Elements of Information Theory
Thomas M. Cover, Joy A. Thomas

Print ISBN 0-471-06259-6 Online ISBN 0-471-20061-1

Chapter 1

Introduction and Preview

This "first and last lecture" chapter goes backwards and forwards through information theory and its naturally related ideas. The full definitions and study of the subject begin in Chapter 2.

Information theory answers two fundamental questions in communication theory: what is the ultimate data compression (answer: the entropy H), and what is the ultimate transmission rate of communication (answer: the channel capacity C). For this reason some consider information theory to be a subset of communication theory. We will argue that it is much more. Indeed, it has fundamental contributions to make in statistical physics (thermodynamics), computer science (Kolmogorov complexity or algorithmic complexity), statistical inference (Occam's Razor: "The simplest explanation is best") and to probability and statistics (error rates for optimal hypothesis testing and estimation).

Figure 1.1 illustrates the relationship of information theory to other fields. As the figure suggests, information theory intersects physics (statistical mechanics), mathematics (probability theory), electrical engineering (communication theory) and computer science (algorithmic complexity). We now describe the areas of intersection in greater detail:

Electrical Engineering (Communication Theory). In the early 1940s, it was thought that increasing the transmission rate of information over a communication channel increased the probability of error. Shannon surprised the communication theory community by proving that this was not true as long as the communication rate was below channel capacity. The capacity can be simply computed from the noise characteristics of the channel. Shannon further argued that random processes such as music and speech have an irreducible

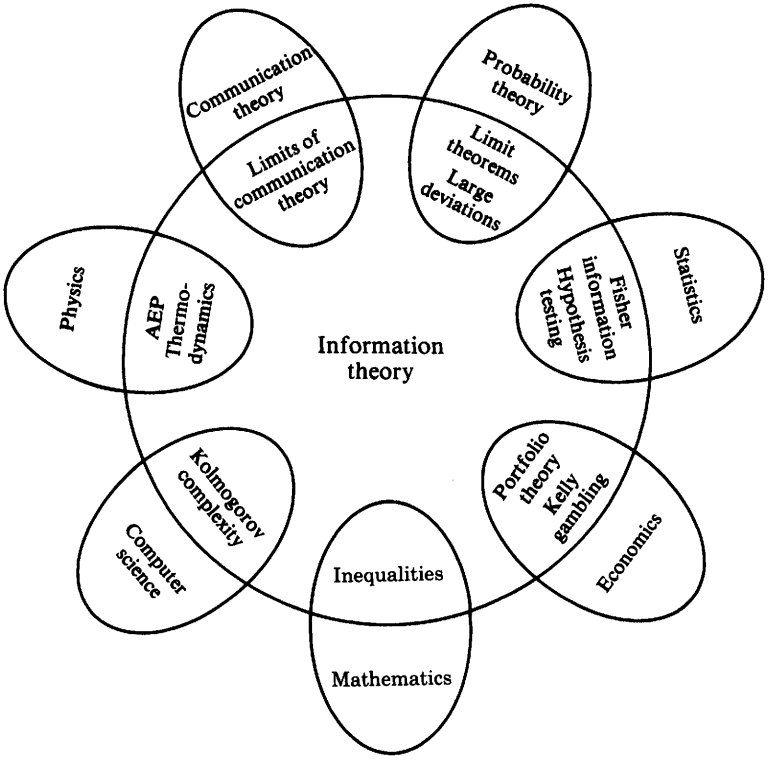

Figure 1.1. The relationship of information theory with other fields.

complexity below which the signal cannot be compressed. This he named the entropy, in deference to the parallel use of this word in thermodynamics, and argued that if the entropy of the source is less than the capacity of the channel, then asymptotically error free communication can be achieved.

Information theory today represents the extreme points of the set of all possible communication schemes, as shown in the fanciful Figure 1.2. The data compression minimum $I(X; \hat{X})$ lies at one extreme of the set of communication ideas. All data compression schemes require description rates at least equal to this minimum. At the other extreme is the data transmission maximum $I(X; Y)$, known as the channel capacity. Thus all

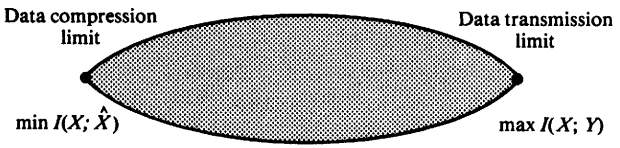

Figure 1.2. Information theoretic extreme points of communication theory.

modulation schemes and data compression schemes lie between these limits.

Information theory also suggests means of achieving these ultimate limits of communication. However, these theoretically optimal communication schemes, beautiful as they are, may turn out to be computationally impractical. It is only because of the computational feasibility of simple modulation and demodulation schemes that we use them rather than the random coding and nearest neighbor decoding rule suggested by Shannon's proof of the channel capacity theorem. Progress in integrated circuits and code design has enabled us to reap some of the gains suggested by Shannon's theory. A good example of an application of the ideas of information theory is the use of error correcting codes on compact discs.

Modern work on the communication aspects of information theory has concentrated on network information theory: the theory of the simultaneous rates of communication from many senders to many receivers in a communication network. Some of the trade-offs of rates between senders and receivers are unexpected, and all have a certain mathematical simplicity. A unifying theory, however, remains to be found.

Computer Science (Kolmogorov Complexity). Kolmogorov, Chaitin and Solomonoff put forth the idea that the complexity of a string of data can be defined by the length of the shortest binary program for computing the string. Thus the complexity is the minimal description length. This definition of complexity turns out to be universal, that is, computer independent, and is of fundamental importance. Thus Kolmogorov complexity lays the foundation for *the* theory of descriptive complexity. Gratifyingly, the Kolmogorov complexity K is approximately equal to the Shannon entropy H if the sequence is drawn at random from a distribution that has entropy H. So the tie-in between information theory and Kolmogorov complexity is perfect. Indeed, we consider Kolmogorov complexity to be more fundamental than Shannon entropy. It is the ultimate data compression and leads to a logically consistent procedure for inference.

There is a pleasing complementary relationship between algorithmic complexity and computational complexity. One can think about computational complexity (time complexity) and Kolmogorov complexity (program length or descriptive complexity) as two axes corresponding to program running time and program length. Kolmogorov complexity focuses on minimizing along the second axis, and computational complexity focuses on minimizing along the first axis. Little work has been done on the simultaneous minimization of the two.

Physics (Thermodynamics). Statistical mechanics is the birthplace of entropy and the second law of thermodynamics. Entropy always

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.

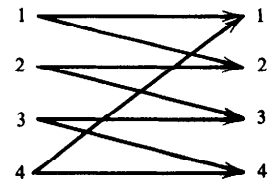

Figure 1.4. A noisy channel.

are approximately disjoint. We can then look at the output sequence and identify the input sequence with a vanishingly low probability of error.

Example 1.1.5 (*Binary symmetric channel*): This is the basic example of a noisy communication system. The channel is illustrated in Figure 1.5.

The channel has a binary input, and its output is equal to the input with probability $1-p$. With probability p, on the other hand, a 0 is received as a 1, and vice versa.

In this case, the capacity of the channel can be calculated to be $C = 1 + p \log p + (1-p)\log(1-p)$ bits per transmission. However, it is no longer obvious how one can achieve this capacity. If we use the channel many times, however, the channel begins to look like the noisy four-symbol channel of the previous example, and we can send information at a rate C bits per transmission with an arbitrarily low probability of error.

The ultimate limit on the rate of communication of information over a channel is given by the channel capacity. The channel coding theorem shows that this limit can be achieved by using codes with a long block length. In practical communication systems, there are limitations on the complexity of the codes that we can use, and therefore we may not be able to achieve capacity.

Mutual information turns out to be a special case of a more general quantity called relative entropy $D(p\|q)$ which is a measure of the "distance" between two probability mass functions p and q. It is defined as

$$D(p\|q) = \sum_x p(x) \log \frac{p(x)}{q(x)} . \tag{1.6}$$

Although relative entropy is not a true metric, it has some of the properties of a metric. In particular, it is always non-negative and is zero if and only if $p = q$. Relative entropy arises as the exponent in the

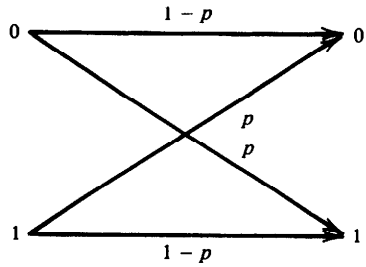

Figure 1.5. Binary symmetric channel.

probability of error in a hypothesis test between distributions p and q. Relative entropy can be used to define a geometry for probability distributions that allows us to interpret many of the results of large deviation theory.

There are a number of parallels between information theory and the theory of investment in a stock market. A stock market is defined by a random vector $\mathbf{X}$ whose elements are non-negative numbers equal to the ratio of the price of a stock at the end of a day to the price at the beginning of the day. For a stock market with distribution $F(\boldsymbol{x})$, we can define the doubling rate W as

$$W = \max_{\mathbf{b}: b_i \geq 0,\ \Sigma b_i = 1} \int \log \mathbf{b}^t \mathbf{x}\, dF(\mathbf{x}) . \tag{1.7}$$

The doubling rate is the maximum asymptotic exponent in the growth of wealth. The doubling rate has a number of properties that parallel the properties of entropy. We explore some of these properties in Chapter 15.

The quantities H, I, C, D, K, W arise naturally in the following areas:

- *Data compression.* The entropy H of a random variable is a lower bound on the average length of the shortest description of the random variable. We can construct descriptions with average length within one bit of the entropy.

  If we relax the constraint of recovering the source perfectly, we can then ask what rates are required to describe the source up to distortion D? And what channel capacities are sufficient to enable the transmission of this source over the channel and its reconstruction with distortion less than or equal to D? This is the subject of rate distortion theory.

  When we try to formalize the notion of the shortest description for non-random objects, we are led to the definition of Kolmogorov complexity K. Later, we will show that Kolmogorov complexity is universal and satisfies many of the intuitive requirements for the theory of shortest descriptions.

- *Data transmission.* We consider the problem of transmitting information so that the receiver can decode the message with a small probability of error. Essentially, we wish to find codewords (sequences of input symbols to a channel) that are mutually far apart in the sense that their noisy versions (available at the output of the channel) are distinguishable. This is equivalent to sphere packing in high dimensional space. For any set of codewords it is possible to calculate the probability the receiver will make an error, i.e., make an incorrect decision as to which codeword was sent. However, in most cases, this calculation is tedious.

Using a randomly generated code, Shannon showed that one can send information at any rate below the capacity C of the channel with an arbitrarily low probability of error. The idea of a randomly generated code is very unusual. It provides the basis for a simple analysis of a very difficult problem. One of the key ideas in the proof is the concept of typical sequences.

- *Network information theory.* Each of the topics previously mentioned involves a single source or a single channel. What if one wishes simultaneously to compress many sources and then put the compressed descriptions together into a joint reconstruction of the sources? This problem is solved by the Slepian-Wolf theorem. Or what if one has many senders independently sending information to a common receiver? What is the channel capacity of this channel? This is the multiple access channel solved by Liao and Ahlswede. Or what if one has one sender and many receivers and wishes to simultaneously communicate (perhaps different) information to each of the receivers? This is the broadcast channel. Finally, what if one has an arbitrary number of senders and receivers in an environment of interference and noise. What is the capacity region of achievable rates from the various senders to the receivers? This is the general network information theory problem. All of the preceding problems fall into the general area of multiple-user or network information theory. Although hopes for a unified theory may be beyond current research techniques, there is still some hope that all the answers involve only elaborate forms of mutual information and relative entropy.
- *Ergodic theory.* The asymptotic equipartition theorem states that most sample n-sequences of an ergodic process have probability about 2^{-nH} and that there are about 2^{nH} such typical sequences.
- *Hypothesis testing.* The relative entropy D arises as the exponent in the probability of error in a hypothesis test between two distributions. It is a natural measure of distance between distributions.
- *Statistical mechanics.* The entropy H arises in statistical mechanics as a measure of uncertainty or disorganization in a physical system. The second law of thermodynamics says that the entropy of a closed system cannot decrease. Later we provide some interpretations of the second law.
- *Inference.* We can use the notion of Kolmogorov complexity K to find the shortest description of the data and use that as a model to predict what comes next. A model that maximizes the uncertainty or entropy yields the maximum entropy approach to inference.
- *Gambling and investment.* The optimal exponent in the growth rate of wealth is given by the doubling rate W. For a horse race

with uniform odds, the sum of the doubling rate W and the entropy H is constant. The mutual information I between a horse race and some side information is an upper bound on the increase in the doubling rate due to the side information. Similar results hold for investment in a stock market.

- *Probability theory.* The asymptotic equipartition property (AEP) shows that most sequences are typical in that they have a sample entropy close to H. So attention can be restricted to these approximately 2^{nH} typical sequences. In large deviation theory, the probability of a set is approximately 2^{-nD}, where D is the relative entropy distance between the closest element in the set and the true distribution.
- *Complexity theory.* The Kolmogorov complexity K is a measure of the descriptive complexity of an object. It is related to, but different from, computational complexity, which measures the time or space required for a computation.

Information theoretic quantities like entropy and relative entropy arise again and again as the answers to the fundamental questions in communication and statistics. Before studying these questions, we shall study some of the properties of the answers. We begin in the next chapter with the definitions and the basic properties of entropy, relative entropy and mutual information.