ELEMENTS OF INFORMATION THEORY

Second Edition

THOMAS M. COVER

JOY A. THOMAS

WILEY-INTERSCIENCE

A JOHN WILEY & SONS, INC., PUBLICATION

ELEMENTS OF INFORMATION THEORY

ELEMENTS OF INFORMATION THEORY

Second Edition

THOMAS M. COVER

JOY A. THOMAS

WILEY-INTERSCIENCE

A JOHN WILEY & SONS, INC., PUBLICATION

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

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

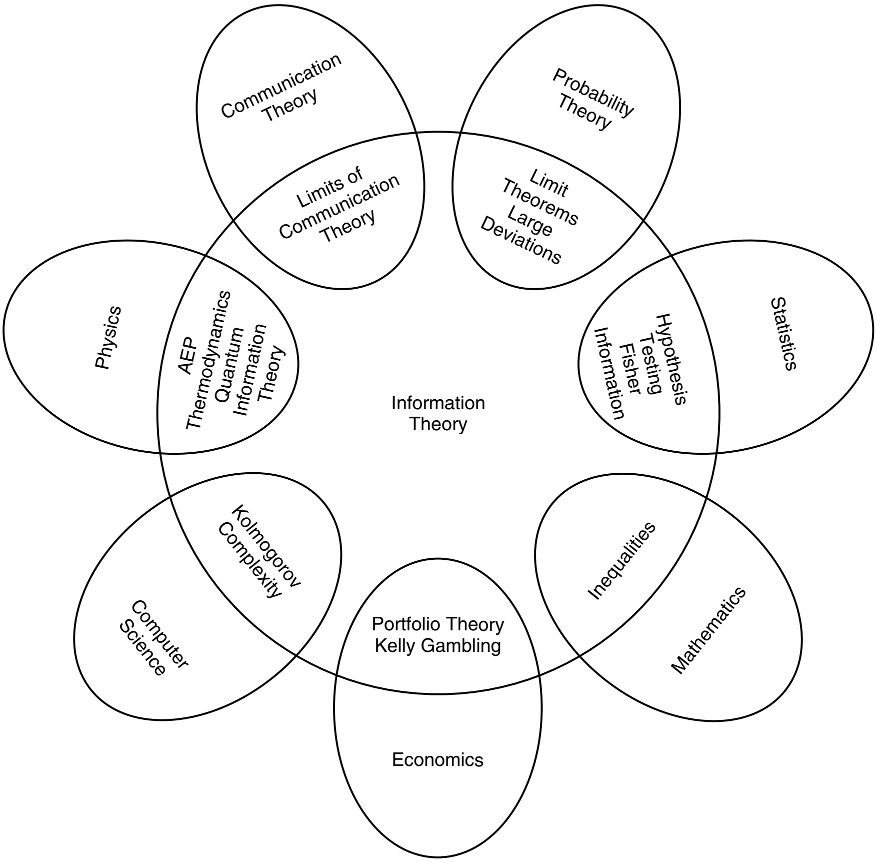

FIGURE 1.1. Relationship of information theory to other fields.

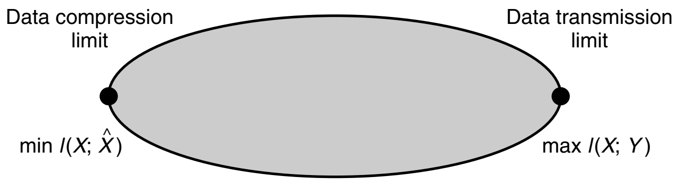

FIGURE 1.2. Information theory as the extreme points of communication theory.

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