increases. Among other things, the second law allows one to dismiss any claims to perpetual motion machines. We briefly discuss the second law in Chapter 2.

Mathematics (Probability Theory and Statistics). The fundamental quantities of information theory—entropy, relative entropy and mutual information—are defined as functionals of probability distributions. In turn, they characterize the behavior of long sequences of random variables and allow us to estimate the probabilities of rare events (large deviation theory) and to find the best error exponent in hypothesis tests.

Philosophy of Science (Occam's Razor). William of Occam said "Causes shall not be multiplied beyond necessity," or to paraphrase it, "The simplest explanation is best". Solomonoff, and later Chaitin, argue persuasively that one gets a universally good prediction procedure if one takes a weighted combination of all programs that explain the data and observes what they print next. Moreover, this inference will work in many problems not handled by statistics. For example, this procedure will eventually predict the subsequent digits of π. When this procedure is applied to coin flips that come up heads with probability 0.7, this too will be inferred. When applied to the stock market, the procedure should essentially find all the "laws" of the stock market and extrapolate them optimally. In principle, such a procedure would have found Newton's laws of physics. Of course, such inference is highly impractical, because weeding out all computer programs that fail to generate existing data will take impossibly long. We would predict what happens tomorrow a hundred years from now.

Economics (Investment). Repeated investment in a stationary stock market results in an exponential growth of wealth. The growth rate of the wealth (called the doubling rate) is a dual of the entropy rate of the stock market. The parallels between the theory of optimal investment in the stock market and information theory are striking. We develop the theory of investment to explore this duality.

Computation vs. Communication. As we build larger computers out of smaller components, we encounter both a computation limit and a communication limit. Computation is communication limited and communication is computation limited. These become intertwined, and thus all of the developments in communication theory via information theory should have a direct impact on the theory of computation.

1.1 PREVIEW OF THE BOOK

The initial questions treated by information theory were in the areas of data compression and transmission. The answers are quantities like entropy and mutual information, which are functions of the probability distributions that underlie the process of communication. A few definitions will aid the initial discussion. We repeat these definitions in Chapter 2.

The entropy of a random variable X with a probability mass function $p(x)$ is defined by

$$H(X) = -\sum p(x) \log_2 p(x) . \tag{1.1}$$

We will use logarithms to base 2. The entropy will then be measured in bits. The entropy is a measure of the average uncertainty in the random variable. It is the number of bits on the average required to describe the random variable.

Example 1.1.1: Consider a random variable which has a uniform distribution over 32 outcomes. To identify an outcome, we need a label that takes on 32 different values. Thus 5-bit strings suffice as labels.

The entropy of this random variable is

$$H(X) = -\sum_{i=1}^{32} p(i) \log p(i) = -\sum_{i=1}^{32} \frac{1}{32} \log \frac{1}{32} = \log 32 = 5 \text{ bits} , \tag{1.2}$$

which agrees with the number of bits needed to describe X. In this case, all the outcomes have representations of the same length.

Now consider an example with a non-uniform distribution.

Example 1.1.2: Suppose we have a horse race with eight horses taking part. Assume that the probabilities of winning for the eight horses are $(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64})$. We can calculate the entropy of the horse race as

$$\begin{aligned} H(X) &= -\frac{1}{2} \log \frac{1}{2} - \frac{1}{4} \log \frac{1}{4} - \frac{1}{8} \log \frac{1}{8} - \frac{1}{16} \log \frac{1}{16} - 4 \frac{1}{64} \log \frac{1}{64} \\ &= 2 \text{ bits} . \end{aligned} \tag{1.3}$$

Suppose that we wish to send a message to another person indicating which horse won the race. One alternative is to send the index of the winning horse. This description requires 3 bits for any of the horses. But the win probabilities are not uniform. It therefore makes sense to use shorter descriptions for the more probable horses, and longer descriptions for the less probable ones, so that we achieve a lower average description length. For example, we could use the following set of bit

strings to represent the eight horses—0, 10, 110, 1110, 111100, 111101, 111110, 111111. The average description length in this case is 2 bits, as opposed to 3 bits for the uniform code. Notice that the average description length in this case is equal to the entropy. In Chapter 5, we show that the entropy of a random variable is a lower bound on the average number of bits required to represent the random variable and also on the average number of questions needed to identify the variable in a game of "twenty questions." We also show how to construct representations that have an average length within one bit of the entropy.

The concept of entropy in information theory is closely connected with the concept of entropy in statistical mechanics. If we draw a sequence of n independent and identically distributed (i.i.d.) random variables, we will show that the probability of a "typical" sequence is about $2^{-nH(X)}$ and that there are about $2^{nH(X)}$ such "typical" sequences. This property (known as the asymptotic equipartition property, or AEP) is the basis of many of the proofs in information theory. We later present other problems for which entropy arises as a natural answer (for example, the number of fair coin flips needed to generate a random variable).

The notion of descriptive complexity of a random variable can be extended to define the descriptive complexity of a single string. The Kolmogorov complexity of a binary string is defined as the length of the shortest computer program that prints out the string. It will turn out that if the string is indeed random, the Kolmogorov complexity is close to the entropy. Kolmogorov complexity is a natural framework in which to consider problems of statistical inference and modeling and leads to a clearer understanding of Occam's Razor "The simplest explanation is best." We describe some simple properties of Kolmogorov complexity in Chapter 7.

Entropy is the uncertainty of a single random variable. We can define conditional entropy, which is the entropy of a random variable, given another random variable. The reduction in uncertainty due to another random variable is called the mutual information. For two random variables X and Y this reduction is

$$I(X; Y) = H(X) - H(X|Y) = \sum_{x,y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)} . \tag{1.4}$$

The mutual information $I(X; Y)$ is a measure of the dependence between the two random variables. It is symmetric in X and Y and always non-negative.

A communication channel is a system in which the output depends probabilistically on its input. It is characterized by a probability transition matrix that determines the conditional distribution of the output given the input. For a communication channel with input X and output Y, we define the capacity C by

$$C = \max_{p(x)} I(X; Y) . \tag{1.5}$$

Later we show that the capacity is the maximum rate at which we can send information over the channel and recover the information at the output with a vanishingly low probability of error. We illustrate this with a few examples.

Example 1.1.3 (*Noiseless binary channel*): For this channel, the binary input is reproduced exactly at the output. This channel is illustrated in Figure 1.3. Here, any transmitted bit is received without error. Hence, in each transmission, we can send 1 bit reliably to the receiver, and the capacity is 1 bit. We can also calculate the information capacity $C = \max I(X; Y) = 1$ bit.

Example 1.1.4 (*Noisy four-symbol channel*): Consider the channel shown in Figure 1.4. In this channel, each input letter is received either as the same letter with probability 1/2 or as the next letter with probability 1/2. If we use all four input symbols, then inspection of the output would not reveal with certainty which input symbol was sent. If, on the other hand, we use only two of the inputs (1 and 3 say), then we can immediately tell from the output which input symbol was sent. This channel then acts like the noiseless channel of the previous example, and we can send 1 bit per transmission over this channel with no errors. We can calculate the channel capacity $C = \max I(X; Y)$ in this case, and it is equal to 1 bit per transmission, in agreement with the analysis above.

In general, communication channels do not have the simple structure of this example, so we cannot always identify a subset of the inputs to send information without error. But if we consider a sequence of transmissions, then all channels look like this example and we can then identify a subset of the input sequences (the codewords) which can be used to transmit information over the channel in such a way that the sets of possible output sequences associated with each of the codewords

Figure 1.3. Noiseless binary channel.

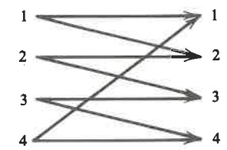

Figure 1.4. A noisy channel.

Chapter 2

Entropy, Relative Entropy and Mutual Information

This chapter introduces most of the basic definitions required for the 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 non-negativity of mutual information, the data processing inequality, and finally investigate the extent to which the second law of thermodynamics holds for Markov processes.

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. We derive some of these properties 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 will first introduce the concept of entropy, which is a measure of uncertainty of a random variable. Let X be a discrete random variable

with alphabet $\mathcal{X}$ and probability mass function $p(x) = \Pr\{X = x\}$, $x \in \mathcal{X}$. 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 $x \to 0$. Thus adding terms of zero probability does not change the entropy.

If the base of the logarithm is b, we will denote the entropy as $H_b(X)$. If the base of the logarithm is e, then 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 shall denote expectation by E. Thus if $X \sim p(x)$, then 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 $\log \frac{1}{p(X)}$, 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 will be 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 a problem at the end of the chapter. We

will not use the axiomatic approach to justify the definition of entropy; instead, we will 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 $\log(1/p(x)) \geq 0$. □

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 $p = 1/2$. 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

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

The entropy of X is

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

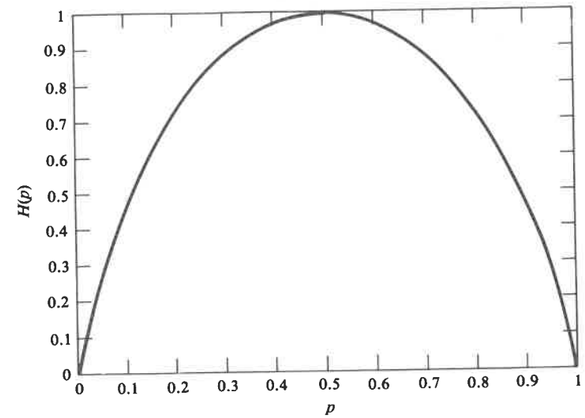

Figure 2.1. $H(p)$ versus p.

Suppose 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, then 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 have defined the entropy of a single random variable in the previous section. 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

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