Entropy and Information Theory

Robert M. Gray

Entropy and Information Theory

Second Edition

Springer

Robert M. Gray
Department of Electrical Engineering
Stanford University
Stanford, CA 94305-9510
USA
rmgray@stanford.edu

ISBN 978-1-4419-7969-8 e-ISBN 978-1-4419-7970-4
DOI 10.1007/978-1-4419-7970-4
Springer New York Dordrecht Heidelberg London

Library of Congress Control Number: 2011920808

to Tim, Lori, Julia, Peter,
Gus, Amy, and Alice

and in memory of Tino

Preface

This book is devoted to the theory of probabilistic information measures and their application to coding theorems for information sources and noisy channels, with a strong emphasis on source coding and stationary codes. The eventual goal is a general development of Shannon's mathematical theory of communication for single user systems, but much of the space is devoted to the tools and methods required to prove the Shannon coding theorems, especially the notions of sources, channels, codes, entropy, information, and the entropy ergodic theorem. These tools form an area common to ergodic theory and information theory and comprise several quantitative notions of the information in random variables, random processes, and dynamical systems. Examples are entropy, mutual information, conditional entropy, conditional information, and relative entropy (discrimination, Kullback-Leibler information, informational divergence), along with the limiting normalized versions of these quantities such as entropy rate and information rate. In addition to information we will be concerned with the distance or distortion between the random objects, that is, the accuracy of the representation of one random object by another or the degree of mutual approximation. Much of the book is concerned with the properties of these quantities, especially the long term asymptotic behavior of average information and distortion, where both sample averages and probabilistic averages are of interest.

The book has been strongly influenced by M. S. Pinsker's classic *Information and Information Stability of Random Variables and Processes* and by the seminal work of A. N. Kolmogorov, I. M. Gelfand, A. M. Yaglom, and R. L. Dobrushin on information measures for abstract alphabets and their convergence properties. The book also has as a major influence the work of D.S. Ornstein on the isomorphism problem in ergodic theory, especially on his ideas of stationary codes mimicking block codes implied by the entropy ergodic theorem and of the d-bar distance between random processes. Many of the results herein are extensions of their

generalizations of Shannon's original results. The mathematical models adopted here are more general than traditional treatments in that nonstationary and nonergodic information processes are treated. The models are somewhat less general than those of the Russian school of information theory in the sense that standard alphabets rather than completely abstract alphabets are considered. This restriction, however, permits many stronger results as well as the extension to nonergodic processes. In addition, the assumption of standard spaces simplifies many proofs and such spaces include as examples virtually all examples of engineering interest.

The information convergence results are combined with ergodic theorems to prove general Shannon coding theorems for sources and channels. The results are not the most general known and the converses are not the strongest available in the literature, but they are sufficiently general to cover most sources and single-user communications systems encountered in applications and they are more general than those encountered in most modern texts. For example, most treatments confine interest to stationary and ergodic sources or even independent identically distributed (IID) sources and memoryless channels; here we consider asymptotic mean stationary sources, both one-sided and two-sided sources, and nonergodic sources. General channels with memory are considered, in particular the class of d-bar continuous channels.

Perhaps more important than the generality of the sources and channels is the variety of code structures considered. Most of the literature and virtually all of the texts on information theory focus exclusively on block codes, while many codes are more naturally described as a stationary or sliding-block code — a time-invariant possibly nonlinear filter, generally with a discrete output. Here the basic results of information theory are described for stationary or sliding-block codes as well as for the traditional block codes and the relationships between the two coding structures are explored in detail. Stationary codes arose in ergodic theory in the context of Ornstein's proof of the isomorphism theorem in the 1970s, and they arise naturally in the communications context of classical information theory, including common coding techniques such as time-invariant convolutional codes, predictive quantization, sigma-delta coding, and wavelet transform based techniques that operate as sliding-window or online filters rather than as block operations. Mathematically, stationary codes preserve many of the statistical properties of the source being coded such as stationarity, ergodicity, and mixing. In practice, stationary codes avoid the introduction of blocking artifacts not present in the original source.

This book can be considered as a sequel to my book *Probability, Random Processes, and Ergodic Properties* [58], as the first edition of this book was a sequel to the first edition [56]. There the prerequisite results on probability, standard spaces, and ordinary ergodic properties

may be found along with a development of the general sources considered (asymptotically mean stationary, not necessarily ergodic) and of the process distortion measures used here. This book is self contained with the exception of common (and a few less common) results which may be found in the first book. Results quoted from the first book are cited for both first and second editions as the numbering system in the two editions differs.

It is my hope that the book will interest engineers in some of the mathematical aspects and general models of the theory and mathematicians in some of the important engineering applications of performance bounds and code design for communication systems.

What's New in the Second Edition

As in the second edition of the companion volume [58], material has been corrected, rearranged, and rewritten in an effort to improve the flow of ideas and the presentation. This volume has been revised to reflect the changes in the companion volume, and citations to specific results are given for both the first and second editions [55, 58]. A significant amount of new material has been added both to expand some of the discussions to include more related topics and to include more recent results on old problems.

More general distortion measures are considered when treating the process distance and distortion measures, consistent with extensions or results in [55] on metric distortion measures to powers of metrics (such as the ubiquitous squared-error distortion) in [58].

Three new chapters have been added: one on the interplay between distortion and entropy, one on the interplay between distortion and information, and one on properties of good source codes — codes that are either optimal or asymptotically optimal in the sense of converging to the Shannon limit.

The chapter on distortion and entropy begins with a classic result treated in the first edition, the Fano inequality and its extensions, but it expands the discussion to consider the goodness of approximation of codes and their relation to entropy rate. Pinsker's classic result relating variation distance between probability measures and the divergence (Kullback-Leibler) distance is now treated along with its recent extension by Marton comparing Ornstein's d-bar process distance to divergence rate. The chapter contains a preliminary special case of the coding theorems to come — the application of the entropy ergodic theorem to the design of both block and sliding-block (stationary) almost lossless codes. The example introduces several basic ideas in a relatively simple context, including the construction of a sliding-block code from a block code in a

way that preserves the essential properties. The example also serves to illustrate the connections between information theory and ergodic theory by means of an interpretation of Ornstein's isomorphism theorem — which is not proved here — in terms of almost lossless stationary coding — which is. The results also provide insight into the close relationships between source coding or data compression and rate-constrained simulation of a stationary and ergodic process, the finding of a simple model based on coin flips that resembles as closely as possible the given process.

The chapter on distortion and information adds considerable material on rate-distortion theory to the treatment of the first edition, specifically on the evaluation of Shannon distortion-rate and rate-distortion functions along with their easy applications to lower bounds on performance in idealized communications systems. The fundamentals of Csiszár's variational approach based on the divergence inequality is described and some of the rarely noted attributes are pointed out. The implied algorithm for the evaluation of rate-distortion functions (originally due to Blahut [18]) is interpreted as an early example of alternating optimization.

An entirely new chapter on properties of good codes provides a development along the lines of Gersho and Gray [50] of the basic properties of optimal block codes originally due to Lloyd [110] and Steinhaus [175] along with the implied iterative design algorithm, another early example of alternating optimization. An incomplete extension of these block code optimality properties to sliding-block codes is described, and a simple example of trellis encoding is used to exemplify basic relations between block, sliding-block, and hybrid codes. The remainder of the chapter comprises recent developments in properties of asymptotically optimal sequences of sliding-block codes as developed by Mao, Gray, and Linder [117]. This material adds to the book's emphasis on stationary and sliding-block codes and adds to the limited literature on the subject.

Along with these major additions, I have added many minor results either because I was annoyed to discover they were not already in the first edition when I looked for them or because they eased the development of results.

The addition of three new chapters was partially balanced by the merging of two old chapters to better relate information rates for finite alphabet and continuous alphabet random processes.

Errors

Typographical and technical errors reported to or discovered by me during the two decades since the publication of the first edition have been

corrected and efforts have been made to improve formatting and appearance of the book. Doubtless with the inclusion of new material new errors have occurred. As I age my frequency of typographical and other errors seems to grow along with my ability to see through them. I apologize for any that remain in the book. I will keep a list of all errors found by me or sent to me at rmgray@stanford.edu and I will post the list at my Web site, http://ee.stanford.edu/~gray/.

Acknowledgments

The research in information theory that yielded many of the results and some of the new proofs for old results in this book was supported by the National Science Foundation. Portions of the research and much of the early writing were supported by a fellowship from the John Simon Guggenheim Memorial Foundation. Recent research and writing on some of these topics has been aided by gifts from Hewlett Packard, Inc.

The book benefited greatly from comments from numerous students and colleagues over many years; including Paul Shields, Paul Algoet, Ender Ayanoglu, Lee Davisson, John Kieffer, Dave Neuhoff, Don Ornstein, Bob Fontana, Jim Dunham, Farivar Saadat, Michael Sabin, Andrew Barron, Phil Chou, Tom Lookabaugh, Andrew Nobel, Bradley Dickinson, and Tamás Linder. I am grateful to Matt Shannon, Ricardo Blasco Serrano, Young-Han Kim, and Christopher Ellison for pointing out typographical errors.

Robert M. Gray
Rockport, Massachusetts
November 2010

Contents

Preface vii

Introduction xvii

1 Information Sources 1
1.1 Probability Spaces and Random Variables 1
1.2 Random Processes and Dynamical Systems 5
1.3 Distributions 7
1.4 Standard Alphabets 12
1.5 Expectation 13
1.6 Asymptotic Mean Stationarity 16
1.7 Ergodic Properties 17

2 Pair Processes: Channels, Codes, and Couplings 21
2.1 Pair Processes 21
2.2 Channels 22
2.3 Stationarity Properties of Channels 25
2.4 Extremes: Noiseless and Completely Random Channels 29
2.5 Deterministic Channels and Sequence Coders 30
2.6 Stationary and Sliding-Block Codes 31
2.7 Block Codes 37
2.8 Random Punctuation Sequences 38
2.9 Memoryless Channels 42
2.10 Finite-Memory Channels 42
2.11 Output Mixing Channels 43
2.12 Block Independent Channels 45
2.13 Conditionally Block Independent Channels 46
2.14 Stationarizing Block Independent Channels 46
2.15 Primitive Channels 48
2.16 Additive Noise Channels 49
2.17 Markov Channels 49

2.18 Finite-State Channels and Codes 50
2.19 Cascade Channels 51
2.20 Communication Systems 52
2.21 Couplings 52
2.22 Block to Sliding-Block: The Rohlin-Kakutani Theorem 53

3 Entropy 61
3.1 Entropy and Entropy Rate 61
3.2 Divergence Inequality and Relative Entropy 65
3.3 Basic Properties of Entropy 69
3.4 Entropy Rate 78
3.5 Relative Entropy Rate 81
3.6 Conditional Entropy and Mutual Information 82
3.7 Entropy Rate Revisited 90
3.8 Markov Approximations 91
3.9 Relative Entropy Densities 93

4 The Entropy Ergodic Theorem 97
4.1 History 97
4.2 Stationary Ergodic Sources 100
4.3 Stationary Nonergodic Sources 106
4.4 AMS Sources 110
4.5 The Asymptotic Equipartition Property 114

5 Distortion and Approximation 117
5.1 Distortion Measures 117
5.2 Fidelity Criteria 120
5.3 Average Limiting Distortion 121
5.4 Communications Systems Performance 123
5.5 Optimal Performance 124
5.6 Code Approximation 124
5.7 Approximating Random Vectors and Processes 129
5.8 The Monge/Kantorovich/Vasershtein Distance 132
5.9 Variation and Distribution Distance 132
5.10 Coupling Discrete Spaces with the Hamming Distance 134
5.11 Process Distance and Approximation 135
5.12 Source Approximation and Codes 141
5.13 d-bar Continuous Channels 142

6 Distortion and Entropy 147
6.1 The Fano Inequality 147
6.2 Code Approximation and Entropy Rate 150
6.3 Pinsker's and Marton's Inequalities 152
6.4 Entropy and Isomorphism 156
6.5 Almost Lossless Source Coding 160
6.6 Asymptotically Optimal Almost Lossless Codes 168

6.7 Modeling and Simulation 169

7 Relative Entropy 173
7.1 Divergence 173
7.2 Conditional Relative Entropy 189
7.3 Limiting Entropy Densities 202
7.4 Information for General Alphabets 204
7.5 Convergence Results 216

8 Information Rates 219
8.1 Information Rates for Finite Alphabets 219
8.2 Information Rates for General Alphabets 221
8.3 A Mean Ergodic Theorem for Densities 225
8.4 Information Rates of Stationary Processes 227
8.5 The Data Processing Theorem 234
8.6 Memoryless Channels and Sources 235

9 Distortion and Information 237
9.1 The Shannon Distortion-Rate Function 237
9.2 Basic Properties 239
9.3 Process Definitions of the Distortion-Rate Function 242
9.4 The Distortion-Rate Function as a Lower Bound 250
9.5 Evaluating the Rate-Distortion Function 252

10 Relative Entropy Rates 265
10.1 Relative Entropy Densities and Rates 265
10.2 Markov Dominating Measures 268
10.3 Stationary Processes 272
10.4 Mean Ergodic Theorems 275

11 Ergodic Theorems for Densities 281
11.1 Stationary Ergodic Sources 281
11.2 Stationary Nonergodic Sources 286
11.3 AMS Sources 290
11.4 Ergodic Theorems for Information Densities 293

12 Source Coding Theorems 295
12.1 Source Coding and Channel Coding 295
12.2 Block Source Codes for AMS Sources 296
12.3 Block Source Code Mismatch 307
12.4 Block Coding Stationary Sources 310
12.5 Block Coding AMS Ergodic Sources 312
12.6 Subadditive Fidelity Criteria 319
12.7 Asynchronous Block Codes 321
12.8 Sliding-Block Source Codes 323
12.9 A Geometric Interpretation 333

13 Properties of Good Source Codes 335
13.1 Optimal and Asymptotically Optimal Codes 335
13.2 Block Codes 337
13.3 Sliding-Block Codes 343

14 Coding for Noisy Channels 359
14.1 Noisy Channels 359
14.2 Feinstein's Lemma 361
14.3 Feinstein's Theorem 364
14.4 Channel Capacity 367
14.5 Robust Block Codes 372
14.6 Block Coding Theorems for Noisy Channels 375
14.7 Joint Source and Channel Block Codes 377
14.8 Synchronizing Block Channel Codes 380
14.9 Sliding-block Source and Channel Coding 384

References 395

Index 405

Introduction

Abstract A brief history of the development of Shannon information theory is presented with an emphasis on its interactions with ergodic theory. The origins and goals of this book are sketched.

Information theory, the mathematical theory of communication, has two primary goals: The first is the development of the fundamental theoretical limits on the achievable performance when communicating a given information source over a given communications channel using coding schemes from within a prescribed class. The second goal is the development of coding schemes that provide performance that is reasonably good in comparison with the optimal performance given by the theory. Information theory was born in a remarkably rich state in the classic papers of Claude E. Shannon [162, 163] which contained the basic results for simple memoryless sources and channels and introduced more general communication systems models, including finite-state sources and channels. The key tools used to prove the original results and many of those that followed were special cases of the ergodic theorem and a new variation of the ergodic theorem which considered sample averages of a measure of the entropy or self information in a process.

Information theory can be viewed as simply a branch of applied probability theory. Because of its dependence on ergodic theorems, however, it can also be viewed as a branch of ergodic theory, the theory of invariant transformations and transformations related to invariant transformations. In order to develop the ergodic theory example of principal interest to information theory, suppose that one has a random process, which for the moment we consider as a sample space or ensemble of possible output sequences together with a probability measure on events

composed of collections of such sequences. The shift is the transformation on this space of sequences that takes a sequence and produces a new sequence by shifting the first sequence a single time unit to the left. In other words, the shift transformation is a mathematical model for the effect of time on a data sequence. If the probability of any sequence event is unchanged by shifting the event, that is, by shifting all of the sequences in the event, then the shift transformation is said to be *invariant* and the random process is said to be *stationary*. Thus the theory of stationary random processes can be considered as a subset of ergodic theory. Transformations that are not actually invariant (random processes which are not actually stationary) can be considered using similar techniques by studying transformations which are almost invariant, which are invariant in an asymptotic sense, or which are dominated or asymptotically dominated in some sense by an invariant transformation. This generality can be important as many real processes are not well modeled as being stationary. Examples are processes with transients, processes that have been parsed into blocks and coded, processes that have been encoded using variable-length codes or finite-state codes, and channels with arbitrary starting states.

Ergodic theory was originally developed for the study of statistical mechanics as a means of quantifying the trajectories of physical or dynamical systems. Hence, in the language of random processes, the early focus was on ergodic theorems: theorems relating the time or sample average behavior of a random process to its ensemble or expected behavior. The work of Hoph [77], von Neumann [190] and others culminated in the pointwise or almost everywhere ergodic theorem of Birkhoff [17].

In the 1940's and 1950's Shannon made use of the ergodic theorem in the simple special case of memoryless processes to characterize the optimal performance possible when communicating an information source over a constrained random medium or *channel* using *codes*. The ergodic theorem was applied in a direct fashion to study the asymptotic behavior of error frequency and time average distortion in a communication system, but a new variation was introduced by defining a mathematical measure of the entropy or information in a random process and characterizing its asymptotic behavior. The results characterizing the optimal performance achievable using codes became known as *coding theorems*. Results describing performance that is actually achievable, at least in the limit of unbounded complexity and time, are known as *positive coding theorems*. Results providing unbeatable bounds on performance are known as *converse coding theorems* or *negative coding theorems.* When the same quantity is given by both positive and negative coding theorems, one has exactly the optimal performance achievable in theory using codes from a given class to communicate through the given communication systems model.

While mathematical notions of information had existed before, it was Shannon who coupled the notion with the ergodic theorem and an ingenious idea known as "random coding" in order to develop the coding theorems and to thereby give operational significance to such information measures. The name "random coding" is a bit misleading since it refers to the random selection of a deterministic code and not a coding system that operates in a random or stochastic manner. The basic approach to proving positive coding theorems was to analyze the average performance over a random selection of codes. If the average is good, then there must be at least one code in the ensemble of codes with performance as good as the average. The ergodic theorem is crucial to this argument for determining such average behavior. Unfortunately, such proofs promise the existence of good codes but give little insight into their construction.

Shannon's original work focused on memoryless sources whose probability distribution did not change with time and whose outputs were drawn from a finite alphabet or the real line. In this simple case the well-known ergodic theorem immediately provided the required result concerning the asymptotic behavior of information. He observed that the basic ideas extended in a relatively straightforward manner to more complicated Markov sources. Even this generalization, however, was a far cry from the general stationary sources considered in the ergodic theorem.

To continue the story requires a few additional words about measures of information. Shannon really made use of two different but related measures. The first was entropy, an idea inherited from thermodynamics and previously proposed as a measure of the information in a random signal by Hartley [75]. Shannon defined the entropy of a discrete time discrete alphabet random process $\{X_n\}$, which we denote by $H(X)$ while deferring its definition, and made rigorous the idea that the the entropy of a process is the amount of information in the process. He did this by proving a coding theorem showing that if one wishes to code the given process into a sequence of binary symbols so that a receiver viewing the binary sequence can reconstruct the original process perfectly (or nearly so), then one needs at least $H(X)$ binary symbols or bits (converse theorem) and one can accomplish the task with very close to $H(X)$ bits (positive theorem). This coding theorem is known as the *noiseless source coding theorem*.

The second notion of information used by Shannon was mutual information. Entropy is really a notion of self information — the information provided by a random process about itself. Mutual information is a measure of the information contained in one process about another process. While entropy is sufficient to study the reproduction of a single process through a noiseless environment, more often one has two or more distinct random processes, e.g., one random process representing an infor-

mation source and another representing the output of a communication medium wherein the coded source has been corrupted by another random process called noise. In such cases observations are made on one process in order to make decisions on another. Suppose that $\{X_n, Y_n\}$ is a random process with a discrete alphabet, that is, taking on values in a discrete set. The coordinate random processes $\{X_n\}$ and $\{Y_n\}$ might correspond, for example, to the input and output of a communication system. Shannon introduced the notion of the average mutual information between the two processes:

$$I(X,Y) = H(X) + H(Y) - H(X,Y), \tag{1}$$

the sum of the two self entropies minus the entropy of the pair. This proved to be the relevant quantity in coding theorems involving more than one distinct random process: the channel coding theorem describing reliable communication through a noisy channel, and the general source coding theorem describing the coding of a source for a user subject to a fidelity criterion. The first theorem focuses on error detection and correction and the second on analog-to-digital conversion and data compression. Special cases of both of these coding theorems were given in Shannon's original work.

Average mutual information can also be defined in terms of *conditional entropy* $H(X|Y) = H(X,Y) - H(Y)$ and hence

$$I(X,Y) = H(X) - H(X|Y) = H(Y) - H(X|Y). \tag{2}$$

In this form the mutual information can be interpreted as the information contained in one process minus the information contained in the process when the other process is known. While elementary texts on information theory abound with such intuitive descriptions of information measures, we will minimize such discussion because of the potential pitfall of using the interpretations to apply such measures to problems where they are not appropriate. (See, e.g., P. Elias' "Information theory, photosynthesis, and religion" in his "Two famous papers" [37].) Information measures are important because coding theorems exist imbuing them with operational significance and not because of intuitively pleasing aspects of their definitions.

We focus on the definition (1) of mutual information since it does not require any explanation of what conditional entropy means and since it has a more symmetric form than the conditional definitions. It turns out that $H(X,X) = H(X)$ (the entropy of a random variable is not changed by repeating it) and hence from (1)

$$I(X,X) = H(X) \tag{3}$$

so that entropy can be considered as a special case of average mutual information.

To return to the story, Shannon's work spawned the new field of information theory and also had a profound effect on the older field of ergodic theory.

Information theorists, both mathematicians and engineers, extended Shannon's basic approach to ever more general models of information sources, coding structures, and performance measures. The fundamental ergodic theorem for entropy was extended to the same generality as the ordinary ergodic theorems by McMillan [123] and Breiman [20] and the result is now known as the Shannon-McMillan-Breiman theorem. Other names are the asymptotic equipartition theorem or AEP, the ergodic theorem of information theory, and the entropy theorem. A variety of detailed proofs of the basic coding theorems and stronger versions of the theorems for memoryless, Markov, and other special cases of random processes were developed, notable examples being the work of Feinstein [39] [40] and Wolfowitz (see, e.g., Wolfowitz [196].) The ideas of measures of information, channels, codes, and communications systems were rigorously extended to more general random processes with abstract alphabets and discrete and continuous time by Khinchine [87], [88] and by Kolmogorov and his colleagues, especially Gelfand, Yaglom, Dobrushin, and Pinsker [49], [104], [101], [32], [150]. (See, for example, "Kolmogorov's contributions to information theory and algorithmic complexity" [23].) In almost all of the early Soviet work, it was average mutual information that played the fundamental role. It was the more natural quantity when more than one process were being considered. In addition, the notion of entropy was not useful when dealing with processes with continuous alphabets since it is generally infinite in such cases. A generalization of the idea of entropy called *discrimination* was developed by Kullback (see, e.g., Kullback [106]) and was further studied by the Soviet school. This form of information measure is now more commonly referred to as relative entropy, cross entropy, or Kullback-Leibler number, or information divergence and it is better interpreted as a measure of similarity or dissimilarity between probability distributions than as a measure of information between random variables. Many results for mutual information and entropy can be viewed as special cases of results for relative entropy and the formula for relative entropy arises naturally in some proofs.

It is the mathematical aspects of information theory and hence the descendants of the above results that are the focus of this book, but the developments in the engineering community have had as significant an impact on the foundations of information theory as they have had on applications. Simpler proofs of the basic coding theorems were developed for special cases and, as a natural offshoot, the rate of convergence to the optimal performance bounds characterized in a variety of important

cases. See, e.g., the texts by Gallager [47], Berger [11], and Csiszàr and Körner [27]. Numerous practicable coding techniques were developed which provided performance reasonably close to the optimum in many cases: from the simple linear error correcting and detecting codes of Slepian [171] to the huge variety of algebraic codes that have been implemented (see, e.g., [12], [192],[109], [113], [19]), the various forms of convolutional, tree, and trellis codes for error correction and data compression (see, e.g., [189, 81]), and the recent codes approaching the Shannon limits based on iterative coding and message passage ideas [126, 156], codes which have their roots in Gallager's PhD thesis on low density parity check codes [48]. Codes for source coding and data compression include a variety of traditional and recent techniques for lossless coding of data and lossy coding of realtime signals such as voice, audio, still images, and video. Techniques range from simple quantization to predictive quantization, adaptive methods, vector quantizers based on linear transforms followed by quantization and lossless codes, subband coders, and model coders such as the linear preditive codes for voice which fit linear models to observed signals for local synthesis. A sampling of the fundamentals through the standards can be found in [50, 160, 144, 178].

The engineering side of information theory through the middle 1970's has been well chronicled by two IEEE collections: *Key Papers in the Development of Information Theory,* edited by D. Slepian [172], and *Key Papers in the Development of Coding Theory,* edited by E. Berlekamp [13] and many papers describing the first fifty years of the field were collected into *Information Theory: 50 Years of Discovery* in 2000 [184]. In addition there have been several survey papers describing the history of information theory during each decade of its existence published in the *IEEE Transactions on Information Theory.*

The influence on ergodic theory of Shannon's work was equally great but in a different direction. After the development of quite general ergodic theorems, one of the principal issues of ergodic theory was the isomorphism problem, the characterization of conditions under which two dynamical systems are really the same in the sense that each could be obtained from the other in an invertible way by coding. Here, however, the coding was not of the variety considered by Shannon — Shannon considered block codes, codes that parsed the data into nonoverlapping blocks or windows of finite length and separately mapped each input block into an output block. The more natural construct in ergodic theory can be called a sliding-block code or stationary code — here the encoder views a block of possibly infinite length and produces a single symbol of the output sequence using some mapping (or code or filter). The input sequence is then shifted one time unit to the left, and the same mapping applied to produce the next output symbol, and so on. This is a smoother operation than the block coding structure since the outputs

are produced based on overlapping windows of data instead of on a completely different set of data each time. Unlike the Shannon codes, these codes will produce stationary output processes if given stationary input processes. It should be mentioned that examples of such sliding-block codes often occurred in the information theory literature: time-invariant convolutional codes or, simply, time-invariant linear filters are sliding-block codes. It is perhaps odd that virtually all of the theory for such codes in the information theory literature was developed by effectively considering the sliding-block codes as very long block codes. Sliding-block codes have proved a useful structure for the design of noiseless codes for constrained alphabet channels such as magnetic recording devices, and techniques from symbolic dynamics have been applied to the design of such codes. See, for example [3, 118].

Shannon's noiseless source coding theorem suggested a solution to the isomorphism problem: If we assume for the moment that one of the two processes is binary, then perfect coding of a process into a binary process and back into the original process requires that the original process and the binary process have the same entropy. Thus a natural conjecture is that two processes are isomorphic if and only if they have the same entropy. A major difficulty was the fact that two different kinds of coding were being considered: stationary sliding-block codes with zero error by the ergodic theorists and either fixed length block codes with small error or variable length (and hence nonstationary) block codes with zero error by the Shannon theorists. While it was plausible that the former codes might be developed as some sort of limit of the latter, this proved to be an extremely difficult problem. It was Kolmogorov [102], [103] who first reasoned along these lines and proved that in fact equal entropy (appropriately defined) was a necessary condition for isomorphism.

Kolmogorov's seminal work initiated a new branch of ergodic theory devoted to the study of entropy of dynamical systems and its application to the isomorphism problem. Most of the original work was done by Soviet mathematicians; notable papers are those by Sinai [168] [169] (in ergodic theory entropy is also known as the Kolmogorov-Sinai invariant), Pinsker [150], and Rohlin and Sinai [157]. An actual construction of a perfectly noiseless sliding-block code for a special case was provided by Meshalkin [124]. While much insight was gained into the behavior of entropy and progress was made on several simplified versions of the isomorphism problem, it was several years before Ornstein [138] proved a result that has since come to be known as the Ornstein isomorphism theorem or the Kolmogorov-Ornstein or Kolmogorov-Sinai-Ornstein isomorphism theorem.

Ornstein showed that if one focused on a class of random processes which we shall call B-processes, then two processes are indeed isomorphic if and only if they have the same entropy. *B*-process are also called

Bernoulli processes in the ergodic theory literature, but this is potentially confusing because of the usage of "Bernoulli process" as a synonym of an independent identically distributed (IID) process in information theory and random process theory. B-processes have several equivalent definitions, perhaps the simplest is that they are processes which can be obtained by encoding a memoryless process using a sliding-block code. This class remains the most general class known for which the isomorphism conjecture holds. In the course of his proof, Ornstein developed intricate connections between block coding and sliding-block coding. He used Shannon-like techniques on the block codes, then imbedded the block codes into sliding-block codes, and then used the stationary structure of the sliding-block codes to advantage in limiting arguments to obtain the required zero error codes. Several other useful techniques and results were introduced in the proof: notions of the distance between processes and relations between the goodness of approximation and the difference of entropy. Ornstein expanded these results into a book [140] and gave a tutorial discussion in the premier issue of the *Annals of Probability* [139]. Several correspondence items by other ergodic theorists discussing the paper accompanied the article.

The origins of this book lie in the tools developed by Ornstein for the proof of the isomorphism theorem rather than with the result itself. During the early 1970's I first become interested in ergodic theory because of joint work with Lee D. Davisson on source coding theorems for stationary nonergodic processes. The ergodic decomposition theorem discussed in Ornstein [139] provided a needed missing link and led to an intense campaign on my part to learn the fundamentals of ergodic theory and perhaps find other useful tools. This effort was greatly eased by Paul Shields' book *The Theory of Bernoulli Shifts* [164] and by discussions with Paul on topics in both ergodic theory and information theory. This in turn led to a variety of other applications of ergodic theoretic techniques and results to information theory, mostly in the area of source coding theory: proving source coding theorems for sliding-block codes and using process distance measures to prove universal source coding theorems and to provide new characterizations of Shannon distortion-rate functions. The work was done with Dave Neuhoff, like me then an apprentice ergodic theorist, and Paul Shields.

With the departure of Dave and Paul from Stanford, my increasing interest led me to discussions with Don Ornstein on possible applications of his techniques to channel coding problems. The interchange often consisted of my describing a problem, his generation of possible avenues of solution, and then my going off to work for a few weeks to understand his suggestions and work them through.

One problem resisted our best efforts–how to synchronize block codes over channels with memory, a prerequisite for constructing sliding-block codes for such channels. In 1975 I had the good fortune to meet and talk

with Roland Dobrushin at the 1975 IEEE/USSR Workshop on Information Theory in Moscow. He observed that some of his techniques for handling synchronization in memoryless channels should immediately generalize to our case and therefore should provide the missing link. The key elements were all there, but it took seven years for the paper by Ornstein, Dobrushin and me to evolve and appear [68].

Early in the course of the channel coding paper, I decided that having the solution to the sliding-block channel coding result in sight was sufficient excuse to write a book on the overlap of ergodic theory and information theory. The intent was to develop the tools of ergodic theory of potential use to information theory and to demonstrate their use by proving Shannon coding theorems for the most general known information sources, channels, and code structures. Progress on the book was disappointingly slow, however, for a number of reasons. As delays mounted, I saw many of the general coding theorems extended and improved by others (often by J. C. Kieffer) and new applications of ergodic theory to information theory developed, such as the channel modeling work of Neuhoff and Shields [133], [136], [135], [134] and design methods for sliding-block codes for input restricted noiseless channels by Adler, Coppersmith, and Hasner [3] and Marcus [118]. Although I continued to work in some aspects of the area, especially with nonstationary and nonergodic processes and processes with standard alphabets, the area remained for me a relatively minor one and I had little time to write. Work and writing came in bursts during sabbaticals and occasional advanced topic seminars. I abandoned the idea of providing the most general possible coding theorems and decided instead to settle for coding theorems that were sufficiently general to cover most applications and which possessed proofs I liked and could understand.

Only one third of this book is actually devoted to Shannon source and channel coding theorems; the remainder can be viewed as a monograph on sources, channels, and codes and on information and distortion measures and their properties, especially their ergodic properties. The sources or random processes considered include asymptotically mean stationary processes with standard alphabets, a subject developed in detail in my earlier book *Probability, Random Processes, and Ergodic Properties*, which was published by Springer-Verlag in 1988 [55] with a second edition published by Springer in 2009. That books treats advanced probability and random processes with an emphasis on processes with standard alphabets, on nonergodic and nonstationary processes, and on necessary and sufficient conditions for the convergence of long term sample averages. Asymptotically mean stationary sources and the ergodic decomposition are there treated in depth and recent simplified proofs of the ergodic theorem due to Ornstein and Weiss [141] and others are incorporated. The next chapter of this book reviews some of the basic notation of the first one in information theoretic terms, but results are

often simply quoted as needed from the first book without any attempt to derive them. The two books together are self-contained in that all supporting results from probability theory and ergodic theory needed here may be found in the first book. This book is self-contained so far as its information theory content, but it should be considered as an advanced text on the subject and not as an introductory treatise to the reader only wishing an intuitive overview. The border between the two books is the beginning of the treatment of entropy.

Here the Shannon-McMillan-Breiman theorem is proved using the coding approach of Ornstein and Weiss [141] (see also Shield's tutorial paper [165]) and hence the treatments of ordinary ergodic theorems in the first book and the ergodic theorems for information measures in this book are consistent. The extension of the Shannon-McMillan-Breiman theorem to densities is proved using the "sandwich" approach of Algoet and Cover [7], which depends strongly on the usual pointwise or Birkhoff ergodic theorem: sample entropy is asymptotically sandwiched between two functions whose limits can be determined from the ergodic theorem. These results are the most general yet published in book form and differ from traditional developments in that martingale theory is not required in the proofs.

A few words are in order regarding topics that are not contained in this book. I have not included the increasingly important and growing area of multiuser information theory because my experience in the area is slight and I believe this topic can be better handled by others.

Traditional noiseless coding theorems and actual codes such as the Huffman codes are not considered in depth because quite good treatments exist in the literature, e.g., [47], [1], [122]. The corresponding ergodic theory result — the Ornstein isomorphism theorem — is also not proved, because its proof is difficult and the result is not needed for the Shannon coding theorems. It is, however, described and many techniques used in its proof are used here for similar and other purposes.

The actual computation of channel capacity and distortion rate functions has not been included because existing treatments [47], [18], [11], [25] [57] are quite adequate. New to the second edition, however, is a partial development of Csiszár's [25] rigorous development of the information-theoretic optimization underlying the evaluation of the rate-distortion function.

This book does not treat code design techniques in any depth, but in this second edition properties of optimal and asymptotically optimal source codes are developed and these properties provide insight into the structure of good codes and can be used to guide code design. The traditional Lloyd optimality properties for vector quantizers are described along with recent results for sliding-block codes which resemble their block coding cousins.

J. C. Kieffer developed a powerful new ergodic theorem that can be used to prove both traditional ergodic theorems and the extended Shannon-McMillan-Brieman theorem [96]. He has used this theorem to prove strong (almost everywhere) versions of the source coding theorem and its converse, that is, results showing that sample average distortion is with probability one no smaller than the distortion-rate function and that there exist codes with sample average distortion arbitrarily close to the distortion-rate function [99, 100].