RECURRENCE AND WAITING TIMES IN STATIONARY PROCESSES, AND THEIR APPLICATIONS IN DATA COMPRESSION

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF ELECTRICAL ENGINEERING
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

By
Ioannis Kontoyiannis
May 1998

I certify that I have read this dissertation and that in my opinion it is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Thomas M. Cover
(Principal Advisor)

I certify that I have read this dissertation and that in my opinion it is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Amir Dembo

I certify that I have read this dissertation and that in my opinion it is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

George C. Papanicolaou

Approved for the University Committee on Graduate Studies:

Abstract

Over the past 25 years, the practical requirement for efficient data compression algorithms has generated a large volume of research covering the whole spectrum from practically implementable algorithms to deep theoretical results. One prominent example is the Lempel-Ziv algorithm for lossless data compression: Not only is it implemented on most computers used today, but also, attempts to analyze its performance have provided new problems in probability, information theory and ergodic theory, whose solutions reveal a series of interesting results about the entropy and the recurrence structure of stationary processes.

The main problems considered in this thesis are those of determining the asymptotic behavior of waiting times and recurrence times in stationary processes. These questions are motivated primarily by their important applications in data compression and the analysis of string matching algorithms in DNA sequence analysis. In particular, solving the waiting times problem also allowed us to solve a long-standing open problem in data compression: That of finding a practical extension of the Lempel-Ziv coding algorithm for lossy compression.

This thesis is divided into three parts. In the first part we generalize one of the central theoretical results in source coding theory: We prove a natural generalization of the celebrated Shannon-McMillan-Breiman theorem (as well as its subsequent refinements by Ibragimov and by Philipp and Stout) for real-valued processes and for the case when distortion is allowed. These results are inspired by, and provide the key technical ingredient in, our asymptotic analysis of recurrence and waiting times, in the second part. The main probabilistic tools used in establishing them are uniform almost-sure approximation, powerful techniques from large deviations, and classical second-moment blocking arguments.

In the second part we consider the problem of waiting times between stationary

processes. We show that waiting times grow exponentially with probability one and, that their rate is given by the solution to an explicit variational problem in terms of the entropies of the underlying processes. Moreover, we show that, properly scaled, the deviations of the waiting times from their limiting exponent are asymptotically Gaussian (with a limiting variance explicitly identified), and we prove finer theorems (e.g., a law of the iterated logarithm and an almost sure invariance principle) that provide the exact rate of convergence in the above limit theorems. Corresponding results are proved for recurrence times, and dual results are stated and proved for certain longest-match lengths between stationary processes.

Finally, in the third part, we use the insight gained by the waiting times results to find a practical extension of the Lempel-Ziv scheme for the case of lossy data compression. We propose a new lossy version of the so-called Fixed-Database Lempel-Ziv coding algorithm, which is of complexity "comparable" to that of the corresponding lossless scheme, and we prove that its compression performance is (asymptotically) optimal.

Acknowledgments

First and foremost, I would like to acknowledge my debt to my advisor, Professor Tom Cover. Being part of his research group during my five years at Stanford has influenced me deeply. I am grateful to him for creating an environment of unusual intellectual breadth, constant curiosity and challenge.

During the past two years I was very fortunate also to work closely with Professor Amir Dembo. Our frequent interactions and our later collaboration were very valuable learning experiences, and I am grateful for them. I would also like to thank Professors Mike Harrison and George Papanicolaou for being on my Orals committee and for all the help they offered me along the way.

My first classes in probability at Stanford were taught by Professor Tze Lai. His enthusiasm for the subject and his later interest (and encouragement) for my work are greatly appreciated.

Even when it did not have a specific purpose, informal interaction with several members of the faculty was a major source of learning and new ideas. In particular, Professors George Papanicolaou and Sam Karlin were always very generous with their time and knowledge.

Very special thanks to Professor Yurii Suhov of Cambridge University. During my Master's year he introduced me to the *real* world of mathematics, and, since then, he has remained an important force of motivation and a deep source of mathematical insight. Thanks also to Stephen Souras for turning my career around by introducing me to Shannon's work; more importantly, thanks for being the great friend he is.

My office-mates and my fellow ISL graduate students played an essential role in my education. I learnt a great deal from Paul Algoet, Vittorio Castelli, Tom Chung, Elza Erkip, Paul Fahn, Garud Iyengar, Don Kimber, Amos Lapidoth, Costis Maglaras, Erik Ordentlich, Joshua Singer, Rick Wesel and Assaf Zeevi. Especially I

want to thank Jim Hwang for helping render my first few months at Stanford tolerable. And many thanks to Denise Cuevas, for single-handedly managing to keep the whole process running smoothly.

I was very lucky to make very good friends while at Stanford. None of this would have been worth it without them. Hilary Teplitz, Christos Tryfonas and Vassilis Vassalos have won their place in these acknowledgments by exhibiting (well, occasionally, at least!) super-human amounts of patience and tolerance for my moods.

I owe very much to my family and my parents. Their love, support and encouragement have been unconditional, constant and boundless. I cannot imagine a better gift than what they have given me.

Financial support for the work reported in this thesis was provided, in part, by the following grants: NSF #NCR-9628193, JSEP #DAAH04-94-G-0058 and ARPA #J-FBI-94-218-2.

This thesis is dedicated to the precious memory of my mother.
Λάμπει μέσα της κείνο που αγνοεί. Μα ωστόσο λάμπει.

Contents

List of Figures

Chapter 1

Introduction

The central problem considered in this thesis is, loosely speaking, that of understanding the behavior of long pattern occurrences in realizations of random processes in discrete time. A typical question we will be asking is the following: *Suppose we observe the outcome of a binary random process; how long does it take until a certain pattern of zeros and ones first appears?* Questions of this type arise naturally in several areas, sometimes because of their theoretical interest and sometimes in applications. Here are four representative examples.

i. *Poincaré recurrence.* Here one asks questions about the *reappearance* of an initial pattern generated by the process. Does it always reappear? When it does, how long does it take? This problem and its ramifications are important in the study of dynamical systems in ergodic theory. In Chapter 3 we will ask what happens when we look for longer and longer such initial patterns – how much longer do we have to wait each time?

ii. *String matching.* Given two finite strings that are generated independently by the same process, what is the length of their longest common (contiguous) substring? This question arises in DNA sequence matching and in string searching algorithms in computer science. As we will see in Chapters 3 and 4, there is a natural "duality" relationship between questions about longest-match lengths, and questions about the first occurrence of random patterns.

iii. *Typicality.* In a long realization of a stationary ergodic process there are "typical" patterns that tend to appear often and "atypical" ones that only appear

rarely. This observation was made by Shannon in his landmark 1948 paper [62]. What is the length and the relative frequency of typical patterns? In Chapter 2 we generalize Shannon's original answers for these questions to real-valued (or more general) processes, and also to the case when distortion is allowed in the patterns.

iv. *Data compression.* Shannon's observation of typical patterns provides a precise way to quantify how much structure there is in a "message" produced by a random "source." How can we take advantage of this structure to do "compression," i.e., to describe long messages efficiently? The celebrated Lempel-Ziv family of data compression algorithms is based on exploiting this structure. In Chapter 5 we extend this idea further to the case of lossy data compression.

This list is by no means exhaustive. Several related questions are mentioned in Section 1.3 below.

As we shall see later, there is a common theme at the heart of all these problems – a strong connection between the geometry along a single realization and the probabilistic structure of the underlying process that produced it, in particular, with the entropy of that process. We can interpret this connection in the "big picture" by saying that it provides yet another snapshot of the sample-path picture of stochastic processes, added to the many other such properties that have come to form a major part of the foundation of modern probability theory over the past 50 years.

1.1 The Question of Recurrence

In order to get a better idea of the flavor of our problems and the ideas involved in solving them, we present here a concrete example of a question that is tackled in detail in Chapter 3. We will try to illustrate three points: (1) the motivation for the problem and the intuition underlying the analysis; (2) the natural way in which the entropy enters when we calculate probabilities of patterns along a realization; (3) the connection between pattern matching and data compression.

1.1.1 Recurrence and Entropy

Suppose we observe a doubly-infinite realization $\boldsymbol{x} = (\ldots, x_{-1}, x_0, x_1, x_2, \ldots)$ produced by a stationary ergodic process $\boldsymbol{X} = \{X_n \,;\, n \in \mathbb{Z}\}$, which takes values in a finite alphabet A. Write x_i^j for the substring of $\boldsymbol{x}$ between positions i and j

$$x_i^j \stackrel{\triangle}{=} (x_i, x_{i+1}, \ldots, x_j), \quad -\infty \leq i \leq j \leq \infty,$$

and similarly X_i^j for the vector of random variables $(X_i, X_{i+1}, \ldots, X_j)$. For a fixed integer n we consider the pattern $x_1^n = (x_1, x_2, \ldots, x_n)$ formed by the first n symbols produced by $\boldsymbol{X}$, and we ask how far back into the past one has to look before seeing the same pattern appear again. More precisely, we define R_n, the *recurrence time* for x_1^n, as the first position $k \geq 1$ for which $x_{-k+1}^{-k+n} = x_1^n$:

$$R_n \;=\; \inf\{k \geq 1 \;:\; x_{-k+1}^{-k+n} = x_1^n\}.$$

If we increase the length of the pattern we are looking for, then, clearly, the time we have to wait will increase, which implies that for every fixed realization $\boldsymbol{x}$ the recurrence time R_n increases with n. Our main question here is: *How fast does R_n increase?*

To gain some intuition we first try to understand what happens in the simplest case. Suppose $\boldsymbol{X}$ is a sequence of independent and identically distributed (i.i.d.) binary random variables, with each $X_n = 1$ with probability p, or $X_n = 0$ with probability $(1-p)$. Below we show an example of a realization from $\boldsymbol{X}$, with two recurring strings x_1^4 and x_1^5 and corresponding recurrence times $R_4 = 14$ and $R_5 = 26$.

$$\cdots \underbrace{0\,0\,1\,0\,1}_{R_5=26} 1\,0\,1\,1\,1\,0\,1 \overbrace{0\,0\,1\,0}^{R_4=14} \,0\,0\,1\,1\,0\,1\,0\,1\,1\,0 \overbrace{0\,0\,1\,0\,1}^{x_1^4} \cdots$$
$$\underbrace{\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad}_{x_1^5}$$

Conditional on the value of x_1, say $x_1 = 1$, the distribution of the recurrence time R_1 is exponential, with mean $1/p$. Thus, R_1 is concentrated around the reciprocal of the probability of the recurring symbol and has exponential tails away from its mean.

What about R_n for general n? Although its distribution is more complicated in this case, it is not hard to show that conditional on the recurring pattern x_1^n, the

mean of R_n is still equal to the reciprocal of the probability of that pattern

$$E(R_n \,|\, X_1^n = x_1^n) = \frac{1}{P(x_1^n)}, \tag{1.1}$$

where P denotes the distribution of $\boldsymbol{X}$. Now what is this probability? If n is large, there will be roughly np ones and $n(1-p)$ zeros in x_1^n, so that $P(x_1^n) \approx p^{np}(1-p)^{n(1-p)}$. Since this decays exponentially with n it suggests that, at least on the average, R_n increases exponentially with n. Moreover, looking at the exponent of decay of $P(x_1^n)$, we see that

$$-\frac{1}{n}\log P(x_1^n) \approx -\frac{1}{n}\log\left(p^{np}(1-p)^{n(1-p)}\right) = H, \tag{1.2}$$

where[1] $H = -p\log p - (1-p)\log(1-p)$ is the *entropy rate* of the process $\boldsymbol{X}$. This, then, suggests that R_n increases exponentially with a rate in the exponent given by the entropy rate of $\boldsymbol{X}$ and, indeed, it is probably not very surprising that the above informal argument can easily be made rigorous to show that

$$\lim_{n\to\infty} \frac{1}{n}\log R_n \;=\; H \quad \text{a.s.} \tag{1.3}$$

What is somewhat remarkable, though, is that each one of the above steps is essentially valid in full generality – for every finite-valued stationary ergodic process: A theorem of Kac from 1947 [34] says that (1.1) remains *verbatim* true for every stationary ergodic $\boldsymbol{X}$. This can be used to conclude (not trivially – see Theorem 3.1 in Chapter 3) that the asymptotic behavior of R_n is the same as that of $1/P(X_1^n)$, in that

$$\lim_{n\to\infty}\left[\frac{1}{n}\log R_n - \frac{1}{n}\log\frac{1}{P(X_1^n)}\right] = \lim_{n\to\infty}\frac{1}{n}\log\left[R_n P(X_1^n)\right] \;= 0 \quad \text{a.s.}, \tag{1.4}$$

and the Shannon-McMillan-Breiman theorem [13] states that (1.2) also remains true in this case

$$\lim_{n\to\infty} -\frac{1}{n}\log P(X_1^n) = H \quad \text{a.s.} \tag{1.5}$$

[1] Here and throughout this thesis log denotes the logarithm taken to base 2, and $\log_e$ denotes the natural logarithm.

where the entropy rate H of $\boldsymbol{X}$ is now defined by $H \stackrel{\triangle}{=} \lim_n E[-\log P(X_1 \,|\, X_{-n}^0]$. Combining (1.4) and (1.5) we recover (1.3) in complete generality!

1.1.2 Second-Order Results

After seeing that the rate in the exponent of the recurrence times R_n converges, with probability one, to a constant (the entropy rate H), there is a natural sequence of further questions we would like to ask, including:

i. What is the *rate of convergence* to the H in (1.3)?

ii. What is the *asymptotic distribution* of the deviations away from H?

iii. What is the *variance* of these deviations?

The way we will answer these questions in Chapter 3 is by refining the steps we took in the strategy that gave us (1.3). The main intuition we gained there was that, in a strong asymptotic sense, R_n, the recurrence time for the pattern X_1^n is close to the reciprocal of the probability $P(X_1^n)$ of that pattern. First we will show that the formal connection between R_n and $1/P(X_1^n)$ given in (1.4) can be strengthened to

$$\lim_{n\to\infty} \frac{1}{\sqrt{n}} \log[R_n P(X_1^n)] \;=\; 0 \quad \text{a.s.} \tag{1.6}$$

Then, looking at $-\log P(X_1^n)$ a little more carefully and assuming for a moment that $\boldsymbol{X}$ is i.i.d., we see that $-\log P(X_1^n)$ can be rewritten as an ordinary random walk

$$-\log P(X_1^n) = \sum_{i=1}^{n} [-\log P(X_i)], \tag{1.7}$$

so that its asymptotic behavior can be described in detail by the classical limit theorems for partial sums of i.i.d. random variables. For example, combining equations (1.6) and (1.7) with the classical central limit theorem immediately yields

$$\frac{\log R_n - nH}{\sqrt{n}} \xrightarrow{\mathcal{D}} N(0, \sigma^2) \tag{1.8}$$

with $\sigma^2 = \mathrm{Var}(-\log P(X_1))$, answering our questions (ii) and (iii) above [" $\xrightarrow{\mathcal{D}}$ " denotes convergence in distribution]. This can be viewed as a central-limit-theorem-type refinement to the strong-law-of-large-numbers statement of (1.3). Similarly, a simple application of the law of the iterated logarithm gives

$$\limsup_{n\to\infty} \frac{\log R_n - nH}{\sqrt{2n \log_e \log_e n}} = \sigma \quad \text{a.s.}, \tag{1.9}$$

providing the pointwise rate of convergence in (1.3) and answering question (i).

In Chapters 2 and 3 we show that the independence assumption can be significantly relaxed, and the same strategy works for a large class of processes with memory.

1.1.3 Recurrence and Data Compression

How did the question of the asymptotic behavior of R_n first arise?

In 1989, in an attempt to understand the exact compression performance of some variants of the Lempel-Ziv data compression algorithm, Wyner and Ziv [69] discovered the connection between recurrence times and entropy described in (1.3). One of the central ideas in their paper was, instead of considering the actual algorithms directly, to introduce and analyze an idealized coding scenario, a simple version of which we describe below.

Suppose an encoder and a decoder, me and you, say, have been communicating for a long time so that presently we share a very long, in fact *infinitely* long, common database $X_{-\infty}^0 = (\ldots, X_{-1}, X_0)$ produced by some stationary ergodic "source" $\boldsymbol{X}$. My task as the encoder is to describe to you the "message" X_1^n consisting of the next n symbols produced by $\boldsymbol{X}$, and I want to find a way to utilize somehow the "common information" $X_{-\infty}^0$ we share in order to describe X_1^n more efficiently.

My idea is, rather than describing X_1^n to you directly, I will look in the database $X_{-\infty}^0$, find the first position R_n where a copy of the message X_1^n appears, and tell you that position. From this information you can easily recover X_1^n by looking in the database and reading off the string $(X_{-R_n+1}, X_{-R_n+2}, \ldots, X_{-R_n+n})$.

Is this a good idea? Since all I have to tell you is R_n, my description consists of approximately $\log R_n$ bits (in general it takes about $\log k$ bits to describe an integer k), and from this you can recover a message of length n symbols, giving a compression

ratio of approximately

$$\frac{\log R_n}{n} \quad \text{bits per symbol.}$$

As we saw in (1.3) this ratio converges to the entropy rate of $\boldsymbol{X}$, implying that the compression performance of this simple-minded scheme is asymptotically optimal!

Although of no practical use in itself, this result provides the main technical ingredient in proving the optimality of the so-called Sliding-Window Lempel-Ziv algorithm [84][71], probably the most popular compression algorithm in use today. Moreover, Wyner and Ziv's idea of reducing the study of a practical algorithm to that of an idealized coding scenario was a very significant contribution to our intuitive understanding of the workings of several Lempel-Ziv schemes. Since then, this reduction has been exploited by a number of authors and has ultimately lead not only to a better understanding of the existing methods, but also to several new, practical data compression algorithms.

In Section 1.2.2 below we will push this connection a little further; we will discuss extensions of the Lempel-Ziv idea to lossy data compression, and motivate our subsequent results in Chapter 5.

1.2 Three More Questions

Next we outline three more questions that are addressed later in this thesis, and we highlight some of our relevant results from Chapters 2–5.

1.2.1 Waiting Times

Consider the following variation of the recurrence times problem: Instead of asking how long it takes before the first *re*appearance of the initial pattern generated by some random process, we ask how long it takes before the first *approximate appearance* of a random pattern generated independently by a different process.

For the sake of simplicity, consider two i.i.d. binary processes $\boldsymbol{X} = \{X_n \,;\, n \in \mathbb{Z}\}$ and $\boldsymbol{Y} = \{Y_n \,;\, n \in \mathbb{Z}\}$, with distributions P and Q, respectively. We will measure the closeness between finite realizations from $\boldsymbol{X}$ and $\boldsymbol{Y}$ by the proportion of positions

where they agree, so we define the *Hamming distortion* between x_1^n and y_1^n by

$$\rho_n(x_1^n, y_1^n) \;=\; \frac{1}{n}\sum_{i=1}^{n} I\{x_i = y_i\}, \quad x_1^n,\; y_1^n \in \{0,1\}^n,\; n \geq 1, \tag{1.10}$$

where $I\{x_i = y_i\}$ is the indicator function of the event $\{x_i = y_i\}$. For any binary string x_1^n and any distortion level $D \in [0,1]$ we let $B(x_1^n, D)$ denote the distortion-ball of radius D around x_1^n:

$$B(x_1^n, D) = \{y_1^n \in \{0,1\}^n \,:\, \rho_n(x_1^n, y_1^n) \leq D\}.$$

Given two realizations of $\boldsymbol{X}$ and $\boldsymbol{Y}$ and a $D \in [0,1]$, our quantity of interest here is the *waiting time* $W_n(D)$ until a D-close version of x_1^n first appears somewhere in y_1^∞:

$$W_n(D) \;=\; \inf\{k \geq 1 \,:\, y_k^{k+n-1} \in B(x_1^n, D)\}.$$

Intuitively, it seems natural to expect that the asymptotic behavior of $W_n(D)$ as $n \to \infty$ would not be very different from that of R_n, so we ask: *To what extent does $W_n(D)$ behave like R_n?*

In Chapter 4 this question is addressed (and answered), and the analysis follows essentially the same strategy as the one employed to analyze the behavior of R_n:

i. First, we prove that the waiting time $W_n(D)$ until we find a D-close match for X_1^n can be approximated by the reciprocal of the probability $Q(B(X_1^n, D))$ of finding such a match (see Theorem 4.1, Chapter 4):

$$\log W_n(D) \approx -\log Q(B(X_1^n, D)).$$

ii. Then we show that, asymptotically, $-\log Q(B(X_1^n, D))$ behaves as a random walk (Theorems 2.4 and 2.5, Chapter 2), just like $-\log P(X_1^n)$ did in the case of R_n.

Although these two steps closely parallel the corresponding recurrence times results in (1.6) and (1.7), the techniques used to prove them had to be different in this case. One of the difficulties can be spotted easily from the fact that we cannot expand $-\log Q(B(X_1^n, D))$ as random walk like we did with $-\log P(X_1^n)$ in (1.7). In fact,

it is not even clear *a priori* that $-\log Q(B(X_1^n, D))$ will have the same asymptotic behavior as $-\log P(X_1^n)$.

Chapter 2 is devoted to showing that the same behavior does indeed persist, in that the probabilities $Q(B(X_1^n, D))$ decay exponentially and their deviations from the limiting exponent are asymptotically those of a random walk. These results provide natural generalizations of the Shannon-McMillan-Breiman theorem and its refinements to general processes (taking more than a discrete set of values) and to the case when distortion is allowed.

Combining, as before, (i) and (ii) with the classical limit theorems for partial sums of i.i.d. random variables we obtain analogs of (1.3), (1.8) and (1.9): From the strong law of large numbers it follows that the waiting times $W_n(D)$ increase exponentially with probability one,

$$\lim_{n\to\infty} \frac{1}{n} \log W_n(D) \;=\; R(P, Q, D) \quad \text{a.s.}, \tag{1.11}$$

where the rate in the exponent $R(P, Q, D)$ can be explicitly identified as the solution to a variational problem in terms of the entropies of $\boldsymbol{X}$ and $\boldsymbol{Y}$. Similarly, using the central limit theorem and the law of the iterated logarithm we get analogs for (1.8) and (1.9), respectively.

1.2.2 Lossy Data Compression

In many engineering applications where large amounts of data are to be stored or transmitted, compression is an important component. Often, in order to reduce the storage or transmission requirements, we are willing to tolerate a certain amount of error in the reconstructed data – for example, think of a large image database where each image is compressed by a factor of, say, 50:1, and can be recovered not perfectly, but with a small amount of visual distortion. The following question will be addressed in Chapter 5: *Is there an easy way to extend the Lempel-Ziv idea to the case when distortion is allowed, to obtain a practical lossy compression scheme based on pattern matching?*

The great success of the Lempel-Ziv family of algorithms has been mainly due to two reasons. First, they are low complexity algorithms that can be simply implemented (they are, for example, implemented on almost every personal computer in

use today). Since efficient string matching has been very well studied by computer scientists over the past several decades, there are, by now, a number of very efficient algorithms that can be readily used in the context of compression.

The second reason for their practical success is that Lempel-Ziv schemes are *universal* – they assume essentially zero prior knowledge about the distribution of the source to be compressed. The trick they employ to overcome this lack of knowledge comes down to the idea of using the message itself as a codebook. For example, in the idealized coding scenario described in relation to recurrence times (Section 1.1.3 above), we assumed that the encoder and decoder shared an infinitely long database that had the same distribution as the source, and that the next part of the message was described by a pointer into that database.

There is, therefore, an implicit assumption that plays a key role in the success of these compression algorithms, namely, that the optimal (lossless) description of some random message is in terms of a codebook with the same distribution as the message itself. Unfortunately, this assumption is *not true* in the lossy case, and one is forced to consider codebooks generated according to different distributions.

To understand the situation better we follow Wyner and Ziv's example [69] and turn to an idealized coding scenario: Consider an encoder and a decoder sharing a common infinite database $Y_1^\infty = (Y_1, Y_2, \ldots)$, generated by some i.i.d. binary process $\boldsymbol{Y}$ with distribution Q. Suppose that the encoder's task is to communicate a message X_1^n, generated by a different i.i.d. binary process $\boldsymbol{X}$ of distribution P, to the decoder, within some prescribed distortion D (with respect, say, to Hamming distortion $\{\rho_n\}$ as defined in (1.10)). The encoder's strategy is, as before, to look through the database until the first time when a D-close match of X_1^n is found, and then tell the decoder the position $W_n(D)$ of this first match. To describe $W_n(D)$ it takes roughly $\log W_n(D)$ bits, so the compression achieved by this simple code equals

$$\frac{\log W_n(D)}{n} \quad \text{bits per symbol.}$$

As we saw in (1.11), this converges to $R(P, Q, D)$, so different choices of the database distribution yield different limiting compression ratios. The bad news here is that, unlike in the case of lossless compression, $R(P, Q, D)$ is *not* in general minimized by choosing the database to be of the same distribution as the source, i.e., taking $Q = P$. On the other hand, the optimal compression ratio for $\boldsymbol{X}$ with respect to Hamming

distortion at level D (given by the *rate-distortion function* $R(D)$ of $\boldsymbol{X}$) satisfies

$$R(D) = \inf_Q R(P, Q, D)$$

so that the problem is that we do not know *a priori* how to choose the best database distribution in order to minimize $R(P, Q, D)$.

In Chapter 5 we describe a new lossy version of Lempel-Ziv coding that gets around this problem by maintaining not just one, but *multiple* databases at the encoder and the decoder, and chooses which one to use at each stage in a "greedy" manner. The new algorithm is demonstrated to have asymptotically optimal compression performance (Theorem 5.2), and we argue that its complexity and redundancy characteristics are comparable to those of its lossless counterpart.

1.2.3 Match Lengths and DNA Template Matching

In the analysis of DNA or protein sequences the following problem is of interest: Suppose we have a template $(X_1, X_2, \ldots)$ and a long but finite database sequence $Y_1^m = (Y_1, Y_2, \ldots, Y_m)$. *What is the length of the longest initial portion X_1^ℓ of the template that matches within distortion D somewhere in the database?* By a "match" here we mean that there exists a contiguous substring $Y_{j+1}^{j+\ell}$ of the database such that the distortion between X_1^ℓ and $Y_{j+1}^{j+\ell}$ is at most D, with respect to, say, Hamming distortion. Given two realizations of the processes $\boldsymbol{X}$ and $\boldsymbol{Y}$ producing the above template and database, respectively, we write $L_m(D)$ for this maximal match-length:

$$L_m(D) = \sup\{\ell \geq 1 \ : \ y_{j+1}^{j+\ell} \in B(x_1^\ell, D), \ \text{ for some } \ j = 0, 1, \ldots, m-1\}.$$

Intuitively it seems that there is some connection between the match lengths $L_m(D)$ and the waiting times $W_n(D)$. We would expect that the database length m is essentially the same as the waiting time for $(X_1, \ldots, X_{L_m(D)})$, that is, if $n = L_m(D)$ then $W_n(D)$ should be approximately equal to m, and vice versa. Taking this analogy a step further, we might be tempted to replace m by $W_n(D)$ and n by $L_m(D)$ in our asymptotic results about waiting times, and hope that they remain valid.

We will see in detail in Chapters 3 and 4, that this intuition is essentially correct but it is not trivial to justify. For example, replacing m by $W_n(D)$ and n by $L_m(D)$

in (1.11) we obtain (see Theorem 4.2 in Chapter 4)

$$\lim_{m\to\infty} \frac{\log m}{L_m(D)} \;=\; R(P,Q,D) \quad \text{a.s.} \tag{1.12}$$

Similarly, all second-order results about $W_n(D)$ give us corresponding results for $L_m(D)$, providing a complete picture of the asymptotic behavior of $L_m(D)$.

1.3 History

Some general remarks about the history of the results we have been discussing are in order here. More detailed references to specific or more recent results are given at appropriate points in the subsequent chapters.

In ergodic theory, the question of what we called Poincaré recurrence was first raised by Poincaré in 1899 [59]. A very nice exposition of the long history of the results that followed, and also of the connection with the infamous H-theorem of Boltzmann, are presented in Petersen's text [55]. Kac's theorem was proved in 1947 [34]; alternative proofs can be found in [55][69].

Within probability theory, recurrence properties have been very important since at least as far back as the late 1930's. Doeblin and Harris both identified recurrence as the key concept in analyzing the asymptotic behavior of Markov processes; see Meyn and Tweedie's book [48] for a modern exposition. In particular, the idea of approximating the waiting time for an event by the reciprocal of its probability appears already in Doeblin's work on continued fractions in 1940 [24], in Bellman and Harris' (1951) work on the Ehrenfest model [10], and also in Harris' (1952) paper [31] on recurrence in Markov chains. At the cost of more restrictive assumptions, these authors go a step further and essentially show that the distribution of the waiting time for a rare event A is approximately exponential, with mean equal to the probability of A. Recent work in this direction is reported by Galves and Schmitt [27] who also provide an extensive list of references.

Closer to our approach, the use of $-\log P(X_1^n)$ or a similar random walk as an approximating sequence was employed by Ibragimov [32] and by Philipp and Stout [57, Chapter 9] in proving refinements to the Shannon-McMillan-Breiman theorem; by Barron [7] in proving the Shannon source coding theorem in the almost sure sense; and

by Algoet and Cover [2] in an elementary proof of the Shannon-McMillan-Breiman theorem.

The notion of typicality was introduced by Shannon in his famous 1948 paper [62] that founded the field of information theory. Our calculation of the probability of a typical sequence that lead to equation (1.2) was taken, essentially *verbatim*, from the discussion preceding Theorem 3 in [62]. There, Shannon showed that for every stationary ergodic Markov chain $\boldsymbol{X}$ with a finite number of states,

$$-\frac{1}{n}\log P(X_1^n) \to H \quad \text{in probability.} \tag{1.13}$$

McMillan [47] showed that (1.13) holds for every stationary ergodic process, and Breiman [13] strengthened McMillan's result to the almost sure convergence result we saw in (1.5). Meanwhile, first Yushkevich [77] in 1953 and then Ibragimov [32] in his well-known 1962 paper proved a central limit theorem refinement of (1.13). More on the history of further work in this direction is given in Chapter 2.

Turning to applications, the first explicit connection between match lengths and entropy seems to have been made in 1985 by Pittel [58], whose results are phrased in terms of path lengths in random trees. Aldous and Shields [1] pointed out the relationship between randomly growing trees and data compression, and Szpankowski [66] made explicit the equivalence between match lengths along random sequences and feasible paths in random trees.

Recurrence times in relation to data compression first appeared in Willems' work [67] and also in Wyner and Ziv's 1989 paper [69], where they (implicitly) introduced the idealized coding scenario we saw in Section 1.1.3. Wyner and Ziv [69] discovered (1.3) and the corresponding result for waiting times (without distortion), and these were formally established by Ornstein and Weiss [53] and by Shields [63], respectively, using methods from ergodic theory. Extensive references to subsequent work of refining and generalizing these results are given in Chapters 3 and 4.

In connection with DNA sequence analysis, results about asymptotics of match lengths arising from string matching problems can found in the work of Karlin and Ost [35], Pevzner, Borodovsky and Mironov [56], Arratia and Waterman [5], and Dembo, Karlin and Zeitouni [20]. Some of these results can be viewed as natural generalizations of the classical Erdös-Rényi laws of large numbers, as discussed by Arratia, Gordon and Waterman in [4]. Finally we mention that related questions

about string searching algorithms in computer science have been studied by Guibas and Odlyzko [29] and Jacquet and Szpankowski [33], among many others.

1.4 About This Thesis

1.4.1 Theory and Applications

Our initial motivation for this work was to gain a better understanding of the workings of the Lempel-Ziv family of data compression algorithms. Our introduction to the problem was through Wyner and Ziv's 1989 paper [69]; there, they isolated two very interesting theoretical questions (the questions about the asymptotic behavior of recurrence and waiting times), and demonstrated that the performance of the practical algorithms can be determined from the answers to these questions. Subsequently, researchers in several communities outside information theory found these problems also to be of theoretical interest and expanded on Wyner and Ziv's work. In the process of generalizing the original results to the case when distortion is allowed, further theoretical questions arose which led to the generalizations of the Shannon-McMillan-Breiman theorem and its refinements that we present in Chapter 2. These results, in turn, provided the intuition that was missing in order to solve an important practical problem, that of finding a practical extension of the Lempel-Ziv idea to the case of lossy compression – see Chapter 5.

In summary, a real practical application gave rise to some interesting theoretical questions, whose solutions may have significant impact in practice.

1.4.2 Organization

The rest of the thesis is organized as follows.

In Chapter 2 we describe the Shannon-McMillan-Breiman theorem, its refinements (by Yushkevich [77], Ibragimov [32], and Philipp and Stout [57]), and their generalizations to the case when distortion is allowed (by Łuczak and Szpankowski [45], Yang and Kieffer [75], and Dembo and Kontoyiannis [21]).

In Chapter 3 we address the problem of recurrence times in stationary processes, and we show the asymptotic behavior of the recurrence times R_n can be deduced from that of the random walk $-\log P(X_1^n)$. This, combined with the results presented in

Chapter 2, gives us a complete asymptotic description of R_n. Corresponding results are proved for certain longest match-lengths M_m along a realization, by exploiting a nice duality relationship between R_n and M_m.

Chapter 4 contains analogous results about waiting times, both with and without distortion. We first show that the behavior of the waiting times $W_n(D)$ can be deduced from that of the Q-probabilities of distortion balls $B(X_1^n, D)$, and then we apply our results from Chapter 2 to read-off the asymptotics of $W_n(D)$. Again, corresponding results are proved for the match lengths $L_m(D)$ via duality.

In Chapter 5 we address the problem of finding an extension of the Lempel-Ziv data compression algorithm that has asymptotically optimal compression performance, and is also implementable in practice. We introduce a new lossy variant of Lempel-Ziv, we prove its asymptotic optimality, and we argue that its complexity and redundancy characteristics are comparable to those of its lossless counterpart.

The contributions of this thesis are briefly summarized in Chapter 6, where we also mention some promising future research directions.

Finally in Appendix A we give the proofs of some of the more technical results from Chapters 2–5.

1.5 Notation

Here we state some notation and definitions that will remain in effect throughout this thesis. Although most of these are repeated (at least once) somewhere else, we also collect them here for easy reference.

- $\boldsymbol{X} = \{X_n \ ; \ n \in \mathbb{Z}\}$ denotes a stationary process with values in some space $(A, \mathcal{A})$, and distribution determined by the measure P on the product space $(A^\infty, \mathcal{A}^\infty)$.

- Similarly, $\boldsymbol{Y} = \{Y_n \ ; \ n \in \mathbb{Z}\}$ denotes a stationary process with values in some space $(\hat{A}, \hat{\mathcal{A}})$, and distribution determined by the measure Q on $(\hat{A}^\infty, \hat{\mathcal{A}}^\infty)$.

- For integers $-\infty \le i \le j \le \infty$, we denote by X_i^j the vector of random variables $(X_i, X_{i+1}, \ldots, X_j)$. Similarly, for a sequence $(x_n)_{n\in\mathbb{Z}}$ of elements from a set A, x_i^j denotes the part of the sequence between positions i and j.

- $\boldsymbol{x}$ denotes an infinite realization $\boldsymbol{x} = x_{-\infty}^{\infty} \in A^{\infty}$ of the process $\boldsymbol{X}$; similarly, $\boldsymbol{y}$ denotes a realization $\boldsymbol{y} = y_{-\infty}^{\infty} \in \hat{A}^{\infty}$ of $\boldsymbol{Y}$.

- "log" denotes the logarithm taken to base 2, and "$\log_e$" denotes the natural logarithm.

- $H(X) \stackrel{\triangle}{=} -\sum_x P(x) \log P(x)$ denotes the entropy (in bits) of the discrete random variable X, distributed according to the probability mass function P.

- $H(P)$ denotes the entropy rate (in bits) of the process $\boldsymbol{X}$ with distribution P, and is defined by
$$H(P) \;=\; \lim_{n\to\infty} \frac{1}{n}\, H(X_1^n).$$
If $\boldsymbol{X}$ is stationary then, equivalently, $H(P) \;=\; \lim_n E[-\log P(X_0 \,|\, X_{-n}^{-1})]$.

- $H(P\|Q)$ denotes the relative entropy (in bits) between the two probability measures P and Q, and is defined by
$$H(P\|Q) = \begin{cases} \int dP \log \frac{dP}{dQ}, & \text{when } \frac{dP}{dQ} \text{ exists} \\ \infty, & \text{otherwise.} \end{cases}$$

- $I(X;Y) \stackrel{\triangle}{=} H(P_{(X,Y)} \| P_X \times P_Y)$ denotes the mutual information (in bits) between the random variables X and Y, where P_X and P_Y denote the marginals of X and Y, respectively, and $P_{(X,Y)}$ is their joint distribution.

- ρ is some fixed measurable function $\rho : A \times \hat{A} \to [0, \infty)$, and $\{\rho_n\}$ is a sequence of single-letter distortion measures $\rho_n : A^n \times \hat{A}^n \to [0, \infty)$ defined by
$$\rho_n(x_1^n, y_1^n) = \frac{1}{n} \sum_{i=1}^{n} \rho(x_i, y_i), \qquad x_1^n \in A^n,\; y_1^n \in \hat{A}^n,\; n \ge 1.$$

- $R(D)$ is the rate-distortion function (in bits) of the process $\boldsymbol{X}$, with respect to the sequence of distortion measures $\{\rho_n\}$ and at distortion level D; it is defined

by

$$R(D) = \lim_{n\to\infty} \frac{1}{n} \inf_{\pi_n \in \mathcal{Q}_n} I(X_1^n; Y_1^n)$$

where $\mathcal{Q}_n$ is the space of all joint distributions π_n for (X_1^n, Y_1^n), such that $\int \rho_n(x_1^n, y_1^n) d\pi_n(x_1^n, y_1^n) \leq D$ and the X_1^n-marginal of π_n is the same as the original distribution of X_1^n.

- $H_e(X)$, $H_e(P)$, $H_e(P\|Q)$, $I_e(X;Y)$ and $R_e(D)$ denote the entropy, entropy rate, relative entropy, mutual information and rate-distortion function in *nats* rather than in bits, i.e., they have the same definitions as the corresponding functionals without the subscript e, but with the logarithms to base 2 replaced with natural logarithms.

Chapter 2

The Shannon-McMillan-Breiman Theorem, Generalizations, and Refinements

In this chapter we collect several theoretical results that we will need in later parts of this thesis. In Section 2.1 we present the Shannon-McMillan-Breiman theorem and its refinements, and in Section 2.2 we give their generalizations to the case when distortion is allowed.

2.1 Known Results

Shannon's 1948 landmark paper [62] contains a remarkably deep observation about "typical" patterns in random processes. Speaking of realizations of long sequences $X_1^N = (X_1, X_2, \ldots, X_N)$ generated by a stationary, ergodic, finite-state Markov chain, Shannon writes, *"... it is possible for most purposes to treat the long sequences as though there were just 2^{HN} of them, each with a probability 2^{-HN}."* Mathematically, this fact is formalized by the statement

$$-\frac{1}{n}\log P(X_1^n) \to H \quad \text{in probability,} \tag{2.1}$$

where H is the entropy rate of the Markov chain $\{X_n\}$ (cf. [62, Theorem 3]).

In general, let $\boldsymbol{X} = \{X_n \,;\, n \in \mathbb{Z}\}$ be a stationary ergodic process, with values in

the finite set A called the *alphabet* of $\boldsymbol{X}$, and with distribution determined by some probability measure P on the space $(A^\infty, \mathcal{A}^\infty)$, where $\mathcal{A}^\infty$ is the σ-field generated by finite-dimensional cylinders. In its general form, (2.1) is known as the Shannon-McMillan-Breiman theorem.

Theorem 2.1 (Shannon-McMillan-Breiman Theorem [62][47][13])
For every finite-valued stationary ergodic process $\boldsymbol{X}$,

$$-\frac{1}{n}\log P(X_1^n) \to H \quad a.s. \tag{2.2}$$

where H *is the* entropy rate *of the process* $\boldsymbol{X}$, *defined by*

$$H \stackrel{\triangle}{=} H(P) \stackrel{\triangle}{=} \lim_{n\to\infty} E[-\log P(X_0 \,|\, X_{-n}^{-1})].$$

Using the same approach as Breiman [13], Chung [15] in 1961 generalized (2.2) to stationary ergodic processes $\boldsymbol{X}$ with countable alphabets, under the assumption that $H(X_1) < \infty$.

In case of Markov chains it is not hard to see why (2.2) is true. We can expand

$$\begin{aligned} -\frac{1}{n}\log P(X_1^n) &= \frac{1}{n}\sum_{i=1}^{n}[-\log P(X_i \,|\, X_{i-1})] + \frac{1}{n}\log\frac{P(X_1 \,|\, X_0)}{P(X_1)} \\ &= \frac{1}{n}\sum_{i=1}^{n} f(\tilde{X}_{i-1}) + \frac{1}{n}C(\tilde{X}_0), \end{aligned} \tag{2.3}$$

where $\tilde{X}_n = (X_n, X_{n+1})$ forms a new Markov chain $\tilde{\boldsymbol{X}} = \{\tilde{X}_n = (X_n, X_{n+1})\ ;\ n \in \mathbb{Z}\}$ with state-space $T = \{(a,b) \in A \times A\ :\ P(X_1 = b \,|\, X_0 = a) > 0\}$, the function $f : T \to \mathbb{R}$ is defined by $f(a,b) = -\log P(X_1 = b \,|\, X_0 = a)$, and $C(\cdot)$ is defined by $C(a,b) \stackrel{\triangle}{=} \log[P(X_1 = b \,|\, X_0 = a)/P(X_1 = b)]$, for $(a,b) \in T$. Therefore (2.3) says that $-\log P(X_1^n)$ behaves like the sequence of partial sums of a bounded function of a Markov chain, up to a bounded term. Since $\boldsymbol{X}$ is stationary and ergodic so is $\tilde{\boldsymbol{X}}$, and the ergodic theorem implies (2.2) upon observing that $Ef(\tilde{X}_i) = E[-\log P(X_1 \,|\, X_0)] = H$, the entropy rate of $\boldsymbol{X}$.

Shortly after Shannon's paper, Yushkevich in 1953 [77], prompted by a question raised by Kolmogorov, observed that normalizing $-\log P(X_1^n)$ by $\sqrt{n}$ instead of n

and applying the central limit theorem for Markov chains yields

$$\frac{-\log P(X_1^n) - nH}{\sqrt{n}} \xrightarrow{\mathcal{D}} N(0, \sigma^2), \tag{2.4}$$

where

$$\sigma^2 = \lim_{n\to\infty} \frac{1}{n}\mathrm{Var}(-\log P(X_1^n)). \tag{2.5}$$

[" $\xrightarrow{\mathcal{D}}$ " denotes convergence in distribution.] Yushkevich's central limit theorem (2.4) was later extended by Ibragimov [32] to more general stationary ergodic processes by noticing that when the "memory" of the process $\boldsymbol{X}$ decays fast enough, $-\log P(X_1^n)$ still behaves like the partial sum of a stationary process, i.e.,

$$-\frac{1}{n}\log P(X_1^n) \;=\; \frac{1}{n}\sum_{i=1}^{n}[-\log P(X_i \,|\, X_1^{i-1})] \;\approx\; \frac{1}{n}\sum_{i=1}^{n}[-\log P(X_i \,|\, X_{-\infty}^{i-1})].$$

Essentially the same idea was used by Philipp and Stout [57, Chapter 9] in proving an almost sure invariance principle for $-\log P(X_1^n)$: Define a continuous time process $\{p(t)\ ;\ t \geq 0\}$ by letting $p(t) = 0$ for $t \in [0,1)$ and $p(t) = [-\log P(X_1^{\lfloor t \rfloor}) - \lfloor t \rfloor H]$ for $t \geq 1$. To quantify the rate at which the memory of $\boldsymbol{X}$ decays we define several mixing coefficients

$$\gamma(n) \;=\; \max_{a\in A} E\left|\log P(X_0 = a \,|\, X_{-\infty}^{-1}) - \log P(X_0 = a \,|\, X_{-n}^{-1})\right| \tag{2.6}$$

$$\alpha(n) \;=\; \sup\left\{|P(C\cap B) - P(C)P(B)| \;:\; B \in \sigma(X_{-\infty}^0),\ C \in \sigma(X_n^\infty)\right\} \tag{2.7}$$

$$\phi(n) \;=\; \sup\left\{|P(C|B) - P(C)| \;:\; B \in \sigma(X_{-\infty}^0),\ C \in \sigma(X_n^\infty)\right\}. \tag{2.8}$$

If $\alpha(n) \to 0$ as $n \to \infty$ $\boldsymbol{X}$ is called *strongly mixing* or *α-mixing*; similarly, if $\phi(n) \to 0$ as $n \to \infty$ $\boldsymbol{X}$ is called *ϕ-mixing*; see [12] for an extensive discussion of the properties of various mixing conditions of this form. The coefficients $\gamma(n)$ were introduced by Ibragimov [32] and they measure how well $\boldsymbol{X}$ can be approximated by finite-order Markov chains.

Theorem 2.2 (Phillip and Stout [57])

For every finite-valued stationary ergodic Markov chain $\boldsymbol{X}$:

(i) The following series converges:

$$\begin{aligned} \sigma^2 &= E[-\log P(X_0 \,|\, X_{-1}) - H]^2 \\ &+ 2\sum_{k=1}^{\infty} E[(-\log P(X_0 \,|\, X_{-1}) - H)(-\log P(X_k \,|\, X_{k-1}) - H)]. \end{aligned} \tag{2.9}$$

(ii) If $\sigma^2 > 0$, then there exists a standard Brownian motion $\{B(t)\,;\, t \geq 0\}$ such that

$$p(t) - \sigma B(t) = o(\sqrt{t}) \quad a.s. \tag{2.10}$$

(iii) Moreover, (i) and (ii) remain true if the ergodic Markov chain assumption is replaced by the assumptions that $\alpha(n) = O(n^{-336})$ and $\gamma(n) = O(n^{-48})$, with σ^2 replaced by

$$\begin{aligned} \sigma^2 &= E[-\log P(X_0 \,|\, X_{-\infty}^{-1}) - H]^2 \\ &+ 2\sum_{k=1}^{\infty} E[(-\log P(X_0 \,|\, X_{-\infty}^{-1}) - H)(-\log P(X_k \,|\, X_{-\infty}^{k-1}) - H)]. \end{aligned} \tag{2.11}$$

As usual, we interpret *(ii)* as saying that, without changing its distribution, $p(t)$ can be redefined on a richer probability space that contains a Brownian motion $\{B(t)\}$ such that (2.10) holds.

The numerous corollaries that can be derived from almost sure invariance principles like the one in (2.10) are well-known and include the central limit theorem (CLT), the law of the iterated logarithm (LIL), as well as their infinite dimensional, functional counterparts (see, e.g., Strassen's original paper [65], or [57, Chapter 1]). Several of these corollaries will be explicitly stated in Chapters 3 and 4, when we actually use (2.10) to obtain corresponding results for waiting times and recurrence times.

In the case of Markov chains the expressions for σ^2 in (2.5), (2.9) and (2.11), of course, all coincide; Yushkevich gave the following characterization of the degenerate case $\sigma^2 = 0$. We supply a proof of a slightly stronger result in the Appendix, by generalizing a formula of Fréchet [26].

Theorem 2.3 (Yushkevich [77], Kontoyiannis [39])

Let $\boldsymbol{X}$ be a finite-valued stationary ergodic Markov chain, with entropy rate H.

The variance σ^2 defined by (2.5) is equal to zero if and only if every string x_1^{n+1} that starts and ends in some fixed state $a \in A$, has probability, conditional on $x_1 = a$, either zero or q^n, for some constant q depending on $\boldsymbol{X}$.

Finally, we will also need the following variation on Theorem 2.2. Let Q be a stationary Markov measure on the same space as P, and assume that for all n large enough, the finite-dimensional marginals P_n of P are dominated by the corresponding marginals Q_n of Q

$$P_n \ll Q_n \quad \text{eventually.} \tag{2.12}$$

The *relative entropy rate* between P and Q is given by

$$H(P\|Q) \stackrel{\triangle}{=} \lim_{n\to\infty} E_P\left[\log \frac{P(X_0 \mid X_{-n}^{-1})}{Q(X_0 \mid X_{-n}^{-1})}\right].$$

and we define a continuous time process $\{q(t)\ ;\ t \geq 0\}$ by letting $q(t) = 0$ for $t \in [0,1)$ and $q(t) = [-\log Q(X_1^{\lfloor t \rfloor}) - \lfloor t \rfloor (H(P) + H(P\|Q))]$ for $t \geq 1$.

Proposition 2.1

Let $\boldsymbol{X}$ be a finite-valued stationary ergodic process, and Q be a stationary Markov measure satisfying (2.12).

(i) We have

$$-\frac{1}{n}\log Q(X_1^n) \to H(P) + H(P\|Q) \quad P-a.s.$$

(ii) If $\boldsymbol{X}$ is also a Markov process then the following limit exists

$$\sigma^2 \quad = \quad = \lim_{n\to\infty} \mathrm{Var}_P(-\log Q(X_1^n)). \tag{2.13}$$

(iii) If, moreover, $\sigma^2 > 0$, there exists a standard Brownian motion $\{B(t)\ ;\ t \geq 0\}$ such that

$$q(t) - \sigma B(t) = o(\sqrt{t}) \quad a.s. \tag{2.14}$$

Proof of Proposition 2.1: Part (i) follows from Barron's generalized Shannon-McMillan-Breiman theorem [8]. Parts (ii) and (iii) follow by an application of Theorem 10.1 of Philipp and Stout [57] to the sequence

$$-\log Q(X_1^n) - n(H(P) + H(P\|Q)) \;=\; \sum_{i=1}^{n-1}[g(\tilde{X}_i) - Eg(\tilde{X}_i)] + [-\log Q(X_1) - (H(P) + H(P\|Q))] \qquad (2.15)$$

where $g : T \to \mathbb{R}$ is the function $g(a,b) = [-\log Q(X_{i+1} = b \,|\, X_i = a)]$. Since $Eg(\tilde{X}_i) = H(P) + H(P\|Q)$ and g is bounded, the right hand side of (2.15) is equal (up to a bounded term) to the partial sum of a zero-mean, bounded function of a Markov chain. □

2.2 Allowing Distortion

As we saw in the introduction, when we consider "approximate matches" or "matches with distortion" between patterns generated by different processes, the quantity that naturally replaces $-\log P(X_1^n)$ is $-\log Q(B(X_1^n, D))$. Here, we will see how the asymptotic results for $-\log P(X_1^n)$ presented in the previous section generalize to the case of $-\log Q(B(X_1^n, D))$.

Little has been done in this direction. Recently, Luczak and Szpankowski [45] showed that, when A and $\hat{A}$ are finite sets, $(1/n)\log Q(B(X_1^n, D))$ converges to some constant R with probability one, and Yang and Kieffer [75] identified R as the solution to a variational problem in terms of relative entropy (see Theorem 2.4 below). Neither of these papers considered the problem of determining the second-order asymptotic properties of $-\log Q(B(X_1^n, D))$, and they also left open the question of whether analogous results can be established for processes taking values in general spaces A, $\hat{A}$. Here, we address both of these issues. The novelty in our approach is the use of large deviations techniques to relate the Q-probability of the ball $B(X_1^n, D)$ around the random center X_1^n to an associated random walk induced by X_1^n.

The typical scenario we will encounter in Chapters 4 and 5 consists of two stationary ergodic processes $\boldsymbol{X} = \{X_n \,;\, n \in \mathbb{Z}\}$ and $\boldsymbol{Y} = \{Y_n \,;\, n \in \mathbb{Z}\}$ with possibly different alphabets: Suppose $\boldsymbol{X}$ and $\boldsymbol{Y}$ take values in the Polish (= complete, seperable, metric) spaces $(A^\infty, \mathcal{A}^\infty)$ and $(\hat{A}^\infty, \hat{\mathcal{A}}^\infty)$, and are distributed according

to the probability measures P and Q, respectively. Given a measurable function $\rho : A \times \hat{A} \to [0, \infty)$, the *distortion* between finite strings $x_1^n \in A^n$ and $y_1^n \in \hat{A}^n$ is measured by:

$$\rho(x_1^n, y_1^n) = \frac{1}{n} \sum_{i=1}^{n} \rho(x_i, y_i). \tag{2.16}$$

For $x_1^n \in A^n$ and $D > 0$ we write $B(x_1^n, D)$ for the ball of radius D around x_1^n, $B(x_1^n, D) = \{y_1^n \in \hat{A}^n : \rho(x_1^n, y_1^n) \leq D\}$.

Throughout this section we will assume, for simplicity, that Q is a product measure, and write Q_1 for its one-dimensional marginal. Let

$$\begin{aligned} D_{\min} &\stackrel{\triangle}{=} E_{P_1}[\operatorname*{ess\,inf}_{Y_1 \sim Q_1} \rho(X_1, Y_1)] \\ D_{\text{av}} &\stackrel{\triangle}{=} E\rho(X_1, Y_1) \end{aligned}$$

and assume that

$$D_{\max} \stackrel{\triangle}{=} \operatorname*{ess\,sup}_{(X_1, Y_1)} \rho(X_1, Y_1) \in (D_{\min}, \infty).$$

Since $\boldsymbol{X}$ is stationary and ergodic, if we take $D > D_{\text{av}}$ then by the ergodic theorem $Q(B(X_1^n, D)) \to 1$ with $P-$probability one, whereas $Q(B(X_1^n, D)) = 0$ eventually $P-$almost surely for any $D < D_{\min}$. Therefore, of interest is the range of distortions D between $D_{\min}$ and D_{av}, where $Q(B(X_1^n, D))$ decays exponentially in a nontrivial manner.

Although the structure of $-\log Q(B(X_1^n, D))$ is no longer that of a random walk, our next two results show that we can relate $-\log Q(B(X_1^n, D))$ to a different random walk on the same probability space, which arises from a functional of the empirical measure $\hat{P}_n = n^{-1} \sum_{i=1}^{n} \delta_{X_i}$ induced on A by X_1^n. Theorem 2.4 is proved in Section 2.3 and Theorem 2.5 is proved in Section 2.4.

Theorem 2.4

Let $\boldsymbol{X}$ be a stationary ergodic process, Q be a product measure, and assume $D \in (D_{\min}, D_{\text{av}})$. Then

$$-\log Q(B(X_1^n, D)) - nR(\hat{P}_n) = o(\sqrt{n}) \quad P - a.s. \tag{2.17}$$

where $R(\hat{P}_n) = R(\hat{P}_n, Q_1, D)$ *is defined by the following variational problem:*

$$R(\hat{P}_n, Q_1, D) \;=\; \inf \int H(\Theta(\cdot|x)\|Q_1(\cdot))d\hat{P}_n(x) \tag{2.18}$$

where the infimum is taken over all probability measures Θ *on* $A \times \hat{A}$ *such that the* A*-marginal of* Θ *is* $\hat{P}_n$ *and* $\int \rho(x,y)\, d\Theta(x,y) \le D$.

A slightly different way to write $R(\hat{P}_n, Q_1, D)$ that will be useful in Chapter 5 is

$$R(\hat{P}_n, Q_1, D) \;=\; \inf[\, I(X;Y) + H(Q_1'\|Q_1)\,], \tag{2.19}$$

where the infimum is over all random variables (X, Y) with values in $A \times \hat{A}$, such that $X \sim \hat{P}_n$, $E\rho(X,Y) \le D$, and Q_1' denotes the marginal of Y. Yet another characterization of $R(\hat{P}_n, Q_1, D)$ is given by Proposition 2.2 in the next section.

Our first use of Theorem 2.4 is to prove the following generalization of Theorem 2.1. Its proof is given in Section 2.3.

Corollary 2.1 (Shannon-McMillan-Breiman Theorem with distortion)
Let $\boldsymbol{X}$ *be a stationary ergodic process, let* Q *be a product measure, and assume* $D \in (D_{\min}, D_{\text{av}})$. *Then,* $R(\hat{P}_n) \to R(P_1)$ *almost surely, and hence*

$$-\frac{1}{n}\log Q(B(X_1^n, D)) \to R(P_1, Q_1, D) \quad a.s. \tag{2.20}$$

(As we already mentioned in the beginning of this section, in the finite-alphabet case (2.20) was proved in [45][75].) Next, we investigate the behavior of the deviations of $R(\hat{P}_n)$ about its asymptotic mean $R(P_1)$ of order $\sqrt{n}$. For any probability measure μ on A and any $\lambda \in \mathbb{R}$, let

$$\Lambda_\mu(\lambda) \;=\; \int \log_e \left\{ \int e^{\lambda\rho(x,y)} dQ_1(y) \right\} d\mu(x).$$

Write $\Lambda(\cdot) = \Lambda_{P_1}(\cdot)$ when $\mu = P_1$, $\Lambda_x(\cdot) = \Lambda_{\delta_x}(\cdot)$ for any $x \in A$, and define the function $h : \mathbb{R} \times A \to [0, \infty)$ by

$$h(\lambda; x) \stackrel{\triangle}{=} (\log e) \left[\Lambda_x(\lambda) - \int \Lambda_{x'}(\lambda) dP_1(x') \right]. \tag{2.21}$$

Theorem 2.5 provides an explicit approximation of $\sqrt{n}[R(\hat{P}_n) - R(P_1)]$ by a random walk induced by X_1^n.

Theorem 2.5

Let $\boldsymbol{X}$ be a stationary process with α-mixing coefficients satisfying $\sum \alpha(n) < \infty$, let Q be a product measure, and $D \in (D_{\min}, D_{\text{av}})$. Then for some $\lambda = \lambda(D) < 0$ such that $\Lambda'(\lambda) = D$,

$$n[R(\hat{P}_n) - R(P_1)] + \sum_{i=1}^{n} h(\lambda; X_i) \;=\; o(\sqrt{n}) \quad a.s. \tag{2.22}$$

where h is defined by (2.21).

We now easily see from Theorems 2.4 and 2.5 that the asymptotic behavior of $-\log Q(B(X_1^n, D))$ is exactly that of a random walk

$$[-\log Q(B(X_1^n, D)) - nR(P_1, Q_1, D)] \;+\; \sum_{i=1}^{n} h(\lambda; X_i) \;=\; o(\sqrt{n}) \quad \text{a.s.} \tag{2.23}$$

where h is a bounded and centered function of the X_i's. The following is an immediate consequence of combining (2.23) with well-known CLT results (see, for example, [54, Theorem 1.7]).

Corollary 2.2 (CLT)

Let $\boldsymbol{X}$ be a stationary process with α-mixing coefficients satisfying $\sum \alpha(n) < \infty$, let Q be a product measure, and $D \in (D_{\min}, D_{\text{av}})$. Then, for $\lambda = \lambda(D)$, the following series converges

$$\sigma^2 \;=\; E_P\left\{h(\lambda; X_1)^2\right\} + 2\sum_{k=2}^{\infty} E_P\left\{h(\lambda; X_1)h(\lambda; X_k)\right\}, \tag{2.24}$$

where h is defined by (2.21). Moreover, when $\sigma^2 > 0$,

$$\frac{-\log Q(B(X_1^n, D)) - nR(P_1)}{\sqrt{n}} \xrightarrow{\mathcal{D}} N(0, \sigma^2),$$

and the sequence of processes

$$\left\{ \frac{q(nt;D)}{\sigma\sqrt{n}} \; ; \; t \in [0,1] \right\}, \quad n \geq 1$$

converges in distribution to standard Brownian motion, where $q(t;D) = 0$ *for* $t \in [0,1)$ *and* $q(t;D) = [-\log Q(B(X_1^{\lfloor t \rfloor}, D)) - \lfloor t \rfloor R(P_1, Q_1, D)]$ *for* $t \geq 1$.

Similarly, Corollary 2.3 is a consequence of (2.23) combined with the LIL [61].

Corollary 2.3 (LIL)

Let $\boldsymbol{X}$ *be a stationary process with* α*-mixing coefficients satisfying* $\sum \alpha(n) < \infty$, *let* Q *be a product measure, and* $D \in (D_{\min}, D_{\text{av}})$. *Then, for* $\sigma^2 > 0$ *as in (2.24), with* $P-$*probability one, the set of limit points of the sequence*

$$\left\{ \frac{-\log Q(B(X_1^n, D)) - nR(P_1)}{\sqrt{2n \log_e \log_e n}} \right\}, \quad n \geq 3$$

coincides with the interval $[-\sigma, \sigma]$. *Moreover, with* $P-$*probability one, the sequence of sample paths*

$$\left\{ \frac{q(nt;D)}{\sqrt{2n \log_e \log_e n}} \; ; \; t \in [0,1] \right\}, \quad n \geq 3,$$

is relatively compact in the topology of uniform convergence, and the set of its limit points is the collection of all absolutely continuous functions $r : [0,1] \to \mathbb{R}$, *such that* $r(0) = 0$ *and* $\int_0^1 (dr/dt)^2 dt \leq \sigma^2$.

Finally, the next Corollary generalizes Theorem 2.2; it follows from (2.23) and an almost sure invariance principle proved by Philipp and Stout [57, Theorem 4.1].

Corollary 2.4 (Almost sure invariance principle)

Let $\boldsymbol{X}$ *be a stationary process with* ϕ*-mixing coefficients satisfying* $\sum \phi(n) < \infty$, *let* Q *be a product measure, and* $D \in (D_{\min}, D_{\text{av}})$. *Then, with* $\sigma^2 > 0$ *as in (2.24), there exists a Brownian motion* $\{B(t) \; ; \; t \geq 0\}$ *such that*

$$q(t;D) - \sigma B(t) \; = \; o(\sqrt{t}) \quad P - a.s. \tag{2.25}$$

2.3 Large Deviations

In this section we give the proofs of Theorem 2.4 and Corollary 2.1.

First we state in Lemma 2.1 some useful technical facts that will be needed in the later proofs, and then we state the alternative characterization of the rate function R in terms of relative entropy in Proposition 2.2. Their proofs are given in the Appendix.

Lemma 2.1

Let μ and ν be arbitrary probability measures on A and $\hat{A}$, respectively. Let

$$\begin{aligned} D_{\min}^{\mu,\nu} &= \int \operatorname*{ess\,inf}_{Y\sim\nu} \rho(x,Y)\, d\mu(x) \\ D_{\text{av}}^{\mu,\nu} &= \int \rho(x,y)\, d\mu(x) d\nu(y) \\ D_{\max}^{\mu,\nu} &= \operatorname*{ess\,sup}_{(X,Y)\sim\mu\times\nu} \rho(X,Y) \end{aligned}$$

and for $\lambda, x \in \mathbb{R}$ define

$$\Lambda_{\mu,\nu}(\lambda) = \int \log_e \left(\int e^{\lambda\rho(x,y)} d\nu(y) \right) d\mu(x)$$

and its Fenchel-Legendre transform

$$\Lambda_{\mu,\nu}^*(x) = \sup_{\lambda\in\mathbb{R}} [\lambda x - \Lambda_{\mu,\nu}(\lambda)].$$

Assume $0 \le D_{\min}^{\mu,\nu} < D_{\text{av}}^{\mu,\nu} \le D_{\max}^{\mu,\nu} < \infty$. Then

(*i*) $|\Lambda_{\mu,\nu}(\lambda)| \le |\lambda| D_{\max}^{\mu,\nu}$.

(*ii*) $\Lambda_{\mu,\nu} \in C^\infty$, $\Lambda'_{\mu,\nu}(0) = D_{\text{av}}^{\mu,\nu}$, $\Lambda''_{\mu,\nu}(\lambda) > 0$ *for all* $\lambda \in \mathbb{R}$, *and* $\Lambda'_{\mu,\nu}(\lambda) \downarrow D_{\min}^{\mu,\nu}$ *as* $\lambda \to -\infty$.

(*iii*) *For each* $D \in (D_{\min}^{\mu,\nu}, D_{\text{av}}^{\mu,\nu})$, *there exists a unique* $\lambda < 0$ *such that* $\Lambda'_{\mu,\nu}(\lambda) = D$ *and* $\Lambda_{\mu,\nu}^*(D) = \lambda D - \Lambda_{\mu,\nu}(\lambda)$. *Therefore,* $\Lambda_{\mu,\nu}^*(D)$ *is finite, continuous and decreasing for* $D \in (D_{\min}^{\mu,\nu}, D_{\text{av}}^{\mu,\nu})$.

(*iv*) *For μ-almost any $x \in A$, $\Lambda_{\delta_x,\nu} \in C^\infty$ and its derivatives are uniformly bounded over μ-almost all $x \in A$ and all λ in a compact subset of $\mathbb{R}$.*

Proposition 2.2

In the notation of Lemma 2.1 with μ and ν being arbitrary probability measures on A and $\hat{A}$, respectively, and $D \in (D_{\min}^{\mu,\nu}, D_{\text{av}}^{\mu,\nu})$, we have,

$$R_e(\mu,\nu,D) \;=\; (\log_e 2)R(\mu,\nu,D) \;=\; \Lambda^*_{\mu,\nu}(D)$$

where $R_e(\mu,\nu,D)$ is defined as in (2.18), but with relative entropy in nats instead of bits.

Proof of Theorem 2.4. In order to simplify the notation, we will prove the statement of the theorem in terms of natural logarithms rather than logarithms to base 2, i.e., we will show that

$$-\log_e Q(B(X_1^n, D)) - nR_e(\hat{P}_n) \;=\; o(\sqrt{n}) \quad P-\text{a.s.} \tag{2.26}$$

Let $D_{\text{av}}^{(n)} = \int \rho(x,y)\, d\hat{P}_n(x) dQ_1(y)$, so that, by the ergodic theorem

$$D_{\text{av}}^{(n)} \to D_{\text{av}} \quad P-\text{a.s.} \tag{2.27}$$

Similarly let $D_{\min}^{(n)} = E_{\hat{P}_n}[\text{ess}\inf_{Y_1}\ \rho(X_1, Y_1)]$, so that

$$D_{\min}^{(n)} \to D_{\min} \quad P-\text{a.s.} \tag{2.28}$$

Given a realization of the $\boldsymbol{X}$ process such that both (2.27) and (2.28) hold, for n large enough the given D will be strictly between $D_{\min}^{(n)}$ and $D_{\text{av}}^{(n)}$. Therefore, by Lemma 2.1 we can choose, for each n, a negative λ_n such that $\Lambda'_{\hat{P}_n}(\lambda_n) = D$, $\Lambda^*_{\hat{P}_n}(D) = \lambda_n D - \Lambda_{\hat{P}_n}(\lambda_n)$, and $\Lambda''_{\hat{P}_n}(\lambda_n) > 0$. We similarly choose $\lambda < 0$ such that $\Lambda'(\lambda) = D$, and we claim that

$$\lambda_n \to \lambda \quad P-\text{a.s.} \tag{2.29}$$

To see this suppose, for example, that $\limsup_{n\to\infty} \lambda_n \le \lambda - \epsilon$, for some $\epsilon > 0$, so that, eventually, $\lambda_n \le \lambda - \epsilon/2$. Then by the ergodic theorem and the strict monotonicity

of Λ' we get a contradiction:

$$\begin{aligned} D &= \limsup_{n\to\infty} \Lambda'_{\hat{P}_n}(\lambda_n) \le \limsup_{n\to\infty} \Lambda'_{\hat{P}_n}(\lambda - \epsilon/2) = \lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} \Lambda'_{X_i}(\lambda - \epsilon/2) \\ &= \Lambda'(\lambda - \epsilon/2) < \Lambda'(\lambda) = D. \end{aligned}$$

The case $\liminf_{n\to\infty} \lambda_n > \lambda$ is ruled out similarly.

Before moving to the main part of the proof, we also show that

$$\Lambda''_{\hat{P}_n}(\lambda_n) \to \Lambda''(\lambda) > 0 \quad P-\text{a.s.} \tag{2.30}$$

Since

$$|\Lambda''_{\hat{P}_n}(\lambda_n) - \Lambda''(\lambda)| \le \frac{1}{n}\sum_{i=1}^{n} |\Lambda''_{X_i}(\lambda_n) - \Lambda''_{X_i}(\lambda)| + \left| \frac{1}{n}\sum_{i=1}^{n} \Lambda''_{X_i}(\lambda) - \Lambda''(\lambda) \right|, \tag{2.31}$$

we can bound the first term above, for any $\epsilon > 0$ and n large enough, by

$$\operatorname*{ess\,sup}_{X_1} |\Lambda''_{X_1}(\lambda_n) - \Lambda''_{X_1}(\lambda)| \le |\lambda_n - \lambda| \operatorname*{ess\,sup}_{X_1} \sup_{\lambda-\epsilon\le\xi\le\lambda+\epsilon} |\Lambda'''_{X_1}(\xi)|$$

and this converges to zero, by (2.29) and Lemma 2.1. As for the second term of (2.31), by the ergodic theorem it converges to zero, $P-$almost surely.

Now we choose and fix a realization $\boldsymbol{x}$ of $\boldsymbol{X}$ such that the statements (2.27), (2.28), (2.29) and (2.30) all hold. Define $\zeta_i = \rho(x_i, Y_i)$, $T_n = \sum_{i=1}^{n} \zeta_i$, and $\hat{T}_n = T_n/n$, with μ_n denoting the law of ζ_1^n. With a slight abuse of notation we write $\hat{P}_n$ for the (non-random, since x_1^n is fixed) empirical measure induced by x_1^n on A. In this notation, $Q(B(x_1^n, D)) = \Pr(\hat{T}_n \le D)$, and, if we define

$$J_n = e^{n\Lambda^*_{\hat{P}_n}(D)} \Pr(\hat{T}_n \le D),$$

then in view of Proposition 2.2 and (2.26) the statement of the theorem can be rephrased as

$$\log_e J_n = o(\sqrt{n}) \quad P-\text{a.s.} \tag{2.32}$$

The upper-bound part of (2.32) follows from

$$\begin{aligned} J_n &= e^{n\Lambda^*_{\hat{P}_n}(D)} E\left\{1_{\{\hat{T}_n \le D\}}\right\} \le e^{n\Lambda^*_{\hat{P}_n}(D)} E\left\{e^{n\lambda_n(\hat{T}_n - D)}\right\} \\ &= e^{n[\Lambda^*_{\hat{P}_n}(D) - \lambda_n D]} E\left\{e^{\lambda_n T_n}\right\} = 1 \end{aligned}$$

(by the choice of λ_n and the definition of $\Lambda_{\hat{P}_n}$).

Turning to the proof of the lower bound, suppose n is large enough so that λ_n exists, and define a new probability measure ν_n by

$$\frac{d\nu_n}{d\mu_n}(z_1^n) = \exp\left\{\lambda_n \sum_{i=1}^n z_i - n\Lambda_{\hat{P}_n}(\lambda_n)\right\}.$$

Let

$$G_n = -\frac{\sum_{i=1}^n [\zeta_i - E_{\nu_n}\zeta_i]}{\sqrt{n\Lambda''_{\hat{P}_n}(\lambda_n)}}, \quad \text{when } \zeta_1^n \sim \nu_n.$$

It is easy to see that G_n is the partial sum of zero mean random variables, normalized so that $\mathrm{Var}(G_n) = 1$. Observe that when ζ_1^n is distributed according to ν_n,

$$\hat{T}_n \stackrel{\mathcal{D}}{=} D - \sqrt{\frac{\Lambda_{\hat{P}_n}{}''(\lambda_n)}{n}}\, G_n,$$

so that we can expand

$$\begin{aligned} J_n &= e^{n\Lambda^*_{\hat{P}_n}(D)} E_{\nu_n}\left\{1_{\{\hat{T}_n \le D\}} e^{-n\lambda_n \hat{T}_n + n\Lambda_{\hat{P}_n}(\lambda_n)}\right\} \\ &= E_{\nu_n}\left\{1_{\{G_n \ge 0\}} e^{\lambda_n \sqrt{n\Lambda''_{\hat{P}_n}(\lambda_n)}\, G_n}\right\} \\ &\ge E_{\nu_n}\left\{1_{\{0 < G_n < \delta\}} e^{-\beta_n \sqrt{n}\, G_n}\right\} \\ &\ge e^{-\beta_n \sqrt{n}\,\delta} \mathrm{Pr}_{\nu_n}(0 < G_n < \delta), \end{aligned} \tag{2.33}$$

for any $\delta > 0$, where $\beta_n = -\lambda_n \sqrt{\Lambda''_{\hat{P}_n}(\lambda_n)} > 0$ and $\beta_n = O(1)$, by (2.29) and (2.30).

Since the random variables ζ_i are uniformly bounded, and also $\Lambda''_{\hat{P}_n}(\lambda_n)$ is bounded away from zero by (2.30), it is easy to check that the Lindeberg condition for the CLT is satisfied by G_n, from which it follows that $\mathrm{Pr}_{\nu_n}(0 < G_n < \delta) \to \rho > 0$ as $n \to \infty$. Now choose $M > 0$ large enough so that $M - \beta_n$ is bounded away from zero, and get

from (2.33) that

$$\liminf_{n\to\infty} \log_e \left[e^{M\sqrt{n}\delta} J_n \right] \geq \log \rho > -\infty,$$

i.e.,

$$\liminf_{n\to\infty} \sqrt{n} \left[M\delta + \frac{1}{\sqrt{n}} \log_e J_n \right] > -\infty$$

from which we conclude that

$$\liminf_{n\to\infty} \frac{1}{\sqrt{n}} \log_e J_n \geq -M\delta.$$

Since $\delta > 0$ was arbitrary and $M > 0$ was chosen independent of δ, letting $\delta \downarrow 0$ completes the proof. □

Proof of Corollary 2.1. The result (2.20) immediately follows from Theorem 2.4, provided we show that $R(\hat{P}_n) \to R(P_1)$ almost surely, or, equivalently (by Proposition 2.2), that $\Lambda^*_{\hat{P}_n}(D) \to \Lambda^*(D)$ almost surely. Recall that for all n large enough $\Lambda^*_{\hat{P}_n}(D) = \lambda_n D - \Lambda_{\hat{P}_n}(\lambda_n)$ and $\Lambda^*(D) = \lambda D - \Lambda(\lambda)$, as in the proof of Theorem 2.4, where $\lambda_n \to \lambda$ almost surely by (2.29). So we only have to show that $\Lambda_{\hat{P}_n}(\lambda_n) \to \Lambda(\lambda)$, which comes from an obvious adaptation of the derivation of (2.30). □

2.4 Uniform Approximation

Proof of Theorem 2.5. Let λ and $\{\lambda_n\}$ be chosen as in the beginning of the proof of Theorem 2.4, so that, in particular, $\Lambda^*(\lambda) = \lambda D - \Lambda(\lambda)$ and $\Lambda''(\lambda) > 0$. By the continuity of Λ'' we can choose constants $\delta, \eta > 0$ such that $\Lambda''(\lambda + \theta) > \eta$ whenever $|\theta| < \delta$. Also, from (2.29), we can pick $N = N(X_1^\infty) < \infty$ $P-$almost surely, such that $|\lambda_n - \lambda| < \delta$ for all $n \geq N$.

In view of Proposition 2.2, it suffices to show that

$$\sqrt{n} \left\{ [\Lambda^*_{\hat{P}_n}(D) - \Lambda^*(D)] - [\Lambda(\lambda) - \Lambda_{\hat{P}_n}(\lambda)] \right\} \to 0. \tag{2.34}$$

From the definition of $\Lambda^*_{\hat{P}_n}$ and our choice of N, $\Lambda^*_{\hat{P}_n}(D)$ is given by the supremum of $[\theta D - \Lambda_{\hat{P}_n}(\theta)]$ over all $\theta \in (\lambda - \delta, \lambda + \delta)$, so (2.34) is the same as

$$\sqrt{n} \sup_{|\theta|<\delta} \left[\theta D - \Lambda_{\hat{P}_n}(\theta + \lambda) + \Lambda_{\hat{P}_n}(\lambda) \right] \to 0. \tag{2.35}$$

Since this supremum is always non-negative (take $\theta = 0$), (2.35) is equivalent to

$$\liminf_{n\to\infty} \sqrt{n} \inf_{|\theta|<\delta} \frac{1}{n} \sum_{i=1}^{n} [f(\theta, X_i) - f(0, X_i)] \geq 0 \ , \tag{2.36}$$

where $f(\theta, x) = \Lambda_x(\lambda+\theta)-(\lambda+\theta)D$. By Taylor's theorem we can expand the function $g(\theta) = n^{-1} \sum_{i=1}^{n} f(\theta, X_i)$ around $\theta = 0$, to obtain

$$\frac{1}{n} \sum_{i=1}^{n} [f(\theta, X_i) - f(0, X_i)] = \theta A_n + \frac{\theta^2}{2} B_n(\theta) \ , \tag{2.37}$$

where $A_n = n^{-1} \sum_{i=1}^{n} f'(0, X_i)$ and $B_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} f''(\xi_n, X_i)$ for some $\xi_n(\theta)$ such that $|\xi_n| < \delta$.

The family of functions $\{f''(\xi, \cdot) \ ; \ \xi \in (-\delta, \delta)\}$ is uniformly bounded and equicontinuous (by Lemma 2.1), so by the uniform ergodic theorem (see, for example, [60, Section 6]),

$$\sup_{|\xi|<\delta} \left| \frac{1}{n} \sum_{i=1}^{n} f''(\xi, X_i) - E_P f''(\xi, X_1) \right| \to 0 \quad P - \text{a.s.}$$

Therefore, $P-$almost surely, by the choice of δ,

$$\begin{aligned} &\liminf_{n\to\infty} \inf_{|\theta|<\delta} B_n(\theta) \\ &\qquad \geq \quad \liminf_{n\to\infty} \left\{ \inf_{|\xi|<\delta} E_P f''(\xi, X_1) - \sup_{|\xi|<\delta} \left| \frac{1}{n} \sum_{i=1}^{n} f''(\xi, X_i) - E_P f''(\xi, X_1) \right| \right\} \\ &\qquad \geq \quad \inf_{|\xi|<\delta} E_P f''(\xi, X_1) = \inf_{|\xi|<\delta} \Lambda''(\lambda + \xi) \geq \eta > 0 \ . \end{aligned} \tag{2.38}$$

By our choice of λ, we have $E_P f'(0, X_1) = \Lambda'(\lambda) - D = 0$, so A_n is the partial sum corresponding to the zero-mean stationary process $\{f'(0, X_n) \ ; \ n \geq 1\}$. Since $\sum \alpha(n) < \infty$ and the random variables $f'(0, X_i)$ are bounded, the LIL [61] implies that $\sqrt{n} A_n^2 \to 0$ $P-$almost surely. Since the infimum over $|\theta| < \delta$ of the right side of (2.37) is bounded below by $-A_n^2 / \inf_{|\theta|<\delta} B_n(\theta)$, combining this with (2.38) gives (2.36) and completes the proof. □

Chapter 3

Recurrence in Stationary Processes

3.1 Introduction and Main Results

As we have seen in the introduction, recurrence properties are important in the study of stationary processes in probability theory, and dynamical systems in ergodic theory. In this chapter we investigate the asymptotic behavior of recurrence times for finite-valued stationary processes, under various mixing conditions.

As before, let $\boldsymbol{X} = \{X_n \ ; \ n \in \mathbb{Z}\}$ be a stationary ergodic process with values in a finite alphabet A and distribution determined by the probability measure P on $(A^\infty, \mathcal{A}^\infty)$, where $\mathcal{A}^\infty$ is the σ-field generated by finite-dimensional cylinders. Given a realization $\boldsymbol{x}$ from $\boldsymbol{X}$, our main quantity of interest here is the *recurrence time* R_n defined as the first time until the opening string x_1^n recurs in the past of $\boldsymbol{x}$:

$$R_n \ = \ \inf\{k \geq 1 \ : \ x_{-k+1}^{-k+n} = x_1^n\}$$

There has been a lot of work on calculating the exact asymptotic behavior of R_n. Wyner and Ziv [69], motivated by coding problems in information theory, drew a deep connection between recurrence times and the entropy rate of the underlying process. They proved that R_n grows exponentially with n and that the limiting rate is equal to the entropy rate $H = H(P) = \lim_n E[-\log P(X_0 \,|\, X_{-n}^{-1})]$ of $\boldsymbol{X}$. Specifically, they showed that for stationary ergodic processes $(1/n)\log R_n$ converges to H in probability, and they suggested that this also holds in the almost sure sense. Indeed, this was later established by Ornstein and Weiss [53] who showed that for stationary

ergodic processes

$$\frac{1}{n}\log R_n \to H \quad \text{a.s.} \tag{3.1}$$

In their analysis, Wyner and Ziv used a theorem of Kac from [34] which can be phrased as follows: If $\boldsymbol{X}$ is stationary ergodic, then for any opening string x_1^n we have $E(R_n \,|\, X_1^n = x_1^n) = 1/P(x_1^n)$. This provides a strong formal connection between R_n and H: Taking logarithms of both sides in Kac's theorem, dividing by n and applying the Shannon-McMillan-Breiman theorem yields

$$\lim_n \frac{1}{n}\log E(R_n \,|\, X_1^n) = \lim_n \frac{1}{n}\log[1/P(X_1^n)] = H \quad \text{a.s.} \tag{3.2}$$

We can therefore rephrase the Wyner-Ziv-Ornstein-Weiss result (3.1) by saying that they strengthened (3.2) by removing the conditional expectation

$$\lim_n \frac{1}{n}\log R_n = \lim_n \frac{1}{n}\log[1/P(X_1^n)] = H \quad \text{a.s.} \tag{3.3}$$

The crucial observation here is that (3.3) can be thought of as a strong approximation result between $\log R_n$ and $-\log P(X_1^n)$:

$$\log[R_n P(X_1^n)] \;=\; o(n) \quad \text{a.s.} \tag{3.4}$$

Our first result is a sharper form of (3.4).

Theorem 3.1 (Strong approximation)

Let $\boldsymbol{X}$ be a finite-valued stationary ergodic process, and $\{c(n)\}$ an arbitrary sequence of non-negative constants such that $\sum n2^{-c(n)} < \infty$. We have,

$$\begin{aligned} &(i) \quad \log[R_n P(X_1^n)] \;\le\; c(n) \quad \textit{eventually a.s.} \\ &(ii) \quad \log[R_n P(X_1^n \,|\, X_{-\infty}^0)] \;\ge\; -c(n) \quad \textit{eventually a.s.} \end{aligned}$$

Theorem 3.1 is proved Section 3.2. Notice that it suffices to take $c(n) \ge 3\log n$ in order to satisfy the condition $\sum n2^{-c(n)} < \infty$. In particular, taking $c(n) = \epsilon n^\beta$ in Theorem 3.1 and letting $\epsilon \downarrow 0$ we obtain the following Corollary (proved in Section 3.2).

Corollary 3.1

(*a*) *For every finite-valued stationary ergodic process* X,

$$\log[R_n P(X_1^n)] \;=\; o(n) \quad a.s.$$

(*b*) *If, moreover,* $\sum \gamma(n) < \infty$ *then for any* $\beta > 0$

$$\log[R_n P(X_1^n)] \;=\; o(n^\beta) \quad a.s.$$

Recall that the coefficients $\gamma(n)$ were defined in (2.6) in Section 2.1, by

$$\gamma(n) \;=\; \max_{a \in A} E\,|\log P(X_0 = a \,|\, X_{-\infty}^{-1}) - \log P(X_0 = a \,|\, X_{-n}^{-1})|,$$

and the $\alpha-$mixing coefficients defined in (2.7) by

$$\alpha(n) \;=\; \sup\left\{|P(C \cap B) - P(C)P(B)| \;:\; B \in \sigma(X_{-\infty}^0),\; C \in \sigma(X_n^\infty)\right\}.$$

We can now use Corollary 3.1 to read off the exact asymptotic behavior of $\log R_n$ from that of $-\log P(X_1^n)$. As we saw in the previous chapter, if $\boldsymbol{X}$ is ergodic, the Shannon-McMillan-Breiman theorem says that $(-1/n)\log P(X_1^n)$ converges almost surely to H, and combining this with Corollary 3.1 we get (3.1). If $\boldsymbol{X}$ is a Markov chain or, more generally, if it satisfies certain conditions on the rate of decay of $\alpha(n)$ and $\gamma(n)$, then $-\log P(X_1^n)$ behaves like the partial sum sequence of a strongly mixing stationary process, so it satisfies a central limit theorem (CLT), a law of the iterated logarithm (LIL), their infinite dimensional (functional) counterparts, as well as an almost sure invariance principle. Combining Theorem 2.2 with Corollary 3.1 gives us an almost sure invariance principle for $\log R_n$: Define a continuous-time process $\{R(t)\,;\; t \geq 0\}$ by letting $R(t) = 0$ for $t \in [0,1)$ and $R(t) = [\log R_{\lfloor t \rfloor} - \lfloor t \rfloor H]$ for $t \geq 1$.

Theorem 3.2 (Almost sure invariance principle)

Let $\boldsymbol{X}$ *be a finite-valued stationary ergodic Markov chain, and let* σ^2 *be defined as in (2.9):*

$$\begin{aligned}\sigma^2 \;&=\; E[-\log P(X_0 \,|\, X_{-1}) - H]^2 \\ &+\; 2\sum_{k=1}^{\infty} E[(-\log P(X_0 \,|\, X_{-1}) - H)(-\log P(X_k \,|\, X_{k-1}) - H)].\end{aligned}$$

(i) If $\sigma^2 > 0$, then there exists a standard Brownian motion $\{B(t)\ ;\ t \geq 0\}$ such that

$$R(t) - \sigma B(t) = o(\sqrt{t}) \quad a.s.$$

(ii) Moreover, (i) remains true if the ergodic Markov chain assumption is replaced by the assumptions that $\alpha(n) = O(n^{-336})$ and $\gamma(n) = O(n^{-48})$, and σ^2 is replaced by the expression in (2.11).

It is now a routine matter (see, e.g., [57, Chapter 1]) to obtain from the almost sure invariance principle of Theorem 3.2 the second-order asymptotic behavior of R_n:

Corollary 3.2

Under the assumptions of Theorem 3.2, if $\sigma^2 > 0$:

(*i*) **CLT:**

$$\frac{\log R_n - nH}{\sigma\sqrt{n}} \xrightarrow{\mathcal{D}} N(0,1)$$

Moreover, the sequence of processes

$$\left\{ \frac{R(nt)}{\sigma\sqrt{n}}\ ;\ t \in [0,1] \right\}, \quad n \geq 1,$$

converges in distribution to standard Brownian motion.

(*ii*) **LIL:** *With probability one, the set of limit points of the sequence*

$$\left\{ \frac{\log R_n - nH}{\sqrt{2n \log_e \log_e n}} \right\}, \quad n \geq 3,$$

coincides with the interval $[-\sigma, \sigma]$. Moreover, with probability one, the sequence of sample paths

$$\left\{ \frac{R(nt)}{\sigma\sqrt{2n \log_e \log_e n}}\ ;\ t \in [0,1] \right\}, \quad n \geq 3,$$

is relatively compact in the topology of uniform convergence, and the set of its limit points is the collection of all absolutely continuous functions $r : [0,1] \to \mathbb{R}$, such that $r(0) = 0$ and $\int_0^1 (dr/dt)^2 dt \leq 1$.

Remark. Recall that, at least in the case of Markov chains, a characterization of the degenerate case $\sigma^2 = 0$ was provided in Chapter 2 by Theorem 2.3.

3.1.1 Match Lengths

The story of the asymptotics of R_n can equivalently be told in terms of match lengths along a realization. Given a realization $\boldsymbol{x}$ from $\boldsymbol{X}$, we define M_m as the length n of the longest string x_1^n starting at x_1 that also appears starting somewhere else in the previous m positions x_{-m+1}^0:

$$M_m \;=\; \sup\{n \geq 1 \;:\; x_1^n = x_{-j+1}^{-j+m}, \;\; \text{for some} \;\; j = 1, 2, \ldots, m\}.$$

Following Wyner and Ziv [69] we observe that there is a nice duality between recurrence times and match lengths, in that

$$M_m \geq n \quad \text{if and only if} \quad R_n \leq m. \tag{3.5}$$

Consequently, all asymptotic results about R_n can be translated into corresponding results about M_m. For example, the almost sure convergence of $(1/n)\log R_n$ to H is equivalent to

$$\frac{M_m}{\log m} \to \frac{1}{H} \quad \text{a.s.} \tag{3.6}$$

The CLT and LIL for $\log R_n$ (Corollary 3.2) translate to:

Corollary 3.3

Under the assumptions of Theorem 3.2:

(i) **CLT:**

$$\frac{M_m - \frac{\log m}{H}}{\sigma H^{-3/2}\sqrt{\log m}} \xrightarrow{\mathcal{D}} N(0,1)$$

(ii) **LIL:**

$$\limsup_{n\to\infty} \frac{M_m - \frac{\log m}{H}}{\sigma H^{-3/2}\sqrt{2\log m \log_e \log_e \log m}} \;=\; 1 \quad \textit{a.s.}$$

3.1.2 Earlier Work

In addition to the historical remarks in Section 1.3, we provide a few comments and references to more recent work and closely related results.

Wyner and Ziv discovered the result in (3.1), which was formally established by

Ornstein and Weiss [53] using methods from ergodic theory. In the Markov case, A.J. Wyner [72] used the Chen-Stein method for Poisson approximation and Markov coupling to prove a one-dimensional CLT for *waiting times*, and he also remarked that his methods can be modified to prove the one-dimensional CLT for R_n in Corollary 3.2 (i). In the memoryless case, Louchard and Szpankowski [44] and Jacquet and Szpankowski [33] proved implicit first- and second-order results for recurrence times, by exploiting a connection between match lengths in realizations of memoryless processes and feasible paths in random trees. We also mention that the first-order results abour recurrence times and match lengths were extended to processes with countably infinite alphabets and to random fields in several dimensions by Kontoyiannis *et al.* in [42].

The approach introduced in this chapter provides a natural probabilistic framework for studying the asymptotic behavior of R_n. From Theorem 3.1 and Corollary 3.1 we can deduce strong results that were not previously known, as well as new proofs of several known results. Moreover, Theorem 3.1 tells us why these results are true; because, in a strong pointwise sense, the recurrence time is asymptotically equal to the reciprocal of the probability of the recurring string.

3.2 Strong Approximation

We deduce Corollary 3.1 from Theorem 3.1 and give the proof of Theorem 3.1.

Proof of Corollary 3.1. For part (b) let $\beta > 0$ arbitrary. Since $\sum n2^{-\epsilon n^\beta} < \infty$ for any $\epsilon > 0$, from (i) and (ii) of Theorem 1 we get

$$\limsup_{n\to\infty} \frac{1}{n^\beta} \log\left[R_n P(X_1^n)\right] \;\leq\; 0 \quad \text{a.s.} \tag{3.7}$$

and

$$\liminf_{n\to\infty} \frac{1}{n^\beta} \log\left[R_n P(X_1^n \,|\, X_{-\infty}^0)\right] \;\geq\; 0 \quad \text{a.s.,} \tag{3.8}$$

so it suffices to show that

$$\log P(X_1^n) - \log P(X_1^n \,|\, X_{-\infty}^0) = O(1) \quad \text{a.s.} \tag{3.9}$$

First, expanding,

$$| \log P(X_1^n) - \log P(X_1^n \,|\, X_{-\infty}^0) | \le \sum_{i=1}^{n} | \log P(X_i \,|\, X_1^{i-1}) - \log P(X_i \,|\, X_{-\infty}^{i-1}) |,$$

and then taking expectations,

$$E| \log P(X_1^n) - \log P(X_1^n \,|\, X_{n+1}^{\infty})| \le \sum_{i=1}^{n} \gamma(i).$$

Now $\sum_{i=1}^{\infty} \gamma(i) < \infty$ implies (3.9).

For part (a), taking $\beta = 1$ in equations (3.7) and (3.8) above, we see that it suffices to show

$$\frac{1}{n} \left[\log P(X_1^n) - \log P(X_1^n \,|\, X_{-\infty}^0) \right] \to 0 \quad \text{a.s.} \tag{3.10}$$

By the Shannon-McMillan-Breiman theorem, the first term converges almost surely to $-H$, and the second term equals $(1/n) \sum_{i=1}^{n} [-\log P(X_i \,|\, X_{-\infty}^{i-1})]$, which converges to $E[-\log P(X_0 \,|\, X_{-\infty}^{-1})] = \lim_n E[-\log P(X_0 \,|\, X_{-n}^{-1})] = H$ almost surely, by the ergodic theorem and the definition of H. This proves (3.10) and completes the proof. □

Proof of Theorem 3.1. Part (i). Given an arbitrary positive constant K, by Markov's inequality and Kac's theorem,

$$P(R_n > K \,|\, X_1^n = x_1^n) \;\le\; \frac{E(R_n \,|\, X_1^n = x_1^n)}{K} \;=\; \frac{1}{K\, P(x_1^n)},$$

for any opening sequence x_1^n with non-zero probability. Since $P(x_1^n)$ is constant with respect to the conditional measure $P(\cdot \,|\, X_1^n = x_1^n)$, we can let $K = 2^{c(n)}/P(x_1^n)$ to get

$$P(\, \log[R_n P(X_1^n)] > c(n) \,|\, X_1^n = x_1^n) \;=\; P\left(R_n > 2^{c(n)}/P(x_1^n) \,|\, X_1^n = x_1^n\right) \;\le\; 2^{-c(n)}.$$

Averaging over all opening patterns $x_1^n \in A^n$, $P(\, \log[R_n P(X_1^n)] > c(n)\,) \le 2^{-c(n)}$, and the Borel-Cantelli lemma gives (i).

Part (ii). We now condition on the infinite past $X_{-\infty}^0$ instead of the opening string X_1^n. Fix any $x_{-\infty}^0$ and consider

$$P\left\{ \log[R_n(X) P(X_1^n \,|\, X_{-\infty}^0)] < -c(n) \,|\, X_{-\infty}^0 = x_{-\infty}^0 \right\} =$$

$$P\left\{z_1^n \in A^n \,:\, P(X_1^n = z_1^n \,|\, X_{-\infty}^0) < \frac{2^{-c(n)}}{R_n(x_{-\infty}^0 * z_1^n)} \;\middle|\; X_{-\infty}^0 = x_{-\infty}^0\right\},$$

where $*$ denotes concatenation of strings. If we let $G_n = G_n(x_{-\infty}^0)$ denote the set

$$\left\{z_1^n \in A^n \,:\, P(z_1^n \,|\, x_{-\infty}^0) < 2^{-c(n)}/R_n(x_{-\infty}^0 * z_1^n)\right\},$$

then the above probability can be written as

$$\begin{aligned}\sum_{z_1^n \in G_n} P(z_1^n \,|\, x_{-\infty}^0) &\le \sum_{z_1^n \in G_n} 2^{-c(n)}/R_n(x_{-\infty}^0 * z_1^n) \\ &\le 2^{-c(n)} \sum_{z_1^n \in A^n} 1/R_n(x_{-\infty}^0 * z_1^n). \qquad (3.11)\end{aligned}$$

Since $x_{-\infty}^0$ is fixed, for each $j \ge 1$ there is exactly one string z_1^n from A^n with $R_n(x_{-\infty}^0 * z_1^n) = j$, so the sum in (3.11) is bounded above by

$$\sum_{z_1^n \in A^n} 1/R_n(x_{-\infty}^0 * z_1^n) \;\le\; \sum_{j=1}^{s^n} 1/j \;\le\; Cn,$$

for some positive constant C, where $s = |A|$ is the cardinality of A. Therefore,

$$P(\log[R_n P(X_1^n \,|\, X_{-\infty}^0)] < -c(n) \,|\, X_{-\infty}^0 = x_{-\infty}^0) \le Cn2^{-c(n)},$$

and since this bound is independent of $x_{-\infty}^0$ and summable over n, from the Borel-Cantelli lemma we deduce (ii). □

Remark. In the proof of (ii), only the stationarity (and not the ergodicity) of $\boldsymbol{X}$ was used.

Chapter 4

Waiting Times Between Stationary Processes

4.1 Motivation

In this chapter, we consider a more general version of the recurrence times problem addressed in Chapter 3. We ask how long it takes before a random pattern generated by some process $\boldsymbol{X}$ first appears in a realization of a (possibly different) process $\boldsymbol{Y}$. Going a step further, we allow for distortion in the patterns, and we ask for the first "approximate" appearance of a random pattern, within some prescribed accuracy.

To be precise, we consider two stationary processes $\boldsymbol{X} = \{X_n \ ; \ n \in \mathbb{Z}\}$ and $\boldsymbol{Y} = \{Y_n \ ; \ n \in \mathbb{Z}\}$ taking values in the Polish (= complete, seperable, metric) spaces $(A^\infty, \mathcal{A}^\infty)$ and $(\hat{A}^\infty, \hat{\mathcal{A}}^\infty)$, and distributed according to the probability measures P and Q, respectively. To measure "closeness," we fix a nonnegative measurable function ρ on $A \times \hat{A}$, and define the *distortion* between two finite strings $x_1^n \in A^n$ and $y_1^n \in \hat{A}^n$ by

$$\rho(x_1^n, y_1^n) = \frac{1}{n} \sum_{i=1}^{n} \rho(x_i, y_i).$$

For $x_1^n \in A^n$ and $D \geq 0$, we write $B(x_1^n, D)$ for the ball of radius D around x_1^n:

$$B(x_1^n, D) = \{y_1^n \in \hat{A}^n \ : \ \rho(x_1^n, y_1^n) \leq D\}.$$

Given $D \geq 0$ and two independent realizations $\boldsymbol{x}$, $\boldsymbol{y}$ from $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively, our main quantity of interest here is the *waiting time* $W_n(D)$ until a D-close version of x_1^n first appears in y_1^∞:

$$W_n(D) \;=\; \inf\{k \geq 1 \,:\, y_k^{k+n-1} \in B(x_1^n, D)\}.$$

In the special case when we look for *exact* matches (corresponding to taking $D = 0$ and ρ being Hamming distortion), we omit the D and write W_n instead of $W_n(D)$.

The asymptotic behavior of waiting times has been studied extensively and very actively during the past ten years, motivated primarily by important applications in the areas of data compression, DNA sequence analysis, and string matching algorithms in computer science. Some examples of these applications were discussed in Chapter 1. In Chapter 5 we will see how the results we obtain here can be interpreted in the context of data compression (Section 5.2.1), and also how they can be extended to prove the optimality of a new data compression algorithm (Section 5.2.2).

Before we move on to our own results, in the next section we briefly describe what is already known in this area.

4.1.1 Earlier Work

In its simplest form, the problem of the asymptotic behavior of waiting times first appeared in Wyner and Ziv's work on data compression [69]: $\boldsymbol{X}$ and $\boldsymbol{Y}$ were assumed to have the same distribution over the same *finite* alphabet, and no distortion was allowed. In that case, Wyner and Ziv showed that $(1/n)\log W_n$ converges, in probability, to the entropy rate H of $\boldsymbol{X}$. Moreover, they suggested that the same result holds in the almost sure sense,

$$\frac{1}{n}\log W_n \to H \quad \text{a.s.} \tag{4.1}$$

and this was later established by Shields [63] using ideas and methods from ergodic theory. These results were extended further, first by Nobel and Wyner [51] who showed that the convergence in probability holds for processes that are α-mixing with a certain rate, and then by Marton and Shields [46] who proved that (4.1) holds for the class of weak Bernoulli processes. Shields [63] also provided a counter-example

to show that (4.1) does not hold in the general ergodic case. Finally, a CLT-refinement to (4.1) was discovered by A.J. Wyner in his Ph.D. thesis [72], where he used Poisson approximation to prove that, when $\boldsymbol{X}$ is a finite-state, stationary ergodic Markov chain,

$$\frac{\log W_n - nH}{\sigma\sqrt{n}} \xrightarrow{\mathcal{D}} N(0,1).$$

Much less is known in the case when distortion is allowed. Recently, Łuczak and Szpankowski [45] showed that, for processes $\boldsymbol{X}$ and $\boldsymbol{Y}$ with a finite alphabet, $(1/n)\log W_n(D)$ converges to some constant R with probability one. Independently and around the same time Yang and Kieffer [75] also proved the same result, and they identified the constant R as the solution to a variational problem in terms of relative entropy (see Theorem 2.4 in Chapter 2, or Corollary 4.7 below).

In this chapter we introduce a natural probabilistic framework which gives us a unified strategy for greatly generalizing and extending these recent results, and also allows us to recover most of the known asymptotic results for waiting times in full generality.

4.1.2 The Strong Approximation Framework

The gist of the approach we took in Chapter 3 to understand recurrence times was to realize that the time R_n until a match for the pattern X_1^n is found is approximately equal to the reciprocal of the probability $P(X_1^n)$ of this pattern. Our main idea here is to extend this intuition to the case of waiting times: We claim that the time $W_n(D)$ until a D-close match for the pattern X_1^n appears can be approximated by the reciprocal of the probability $Q(B(X_1^n, D))$ of finding such a match:

$$W_n(D) \approx \frac{1}{Q(B(X_1^n, D))}.$$

This claim is made precise in our first result, Theorem 4.1, which enables us to deduce the asymptotic properties of the waiting times $W_n(D)$ from the corresponding properties of the probabilities $Q(B(X_1^n, D))$. In view of Chapter 2 and the detailed study of $Q(B(X_1^n, D))$ it contains, this is a very pleasant position to be in!

The power of this formulation is amply demonstrated in the next three subsections, where we state *nine* non-trivial, immediate corollaries of Theorem 4.1, providing a

complete description of the asymptotic behavior of $W_n(D)$.

Theorem 4.1 (Strong approximation)

Let $\boldsymbol{X}$ be a stationary ergodic process, let $\boldsymbol{Y}$ be a stationary process with ϕ-mixing coefficients that satisfy $\sum \phi(n) < \infty$, and assume that $Q(B(X_1^n, D)) > 0$, eventually $P-$almost surely. For any sequence of nonnegative constants $\{c(n)\}$ such that $\sum ne^{-c(n)} < \infty$ we have,

$$|\log[W_n(D)Q(B(X_1^n, D))]| \leq c(n) \quad \textit{eventually } P \times Q-\textit{a.s.}$$

[Recall from (2.8) of Chapter 2 that the ϕ-mixing coefficients of the process $\boldsymbol{Y} \sim Q$ are defined by $\phi(n) = \sup\{|Q(C|B) - Q(C)| : B \in \sigma(Y_{-\infty}^0),\ C \in \sigma(Y_n^\infty)\}$.]

Taking $c(n) = \epsilon\sqrt{n}$ in Theorem 4.1 and letting $\epsilon \downarrow 0$ we obtain

$$\log W_n(D) - \log[1/Q(B(X_1^n, D))] = o(\sqrt{n}) \quad P \times Q - \text{a.s.} \tag{4.2}$$

Now we can combine (4.2) with the various results of Chapter 2 about $Q(B(X_1^n, D))$ to harvest the fruits of our labor there, in the form of a series of interesting corollaries about waiting times.

4.2 Waiting Times Results

We consider three separate cases.

4.2.1 Waiting Times With No Distortion

When $\boldsymbol{X}$ and $\boldsymbol{Y}$ have the same distribution P and take values in the same finite alphabet A, it is clear that $Q(B(X_1^n, D))$ is just $P(X_1^n)$, and, of course, $P(X_1^n) > 0$ with $P-$probability one. Therefore, we can apply Theorem 4.1, and rewrite (4.2) as

$$\log W_n - \log[1/P(X_1^n)] = o(\sqrt{n}) \quad P \times P - \text{a.s.} \tag{4.3}$$

Combining this with the Shannon-McMillan-Breiman theorem (Theorem 2.1):

Corollary 4.1 (SLLN; Marton & Shields [46])

Let $\boldsymbol{X}$ and $\boldsymbol{Y}$ be finite-valued stationary processes, with the same distribution P, entropy rate H, and ϕ-mixing coefficients that satisfy $\sum \phi(n) < \infty$. We have,

$$\frac{1}{n} \log W_n \;\to\; H \quad P \times P - a.s.$$

Similarly, combining (4.3) with the almost sure invariance principle of Theorem 2.2, gives us an almost sure invariance principle for W_n. Define a continuous-time process $\{w(t)\ ;\ t \geq 0\}$ by $w(t) = 0$, $t \in [0,1)$, and $w(t) = [\log W_{\lfloor t \rfloor} - \lfloor t \rfloor H]$, $t \geq 1$. Recall that the α- and γ-mixing coefficients for $\boldsymbol{X} \sim P$ were defined in (2.7) and (2.6) by $\alpha(n) = \sup\big\{|P(C \cap B) - P(C)P(B)| \ :\ B \in \sigma(X_{-\infty}^0),\ C \in \sigma(X_n^\infty)\big\}$ and $\gamma(n) = \max_{a \in A} E|\log P(X_0 = a \,|\, X_{-\infty}^{-1}) - \log P(X_0 = a \,|\, X_{-n}^{-1})|$, respectively

Corollary 4.2 (Almost sure invariance principle)

Suppose $\boldsymbol{X}$ and $\boldsymbol{Y}$ are finite-valued stationary ergodic Markov chains with the same distribution P and entropy rate H, and let σ^2 be defined as in (2.9):

$$\begin{aligned} \sigma^2 &= E[-\log P(X_0 \,|\, X_{-1}) - H]^2 \\ &+ 2\sum_{k=1}^{\infty} E[(-\log P(X_0 \,|\, X_{-1}) - H)(-\log P(X_k \,|\, X_{k-1}) - H)]. \end{aligned}$$

(i) If $\sigma^2 > 0$, then there exists a standard Brownian motion $\{B(t)\ ;\ t \geq 0\}$ such that

$$w(t) - \sigma B(t) = o(\sqrt{t}) \quad a.s.$$

(ii) Moreover, (i) remains true if the ergodic Markov chain assumption is replaced by the assumptions that $\alpha(n) = O(n^{-336})$, $\sum \phi(n) < \infty$, and $\gamma(n) = O(n^{-48})$, and σ^2 replaced by the expression in (2.11).

As in the case of recurrence times in Chapter 3, from the above almost sure invariance principle we immediately conclude that $\log W_n$ satisfies a central limit theorem, a law of the iterated logarithm, and their functional counterparts:

Corollary 4.3

Under the assumptions of Corollary 4.2, if $\sigma^2 > 0$:

(i) **CLT:**

$$\frac{\log W_n - nH}{\sigma\sqrt{n}} \xrightarrow{\mathcal{D}} N(0,1)$$

Moreover, the sequence of processes

$$\left\{ \frac{w(nt)}{\sigma\sqrt{n}} \; ; \; t \in [0,1] \right\}, \quad n \geq 1,$$

converges in distribution to standard Brownian motion.
(ii) **LIL:** *With probability one, the set of limit points of the sequence*

$$\left\{ \frac{\log W_n - nH}{\sqrt{2n \log_e \log_e n}} \right\}, \quad n \geq 3,$$

coincides with the interval $[-\sigma, \sigma]$. *Moreover, with probability one, the sequence of sample paths*

$$\left\{ \frac{w(nt)}{\sigma\sqrt{2n \log_e \log_e n}} \; ; \; t \in [0,1] \right\}, \quad n \geq 3,$$

is relatively compact in the topology of uniform convergence, and the set of its limit points is the collection of all absolutely continuous functions $r : [0,1] \to \mathbb{R}$, *such that* $r(0) = 0$ *and* $\int_0^1 (dr/dt)^2 dt \leq 1$.

[As we already mentioned, in the Markov case, the one-dimensional version of the CLT in (i) was proved by A.J. Wyner in his Ph.D. thesis [72].]

4.2.2 Waiting Times Between Different Processes

Now suppose that the processes $\boldsymbol{X}$ and $\boldsymbol{Y}$ take values in the same finite alphabet, but they have different distributions given by the measures P and Q, respectively. Throughout this section we assume that, for all n large enough, the finite dimensional marginals of $\boldsymbol{X}$ are dominated by the corresponding marginals of $\boldsymbol{Y}$,

$$P_n \ll Q_n \quad \text{eventually,}$$

otherwise the waiting times W_n will be infinite with positive probability. This assumption clearly implies that $Q(B(X_1^n, D)) = Q(X_1^n) > 0$, eventually $P-$almost surely, so

we can invoke Theorem 4.1 to get, via (4.2), that

$$\log W_n - \log[1/Q(X_1^n)] \;=\; o(\sqrt{n}) \quad P \times Q - \text{a.s.} \tag{4.4}$$

The behavior of the waiting times can now be deduced from combining this with Proposition 2.1.

Corollary 4.4 (SLLN)

Let $\boldsymbol{X}$ be a finite-valued stationary ergodic process with distribution P, $\boldsymbol{Y}$ be a stationary ergodic Markov chain with distribution Q, and assume that $P_n \ll Q_n$ eventually. We have,

$$\frac{1}{n}\log W_n \to H(P) + H(P\|Q) \quad a.s.$$

where $H(P)$ is the entropy rate of $\boldsymbol{X}$, and $H(P\|Q)$ is the relative entropy rate between $\boldsymbol{X}$ and $\boldsymbol{Y}$,

$$H(P\|Q) \stackrel{\triangle}{=} \lim_{n\to\infty} E_P\left[\log \frac{P(X_0 \mid X_{-n}^{-1})}{Q(X_0 \mid X_{-n}^{-1})}\right].$$

Corollary 4.5 (Almost sure invariance principle)

Let $\boldsymbol{X}$ and $\boldsymbol{Y}$ be finite-valued stationary ergodic Markov chains with distribution P and Q, respectively, assume that $P_n \ll Q_n$ eventually, and let σ^2 be defined as in equation (2.13):

$$\sigma^2 \quad = \quad = \lim_{n\to\infty} \mathrm{Var}_P(-\log Q(X_1^n)). \tag{4.5}$$

If $\sigma^2 > 0$, then there exists a standard Brownian motion $\{B(t)\,;\, t \geq 0\}$ such that

$$\tilde{w}(t) - \sigma B(t) = o(\sqrt{t}) \quad a.s.$$

where $\{\tilde{w}(t)\,;\, t \geq 0\}$ is the continuous-time process defined by letting $\tilde{w}(t) = 0$ for $t \in [0,1)$ and $\tilde{w}(t) = [\log W_{\lfloor t \rfloor} - \lfloor t \rfloor (H(P) + H(P\|Q))]$ for $t \geq 1$.

As before, this immediately implies:

Corollary 4.6

Under the assumptions of Corollary 4.5, the CLT and LIL (and their functional

counterparts) of Corollary 4.3 remain valid in this case, with σ^2 *defined as in (4.5),* H *replaced by* $[H(P) + H(P\|Q)]$*, and* $w(\cdot)$ *replaced by* $\tilde{w}(\cdot)$*.*

4.2.3 Waiting Times Allowing Distortion

Next we turn to the most interesting case, the case when distortion is allowed. As in Chapter 2, we define

$$\begin{aligned} D_{\min} &\stackrel{\triangle}{=} E_{P_1}[\operatorname*{ess\,inf}_{Y_1 \sim Q_1} \rho(X_1, Y_1)] \\ D_{\text{av}} &\stackrel{\triangle}{=} E\rho(X_1, Y_1), \end{aligned}$$

and we assume that

$$D_{\max} \stackrel{\triangle}{=} \operatorname*{ess\,sup}_{(X_1,Y_1)} \rho(X_1, Y_1) \in (D_{\min}, \infty).$$

For simplicity, we will also assume throughout this section that the process $\boldsymbol{Y}$ is a sequence of independent and identically distributed random variables (an "i.i.d. process").

Since $\boldsymbol{X}$ is stationary ergodic, if $D < D_{\min}$ we will have $W_n(D) = \infty$, eventually almost surely, by the ergodic theorem. Similarly, if $D > D_{\text{av}}$ then $W_n(D) = 1$, eventually almost surely. We, therefore, concentrate on the range of interesting distortion values between $D_{\min}$ and D_{av}, where $W_n(D)$ increases exponentially. In that range we have $Q(B(X_1^n, D)) > 0$, eventually $P-$almost surely, so we can apply Theorem 4.1. Our first result comes from combining (4.2) with Corollary 2.1 and (2.19):

Corollary 4.7 (SLLN)

Let $\boldsymbol{X}$ *be a stationary ergodic process,* $\boldsymbol{Y}$ *be and i.i.d. process, and assume* $D \in (D_{\min}, D_{\text{av}})$*. We have,*

$$\frac{1}{n} \log W_n(D) \to R(P_1, Q_1, D) \quad P \times Q-a.s.$$

where

$$R(P_1, Q_1, D) = \inf \int H(\Theta(\cdot|x)\|Q_1(\cdot)) d\hat{P}_n(x) = \inf[\, I(X;Y) + H(Q_1'\|Q_1)\,] \quad (4.6)$$

where the infimum is taken over all random variables (X,Y) *such that* $E\rho(X,Y) \leq D$, $X \sim P_1$, Θ *denotes the conditional distribution of* Y *given* X *and* $Y \sim Q_1'$.

(Recall that an alternative characterization of $R(P_1,Q_1,D)$ was given in Proposition 2.2.) Before moving on to the corresponding second-order results, we recall from Chapter 2 the averaged logarithmic moment generating function $\Lambda(\cdot)$,

$$\Lambda(\lambda) \;=\; \int \log_e \left\{ \int e^{\lambda\rho(x,y)} dQ_1(y) \right\} dP_1(x),$$

the function $h(\cdot;\cdot)$ defined by

$$h(\lambda;x) \stackrel{\triangle}{=} \log \left\{ \int e^{\lambda\rho(x,y)} dQ_1(y) \right\} - (\log e)\Lambda(\lambda), \tag{4.7}$$

and that for any $D \in (D_{\min}, D_{\rm av})$ we can choose a $\lambda = \lambda(D) < 0$ such that $\Lambda'(\lambda) = D$ (by Lemma 2.1). Now we can combine (4.2) with Corollaries 2.2, and 2.3, to obtain the CLT and LIL for the waiting times $W_n(D)$:

Corollary 4.8

Let $\boldsymbol{X}$ *be a stationary process with* α*-mixing coefficients satisfying* $\sum \alpha(n) < \infty$ *and let* $\boldsymbol{Y}$ *be an i.i.d. process. Given* $D \in (D_{\min}, D_{\rm av})$, *let* σ^2 *be defined as in (2.24),*

$$\sigma^2 \;=\; E_P\left\{h(\lambda(D);X_1)^2\right\} + 2\sum_{k=2}^{\infty} E_P\left\{h(\lambda(D);X_1)h(\lambda(D);X_k)\right\}, \tag{4.8}$$

with h *given by (4.7). If* $\sigma^2 > 0$, *we have:*
(i) **CLT:**

$$\frac{\log W_n(D) - nR(P_1)}{\sqrt{n}} \xrightarrow{\mathcal{D}} N(0,\sigma^2).$$

Moreover, the sequence of processes

$$\left\{ \frac{w(nt;D)}{\sigma\sqrt{n}} \;;\; t \in [0,1] \right\}, \quad n \geq 1$$

converges in distribution to standard Brownian motion, where $\{w(t;D) \;;\; t \geq 0\}$ *is the continuous-time process defined by letting* $w(t;D) = 0$ *for* $t \in [0,1)$ *and* $w(t;D) = [\log W_{\lfloor t \rfloor}(D) - \lfloor t \rfloor R(P_1,Q_1,D)]$ *for* $t \geq 1$.

(*ii*) **LIL:** *With $P \times Q-$probability one, the set of limit points of the sequence*

$$\left\{ \frac{\log W_n(D) - nR(P_1)}{\sqrt{2n \log_e \log_e n}} \right\}, \quad n \geq 3$$

coincides with the interval $[-\sigma, \sigma]$. Moreover, with $P \times Q-$probability one, the sequence of sample paths

$$\left\{ \frac{w(nt; D)}{\sqrt{2n \log_e \log_e n}} \; ; \; t \in [0,1] \right\}, \quad n \geq 3,$$

is relatively compact in the topology of uniform convergence, and the set of its limit points is the collection of all absolutely continuous functions $r : [0,1] \to \mathbb{R}$, such that $r(0) = 0$ and $\int_0^1 (dr/dt)^2 dt \leq \sigma^2$.

Finally, combining (4.2) and Corollary 2.4, yields:

Corollary 4.9 (Almost sure invariance principle)

Let $\boldsymbol{X}$ be a stationary process with ϕ-mixing coefficients satisfying $\sum \phi(n) < \infty$ and let $\boldsymbol{Y}$ be an i.i.d. process. Given $D \in (D_{\min}, D_{\text{av}})$, let σ^2 be defined as in (4.8). If $\sigma^2 > 0$, then there exists a standard Brownian motion $\{B(t) \; ; \; t \geq 0\}$ such that

$$w(t; D) - \sigma B(t) = o(\sqrt{t}) \quad a.s.$$

4.3 Match Lengths Results

As in the case of recurrence times, here also the waiting times story can equivalently be told in terms of match lengths between realizations: Given an integer $m \geq 1$, a distortion level D, and two independent realizations $\boldsymbol{x}$ and $\boldsymbol{y}$ from the processes $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively, we look for the *longest* string x_1^ℓ that matches, within distortion D, somewhere in y_1^m. The length $L_m(D)$ of this longest match is of interest here:

$$L_m(D) \; = \; \sup\{n \geq 1 \; : \; y_j^{j+n-1} \in B(x_1^n, D), \;\; \text{for some} \;\; j = 1, 2, \ldots, m\}.$$

When no distortion is allowed, we omit the D and write L_m instead of $L_m(D)$.

As we discussed in Chapter 1 (Section 1.2.3), intuitively we expect that there is some sort of relationship between match lengths and waiting times; it seems plausible that long match lengths should imply short waiting times, and vice versa. In this section we show that this intuition can be made precise, and that we can use it to translate our waiting times results into corresponding results for match lengths. Note, however, that when we allow for matches with distortion, the duality relationship between $W_n(D)$ and $L_m(D)$ becomes a more complex one, so that there is some work to be done in the "translation" from $W_n(D)$ to $L_m(D)$.

Let us begin again with simplest case. Suppose that no distortion is allowed, and the processes $\boldsymbol{X}$ and $\boldsymbol{Y}$ have the same distribution P over the same finite alphabet A. Here the duality between L_m and W_n is manifested in precisely the same way as in the context of recurrence times (cf. (3.5)):

$$L_m \geq n \quad \text{if and only if} \quad W_n \leq m. \tag{4.9}$$

Therefore, just as in Section 3.1.1, all asymptotic results about waiting times immediately give us corresponding results about match lengths:

Corollary 4.10 (Match lengths without distortion)

Under the assumptions of Corollary 4.1, we have:

(i) **SLLN:**

$$\frac{L_m}{\log m} \to \frac{1}{H} \quad P \times P - a.s.$$

where $H = H(P)$ is the entropy rate of $\boldsymbol{X}$. Moreover, under the assumptions Corollary 4.3 we have:

(ii) **CLT:**

$$\frac{L_m - \frac{\log m}{H}}{\sqrt{\log m}} \xrightarrow{\mathcal{D}} N(0, \sigma^2 H^{-3})$$

(iii) **LIL:**

$$\limsup_{m\to\infty} \frac{L_m - \frac{\log m}{H}}{\sqrt{2 \log m \, \log_e \log_e \log m}} = \sigma H^{-3/2} \quad P \times P - a.s.$$

where σ^2 is defined by (2.11).

We note that the almost sure convergence in (i) was discovered by Wyner and Ziv [69], and, in the Markov case, the CLT in (ii) was first proved by A.J. Wyner in his Ph.D. thesis [72].

Now, if $\boldsymbol{X}$ and $\boldsymbol{Y}$ have different distributions given by the measures P and Q, respectively, over the finite alphabet A:

Corollary 4.11 (Match lengths between different processes)
Under the assumptions of of Corollary 4.4 we have:
(i) **SLLN:**

$$\frac{L_m}{\log m} \to \frac{1}{\tilde{H}} \quad P \times Q - a.s.$$

where $\tilde{H} \stackrel{\triangle}{=} H(P) + H(P\|Q)$. Moreover, under the assumptions of Corollary 4.6 we have:
(ii) **CLT:**

$$\frac{L_m - \frac{\log m}{\tilde{H}}}{\sqrt{\log m}} \stackrel{\mathcal{D}}{\longrightarrow} N(0, \sigma^2 \tilde{H}^{-3})$$

(iii) **LIL:**

$$\limsup_{m\to\infty} \frac{L_m - \frac{\log m}{\tilde{H}}}{\sqrt{2 \log m \, \log_e \log_e \log m}} = \sigma \tilde{H}^{-3/2} \quad P \times Q - a.s.$$

where σ^2 is defined by (4.5).

Coming back to the general case, suppose that the processes $\boldsymbol{X} \sim P$ and $\boldsymbol{Y} \sim Q$ take values in the general alphabets A and $\hat{A}$, respectively, and that nonzero distortion is allowed (we still assume that $\boldsymbol{X}$, $\boldsymbol{Y}$ and the distortion measure ρ are defined as in Section 4.2.3, above). Here, although the general intuition of the duality relationship between waiting times and match length remains true, its mathematical form has to be modified to:

$$L_m(D) \geq n \quad \text{if and only if} \quad [W_k(D) \leq m \ \text{ for some } k \geq n]. \tag{4.10}$$

The reason why here we need to consider all possible waiting times $W_k(D)$, $k \geq n$, is that, unlike the case of no distortion, here the sequence $\{W_k(D)\ ;\ k \geq n\}$ is no longer monotonically increasing, as can be seen from the following simple binary example: If we let ρ be Hamming distortion, and set $D = 0.4$, then for the realizations

$$\begin{aligned} x_1^\infty &= 0\,0\,1\,\cdots \\ y_1^\infty &= 1\,0\,1\,0\,0\,1\,1\,\cdots \end{aligned}$$

we have $W_2(D) = 4$ but $W_3(D) = 1$.

The relationship (4.10) is exploited in Section 4.5 to show how we can still recover the asymptotic behavior of $L_m(D)$ from that of $W_n(D)$:

Theorem 4.2 (Match lengths with distortion)

Under the assumptions of of Corollary 4.7 we have:

(i) **SLLN:**

$$\frac{L_m(D)}{\log m} \to \frac{1}{R} \quad P \times Q - a.s.$$

where $R = R(P_1, Q_1, D)$. Moreover, under the assumptions of Corollary 4.8 we have:

(ii) **CLT:**

$$\frac{L_m(D) - \frac{\log m}{R}}{\sqrt{\log m}} \xrightarrow{\mathcal{D}} N(0, \sigma^2 R^{-3})$$

(iii) **LIL:**

$$\limsup_{m\to\infty} \frac{L_m(D) - \frac{\log m}{R}}{\sqrt{2 \log m \, \log_e \log_e \log m}} = \sigma R^{-3/2} \quad P \times Q - a.s.$$

where σ^2 is defined by (4.8).

4.4 Strong Approximation

Proof of Theorem 4.1. Write $\mathbb{P}$ for the product measure $P \times Q$, and for each integer $m \geq 1$ let $G_m = \{\boldsymbol{x} : \ Q(B(x_1^n, D)) > 0 \text{ for all } n \geq m\}$. For the upper bound we use a standard second-moment blocking argument.

Choose and fix any integer $m \geq 1$, pick an arbitrary $\boldsymbol{x} \in G_m$, and let $n \geq m$ be large enough so that $e^{c(n)} \geq n+1$. Let $K \geq n+1$ and write $S_n = \sum_{j=0}^{V(K,n)} I_n(j)$, where $I_n(j)$ is the indicator function of $\{Y_{jn+1}^{(j+1)n} \in B(x_1^n, D)\}$ and $V(K,n) = \lfloor (K-1)/n \rfloor$. Then

$$\mathbb{P}(W_n(D) > K \,|\, X_1^n = x_1^n) \ \leq \ Q(S_n = 0) \ \leq \ \frac{\mathrm{Var}_Q(S_n)}{(E_Q S_n)^2}. \tag{4.11}$$

By stationarity,

$$E_Q S_n = [V(K,n)+1] Q(B(x_1^n, D)), \tag{4.12}$$

and $E_Q(I_n(0) I_n(j)) \leq Q(B(x_1^n, D))[\phi((j-1)n+1) + Q(B(x_1^n, D))]$, so that

$$\begin{aligned} \mathrm{Var}_Q(S_n) \ &= \ \sum_{j,k=0}^{V(K,n)} \mathrm{Cov}_Q(I_n(j) I_n(k)) \\ &\leq \ [V(K,n)+1] Q(B(x_1^n, D)) \left[1 + 2 \sum_{j=1}^{V(K,n)} \phi((j-1)n+1) \right]. \end{aligned} \tag{4.13}$$

Writing $\Phi = 1 + 2\sum \phi(k)$ and substituting (4.12) and (4.13) in (4.11) we get

$$\mathbb{P}(W_n(D) > K \,|\, X_1^n = x_1^n) \ \leq \ \frac{\Phi}{[V(K,n)+1]\, Q(B(x_1^n, D))}. \tag{4.14}$$

Choosing $K = e^{c(n)}/Q(B(x_1^n, D))$ we have $[V(K,n)+1]Q(B(x_1^n, D)) > e^{c(n)}/2n$, and (4.14) yields

$$\mathbb{P}(\log[W_n(D) Q(B(X_1^n, D))] > c(n) \,|\, X_1^n = x_1^n) \ \leq \ 2\Phi n e^{-c(n)}.$$

Since the above bound is uniform over $\boldsymbol{x} \in G_m$ and summable, by the Borel-Cantelli

lemma we obtain that, for $P \times Q-$ almost all $(\boldsymbol{x}, \boldsymbol{y}) \in G_m \times \hat{A}^\infty$,

$$\log[W_n(D)Q(B(x_1^n, D))] \leq c(n) \quad \text{eventually.} \tag{4.15}$$

For the lower bound, observe that for an any constant $K > 1$ and any $\boldsymbol{x} \in G_m$,

$$\mathbb{P}(W_n(D) < K \mid X_1^n = x_1^n) \leq \sum_{j=1}^{\lfloor K \rfloor} Q(Y_j^{j+n-1} \in B(x_1^n, D)) \leq K\, Q(B(x_1^n, D)). \tag{4.16}$$

Since $W_n(D) \geq 1$, this inequality holds also for $K \in [0, 1]$. In particular, setting $K = e^{-c(n)}/Q(B(x_1^n, D))$ gives

$$\mathbb{P}(\log[W_n(D)Q(B(X_1^n, D))] < -c(n) \mid X_1^n = x_1^n) \leq e^{-c(n)},$$

and summing this over n, by the Borel-Cantelli lemma we get that, for $P \times Q-$ almost all $(\boldsymbol{x}, \boldsymbol{y}) \in G_m \times \hat{A}^\infty$,

$$\log[W_n(D)Q(B(X_1^n, D))] \;\geq\; -c(n) \quad \text{eventually.} \tag{4.17}$$

Finally, combining (4.15) and (4.17) with the assumption that $P\{\cup_m G_m\} = 1$ completes the proof. □

4.5 Duality: Match Lengths

Let R denote $R(P_1, Q_1, D)$, and define, for $n \geq 1$, $T_n(D) = \inf_{k \geq n} W_k(D)$, and $\tilde{T}_n(D) = \min_{n \leq k \leq 2n} W_k(D)$. The duality relationship (4.10) between $W_n(D)$ and $L_m(D)$ can be restated as:

$$L_m(D) \geq n \quad \Longleftrightarrow \quad T_n(D) \leq m\ . \tag{4.18}$$

When combined with Lemma 4.1 below, (4.18) allows us to deduce (i), (ii) and (iii) in Theorem 4.2 from corresponding results for $\tilde{T}_n(D)$, namely, in the notation and under the corresponding assumptions of Theorem 4.2:

$$(i') \qquad \frac{\log \tilde{T}_n(D)}{n} \to R \quad P \times Q - \text{a.s.}$$

$$(ii') \qquad \frac{\log \tilde{T}_n(D) - nR}{\sqrt{n}} \xrightarrow{\mathcal{D}} N(0, \sigma^2)$$

$$(iii') \qquad \liminf_{n\to\infty} \frac{\log \tilde{T}_n(D) - nR}{\sqrt{2n \log_e \log_e n}} = -\sigma \quad P \times Q - \text{a.s.}$$

Lemma 4.1

Assume that $\boldsymbol{X}$ *is stationary ergodic,* $\boldsymbol{Y}$ *is an i.i.d. process, and* $D \in (D_{\min}, D_{\text{av}})$. *Then* $T_n(D) = \tilde{T}_n(D)$, *eventually* $P \times Q$*-almost surely.*

Proof of Lemma 4.1. Note that $T_n(D) \le \tilde{T}_n(D) \le W_n(D)$, and that whenever $T_{2n}(D) > W_n(D)$, we have $T_n(D) = \tilde{T}_n(D)$. Therefore, if we can show

$$\liminf_{n\to\infty} n^{-1} \log T_{2n}(D) \ge \frac{4R}{3} \quad P \times Q - \text{a.s.} \tag{4.19}$$

then, by Corollary 4.7, $T_n(D) = \tilde{T}_n(D)$ eventually $P \times Q$-almost surely.

For any $x_1^\infty \in A^\infty$, any positive integer m, and any n large enough, by the union bound and (4.16) we have

$$\begin{aligned} \mathbb{P}(T_{2n}(D) \le m \,|\, X_1^\infty = x_1^\infty) &\le \sum_{k \ge 2n} \mathbb{P}(W_k(D) \le m \,|\, X_1^k = x_1^k) \\ &\le m \sum_{k \ge 2n} Q(B(x_1^k, D)). \end{aligned} \tag{4.20}$$

Since, by Corollary 2.1, $\lim_{k\to\infty} k^{-1} \log Q(B(X_1^k, D)) = -R$, with $P-$probability one, we must have

$$\sup_{k \ge n} k^{-1} \log Q(B(x_1^k, D)) \le -3R/4 \quad \text{eventually } P\text{–a.s.}$$

Substituting this in (4.20) with $m = \exp(4Rn/3)$ gives

$$\mathbb{P}(T_{2n}(D) \le \exp(4Rn/3) \,|\, X_1^\infty = x_1^\infty) \le Ce^{-nR/6} \quad \text{eventually } P\text{–a.s.}$$

for some fixed $C < \infty$, and by the Borel-Cantelli lemma,

$$T_{2n}(D) > \exp(4Rn/3) \quad \text{eventually } P \times Q\text{–a.s.}$$

implying (4.19) and the conclusion of the Lemma. □

Proof of Theorem 4.2. As already stated, it suffices to prove (i') – (iii'). To this end, first observe that combining Theorem 4.1 and Theorem 2.4,

$$\liminf_{n\to\infty} \frac{1}{n} \min_{n\le k\le 2n} \left[\log W_k(D) - kR(\hat{P}_k)\right] \ge 0 \quad P\times Q - \text{a.s.} \tag{4.21}$$

and from Corollary 2.1 it follows that

$$\frac{1}{n} \min_{n\le k\le 2n} kR(\hat{P}_k) \to R \quad P\times Q - \text{a.s.} \tag{4.22}$$

By (4.21) and (4.22) we have

$$\frac{1}{n}\log \tilde{T}_n(D) \ge \frac{1}{n} \min_{n\le k\le 2n} \left[\log W_k(D) - kR(\hat{P}_k)\right] + \frac{1}{n} \min_{n\le k\le 2n} kR(\hat{P}_k) \to R,$$

with $P\times Q-$probability one. Since $\tilde{T}_n(D) \le W_n(D)$, the corresponding upper bound also holds by Corollary 4.7, proving (i').

Next let $\epsilon > 0$ arbitrary, so that in the notation of Corollary 4.7,

$$\mathbb{P}\left\{\frac{\log \tilde{T}_n(D)}{\sqrt{n}} - \frac{\log W_n(D)}{\sqrt{n}} < -\epsilon\right\} = \mathbb{P}\left\{\inf_{1\le t\le 2}\left[\frac{w(nt;D)}{\sigma\sqrt{n}} - \frac{w(n;D)}{\sigma\sqrt{n}} + \left(\frac{\lfloor nt\rfloor - n}{\sigma\sqrt{n}}\right)R\right] \le -\frac{\epsilon}{\sigma}\right\}.$$

For any $\delta > 0$ and n large enough this is bounded above by

$$\mathbb{P}\left\{\inf_{1\le t\le 1+\delta}\left[\frac{w(nt;D)}{\sigma\sqrt{n}} - \frac{w(n;D)}{\sigma\sqrt{n}}\right] \le -\frac{\epsilon}{\sigma}\right\} + \mathbb{P}\left\{\inf_{1+\delta\le t\le 2}\left[\frac{w(nt;D)}{\sigma\sqrt{n}} - \frac{w(n;D)}{\sigma\sqrt{n}}\right] \le -\frac{\epsilon}{\sigma} - K\sqrt{n}\right\}, \tag{4.23}$$

where $K = \delta R/(2\sigma)$. By the functional CLT of Corollary 4.8 (extended in the obvious way to $t \in [0,2]$), the first term of (4.23) converges to $\Pr\{\inf_{0\le t\le\delta} B_t \le -\epsilon/\sigma\}$ as $n\to\infty$, where $\{B_t\}$ is standard Brownian motion, and this can be made arbitrarily small by taking δ small enough. Similarly for any $C > 0$ the second term in (4.23) is asymptotically bounded above by $\Pr\{\inf_{0\le t\le 1} B_t \le -C\}$ which can also be made

arbitrarily small by taking C large enough. Combining these with the fact that $\tilde{T}_n(D) \le W_n(D)$ implies that $[\log \tilde{T}_n(D) - \log W_n(D)] = o(\sqrt{n})$ in probability, which, together with part (i) of Corollary 4.8, gives (ii').

We similarly obtain (iii') by applying the functional LIL instead of the functional CLT: Set $s_n = \sigma\sqrt{2n \log_e \log_e n}$ noting that,

$$\frac{\log \tilde{T}_n(D)}{s_n} - \frac{\log W_n(D)}{s_n} = \inf_{1 \le t \le 2} \left[\frac{w(nt;D)}{s_n} - \frac{w(n;D)}{s_n} + \left(\frac{\lfloor nt \rfloor - n}{s_n} \right) R \right].$$

For any $\delta > 0$ and n large enough this is bounded below by

$$\min \left\{ \inf_{1 \le t \le 1+\delta} \left[\frac{w(nt;D)}{s_n} - \frac{w(n;D)}{s_n} \right], \inf_{1+\delta \le t \le 2} \left[\frac{w(nt;D)}{s_n} - \frac{w(n;D)}{s_n} \right] + \frac{K\sigma\sqrt{2n}}{s_n} \right\}$$

By the functional LIL of Corollary 4.8 (extended in the obvious way to $t \in [0,2]$), the first term in the above minimum is asymptotically $P \times Q$-almost surely bounded below by

$$\inf_r \inf_{1 \le t \le 1+\delta} [r(t) - r(1)] \ge -\sqrt{\delta},$$

where the outermost infimum is taken over all absolutely continuous functions r with $\int_0^2 (dr/dt)^2 dt \le 1$ and $r(0) = 0$. Similarly, with $P \times Q-$probability one,

$$\liminf_{n \to \infty} \inf_{1+\delta \le t \le 2} \left[\frac{w(nt;D)}{s_n} - \frac{w(n;D)}{s_n} \right] \ge \inf_r \inf_{1+\delta \le t \le 2} [r(t) - r(1)] \ge -\sqrt{1-\delta},$$

so that the second term in the above minimum converges to $+\infty$ with probability one, and, hence,

$$\liminf_{n \to \infty} \frac{\log \tilde{T}_n(D)}{\sigma\sqrt{2n \log_e \log_e n}} - \frac{\log W_n(D)}{\sigma\sqrt{2n \log_e \log_e n}} \ge -\sqrt{\delta} \quad P \times Q - \text{a.s.}$$

Letting $\delta \downarrow 0$, recalling that $\tilde{T}_n(D) \le W_n(D)$ and applying part (ii) of Corollary 4.8 gives (iii') and completes the proof. □

Chapter 5

Efficient, Universal, Lossy Data Compression

In this chapter we bring together many of the ideas we encountered in the previous four chapters, in order to tackle an important practical problem in data compression – that of finding an efficient extension of the celebrated Lempel-Ziv algorithm to the case of *lossy* compression. As we will see, the waiting times results of Chapter 4 provide the key insight for our proposed solution.

In the next section we give a general introduction to the problem and we briefly review some of the relevant literature. In Section 5.2 we recall (from Sections 1.1.3 and 1.2.2) the connection between waiting times and Lempel-Ziv coding. We interpret the waiting times results of Chapter 4 in this framework and show that they can be extended (Theorem 5.1) to achieve optimal compression in the lossy case. This motivates us to introduce, in the following section, a new practical lossy compression algorithm. In Section 5.3 the algorithm is described in detail, and our main result of this chapter (Theorem 5.2) is stated, establishing its asymptotic optimality. The proof of Theorem 5.2 is given in Section 5.4, and it is based, in part, on Theorem 5.1. In Section 5.5 we discuss some implementation issues and present brief simulation results illustrating the performance of the algorithm. In Section 5.6 we describe extensions along several directions, and in Section 5.7 we give the proofs of the theoretical results from Sections 5.2 and 5.4.

5.1 Introduction: Data Compression

Over the past 25 years, the practical requirement for efficient data compression methods has become apparent in almost every engineering application where large amounts of data are transmitted or stored.

In applications where the data needs to be perfectly reconstructed from its compressed form (*lossless* coding), the most prominent example of a successful practical scheme is probably the Lempel-Ziv data compression algorithm: Some variation of the original algorithm [84][85] can be found on virtually every personal computer in use today. Although in terms of compression performance they have been shown to be asymptotically optimal and to achieve optimality universally over several general classes of data sources (i.e., without prior knowledge of the source) [84][82][85][70][53][71], their practical success is perhaps mainly due to the fact that they provide low-complexity algorithms that offer themselves to easy on-line implementations. (A comprehensive introduction to several lossless Lempel-Ziv schemes and their implementations is given in the recent text [30]; see also [9] for numerous variants.)

On the other hand, there are several applications in which the requirement for perfect reconstruction of the data can be relaxed (*lossy* coding), for example, when images are transmitted over the World Wide Web. In this case the story has been somewhat less successful. From rate-distortion theory [11] we know that one can achieve a sometimes dramatic improvement in compression by allowing some amount of error in the reconstructed data. In fact, it has been demonstrated that there exist universal algorithms for lossy data compression that asymptotically achieve optimal performance, and, moreover, there are explicit constructions of such universal codes; see [36], the references therein, and the more recent work of Zhang and Wei [80][81]. Typically, these constructions either involve exhaustive searches over the space of all possible codebooks (as, for example, in [83] and [52]), or are of exponential complexity at the encoder and therefore cannot be realistically implemented in practice (as in [50] and [74]). More practical algorithms have been recently proposed by Yang, Zhang and Berger [76], who suggest a way to circumvent the exponential encoding complexity of earlier schemes (party expanding on the ideas of Muramatsu and Kanaya [50]).

Motivated by the success of the lossless Lempel-Ziv schemes, several attempts were made to extend them to the case of lossy coding, most notably by Morita and

Kobayashi [49] and by Steinberg and Gutman [64]. Unfortunately, these schemes have strictly suboptimal compression performance, as we will see in Section 5.2.

In this chapter we propose a new extension of Lempel-Ziv coding to the lossy case: In Section 5.3 we present a universal algorithm for encoding memoryless sources at a fixed distortion level, which arises as a generalization of the Fixed-Database Lempel-Ziv (FDLZ) lossless compression algorithm [70]. In the following four sections we show that its compression performance is asymptotically optimal with respect to bounded single-letter distortion measures and argue that it is of reasonable encoding complexity: On the one hand, in its naive implementation this algorithm has complexity of the same order as the corresponding implementations of the lossless FDLZ, and, on the other hand, there is a wealth of efficient approximate string matching algorithms that allow more practical implementations (see [18][3][6][17] and the references therein).

In terms of its compression performance, a heuristic argument given in Section 5.5 suggests that the algorithm's redundancy rate is of the same order as the redundancy of its lossless counterpart (FDLZ), and we also present simulation results that agree well with this rate.

The main novelty of our approach is that, instead of doing the encoding with respect to a database generated by the same distribution as the data, the encoder is allowed to have *multiple databases* simultaneously available, and to adaptively choose which one to use at each step in a "greedy" way. As the database length grows, the number of available databases also grows so that, in effect, codebooks are generated according to all possible reproduction distributions. By controlling the rate at which the number of databases grows, we can make sure that reasonable complexity is maintained at the encoder while at the same time the set of possible codebook distributions is refined to cover an asymptotically dense set.

The reason why this algorithm compresses optimally is intuitively clear: We know from rate-distortion theory that, unlike in the case of lossless coding, when distortion is allowed, the optimal codebook distribution is typically different from the distribution of the source. The most straightforward way to fix this mismatch between a fixed database and the optimum one is to maintain multiple databases at the encoder and decoder so that a good enough match can always be found. In this way, two objectives are simultaneously achieved:

(i) Universality; the same algorithm with the same set of databases works for any

memoryless source.

(*ii*) Reasonable complexity; like FDLZ in the lossless case, what makes this algorithm attractive for applications is that it provides a sequence of suboptimal coding schemes, indexed by the database length and the number of available databases, that offers a handle on the complexity/redundancy trade-off: Using few, short databases, we get efficient, easily implementable algorithms, with high redundancy. On the other hand, increasing the length and the number of databases, provides algorithms whose compression performance can be made arbitrarily close to being optimal at the cost of increasing the encoding complexity.

As discussed in Chapter 1 (Sections 1.1.3 and 1.2.2), Wyner and Ziv [69] showed that several variants of Lempel-Ziv coding can be analyzed by studying an idealized coding scenario in terms of waiting times. This connection between data compression and waiting times has been exploited by a number of authors since then, including [70][71][64][66][68], among many others (it is also described in detail in [37]). In the next section we follow along the same path. We introduce an idealized coding scenario and interpret the waiting times results of Chapter 4 in that context. This interpretation suggests a natural generalization of the idealized coding scheme, corresponding to a new result about waiting times (Theorem 5.1). This result, in turn, motivates the new practical algorithm introduced in Section 5.3, and its optimality is established in Section 5.4 using Theorem 5.1.

It is worth noting here that the overall strategy for proving the algorithm's optimality is, by now, a familiar one: First, the waiting times of Section 5.2 are approximated by a sequence of large deviation probabilities (Lemma 5.1). Then the exponent of decay of these probabilities is identified using large deviations (Lemma 5.2), giving us the exponent of growth of the waiting times. Finally, using duality, this waiting times result is translated into a result about match lengths (Corollary 5.3), and this provides the main ingredient in the proof of Theorem 5.2.

Before moving on to the new results, a few words about some earlier work are in order here. The notion of using multiple codebooks for source coding is well-known in information theory, although multiple codebook algorithms typically involve a training stage or a large search over (essentially) all possible codebooks. For example, Chou, Effros and Gray's [14] vector-quantization interpretation of universal lossy

source codes is in terms of two-pass (or "two-stage") weighted universal codes. Another family of two-pass lossy compression algorithms is that of empirically designed vector quantizers, discussed by Linder, Lugosi and Zeger [43] among others. (More pointers to the large literature on vector quantization can be found in the recent review paper by Gray and Neuhoff [28].) Preliminary results from a work closer in spirit to our approach were recently reported by Zamir and Rose in [78][79].

5.2 Lempel-Ziv Coding and Waiting Times

The first extensions of the Lempel-Ziv algorithm to the lossy case [49][64] had suggested using a database of the same distribution as the source, but, as it was recently shown [45][75][21], these schemes generally achieve strictly suboptimal compression.

In this section we illustrate how their performance can be understood by studying an idealized coding scenario in terms of waiting times and show how this idealized scenario can be modified to achieve asymptotically optimal compression.

5.2.1 The Idealized Coding Scenario

Let $\boldsymbol{X} = \{X_n \ ; \ n \geq 1\}$ be a memoryless source with values in the source alphabet A, where A is a Polish (= complete, seperable, metric) space equipped with its Borel σ-field $\mathcal{A}$. The word "source" in this chapter is used interchangeably with the phrase "random process," and "memoryless" means that the distribution of $\boldsymbol{X}$ is determined by specifying that the random variables $\{X_n\}$ are independent and identically distributed (i.i.d.) according to some fixed measure p on $(A, \mathcal{A})$. As before, we write P for the measure on $(A^\infty, \mathcal{A}^\infty)$ describing the distribution of $\boldsymbol{X}$, so that here $P = p^\infty$. We write $\boldsymbol{x} = \{x_n \ ; \ n \geq 1\}$ for an infinite realization generated by $\boldsymbol{X}$, and we refer to $\boldsymbol{x}$ (or any subsequence of it) as a *message* produced by the *source* $\boldsymbol{X}$.

Suppose now that an encoder and a decoder both have available to them an infinite "database" $\boldsymbol{Y} = \{Y_n \ ; \ n \geq 1\}$ taking values in the finite set $\hat{A}$, the *reproduction* alphabet. We assume that $\boldsymbol{Y}$ is also memoryless, and its distribution Q is determined by the probability mass function (p.m.f.) q on $\hat{A}$.

The encoder's task is to describe the message X_1^n produced by $\boldsymbol{X}$ to the decoder, within some prescribed distortion D. As in Chapters 2 and 4, distortion here is

measured with respect to a sequence $\{\rho_n\}$ of *single-letter* distortion measures,

$$\rho_n(x_1^n, y_1^n) \stackrel{\triangle}{=} \frac{1}{n}\sum_{i=1}^{n} \rho(x_i, y_i), \qquad x_1^n \in A^n,\ y_1^n \in \hat{A}^n, \tag{5.1}$$

for some fixed $\rho : A \times \hat{A} \to [0, \infty)$.

In order to take advantage of the common information $\boldsymbol{Y}$, the encoder looks for the first position W in the database where X_1^n appears within distortion D, and communicates W to the decoder. Formally,

$$W_n^{(Q)}(D) \ = \ \inf\{k \geq 1 \ : \ Y_k^{k+n-1} \in B(X_1^n, D)\},$$

where, as before, $B(X_1^n, D)$ denotes the ball of distortion-radius D, centered at X_1^n:

$$B(X_1^n, D) = \{y_1^n \in \hat{A}^n \ : \ \rho_n(X_1^n, y_1^n) \leq D\}.$$

From this information the decoder can easily recover the string Y_W^{W+n-1}, which is guaranteed to be within distortion D of X_1^n.

Since [25][71] it takes approximately $\log W$ bits to describe an integer W, the rate of this code is, to first order

$$\approx \frac{\log W_n^{(Q)}(D)}{n} \quad \text{bits per symbol.}$$

As we saw in Chapter 4 (Corollary 4.7 and equation (4.6)), this ratio converges to $R(p, q, D)$ with probability one, where

$$R(p, q, D) \ = \ \inf \int H(\Theta(\cdot|x) \| q(\cdot)) dp(x) \tag{5.2}$$

$$= \ \inf\left[I(X;Y) + H(q' \| q)\right] \tag{5.3}$$

with the infimum taken over all random variables (X, Y) taking values in $A \times \hat{A}$, such that $X \sim p$, $E\rho(X, Y) \leq D$, Θ denoting the conditional distribution of Y given X, and q' denoting the marginal of Y.

How good is this rate? Recall that the best possible rate, the *rate-distortion*

function $R(D)$ of $\boldsymbol{X}$ with respect to $\{\rho_n\}$, is given by

$$R(D) \;=\; \inf\, I(X;Y) \tag{5.4}$$

where the infimum is taken over the same class of random variables (X,Y) as in the definition of $R(p,q,D)$ above. Comparing (5.3) with (5.4) it is immediately clear that the asymptotic rate $R(p,q,D)$ of this idealized code is generally *strictly* suboptimal. On the other hand, it is not hard to see that

$$R(D) \;=\; \inf_q\; R(p,q,D) \tag{5.5}$$

where the infimum is over all p.m.f.s q on $\hat{A}$, so that, intuitively, the problem is that we do not know *a priori* how to choose the "right" database distribution that achieves the infimum in (5.5).

The main idea behind the algorithm we will describe in Section 5.3, is to use *multiple* databases: In the waiting times framework this corresponds to generating one memoryless database for each n-type on $\hat{A}$, and encode using the "best" database, i.e., the one for which X_1^n has the shortest waiting time. The additional coding cost incurred is that of identifying which database was used, but since there are only polynomially many n-types, this extra cost is asymptotically negligible.

5.2.2 Waiting Times with Multiple Databases

Given an integer k, a p.m.f. q on $\hat{A}$ is called a *k-type*, if, for every $y \in \hat{A}$, $q(y)$ is of the form j/k for some nonnegative integer $j \le k$.

Let $\{s(n)\}$ be a nondecreasing sequence of positive integers. For each n, let $S(n)$ be the number of $s(n)$-types on $\hat{A}$ and write $q^{(j)}$, $1 \le j \le S(n)$, for each one of these $s(n)$-types. Assume that for each n we have $S(n)$ processes $\boldsymbol{Y}^{(j)}$, $1 \le j \le S(n)$, where $\boldsymbol{Y}^{(j)}$ is independent of $\boldsymbol{X}$ and distributed i.i.d. according to $q^{(j)}$. For each j let $W_n^{(j)}(D)$ be the waiting time until X_1^n appears in $\boldsymbol{Y}^{(j)}$ within distortion D,

$$W_n^{(j)}(D) \;=\; \inf\{i \ge 1 \,:\, (Y_i^{(j)}, Y_{i+1}^{(j)}, \ldots, Y_{i+n-1}^{(j)}) \in B(X_1^n, D)\},$$

and write $W_n^*(D)$ for the shortest one of these waiting times,

$$W_n^*(D) \;=\; \min_{1\le j\le S(n)} W_n^{(j)}(D).$$

Theorem 5.1 (Waiting Times)

Let $0 < D < \rho_{\max}$. If $s(n) = n$ for all n, we have

$$\limsup_{n\to\infty} \frac{\log W_n^*(D)}{n} \le R(D) \quad a.s.$$

Moreover, this remains true for any nondecreasing integer sequence $s(n) \to \infty$ as $n \to \infty$.

Before the proof of the theorem, we need to introduce some notation and definitions. First, let $R_e(D)$ denote the rate-distortion function of $\boldsymbol{X}$ in nats rather than bits, and similarly write $R_e(p, q, D)$ for the function defined as in (5.2) but with relative entropy in nats rather than in bits, i.e., with $H(\cdot\|\cdot)$ replaced by $H_e(\cdot\|\cdot)$. Equation (5.5) is equivalent to

$$R_e(D) \;=\; \inf_q \; R_e(p, q, D),$$

and we write q^* for the p.m.f. on $\hat{A}$ that achieves the infimum. [The fact that there does exist an achieving q^* is easy to see: Let $\{q_n\}$ be a sequence of p.m.f.s such that $R_e(p, q_n, D) \to R_e(D)$. Since the simplex of p.m.f.s on $\hat{A}$ is compact set (in the Euclidean topology induced by $\mathbb{R}^{|\hat{A}|}$), $\{q_n\}$ has a convergent subsequence $\{q_n'\}$ with $q_n' \to$ some q^*. But $R_e(p, q, D)$ is continuous in q (this follows easily from Proposition 2.2 and Lemma 2.1 of Section 2.3), and $\{q_n'\}$ is a subsequence of $\{q_n\}$ so we must have $R_e(D) = R_e(p, q^*, D)$.]

For each n sufficiently large, we can choose an $s(n)$-type q_n on $\hat{A}$ such that

$$|q_n(y) - q^*(y)| \le \frac{|\hat{A}|}{s(n)}, \quad \text{for all } y \in \hat{A}, \tag{5.6}$$

and $q_n(y) > 0$ for all $y \in \hat{A}$ (this is outlined in the Appendix). From now and until the end of this section we assume that n is large enough so that q_n can be chosen as above. Write $\widetilde{W}_n(D)$ for the waiting time until a D-close version of X_1^n appears in the $\boldsymbol{Y}$-process distributed according to q_n, and write $\mathbb{Q}^{(n)}$ for the product measure

$(q_n)^\infty$ on $(\hat{A}^\infty, \hat{\mathcal{A}}^\infty)$, where $\hat{\mathcal{A}}^\infty$ is the σ-field on $\hat{A}^\infty$ generated by finite-dimensional cylinders.

Proof of Theorem 5.1: Theorem 5.1 follows by combining Lemmas 5.1 and 5.2, below, together with the trivial observation that $W_n^*(D) \le \widetilde{W}_n(D)$ with probability one. Lemma 5.1 shows that, asymptotically, the waiting time $\widetilde{W}_n(D)$ for a D-close match of X_1^n into $\boldsymbol{Y} \sim \mathbb{Q}^{(n)}$ cannot be significantly larger than the reciprocal of the probability $\mathbb{Q}^{(n)}(B(X_1^n, D))$ of the event that such a match occurs. Its proof parallels those of the corresponding strong approximation theorems in Chapters 3 and 4 (Theorems 3.1 and 4.1).

Lemma 5.1 (Strong Approximation)

$$\limsup_{n\to\infty} \frac{1}{n} \log[\widetilde{W}_n(D)\mathbb{Q}^{(n)}(B(X_1^n, D))] \le 0 \quad a.s.$$

Lemma 5.2 is a large deviations result; it will follow by an application of the Gärtner-Ellis Theorem [22, Theorem 2.3.6].

Lemma 5.2 (Large Deviations)

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{Q}^{(n)}(B(X_1^n, D)) \ge -R(D) \quad a.s.$$

Lemmas 5.1 and 5.2 are proved in Sections 5.7.1 and 5.7.2, respectively. □

5.3 Description of the Algorithm

After some preliminary definitions, we describe the compression algorithm in its simplest form, and we state our first result, Theorem 5.2, which establishes its asymptotic optimality. The algorithm is a lossy source coding scheme for encoding memoryless sources at a fixed distortion level, with respect to single-letter distortion measures. Extensions of the use of the algorithm to more general situations are discussed in Section 5.6.

We follow the notation introduced in the previous section: Let $\boldsymbol{X} = \{X_n\ ;\ n \ge 1\}$ be a memoryless source with distribution $P = p^\infty$ on the the alphabet A, where A is a Polish space and $\mathcal{A}$ is its Borel σ-field. Also let $\hat{A}$ denote the reproduction

alphabet and assume that it is finite (the extension to general alphabets is discussed in Section 5.6). Distortion will be measured with respect to a sequence $\{\rho_n\}$ of single-letter distortion measures, defined as in (5.1), with respect to a fixed, nonnegative (measurable) function $\rho : A \times \hat{A} \to [0, \infty)$. Given $D \geq 0$ and a string $x_1^n \in A^n$, we write, as before, $B(x_1^n, D)$ for the the ρ_n-ball of radius D around x_1^n. Without loss of generality, throughout this chapter we assume, as usual [11], that

$$\sup_{x \in A} \min_{y \in \hat{A}} \rho(x, y) = 0 \tag{5.7}$$

and we also assume that ρ is bounded on the support of the source, i.e.,

$$M \stackrel{\triangle}{=} \max_{y \in \hat{A}} \operatorname*{ess\,sup}_{X \sim p} \rho(X, y) < \infty. \tag{5.8}$$

Define

$$\rho_{\max} = \inf_{q} \int \rho(x, y) dp(x) dq(y)$$

where the infimum is taken over all p.m.f.s q on $\hat{A}$, and assume that $\rho_{\max} > 0$. Finally, given $D \geq 0$, we write $R(D)$ for the rate-distortion function of $\boldsymbol{X}$ with respect to $\{\rho_n\}$, defined by (5.4). It is easy to check that $R(D) = 0$ for $D \geq \rho_{\max}$, so we restrict our attention to the interesting range of allowable distortion values $D \in (0, \rho_{\max})$.

5.3.1 The Algorithm

Let $X_1^N = (X_1, X_2, \ldots, X_N)$ be a message of length N generated by some memoryless source $\boldsymbol{X}$ of unknown distribution p on A, and let a distortion level $D \in (0, \rho_{\max})$ be fixed. Let $\{t(m)\}$ be a nondecreasing sequence of integers, write $T(m)$ for the number of $t(m)$-types on $\hat{A}$, and recall [19] that $T(m)$ is roughly polynomial in $t(m)$

$$T(m) \leq [t(m) + 1]^{|\hat{A}|}. \tag{5.9}$$

For each m, we describe an encoding algorithm that uses $T(m)$ databases of length m. So let us choose and fix an m for now. Assume that the encoder and decoder both have access to $T(m)$ memoryless databases

$$Y_1^{(1)},\ Y_2^{(1)},\ \ldots,\ Y_m^{(1)} \qquad \text{i.i.d. } \sim q^{(1)}$$

$$
\begin{aligned}
&Y_1^{(2)},\ Y_2^{(2)},\ \ldots,\ Y_m^{(2)} \qquad \text{i.i.d.} \sim q^{(2)} \\
&\vdots \\
&Y_1^{(T(m))},\ Y_2^{(T(m))},\ \ldots,\ Y_m^{(T(m))} \quad \text{i.i.d.} \sim q^{(T(m))}
\end{aligned}
$$

where each database has the same length m, they are all generated independently of the message X_1^N, and each one is i.i.d. according to a $t(m)$-type $q^{(j)}$ on $\hat{A}$, for $1 \leq j \leq T(m)$. Figure 1 shows schematically the set of all $t(m)$-types for the specific choice of $t(m) = \lceil \log m \rceil$.

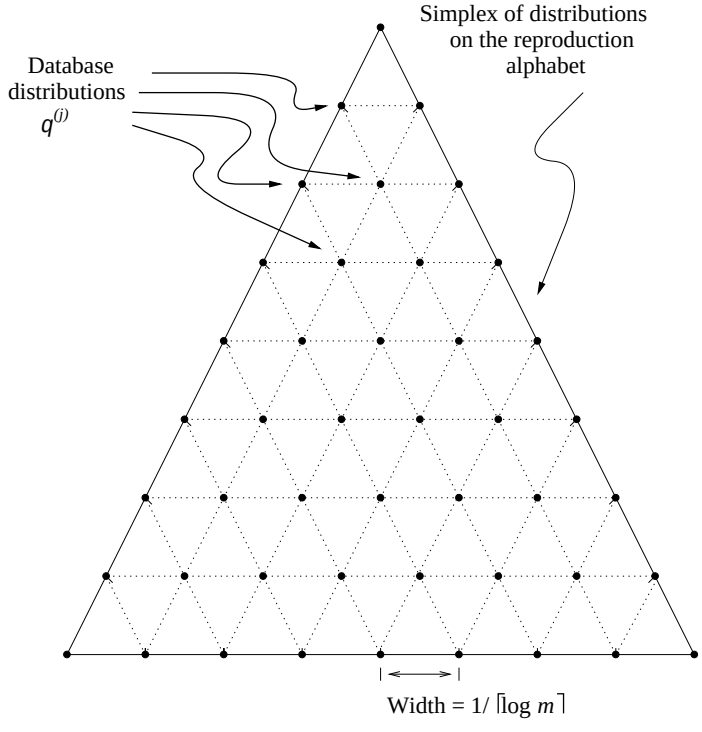

Figure 5.1: The set of all $\lceil \log m \rceil$-types, corresponding to the vertices of a uniform grid of width $1/\lceil \log m \rceil$ placed on the simplex of p.m.f.s on $\hat{A}$.

We can either assume that these databases are available to the encoder and decoder before the coding process begins, or that they are generated at the encoder and transmitted to the decoder using an overhead of

$$\lceil m\, T(m) \log |\hat{A}| \rceil \quad \text{bits.} \tag{5.10}$$

The encoding algorithm is as follows: First, the encoder calculates the length of the longest match of an initial portion of the message, within distortion D, in any one of the databases. Let $L_{m,1}(D)$ denote the length of this longest match,

$$L_{m,1}(D) = \max\{k \geq 1 : \rho_k((Y_i^{(j)}, \ldots, Y_{i+k-1}^{(j)}), X_1^k) \leq D \text{ for some } i \leq m-k+1,\ j \leq T(m)\}$$

and let $Z^{(1)}$ denote the initial phrase of length $L_{m,1}(D)$ in X_1^N :

$$Z^{(1)} \stackrel{\triangle}{=} (X_1, X_2, \ldots, X_{L_{m,1}(D)}).$$

Observe that $L_{m,1}(D) \geq 1$ by assumption (5.7). Then the encoder describes to the decoder:

(a) the length $L_{m,1}(D)$; this takes at most $C \log(L_{m,1}(D)+1)$ bits, where C is a constant (cf. [25][71]);

(b) the index j of the database in which this longest match was found; this takes $\lceil \log T(m) \rceil$ bits;

(c) the position i in database j where the match occurs; this takes $\lceil \log m \rceil$ bits.

Clearly, from (a), (b) and (c) the decoder can easily recover the string

$$\hat{Z}^{(1)} = (Y_i^{(j)}, Y_{i+1}^{(j)}, \ldots, Y_{i+L_{m,1}(D)-1}^{(j)}),$$

which is within distortion D of $Z^{(1)}$. The description length of (a), (b) and (c) is bounded above by

$$C \log(L_{m,1}(D)+1) + \log T(m) + \log m + 2 \quad \text{bits.} \tag{5.11}$$

Alternatively, $\hat{Z}^{(1)}$ can be described by first describing its length $L_{m,1}(D)$ as before, and then describing $\hat{Z}^{(1)}$ directly using

$$\lceil L_{m,1}(D) \log |\hat{A}| \rceil \quad \text{bits.} \tag{5.12}$$

The encoder uses whichever one of the two descriptions is shorter, together with a one-bit flag to indicate which one was chosen. Therefore, from (5.11), (5.12) and

(5.9), the length of the description of $Z^{(1)}$ is bounded above by

$$\min\{C_1 \log(L_{m,1}(D)+1) + C_2 \log(t(m)+1) + \log m,\ C_3 L_{m,1}(D)\} \quad \text{bits}, \quad (5.13)$$

for some fixed constants C_1, C_2, and C_3, independent of m, N, and of the message X_1^N.

After $Z^{(1)}$ has been described within distortion D, the same process is repeated to encode the rest of the message: The encoder finds the length $L_{m,2}(D)$ of the longest string starting at position $(L_{m,1}(D)+1)$ in X_1^N that matches within distortion D into any one of the databases, and describes

$$Z^{(2)} \stackrel{\triangle}{=} (X_{L_{m,1}(D)+1}, X_{L_{m,1}(D)+2}, \ldots, X_{L_{m,1}(D)+L_{m,2}(D)})$$

to the decoder by repeating the above steps.

The algorithm is terminated, in the natural way, when the entire string X_1^N has been exhausted. At that point, X_1^N has been parsed into $\Pi_m = \Pi_m(X_1^N, D)$ distinct phrases $Z^{(k)}$, each of length $L_{m,k}(D)$,

$$X_1^N \ = \ Z^{(1)} Z^{(2)} \cdots Z^{(\Pi_m)},$$

with the possible exception of the last phrase, which may be shorter. Since each substring $Z^{(k)}$ is described within distortion D, also the concatenation of all the reproduction strings,

$$\hat{Z}^{(1)} \hat{Z}^{(2)} \cdots \hat{Z}^{(\Pi_m)},$$

will be within distortion D of X_1^N.

Let $\ell_m(X_1^N) = \ell_m(X_1^N, D)$ denote the overall description length for X_1^N using this algorithm. From (5.10) and (5.13), $\ell_m(X_1^N, D)$ is bounded above by

$$\lceil m\, T(m) \log |\hat{A}| \rceil + \sum_{k=1}^{\Pi_m} \min\{C_1 \log(L_{m,k}(D)+1) + C_2 \log(t(m)+1) + \log m,\ C_3 L_{m,k}(D)\} \text{ bits}. \quad (5.14)$$

The following result establishes the asymptotic optimality of the algorithm by showing that, for long messages ($N \to \infty$), the expected compression ratio achieved does not

exceed the rate-distortion function $R(D)$, as m tends to infinity. In fact, a somewhat stronger result is proved, namely, that for (almost) any message emitted by the source, the compression ratio achieved, averaged over all possible databases, is asymptotically no larger than $R(D)$. Theorem 5.2 is proved in Section 5.7.4.

Theorem 5.2 (Algorithm Optimality)

Let $0 < D < \rho_{\max}$. If the rate at which the databases are refined is $t(m) = \lceil \log m \rceil$, we have

$$\limsup_{m\to\infty} \limsup_{N\to\infty} E\left\{ \frac{\ell_m(X_1^N, D)}{N} \,\middle|\, X_1^N \right\} \;\le\; R(D) \quad a.s., \tag{5.15}$$

and, therefore,

$$\limsup_{m\to\infty} \limsup_{N\to\infty} E\left\{ \frac{\ell_m(X_1^N, D)}{N} \right\} \;\le\; R(D). \tag{5.16}$$

Moreover, (5.15) and (5.16) remain valid for any choice of $t(m)$ such that $t(m) \to \infty$ while $(\log t(m))/\log m \to 0$, as $m \to \infty$.

Remark. The case of lossless compression can be regarded as a special case of the above algorithm, where the encoder looks for exact matches between the source and the database. In fact, implicit in the proof of Theorem 5.2 is a proof that the compression ratio achieved by the lossless FDLZ algorithm [70] applied to a memoryless source $\boldsymbol{X}$ converges to the entropy rate H of $\boldsymbol{X}$, for almost all source messages:

Corollary 5.1 (Strong Optimality of Lossless FDLZ)

Let $\boldsymbol{X}$ be a discrete memoryless source of entropy rate H, and let $\tilde{\ell}_m(X_1^N)$ denote the description length for X_1^N using the FDLZ algorithm. We have:

$$\limsup_{m\to\infty} \limsup_{N\to\infty} E\left\{ \frac{\tilde{\ell}_m(X_1^N)}{N} \,\middle|\, X_1^N \right\} \;\le\; H \quad a.s.$$

5.4 Algorithm Optimality

We use the waiting times results of Section 5.2 to prove Theorem 5.2, establishing the optimality of the algorithm.

First we observe that, as we already saw in Chapter 4, there is a duality relationship between waiting times and match lengths. Here, since we only need to prove an upper bound for the waiting times $W_n^*(D)$, it suffices to state this duality in one direction:

$$W_n^*(D) \leq m - n + 1 \;\Rightarrow\; L_{m,1}(D) \geq n. \tag{5.17}$$

Strictly speaking, since the definitions of $W_n^*(D)$ and $L_{m,1}(D)$ depend on the choices of the underlying sequences $\{s(i)\}$ and $\{t(j)\}$, respectively, we should say that: *If $W_n^*(D)$ defined with respect to a fixed sequence $\{s(i)\}$ satisfies $W_n^*(D) \leq m - n + 1$, then, $L_{m,1}(D)$ defined with respect to a different sequence $\{t(j)\}$ such that $s(n) = t(m)$ satisfies $L_{m,1}(D) \geq n$.*

Using (5.17) we can now easily translate the asymptotic upper bound for $W_n^*(D)$ of Theorem 5.1 to an asymptotic lower bound for $L_{m,1}(D)$:

Corollary 5.2 (Match Lengths)
Let $0 < D < \rho_{\max}$. If $t(m) \to \infty$ as $m \to \infty$, we have

$$\liminf_{m\to\infty} \frac{L_{m,1}(D)}{\log m} \geq \frac{1}{R(D)} \quad a.s.$$

The proof of Corollary 5.2 is a straightforward but tedious calculation, and therefore omitted here. The optimality of the algorithm (proof of Theorem 5.2 below) essentially follows from the fact that the match lengths grow like $(\log m)/R(D)$. This is similar, at least in spirit, to the lossless case, where the optimality of FDLZ follows from the fact that the lengths L_m of the longest exact matches grow like $(\log m)/H$. Unfortunately, the elegant combinatorial argument used by Wyner and Ziv in [69][71] no longer works when distortion is allowed. For that reason, in the proof of Theorem 5.2 we need a stronger bound on the (conditional) lower tails of $L_{m,1}(D)$.

Corollary 5.3 (Tails of Match Lengths)
Let $0 < D < \rho_{\max}$. If $t(m) \to \infty$ as $m \to \infty$, then for any $\epsilon > 0$ we have

$$(\log m)\,\Pr\left\{ L_{m,1}(D) \leq \frac{\log m}{R(D)+\epsilon} \;\middle|\; X_1^\infty \right\} \;\to\; 0 \quad a.s.$$

Corollary 5.3 is proved in Section 5.7.3.

5.5 Redundancy, Complexity, Implementation

Perhaps the most attractive feature of the algorithm is that it provides an active handle in balancing the trade-off of complexity vs. redundancy, depending on the requirements of particular applications. This trade-off is discussed in some detail below; a heuristic argument is presented suggesting that, if the rate at which that databases are being refined is chosen appropriately, then the redundancy of the algorithm is of the same order as that of the lossless FDLZ. (To be precise, "redundancy" here means the difference between the expected compression ratio achieved by the algorithm, and the entropy of the source being encoded.) This heuristic rate is also confirmed by brief simulation results presented in Section 5.5.2.

5.5.1 The Complexity-Redundancy Trade-off

There are three "terms" contributing to the redundancy of the algorithm, due to three different reasons:

(*i*) *Finite-length databases.* Since the databases used by the algorithm are finite, we expect that the compression will not be optimal even if we encode with respect to a database with the optimal distribution. As with FDLZ in the lossless case, we expect that the penalty incurred by using a database of finite length m will be of the order of

$$O\left(\frac{\log\log m}{\log m}\right).$$

The main ingredient in deriving this rate for FDLZ [73] is the fact that the expectations of the exact match lengths L_m grow like $(\log m)/H + O(1)$. We expect that the same behavior persists in the case when distortion is allowed, and that when only one database of distribution $Q = q^\infty$ is used, we have

$$EL_m(D) = \frac{\log m}{R(p,q,D)} + O(1), \quad \text{as } m \to \infty$$

(under some regularity assumptions on the distortion measure ρ). This should not come as surprise, particularly in view of the match length results of Chapter 4 where it is demonstrated that, in addition to their first-order behavior, all

of the second-order properties of $L_m(D)$ are exactly analogous to those obtained in the lossless case (compare Corollary 4.10 with Theorem 4.2 in Section 4.3).

(*ii*) *Several databases.* If the rate $t(m)$ at which the databases are refined is polynomial in $(\log m)$, then the coding cost of identifying which database was used is also of the order of

$$O\left(\frac{\log\log m}{\log m}\right).$$

This can be verified easily by reading through the proof of Theorem 5.2 in Section 5.7.4, and it is also intuitively clear since we use $O(\log\log m)$ bits to identify one of the databases each time we describe a string of length $O(\log m)$. In general, if $t(m)$ grows at a different rate, the contribution to the redundancy is of the order of $(\log t(m))/\log m$.

(*iii*) *Wrong database.* Finally, there is an error associated with the fact that for finite m the optimal database is (typically) not included among the databases currently available to the algorithm, so that the data is encoded with respect to a $(\log m)-$type approximation to the optimal database. In the idealized scenario of Section 5.2.1, this corresponds to comparing the exponent of $\mathbb{Q}^{(n)}(B(X_1^n, D))$ with that of $(q^*)^n(B(X_1^n, D))$, and (5.6) indicates that this difference should be $O(1)$ with probability one. Therefore, it is plausible to expect an additional redundancy term of order

$$O\left(\frac{1}{\log m}\right).$$

Combining (*i*) (*ii*) and (*iii*) suggests that the leading term in the redundancy of the algorithm is of the order of $(\log\log m)/\log m$, just like in the lossless case [73]. In particular, it should now be clear why the choice $t(m) = \lceil \log m \rceil$ was singled out in Theorem 5.2; because it makes the contribution of (*ii*) comparable to that of (*i*).

5.5.2 Implementation and Simulation Results

As stated in Theorem 5.2, the algorithm converges to optimality as long as the rate $t(m)$ at which the databases are refined tends to infinity, while $(\log t(m))/\log m$ tends to zero. More generally, from the proof of Theorem 5.2 it is clear that any

asymptotically dense set of database distributions will work, as long as the number $T(m)$ of available databases of length m does not grow too fast, namely, as long as $(\log T(m))/\log m$ tends to zero as $m \to \infty$. So, in practice, we have the freedom to choose any set of database distributions that fit the specific application better, instead of uniformly covering all possible distributions (as shown in Figure 1). For example, prior knowledge about the distribution of the source can easily be incorporated into the structure of the algorithm.

To illustrate its performance, we chose the simple example of lossy compression of a binary memoryless source with respect to Hamming distortion. We pseudo-randomly generated binary Bernoulli(0.4) data, and implemented the algorithm as described in Section 5.3.1, with some minor practical modifications (described below).

Figure 2 shows its compression performance on a sequence of 524288 bits, with the distortion level D set to 0.22, and for a total of 15 databases of lengths $m = 2^9, 2^{10}, \ldots, 2^{18}$ bits each. For reference, we note that typical values of m in current implementations of lossless versions of Lempel-Ziv are around $m = 2^{15}$ bits (for example, m corresponding to the window-size used by LZ77 as implemented in the Unix command `gzip`; see [30]).

As in several current implementations of lossless versions of Lempel-Ziv coding, we set a maximum possible match length of 128 bits. With this restriction we can describe the $L_{m,1}(D)$'s using a fixed 7 bits rather than the $C \log(L_{m,1}(D)+1)$ bits suggested in Section 5.3.1. We also mention that, although we did not go to great efforts in order to optimize the speed of our implementation, there is extensive literature devoted to approximate string matching algorithms: Implementation details and algorithmic issues relating to efficient, approximate string-matching are discussed in the text [18], and, in the context of data compression, in [3][6][17].

5.6 Extensions

A Fixed-Rate Version

We informally outline how the algorithm can be modified to provide fixed-rate lossy compression for memoryless sources. The main difference is that instead of looking for the *longest* match with distortion smaller than a fixed D, here we look for the *most accurate* match with length greater than some fixed length M.

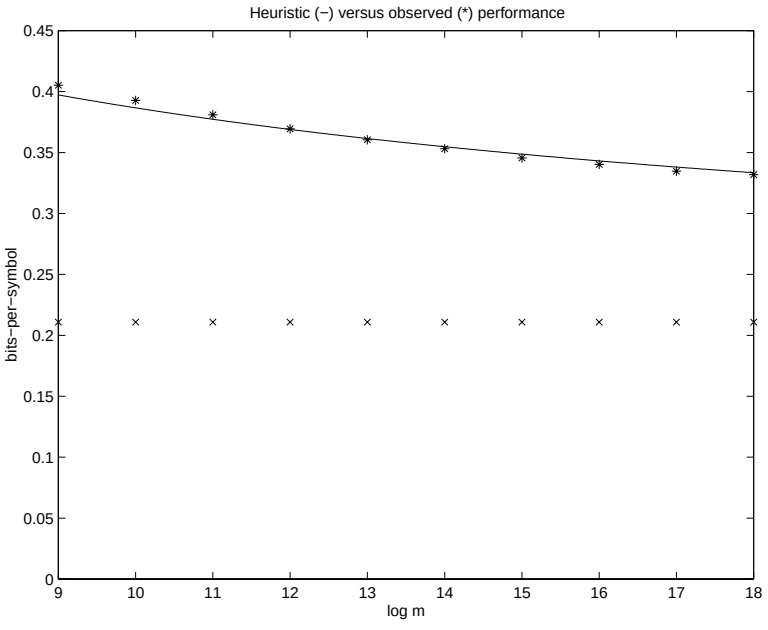

Figure 5.2: Compression performance on a memoryless Bernoulli(0.4) source, with respect to Hamming distortion and $D = 0.22$. The compression ratios achieved by the algorithm for different database sizes m are denoted by $(*)$; the ideal compression ratio (rate-distortion function) is shown as (x); the performance suggested by the heuristic argument in Section 5.5.1, namely, $R(D) + C(\log\log m)/\log m$, is shown as a solid line, with the constant $C \approx 0.53$ empirically fitted to the data.

Let R be the target rate, and recall from (5.13) that a string of length L in the message that matches somewhere in one of the databases, can be encoded using

$$\psi_m(L) \stackrel{\triangle}{=} \min\{C_1 \log(L+1) + C_2 \log(t(m)+1) + \log m,\ C_3 L\} \quad \text{bits.} \tag{5.18}$$

To guarantee an encoding rate below R bits per symbol, we consider initial strings X_1^L of the message X_1^N of lengths L large enough so that $\psi_m(L)/L \le R$, i.e., $L \ge M_m(R)$ where

$$M_m(R) \stackrel{\triangle}{=} \min\left\{1 \le L \le m \ :\ \frac{\psi_m(L)}{L} \le R\right\}$$

(since $\psi_m(L)/L$ is nonincreasing in L, having $L \ge M_m(R)$ implies $\psi_m(L)/L \le R$). Of all such strings X_1^L, choose the one that matches somewhere into one of the databases

with minimal distortion; let

$$D_{m,1}(R) \;=\; \min\{\rho_L(X_1^L, (Y_i^{(j)}, \ldots, Y_{i+L-1}^{(j)})) \;:\; M_m(R) \le L \le m,\; i \le m-L+1,\; j \le T(m)\},$$

and write $\Lambda_{m,1}(R)$ for the achieving L in the above definition. Then the initial string in X_1^N of length $\Lambda_{m,1}(R) \ge M_m(R)$ can be encoded, within distortion $D_{m,1}(R)$, using

$$\frac{\psi_m(\Lambda_{m,1}(R))}{\Lambda_{m,1}(R)} \le R \quad \text{bits pre symbol.} \tag{5.19}$$

The same process can be repeated iteratively until the entire message has been encoded, yielding a total of Π substrings of X_1^N, of lengths $a_i \stackrel{\triangle}{=} \Lambda_{m,i}(R)$, and corresponding description-lengths $b_i \stackrel{\triangle}{=} \psi_m(\Lambda_{m,i}(R))$. By (5.19) and the log-sum inequality [16, Theorem 2.7.1] it follows that

$$\log\left[\frac{\sum_{i=1}^{\Pi} a_i}{\sum_{i=1}^{\Pi} b_i}\right] \;\le\; \left(\sum_{i=1}^{\Pi} a_i\right)^{-1} \sum_{i=1}^{\Pi}\left(a_i \log\frac{a_i}{b_i}\right) \;\le\; \log R,$$

so the overall encoding rate of X_1^N is

$$\frac{\sum_{i=1}^{\Pi} a_i}{\sum_{i=1}^{\Pi} b_i} \;\le\; R \quad \text{bits per symbol.}$$

Now let us look at the distortion achieved. From the definition of ψ_m it is clear that the dominant term in the right hand side of (5.18) is the $(\log m)$-term, which means that, for large m, $\psi_m(L)/L \approx (\log m)/L$ and $M_m(R) \approx (\log m)/R$. Therefore, $D_{m,1}(R)$ is the minimal distortion that can be achieved between the source and any one of the databases by strings of lengths longer than $(\log m)/R$. But from Corollary 5.2 we know that there exist D-close matches of length at least $(\log m)/R(D)$, which suggests that

$$\limsup_{m\to\infty} D_{m,1}(R) \le D(R) \quad \text{a.s.,} \tag{5.20}$$

with $D(R)$ denoting the distortion-rate function of the source. So, in the same way that Corollary 5.2 is the essential ingredient in proving Theorem 5.2, it is plausible

that the optimality of the above scheme (i.e., that the overall description of the message X_1^N is asymptotically within distortion $D(R)$) will similarly follow from (5.20).

Sources with Memory

A simple inspection of the proofs immediately reveals that all the results from Sections 5.2, and 5.4 remain true in the case when the assumption that $\boldsymbol{X}$ is memoryless is replaced with the assumption that it is a stationary ergodic process. In particular, the asymptotic compression ratio achieved by the algorithm is equal to the first-order approximation to its rate-distortion function, which is, in general, larger than the rate-distortion function itself.

Unbounded Distortion Measures

The assumption that ρ is bounded is merely a technical assumption that can be significantly relaxed at the price of more complex proofs. We expect that the algorithm optimality, as well as the waiting times results of Section 5.2, remain valid for a much more general class of distortion measures, satisfying only some mild moment conditions.

General Reproduction Alphabets

As already mentioned in Section 5.5, the algorithm optimality does not depend on the exact form of the database-distributions chosen, as long as (1) they are asymptotically dense, and (2) their number $T(m)$ satisfies $(\log T(m))/\log m \to 0$ as $m \to \infty$. In the case of general reproduction alphabets, the algorithm can be extended in a straightforward way, by including several databases uniformly covering the space of all possible reproduction distributions. Such asymptotically dense finite covers should be possible to construct in a systematic manner, at least as long as the space of database distributions is "compact," in a natural sense.

5.7 Proofs

5.7.1 Proof of Lemma 5.1

Fix $D \in (0, \rho_{\max})$, write $\hat{P}_n$ for the empirical measure induced by X_1^n on A,

$$\hat{P}_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i},$$

and define

$$D_{\min}^{(n)} = E_{\hat{P}_n}[\min_{y \in \hat{A}} \rho(X, y)].$$

Recall that the $s(n)$-types q_n were chosen such that $q_n(y) > 0$ for all $y \in \hat{A}$, so that by (5.7) we have

$$D_{\min}^{(n)} \;=\; \frac{1}{n} \sum_{i=1}^{n} \min_{y \in \hat{A}} \rho(X_i, y) \;\leq\; \sup_{x \in A} \min_{y \in \hat{A}} \rho(x, y) \;=\; 0,$$

i.e., $D_{\min}^{(n)} = 0$ for all n. Therefore, $D > D_{\min}^{(n)}$, and so

$$\mathbb{Q}^{(n)}(B(X_1^n, D)) > 0 \quad P - \text{a.s.} \tag{5.21}$$

Let G_r denote the collection of all infinite realizations $\boldsymbol{x}$ from $\boldsymbol{X}$ that satisfy, for all $n \geq r$, $\mathbb{Q}^{(n)}(B(x_1^n, D)) > 0$, so that

$$P\left(\cup_r G_r\right) = 1 \tag{5.22}$$

by (5.21). Choose and fix any $r \geq 1$, pick $\boldsymbol{x} \in G_r$ and $\epsilon > 0$ arbitrary, and let $K \geq 1$ be a fixed constant. For any $n \geq r$ large enough so that $e^{n\epsilon} \geq 2(n+1)$, we have

$$\begin{aligned} P \times \mathbb{Q}^{(n)}\{\widetilde{W}_n(D) > K \,|X_1^n = x_1^n\} & \\ &\leq \mathbb{Q}^{(n)}\left\{Y_{in+1}^{(i+1)n} \notin B(x_1^n, D), \text{ for all } i = 0, 1, \ldots, \left\lfloor \frac{K-1}{n} \right\rfloor \right\} \\ &= \left[1 - \mathbb{Q}^{(n)}(B(x_1^n, D))\right]^{\lfloor \frac{K-1}{n} \rfloor} \end{aligned}$$

(since $\widetilde{W}_n(D) \geq 1$ by definition we need not consider values of $K < 1$). Letting $K = 2^{n\epsilon}/\mathbb{Q}^{(n)}(B(x_1^n, D))$ above, and noting that $(1-z)^R \leq 1/(Rz)$ for all $z \in (0,1)$ and $R > 0$, yields

$$\begin{aligned} P \times \mathbb{Q}^{(n)} \left\{ \frac{1}{n} \log[\widetilde{W}_n(D)\mathbb{Q}^{(n)}(B(x_1^n, D))] > \epsilon \;\middle|\; X_1^n = x_1^n \right\} & \\ \leq & \left[\mathbb{Q}^{(n)}(B(x_1^n, D)) \left\lfloor \frac{\frac{2^{n\epsilon}}{\mathbb{Q}^{(n)}(B(x_1^n, D))} - 1}{n} \right\rfloor \right]^{-1} \\ \leq & \left[\frac{2^{n\epsilon} - \mathbb{Q}^{(n)}(B(x_1^n, D))}{n} - \mathbb{Q}^{(n)}(B(x_1^n, D)) \right]^{-1} \\ \leq & \; 2n2^{-n\epsilon}. \end{aligned} \tag{5.23}$$

By the Borel-Cantelli Lemma it now follows that

$$\limsup_{n\to\infty} \frac{1}{n} \log[\widetilde{W}_n(D)\mathbb{Q}^{(n)}(x_1^n, D)] \leq 0 \quad \text{for } P \times \mathbb{Q}^{(n)}\text{-almost all } (\boldsymbol{x}, \boldsymbol{y}) \in G_r \times \hat{A}^\infty,$$

and combining this with (5.22) completes the proof. □

5.7.2 Proof of Lemma 5.2

To avoid cumbersome notation, we prove Lemma 5.2 in terms of natural logarithms instead of logarithms taken to base 2, i.e., we will show that

$$\liminf_{n\to\infty} \frac{1}{n} \log_e \mathbb{Q}^{(n)}(B(X_1^n, D)) \geq -R_e(D) \quad \text{a.s.} \tag{5.24}$$

Proof of Lemma 5.2: For all $x_1^n \in A^n$ and $\lambda \in \mathbb{R}$ define

$$\Lambda_{x_1^n}(\lambda) \;=\; \log_e \left\{ \int e^{\lambda \rho_n(x_1^n, y_1^n)} (q_n)^n (dy_1^n) \right\}$$

so that, by expanding ρ_n as a sum and using independence,

$$\frac{1}{n}\Lambda_{x_1^n}(\lambda n) \;=\; \frac{1}{n} \sum_{i=1}^{n} f_n(x_i),$$

where

$$f_n(x) \;=\; \log_e \left(\int e^{\lambda \rho(x,y)} dq_n(y) \right), \quad x \in A.$$

If we define $f(\cdot)$ on A like $f_n(\cdot)$, but with q_n replaced with q^*, then

$$\begin{aligned} \left| \frac{1}{n} \Lambda_{X_1^n}(\lambda n) - \frac{1}{n} \sum_{i=1}^{n} f(X_i) \right| &\leq \frac{1}{n} \sum_{i=1}^{n} |f_n(X_i) - f(X_i)| \\ &\leq \operatorname*{ess\,sup}_{X_1} |f_n(X_1) - f(X_1)| \\ &\leq \operatorname*{ess\,sup}_{X_1} |\log(1 + \epsilon_n(X_1))|, \end{aligned}$$

where

$$\epsilon_n(x) = \frac{\sum_{y \in \hat{A}} [q_n(y) - q^*(y)] e^{\lambda \rho(x,y)}}{\sum_{y \in \hat{A}} q^*(y) e^{\lambda \rho(x,y)}}.$$

But from (5.6) and (5.8),

$$|\epsilon_n(x)| \leq \frac{|\hat{A}|}{s(n)} e^{2|\lambda| M} \to 0 \quad \text{for } p\text{-almost all } x \in A,$$

which implies that

$$\left| \frac{1}{n} \Lambda_{X_1^n}(\lambda n) - \frac{1}{n} \sum_{i=1}^{n} f(X_i) \right| \to 0 \quad \text{a.s.} \tag{5.25}$$

Also, since ρ is bounded (by assumption), so is f, and by the ergodic theorem

$$\frac{1}{n} \sum_{i=1}^{n} f(X_i) \to E(f(X_1)) \;=\; \Lambda_{p,q^*}(\lambda) \quad \text{a.s.} \tag{5.26}$$

From (5.25) and (5.26) we get that

$$\frac{1}{n} \Lambda_{X_1^n}(\lambda n) \to \Lambda_{p,q^*}(\lambda) \quad \text{a.s.}$$

From this combined with Lemma 2.1 it follows that we can apply the Gärtner-Ellis Theorem [22, Theorem 2.3.6] along (almost) every realization of $\boldsymbol{X}$, to obtain that,

with P-probability one,

$$\begin{aligned}
&\liminf_{n\to\infty} \frac{1}{n} \log_e \mathbb{Q}^{(n)}(B(X_1^n, D)) \\
&\qquad = \liminf_{n\to\infty} \frac{1}{n} \log_e P \times \mathbb{Q}^{(n)} \left\{ \frac{1}{n} \sum_{i=1}^{n} \rho(X_i, Y_i^{(n)}) \le D \;\middle|\; X_1^n \right\} \\
&\qquad \ge -\inf_{z\in(0,D)} \Lambda^*_{p,q^*}(z) \qquad \text{(by the Gärtner-Ellis Theorem)} \\
&\qquad = -\Lambda^*_{p,q^*}(D) \qquad \text{(by Lemma 2.1)} \\
&\qquad = -R_e(p, q^*, D) \qquad \text{(by Proposition 2.2)} \\
&\qquad = -R_e(D), \qquad \text{(by the definition of } q^*)
\end{aligned}$$

and this proves (5.24) and the Lemma. □

5.7.3 Proof of Corollary 5.3

We follow the notation in the proofs of Theorem 5.1 and Lemma 5.1.

Let $\epsilon > 0$ be given. Pick one of the (almost all) realizations $\boldsymbol{x}$ of $\boldsymbol{X}$ such that $\boldsymbol{x} \in \cup_r G_r$, and also the result of Lemma 5.2 holds. By Lemma 5.2, we can choose N_0 (depending on $\boldsymbol{x}$) large enough so that

$$\frac{1}{n} \log \mathbb{Q}^{(n)}(B(x_1^n, D)) \;>\; -R(D) - \frac{\epsilon}{4} \quad \text{for all } n \ge N_0. \tag{5.27}$$

Then, by the duality relationship (5.17) and the fact that $W_n^*(D) \le \widetilde{W}_n(D)$,

$$\begin{aligned}
&\Pr\left\{ L_{m,1}(D) \le \frac{\log m}{R(D)+\epsilon} \;\middle|\; \boldsymbol{X} = \boldsymbol{x} \right\} \\
&\qquad \le P \times \mathbb{Q}^{(n)} \left\{ \widetilde{W}_n(D) \ge m - n + 1 \;\middle|\; \boldsymbol{X} = \boldsymbol{x} \right\} \\
&\qquad = P \times \mathbb{Q}^{(n)} \left\{ \frac{\log \widetilde{W}_n(D)}{n} \ge \frac{\log(m-n+1)}{n} \;\middle|\; \boldsymbol{X} = \boldsymbol{x} \right\}
\end{aligned}$$

where $n = \lceil (\log m)/(R(D)+\epsilon) \rceil$. If we take m large enough, say $m \ge M_0$, so that

$n \geq N_0$ and $[\log(m-n+1)]/n \geq R(D)+\epsilon/2$, then this is bounded above by

$$P\times\mathbb{Q}^{(n)}\left\{\frac{\log\widetilde{W}_n(D)}{n}\geq R(D)+\frac{\epsilon}{2}\;\Big|\;\boldsymbol{X}=\boldsymbol{x}\right\}\leq$$

$$P\times\mathbb{Q}^{(n)}\left\{\frac{\log[\widetilde{W}_n(D)\mathbb{Q}^{(n)}(B(x_1^n,D))]}{n}\geq R(D)+\frac{\log\mathbb{Q}^{(n)}(B(x_1^n,D))}{n}+\frac{\epsilon}{2}\Big|\boldsymbol{X}=\boldsymbol{x}\right\}$$

and by (5.27) this is bounded above by

$$P\times\mathbb{Q}^{(n)}\left\{\frac{\log[\widetilde{W}_n(D)\mathbb{Q}^{(n)}(B(x_1^n,D))]}{n}\geq\frac{\epsilon}{4}\;\Big|\;\boldsymbol{X}=\boldsymbol{x}\right\}.$$

Finally take $m\geq M_0$ sufficiently large to make the corresponding n large enough so that the bound (5.23) from the proof of Lemma 5.1 applies. Combining (5.23) with the above bounds yields

$$\Pr\left\{L_{m,1}(D)\leq\frac{\log m}{R(D)+\epsilon}\;\Big|\;\boldsymbol{X}=\boldsymbol{x}\right\}\leq 2n2^{-\epsilon n/4}\leq\alpha\,m^{-\beta}\log m,$$

for some fixed constants α, $\beta>0$; since this argument holds for P-almost any $\boldsymbol{x}$, the result of Corollary 5.3 follows. □

5.7.4 Proof of Theorem 5.2

Let $\epsilon>0$ be given, and choose and fix one of the (almost all) realizations $\boldsymbol{x}$ of $\boldsymbol{X}$ such that Corollary 5.3 holds. Recall that the encoding algorithm parses up X_1^N into Π_m distinct words $Z^{(k)}$, each of length $L_{m,k}(D)$. Let $n=(\log m)/(R(D)+\epsilon)$. Following [70], we assume, without loss of generality, that n is an integer, and that the last phrase in the parsing of X_1^N is complete, i.e.,

$$Z^{(\Pi_m)}\;\text{ has length }\;L_{m,\Pi_m}(D).$$

We call a phrase $Z^{(k)}$ *short* if its length satisfies $L_{m,k}(D)\leq n$; otherwise $Z^{(k)}$ is called *long*.

We break the upper bound (5.14) for the description length $\ell_m(X_1^N)$ into three

parts:

$$\ell_m(X_1^N) \;\leq\; \lceil m\,T(m)\log|\hat{A}|\rceil \;+\; C_3 \sum_{k:\,Z^{(k)}\text{ is short}} L_{m,k}(D) + \sum_{k:\,Z^{(k)}\text{ is long}} [C_1\log(L_{m,k}(D)+1) + C_2\log(t(m)+1) + \log m]. \quad (5.28)$$

The first term is non-random and independent of N, so that dividing by N and letting $N \to \infty$ it tends to zero. For the second term, after taking its conditional expectation, it can be bounded above as:

$$\begin{aligned} E\left\{ C_3 \sum_{k:\,Z^{(k)}\text{ is short}} L_{m,k}(D) \,\middle|\, X_1^N \right\} & \\ &\leq\; C_3\,\frac{\log m}{R(D)+\epsilon}\,E\left\{ \sum_{k:\,Z^{(k)}\text{ is short}} \mathbb{I}_{\{L_{m,k}(D)\leq n\}} \,\middle|\, X_1^N \right\} \\ &\leq\; C_4 \log m\, N\,\Pr\left\{ L_{m,1}(D) \leq \frac{\log m}{R(D)+\epsilon} \,\middle|\, X_1^N \right\}, \end{aligned}$$

where the first inequality follows from the definition of being "short," the constant $C_4 = C_3/(R(D)+\epsilon)$, $\mathbb{I}_F$ denotes the indicator function of the event F, and the second inequality follows by considering not just all k's, but all the possible positions on X_1^N where a short match can occur. We can now divide by N, let $N \to \infty$, and apply Corollary 5.3 to see that the conditional expectation of the second term in (5.28) also converges to zero, P-almost surely.

Finally, we analyze the third – and dominant – term in (5.28). By the assumptions of Theorem 5.2, for all m large enough (independently of N and X_1^N) we have

$$C_2\,\frac{\log t(m)}{\log m} < \epsilon. \quad (5.29)$$

From now and until the end of the proof we assume that m is large enough for (5.29) to hold. Also, let Π'_m be the number of long phrases $Z^{(k)}$. Since each long $Z^{(k)}$ has length $L_{m,k}(D) \geq n$, we must have

$$\Pi'_m n \leq N. \quad (5.30)$$

Now, as in the lossless case [70], we can bound above the third term in (5.28) by

$$C_1\Pi'_m \sum_{k:\, Z^{(k)} \text{ is long}} \left[\frac{1}{\Pi'_m}\log(L_{m,k}(D)+1)\right] \;+\; \Pi'_m\left(1+C_2\frac{\log t(m)}{\log m}\right)\log m,$$

which, applying Jensen's inequality and (5.29), is bounded above by

$$\begin{aligned}
&C_1\Pi'_m \log\left(\frac{1}{\Pi'_m}\sum_{k:\, Z^{(k)} \text{ is long}}(L_{m,k}(D)+1)\right) \;+\; \Pi'_m(1+\epsilon)\log m \\
&\qquad\overset{(a)}{\leq}\quad C_1\Pi'_m\log\left(1+\frac{N}{\Pi'_m}\right) \;+\; \Pi'_m(1+\epsilon)\log m \\
&\qquad\overset{(b)}{\leq}\quad C_1 N\frac{\Pi'_m}{N}\log\left(1+\frac{N}{\Pi'_m}\right) \;+\; \frac{N}{n}(1+\epsilon)\log m \\
&\qquad\overset{(c)}{\leq}\quad C_1 N\frac{1}{n}\log(1+n) \;+\; N(1+\epsilon)(R(D)+\epsilon) \\
&\qquad\overset{(d)}{=}\quad N\left[(R(D)+\epsilon)(1+\epsilon) \;+\; C_5\frac{\log\log m}{\log m}\right],
\end{aligned}$$

where (a) follows by the fact that the sum of the lengths of long phrases cannot exceed N; (b) follows from (5.30); (c) follows from (5.30) together with the fact that the function $x\log(1+1/x)$ is increasing for all $x > 0$; and (d) follows from the definition of n in terms of m, with $C_5 = 2C_1(R(D)+\epsilon)$. Combining this with the fact that the first two terms in (5.28) vanish, immediately yields

$$\limsup_{m\to\infty}\limsup_{N\to\infty} E\left\{\frac{\ell_m(X_1^N, D)}{N}\;\middle|\; X_1^N\right\} \;\leq\; (R(D)+\epsilon)(1+\epsilon) \quad \text{a.s.}$$

and since $\epsilon > 0$ was arbitrary we get (5.15). Finally, (5.16) follows from (5.15) and Fatou's lemma. □

Chapter 6

Concluding Remarks

We summarize the main contributions of this thesis in Section 6.1, and in Section 6.2 we briefly discuss some promising directions along which the results presented in Chapters 2–5 may be extended.

6.1 Summary of Contributions

The two main contributions of this thesis are *(i)* the strong approximation framework for analyzing the asymptotic behavior of recurrence and waiting times (Chapters 2–4); and *(ii)* the new lossy version of the Lempel-Ziv algorithm presented in Chapter 5.

Strong approximation

In general terms, we can think of recurrence and waiting times as hitting times for certain rare events. For example, given a random pattern $(X_1, X_2, \ldots, X_n)$ generated by some process $\boldsymbol{X}$, the waiting time $W_n(D)$ is the time until the first occurrence of the rare event that a D-close version of $(X_1, X_2, \ldots, X_n)$ appears in a realization of a different process $\boldsymbol{Y}$. The approach taken in Chapters 3 and 4 can be summarized by saying that these waiting times (or recurrence times) can be approximated by the reciprocal of the probability of the rare event at hand. In the above example, $W_n(D)$ can be approximated by the reciprocal of the probability $Q(B(X_1^n, D))$ of the event that a D-close match for $(X_1, X_2, \ldots, X_n)$ occurs in $\boldsymbol{Y}$. More precisely, in

Theorem 4.1 we showed that, with probability one, the difference

$$\log W_n(D) - \log\left[\frac{1}{Q(B(X_1^n, D))}\right]$$

does not grow faster than $O(\log n)$ as $n \to \infty$, and, therefore, the asymptotic behavior of the waiting times $W_n(D)$ can be deduced from that of the probabilities $Q(B(X_1^n, D))$, which were studied extensively in Chapter 2 using techniques from large deviations.

This strategy of first approximating waiting times (or recurrence times) by appropriate large deviation probabilities and then using large deviations to determine exactly how these probabilities decay, provides a natural unified framework for deducing various asymptotic results: In Chapters 3 and 4 it allowed us to prove a series of strong new results, and it also allowed us to recover most of the known results in this area.

The strong approximation idea was introduced in [39] in the context of recurrence and waiting times without distortion. There, asymptotic results were proved by utilizing the Shannon-McMillan-Breiman theorem and its classical refinements by Yushkevich [77], Ibragimov [32], and Philipp and Stout [57]. In [21] it was extended to the case of waiting times allowing distortion, and in [41] the same strategy was employed to prove the waiting times results of Chapter 5 that led to establishing the optimality of a new lossy data compression algorithm:

Lossy Lempel-Ziv coding

In Chapter 5 we proposed a solution to a long-standing open problem in data compression: We introduced a new lossy version of the Lempel-Ziv data compression algorithm, for encoding memoryless sources at a fixed distortion level. This algorithm is easily implementable in practice – preliminary simulation results were presented in Section 5.5.2 demonstrating its performance on binary data. We also proved (Theorem 5.2) that its compression is asymptotically optimal with respect to single-letter distortion measures. This was done by first studying an idealized coding scenario in terms of waiting times and then using the corresponding waiting times results to prove the optimality of the practical scheme.

6.2 Extensions and Future Directions

Theory

There are several natural questions to ask about the theoretical results presented in Chapters 2–4, and they point to several directions for generalizations, three of which are mentioned below.

First, we note that most of the results from Chapters 2 and 4 can be extended to the general case of weakly dependent processes. The main tools will again be provided by the theory of large deviations and uniform pointwise approximation, but the characterization of the limiting rate function $R(P, Q, D)$ will be in terms of an infinite-dimensional variational problem.

Second, the strong approximation approach of Chapters 3 and 4 naturally extends to random fields on the integer lattice (as well as to several, more general group actions), although new subtleties arise in this case regarding the conditional structure of the measures and their mixing rates.

Finally, it is interesting to ask if there are simple analogs of the results in Chapters 2, 3 and 4 in the case of continuous-time processes. We expect that for reasonably rich classes of stationary ergodic processes (such as "nice" classes of exponentially mixing diffusions), there will be natural counterparts to most of the results in Chapters 2–4.

Applications

In the case of lossless compression, the classical refinements to the Shannon-McMillan-Breiman theorem were used in [38] to prove second-order lossless source coding theorems. Similarly, we expect that the corresponding results in the case when distortion is allowed (Corollaries 2.1, 2.2 and 2.3) can be used to prove second-order refinements to Shannon's *lossy* source coding theorem.

In terms of practical data compression, it is important to determine how successful the algorithm presented in Chapter 5 can be when applied to real data. In Chapter 5 we showed that this algorithm has several desirable theoretical properties, and currently we are testing to see how effectively it can be put to practical use. The limits of its applicability will essentially be determined by how efficiently we can implement the string-matching part of the algorithm.

As a specific application, we are interested in seeing how this scheme can be combined with existing methods (such as transform coding) to yield efficient *image* compression. Preliminary results in this direction seem promising.

In closing, we mention that we are currently in the process of providing [40] a modification of the Fixed-Database Lempel-Ziv algorithm, different from the one presented in Chapter 5, which achieves optimal compression for a wide class of processes with memory.

Appendix A

Some Technical Points

A.1 Proof of Theorem 2.3

Here we prove the following strengthened version of Theorem 2.3: $\sigma^2 = 0$ if and only if all the nonzero transition probabilities from state a to state b are of the form $2^{-H} v_a / v_b$, for some positive constants v_a, $a \in A$. Theorem 2.3 follows from this with $q = 2^{-H}$.

We begin by deriving a generalization of a formula due to Fréchet [26] for the asymptotic variance of Markov chains. Let $\boldsymbol{Z} = \{Z_n \ ; \ n \in \mathbb{Z}\}$ be a stationary irreducible aperiodic Markov chain with finite state-space T, stationary distribution $(q_i)_{i \in T}$, and kth order transition probabilities $(q_{ij}^{(k)})_{i,j \in T}$. Let f be a real-valued function on T and write $\bar{f}(\cdot)$ for $f(\cdot) - Ef(X_1)$. Define

$$\begin{aligned} \Sigma^2 = \lim_{n \to \infty} \frac{1}{n} \operatorname{Var} \left(\sum_{i=1}^{n} \bar{f}(Z_i) \right) &= E(\bar{f}(Z_1))^2 + 2 \sum_{k=1}^{\infty} E(\bar{f}(Z_1) \bar{f}(Z_{k+1})) \\ &= \sum_{j \in T} \bar{f}(j)^2 q_j + 2 \sum_{k=1}^{\infty} \sum_{i,j \in T} q_i q_{ij}^{(k)} \bar{f}(i) \bar{f}(j). \end{aligned} \tag{A.1}$$

Letting $s_{ij} = \sum_{k=1}^{\infty} \left[q_{ij}^{(k)} - q_j \right] < \infty$ (for $i, j \in T$), the second term above becomes

$$2 \sum_{i,j} q_i s_{ij} \bar{f}(i) \bar{f}(j) = 2 \sum_{i} q_i \bar{f}(i) \theta_i,$$

where $\theta_i = \sum_j s_{ij}\bar{f}(j)$ (for $j \in T$), and substituting this in (A.1) gives

$$\Sigma^2 \;=\; \sum_i q_i \left[\bar{f}(i) + \theta_i\right]^2 - \sum_i q_i \theta_i^2 \;=\; \sum_j q_j \left[\sum_i q_{ji}(\bar{f}(i) + \theta_i)^2 - \theta_j^2\right]. \qquad \text{(A.2)}$$

Expanding,

$$\begin{aligned}
\sum_i q_{ji}\theta_i &= \sum_i q_{ji} \sum_m s_{im}\bar{f}(m) \\
&= \sum_m \bar{f}(m) \sum_i q_{ji} \sum_{k\geq 1} (q_{im}^{(k)} - q_m) \\
&= \sum_m \bar{f}(m) \sum_{k\geq 1} (q_{jm}^{(k+1)} - q_m) \\
&= \sum_m \bar{f}(m) \left[\sum_{k\geq 1} (q_{jm}^{(k)} - q_m) - (q_{jm} - q_m)\right] \\
&= \sum_m s_{jm}\bar{f}(m) - \sum_m q_{jm}\bar{f}(m) \\
&= \theta_j - \sum_m q_{jm}\bar{f}(m), \qquad \text{(A.3)}
\end{aligned}$$

so that

$$\begin{aligned}
\sum_i q_{ji}(\bar{f}(i) + \theta_i)^2 &= \sum_i q_{ji}[(\bar{f}(i) + \theta_i - \theta_j) + \theta_j]^2 \\
&= \sum_i q_{ji}[(\bar{f}(i) + \theta_i - \theta_j)^2 + \theta_j^2], \qquad \text{(A.4)}
\end{aligned}$$

since by (A.3) the cross terms vanish:

$$\begin{aligned}
\sum_i q_{ji} 2\theta_j(\bar{f}(i) + \theta_i - \theta_j) &= 2\theta_j \left(\sum_i q_{ji}\bar{f}(i) - \theta_j + \sum_i q_{ji}\theta_i\right) \\
&= 2\theta_j \left(\sum_i q_{ji}\bar{f}(i) - \theta_j + \theta_j - \sum_m q_{jm}\bar{f}(m)\right) \;=\; 0.
\end{aligned}$$

Substituting (A.4) into (A.2) and interchanging i and j yields

$$\Sigma^2 = \sum_j q_j \sum_i q_{ji}(\bar{f}(i) + \theta_i - \theta_j)^2, \qquad \text{(A.5)}$$

which is the generalization of Fréchet's formula for the variance.

Now consider the chain $\tilde{\boldsymbol{X}}$ defined in Section 2.1. For all $a, b \in A$ we write $p_a = P(X_1 = a)$ and $p_{ab} = P(X_1 = b \,|\, X_0 = a)$, so that $\tilde{\boldsymbol{X}}$ has stationary distribution $(q_{ab}) = (p_a p_{ab})$ and transition probabilities $(q_{ab,cd}) = (\delta_{bc} p_{cd})$. Let f be defined as in Section 2.1. Since here $\theta_{ab} = \theta_b$ is independent of a, using (A.5) we get

$$\begin{aligned} \sigma^2 &= \sum_{(a,b)\in T} p_a p_{ab} \sum_{(c,d)\in T} \delta_{bc} p_{cd} (\bar{f}(c,d) + \theta_{cd} - \theta_{ab})^2 \\ &= \sum_{(a,b)\in T} p_a p_{ab} \sum_{d\in A: p_{bd}>0} p_{bd} (\bar{f}(b,d) + \theta_d - \theta_b)^2. \end{aligned}$$

For any $(b,d) \in T$ we have $p_{bd} > 0$, so $\sigma^2 = 0$ if and only if

$$\bar{f}(b,d) = \theta_b - \theta_b, \quad \text{for all } (b,d) \in T,$$

and the result stated in the beginning of this section follows upon setting $v_a \stackrel{\triangle}{=} 2^{-\theta_a}$, $a \in A$. □

A.2 Proof of Lemma 2.1

Part (i) follows immediately from the definitions of $\Lambda_{\mu,\nu}$ and $D^{\mu,\nu}_{\max}$.

For part (ii): First, since all the random variables involved in the definition of $\Lambda_{\mu,\nu}$ are bounded, its differentiability with respect to λ can be checked easily using the dominated convergence theorem. In particular, we can differentiate under the integral sign to obtain

$$\Lambda'_{\mu,\nu}(\lambda) = \int \left[\frac{\int \rho(x,y) e^{\lambda\rho(x,y)} d\nu(y)}{\int e^{\lambda\rho(x,z)} d\nu(z)} \right] d\mu(x)$$

and for $\lambda = 0$ this gives $\Lambda'_{\mu,\nu}(0) = D^{\mu,\nu}_{\text{av}}$. Differentiating once more,

$$\Lambda''_{\mu,\nu}(\lambda) = \int \left[\frac{\int \rho^2(x,y) e^{\lambda\rho(x,y)} d\nu(y) \int e^{\lambda\rho(x,y)} d\nu(y) - \left(\int \rho(x,y) e^{\lambda\rho(x,y)} d\nu(y) \right)^2}{\left(\int e^{\lambda\rho(x,z)} d\nu(z) \right)^2} \right] d\mu(x)$$

and this is easily seen to be nonnegative for any λ by applying Hölder's inequality to the numerator of the integrand. Moreover, since we assume $D^{\mu,\nu}_{\min} < D^{\mu,\nu}_{\text{av}}$, $\rho(x,y)$

is not almost surely constant, and the above expression is strictly positive. Next we outline a standard calculation which shows that $\Lambda'_{\mu,\nu}(\lambda) \to D_{\min}^{\mu,\nu}$ as $\lambda \to -\infty$. For any fixed $x \in A$ let $\alpha_x = \operatorname{ess\,inf}_{Y\sim\nu} \rho(x,Y)$, and, given $\epsilon > 0$ arbitrary, define

$$B_x(\epsilon) \;=\; \{y \in \hat{A} \;:\; \rho(x,y) \le \alpha_x + \epsilon\},$$

so that $\nu(B_x(\epsilon)) > 0$. Then,

$$\begin{aligned}\Lambda'_{\mu,\nu}(\lambda) = \int d\mu(x) \int_{B_x(\epsilon)} d\nu(y) \frac{\rho(x,y)e^{\lambda\rho(x,y)}}{\int e^{\lambda\rho(x,z)} d\nu(z)} \\ + \int d\mu(x) \int_{\hat{A}-B_x(\epsilon)} d\nu(y) \frac{\rho(x,y)e^{\lambda\rho(x,y)}}{\int e^{\lambda\rho(x,z)} d\nu(z)},\end{aligned} \tag{A.6}$$

where the integral over $(\hat{A} - B_x(\epsilon))$ is bounded above by

$$D_{\max}\left[\frac{\int_{\hat{A}-B_x(\epsilon)} d\nu(y) e^{\lambda(\rho(x,y)-\alpha_x)}}{\int_{B_x(\epsilon/2)} d\nu(y) e^{\lambda(\rho(x,y)-\alpha_x)}}\right] \le D_{\max}\left[\frac{\nu(\hat{A} - B_x(\epsilon))e^{\lambda\epsilon}}{\nu(B_x(\epsilon/2))e^{\lambda\epsilon/2}}\right] \le Ce^{\lambda\epsilon/2} \tag{A.7}$$

where C is a nonnegative constant, independent of λ. Therefore, by the dominated convergence theorem the second term in (A.6) converges to 0 as $\lambda \to -\infty$. Similarly, the integral over $B_x(\epsilon)$ in (A.6) is easily seen to be bounded below by

$$\begin{aligned}\frac{\int_{B_x(\epsilon/2)} d\nu(y)\rho(x,y)e^{\lambda\rho(x,y)}}{\int_{B_x(\epsilon/2)} d\nu(y)e^{\lambda\rho(x,y)} + \int_{\hat{A}-B_x(\epsilon)} d\nu(y)e^{\lambda\rho(x,y)}} &= \alpha_x\left[1 + \frac{\int_{\hat{A}-B_x(\epsilon)} d\nu(y)e^{\lambda\rho(x,y)}}{\int_{B_x(\epsilon/2)} d\nu(y)e^{\lambda\rho(x,y)}}\right]^{-1} \\ &\ge \alpha_x\left[1 + \frac{\nu(\hat{A} - B_x(\epsilon))e^{\lambda\epsilon}}{\nu(B_x(\epsilon/2))e^{\lambda\epsilon/2}}\right]^{-1},\end{aligned}$$

which is seen to converge to α_x as $\lambda \to -\infty$. Observing that that the integral over $B_x(\epsilon)$ in (A.6) bounded above by $(\alpha_x+\epsilon)$, combining this with the above lower bound, and letting $\epsilon \downarrow 0$, implies that the integral over $B_x(\epsilon)$ in (A.6) converges to α_x. This, in turn, together with (A.6), (A.7), and the definition of α_x, shows that $\Lambda'_{\mu,\nu}(\lambda) \to D_{\min}^{\mu,\nu}$ as $\lambda \to -\infty$.

Part (*iii*) is a straightforward application of (*ii*) and elementary calculus.

For part (iv), the infinite differentiability of

$$\Lambda_{\delta_x,\nu}(\lambda) \;=\; \log_e\left(\int e^{\lambda\rho(x,y)}d\nu(y)\right)$$

with respect to λ is established by a standard application of the dominated convergence theorem and induction, and the boundedness of the derivatives follows easily from the boundedness of ρ. □

A.3 Proof of Proposition 2.2

It suffices to show that $R_e(\mu,\nu,D) = \Lambda^*_{\mu,\nu}(D)$, i.e.,

$$\inf\int H_e(\Theta(\cdot|x)\|\nu(\cdot))d\mu(x) \;=\; \sup_{\lambda\in\mathbb{R}}[\lambda x - \Lambda_{\mu,\nu}(\lambda)] \tag{A.8}$$

where the infimum is taken over all probability measures Θ on $A\times\hat{A}$ such that the A-marginal of Θ is μ and $\int\rho(x,y)\,d\Theta(x,y)\le D$.

By Lemma 2.1 we may fix $\lambda<0$ for which the supremum on the right side of (A.8) is achieved. Consider the probability measure Θ defined by

$$\frac{d\Theta(x,y)}{d\mu\times\nu} \;=\; \frac{e^{\lambda\rho(x,y)}}{\int e^{\lambda\rho(x,z)}d\nu(z)}$$

in the left side of (A.8). The A-marginal of Θ is μ, $\int\rho(x,y)d\Theta(x,y) = \Lambda'_{\mu,\nu}(\lambda) = D$, and

$$\int H_e(\Theta(\cdot|x)\|\nu(\cdot))d\mu(x) \;=\; \lambda D - \int\log_e\left[\int e^{\lambda\rho(x,y)}d\nu(y)\right]d\mu(x) \;=\; \Lambda^*_{\mu,\nu}(D),$$

so the left side of (A.8) is no greater than $\Lambda^*_{\mu,\nu}(D)$. To prove the reverse inequality we recall that for any probability measure Θ and any bounded measurable function $\phi:\hat{A}\to\mathbb{R}$,

$$H_e(\Theta(\cdot|x)\|\nu(\cdot)) \ge \int\phi(y)d\Theta(y|x) - \log_e\left\{\int e^{\phi(y)}d\nu(y)\right\}$$

(c.f. [23, Lemma 3.2.13]). In particular, choosing $\phi(\cdot)=\lambda\rho(x,\cdot)$ and integrating both

sides with respect to μ yields the required inequality and completes the proof. □

A.4 Choice of $s(n)$-types

Since $s(n) \to \infty$ and it is nondecreasing, for all n large enough we have

$$s(n) > |\hat{A}| \max\{1/q^*(y) \ : \ y \in \hat{A} \text{ with } q^*(y) > 0\}.$$

Then pick $y_o \in \hat{A}$ with $q^*(y_o) > 0$, and define

$$q_n(y) = \begin{cases} \frac{\lceil s(n)q^*(y) \rceil}{s(n)} & \text{if } y \neq y_o \text{ and } q^*(y) > 0 \\ \frac{1}{s(n)} & \text{if } q^*(y) = 0 \\ 1 - \sum_{y \in \hat{A},\, y \neq y_o} q_n(y) & \text{if } y = y_o. \end{cases}$$

It is now trivial to check that q_n has the required properties.

Bibliography

[1] D. Aldous and P.C. Shields. A diffusion limit for a class of randomly-growing binary trees. *Prob. Th. Rel. Fields*, 79:509–542, 1988.

[2] P.H. Algoet and T.M. Cover. A sandwich proof of the Shannon-Mcmillan-Breiman theorem. *Ann. Probab.*, 16:876–898, 1988.

[3] D. Arnaud and W. Szpankowski. Pattern matching image compression with prediction loop: Preliminary experimental results. Technical Report CSD-TR-96-069, Department of Computer Sciences, Purdue University, 1996.

[4] R. Arratia, L. Gordon, and M.S. Waterman. The Erdös-Rényi law in distribution for coin tossing and sequence matching. *Ann. Stat.*, 18:539–570, 1990.

[5] R. Arratia and M.S. Waterman. A phase transition for the score in matching random sequences allowing deletions. *Ann. Appl. Probab.*, 4:200–225, 1994.

[6] M. Atallah, Y. Génin, and W. Szpankowski. Pattern matching image compression: Algorithmic and empirical results. Technical Report CSD-TR-95-083, Department of Computer Sciences, Purdue University, 1995.

[7] A.R. Barron. *Logically Smooth Density Estimation.* PhD thesis, Dept. of Electrical Engineering, Stanford University, 1985.

[8] A.R. Barron. The strong ergodic theorem for densities: generalized Shannon-Mcmillan-Breiman theorem. *Ann. Probab.*, 13:1292–1303, 1985.

[9] J.G. Bell, T.C. Cleary and I.H. Witten. *Text Compression.* Prentice Hall, New Jersey, 1990.

[10] R. Bellman and Harris T.E. Recurrence times for the Ehrenfest model. *Pacific J. Math.*, 1:179–193, 1951.

[11] T. Berger. *Rate Distortion Theory: A Mathematical Basis for Data Compression.* Prentice-Hall Inc., Englewood Cliffs, NJ, 1971.

[12] B. C. Bradley. Basic properties of strong mixing conditions. In E. Eberlein and M.S. Taqqu, editors, *Dependence in Probability and Statistics*, pages 165–192, 1986.

[13] L. Breiman. The individual ergodic theorem for information theory. *Ann. Math. Stat.*, 28:809–811, 1957. (See also the correction: *Ann. Math. Stat.*, Vol. 31, pp. 809-810, 1960).

[14] P.A. Chou, M. Effros, and R.M. Gray. A vector quantization approach to universal noiseless coding and quantizations. *IEEE Trans. Inform. Theory*, 42(4):1109–1138, 1975.

[15] K.L. Chung. A note on the ergodic theorem of information theory. *Ann. Math. Stat.*, 32:612–614, 1961.

[16] T.M. Cover and J.A. Thomas. *Elements of Information Theory.* J. Wiley, New York, 1991.

[17] M. Crochemore and T. Lecroq. Pattern-matching and text-compression algorithms. *ACM Computing Surveys*, 28(1):39–41, 1996.

[18] M. Crochemore and W. Rytter. *Text Algorithms.* Oxford University Press, New York, 1994.

[19] I. Csiszár and J. Körner. *Information Theory: Coding Theorems for Discrete Memoryless Systems.* Academic Press, New York, 1981.

[20] A. Dembo, S. Karlin, and O. Zeitouni. Critical pehonomena for sequence matching with scoring. *Ann. Probab.*, 22(4):1993–2021, 1994.

[21] A. Dembo and I. Kontoyiannis. The asymptotics of waiting times between stationary processes, allowing distortion. *NSF Technical Report no. 96, Dept. of Statistics, Stanford University; submitted for publication*, June 1997.

[22] A. Dembo and O. Zeitouni. *Large Deviations Techniques And Applications.* Jones and Bartlett, London, 1992. (Second edition, 1998).

[23] J.D. Deuschel and D.W. Stroock. *Large Deviations.* Academic Press, Boston, 1989.

[24] W. Doeblin. Remarques sur la théorie métrique des fractions continues. *Composition Math.*, 7:353–371, 1940. (French).

[25] P. Elias. Universal codeword sets and representations of the integers. *IEEE Trans. Inform. Theory*, 21:194–203, 1975.

[26] M. Fréchet. *Reserches théoriques modernes sur le calcul des probabilités; II; Théorie des événements en chaîne dans le cas d'un nombre fini d'états possibles.* Gauthiers-Villars, Paris, 1938. (French).

[27] A. Galves and B. Schmitt. Inequalities for hitting times in mixing dynamical systems. *Random Comput. Dynam.*, 5(4):337–347, 1997.

[28] R.M. Gray and D.L. Neuhoff. Quantization. *To appear, IEEE Trans. Inform. Theory*, 1998.

[29] L. Guibas and A.M. Odlyzko. Periods in strings. *J. Combin. Theory Ser. A*, 31:19–42, 1981.

[30] D. Hankerson, G.A. Harris, and P.D. Johnson, Jr. *Introduction to Information Theory and Data Compression.* CRC Press LLC, 1998.

[31] T.E. Harris. First passage and recurrence distributions. *Trans. Amer. Math. Soc.*, 73:471–486, 1952.

[32] I.A. Ibragimov. Some limit theorems for stationary processes. *Theory Prob. Appl.*, 7:349–382, 1962.

[33] P. Jacquet and W. Szpankowski. Autocorrelation of words and its applications. *J. Combin. Theory Ser. A*, 66:237–269, 1994.

[34] M. Kac. On the notion of recurrence in discrete stochastic processes. *Bull. Amer. Math. Soc.*, 53:1002–1010, 1947.

[35] S. Karlin and F. Ost. Maximal length of common words among random letter sequences. *Ann. Probab.*, 16:535–563, 1988.

[36] J.C. Kieffer. A survey of the theory of source coding with a fidelity criterion. *IEEE Trans. Inform. Theory*, 39(5):1473–1490, 1993.

[37] I. Kontoyiannis. Second-order analysis of lossless and lossy versions of Lempel-Ziv codes. In *31st Asilomar Conference on Signals, Systems and Computers, Monterey, CA*, November 1997.

[38] I. Kontoyiannis. Second-order noisless source coding theorems. *IEEE Trans. Inform. Theory*, 43(4):1339–1341, July 1997.

[39] I. Kontoyiannis. Asymptotic recurrence and waiting times for stationary processes. *To appear, J. Theoret. Probab.*, 1998.

[40] I. Kontoyiannis. An implementable lossy version of the Lempel-Ziv algorithm – Part II: Optimality for sources with memory. *In preparation*, 1998.

[41] I. Kontoyiannis. An implementable lossy version of the Lempel-Ziv algorithm that is asymptotically optimal – Part I: Memoryless sources. *NSF Technical Report no. 99, Dept. of Statistics, Stanford University; submitted for publication*, April 1998.

[42] I. Kontoyiannis, P.H. Algoet, Yu.M. Suhov, and A.J. Wyner. Nonparametric entropy estimation for stationary processes and random fields, with applications to English text. *IEEE Trans. Inform. Theory*, 44(3):1319–1327, May 1998.

[43] T. Linder, G. Lugosi, and K. Zeger. Rates of convergence in the source coding theorem, in empirical quantizer design, and in universal lossy source coding. *IEEE Trans. Inform. Theory*, 40(6):1728–1740, 1994.

[44] G. Louchard and W. Szpankowski. Average profile and limiting distribution for a phrase size in the Lempel-Ziv parsing algorithms. *IEEE Trans. Inform. Theory*, 41(2):478–488, 1995.

[45] T. Łuczak and W. Szpankowski. A suboptimal lossy data compression algorithm based on approximate pattern matching. *IEEE Trans. Inform. Theory*, 43(5):1439–1451, 1997.

[46] K. Marton and P.C. Shields. Almost sure waiting time results for weak and very weak Bernoulli processes. *Ergod. Th. & Dynam. Sys.*, 15:951–960, 1995.

[47] B. McMillan. The basic theorems of information theory. *Ann. Math. Stat.*, 24:196–219, 1953.

[48] S.P. Meyn and R.L. Tweedie. *Markov Chains and Stochastic Stability.* Springer-Verlag, London, 1993.

[49] H. Morita and K. Kobayashi. An extension of LZW coding algorithm to source coding subject to a fidelity criterion. In *4th Joint Sweedinsh-Soviet Int. Workshop on Inform. Theory*, pages 105–109, Gotland, Sweeden, 1989.

[50] J. Muramatsu and F. Kanaya. Distortion-complexity and rate-distortion function. *IEICE Trans. Fundamentals*, E77-A:1224–1229, 1994.

[51] A. Nobel and A.D. Wyner. A recurrence theorem for dependent processes with applications to data compression. *IEEE Trans. Inform. Theory*, 38(5):1561–1564, 1992.

[52] D. Ornstein and P.C. Shields. Universal almost sure data compression. *Ann. Probab.*, 18:441–452, 1990.

[53] D. Ornstein and B. Weiss. Entropy and data compression schemes. *IEEE Trans. Inform. Theory*, 39(1):78–83, 1993.

[54] M. Peligrad. Recent advances in the central limit theorem and its weak invariance principle for mixing sequences of random variables (a survey). In E. Eberlein and M.S. Taqqu, editors, *Dependence in Probability and Statistics*, pages 193–223, 1986.

[55] K. Petersen. *Ergodic Theory.* Cambridge University Press, Cambridge, 1983.

[56] P. Pevzner, M. Borodovsky, and A. Mironov. Linguistic of nucleotide sequences: the significance of deviations from mean statistical characteristics and prediction of the frequency of occurrence of words. *J. Biomol. Struct. Dynam.*, 6:1013–1026, 1991.

[57] W. Philipp and W. Stout. *Almost Sure Invariance Principles for Partial Sums of Weakly Dependent Random Variables.* Memoirs of the AMS, 1975. vol. 2, issue 2, no. 161.

[58] B. Pittel. Asymptotical growth of a class of random trees. *Ann. Probab.*, 13:414–427, 1985.

[59] H. Poincaré. *Les Méthodes Nouvelles de la Mécanique Céleste, III.* Gauthiers-Villars, Paris, 1899. (French).

[60] R.R. Rao. Relations between weak and uniform convergence of measures with applications. *Ann. Math. Stat.*, 33:659–680, 1962.

[61] E. Rio. The functional law of the iterated logarithm for stationary strongly mixing sequences. *Ann. Probab.*, 23:1188–1203, 1995.

[62] C.E. Shannon. A mathematical theory of communication. *Bell System Technical J.*, 27:379–423, 623–656, 1948. Reprinted in C.E. Shannon and W. Weaver, *The Mathematical Theory of Communication*, Univ. of Illinois Press, Urbana, Ill., 1949.

[63] P.C. Shields. Waiting times: Positive and negative results on the Wyner-Ziv problem. *Journal of Theortical Probability*, 6(3):499–519, 1993.

[64] Y. Steinberg and M. Gutman. An algorithm for source coding subject to a fidelity criterion, based upon string matching. *IEEE Trans. Inform. Theory*, 39(3):877–886, 1993.

[65] V. Strassen. An almost sure invariance principle for the law of the iterated logarithm. *Z. Wahrsch. Verw. Gabiete*, 3:23–32, 1964.

[66] W. Szpankowski. Asymptotic properties of data compression and suffix trees. *IEEE Trans. Inform. Theory*, 39(5):1647–1659, 1993.

[67] F.M.J. Willems. Universal data compression and repetition times. *IEEE Trans. Inform. Theory*, 35(1):54–58, 1989.

[68] A.D. Wyner and A.J. Wyner. Improved redundancy of a version of the Lempel-Ziv algorithm. *IEEE Trans. Inform. Theory*, 35(3):723–731, 1995.

[69] A.D. Wyner and J. Ziv. Some asymptotic properties of the entropy of a stationary ergodic data source with applications to data compression. *IEEE Trans. Inform. Theory*, 35(6):1250–1258, 1989.

[70] A.D. Wyner and J. Ziv. Fixed data base version of the Lempel-Ziv data compression algorithm. *IEEE Trans. Inform. Theory*, 37(3):878–880, 1991.

[71] A.D. Wyner and J. Ziv. The sliding-window Lempel-Ziv algorithm is asymptotically optimal. *Proc. IEEE*, 82(6):872–877, 1994.

[72] A.J. Wyner. *String Matching Theorems and Applications to Data Compression and Statistics.* PhD thesis, Dept. of Statistics, Stanford University, 1993.

[73] A.J. Wyner. The redundancy and distribution of the phrase lengths of the Fixed-Database Lempel-Ziv algorithm. *IEEE Trans. Inform. Theory*, 43(5):1452–1464, 1997.

[74] E.-h. Yang and J.C. Kieffer. Simple universal lossy data data compression schemes derived from the Lempel-Ziv algorithm. *IEEE Trans. Inform. Theory*, 42(1):239–245, 1996.

[75] E.-h. Yang and J.C. Kieffer. On the performance of data compression algorithms based upon string matching. *IEEE Trans. Inform. Theory*, 44(1):47–65, 1998.

[76] E.-h. Yang, Z. Zhang, and T. Berger. Fixed-slope universal lossy data compression. *IEEE Trans. Inform. Theory*, 43(5):1465–1476, 1997.

[77] A.A. Yushkevich. On limit theorems connected with the concept of the entropy of Markov chains. *Uspehi Mat. Nauk*, 8:177–180, 1953. (Russian).

[78] R. Zamir and K. Rose. Towards lossy Lempel-Ziv: Natural type selection. In *Proc. of the Inform. Theory Workshop*, page 58, Haifa, Israel, June 1996.

[79] R. Zamir and K. Rose. A type generator model for adaptive lossy compression. In *Proc. of the IEEE Internation Symposium on Inform. Theory*, page 186, Ulm, Germany, June/July 1997.

[80] Z. Zhang and V.K. Wei. An on-line universal lossy data compression algorithm by continuous codebook refinement – Part I: Basic reults. *IEEE Trans. Inform. Theory*, 42(3):803–821, 1996.

[81] Z. Zhang and V.K. Wei. An on-line universal lossy data compression algorithm by continuous codebook refinement – Part II: Optimality for phi-mixing models. *IEEE Trans. Inform. Theory*, 42(3):822–836, 1996.

[82] J. Ziv. Coding theorems for individual sequences. *IEEE Trans. Inform. Theory*, 24(4):405–412, 1978.

[83] J. Ziv. Distortion-rate theory for individual sequences. *IEEE Trans. Inform. Theory*, 26(2):137–143, 1980.

[84] J. Ziv and A. Lempel. A universal algorithm for sequential data compression. *IEEE Trans. Inform. Theory*, 23(3):337–343, 1977.

[85] J. Ziv and A. Lempel. Compression of individual sequences by variable rate coding. *IEEE Trans. Inform. Theory*, 24(5):530–536, 1978.