ELEMENTS OF INFORMATION THEORY

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada.

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information on our other products and services or for technical support, please contact our Customer Care Department within the United States at (800) 762-2974, outside the United States at (317) 572-3993 or fax (317) 572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic formats. For more information about Wiley products, visit our web site at www.wiley.com.

Library of Congress Cataloging-in-Publication Data:

Cover, T. M., 1938–
Elements of information theory/by Thomas M. Cover, Joy A. Thomas.–2nd ed.
p. cm.
"A Wiley-Interscience publication."
Includes bibliographical references and index.
ISBN-13 978-0-471-24195-9
ISBN-10 0-471-24195-4
1. Information theory. I. Thomas, Joy A. II. Title.
Q360.C68 2005
003′.54–dc22
2005047799

Printed in the United States of America.

10 9 8 7 6 5 4 3 2 1

CONTENTS

PREFACE TO THE SECOND EDITION

In the years since the publication of the first edition, there were many aspects of the book that we wished to improve, to rearrange, or to expand, but the constraints of reprinting would not allow us to make those changes between printings. In the new edition, we now get a chance to make some of these changes, to add problems, and to discuss some topics that we had omitted from the first edition.

The key changes include a reorganization of the chapters to make the book easier to teach, and the addition of more than two hundred new problems. We have added material on universal portfolios, universal source coding, Gaussian feedback capacity, network information theory, and developed the duality of data compression and channel capacity. A new chapter has been added and many proofs have been simplified. We have also updated the references and historical notes.

The material in this book can be taught in a two-quarter sequence. The first quarter might cover Chapters 1 to 9, which includes the asymptotic equipartition property, data compression, and channel capacity, culminating in the capacity of the Gaussian channel. The second quarter could cover the remaining chapters, including rate distortion, the method of types, Kolmogorov complexity, network information theory, universal source coding, and portfolio theory. If only one semester is available, we would add rate distortion and a single lecture each on Kolmogorov complexity and network information theory to the first semester. A web site, http://www.elementsofinformationtheory.com, provides links to additional material and solutions to selected problems.

In the years since the first edition of the book, information theory celebrated its 50th birthday (the 50th anniversary of Shannon's original paper that started the field), and ideas from information theory have been applied to many problems of science and technology, including bioinformatics, web search, wireless communication, video compression, and others. The list of applications is endless, but it is the elegance of the fundamental mathematics that is still the key attraction of this area. We hope that this book will give some insight into why we believe that this is one of the most interesting areas at the intersection of mathematics, physics, statistics, and engineering.

Tom Cover
Joy Thomas

Palo Alto, California
January 2006

PREFACE TO THE FIRST EDITION

This is intended to be a simple and accessible book on information theory. As Einstein said, "*Everything should be made as simple as possible, but no simpler.*" Although we have not verified the quote (first found in a fortune cookie), this point of view drives our development throughout the book. There are a few key ideas and techniques that, when mastered, make the subject appear simple and provide great intuition on new questions.

This book has arisen from over ten years of lectures in a two-quarter sequence of a senior and first-year graduate-level course in information theory, and is intended as an introduction to information theory for students of communication theory, computer science, and statistics.

There are two points to be made about the simplicities inherent in information theory. First, certain quantities like entropy and mutual information arise as the answers to fundamental questions. For example, entropy is the minimum descriptive complexity of a random variable, and mutual information is the communication rate in the presence of noise. Also, as we shall point out, mutual information corresponds to the increase in the doubling rate of wealth given side information. Second, the answers to information theoretic questions have a natural algebraic structure. For example, there is a chain rule for entropies, and entropy and mutual information are related. Thus the answers to problems in data compression and communication admit extensive interpretation. We all know the feeling that follows when one investigates a problem, goes through a large amount of algebra, and finally investigates the answer to find that the entire problem is illuminated not by the analysis but by the inspection of the answer. Perhaps the outstanding examples of this in physics are Newton's laws and Schrödinger's wave equation. Who could have foreseen the awesome philosophical interpretations of Schrödinger's wave equation?

In the text we often investigate properties of the answer before we look at the question. For example, in Chapter 2, we define entropy, relative entropy, and mutual information and study the relationships and a few interpretations of them, showing how the answers fit together in various ways. Along the way we speculate on the meaning of the second law of thermodynamics. Does entropy always increase? The answer is yes and no. This is the sort of result that should please experts in the area but might be overlooked as standard by the novice.

In fact, that brings up a point that often occurs in teaching. It is fun to find new proofs or slightly new results that no one else knows. When one presents these ideas along with the established material in class, the response is "sure, sure, sure." But the excitement of teaching the material is greatly enhanced. Thus we have derived great pleasure from investigating a number of new ideas in this textbook.

Examples of some of the new material in this text include the chapter on the relationship of information theory to gambling, the work on the universality of the second law of thermodynamics in the context of Markov chains, the joint typicality proofs of the channel capacity theorem, the competitive optimality of Huffman codes, and the proof of Burg's theorem on maximum entropy spectral density estimation. Also, the chapter on Kolmogorov complexity has no counterpart in other information theory texts. We have also taken delight in relating Fisher information, mutual information, the central limit theorem, and the Brunn–Minkowski and entropy power inequalities. To our surprise, many of the classical results on determinant inequalities are most easily proved using information theoretic inequalities.

Even though the field of information theory has grown considerably since Shannon's original paper, we have strived to emphasize its coherence. While it is clear that Shannon was motivated by problems in communication theory when he developed information theory, we treat information theory as a field of its own with applications to communication theory and statistics. We were drawn to the field of information theory from backgrounds in communication theory, probability theory, and statistics, because of the apparent impossibility of capturing the intangible concept of information.

Since most of the results in the book are given as theorems and proofs, we expect the elegance of the results to speak for themselves. In many cases we actually describe the properties of the solutions before the problems. Again, the properties are interesting in themselves and provide a natural rhythm for the proofs that follow.

One innovation in the presentation is our use of long chains of inequalities with no intervening text followed immediately by the explanations. By the time the reader comes to many of these proofs, we expect that he or she will be able to follow most of these steps without any explanation and will be able to pick out the needed explanations. These chains of inequalities serve as pop quizzes in which the reader can be reassured of having the knowledge needed to prove some important theorems. The natural flow of these proofs is so compelling that it prompted us to flout one of the cardinal rules of technical writing; and the absence of verbiage makes the logical necessity of the ideas evident and the key ideas perspicuous. We hope that by the end of the book the reader will share our appreciation of the elegance, simplicity, and naturalness of information theory.

Throughout the book we use the method of weakly typical sequences, which has its origins in Shannon's original 1948 work but was formally developed in the early 1970s. The key idea here is the asymptotic equipartition property, which can be roughly paraphrased as "Almost everything is almost equally probable."

Chapter 2 includes the basic algebraic relationships of entropy, relative entropy, and mutual information. The asymptotic equipartition property (AEP) is given central prominence in Chapter 3. This leads us to discuss the entropy rates of stochastic processes and data compression in Chapters 4 and 5. A gambling sojourn is taken in Chapter 6, where the duality of data compression and the growth rate of wealth is developed.

The sensational success of Kolmogorov complexity as an intellectual foundation for information theory is explored in Chapter 14. Here we replace the goal of finding a description that is good on the average with the goal of finding the universally shortest description. There is indeed a universal notion of the descriptive complexity of an object. Here also the wonderful number Ω is investigated. This number, which is the binary expansion of the probability that a Turing machine will halt, reveals many of the secrets of mathematics.

Channel capacity is established in Chapter 7. The necessary material on differential entropy is developed in Chapter 8, laying the groundwork for the extension of previous capacity theorems to continuous noise channels. The capacity of the fundamental Gaussian channel is investigated in Chapter 9.

The relationship between information theory and statistics, first studied by Kullback in the early 1950s and relatively neglected since, is developed in Chapter 11. Rate distortion theory requires a little more background than its noiseless data compression counterpart, which accounts for its placement as late as Chapter 10 in the text.

The huge subject of network information theory, which is the study of the simultaneously achievable flows of information in the presence of noise and interference, is developed in Chapter 15. Many new ideas come into play in network information theory. The primary new ingredients are interference and feedback. Chapter 16 considers the stock market, which is the generalization of the gambling processes considered in Chapter 6, and shows again the close correspondence of information theory and gambling.

Chapter 17, on inequalities in information theory, gives us a chance to recapitulate the interesting inequalities strewn throughout the book, put them in a new framework, and then add some interesting new inequalities on the entropy rates of randomly drawn subsets. The beautiful relationship of the Brunn–Minkowski inequality for volumes of set sums, the entropy power inequality for the effective variance of the sum of independent random variables, and the Fisher information inequalities are made explicit here.

We have made an attempt to keep the theory at a consistent level. The mathematical level is a reasonably high one, probably the senior or first-year graduate level, with a background of at least one good semester course in probability and a solid background in mathematics. We have, however, been able to avoid the use of measure theory. Measure theory comes up only briefly in the proof of the AEP for ergodic processes in Chapter 16. This fits in with our belief that the fundamentals of information theory are orthogonal to the techniques required to bring them to their full generalization.

The essential vitamins are contained in Chapters 2, 3, 4, 5, 7, 8, 9, 11, 10, and 15. This subset of chapters can be read without essential reference to the others and makes a good core of understanding. In our opinion, Chapter 14 on Kolmogorov complexity is also essential for a deep understanding of information theory. The rest, ranging from gambling to inequalities, is part of the terrain illuminated by this coherent and beautiful subject.

Every course has its first lecture, in which a sneak preview and overview of ideas is presented. Chapter 1 plays this role.

Tom Cover
Joy Thomas

Palo Alto, California
June 1990

ACKNOWLEDGMENTS FOR THE SECOND EDITION

Since the appearance of the first edition, we have been fortunate to receive feedback, suggestions, and corrections from a large number of readers. It would be impossible to thank everyone who has helped us in our efforts, but we would like to list some of them. In particular, we would like to thank all the faculty who taught courses based on this book and the students who took those courses; it is through them that we learned to look at the same material from a different perspective.

In particular, we would like to thank Andrew Barron, Alon Orlitsky, T. S. Han, Raymond Yeung, Nam Phamdo, Franz Willems, and Marty Cohn for their comments and suggestions. Over the years, students at Stanford have provided ideas and inspirations for the changes—these include George Gemelos, Navid Hassanpour, Young-Han Kim, Charles Mathis, Styrmir Sigurjonsson, Jon Yard, Michael Baer, Mung Chiang, Suhas Diggavi, Elza Erkip, Paul Fahn, Garud Iyengar, David Julian, Yiannis Kontoyiannis, Amos Lapidoth, Erik Ordentlich, Sandeep Pombra, Jim Roche, Arak Sutivong, Joshua Sweetkind-Singer, and Assaf Zeevi. Denise Murphy provided much support and help during the preparation of the second edition.

Joy Thomas would like to acknowledge the support of colleagues at IBM and Stratify who provided valuable comments and suggestions. Particular thanks are due Peter Franaszek, C. S. Chang, Randy Nelson, Ramesh Gopinath, Pandurang Nayak, John Lamping, Vineet Gupta, and Ramana Venkata. In particular, many hours of dicussion with Brandon Roy helped refine some of the arguments in the book. Above all, Joy would like to acknowledge that the second edition would not have been possible without the support and encouragement of his wife, Priya, who makes all things worthwhile.

Tom Cover would like to thank his students and his wife, Karen.

ACKNOWLEDGMENTS FOR THE FIRST EDITION

We wish to thank everyone who helped make this book what it is. In particular, Aaron Wyner, Toby Berger, Masoud Salehi, Alon Orlitsky, Jim Mazo and Andrew Barron have made detailed comments on various drafts of the book which guided us in our final choice of content. We would like to thank Bob Gallager for an initial reading of the manuscript and his encouragement to publish it. Aaron Wyner donated his new proof with Ziv on the convergence of the Lempel-Ziv algorithm. We would also like to thank Normam Abramson, Ed van der Meulen, Jack Salz and Raymond Yeung for their suggested revisions.

Certain key visitors and research associates contributed as well, including Amir Dembo, Paul Algoet, Hirosuke Yamamoto, Ben Kawabata, M. Shimizu and Yoichiro Watanabe. We benefited from the advice of John Gill when he used this text in his class. Abbas El Gamal made invaluable contributions, and helped begin this book years ago when we planned to write a research monograph on multiple user information theory. We would also like to thank the Ph.D. students in information theory as this book was being written: Laura Ekroot, Will Equitz, Don Kimber, Mitchell Trott, Andrew Nobel, Jim Roche, Erik Ordentlich, Elza Erkip and Vittorio Castelli. Also Mitchell Oslick, Chien-Wen Tseng and Michael Morrell were among the most active students in contributing questions and suggestions to the text. Marc Goldberg and Anil Kaul helped us produce some of the figures. Finally we would like to thank Kirsten Goodell and Kathy Adams for their support and help in some of the aspects of the preparation of the manuscript.

Joy Thomas would also like to thank Peter Franaszek, Steve Lavenberg, Fred Jelinek, David Nahamoo and Lalit Bahl for their encouragment and support during the final stages of production of this book.

CHAPTER 1

INTRODUCTION AND PREVIEW

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 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 exponents for optimal hypothesis testing and estimation).

This "first lecture" chapter goes backward and forward through information theory and its naturally related ideas. The full definitions and study of the subject begin in Chapter 2. 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 to be impossible to send information at a positive rate with negligible probability of error. Shannon surprised the communication theory community by proving that the probability of error could be made nearly zero for all communication rates below channel capacity. The capacity can be computed simply from the noise characteristics of the channel. Shannon further argued that random processes such as music and speech have an irreducible 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

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

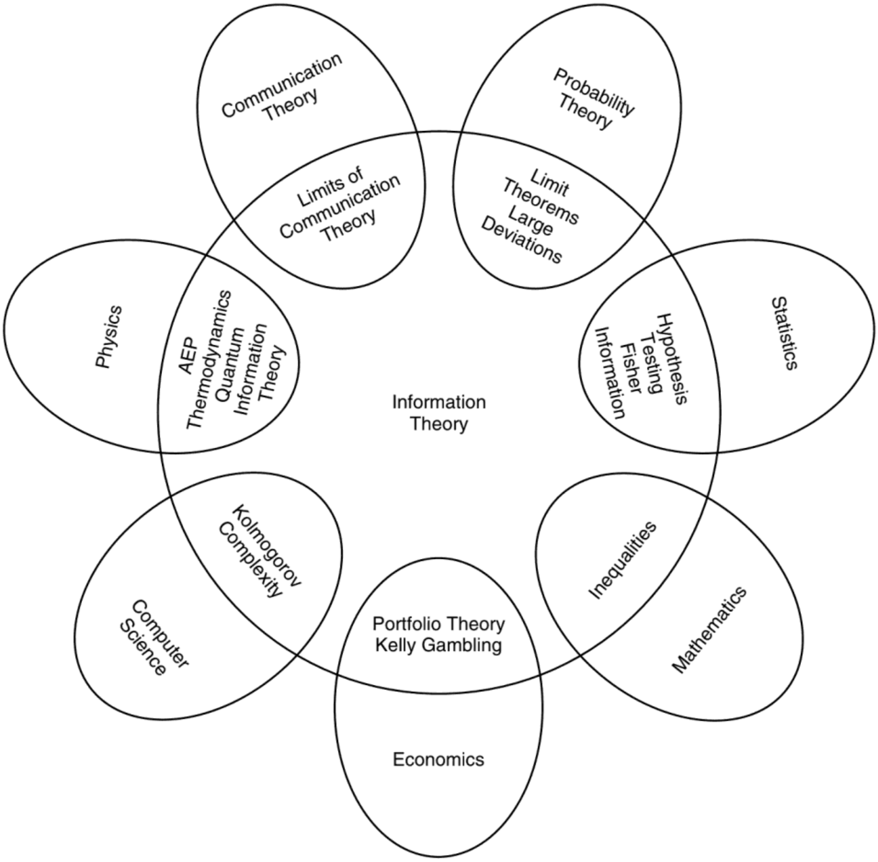

Communication Theory
Probability Theory
Limits of Communication Theory
Limit Theorems Large Deviations
Physics
AEP Thermodynamics Quantum Information Theory
Information Theory
Hypothesis Testing Fisher Information
Statistics
Kolmogorov Complexity
Computer Science
Inequalities
Mathematics
Portfolio Theory Kelly Gambling
Economics

FIGURE 1.1. Relationship of information theory to other fields.

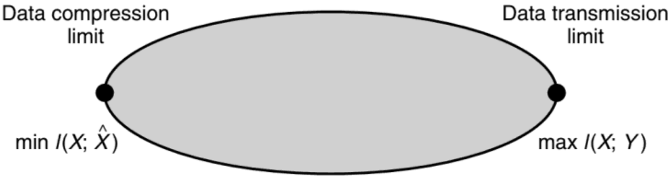


FIGURE 1.2. Information theory as the extreme points of communication theory.

source is less than the capacity of the channel, 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 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. Computational practicality was finally achieved by the advent of turbo codes. A good example of an application of the ideas of information theory is the use of error-correcting codes on compact discs and DVDs.

Recent 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 the presence of interference and noise. 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 computer 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 discuss the second law briefly in Chapter 4.

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 Chaitin argued 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 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 lay in the areas of data compression and transmission. The answers are quantities such as 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

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

We 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 average required to describe the random variable.

Example 1.1.1 Consider a random variable that 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 nonuniform distribution.

Example 1.1.2 Suppose that we have a horse race with eight horses taking part. Assume that the probabilities of winning for the eight horses are $\left(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}\right)$. 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 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 "20 questions." We also show how to construct representations that have an average length within 1 bit of the entropy.

The concept of entropy in information theory is related to 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* (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 (e.g., 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 1.

Entropy is the uncertainty of a single random variable. We can define conditional entropy $H(X|Y)$, which is the entropy of a random variable conditional on the knowledge of 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 the mutual information

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

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 nonnegative and is equal to zero if and only if X and Y are independent.

A *communication channel* is a system in which the output depends probabilistically on its input. It is characterized by a probability transition matrix $p(y|x)$ that determines the conditional distribution of the output given the input. For a communication channel with input X and output Y, we can 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 $\frac{1}{2}$ or as the next letter with probability $\frac{1}{2}$. If we use all four input symbols, inspection of the output would not reveal with certainty which input symbol was sent. If, on the other hand, we use

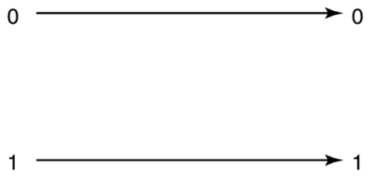


FIGURE 1.3. Noiseless binary channel. $C = 1$ bit.

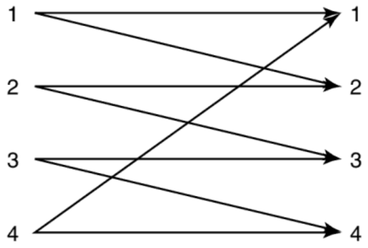


FIGURE 1.4. Noisy channel.

only two of the inputs (1 and 3, say), we can tell immediately from the output which input symbol was sent. This channel then acts like the noiseless channel of Example 1.1.3, 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, all channels look like this example and we can then identify a subset of the input sequences (the codewords) that 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 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.

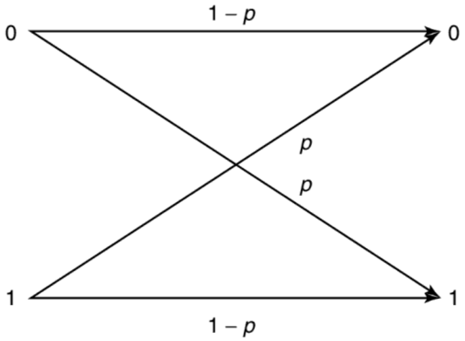


FIGURE 1.5. Binary symmetric channel.

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 Example 1.1.4, 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 nonnegative and is zero if and only if $p = q$. Relative entropy arises as the exponent in the 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 nonnegative 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 $F(\mathbf{x})$, we can define the doubling rate W as

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

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 16.

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 1 bit of the entropy. If we relax the constraint of recovering the source perfectly, we can then ask what communication 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 nonrandom objects, we are led to the definition of Kolmogorov complexity K. Later, we 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 that 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. The capacity C is the logarithm of the number of distinguishable input signals.
- *Network information theory*. Each of the topics mentioned previously involves a single source or a single channel. What if one wishes to compress each of 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 sending information independently 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 communicate (perhaps different) information simultaneously 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 comprehensive theory for networks 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. Roughly speaking, the entropy is the logarithm of the number of ways in which the physical system can be configured. The second law of thermodynamics says that the entropy of a closed system cannot decrease. Later we provide some interpretations of the second law.
- *Quantum mechanics*. Here, von Neumann entropy $S = \mathrm{tr}(\rho \ln \rho) = \sum_i \lambda_i \log \lambda_i$ plays the role of classical Shannon–Boltzmann entropy $H = -\sum_i p_i \log p_i$. Quantum mechanical versions of data compression and channel capacity can then be found.
- *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 increase in the doubling rate due to side information is equal to the mutual information I between a horse race and the side information. Similar results hold for investment in the 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 such as 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 Chapter 2 with the definitions and basic properties of entropy, relative entropy, and mutual information.

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

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

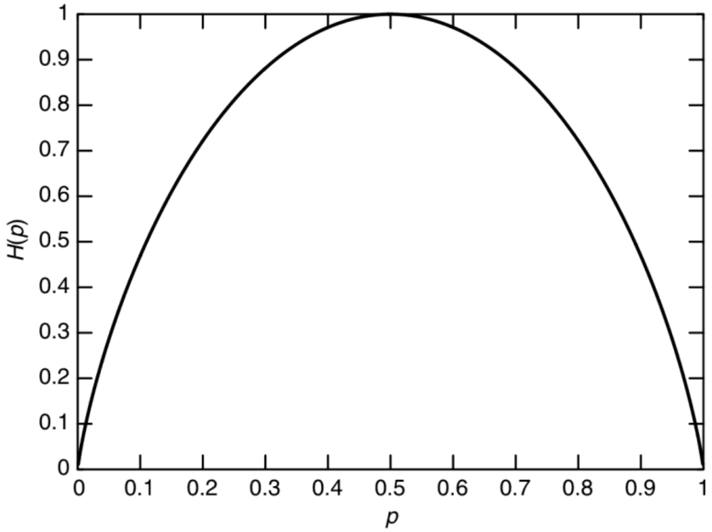


FIGURE 2.1. $H(p)$ vs. p.

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

$$\begin{aligned}
H(Y|X) &= \sum_{x\in\mathcal{X}} p(x) H(Y|X = x) && (2.10)\\
&= -\sum_{x\in\mathcal{X}} p(x) \sum_{y\in\mathcal{Y}} p(y|x) \log p(y|x) && (2.11)\\
&= -\sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p(x, y) \log p(y|x) && (2.12)\\
&= -E \log p(Y|X). && (2.13)
\end{aligned}$$

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

$$\begin{aligned}
H(X, Y) &= -\sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p(x, y) \log p(x, y) && (2.15)\\
&= -\sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p(x, y) \log p(x) p(y|x) && (2.16)\\
&= -\sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p(x, y) \log p(x) - \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p(x, y) \log p(y|x) && (2.17)\\
&= -\sum_{x\in\mathcal{X}} p(x) \log p(x) - \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p(x, y) \log p(y|x) && (2.18)\\
&= H(X) + H(Y|X). && (2.19)
\end{aligned}$$

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:

X Y	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*)

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

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

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

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

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

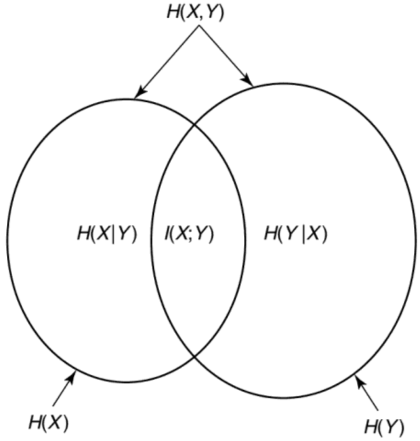


FIGURE 2.2. Relationship between entropy and mutual information.

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

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

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

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

$$\vdots$$

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

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

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

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

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

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

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

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

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

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

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

$$\begin{aligned}
&D(p(x, y)||q(x, y)) \\
&= \sum_x \sum_y p(x, y) \log \frac{p(x, y)}{q(x, y)} && (2.68) \\
&= \sum_x \sum_y p(x, y) \log \frac{p(x)p(y|x)}{q(x)q(y|x)} && (2.69) \\
&= \sum_x \sum_y p(x, y) \log \frac{p(x)}{q(x)} + \sum_x \sum_y p(x, y) \log \frac{p(y|x)}{q(y|x)} && (2.70) \\
&= D(p(x)||q(x)) + D(p(y|x)||q(y|x)). \quad \square && (2.71)
\end{aligned}$$

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 $0 \leq \lambda \leq 1$,

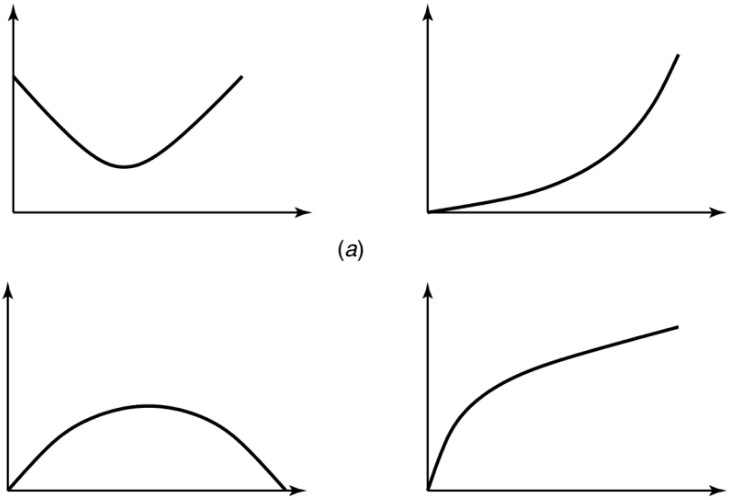
$$f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2). \qquad (2.72)$$

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 $x \geq 0$), and so on. Examples of concave functions include $\log x$ and $\sqrt{x}$ for $x \geq 0$. 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 non-negative (positive) over an interval, the function is convex (strictly convex) over that interval.*



(*a*)

(*b*)

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

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

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

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

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,

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

and the distribution of Y is also a mixture,

$$p_\lambda(y) = \lambda p_1(y) + (1 - \lambda)p_2(y). \quad (2.114)$$

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

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

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

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

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 $X \to Y \to Z$) 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 $X \to Y \to Z$ 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:

- $X \to Y \to Z$ 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.
- $X \to Y \to Z$ implies that $Z \to Y \to X$. Thus, the condition is sometimes written $X \leftrightarrow Y \leftrightarrow Z$.
- If $Z = f(Y)$, then $X \to Y \to Z$.

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 $X \to Y \to Z$, then $I(X; Y) \ge I(X; Z)$.*

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

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

   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

$$\theta \to T(X) \to U(X) \to X. \tag{2.128}$$

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:

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

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

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

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

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

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

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

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

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

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

$$\begin{aligned} H(X|E, \hat{X}) &= \Pr(E = 0)H(X|\hat{X}, E = 0) + \Pr(E = 1)H(X|\hat{X}, E = 1) \qquad (2.141) \\ &\le (1 - P_e)0 + P_e \log(|\mathcal{X}| - 1), \qquad (2.142) \end{aligned}$$

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

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

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

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

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

   $$H(X|Y) \le H(X) \tag{2.157}$$

   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.

$p(x_1, x_2, x_3) = p(x_1)p(x_2|x_1)p(x_3|x_2)$, 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 $I(X_1; X_3) \le \log k$.

(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:

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

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

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

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):

X \ Y	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?

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].

CHAPTER 3

ASYMPTOTIC EQUIPARTITION PROPERTY

In information theory, the analog of the law of large numbers is the asymptotic equipartition property (AEP). It is a direct consequence of the weak law of large numbers. The *law of large numbers* states that for independent, identically distributed (i.i.d.) random variables, $\frac{1}{n}\sum_{i=1}^{n} X_i$ is close to its expected value EX for large values of n. The AEP states that $\frac{1}{n}\log\frac{1}{p(X_1,X_2,\ldots,X_n)}$ is close to the entropy H, where $X_1, X_2, \ldots, X_n$ are i.i.d. random variables and $p(X_1, X_2, \ldots, X_n)$ is the probability of observing the sequence $X_1, X_2, \ldots, X_n$. Thus, the probability $p(X_1, X_2, \ldots, X_n)$ assigned to an observed sequence will be close to 2^{-nH}.

This enables us to divide the set of all sequences into two sets, the *typical set*, where the sample entropy is close to the true entropy, and the nontypical set, which contains the other sequences. Most of our attention will be on the typical sequences. Any property that is proved for the typical sequences will then be true with high probability and will determine the average behavior of a large sample.

First, an example. Let the random variable $X \in \{0, 1\}$ have a probability mass function defined by $p(1) = p$ and $p(0) = q$. If $X_1, X_2, \ldots, X_n$ are i.i.d. according to $p(x)$, the probability of a sequence $x_1, x_2, \ldots, x_n$ is $\prod_{i=1}^{n} p(x_i)$. For example, the probability of the sequence $(1, 0, 1, 1, 0, 1)$ is $p^{\sum X_i} q^{n-\sum X_i} = p^4 q^2$. Clearly, it is not true that all 2^n sequences of length n have the same probability.

However, we might be able to predict the probability of the sequence that we actually observe. We ask for the probability $p(X_1, X_2, \ldots, X_n)$ of the outcomes $X_1, X_2, \ldots, X_n$, where $X_1, X_2, \ldots$ are i.i.d. $\sim p(x)$. This is insidiously self-referential, but well defined nonetheless. Apparently, we are asking for the probability of an event drawn according to the same

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

probability distribution. Here it turns out that $p(X_1, X_2, \ldots, X_n)$ is close to 2^{-nH} with high probability.

We summarize this by saying, "Almost all events are almost equally surprising." This is a way of saying that

$$\Pr\left\{(X_1, X_2, \ldots, X_n) : p(X_1, X_2, \ldots, X_n) = 2^{-n(H \pm \epsilon)}\right\} \approx 1 \tag{3.1}$$

if $X_1, X_2, \ldots, X_n$ are i.i.d. $\sim p(x)$.

In the example just given, where $p(X_1, X_2, \ldots, X_n) = p^{\sum X_i} q^{n - \sum X_i}$, we are simply saying that the number of 1's in the sequence is close to np (with high probability), and all such sequences have (roughly) the same probability $2^{-nH(p)}$. We use the idea of convergence in probability, defined as follows:

Definition (*Convergence of random variables*). Given a sequence of random variables, $X_1, X_2, \ldots$, we say that the sequence $X_1, X_2, \ldots$ converges to a random variable X:

1. *In probability* if for every $\epsilon > 0$, $\Pr\{|X_n - X| > \epsilon\} \to 0$
2. *In mean square* if $E(X_n - X)^2 \to 0$
3. *With probability 1* (also called *almost surely*) if $\Pr\{\lim_{n\to\infty} X_n = X\} = 1$

3.1 ASYMPTOTIC EQUIPARTITION PROPERTY THEOREM

The asymptotic equipartition property is formalized in the following theorem.

Theorem 3.1.1 (*AEP*) *If $X_1, X_2, \ldots$ are i.i.d. $\sim p(x)$, then*

$$-\frac{1}{n} \log p(X_1, X_2, \ldots, X_n) \to H(X) \qquad \textit{in probability.} \tag{3.2}$$

Proof: Functions of independent random variables are also independent random variables. Thus, since the X_i are i.i.d., so are $\log p(X_i)$. Hence, by the weak law of large numbers,

$$-\frac{1}{n} \log p(X_1, X_2, \ldots, X_n) = -\frac{1}{n} \sum_i \log p(X_i) \tag{3.3}$$

$$\to -E \log p(X) \qquad \text{in probability} \tag{3.4}$$

$$= H(X), \tag{3.5}$$

which proves the theorem. □

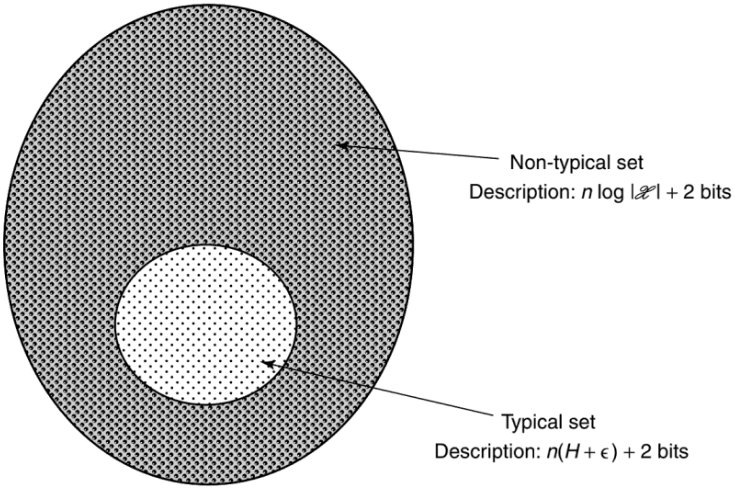

Non-typical set
Description: $n \log |\mathcal{X}| + 2$ bits
Typical set
Description: $n(H + \epsilon) + 2$ bits

FIGURE 3.2. Source code using the typical set.

We order all elements in each set according to some order (e.g., lexicographic order). Then we can represent each sequence of $A_\epsilon^{(n)}$ by giving the index of the sequence in the set. Since there are $\le 2^{n(H+\epsilon)}$ sequences in $A_\epsilon^{(n)}$, the indexing requires no more than $n(H+\epsilon)+1$ bits. [The extra bit may be necessary because $n(H+\epsilon)$ may not be an integer.] We prefix all these sequences by a 0, giving a total length of $\le n(H+\epsilon)+2$ bits to represent each sequence in $A_\epsilon^{(n)}$ (see Figure 3.2). Similarly, we can index each sequence not in $A_\epsilon^{(n)}$ by using not more than $n \log |\mathcal{X}| + 1$ bits. Prefixing these indices by 1, we have a code for all the sequences in $\mathcal{X}^n$.

Note the following features of the above coding scheme:

- The code is one-to-one and easily decodable. The initial bit acts as a flag bit to indicate the length of the codeword that follows.
- We have used a brute-force enumeration of the atypical set $A_\epsilon^{(n)^c}$ without taking into account the fact that the number of elements in $A_\epsilon^{(n)^c}$ is less than the number of elements in $\mathcal{X}^n$. Surprisingly, this is good enough to yield an efficient description.
- The typical sequences have short descriptions of length $\approx nH$.

We use the notation x^n to denote a sequence $x_1, x_2, \ldots, x_n$. Let $l(x^n)$ be the length of the codeword corresponding to x^n. If n is sufficiently large so that $\Pr\{A_\epsilon^{(n)}\} \ge 1-\epsilon$, the expected length of the codeword is

$$E(l(X^n)) = \sum_{x^n} p(x^n) l(x^n) \tag{3.17}$$

$$= \sum_{x^n \in A_\epsilon^{(n)}} p(x^n) l(x^n) + \sum_{x^n \in A_\epsilon^{(n)^c}} p(x^n) l(x^n) \tag{3.18}$$

$$\le \sum_{x^n \in A_\epsilon^{(n)}} p(x^n)(n(H+\epsilon)+2)$$

$$+ \sum_{x^n \in A_\epsilon^{(n)^c}} p(x^n)(n \log |\mathcal{X}| + 2) \tag{3.19}$$

$$= \Pr\left\{A_\epsilon^{(n)}\right\}(n(H+\epsilon)+2) + \Pr\left\{A_\epsilon^{(n)^c}\right\}(n \log |\mathcal{X}| + 2) \tag{3.20}$$

$$\le n(H+\epsilon) + \epsilon n(\log |\mathcal{X}|) + 2 \tag{3.21}$$

$$= n(H+\epsilon'), \tag{3.22}$$

where $\epsilon' = \epsilon + \epsilon \log |\mathcal{X}| + \frac{2}{n}$ can be made arbitrarily small by an appropriate choice of ϵ followed by an appropriate choice of n. Hence we have proved the following theorem.

Theorem 3.2.1 *Let X^n be i.i.d. $\sim p(x)$. Let $\epsilon > 0$. Then there exists a code that maps sequences x^n of length n into binary strings such that the mapping is one-to-one (and therefore invertible) and*

$$E\left[\frac{1}{n} l(X^n)\right] \le H(X) + \epsilon \tag{3.23}$$

for n sufficiently large.

Thus, we can represent sequences X^n using $nH(X)$ bits on the average.

3.3 HIGH-PROBABILITY SETS AND THE TYPICAL SET

From the definition of $A_\epsilon^{(n)}$, it is clear that $A_\epsilon^{(n)}$ is a fairly small set that contains most of the probability. But from the definition, it is not clear whether it is the smallest such set. We will prove that the typical set has essentially the same number of elements as the smallest set, to first order in the exponent.

Definition For each $n = 1, 2, \ldots,$ let $B_\delta^{(n)} \subset \mathcal{X}^n$ be the smallest set with

$$\Pr\{B_\delta^{(n)}\} \ge 1 - \delta. \tag{3.24}$$

We argue that $B_\delta^{(n)}$ must have significant intersection with $A_\epsilon^{(n)}$ and therefore must have about as many elements. In Problem 3.3.11, we outline the proof of the following theorem.

Theorem 3.3.1 *Let $X_1, X_2, \ldots, X_n$ be i.i.d. $\sim p(x)$. For $\delta < \frac{1}{2}$ and any $\delta' > 0$, if $\Pr\{B_\delta^{(n)}\} > 1 - \delta$, then*

$$\frac{1}{n} \log |B_\delta^{(n)}| > H - \delta' \qquad \textit{for n sufficiently large.} \tag{3.25}$$

Thus, $B_\delta^{(n)}$ must have at least 2^{nH} elements, to first order in the exponent. But $A_\epsilon^{(n)}$ has $2^{n(H\pm\epsilon)}$ elements. Therefore, $A_\epsilon^{(n)}$ is about the same size as the smallest high-probability set.

We will now define some new notation to express equality to first order in the exponent.

Definition The notation $a_n \doteq b_n$ means

$$\lim_{n\to\infty} \frac{1}{n} \log \frac{a_n}{b_n} = 0. \tag{3.26}$$

Thus, $a_n \doteq b_n$ implies that a_n and b_n are equal to the first order in the exponent.

We can now restate the above results: If $\delta_n \to 0$ and $\epsilon_n \to 0$, then

$$|B_{\delta_n}^{(n)}| \doteq |A_{\epsilon_n}^{(n)}| \doteq 2^{nH}. \tag{3.27}$$

To illustrate the difference between $A_\epsilon^{(n)}$ and $B_\delta^{(n)}$, let us consider a Bernoulli sequence $X_1, X_2, \ldots, X_n$ with parameter $p = 0.9$. [A Bernoulli(θ) random variable is a binary random variable that takes on the value 1 with probability θ.] The typical sequences in this case are the sequences in which the proportion of 1's is close to 0.9. However, this does not include the most likely single sequence, which is the sequence of all 1's. The set $B_\delta^{(n)}$ includes all the most probable sequences and therefore includes the sequence of all 1's. Theorem 3.3.1 implies that $A_\epsilon^{(n)}$ and $B_\delta^{(n)}$ must both contain the sequences that have about 90% 1's, and the two sets are almost equal in size.

SUMMARY

AEP. “Almost all events are almost equally surprising.” Specifically, if $X_1, X_2, \ldots$ are i.i.d. $\sim p(x)$, then

$$-\frac{1}{n}\log p(X_1, X_2, \ldots, X_n) \to H(X) \text{ in probability.} \tag{3.28}$$

Definition. The *typical set* $A_\epsilon^{(n)}$ is the set of sequences $x_1, x_2, \ldots, x_n$ satisfying

$$2^{-n(H(X)+\epsilon)} \le p(x_1, x_2, \ldots, x_n) \le 2^{-n(H(X)-\epsilon)}. \tag{3.29}$$

Properties of the typical set

1. If $(x_1, x_2, \ldots, x_n) \in A_\epsilon^{(n)}$, then $p(x_1, x_2, \ldots, x_n) = 2^{-n(H\pm\epsilon)}$.
2. $\Pr\left\{A_\epsilon^{(n)}\right\} > 1-\epsilon$ for n sufficiently large.
3. $\left|A_\epsilon^{(n)}\right| \le 2^{n(H(X)+\epsilon)}$, where $|A|$ denotes the number of elements in set A.

Definition. $a_n \doteq b_n$ means that $\frac{1}{n}\log\frac{a_n}{b_n} \to 0$ as $n \to \infty$.

Smallest probable set. Let $X_1, X_2, \ldots, X_n$ be i.i.d. $\sim p(x)$, and for $\delta < \frac{1}{2}$, let $B_\delta^{(n)} \subset \mathcal{X}^n$ be the smallest set such that $\Pr\{B_\delta^{(n)}\} \ge 1-\delta$. Then

$$|B_\delta^{(n)}| \doteq 2^{nH}. \tag{3.30}$$

PROBLEMS

3.1 *Markov's inequality and Chebyshev's inequality*

(a) (*Markov's inequality*) For any nonnegative random variable X and any $t > 0$, show that

$$\Pr\{X \ge t\} \le \frac{EX}{t}. \tag{3.31}$$

Exhibit a random variable that achieves this inequality with equality.

(b) (*Chebyshev's inequality*) Let Y be a random variable with mean μ and variance σ^2. By letting $X = (Y-\mu)^2$, show that for any $\epsilon > 0$,

$$\Pr\{|Y-\mu| > \epsilon\} \le \frac{\sigma^2}{\epsilon^2}. \tag{3.32}$$

(c) (*Weak law of large numbers*) Let $Z_1, Z_2, \ldots, Z_n$ be a sequence of i.i.d. random variables with mean μ and variance σ^2. Let $\overline{Z}_n = \frac{1}{n}\sum_{i=1}^{n} Z_i$ be the sample mean. Show that

$$\Pr\left\{\left|\overline{Z}_n - \mu\right| > \epsilon\right\} \le \frac{\sigma^2}{n\epsilon^2}. \tag{3.33}$$

Thus, $\Pr\left\{\left|\overline{Z}_n - \mu\right| > \epsilon\right\} \to 0$ as $n \to \infty$. This is known as the *weak law of large numbers*.

3.2 *AEP and mutual information.* Let (X_i, Y_i) be i.i.d. $\sim p(x,y)$. We form the log likelihood ratio of the hypothesis that X and Y are independent vs. the hypothesis that X and Y are dependent. What is the limit of

$$\frac{1}{n}\log\frac{p(X^n)p(Y^n)}{p(X^n,Y^n)}?$$

3.3 *Piece of cake.*
A cake is sliced roughly in half, the largest piece being chosen each time, the other pieces discarded. We will assume that a random cut creates pieces of proportions

$$P = \begin{cases} (\frac{2}{3}, \frac{1}{3}) & \text{with probability } \frac{3}{4} \\ (\frac{2}{5}, \frac{3}{5}) & \text{with probability } \frac{1}{4} \end{cases}$$

Thus, for example, the first cut (and choice of largest piece) may result in a piece of size $\frac{3}{5}$. Cutting and choosing from this piece might reduce it to size $\left(\frac{3}{5}\right)\left(\frac{2}{3}\right)$ at time 2, and so on. How large, to first order in the exponent, is the piece of cake after n cuts?

3.4 *AEP.* Let X_i be iid $\sim p(x)$, $x \in \{1, 2, \ldots, m\}$. Let $\mu = EX$ and $H = -\sum p(x)\log p(x)$. Let $A^n = \{x^n \in \mathcal{X}^n : |-\frac{1}{n}\log p(x^n) - H| \le \epsilon\}$. Let $B^n = \{x^n \in \mathcal{X}^n : |\frac{1}{n}\sum_{i=1}^{n} X_i - \mu| \le \epsilon\}$.

(a) Does $\Pr\{X^n \in A^n\} \longrightarrow 1$?

(b) Does $\Pr\{X^n \in A^n \cap B^n\} \longrightarrow 1$?

3.12 *Monotonic convergence of the empirical distribution.* Let $\hat{p}_n$ denote the empirical probability mass function corresponding to $X_1, X_2, \ldots, X_n$ i.i.d. $\sim p(x), x \in \mathcal{X}$. Specifically,

$$\hat{p}_n(x) = \frac{1}{n} \sum_{i=1}^{n} I(X_i = x)$$

is the proportion of times that $X_i = x$ in the first n samples, where I is the indicator function.

(a) Show for $\mathcal{X}$ binary that

$$ED(\hat{p}_{2n} \parallel p) \leq ED(\hat{p}_n \parallel p).$$

Thus, the expected relative entropy "distance" from the empirical distribution to the true distribution decreases with sample size. (*Hint:* Write $\hat{p}_{2n} = \frac{1}{2}\hat{p}_n + \frac{1}{2}\hat{p}'_n$ and use the convexity of D.)

(b) Show for an arbitrary discrete $\mathcal{X}$ that

$$ED(\hat{p}_n \parallel p) \leq ED(\hat{p}_{n-1} \parallel p).$$

(*Hint:* Write $\hat{p}_n$ as the average of n empirical mass functions with each of the n samples deleted in turn.)

3.13 *Calculation of typical set.* To clarify the notion of a typical set $A_\epsilon^{(n)}$ and the smallest set of high probability $B_\delta^{(n)}$, we will calculate the set for a simple example. Consider a sequence of i.i.d. binary random variables, $X_1, X_2, \ldots, X_n$, where the probability that $X_i = 1$ is 0.6 (and therefore the probability that $X_i = 0$ is 0.4).

(a) Calculate $H(X)$.

(b) With $n = 25$ and $\epsilon = 0.1$, which sequences fall in the typical set $A_\epsilon^{(n)}$? What is the probability of the typical set? How many elements are there in the typical set? (This involves computation of a table of probabilities for sequences with k 1's, $0 \leq k \leq 25$, and finding those sequences that are in the typical set.)

(c) How many elements are there in the smallest set that has probability 0.9?

(d) How many elements are there in the intersection of the sets in parts (b) and (c)? What is the probability of this intersection?

k	$\binom{n}{k}$	$\binom{n}{k} p^k (1-p)^{n-k}$	$-\frac{1}{n} \log p(x^n)$
0	1	0.000000	1.321928
1	25	0.000000	1.298530
2	300	0.000000	1.275131
3	2300	0.000001	1.251733
4	12650	0.000007	1.228334
5	53130	0.000054	1.204936
6	177100	0.000227	1.181537
7	480700	0.001205	1.158139
8	1081575	0.003121	1.134740
9	2042975	0.013169	1.111342
10	3268760	0.021222	1.087943
11	4457400	0.077801	1.064545
12	5200300	0.075967	1.041146
13	5200300	0.267718	1.017748
14	4457400	0.146507	0.994349
15	3268760	0.575383	0.970951
16	2042975	0.151086	0.947552
17	1081575	0.846448	0.924154
18	480700	0.079986	0.900755
19	177100	0.970638	0.877357
20	53130	0.019891	0.853958
21	12650	0.997633	0.830560
22	2300	0.001937	0.807161
23	300	0.999950	0.783763
24	25	0.000047	0.760364
25	1	0.000003	0.736966

HISTORICAL NOTES

The asymptotic equipartition property (AEP) was first stated by Shannon in his original 1948 paper [472], where he proved the result for i.i.d. processes and stated the result for stationary ergodic processes. McMillan [384] and Breiman [74] proved the AEP for ergodic finite alphabet sources. The result is now referred to as the AEP or the Shannon–McMillan–Breiman theorem. Chung [101] extended the theorem to the case of countable alphabets and Moy [392], Perez [417], and Kieffer [312] proved the $\mathcal{L}_1$ convergence when $\{X_i\}$ is continuous valued and ergodic. Barron [34] and Orey [402] proved almost sure convergence for real-valued ergodic processes; a simple sandwich argument (Algoet and Cover [20]) will be used in Section 16.8 to prove the general AEP.

CHAPTER 4

ENTROPY RATES OF A STOCHASTIC PROCESS

The asymptotic equipartition property in Chapter 3 establishes that $nH(X)$ bits suffice on the average to describe n independent and identically distributed random variables. But what if the random variables are dependent? In particular, what if the random variables form a stationary process? We will show, just as in the i.i.d. case, that the entropy $H(X_1, X_2, \ldots, X_n)$ grows (asymptotically) linearly with n at a rate $H(\mathcal{X})$, which we will call the *entropy rate* of the process. The interpretation of $H(\mathcal{X})$ as the best achievable data compression will await the analysis in Chapter 5.

4.1 MARKOV CHAINS

A stochastic process $\{X_i\}$ is an indexed sequence of random variables. In general, there can be an arbitrary dependence among the random variables. The process is characterized by the joint probability mass functions $\Pr\{(X_1, X_2, \ldots, X_n) = (x_1, x_2, \ldots, x_n)\} = p(x_1, x_2, \ldots, x_n)$, $(x_1, x_2, \ldots, x_n) \in \mathcal{X}^n$ for $n = 1, 2, \ldots$.

Definition A stochastic process is said to be *stationary* if the joint distribution of any subset of the sequence of random variables is invariant with respect to shifts in the time index; that is,

$$\begin{aligned}\Pr\{X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n\}\\ = \Pr\{X_{1+l} = x_1, X_{2+l} = x_2, \ldots, X_{n+l} = x_n\} \quad (4.1)\end{aligned}$$

for every n and every shift l and for all $x_1, x_2, \ldots, x_n \in \mathcal{X}$.

Elements of Information Theory, Second Edition, By Thomas M. Cover and Joy A. Thomas
Copyright © 2006 John Wiley & Sons, Inc.

A simple example of a stochastic process with dependence is one in which each random variable depends only on the one preceding it and is *conditionally* independent of all the other preceding random variables. Such a process is said to be Markov.

Definition A discrete stochastic process $X_1, X_2, \ldots$ is said to be a *Markov chain* or a *Markov process* if for $n = 1, 2, \ldots,$

$$\begin{aligned}\Pr(X_{n+1} = x_{n+1} | X_n = x_n, X_{n-1} = x_{n-1}, \ldots, X_1 = x_1) \\ = \Pr(X_{n+1} = x_{n+1} | X_n = x_n)\end{aligned} \tag{4.2}$$

for all $x_1, x_2, \ldots, x_n, x_{n+1} \in \mathcal{X}$.

In this case, the joint probability mass function of the random variables can be written as

$$p(x_1, x_2, \ldots, x_n) = p(x_1)p(x_2|x_1)p(x_3|x_2)\cdots p(x_n|x_{n-1}). \tag{4.3}$$

Definition The Markov chain is said to be *time invariant* if the conditional probability $p(x_{n+1}|x_n)$ does not depend on n; that is, for $n = 1, 2, \ldots,$

$$\Pr\{X_{n+1} = b|X_n = a\} = \Pr\{X_2 = b|X_1 = a\} \quad \text{for all } a, b \in \mathcal{X}. \tag{4.4}$$

We will assume that the Markov chain is time invariant unless otherwise stated.

If $\{X_i\}$ is a Markov chain, X_n is called the *state* at time n. A time-invariant Markov chain is characterized by its initial state and a *probability transition matrix* $P = [P_{ij}]$, $i, j \in \{1, 2, \ldots, m\}$, where $P_{ij} = \Pr\{X_{n+1} = j|X_n = i\}$.

If it is possible to go with positive probability from any state of the Markov chain to any other state in a finite number of steps, the Markov chain is said to be *irreducible*. If the largest common factor of the lengths of different paths from a state to itself is 1, the Markov chain is said to *aperiodic*.

If the probability mass function of the random variable at time n is $p(x_n)$, the probability mass function at time $n + 1$ is

$$p(x_{n+1}) = \sum_{x_n} p(x_n)P_{x_n x_{n+1}}. \tag{4.5}$$

A distribution on the states such that the distribution at time $n + 1$ is the same as the distribution at time n is called a *stationary distribution*. The where $p_i = P(X_i = 1)$ is not constant but a function of i, chosen carefully so that the limit in (4.10) does not exist. For example, let

$$p_i = \begin{cases} 0.5 & \text{if } 2k < \log\log i \le 2k+1, \\ 0 & \text{if } 2k+1 < \log\log i \le 2k+2 \end{cases} \tag{4.13}$$

for $k = 0, 1, 2, \ldots$.

Then there are arbitrarily long stretches where $H(X_i) = 1$, followed by exponentially longer segments where $H(X_i) = 0$. Hence, the running average of the $H(X_i)$ will oscillate between 0 and 1 and will not have a limit. Thus, $H(\mathcal{X})$ is not defined for this process.

We can also define a related quantity for entropy rate:

$$H'(\mathcal{X}) = \lim_{n\to\infty} H(X_n|X_{n-1}, X_{n-2}, \ldots, X_1) \tag{4.14}$$

when the limit exists.

The two quantities $H(\mathcal{X})$ and $H'(\mathcal{X})$ correspond to two different notions of entropy rate. The first is the per symbol entropy of the n random variables, and the second is the conditional entropy of the last random variable given the past. We now prove the important result that for stationary processes both limits exist and are equal.

Theorem 4.2.1 *For a stationary stochastic process, the limits in (4.10) and (4.14) exist and are equal:*

$$H(\mathcal{X}) = H'(\mathcal{X}). \tag{4.15}$$

We first prove that $\lim H(X_n|X_{n-1}, \ldots, X_1)$ exists.

Theorem 4.2.2 *For a stationary stochastic process, $H(X_n|X_{n-1}, \ldots, X_1)$ is nonincreasing in n and has a limit $H'(\mathcal{X})$.*

Proof

$$H(X_{n+1}|X_1, X_2, \ldots, X_n) \le H(X_{n+1}|X_n, \ldots, X_2) \tag{4.16}$$

$$= H(X_n|X_{n-1}, \ldots, X_1), \tag{4.17}$$

where the inequality follows from the fact that conditioning reduces entropy and the equality follows from the stationarity of the process. Since $H(X_n|X_{n-1}, \ldots, X_1)$ is a decreasing sequence of nonnegative numbers, it has a limit, $H'(\mathcal{X})$. □

We now use the following simple result from analysis.

Theorem 4.2.3 (*Cesáro mean*) *If $a_n \to a$ and $b_n = \frac{1}{n}\sum_{i=1}^{n} a_i$, then $b_n \to a$.*

Proof: (*Informal outline*). Since most of the terms in the sequence $\{a_k\}$ are eventually close to a, then b_n, which is the average of the first n terms, is also eventually close to a.

Formal Proof: Let $\epsilon > 0$. Since $a_n \to a$, there exists a number $N(\epsilon)$ such that $|a_n - a| \le \epsilon$ for all $n \ge N(\epsilon)$. Hence,

$$|b_n - a| = \left|\frac{1}{n}\sum_{i=1}^{n}(a_i - a)\right| \tag{4.18}$$

$$\le \frac{1}{n}\sum_{i=1}^{n}|(a_i - a)| \tag{4.19}$$

$$\le \frac{1}{n}\sum_{i=1}^{N(\epsilon)}|a_i - a| + \frac{n - N(\epsilon)}{n}\epsilon \tag{4.20}$$

$$\le \frac{1}{n}\sum_{i=1}^{N(\epsilon)}|a_i - a| + \epsilon \tag{4.21}$$

for all $n \ge N(\epsilon)$. Since the first term goes to 0 as $n \to \infty$, we can make $|b_n - a| \le 2\epsilon$ by taking n large enough. Hence, $b_n \to a$ as $n \to \infty$. □

Proof of Theorem 4.2.1: By the chain rule,

$$\frac{H(X_1, X_2, \ldots, X_n)}{n} = \frac{1}{n}\sum_{i=1}^{n} H(X_i|X_{i-1}, \ldots, X_1), \tag{4.22}$$

that is, the entropy rate is the time average of the conditional entropies. But we know that the conditional entropies tend to a limit H'. Hence, by Theorem 4.2.3, their running average has a limit, which is equal to the limit H' of the terms. Thus, by Theorem 4.2.2,

$$\begin{aligned} H(\mathcal{X}) &= \lim \frac{H(X_1, X_2, \ldots, X_n)}{n} = \lim H(X_n|X_{n-1}, \ldots, X_1) \\ &= H'(\mathcal{X}). \qquad \square \end{aligned} \tag{4.23}$$

The significance of the entropy rate of a stochastic process arises from the AEP for a stationary ergodic process. We prove the general AEP in Section 16.8, where we show that for any stationary ergodic process,

$$-\frac{1}{n}\log p(X_1, X_2, \ldots, X_n) \rightarrow H(\mathcal{X}) \tag{4.24}$$

with probability 1. Using this, the theorems of Chapter 3 can easily be extended to a general stationary ergodic process. We can define a typical set in the same way as we did for the i.i.d. case in Chapter 3. By the same arguments, we can show that the typical set has a probability close to 1 and that there are about $2^{nH(\mathcal{X})}$ typical sequences of length n, each with probability about $2^{-nH(\mathcal{X})}$. We can therefore represent the typical sequences of length n using approximately $nH(\mathcal{X})$ bits. This shows the significance of the entropy rate as the average description length for a stationary ergodic process.

The entropy rate is well defined for all stationary processes. The entropy rate is particularly easy to calculate for Markov chains.

Markov Chains. For a stationary Markov chain, the entropy rate is given by

$$\begin{aligned} H(\mathcal{X}) = H'(\mathcal{X}) &= \lim H(X_n|X_{n-1}, \ldots, X_1) = \lim H(X_n|X_{n-1}) \\ &= H(X_2|X_1), \end{aligned} \tag{4.25}$$

where the conditional entropy is calculated using the given stationary distribution. Recall that the stationary distribution μ is the solution of the equations

$$\mu_j = \sum_i \mu_i P_{ij} \quad \text{for all } j. \tag{4.26}$$

We express the conditional entropy explicitly in the following theorem.

Theorem 4.2.4 *Let $\{X_i\}$ be a stationary Markov chain with stationary distribution μ and transition matrix P. Let $X_1 \sim \mu$. Then the entropy rate is*

$$H(\mathcal{X}) = -\sum_{ij} \mu_i P_{ij} \log P_{ij}. \tag{4.27}$$

Proof: $H(\mathcal{X}) = H(X_2|X_1) = \sum_i \mu_i \left(\sum_j -P_{ij}\log P_{ij}\right)$. □

Example 4.2.1 (*Two-state Markov chain*) The entropy rate of the two-state Markov chain in Figure 4.1 is

$$H(\mathcal{X}) = H(X_2|X_1) = \frac{\beta}{\alpha+\beta}H(\alpha) + \frac{\alpha}{\alpha+\beta}H(\beta). \tag{4.28}$$

Remark If the Markov chain is irreducible and aperiodic, it has a unique stationary distribution on the states, and any initial distribution tends to the stationary distribution as $n \to \infty$. In this case, even though the initial distribution is not the stationary distribution, the entropy rate, which is defined in terms of long-term behavior, is $H(\mathcal{X})$, as defined in (4.25) and (4.27).

4.3 EXAMPLE: ENTROPY RATE OF A RANDOM WALK ON A WEIGHTED GRAPH

As an example of a stochastic process, let us consider a random walk on a connected graph (Figure 4.2). Consider a graph with m nodes labeled $\{1, 2, \ldots, m\}$, with weight $W_{ij} \geq 0$ on the edge joining node i to node j. (The graph is assumed to be undirected, so that $W_{ij} = W_{ji}$. We set $W_{ij} = 0$ if there is no edge joining nodes i and j.)

A particle walks randomly from node to node in this graph. The random walk $\{X_n\}$, $X_n \in \{1, 2, \ldots, m\}$, is a sequence of vertices of the graph. Given $X_n = i$, the next vertex j is chosen from among the nodes connected to node i with a probability proportional to the weight of the edge connecting i to j. Thus, $P_{ij} = W_{ij}/\sum_k W_{ik}$.

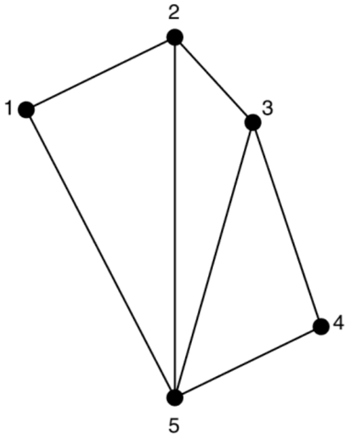


FIGURE 4.2. Random walk on a graph.

In this case, the stationary distribution has a surprisingly simple form, which we will guess and verify. The stationary distribution for this Markov chain assigns probability to node i proportional to the total weight of the edges emanating from node i. Let

$$W_i = \sum_j W_{ij} \tag{4.29}$$

be the total weight of edges emanating from node i, and let

$$W = \sum_{i,j:j>i} W_{ij} \tag{4.30}$$

be the sum of the weights of all the edges. Then $\sum_i W_i = 2W$.

We now guess that the stationary distribution is

$$\mu_i = \frac{W_i}{2W}. \tag{4.31}$$

We verify that this is the stationary distribution by checking that $\mu P = \mu$. Here

$$\begin{aligned}
\sum_i \mu_i P_{ij} &= \sum_i \frac{W_i}{2W}\frac{W_{ij}}{W_i} && (4.32)\\
&= \sum_i \frac{1}{2W} W_{ij} && (4.33)\\
&= \frac{W_j}{2W} && (4.34)\\
&= \mu_j. && (4.35)
\end{aligned}$$

Thus, the stationary probability of state i is proportional to the weight of edges emanating from node i. This stationary distribution has an interesting property of locality: It depends only on the total weight and the weight of edges connected to the node and hence does not change if the weights in some other part of the graph are changed while keeping the total weight constant. We can now calculate the entropy rate as

$$\begin{aligned}
H(\mathcal{X}) &= H(X_2|X_1) && (4.36)\\
&= -\sum_i \mu_i \sum_j P_{ij} \log P_{ij} && (4.37)
\end{aligned}$$

$$= D(p(x_{n+1})||q(x_{n+1}))$$
$$+ D(p(x_n|x_{n+1})||q(x_n|x_{n+1})).$$

Since both p and q are derived from the Markov chain, the conditional probability mass functions $p(x_{n+1}|x_n)$ and $q(x_{n+1}|x_n)$ are both equal to $r(x_{n+1}|x_n)$, and hence $D(p(x_{n+1}|x_n)||q(x_{n+1}|x_n)) = 0$. Now using the nonnegativity of $D(p(x_n|x_{n+1})||q(x_n|x_{n+1}))$ (Corollary to Theorem 2.6.3), we have

$$D(p(x_n)||q(x_n)) \geq D(p(x_{n+1})||q(x_{n+1})) \tag{4.44}$$

or

$$D(\mu_n||\mu'_n) \geq D(\mu_{n+1}||\mu'_{n+1}). \tag{4.45}$$

Consequently, the distance between the probability mass functions is decreasing with time n for any Markov chain.

An example of one interpretation of the preceding inequality is to suppose that the tax system for the redistribution of wealth is the same in Canada and in England. Then if μ_n and μ'_n represent the distributions of wealth among people in the two countries, this inequality shows that the relative entropy distance between the two distributions decreases with time. The wealth distributions in Canada and England become more similar.

2. *Relative entropy $D(\mu_n||\mu)$ between a distribution μ_n on the states at time n and a stationary distribution μ decreases with n.* In (4.45), μ'_n is any distribution on the states at time n. If we let μ'_n be any stationary distribution μ, the distribution μ'_{n+1} at the next time is also equal to μ. Hence,

$$D(\mu_n||\mu) \geq D(\mu_{n+1}||\mu), \tag{4.46}$$

which implies that any state distribution gets closer and closer to each stationary distribution as time passes. The sequence $D(\mu_n||\mu)$ is a monotonically nonincreasing nonnegative sequence and must therefore have a limit. The limit is zero if the stationary distribution is unique, but this is more difficult to prove.

3. *Entropy increases if the stationary distribution is uniform.* In general, the fact that the relative entropy decreases does not imply that the entropy increases. A simple counterexample is provided by any Markov chain with a nonuniform stationary distribution. If we start

this Markov chain from the uniform distribution, which already is the maximum entropy distribution, the distribution will tend to the stationary distribution, which has a lower entropy than the uniform. Here, the entropy decreases with time.

If, however, the stationary distribution is the uniform distribution, we can express the relative entropy as

$$D(\mu_n||\mu) = \log|\mathcal{X}| - H(\mu_n) = \log|\mathcal{X}| - H(X_n). \tag{4.47}$$

In this case the monotonic decrease in relative entropy implies a monotonic increase in entropy. This is the explanation that ties in most closely with statistical thermodynamics, where all the microstates are equally likely. We now characterize processes having a uniform stationary distribution.

Definition A probability transition matrix $[P_{ij}]$, $P_{ij} = \Pr\{X_{n+1} = j|X_n = i\}$, is called *doubly stochastic* if

$$\sum_i P_{ij} = 1, \qquad j = 1, 2, \ldots \tag{4.48}$$

and

$$\sum_j P_{ij} = 1, \qquad i = 1, 2, \ldots. \tag{4.49}$$

Remark The uniform distribution is a stationary distribution of P if and only if the probability transition matrix is doubly stochastic (see Problem 4.1).

4. *The conditional entropy $H(X_n|X_1)$ increases with n for a stationary Markov process*. If the Markov process is stationary, $H(X_n)$ is constant. So the entropy is nonincreasing. However, we will prove that $H(X_n|X_1)$ increases with n. Thus, the conditional uncertainty of the future increases. We give two alternative proofs of this result. First, we use the properties of entropy,

$$H(X_n|X_1) \geq H(X_n|X_1, X_2) \quad \text{(conditioning reduces entropy)} \tag{4.50}$$

$$= H(X_n|X_2) \quad \text{(by Markovity)} \tag{4.51}$$

$$= H(X_{n-1}|X_1) \quad \text{(by stationarity).} \tag{4.52}$$

Alternatively, by an application of the data-processing inequality to the Markov chain $X_1 \to X_{n-1} \to X_n$, we have

$$I(X_1; X_{n-1}) \geq I(X_1; X_n). \tag{4.53}$$

Expanding the mutual informations in terms of entropies, we have

$$H(X_{n-1}) - H(X_{n-1}|X_1) \geq H(X_n) - H(X_n|X_1). \tag{4.54}$$

By stationarity, $H(X_{n-1}) = H(X_n)$, and hence we have

$$H(X_{n-1}|X_1) \leq H(X_n|X_1). \tag{4.55}$$

[These techniques can also be used to show that $H(X_0|X_n)$ is increasing in n for any Markov chain.]

5. *Shuffles increase entropy*. If T is a shuffle (permutation) of a deck of cards and X is the initial (random) position of the cards in the deck, and if the choice of the shuffle T is independent of X, then

$$H(TX) \geq H(X), \tag{4.56}$$

where TX is the permutation of the deck induced by the shuffle T on the initial permutation X. Problem 4.3 outlines a proof.

4.5 FUNCTIONS OF MARKOV CHAINS

Here is an example that can be very difficult if done the wrong way. It illustrates the power of the techniques developed so far. Let $X_1, X_2, \ldots, X_n, \ldots$ be a stationary Markov chain, and let $Y_i = \phi(X_i)$ be a process each term of which is a function of the corresponding state in the Markov chain. What is the entropy rate $H(\mathcal{Y})$? Such functions of Markov chains occur often in practice. In many situations, one has only partial information about the state of the system. It would simplify matters greatly if $Y_1, Y_2, \ldots, Y_n$ also formed a Markov chain, but in many cases, this is not true. Since the Markov chain is stationary, so is $Y_1, Y_2, \ldots, Y_n$, and the entropy rate is well defined. However, if we wish to compute $H(\mathcal{Y})$, we might compute $H(Y_n|Y_{n-1}, \ldots, Y_1)$ for each n and find the limit. Since the convergence can be arbitrarily slow, we will never know how close we are to the limit. (We can't look at the change between the values at n and $n + 1$, since this difference may be small even when we are far away from the limit—consider, for example, $\sum \frac{1}{n}$.)

It would be useful computationally to have upper and lower bounds converging to the limit from above and below. We can halt the computation when the difference between upper and lower bounds is small, and we will then have a good estimate of the limit.

We already know that $H(Y_n|Y_{n-1},\ldots,Y_1)$ converges monotonically to $H(\mathcal{Y})$ from above. For a lower bound, we will use $H(Y_n|Y_{n-1},\ldots,Y_1,X_1)$. This is a neat trick based on the idea that X_1 contains as much information about Y_n as $Y_1, Y_0, Y_{-1}, \ldots$.

Lemma 4.5.1

$$H(Y_n|Y_{n-1},\ldots,Y_2,X_1) \le H(\mathcal{Y}). \tag{4.57}$$

Proof: We have for $k = 1, 2, \ldots,$

$$H(Y_n|Y_{n-1},\ldots,Y_2,X_1) \overset{(a)}{=} H(Y_n|Y_{n-1},\ldots,Y_2,Y_1,X_1) \tag{4.58}$$

$$\overset{(b)}{=} H(Y_n|Y_{n-1},\ldots,Y_1,X_1,X_0,X_{-1},\ldots,X_{-k}) \tag{4.59}$$

$$\overset{(c)}{=} H(Y_n|Y_{n-1},\ldots,Y_1,X_1,X_0,X_{-1},\ldots,X_{-k},Y_0,\ldots,Y_{-k}) \tag{4.60}$$

$$\overset{(d)}{\le} H(Y_n|Y_{n-1},\ldots,Y_1,Y_0,\ldots,Y_{-k}) \tag{4.61}$$

$$\overset{(e)}{=} H(Y_{n+k+1}|Y_{n+k},\ldots,Y_1), \tag{4.62}$$

where (a) follows from that fact that Y_1 is a function of X_1, and (b) follows from the Markovity of X, (c) follows from the fact that Y_i is a function of X_i, (d) follows from the fact that conditioning reduces entropy, and (e) follows by stationarity. Since the inequality is true for all k, it is true in the limit. Thus,

$$H(Y_n|Y_{n-1},\ldots,Y_1,X_1) \le \lim_k H(Y_{n+k+1}|Y_{n+k},\ldots,Y_1) \tag{4.63}$$

$$= H(\mathcal{Y}). \qquad \square \tag{4.64}$$

The next lemma shows that the interval between the upper and the lower bounds decreases in length.

Lemma 4.5.2

$$H(Y_n|Y_{n-1},\ldots,Y_1) - H(Y_n|Y_{n-1},\ldots,Y_1,X_1) \to 0. \tag{4.65}$$

Proof: The interval length can be rewritten as

$$\begin{aligned} & H(Y_n|Y_{n-1}, \ldots, Y_1) - H(Y_n|Y_{n-1}, \ldots, Y_1, X_1) \\ & \quad = I(X_1; Y_n|Y_{n-1}, \ldots, Y_1). \end{aligned} \tag{4.66}$$

By the properties of mutual information,

$$I(X_1; Y_1, Y_2, \ldots, Y_n) \le H(X_1), \tag{4.67}$$

and $I(X_1; Y_1, Y_2, \ldots, Y_n)$ increases with n. Thus, $\lim I(X_1; Y_1, Y_2, \ldots, Y_n)$ exists and

$$\lim_{n\to\infty} I(X_1; Y_1, Y_2, \ldots, Y_n) \le H(X_1). \tag{4.68}$$

By the chain rule,

$$\begin{aligned} H(X) &\ge \lim_{n\to\infty} I(X_1; Y_1, Y_2, \ldots, Y_n) && (4.69) \\ &= \lim_{n\to\infty} \sum_{i=1}^{n} I(X_1; Y_i|Y_{i-1}, \ldots, Y_1) && (4.70) \\ &= \sum_{i=1}^{\infty} I(X_1; Y_i|Y_{i-1}, \ldots, Y_1). && (4.71) \end{aligned}$$

Since this infinite sum is finite and the terms are nonnegative, the terms must tend to 0; that is,

$$\lim I(X_1; Y_n|Y_{n-1}, \ldots, Y_1) = 0, \tag{4.72}$$

which proves the lemma. □

Combining Lemmas 4.5.1 and 4.5.2, we have the following theorem.

Theorem 4.5.1 *If $X_1, X_2, \ldots, X_n$ form a stationary Markov chain, and $Y_i = \phi(X_i)$, then*

$$H(Y_n|Y_{n-1}, \ldots, Y_1, X_1) \le H(\mathcal{Y}) \le H(Y_n|Y_{n-1}, \ldots, Y_1) \tag{4.73}$$

and

$$\lim H(Y_n|Y_{n-1}, \ldots, Y_1, X_1) = H(\mathcal{Y}) = \lim H(Y_n|Y_{n-1}, \ldots, Y_1). \tag{4.74}$$

In general, we could also consider the case where Y_i is a stochastic function (as opposed to a deterministic function) of X_i. Consider a Markov

In other words, the present has a conditional entropy given the past equal to the conditional entropy given the future. This is true even though it is quite easy to concoct stationary random processes for which the flow into the future looks quite different from the flow into the past. That is, one can determine the direction of time by looking at a sample function of the process. Nonetheless, given the present state, the conditional uncertainty of the next symbol in the future is equal to the conditional uncertainty of the previous symbol in the past.

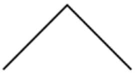
4.3 *Shuffles increase entropy*. Argue that for any distribution on shuffles T and any distribution on card positions X that

$$\begin{aligned} H(TX) &\geq H(TX|T) && (4.82)\\ &= H(T^{-1}TX|T) && (4.83)\\ &= H(X|T) && (4.84)\\ &= H(X) && (4.85)\end{aligned}$$

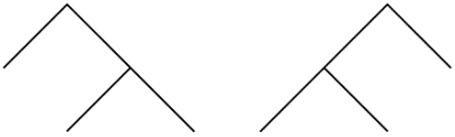
if X and T are independent.

4.4 *Second law of thermodynamics*. Let $X_1, X_2, X_3, \ldots$ be a stationary first-order Markov chain. In Section 4.4 it was shown that $H(X_n \mid X_1) \geq H(X_{n-1} \mid X_1)$ for $n = 2, 3, \ldots$. Thus, conditional uncertainty about the future grows with time. This is true although the unconditional uncertainty $H(X_n)$ remains constant. However, show by example that $H(X_n|X_1 = x_1)$ does not necessarily grow with n for every x_1.

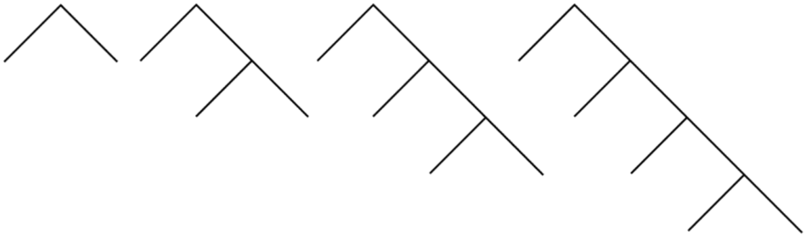
4.5 *Entropy of a random tree*. Consider the following method of generating a random tree with n nodes. First expand the root node:



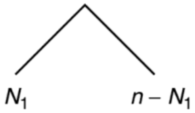
Then expand one of the two terminal nodes at random:



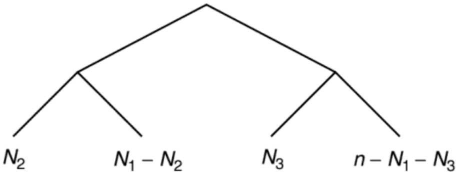
At time k, choose one of the $k - 1$ terminal nodes according to a uniform distribution and expand it. Continue until n terminal nodes have been generated. Thus, a sequence leading to a five-node tree might look like this:



Surprisingly, the following method of generating random trees yields the same probability distribution on trees with n terminal nodes. First choose an integer N_1 uniformly distributed on $\{1, 2, \ldots, n-1\}$. We then have the picture



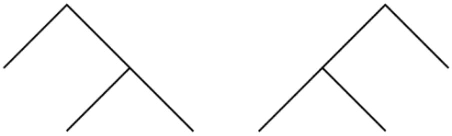
Then choose an integer N_2 uniformly distributed over $\{1, 2, \ldots, N_1 - 1\}$, and independently choose another integer N_3 uniformly over $\{1, 2, \ldots, (n - N_1) - 1\}$. The picture is now



Continue the process until no further subdivision can be made. (The equivalence of these two tree generation schemes follows, for example, from Polya's urn model.)

Now let T_n denote a random n-node tree generated as described. The probability distribution on such trees seems difficult to describe, but we can find the entropy of this distribution in recursive form.

First some examples. For $n = 2$, we have only one tree. Thus, $H(T_2) = 0$. For $n = 3$, we have two equally probable trees:



Thus, $H(T_3) = \log 2$. For $n = 4$, we have five possible trees, with probabilities $\frac{1}{3}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}$.
Now for the recurrence relation. Let $N_1(T_n)$ denote the number of terminal nodes of T_n in the right half of the tree. Justify each of the steps in the following:

$$\begin{aligned}
H(T_n) &\overset{\text{(a)}}{=} H(N_1, T_n) &(4.86)\\
&\overset{\text{(b)}}{=} H(N_1) + H(T_n|N_1) &(4.87)\\
&\overset{\text{(c)}}{=} \log(n-1) + H(T_n|N_1) &(4.88)\\
&\overset{\text{(d)}}{=} \log(n-1) + \frac{1}{n-1}\sum_{k=1}^{n-1}(H(T_k) + H(T_{n-k})) &(4.89)\\
&\overset{\text{(e)}}{=} \log(n-1) + \frac{2}{n-1}\sum_{k=1}^{n-1} H(T_k) &(4.90)\\
&= \log(n-1) + \frac{2}{n-1}\sum_{k=1}^{n-1} H_k. &(4.91)
\end{aligned}$$

(f) Use this to show that

$$(n-1)H_n = nH_{n-1} + (n-1)\log(n-1) - (n-2)\log(n-2) \tag{4.92}$$

or

$$\frac{H_n}{n} = \frac{H_{n-1}}{n-1} + c_n \tag{4.93}$$

for appropriately defined c_n. Since $\sum c_n = c < \infty$, you have proved that $\frac{1}{n}H(T_n)$ converges to a constant. Thus, the expected number of bits necessary to describe the random tree T_n grows linearly with n.

4.6 *Monotonicity of entropy per element*. For a stationary stochastic process $X_1, X_2, \ldots, X_n$, show that

(a)

$$\frac{H(X_1, X_2, \ldots, X_n)}{n} \le \frac{H(X_1, X_2, \ldots, X_{n-1})}{n-1}. \tag{4.94}$$

(b)

$$\frac{H(X_1, X_2, \ldots, X_n)}{n} \ge H(X_n|X_{n-1}, \ldots, X_1). \tag{4.95}$$

4.7 *Entropy rates of Markov chains*

(a) Find the entropy rate of the two-state Markov chain with transition matrix

$$P = \begin{bmatrix} 1 - p_{01} & p_{01} \\ p_{10} & 1 - p_{10} \end{bmatrix}.$$

(b) What values of p_{01}, p_{10} maximize the entropy rate?

(c) Find the entropy rate of the two-state Markov chain with transition matrix

$$P = \begin{bmatrix} 1 - p & p \\ 1 & 0 \end{bmatrix}.$$

(d) Find the maximum value of the entropy rate of the Markov chain of part (c). We expect that the maximizing value of p should be less than $\frac{1}{2}$, since the 0 state permits more information to be generated than the 1 state.

(e) Let $N(t)$ be the number of allowable state sequences of length t for the Markov chain of part (c). Find $N(t)$ and calculate

$$H_0 = \lim_{t \to \infty} \frac{1}{t} \log N(t).$$

[*Hint*: Find a linear recurrence that expresses $N(t)$ in terms of $N(t-1)$ and $N(t-2)$. Why is H_0 an upper bound on the entropy rate of the Markov chain? Compare H_0 with the maximum entropy found in part (d).]

4.8 *Maximum entropy process*. A discrete memoryless source has the alphabet $\{1, 2\}$, where the symbol 1 has duration 1 and the symbol 2 has duration 2. The probabilities of 1 and 2 are p_1 and p_2, respectively. Find the value of p_1 that maximizes the source entropy per unit time $H(\mathcal{X}) = \frac{H(X)}{ET}$. What is the maximum value $H(\mathcal{X})$?

4.9 *Initial conditions*. Show, for a Markov chain, that

$$H(X_0|X_n) \geq H(X_0|X_{n-1}).$$

Thus, initial conditions X_0 become more difficult to recover as the future X_n unfolds.

4.10 *Pairwise independence*. Let $X_1, X_2, \ldots, X_{n-1}$ be i.i.d. random variables taking values in $\{0, 1\}$, with $\Pr\{X_i = 1\} = \frac{1}{2}$. Let $X_n = 1$ if $\sum_{i=1}^{n-1} X_i$ is odd and $X_n = 0$ otherwise. Let $n \geq 3$.