

9463340

# Elements of Information Theory



0250  
C873

9463340

---

---

# Elements of Information Theory

---

---

THOMAS M. COVER

*Stanford University  
Stanford, California*

JOY A. THOMAS

*IBM T. J. Watson Research Center  
Yorktown Heights, New York*



E9463340

rsience Publication

Y & SONS, INC.

Chichester / Brisbane / Toronto / Singapore

In recognition of the importance of preserving what has been written, it is a policy of John Wiley & Sons, Inc., to have books of enduring value published in the United States printed on acid-free paper, and we exert our best efforts to that end.

Copyright © 1991 by John Wiley & Sons, Inc.

All rights reserved. Published simultaneously in Canada.

Reproduction or translation of any part of this work beyond that permitted by Section 107 or 108 of the 1976 United States Copyright Act without the permission of the copyright owner is unlawful. Requests for permission or further information should be addressed to the Permissions Department, John Wiley & Sons, Inc.

***Library of Congress Cataloging in Publication Data:***

Cover, T. M., 1938-

Elements of Information theory / Thomas M. Cover, Joy A. Thomas.

p. cm. -- (Wiley series in telecommunications)

"A Wiley-Interscience publication."

Includes bibliographical references and index.

ISBN 0-471-06259-6

1. Information theory. I. Thomas, Joy A. II. Title.

III. Series.

Q360.C68 1991

003'54--dc20

90-45484

CIP

Printed in the United States of America

10 9 8 7 6 5 4 3 2

Printed and bound by Courier Companies, Inc.

---

---

# **Elements of Information Theory**

---

---

0488340

# WILEY SERIES IN TELECOMMUNICATIONS

Donald L. Schilling, Editor  
City College of New York

*Digital Telephony, 2nd Edition*

John Bellamy

*Elements of Information Theory*

Thomas M. Cover and Joy A. Thomas

*Telecommunication System Engineering, 2nd Edition*

Roger L. Freeman

*Telecommunication Transmission Handbook, 3rd Edition*

Roger L. Freeman

*Introduction to Communications Engineering, 2nd Edition*

Robert M. Gagliardi

*Expert System Applications to Telecommunications*

Jay Liebowitz

*Synchronization in Digital Communications, Volume 1*

Heinrich Meyr and Gerd Ascheid

*Synchronization in Digital Communications, Volume 2*

Heinrich Meyr and Gerd Ascheid (in preparation)

*Computational Methods of Signal Recovery and Recognition*

Richard J. Mammone (in preparation)

*Business Earth Stations for Telecommunications*

Walter L. Morgan and Denis Rouffet

*Satellite Communications: The First Quarter Century of Service*

David W. E. Rees

*Worldwide Telecommunications Guide for the Business Manager*

Walter L. Vignault

---

---

# Preface

---

---

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 text book.

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, 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 theory.

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 introducing 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 so-called asymptotic equipartition property, which can be roughly paraphrased as "Almost everything is almost equally probable."

Chapter 2, which is the true first chapter of the subject, includes the basic algebraic relationships of entropy, relative entropy and mutual information as well as a discussion of the second law of thermodynamics and sufficient statistics. 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 fundamental idea of Kolmogorov complexity as an intellectual foundation for information theory is explored in Chapter 7. 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  $\Omega$  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, which is the fundamental theorem in information theory, is established in Chapter 8. The necessary material on differential entropy is developed in Chapter 9, 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 10.

The relationship between information theory and statistics, first studied by Kullback in the early 1950s, and relatively neglected since, is developed in Chapter 12. Rate distortion theory requires a little more background than its noiseless data compression counterpart, which accounts for its placement as late as Chapter 13 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 14. Many new ideas come into play in network information theory. The primary new ingredients are interference and feedback. Chapter 15 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 16, 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 senior year 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 15. 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.

Each chapter ends with a brief telegraphic summary of the key results. These summaries, in equation form, do not include the qualifying conditions. At the end of each we have included a variety of problems followed by brief historical notes describing the origins of the main results. The bibliography at the end of the book includes many of the key papers in the area and pointers to other books and survey papers on the subject.

The essential vitamins are contained in Chapters 2, 3, 4, 5, 8, 9, 10, 12, 13 and 14. This subset of chapters can be read without reference to the others and makes a good core of understanding. In our opinion, Chapter 7 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

---

---

# Acknowledgments

---

---

We wish to thank everyone who helped make this book what it is. In particular, 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. We were pleased to use twelve of his problems in the text. Aaron Wyner donated his new proof with Ziv on the convergence of the Lempel-Ziv algorithm. We would also like to thank Norman Abramson, Ed van der Meulen, Jack Salz and Raymond Yeung for their suggestions.

Certain key visitors and research associates contributed as well, including Amir Dembo, Paul Algoet, Hirosuke Yamamoto, Ben Kawabata, Makoto 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 the 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 encouragement and support during the final stages of production of this book.

TOM COVER  
JOY THOMAS

---

---

# Contents

---

---

<b>List of Figures</b>	<b>xix</b>
<b>1 Introduction and Preview</b>	<b>1</b>
1.1 Preview of the book / 5	
<b>2 Entropy, Relative Entropy and Mutual Information</b>	<b>12</b>
2.1 Entropy / 12	
2.2 Joint entropy and conditional entropy / 15	
2.3 Relative entropy and mutual information / 18	
2.4 Relationship between entropy and mutual information / 19	
2.5 Chain rules for entropy, relative entropy and mutual information / 21	
2.6 Jensen's inequality and its consequences / 23	
2.7 The log sum inequality and its applications / 29	
2.8 Data processing inequality / 32	
2.9 The second law of thermodynamics / 33	
2.10 Sufficient statistics / 36	
2.11 Fano's inequality / 38	
Summary of Chapter 2 / 40	
Problems for Chapter 2 / 42	
Historical notes / 49	
<b>3 The Asymptotic Equipartition Property</b>	<b>50</b>
3.1 The AEP / 51	

3.2	Consequences of the AEP: data compression / 53	
3.3	High probability sets and the typical set / 55	
	Summary of Chapter 3 / 56	
	Problems for Chapter 3 / 57	
	Historical notes / 59	
<b>4</b>	<b>Entropy Rates of a Stochastic Process</b>	<b>60</b>
4.1	Markov chains / 60	
4.2	Entropy rate / 63	
4.3	Example: Entropy rate of a random walk on a weighted graph / 66	
4.4	Hidden Markov models / 69	
	Summary of Chapter 4 / 71	
	Problems for Chapter 4 / 72	
	Historical notes / 77	
<b>5</b>	<b>Data Compression</b>	<b>78</b>
5.1	Examples of codes / 79	
5.2	Kraft inequality / 82	
5.3	Optimal codes / 84	
5.4	Bounds on the optimal codelength / 87	
5.5	Kraft inequality for uniquely decodable codes / 90	
5.6	Huffman codes / 92	
5.7	Some comments on Huffman codes / 94	
5.8	Optimality of Huffman codes / 97	
5.9	Shannon-Fano-Elias coding / 101	
5.10	Arithmetic coding / 104	
5.11	Competitive optimality of the Shannon code / 107	
5.12	Generation of discrete distributions from fair coins / 110	
	Summary of Chapter 5 / 117	
	Problems for Chapter 5 / 118	
	Historical notes / 124	
<b>6</b>	<b>Gambling and Data Compression</b>	<b>125</b>
6.1	The horse race / 125	
6.2	Gambling and side information / 130	
6.3	Dependent horse races and entropy rate / 131	
6.4	The entropy of English / 133	
6.5	Data compression and gambling / 136	

- 6.6 Gambling estimate of the entropy of English / 138
- Summary of Chapter 6 / 140
- Problems for Chapter 6 / 141
- Historical notes / 143

## 7 Kolmogorov Complexity

144

- 7.1 Models of computation / 146
- 7.2. Kolmogorov complexity: definitions and examples / 147
- 7.3 Kolmogorov complexity and entropy / 153
- 7.4 Kolmogorov complexity of integers / 155
- 7.5 Algorithmically random and incompressible sequences / 156
- 7.6 Universal probability / 160
- 7.7 The halting problem and the non-computability of Kolmogorov complexity / 162
- 7.8  $\Omega$  / 164
- 7.9 Universal gambling / 166
- 7.10 Occam's razor / 168
- 7.11 Kolmogorov complexity and universal probability / 169
- 7.12 The Kolmogorov sufficient statistic / 175
- Summary of Chapter 7 / 178
- Problems for Chapter 7 / 180
- Historical notes / 182

## 8 Channel Capacity

183

- 8.1 Examples of channel capacity / 184
- 8.2 Symmetric channels / 189
- 8.3 Properties of channel capacity / 190
- 8.4 Preview of the channel coding theorem / 191
- 8.5 Definitions / 192
- 8.6 Jointly typical sequences / 194
- 8.7 The channel coding theorem / 198
- 8.8 Zero-error codes / 203
- 8.9 Fano's inequality and the converse to the coding theorem / 204
- 8.10 Equality in the converse to the channel coding theorem / 207
- 8.11 Hamming codes / 209
- 8.12 Feedback capacity / 212

8.13	The joint source channel coding theorem / 215	
	Summary of Chapter 8 / 218	
	Problems for Chapter 8 / 220	
	Historical notes / 222	
<b>9</b>	<b>Differential Entropy</b>	<b>224</b>
9.1	Definitions / 224	
9.2	The AEP for continuous random variables / 225	
9.3	Relation of differential entropy to discrete entropy / 228	
9.4	Joint and conditional differential entropy / 229	
9.5	Relative entropy and mutual information / 231	
9.6	Properties of differential entropy, relative entropy and mutual information / 232	
9.7	Differential entropy bound on discrete entropy / 234	
	Summary of Chapter 9 / 236	
	Problems for Chapter 9 / 237	
	Historical notes / 238	
<b>10</b>	<b>The Gaussian Channel</b>	<b>239</b>
10.1	The Gaussian channel: definitions / 241	
10.2	Converse to the coding theorem for Gaussian channels / 245	
10.3	Band-limited channels / 247	
10.4	Parallel Gaussian channels / 250	
10.5	Channels with colored Gaussian noise / 253	
10.6	Gaussian channels with feedback / 256	
	Summary of Chapter 10 / 262	
	Problems for Chapter 10 / 263	
	Historical notes / 264	
<b>11</b>	<b>Maximum Entropy and Spectral Estimation</b>	<b>266</b>
11.1	Maximum entropy distributions / 266	
11.2	Examples / 268	
11.3	An anomalous maximum entropy problem / 270	
11.4	Spectrum estimation / 272	
11.5	Entropy rates of a Gaussian process / 273	
11.6	Burg's maximum entropy theorem / 274	
	Summary of Chapter 11 / 277	
	Problems for Chapter 11 / 277	
	Historical notes / 278	

<b>12</b>	<b>Information Theory and Statistics</b>	<b>279</b>
12.1	The method of types / 279	
12.2	The law of large numbers / 286	
12.3	Universal source coding / 288	
12.4	Large deviation theory / 291	
12.5	Examples of Sanov's theorem / 294	
12.6	The conditional limit theorem / 297	
12.7	Hypothesis testing / 304	
12.8	Stein's lemma / 309	
12.9	Chernoff bound / 312	
12.10	Lempel-Ziv coding / 319	
12.11	Fisher information and the Cramér-Rao inequality / 326	
	Summary of Chapter 12 / 331	
	Problems for Chapter 12 / 333	
	Historical notes / 335	
<b>13</b>	<b>Rate Distortion Theory</b>	<b>336</b>
13.1	Quantization / 337	
13.2	Definitions / 338	
13.3	Calculation of the rate distortion function / 342	
13.4	Converse to the rate distortion theorem / 349	
13.5	Achievability of the rate distortion function / 351	
13.6	Strongly typical sequences and rate distortion / 358	
13.7	Characterization of the rate distortion function / 362	
13.8	<u>Computation of channel capacity and the rate distortion function</u> / 364	
	Summary of Chapter 13 / 367	
	Problems for Chapter 13 / 368	
	Historical notes / 372	
<b>14</b>	<b>Network Information Theory</b>	<b>374</b>
14.1	Gaussian multiple user channels / 377	
14.2	Jointly typical sequences / 384	
14.3	The multiple access channel / 388	
14.4	Encoding of correlated sources / 407	
14.5	Duality between Slepian-Wolf encoding and multiple access channels / 416	
14.6	The broadcast channel / 418	
14.7	The relay channel / 428	



14.8	Source coding with side information / 432	
14.9	Rate distortion with side information / 438	
14.10	General multiterminal networks / 444	
	Summary of Chapter 14 / 450	
	Problems for Chapter 14 / 452	
	Historical notes / 457	
<b>15</b>	<b>Information Theory and the Stock Market</b>	<b>459</b>
15.1	The stock market: some definitions / 459	
15.2	Kuhn-Tucker characterization of the log-optimal portfolio / 462	
15.3	Asymptotic optimality of the log-optimal portfolio / 465	
15.4	Side information and the doubling rate / 467	
15.5	Investment in stationary markets / 469	
15.6	Competitive optimality of the log-optimal portfolio / 471	
15.7	The Shannon-McMillan-Breiman theorem / 474	
	Summary of Chapter 15 / 479	
	Problems for Chapter 15 / 480	
	Historical notes / 481	
<b>16</b>	<b>Inequalities in Information Theory</b>	<b>482</b>
16.1	Basic inequalities of information theory / 482	
16.2	Differential entropy / 485	
16.3	Bounds on entropy and relative entropy / 488	
16.4	Inequalities for types / 490	
16.5	Entropy rates of subsets / 490	
16.6	Entropy and Fisher information / 494	
16.7	The entropy power inequality and the Brunn-Minkowski inequality / 497	
16.8	Inequalities for determinants / 501	
16.9	Inequalities for ratios of determinants / 505	
	Overall Summary / 508	
	Problems for Chapter 16 / 509	
	Historical notes / 509	
	<b>Bibliography</b>	<b>510</b>
	<b>List of Symbols</b>	<b>526</b>
	<b>Index</b>	<b>529</b>