Now suppose that two i.i.d. questions $Q_1, Q_2, \sim r(q)$ are asked, eliciting answers A_1 and A_2. Show that two questions are less valuable than twice a single question in the sense that $I(X; Q_1, A_1, Q_2, A_2) \leq 2I(X; Q_1, A_1)$.

2.42 *Inequalities.* Which of the following inequalities are generally $\geq, =, \leq$? Label each with $\geq, =,$ or $\leq$.

(a) $H(5X)$ vs. $H(X)$

(b) $I(g(X); Y)$ vs. $I(X; Y)$

(c) $H(X_0|X_{-1})$ vs. $H(X_0|X_{-1}, X_1)$

(d) $H(X, Y)/(H(X) + H(Y))$ vs. 1

2.43 *Mutual information of heads and tails*

(a) Consider a fair coin flip. What is the mutual information between the top and bottom sides of the coin?

(b) A six-sided fair die is rolled. What is the mutual information between the top side and the front face (the side most facing you)?

2.44 *Pure randomness.* We wish to use a three-sided coin to generate a fair coin toss. Let the coin X have probability mass function

$$X = \begin{cases} A, & p_A \\ B, & p_B \\ C, & p_C, \end{cases}$$

where p_A, p_B, p_C are unknown.

(a) How would you use two independent flips X_1, X_2 to generate (if possible) a Bernoulli($\frac{1}{2}$) random variable Z?

(b) What is the resulting maximum expected number of fair bits generated?

2.45 *Finite entropy.* Show that for a discrete random variable $X \in \{1, 2, \ldots\}$, if $E \log X < \infty$, then $H(X) < \infty$.

2.46 *Axiomatic definition of entropy (Difficult).* If we assume certain axioms for our measure of information, we will be forced to use a logarithmic measure such as entropy. Shannon used this to justify his initial definition of entropy. In this book we rely more on the other properties of entropy rather than its axiomatic derivation to justify its use. The following problem is considerably more difficult than the other problems in this section.

If a sequence of symmetric functions $H_m(p_1, p_2, \ldots, p_m)$ satisfies the following properties:

- Normalization: $H_2\left(\frac{1}{2}, \frac{1}{2}\right) = 1$,
- Continuity: $H_2(p, 1-p)$ is a continuous function of p,
- Grouping: $H_m(p_1, p_2, \ldots, p_m) = H_{m-1}(p_1 + p_2, p_3, \ldots, p_m) + (p_1 + p_2)H_2\left(\frac{p_1}{p_1+p_2}, \frac{p_2}{p_1+p_2}\right)$,

prove that H_m must be of the form

$$H_m(p_1, p_2, \ldots, p_m) = -\sum_{i=1}^{m} p_i \log p_i, \qquad m = 2, 3, \ldots. \tag{2.181}$$

There are various other axiomatic formulations which result in the same definition of entropy. See, for example, the book by Csiszár and Körner [149].

2.47 *Entropy of a missorted file.* A deck of n cards in order $1, 2, \ldots, n$ is provided. One card is removed at random, then replaced at random. What is the entropy of the resulting deck?

2.48 *Sequence length.* How much information does the length of a sequence give about the content of a sequence? Suppose that we consider a Bernoulli $(\frac{1}{2})$ process $\{X_i\}$. Stop the process when the first 1 appears. Let N designate this stopping time. Thus, X^N is an element of the set of all finite-length binary sequences $\{0, 1\}^* = \{0, 1, 00, 01, 10, 11, 000, \ldots\}$.

(a) Find $I(N; X^N)$.

(b) Find $H(X^N|N)$.

(c) Find $H(X^N)$.

Let's now consider a different stopping time. For this part, again assume that $X_i \sim \text{Bernoulli}(\frac{1}{2})$ but stop at time $N = 6$, with probability $\frac{1}{3}$ and stop at time $N = 12$ with probability $\frac{2}{3}$. Let this stopping time be independent of the sequence $X_1 X_2 \cdots X_{12}$.

(d) Find $I(N; X^N)$.

(e) Find $H(X^N|N)$.

(f) Find $H(X^N)$.

HISTORICAL NOTES

The concept of entropy was introduced in thermodynamics, where it was used to provide a statement of the second law of thermodynamics. Later, statistical mechanics provided a connection between thermodynamic entropy and the logarithm of the number of microstates in a macrostate of the system. This work was the crowning achievement of Boltzmann, who had the equation $S = k \ln W$ inscribed as the epitaph on his gravestone [361].

In the 1930s, Hartley introduced a logarithmic measure of information for communication. His measure was essentially the logarithm of the alphabet size. Shannon [472] was the first to define entropy and mutual information as defined in this chapter. Relative entropy was first defined by Kullback and Leibler [339]. It is known under a variety of names, including the Kullback–Leibler distance, cross entropy, information divergence, and information for discrimination, and has been studied in detail by Csiszár [138] and Amari [22].

Many of the simple properties of these quantities were developed by Shannon. Fano's inequality was proved in Fano [201]. The notion of sufficient statistic was defined by Fisher [209], and the notion of the minimal sufficient statistic was introduced by Lehmann and Scheffé [350]. The relationship of mutual information and sufficiency is due to Kullback [335]. The relationship between information theory and thermodynamics has been discussed extensively by Brillouin [77] and Jaynes [294].

The physics of information is a vast new subject of inquiry spawned from statistical mechanics, quantum mechanics, and information theory. The key question is how information is represented physically. Quantum channel capacity (the logarithm of the number of distinguishable preparations of a physical system) and quantum data compression [299] are well-defined problems with nice answers involving the von Neumann entropy. A new element of quantum information arises from the existence of quantum entanglement and the consequences (exhibited in Bell's inequality) that the observed marginal distribution of physical events are not consistent with any joint distribution (no local realism). The fundamental text by Nielsen and Chuang [395] develops the theory of quantum information and the quantum counterparts to many of the results in this book. There have also been attempts to determine whether there are any fundamental physical limits to computation, including work by Bennett [47] and Bennett and Landauer [48].