Source Coding, Large Deviations, and Approximate Pattern Matching

Amir Dembo and Ioannis Kontoyiannis, *Member, IEEE*

Invited Paper

Dedicated to the memory of Aaron Wyner, a valued friend and colleague.

***Abstract*—In this review paper, we present a development of parts of rate-distortion theory and pattern-matching algorithms for lossy data compression, centered around a lossy version of the asymptotic equipartition property (AEP). This treatment closely parallels the corresponding development in lossless compression, a point of view that was advanced in an important paper of Wyner and Ziv in 1989. In the lossless case, we review how the AEP underlies the analysis of the Lempel–Ziv algorithm by viewing it as a random code and reducing it to the idealized Shannon code. This also provides information about the redundancy of the Lempel–Ziv algorithm and about the asymptotic behavior of several relevant quantities. In the lossy case, we give various versions of the statement of the generalized AEP and we outline the general methodology of its proof via large deviations. Its relationship with Barron and Orey's generalized AEP is also discussed. The lossy AEP is applied to i) prove strengthened versions of Shannon's direct source-coding theorem and universal coding theorems; ii) characterize the performance of "mismatched" codebooks in lossy data compression; iii) analyze the performance of pattern-matching algorithms for lossy compression (including Lempel–Ziv schemes); and iv) determine the first-order asymptotic of waiting times between stationary processes. A refinement to the lossy AEP is then presented, and it is used to i) prove second-order (direct and converse) lossy source-coding theorems, including universal coding theorems; ii) characterize which sources are quantitatively easier to compress; iii) determine the second-order asymptotic of waiting times between stationary processes; and iv) determine the precise asymptotic behavior of longest match-lengths between stationary processes. Finally, we discuss extensions of the above framework and results to random fields.**

***Index Terms*—Data compression, large deviations, pattern-matching, rate-distortion theory.**

I. Introduction

A. Lossless Data Compression

It is probably only a slight exaggeration to say that the central piece of mathematics in the proof of almost any lossless coding theorem is provided by the asymptotic equipartition property (AEP). Suppose we want to (losslessly) compress a message $X_1^n = (X_1, X_2, \ldots, X_n)$ generated by a stationary memoryless source $\boldsymbol{X} = \{X_n; n \geq 1\}$ where each X_i takes values in the finite alphabet A (much more general situations will be considered later). For this source, the AEP states that as $n \to \infty$

$$-\frac{1}{n}\log_2 P^n(X_1^n) \to H \qquad \text{in probability} \tag{1}$$

where P is the common distribution of the independent and identically distributed (i.i.d.) random variables X_i, P^n denotes the (product) joint distribution of X_1^n, and

$$H = E[-\log_2 P(X_1)]$$

is the entropy rate of the source—see Shannon's original paper [74, Theorem 3] or Cover and Thomas' text [24, Ch. 4]. [Here, and throughout the paper, $\log_2$ denotes the logarithm taken to base 2, and $\log$ denotes the natural logarithm.] From (1), we can immediately extract some useful information. It implies that when n is large, the message X_1^n will most likely have probability at least as high as $2^{-n(H+\epsilon)}$

$$P^n(X_1^n) \geq 2^{-n(H+\epsilon)} \qquad \text{with high probability.} \tag{2}$$

But there cannot be many high-probability messages. In fact, there can be at most $2^{n(H+\epsilon)}$ messages with

$$P^n(X_1^n) \geq 2^{-n(H+\epsilon)}$$

so we need approximately 2^{nH} representative messages from the source $\boldsymbol{X}$ in order to cover our bets (with high probability). If we let $\mathcal{T}_n$ be the set of high-probability strings $x_1^n \in A^n$ having $P^n(x_1^n) \geq 2^{-n(H+\epsilon)}$, then we can correctly represent the source output X_1^n by an element of $\mathcal{T}_n$ (with high probability). Since there are no more than $2^{n(H+\epsilon)}$ of them, we need no more than nH bits to encode X_1^n.

Shannon's Random Code: Another way to extract information from (1) is as follows. The fact that for large n we typically have $P^n(X_1^n) \approx 2^{-nH}$ also means that if we independently generate another random string, say Y_1^n, from the same distribution as the source, the probability that X_1^n is the same as Y_1^n

Manuscript received February 28, 2001; revised November 21, 2001. The work of A. Dembo was supported in part by NSF under Grant DMS-0072331. The work of I. Kontoyiannis was supported in part by NSF under Grants 0073378-CCR and DMS-9615444.

A. Dembo is with the Departments of Mathematics and Statistics, Stanford University, Stanford, CA 94305 USA (e-mail: amir@stat.stanford.edu).

I. Kontoyiannis is with the Division of Applied Mathematics and the Department of Computer Science, Brown University, Providence, RI 02912 USA (e-mail: yiannis@dam.brown.edu).

Communicated by S. Shamai, Guest Editor.

Publisher Item Identifier S 0018-9448(02)04010-5.

is about 2^{-nH}. Suppose that instead of using the strings in $\mathcal{T}_n$ above as our representatives for the source, we decided to independently generate a collection of random strings Y_1^n from the distribution P^n; how many would we need? Given a source string X_1^n, the probability that any one of the Y_1^n matches it is $\approx 2^{-nH}$, so in order to have high probability of success in representing X_1^n without error we should choose approximately $2^{n(H+\epsilon)}$ random strings Y_1^n. Therefore, whether we choose the set of representatives systematically or randomly, we always need about 2^{nH} strings in order to be able to encode X_1^n losslessly with high probability. Note that the randomly generated set $\mathcal{T}_n$ is nothing but Shannon's random codebook [75] specialized to the case of lossless compression.

Idealized Lempel–Ziv Coding: In 1989, in a very influential paper [84], Wyner and Ziv took the above argument several steps further. Aiming to "obtain insight into the workings of [...] the Lempel–Ziv data compression algorithm," they considered the following coding scenario. Suppose that an encoder and a decoder both have available to them a long database, say an infinitely long string $Y_1^\infty = (Y_1, Y_2, \ldots)$ that is independently generated from the same distribution as the source. Given a source string X_1^n to be transmitted, the encoder looks for the first appearance of X_1^n in the database (assuming, for now, that it does appear somewhere). Let W denote the position of this first appearance, that is, let W be the smallest integer for which $Y_W^{W+n-1} = (Y_W, Y_{W+1}, \ldots, Y_{W+n-1})$ is equal to X_1^n. Then all the encoder has to do is it to tell the decoder the value of W; the decoder can read off the string Y_W^{W+n-1} and recover X_1^n perfectly. This description can be given using (cf. [32], [86]) no more than

$$\ell(X_1^n) = \log_2 W + O(\log_2 \log_2 W) \text{ bits.} \tag{3}$$

How good is this scheme? First note that, for any given source string X_1^n, the random variable W records the first "success" in a sequence of trials ("Is $Y_1^n = X_1^n$?," "Is $Y_2^{n+1} = X_1^n$?," and so on), each of which has probability of success $p = P^n(X_1^n)$. Although these trials are not independent, for large n they are almost independent (in a sense that will be made precise below), so the distribution of W is close to a geometric with parameter $p = P^n(X_1^n)$. For long strings X_1^n, p is small and W is typically close to its expected value, which is approximately equal to the mean of a geometric random variable with parameter p, namely, $1/p$. But the AEP tells us that, when n is large, $p = P^n(X_1^n) \approx 2^{-nH}$, so we expect W to be typically around 2^{nH}. Hence, from (3), the description length $\ell(X_1^n)$ of X_1^n will be, to first order

$$\ell(X_1^n) \approx -\log_2 P^n(X_1^n) \approx nH \text{ bits, with high probability.}$$

This shows that the above scheme is asymptotically optimal, in that its limiting compression ratio is equal to the entropy.[1]

Practical Lempel–Ziv Coding: The Lempel–Ziv algorithm [98], [99] and its many variants (see, e.g., [7, Ch. 8]) are some of the most successful data compression algorithms used in practice. Roughly speaking, the main idea behind these algorithms is to use the message's own past as a database for future encoding. Instead of looking for the first match in an infinitely long database, in practice, the encoder looks for the longest match in a database of fixed length. The analysis in [84] of the idealized scheme described above was the first step in providing a probabilistic justification for the optimality of the actual practical algorithms. Subsequently, in [85] and [86], Wyner and Ziv established the asymptotic optimality of the sliding-window (SWLZ) and the fixed-database (FDLZ) versions of the algorithm.

[1]We should also mention that around the same time a similar connection between data compression and waiting times was made by Willems in [81].

B. Lossy Data Compression

A similar development to the one outlined above can be given in the case of lossy data compression, this time centered around a lossy analog of the AEP [52]; see also [60], [88]. To motivate this discussion we look at Shannon's original random coding proof of the (direct) lossy source-coding theorem [75].

Shannon's Random Code: Suppose we want to describe the output X_1^n of a memoryless source, with distortion D or less with respect to a family of single-letter distortion measures $\{\rho_n\}$. Let Q_n^* be the optimum reproduction distribution on $\hat{A}^n$, where $\hat{A}$ is the reproduction alphabet. Shannon's random coding argument says that we should construct a codebook $\mathcal{T}_n$ of $2^{n(R(D)+\epsilon)}$ codewords Y_1^n generated i.i.d. from Q_n^*, where $R(D)$ is the rate-distortion function of the source (in bits). The proof that $2^{n(R(D)+\epsilon)}$ codewords indeed suffice is based on the following result [75, Lemma 1].

Shannon's "Lemma 1": For $x_1^n \in A^n$ let $B(x_1^n, D)$ denote the distortion-ball of radius D around x_1^n, i.e., the collection of all reproduction strings $y_1^n \in \hat{A}^n$ with $\rho_n(x_1^n, y_1^n) \le D$. When n is large[2]

$$Q_n^*(B(X_1^n, D)) \ge 2^{-n(R(D)+\epsilon)} \text{ with high probability.} \tag{4}$$

In the proof of the coding theorem this lemma plays the same role that the AEP played in the lossless case; notice the similarity between (4) and its analog (2) in the lossless case. Let us fix a source string X_1^n to be encoded. The probability that X_1^n matches any one of the codewords Y_1^n in $\mathcal{T}_n$ is

$$\begin{aligned}\Pr\{\rho_n(X_1^n, Y_1^n) \le D | X_1^n\} &= \Pr\{Y_1^n \in B(X_1^n, D) | X_1^n\} \\ &= Q_n^*(B(X_1^n, D))\end{aligned}$$

and by the lemma this probability is at least $2^{-n(R(D)+\epsilon)}$. Therefore, with $2^{n(R(D)+\epsilon)}$ independent codewords to choose from, we have a good chance for finding a match with distortion D or less.

The Generalized AEP: A stronger and more general version of Lemma 1 will be our starting point in this paper. In the following section, we will prove a *generalized AEP*. For any product measure Q^n on $\hat{A}^n$

$$-\frac{1}{n}\log Q^n(B(X_1^n, D)) \to R_1(P, Q, D) \quad \text{w.p. 1} \tag{5}$$

[2]The notation in Shannon's statement is slightly different, and he considers the more general case of ergodic sources. For the sake of clarity, we restrict our attention here to the i.i.d. case. It is also worth mentioning that the outline of the random coding argument is already given in Shannon's 1948 paper [74, Sec. 28], where the content of Lemma 1 is described in words. Paraphrasing that passage in our present notation: "A calculation [...] shows that with large n almost all of the X_1^n.s are covered within distortion D by [...] the chosen Y_1^n's."

where $R_1(P, Q, D)$ is a (nonrandom) function of the distributions P and Q and of the distortion level D. Note that in the case of lossless compression (with $P = Q$ and $D = 0$), the generalized AEP in (5) reduces to the classical AEP in (1). In fact, in our subsequent development of lossy data compression, the generalized AEP will play essentially the same role that the AEP played in the lossless case.

But also there is an essential difference between the two. Although the natural abstract formulation of the classical AEP is as an ergodic theorem [13], [14], the natural mathematical framework for understanding and proving the generalized AEP is the theory of large deviations. To see this, let us fix a realization x_1^n of the random variables X_1^n, and suppose that distortion is measured with respect to a single-letter distortion measure $\rho(x, y)$. Then, the Q^n-probability of the ball $B(x_1^n, D)$ can be written as

$$Q^n(B(x_1^n, D)) = \Pr\{Y_1^n \in B(x_1^n, D)\} = \Pr\left\{\frac{1}{n}\sum_{i=1}^{n} \rho(x_i, Y_i) \le D\right\}$$

where $Y_1^n = (Y_1, Y_2, \ldots, Y_n)$ denote n i.i.d. random variables with distribution Q. As we will see, the range of interesting distortion values is when D is smaller than the average distortion $E[\rho(X, Y)]$, in which case $Q^n(B(x_1^n, D))$ can be thought of as large deviations probability for the lower tail of the partial sums $\sum_{i=1}^n Z_i$ of the independent (but not identically distributed) random variables $Z_i = \rho(x_i, Y_i)$. Therefore, it is natural to expect that the probabilities $Q^n(B(X_1^n, D))$ will indeed decrease at some exponential rate, and the natural tools for proving this exponential convergence (i.e., the generalized AEP in (5)) will come from large deviations. For example, the proof of (5) in Theorem 1 is a direct application of the Gärtner–Ellis theorem. Similarly, more elaborate large deviations techniques will be employed to prove several variants of (5) under much weaker assumptions.

Aaron Wyner's Influence: Like the AEP in the lossless case, the generalized AEP and its refinements find numerous applications in data compression, universal data compression, and in general pattern matching questions. Many of these applications were inspired by the treatment in Wyner and Ziv's 1989 paper [84]. A (very incomplete) sample of subsequent work in the Wyner–Ziv spirit includes the work in [69], [78] elaborating on the Wyner–Ziv results, the papers [77], [60], [88], [53] on lossy data compression, and [60], [27], [3], [91] on pattern matching; see also the recent text [79].

Aaron Wyner himself remained active in this field for the following ten years, and his last paper [87], coauthored with J. Ziv and A. J. Wyner, was a review paper on this subject. In the present paper, we review the corresponding developments in the lossy case, and in the process we add new results (and some new proofs of recent results) in an attempt to present a more complete picture.

C. Central Themes, Paper Outline

In Section II, we give an extensive discussion of the generalized AEP. By now there are numerous different proofs under different assumptions, and we offer a streamlined approach to the most general versions using techniques from large deviation theory (cf. [88], [27], [20], [21] and Bucklew's earlier work [16], [17]). We also discuss the relationship of the generalized AEP with the classical extensions of the AEP (due to [6], [66]) to processes with densities. We establish a formal connection between these two by looking at the limit of the distortion level $D \downarrow 0$.

In Section III, we develop applications of the generalized AEP to a number of related problems. We show how the generalized AEP can be used to determine the asymptotic behavior of Shannon's random coding scheme, and we discuss the role of mismatch in lossy data compression. We also determine the first-order asymptotic behavior of waiting times and longest match-lengths between stationary processes. The main ideas used here are strong approximation [51] and duality [84]. We present strengthened versions of Shannon's direct lossy source-coding theorem (and of a corresponding universal coding theorem), showing that *almost all* random codebooks achieve essentially the same compression performance. A lossy version of the Lempel–Ziv algorithm is recalled, which achieves optimal compression performance (asymptotically) as well as polynomial complexity at the encoder. We also briefly mention how the classical source-coding problem can be generalized to a question about weighted sphere-covering. The answer to this question gives, as corollaries, Shannon's coding theorems, Stein's lemma in hypothesis testing, and some converse concentration inequalities.

Section IV is devoted to second-order refinements of the AEP and the generalized AEP. It is shown, for example, that under certain conditions, $-\log Q^n(B(X_1^n, D))$ are asymptotically Gaussian. The main idea is to refine the generalized AEP in (5) by showing that the quantities $-\log Q^n(B(X_1^n, D))$ are asymptotically very close to a sequence of partial sums of i.i.d. random variables, namely,

$$-\log Q^n(B(X_1^n, D)) \approx nR_1(P, Q, D) + \sum_{i=1}^{n} g(X_i)$$

where $\approx$ denotes asymptotic equality with probability one up to terms of order $O(\log n)$, and g is an explicitly defined function with $E_P[g(X)] = 0$; see Corollary 17 for more details.

These refinements are used in Section V to provide corresponding second-order results (such as central limit theorems) for the applications considered in Section III. We prove second-order asymptotic results for waiting times and longest match-lengths. Precise redundancy rates are given for Shannon's random code, and converse coding theorems show that the random code achieves the optimal pointwise redundancy, up to terms of order $O(\log n)$. For i.i.d. sources, the pointwise redundancy is typically of order $\sigma\sqrt{n}$, where σ is the minimal coding variance of the source. When $\sigma = 0$, these fluctuations disappear, and the best pointwise redundancy is of order $O(\log n)$. The question of exactly when σ can be equal to zero is briefly discussed.

Finally, Sections VI and VII contain generalizations of some of the above results to random fields. All the results stated there are new, although most of them are straightforward generalizations of corresponding one-dimensional results.

II. The Generalized AEP

A. Notation and Definitions

We begin by introducing some basic definitions and notation that will remain in effect for the rest of the paper. We will consider a stationary-ergodic process $\boldsymbol{X} = \{X_n; n \in \mathbb{Z}\}$ taking values in a general alphabet A.[3] When talking about data compression, $\boldsymbol{X}$ will be our source and A will be called the source alphabet. We write X_i^j for the vector of random variables $X_i^j = (X_i, X_{i+1}, \ldots, X_j)$, and similarly $x_i^j = (x_i, x_{i+1}, \ldots, x_j) \in A^{j-i+1}$ for a realization of these random variables, $-\infty \le i \le j \le \infty$. We let P_n denote the marginal distribution of X_1^n on A^n $(n \ge 1)$, and write $\mathbb{P}$ for the distribution of the whole process. Similarly, we take $\boldsymbol{Y} = \{Y_n; n \in \mathbb{Z}\}$ to be a stationary-ergodic process taking values in the (possibly different) alphabet $\hat{A}$ (see footnote 2). In the context of data compression, $\hat{A}$ is the reproduction alphabet and $\boldsymbol{Y}$ has the "codebook" distribution. We write Q_n for the marginal distribution of Y_1^n on $\hat{A}^n$, $n \ge 1$, and $\mathbb{Q}$ for the distribution of the whole process $\boldsymbol{Y}$. We will always assume that the process $\boldsymbol{Y}$ is independent of $\boldsymbol{X}$.

Let ρ: $A \times \hat{A} \to [0, \infty)$ be an arbitrary nonnegative (measurable) function, and define a sequence of single-letter distortion measures ρ_n: $A^n \times \hat{A}^n \to [0, \infty)$ by

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

Given $D \ge 0$ and $x_1^n \in A^n$, we write $B(x_1^n, D)$ for the distortion-ball of radius D around x_1^n

$$B(x_1^n, D) = \{y_1^n \in \hat{A}^n \colon \rho_n(x_1^n, y_1^n) \le D\}.$$

Throughout the paper, $\log$ denotes the natural logarithm and $\log_2$ the logarithm to base 2. Unless otherwise mentioned, all familiar information-theoretic quantities (such as the entropy, mutual information, and so on) are assume to be defined in terms of natural logarithms (and are therefore given in nats).

B. Generalized AEP When $\boldsymbol{Y}$ Is i.i.d.

In the case when A is finite, the classical AEP, also known as the Shannon–McMillan–Breiman theorem (see [24, Ch. 15] or the original papers [74], [62], [13], [14]), states that as $n \to \infty$

$$-\frac{1}{n} \log P_n(X_1^n) \to H(\mathbb{P}) \quad \text{w.p. 1} \tag{6}$$

where

$$H(\mathbb{P}) \triangleq \lim_{n\to\infty} \frac{1}{n} H(X_1^n)$$

is the entropy rate of the process $\boldsymbol{X}$ (in nats, since we are taking logarithms to base e). As we saw in the Introduction, in lossy data compression the role of the AEP is taken up by the result of Shannon's "Lemma 1" and, more generally, by statements of the form

$$-\frac{1}{n} Q_n(B(X_1^n, D)) \to R(\mathbb{P}, \mathbb{Q}, D) \quad \text{w.p. 1}$$

for some nonrandom "rate-function" $R(\mathbb{P}, \mathbb{Q}, D)$.

First, we consider the simplest case where $\boldsymbol{Y}$ is assumed to be an i.i.d. process. We write $Q = Q_1$ for its first-order marginal, so that $Q_n = Q^n$, for $n \ge 1$. Similarly, we write $P = P_1$ for the first-order marginal of $\boldsymbol{X}$. Let

$$D_{\min} \triangleq E_P[\operatorname*{ess\,inf}_{Y\sim Q} \rho(X, Y)] \tag{7}$$

$$D_{\text{av}} \triangleq E_{P\times Q}[\rho(X, Y)]. \tag{8}$$

[Recall that the essential infimum of a function $g(Y)$ of the random variable Y with distribution Q is defined as $\operatorname{ess\,inf}_{Y\sim Q} g(Y) = \sup\{t \in \mathbb{R}\colon Q\{g(Y) > t\} = 1\}$.]

Clearly, $0 \le D_{\min} \le D_{\text{av}}$. To avoid the trivial case when $\rho(x, y)$ is essentially constant for (P-almost) all $x \in A$, we assume that with positive P-probability $\rho(x, y)$ is not essentially constant in y, that is,

$$D_{\min} < D_{\text{av}}. \tag{9}$$

Note also that for D greater than D_{av}, the probability $Q^n(B(X_1^n, D)) \to 1$ as $n \to \infty$ (this is easy to see by the ergodic theorem), so we restrict our attention to distortion levels $D < D_{\text{av}}$.

Theorem 1: Generalized AEP when $\boldsymbol{Y}$ is i.i.d.: Let $\boldsymbol{X}$ be a stationary-ergodic process and $\boldsymbol{Y}$ be i.i.d. with marginal distribution Q on $\hat{A}$. Assume that $D_{\text{av}} = E_{P\times Q}[\rho(X, Y)]$ is finite. Then for any $D \in (D_{\min}, D_{\text{av}})$

$$-\frac{1}{n} \log Q^n(B(X_1^n, D)) \to R_1(P, Q, D) \quad \text{w.p. 1.}$$

The rate-function $R_1(P, Q, D)$ is defined as

$$R_1(P, Q, D) = \inf_W H(W\|P \times Q)$$

where $H(W\|V)$ denotes the relative entropy between two distributions W and V

$$H(W\|V) \triangleq \begin{cases} E_W\left[\log \frac{dW}{dV}\right], & \text{if the density } \frac{dW}{dV} \text{ exists} \\ \infty, & \text{otherwise} \end{cases}$$

and the infimum is taken over all joint distributions W on $A \times \hat{A}$ such that the first marginal of W is P and $E_W[\rho(X, Y)] \le D$.

Note that Theorem 1, as well as most of our subsequent results, is only stated for distortion levels D that are *strictly* greater than $D_{\min}$. This means that, despite the fact that we think of Theorem 1 as a generalization of the AEP in the lossless case, here the lossless case (corresponding to $D = D_{\min} = 0$) is excluded. There are two reasons for this. First, the large deviations techniques used to prove many of our main results do not immediately yield any information on what happens at the boundary point $D = D_{\min}$. And second, the results themselves do not always remain valid in that case; for example, although $R_1(P, Q, D)$ is finite for all $D > D_{\min}$, it may be infinite at $D = D_{\min}$ as Example 1 illustrates. A more dramatic difference

[3] To avoid uninteresting technicalities, we will assume throughout that A is a complete, separable metric space, equipped with its associated Borel σ-field $\mathcal{A}$. Similarly, we take $(\hat{A}, \hat{\mathcal{A}})$ to be the Borel measurable space corresponding to a complete, separable metric space $\hat{A}$.

between the lossless $(D = 0)$ and lossy $(D > 0)$ case is the result of Theorem 15, where the term $(1/2)(\log n)$ disappears in the lossless case.

Example 1: The rate-function $R_1(P, Q, D)$ when Q is Gaussian: Although in general the rate-function $R_1(P, Q, D)$ cannot be evaluated explicitly, here we show that it is possible to obtain an exact expression for $R_1(P, Q, D)$ in the special case when $\rho(x, y) = (x - y)^2$, $\boldsymbol{X}$ is a real-valued process, and Q is a Gaussian measure on $\mathbb{R}$. Specifically, assume that $\boldsymbol{X}$ is a zero-mean, stationary-ergodic process with finite variance $\sigma^2 = \mathrm{Var}(X_1) < \infty$, and take Q to be a zero-mean Gaussian measure with variance τ^2, i.e., $Q \sim N(0, \tau^2)$. Under these assumptions, it is easy to see that $D_{\min} = 0$ and $D_{\mathrm{av}} = \sigma^2 + \tau^2$. Moreover, with the help of Theorem 2 that follows, $R_1(P, Q, D)$ can be explicitly evaluated as

$$R_1(P, Q, D) = \begin{cases} \infty, & D = 0 \\ \frac{1}{2}\log\left(\frac{v}{D}\right) - \frac{(v-D)(v-\sigma^2)}{2v\tau^2}, & 0 < D < \sigma^2 + \tau^2 \\ 0, & D \geq \sigma^2 + \tau^2 \end{cases}$$

where

$$v \triangleq \frac{1}{2}\left[\tau^2 + \sqrt{\tau^4 + 4D\sigma^2}\right].$$

We will come back to this example when considering mismatched rate-distortion codebooks in Section III-B.

Remark 1: In more familiar information-theoretic terms, the rate-function $R_1(P, Q, D)$ can equivalently be defined as (cf. [88])

$$R_1(P, Q, D) = \inf_{(X,Y)} [I(X; Y) + H(Q_Y \| Q)]$$

where $I(X; Y)$ denotes the mutual information (in nats) between the random variables X and Y, and the infimum is over all jointly distributed random variables (X, Y) with values in $A \times \hat{A}$ such that X has distribution P, $E[\rho(X, Y)] \leq D$, and Q_Y denotes the distribution of Y.

Remark 2: The assumption that $\boldsymbol{Y}$ is i.i.d. is clearly restrictive and it will be relaxed below. On the other hand, the assumptions on the distortion measure ρ seem to be minimal; we simply assume that ρ has finite expectation (in the more general results below ρ is assumed to be bounded). In this form, the result of Theorem 1 is new.

Proof Outline: As discussed in the Introduction, Theorem 1 will be proved by an application of the Gärtner–Ellis theorem from large deviations; see [29, Theorem 2.3.6]. Choose and fix a realization x_1^∞ of $\boldsymbol{X}$ and define the random variables $Z_i = \rho(x_i, Y_i)$. Let

$$S_n = \frac{1}{n}\sum_{i=1}^{n} Z_i$$

and define the log-moment generating functions of the normalized partial sums $(1/n)S_n$ by

$$\Lambda_n(\lambda) \triangleq \log E_{Q^n}(e^{\lambda S_n/n}), \qquad \lambda \leq 0.$$

Then for any $\lambda \leq 0$, by the ergodic theorem we have that

$$\frac{1}{n}\Lambda_n(n\lambda) = \frac{1}{n}\sum_{i=1}^{n} \log E_Q\left(e^{\lambda\rho(x_i, Y_i)}\right) \to \Lambda(\lambda) \triangleq E_P\left[\log E_Q\left(e^{\lambda\rho(X,Y)}\right)\right] \tag{10}$$

for $\mathbb{P}$-almost any realization x_1^∞. Now we would like to apply the Gärtner–Ellis theorem, but first we need to check some simple properties of the function $\Lambda(\lambda)$. Note that $\Lambda(\lambda) \leq 0$ and also (by Jensen's inequality) $\Lambda(\lambda) \geq \lambda D_{\mathrm{av}} > -\infty$, for all $\lambda \leq 0$. Moreover, $\Lambda(\lambda)$ is twice differentiable in λ with

$$\Lambda'(\lambda) = E_{P\times Q}\left(\rho(X, Y)\frac{e^{\lambda\rho(X,Y)}}{E_Q\left[e^{\lambda\rho(X,Y')}\right]}\right)$$

and

$$\Lambda''(\lambda) = E_P\left[E_Q\left\{\rho^2(X, Y)\frac{e^{\lambda\rho(X,Y)}}{E_Q\left[e^{\lambda\rho(X,Y')}\right]}\right\} - \left(E_Q\left\{\rho(X, Y)\frac{e^{\lambda\rho(X,Y)}}{E_Q\left[e^{\lambda\rho(X,Y')}\right]}\right\}\right)^2\right]$$

(this differentiability is easily verified by an application of the dominated convergence theorem). By the Cauchy–Schwarz inequality $\Lambda''(\lambda) \geq 0$ for all $\lambda < 0$, and in fact $\Lambda''(\lambda)$ is strictly positive due to assumption (9). Also it is not hard to verify that $\lim_{\lambda\uparrow 0}\Lambda'(\lambda) = D_{\mathrm{av}}$ and

$$\lim_{\lambda\downarrow -\infty}\Lambda'(\lambda) = D_{\min}. \tag{11}$$

Since $D \in (D_{\min}, D_{\mathrm{av}})$, there exists a unique $\lambda^* < 0$ with $\Lambda'(\lambda^*) = D$, and, therefore, the Fenchel–Legendre transform of $\Lambda(\lambda)$ evaluated at D is

$$\Lambda^*(D) \triangleq \sup_{\lambda\leq 0}[\lambda D - \Lambda(\lambda)] = \lambda^* D - \Lambda(\lambda^*).$$

Now we can apply the Gärtner–Ellis theorem [29, Theorem 2.3.6] to deduce from (10) that with $\mathbb{P}$-probability one

$$-\frac{1}{n}\log Q^n(B(X_1^n, D)) \to \Lambda^*(D).$$

The proof is complete upon noticing that $\Lambda^*(D)$ is nothing but $R_1(P, Q, D)$. This is stated and proved in the following theorem. □

Theorem 2—Characterization of the Rate Function: In the notation of the proof of Theorem 1,

$$\Lambda^*(D) = R_1(P, Q, D), \qquad \text{for } D \in (D_{\min}, D_{\mathrm{av}}).$$

Proof Outline: Under additional assumptions on the distortion measure ρ this has appeared in various papers (see, e.g., [27], [90]). For completeness, we offer a proof sketch here.

In the notation of the above proof, consider the measure W on $A \times \hat{A}$ defined by

$$\frac{dW(x, y)}{d(P \times Q)} = \frac{e^{\lambda^*\rho(x,y)}}{E_Q\left[e^{\lambda^*\rho(x,Y)}\right]}.$$

Obviously, the first marginal of W is P and it is easy to check that $E_W[\rho(X, Y)] = \Lambda'(\lambda^*) = D$. Therefore, by the definitions of $R_1(P, Q, D)$ and W, and by the choice of λ^*

$$\begin{aligned} R_1(P, Q, D) &\le H(W\|P \times Q) \\ &= \lambda^* D - \Lambda(\lambda^*) = \Lambda^*(D). \end{aligned} \tag{12}$$

To prove the corresponding lower bound we first claim that for any measurable function $\phi: \hat{A} \to (-\infty, 0]$, and any probability measure Q' on $\hat{A}$

$$H(Q'\|Q) \ge E_{Q'}(\phi(Y)) - \log E_Q(e^{\phi(Y)}). \tag{13}$$

Let Q_ϕ denote the probability measure on $\hat{A}$ such that $dQ_\phi/dQ = e^\phi / E_Q(e^{\phi(Y)})$. Clearly, it suffices to prove (13) in case dQ'/dQ exists, in which case the difference between the left- and right-hand sides is

$$E_{Q'}\left\{\log \frac{dQ'}{dQ}\right\} - E_{Q'}\left\{\log\left(\frac{e^\phi}{E_Q(e^\phi)}\right)\right\} = H(Q'\|Q_\phi) \ge 0.$$

Given an arbitrary candidate W as in the definition of $R_1(P, Q, D)$ and any $x \in A$, we take $Q' = W(\cdot|x)$ and $\phi(y) = \lambda^* \rho(x, y)$ in (13) to get that

$$H(W(\cdot|x)\|Q(\cdot)) \ge \lambda^* E_{W(\cdot|x)}[\rho(x, Y)] - \log E_Q(e^{\lambda^* \rho(x, Y)}).$$

Substituting X for x, taking expectations of both sides with respect to P, and recalling that $\lambda^* < 0$ and $E_W[\rho(X, Y)] \le D$, we get

$$H(W\|P \times Q) \ge \lambda^* D - \Lambda(\lambda^*) = \Lambda^*(D).$$

Since W was arbitrary, it follows that $R_1(P, Q, D) \ge \Lambda^*(D)$, and together with (12) this completes the proof. □

C. Generalized AEP When $\boldsymbol{Y}$ Is Not i.i.d.

Next we present two versions of the generalized AEP that hold when $\boldsymbol{Y}$ is a stationary dependent process, under some additional conditions.

Throughout this section we will assume that the distortion measure is *essentially bounded*, i.e.,

$$D_{\max} \triangleq \operatorname*{ess\,sup}_{(X_1, Y_1) \sim P_1 \times Q_1} \rho(X_1, Y_1) < \infty. \tag{14}$$

We let D_{av} be defined as earlier, $D_{\text{av}} = E_{P_1 \times Q_1}[\rho(X_1, Y_1)]$, and for $n \ge 1$ we let

$$D_{\min}^{(n)} \triangleq E_{P_n}\left[\operatorname*{ess\,inf}_{Y_1^n \sim Q_n} \rho_n(X_1^n, Y_1^n)\right].$$

It is easy to see that $nD_{\min}^{(n)}$ is a finite, superadditive sequence, and therefore we can also define

$$D_{\min} = \lim_{n \to \infty} D_{\min}^{(n)} = \sup_{n \ge 1} D_{\min}^{(n)}.$$

As before, we will assume that the distortion measure ρ is not essentially constant, that is, $D_{\min} < D_{\text{av}}$.

We first state a version of the generalized AEP that was recently proved by Chi [20], for processes $\boldsymbol{Y}$ satisfying a rather strong mixing condition. We say that the stationary process $\boldsymbol{Y}$ is $\psi^{\pm}$-*mixing*, if for all d large enough there is a finite constant c_d such that

$$c_d^{-1} \mathbb{Q}(A)\mathbb{Q}(B) < \mathbb{Q}(A \cap B) < c_d \mathbb{Q}(A)\mathbb{Q}(B)$$

for all events $A \in \sigma(Y_{-\infty}^0)$ and $B \in \sigma(Y_d^\infty)$, where $\sigma(Y_i^j)$ denotes the σ-field generated by Y_i^j. Recall the usual definition according to which $\boldsymbol{Y}$ is called ψ-*mixing* if in fact the constants $c_d \to 1$ as $d \to \infty$; see [12] for more details. Clearly, $\psi^{\pm}$-mixing is weaker than ψ-*mixing*.

Theorem 3—Generalized AEP When $\boldsymbol{Y}$ is $\psi^{\pm}$-Mixing [20]: Let $\boldsymbol{X}$ and $\boldsymbol{Y}$ be stationary-ergodic processes. Assume that $\boldsymbol{Y}$ is $\psi^{\pm}$-mixing, and that the distortion measure ρ is essentially bounded, $D_{\max} < \infty$. Then for all $D \in (D_{\min}, D_{\text{av}})$

$$-\frac{1}{n} \log Q_n(B(X_1^n, D)) \to R(\mathbb{P}, \mathbb{Q}, D) \quad \text{w.p. 1} \tag{15}$$

where $R(\mathbb{P}, \mathbb{Q}, D)$ is the rate-function defined by

$$R(\mathbb{P}, \mathbb{Q}, D) = \lim_{n \to \infty} R_n(P_n, Q_n, D) \tag{16}$$

where, for $n \ge 1$

$$R_n(P_n, Q_n, D) \triangleq \inf_{V_n} n^{-1} H(V_n\|P_n \times Q_n)$$

and the infimum is taken over all joint distributions V_n on $A^n \times \hat{A}^n$ such that the A^n-marginal of V_n is P_n and

$$E_{V_n}[\rho_n(X_1^n, Y_1^n)] \le D.$$

As we discussed in the previous section, the proof of most versions of the generalized AEP consists of two steps. First, a "conditional large deviations" result is proved for the random variables $\{\rho_n(x_1^n, Y_1^n);\ n \ge 1\}$, where x_1^∞ is a fixed realization of the process $\boldsymbol{X}$. Second, the rate-function $R(\mathbb{P}, \mathbb{Q}, D)$ is characterized as the limit of a sequence of minimizations in terms of relative entropy.

In a subsequent paper, Chi [21] showed that the first of these steps (the large deviations part) remains valid under a condition weaker than $\psi^{\pm}$-mixing, condition (S) of [15]. In the following theorem, we give a general version of the second step; we prove that the generalized AEP (15) and the formula (16) for the rate-function remain valid as long as the random variables $\{\rho_n(x_1^n, Y_1^n);\ n \ge 1\}$ satisfy a large deviations principle (LDP) with some *deterministic*, convex rate-function (see [29] for the precise meaning of this statement).

Theorem 4: Let $\boldsymbol{X}$ and $\boldsymbol{Y}$ be stationary processes. Assume that ρ is essentially bounded, i.e., $D_{\max} < \infty$, and that with $\mathbb{P}$-probability one, conditional on $X_1^\infty = x_1^\infty$, the random variables $\{\rho_n(x_1^n, Y_1^n);\ n \ge 1\}$ satisfy a large deviations principle with some deterministic, convex rate-function. Then, both (15) and (16) hold for any $D \in (D_{\min}, D_{\text{av}})$, except possibly at the point $D = D_{\min}^{(\infty)}$, where

$$D_{\min}^{(\infty)} \triangleq \inf\left\{D \ge 0:\ \sup_{n \ge 1} R_n(P_n, Q_n, D) < \infty\right\}. \tag{17}$$

Since Theorem 4 has an exact analog in the case of random fields, we postpone its proof until the proof of the corresponding result (Theorem 25) in Section VI.

As will be seen in the proof, the rate-function $R(\mathbb{P}, \mathbb{Q}, D) = \infty$ for $D < D_{\min}^{(\infty)}$ and it is finite for $D > D_{\min}^{(\infty)}$. Recall that, similarly, each $R_n(P_n, Q_n, D)$ is finite when $D > D_{\min}$, but without additional assumptions on $\mathbb{Q}$ it is now possible that there are distortion values D such that $R_n(P_n, Q_n, D) < \infty$ for all n, but $R(\mathbb{P}, \mathbb{Q}, D) = \infty$. Note that this difficulty was not present in Theorem 3, where the mixing property of $\mathbb{Q}$ was used to show that indeed $D_{\min} = D_{\min}^{(\infty)}$.

Remark 3: Suppose that the joint process $(\boldsymbol{X}, \boldsymbol{Y})$ is stationary, and that it satisfies a "process-level large deviations principle" (see Remark 6 in Section VI for a somewhat more detailed statement) on the space of stationary probability measures on $(A^\infty \times \hat{A}^\infty)$ equipped with the topology of weak convergence. Assume, moreover, that this LDP holds with a convex, good rate-function $I(\cdot)$. (See [26], [30, Sec. 5.3, 5.4], [29, Sec. 6.5.3], [15] for a general discussion as well as specific examples of processes for which the above conditions hold. Apart from the i.i.d. case, these examples also include all ergodic finite-state Markov chains, among many others.)

It is easy to check that, when ρ is bounded and continuous on $A \times \hat{A}$, then with $\mathbb{P}$-probability one, conditional on x_1^∞, the random variables $\{\rho_n(x_1^n, Y_1^n)\}$ satisfy the LDP upper bound with respect to the deterministic, convex rate-function $J(D) = \inf I(\nu)$, where the infimum is over all stationary probability measures ν on $A^\infty \times \hat{A}^\infty$ such that the A^∞-marginal of ν is $\mathbb{P}$ and $E_\nu[\rho(X_1, Y_1)] = D$. Indeed, Comets [23] provides such an argument when $\boldsymbol{X}$ and $\boldsymbol{Y}$ are both i.i.d. Moreover, he shows that in that case, the corresponding LDP lower bound also holds, and hence Theorem 4 applies. Unfortunately, the conditional LDP lower bound has to be verified on a case-by-case basis.

Remark 4: Although quite strong, the $\psi^{\pm}$-mixing condition of Theorem 3, and the (S)-mixing condition of [21], probably cannot be significantly relaxed. For example, in the special case when $\boldsymbol{X}$ is a constant process taking on just a single value, if Theorem 3 were to hold (for any bounded distortion measure) with a strictly monotone rate-function, then necessarily the empirical measures of Y_1^n would satisfy the LDP in the space $\mathcal{P}_a(\hat{A})$ (see [15] for details). But [15, Example 1] illustrates that this LDP may fail even when $\boldsymbol{Y}$ is a stationary-ergodic Markov chain with discrete alphabet $\hat{A}$. In particular, the example in [15] has an exponential ϕ-mixing rate.

D. Generalized AEP for Optimal Lossy Compression

Here we present a version of the generalized AEP that is useful in proving direct coding theorems. Let $\boldsymbol{X}$ be a stationary-ergodic process. For the distortion measure ρ we adopt two simple regularity conditions. We assume the existence of a *reference letter*, i.e., an $\hat{a} \in \hat{A}$ such that

$$E_{P_1}[\rho(X_1, \hat{a})] < \infty.$$

Also, following [48], we require that for any distortion level $D > 0$ there is a scalar quantizer for $\boldsymbol{X}$ with finite rate.

Quantization Condition: For each $D > 0$, there is a "quantizer" $q: A \to B$ for some countable (finite or infinite) subset $B \subset \hat{A}$, such that

i) $\rho(x, q(x)) \le D$ for all $x \in A$;

ii) the entropy $H(q(X_1)) < \infty$.

The following was implicitly proved in [48]; see also [56] for details.

Theorem 5—Generalized AEP for Optimal Lossy Compression [48]: Let $\boldsymbol{X}$ be a stationary-ergodic process. Assume that the distortion measure ρ satisfies the quantization condition, that a reference letter exists, and that for each $n \ge 1$ the infimum of

$$E_{P_n}[-\log Q_n(B(X_1^n, D))]$$

over all probability measures Q_n on $\hat{A}^n$ is achieved by some $\tilde{Q}_n$. Then for any $D > 0$

$$-\frac{1}{n} \log \tilde{Q}_n(B(X_1^n, D)) \to R(D) \quad \text{w.p. } 1 \tag{18}$$

where $R(D)$ is the rate-distortion function of the process $\boldsymbol{X}$.

Historical Remarks: The relevance of the quantities $-\log Q_n(B(X_1^n, D))$ to information theory was first suggested implicitly by Kieffer [48] and more explicitly by Łuczak and Szpankowski [60]. Since then, many papers have appeared proving the generalized AEP under different conditions; we mention here a subset of those proving some of the more general results. The case of finite alphabet processes was considered by Yang and Kieffer [88]. The generalized AEP for processes with general alphabets and $\boldsymbol{Y}$ i.i.d. was proved by Dembo and Kontoyiannis [27] and by Yang and Zhang [90]. Finally, the case when $\boldsymbol{Y}$ is not i.i.d. was (Theorem 3) treated by Chi [20], [21]. The observations of Theorem 4 about the rate-function $R(\mathbb{P}, \mathbb{Q}, D)$ are new. Theorem 5 essentially comes from Kieffer's work [48]; see also [56]. A different version of the generalized AEP (based on fixed-composition codebooks) was recently utilized in [96] in the context of adaptive lossy compression. We should also mention that, in a somewhat different context, the intimate relationship between the AEP and large deviations is discussed in some detail by Orey in [67].

E. Densities Versus Balls

Let us recall the classical generalization of the AEP, due to Barron [6] and Orey [66], to processes with values in general alphabets. Suppose $\boldsymbol{X}$ as above is a general stationary-ergodic process with marginals $\{P_n\}$ that are absolutely continuous with respect to the sequence of measures $\mathbb{M} = \{M_n\}$.

Theorem 6—AEP for Processes With Densities [6], [66]: Let $\boldsymbol{X}$ be a stationary-ergodic process whose marginals P_n have densities $f_n = dP_n/dM_n$ with respect to the σ-finite measures M_n, $n \ge 1$. Assume that the sequence $\mathbb{M}$ of dominating measures is Markov of finite order, with a stationary transition measure, and that the relative entropies

$$H_n \triangleq E_{P_n}\left[\log \frac{f_n(X_1^n)}{f_{n-1}(X_1^{n-1})}\right], \qquad n \ge 2$$

have $H_n > -\infty$ eventually. Then

$$-\frac{1}{n}\log\frac{dP_n}{dM_n}(X_1^n) \to -H(\mathbb{P}\|\mathbb{M}) \quad \text{w.p. 1} \tag{19}$$

where $H(\mathbb{P}\|\mathbb{M})$ is the relative entropy rate defined as

$$H(\mathbb{P}\|\mathbb{M}) = \lim_{n\to\infty} H_n = \inf_{n\geq 1} H_n.$$

The AEP for processes with densities is also know to hold when the reference measures M_n do not form a Markov sequence, under some additional mixing conditions (see [66], where M_n are taken to be non-Markov measures satisfying an additional mixing condition, and the more recent extension in [18] where the M_n are taken to be discrete Gibbs measures). Moreover, Kieffer [46], [47] has given counterexamples illustrating that without some mixing conditions on $\{M_n\}$ the AEP (19) fails to hold.

There is a tempting analogy between the generalized AEP (15) and the AEP for processes with densities (19). The formal similarity between the two suggests that, if we identify the measures Q_n with the reference measures M_n, corresponding results should hold in the two cases. Indeed, this does in general appear to be the case, as is illustrated by the various generalized AEPs stated above. Moreover, we can interpret the result of Theorem 5 as the natural analog of the classical discrete AEP (6) to the case of lossy data compression. As we argued in the Introduction, the generalized AEPs of the previous sections play analogous roles in the proofs of the corresponding direct coding theorems.

Taking this analogy further indicates that there might be a relationship between these two different generalizations. In particular, when n is large and the distortion level D is small, the following heuristic calculation seems compelling. Assuming for a moment that A and $\hat{A}$ are the same

$$\begin{aligned}
-H(\mathbb{P}\|\mathbb{Q}) &\overset{(a)}{\approx} -\frac{1}{n}\log\frac{dP_n}{dQ_n}(X_1^n)\\
&\overset{(b)}{\approx} -\frac{1}{n}\log\frac{P_n(B(X_1^n,D))}{Q_n(B(X_1^n,D))}\\
&= -\frac{1}{n}\log P_n(B(X_1^n,D)) + \frac{1}{n}\log Q_n(B(X_1^n,D))\\
&\overset{(c)}{\approx} R(\mathbb{P},\mathbb{P},D) - R(\mathbb{P},\mathbb{Q},D)\\
&\overset{(d)}{\approx} -H(\mathbb{P}\|\mathbb{Q})
\end{aligned}$$

where (a) holds in the limit as $n\to\infty$ by Theorem 6, (b) should hold when D is small by the assumption that P_n has a density with respect to Q_n, (c) would follow in the limit as $n\to\infty$ by an application of the generalized AEP, and it is natural to conjecture that (d) holds in the limits as $D\downarrow 0$ by reading the above calculation backward.

We next formalize this heuristic argument in two special cases. First, when $\boldsymbol{X}$ is a discrete process taking values in a finite alphabet, and second when $\boldsymbol{X}$ is a continuous process taking values in $\mathbb{R}^d$.

1) Discrete Case: Here we take $\boldsymbol{X}$ to be a stationary-ergodic process taking values in a finite alphabet A, and $\boldsymbol{Y}$ to be i.i.d. with first-order marginal distribution $Q = Q_1$ on the same alphabet $A = \hat{A}$. Similarly, we write $P = P_1$ for the first-order marginal of $\boldsymbol{X}$. In Theorem 7, we justify the above calculation by showing that the limits as $D\downarrow 0$ and as $n\to\infty$ can indeed be taken together in any fashion. We show that the double limit of the central expression

$$r_n(X_1^n, D) \triangleq \frac{1}{n}\log\frac{P_n(B(X_1^n,D))}{Q_n(B(X_1^n,D))} \tag{20}$$

is equal to $H(\mathbb{P}\|\mathbb{Q})$ with probability 1, independently of how n grows and D decreases to zero. Its proof is given in Appendix A.

Theorem 7—Densities Versus Balls in the Discrete Case: Let $\boldsymbol{X}$ be a stationary-ergodic process and $\boldsymbol{Y}$ be i.i.d., both on the finite alphabet A. Assume that $\rho(x, y) = 0$ if and only if $x = y$, and $Q(x) > 0$ for all x. Then the following double limit exists:

$$\lim_{\substack{n\to\infty\\ D\downarrow 0}} \frac{1}{n}\log\frac{P_n(B(X_1^n,D))}{Q^n(B(X_1^n,D))} = H(\mathbb{P}\|\mathbb{Q}) \quad \text{w.p. 1.}$$

In particular, the repeated limit $\lim_n \lim_D$ exists with probability one and is equal to $H(\mathbb{P}\|\mathbb{Q})$.

2) Continuous Case: Here we state a weaker version of Theorem 7 in the case when $A = \hat{A} = \mathbb{R}^d$ for some $d \geq 1$, and when $\boldsymbol{X}$ is an $\mathbb{R}^d$-valued, stationary-ergodic process. Suppose that the marginals $\{P_n\}$ of $\boldsymbol{X}$ are absolutely continuous with respect to a sequence of reference measures $\{Q_n\}$. Throughout this section we take the Q_n to be product measures $Q_n = Q^n$ for some fixed Borel probability measure Q on $\mathbb{R}^d$. A typical example to keep in mind is when Q a Gaussian measure on $\mathbb{R}$ and $\boldsymbol{X}$ a real-valued stationary-ergodic process all of whose marginals P_n have continuous densities with respect to Lebesgue measure.

For simplicity, we take ρ to be squared-error distortion $\rho(x, y) = (x - y)^2$, although the proof of Theorem 8, given in Appendix B, may easily be adapted to apply for somewhat more general difference distortion measures.

Theorem 8—Densities Versus Balls in the Continuous Case: Let $\boldsymbol{X}$ be an $\mathbb{R}^d$-valued stationary-ergodic process, whose marginals P_n have densities $f_n = dP_n/dQ_n$ with respect to a sequence of product measures $Q_n = Q^n$, $n \geq 1$, for a given probability measure Q on $\mathbb{R}^d$. Let $\rho(x, y) = (x - y)^2$ for any $x, y \in \mathbb{R}^d$.

a) The following repeated limit holds:

$$\lim_{n\to\infty}\lim_{D\downarrow 0}\frac{1}{n}\log\frac{P_n(B(X_1^n,D))}{Q_n(B(X_1^n,D))} = H(\mathbb{P}\|\mathbb{Q}) \quad \text{w.p. 1.}$$

b) Assume, moreover, that $\boldsymbol{X}$ is i.i.d. with marginal distribution $P_1 = P$ on $\mathbb{R}^d$, and that the following conditions are satisfied. Both $E_{P\times Q}[\rho(X, Y)]$ and $E_{P\times P}[\rho(X, Y)]$ are finite and nonzero, the expectation

$$E_P[-\log Q(B(X, D))] \text{ is finite for all } D > 0$$

and a $\delta > 0$ exists for which

$$E_P\left[\sup_{0<D<\delta}\left|\log\frac{P(B(X,D))}{Q(B(X,D))}\right|\right] < \infty. \tag{21}$$

Then, the reverse repeated limit also holds:

$$\lim_{D\downarrow 0}\lim_{n\to\infty}\frac{1}{n}\log\frac{P_n(B(X_1^n,D))}{Q_n(B(X_1^n,D))}=H(\mathbb{P}\|\mathbb{Q})\quad\text{w.p. 1.}$$

It is easy to check that all conditions of the theorem hold when Q is a Gaussian measure on $\mathbb{R}$ and P has finite variance and a probability density function g (with respect to Lebesgue measure) such that $E_P(\sup_{|y-X|<\delta}|\log g(y)|)<\infty$ for some $\delta>0$. For example, this is the case when both P and Q are Gaussian distributions on $\mathbb{R}$.

As will be seen from the proof of the theorem, although we are primarily interested in the case when the relative entropy rate $H(\mathbb{P}\|\mathbb{Q})$ is finite, the result remains true when $H(\mathbb{P}\|\mathbb{Q})=\infty$, and in that case assumption (21) can be relaxed to

$$E_P\left[\sup_{0<D<\delta}\log\frac{Q(B(X,D))}{P(B(X,D))}\right]<\infty.$$

Finally, we note that, in the context of ergodic theory, Feldman [34] developed a different version of the generalized AEP, and also discussed the relationship between the two types of asymptotics (as $n\to\infty$, and as $D\downarrow 0$).

III. Applications of the Generalized AEP

As outlined in the Introduction, the generalized AEP can be applied to a number of problems in data compression and pattern matching. Following along the lines of the corresponding applications in the lossless case, in the following we present applications of the results of the previous section to: 1) Shannon's random coding schemes; 2) mismatched codebooks in lossy data compression; 3) waiting times between stationary processes (corresponding to idealized Lempel–Ziv coding); 4) practical lossy Lempel–Ziv coding for memoryless sources; and 5) weighted codebooks in rate-distortion theory.

A. Shannon's Random Codes

Shannon's well-known construction of optimal codes for lossy data compression is based on the idea of generating a random codebook. We review here a slightly modified version of his construction [75] and describe how the performance of the resulting random code can be analyzed using the generalized AEP.

Given a sequence of probability distributions Q_n on $\hat{A}^n$, $n\geq 1$, we generate a *random codebook according to the measures* Q_n as an infinite sequence of i.i.d. random vectors

$$Y_1^n(i),\qquad i\geq 1$$

with each $Y_1^n(i)$ having distribution Q_n on $\hat{A}^n$. Suppose that, for a fixed n, this codebook is available to both the encoder and decoder. Given a source string X_1^n to be described with distortion D or less, the encoder looks for a D-close match of X_1^n into the codebook $\{Y_1^n(i);\,i\geq 1\}$. Let i_n be the position of the first such match

$$i_n\stackrel{\Delta}{=}\inf\{i\geq 1\colon \rho_n(X_1^n,Y_1^n(i))\leq D\}$$

with the convention that the infimum of the empty set equals $+\infty$. If a match is found, then the encoder describes to the decoder the position i_n using Elias' code for the integers [32]. This takes no more than

$$\log_2 i_n+2\log_2\log_2 i_n+\text{const.}\quad\text{bits.}\tag{22}$$

If no match is found (something that asymptotically will *not* happen, with probability one), then the encoder describes X_1^n with distortion D or less using some other default scheme.

Let $\ell_n(X_1^n)$ denote the overall description length of the algorithm just described. In view of (22), in order to understand its compression performance, that is, to understand the asymptotic behavior of $\ell_n(X_1^n)$, it suffices to understand the behavior of the quantity

$$\log_2 i_n,\qquad\text{for large } n.$$

Suppose that the probability $Q_n(B(X_1^n,D))$ of finding a D-close match for X_1^n in the codebook is nonzero. Then, conditional on the source string X_1^n, the distribution of i_n is geometric with parameter $Q_n(B(X_1^n,D))$. From this observation, it is easy to deduce that the behavior of i_n is closely related to the behavior of the quantity $1/Q_n(B(X_1^n,D))$. The next theorem is an easy consequence of this fact so it is stated here without proof; see the corresponding arguments in [54], [56].

Theorem 9—Strong Approximation [4]*:* Let $\boldsymbol{X}$ be an arbitrary process and let $\{Q_n\}$ be a given sequence of codebook distributions. If $Q_n(B(X_1^n,D))>0$ eventually with probability one, then for any $\epsilon>0$

$$\log_2 i_n\leq-\log_2 Q_n(B(X_1^n,D))+\log_2\log_2 n+3\qquad\text{eventually, w.p. 1}$$

and

$$\log_2 i_n\geq-\log_2 Q_n(B(X_1^n,D))-\log_2 n-(1+\epsilon)\log_2\log_2 n\qquad\text{eventually, w.p. 1.}$$

The above estimates can now be combined with the results of the generalized AEP in the previous section to determine the performance of codes based on random codebooks with respect to the "optimal" measures Q_n. To illustrate this approach, we consider the special case of memoryless sources and finite reproduction alphabets, and show that the random code with respect to (almost) any random codebook realization is asymptotically optimal, with probability one. Note that corresponding results can be proved, in exactly the same way, under much more general assumptions. For example, utilizing Theorem 5 instead of Theorem 1 we can prove the analog of Theorem 10 below for arbitrary stationary-ergodic sources.

Let $\boldsymbol{X}$ be an i.i.d. source with marginal distribution $P_1=P$ on A, and take the reproduction alphabet $\hat{A}$ to be finite. It is assumed throughout this section that $\hat{A}$ is known to both the encoder and the decoder (note that it is not necessarily assumed to be the optimal reproduction alphabet). For simplicity,

[4]The name "strong approximation" comes from the probabilistic terminology where the adjective "strong" often refers to asymptotic results with probability one (such as the strong law of large numbers), as opposed to results about convergence in probability or in distribution.

we will assume that the distortion measure ρ is bounded, i.e., $\sup_{x,y} \rho(x, y) < \infty$, and we also make the customary assumption that

$$\sup_{x \in A} \min_{y \in \hat{A}} \rho(x, y) = 0. \tag{23}$$

[See the remark at the end of Section V-A1 for a discussion of this condition and when it can be relaxed.] As usual, we define the rate-distortion function of the memoryless source $\boldsymbol{X}$ by

$$R(D) = \inf_{(X,Y)} I(X; Y)$$

where the infimum is over all jointly distributed random variables (X, Y) with values in $A \times \hat{A}$, such that X has distribution P and $E[\rho(X, Y)] \leq D$. Let

$$\overline{D} \triangleq \min_{y \in \hat{A}} E_P[\rho(X, y)] \tag{24}$$

and note that $R(D) = 0$ for $D \geq \overline{D}$. To avoid the trivial case when $R(D) = 0$ for all D, we assume that $\overline{D} > 0$ and we restrict our attention to the interesting range of values $D \in (0, \overline{D})$. Recall [90], [54] that for any such D, $R(D)$ can alternatively be written as

$$R(D) = \inf_{Q} R_1(P, Q, D)$$

where the infimum is over all probability distributions Q on $\hat{A}$. Since we take $\hat{A}$ to be finite, this infimum is always achieved (see [54]) by a probability distribution $Q = Q^*$, although Q^* may not be unique. To avoid cumbersome notation in the statements of the coding theorems given next and also in later parts of the paper, we also write $\mathcal{R}(D)$ for the rate-distortion function of the source $\boldsymbol{X}$ expressed in *bits* rather than in nats

$$\mathcal{R}(D) \triangleq (\log_2 e) R(D).$$

Finally, we write Q_n^* for the product measures $(Q^*)^n$ and, although as mentioned Q^* may not be unique, with a slight abuse of terminology we call $\{Q_n^*\}$ the *optimal reproduction distributions at distortion level* D.

Combining Theorem 9 with the generalized AEP of Theorem 1 implies the following strengthened direct coding theorem.

Theorem 10—Pointwise Coding Theorem for i.i.d. Sources [54]: Let $\boldsymbol{X}$ be an i.i.d. source with distribution P on A, and let Q_n^* denote the optimal reproduction distributions at distortion level $D \in (0, \overline{D})$. Then the codes based on almost any realization of the Shannon random codebooks according to the measures $\{Q_n^*\}$ have code lengths $\ell_n(X_1^n)$ satisfying

$$\lim_{n \to \infty} \frac{1}{n} \ell_n(X_1^n) = \mathcal{R}(D) \quad \text{bits per symbol, w.p. 1.}$$

A simple modification of the above scheme can be used to obtain *universal* codebooks that achieve optimal compression for any memoryless source. When the source distribution is not known so that we have no way of knowing *a priori* the optimal reproduction distribution Q^*, we generate *multiple codebooks* according to an asymptotically dense set of probability measures. Specifically, given a fixed block length n, we consider the collection of all n-types on $\hat{A}$, namely, all distributions Q of the form $Q(\hat{a}) = j/n$, $0 \leq j \leq n$, for $\hat{a} \in \hat{A}$. Instead of generating a single random codebook according to the optimal distribution Q_n^*, we generate a different codebook for each product measure Q^n corresponding to an n-type Q on $\hat{A}$. Then we (as the encoder) adopt a greedy coding strategy. We find the first D-close match for X_1^n in each of the codebooks, and pick the one in which the match appears the earliest. To describe X_1^n to the decoder with distortion D or less we then describe two things: a) the index of the codebook in which the earliest match was found, and b) the position i_n of this earliest match. Since there are at most polynomially many n-types (cf. [25], [24]), the rate of the description of a) is asymptotically negligible. Moreover, since the set of n-types is asymptotically dense among probability measures on $\hat{A}$, we eventually do as well as if we were using the optimum codebook distribution Q_n^*.

Theorem 11—Pointwise Universal Coding Theorem [54]: Let $\boldsymbol{X}$ be an arbitrary i.i.d. source with distribution P on A, let $R(D)$ be the rate-distortion function of this source at distortion level $D \in (0, \overline{D})$, and let $\mathcal{R}(D)$ denote its rate-distortion function in bits. The codes based on almost any realization of the universal Shannon random codebooks have code lengths $\ell_n(X_1^n)$ satisfying

$$\lim_{n \to \infty} \frac{1}{n} \ell_n(X_1^n) = \mathcal{R}(D) \quad \text{bits per symbol, w.p. 1.}$$

B. Mismatched Codebooks

In the preceding subsection we described how, for memoryless sources, the Shannon random codebooks with respect to the optimal reproduction distributions can be used to achieve asymptotically optimal compression performance. In this subsection, we briefly consider the question of determining the rate achieved when an arbitrary (stationary-ergodic) source $\boldsymbol{X}$ is encoded using a random codebook according to the i.i.d. distributions Q^n for an arbitrary distribution Q on $\hat{A}$. For further discussion of the problem of mismatched codebooks see [72], [73], [58], [43], [44], and the references therein. Also see [94] for an application of the generalized AEP to a different version of the mismatched-codebooks problem.

The following theorem is an immediate consequence of combining Theorem 1 with Theorem 9 and the discussion in Section III-A (see also Example 1 in Section II-B).

Theorem 12—Mismatched Coding Rate: Let $\boldsymbol{X}$ be a stationary-ergodic process with marginal distribution $P_1 = P$ on A, let Q be an arbitrary distribution on $\hat{A}$, and define $D_{\min}$ and D_{av} as in Section II-B.

a) *Arbitrary i.i.d. Codebooks:* For any distortion level $D \in (D_{\min}, D_{\text{av}})$, the codes based on almost any realization of the Shannon random codebooks according to the measures $\{Q^n\}$ have code lengths $\ell_n(X_1^n)$ satisfying

$$\lim_{n \to \infty} \frac{1}{n} \ell_n(X_1^n) = (\log_2 e) R_1(P, Q, D)$$

bits per symbol, w.p. 1.

b) *i.i.d. Gaussian Codebooks:* Suppose $\rho(x, y) = (x - y)^2$ and $\boldsymbol{X}$ is a real-valued process with finite variance $\sigma^2 = \mathrm{Var}(X_1)$. Let Q be the $N(0, \tau^2)$ distribution on $\mathbb{R}$. Then for any distortion level $D \in (0, \sigma^2+\tau^2)$, the codes based on almost any realization of the Gaussian codebooks according to the measures $\{Q^n\}$ have code lengths $\ell_n(X_1^n)$ satisfying

$$\lim_{n\to\infty} \frac{1}{n}\,\ell_n(X_1^n) = \frac{1}{2}\log_2\left(\frac{v}{D}\right) - (\log_2 e)\frac{(v-D)(v-\sigma^2)}{2v\tau^2} \quad \text{bits per symbol, w.p. 1}$$

where

$$v \triangleq \frac{1}{2}\left[\tau^2 + \sqrt{\tau^4 + 4D\sigma^2}\right].$$

Lossless Versus Lossy Mismatch: Recall that, in the case of lossless data compression, if instead of the true source distribution P a different coding distribution Q is used, then the coderate achieved is

$$H(P) + H(P\|Q). \tag{25}$$

Similarly, in the current setting of lossy data compression, if instead of the optimal reproduction distribution Q^* we use a different codebook distribution Q, the rate we achieve is $R_1(P, Q, D)$. An upper bound for $R_1(P, Q, D)$ is obtained by taking (X, Y) in the expression of Remark 1 to be the jointly distributed random variables that achieve the infimum in the definition of the rate-distortion function of P. Then the (mismatched) rate of the random code based on Q instead of Q^* is

$$R_1(P, Q, D) \le R(D) + H(Q^*\|Q). \tag{26}$$

Equations (25) and (26) illustrate the analogy between the penalty terms in the lossless and lossy case due to mismatch.

Next we discuss two special cases of part b) of the theorem that are of particular interest.

Example 2—Gaussian Codebook With Mismatched Distribution: Consider the following coding scenario. We want to encode data generated by an i.i.d. Gaussian process with $N(0, \sigma^2)$ distribution, with squared-error distortion D or less. In this case, it is well known [9], [24] that for any $D \in (0, \sigma^2)$ the optimal reproduction distribution Q^* is the $N(0, \sigma^2 - D)$ distribution, so we construct random codebooks according to the i.i.d. distributions $Q_n^* = (Q^*)^n$.

But suppose that, instead of an i.i.d. Gaussian, the source turns out to be some arbitrary stationary-ergodic $\boldsymbol{X}$ with zero mean and variance σ^2. Theorem 12 b) implies that the asymptotic rate achieved by our i.i.d. Gaussian codebook is equal to

$$\frac{1}{2}\log_2\left(\frac{\sigma^2}{D}\right) \qquad \text{bits per symbol.}$$

Since this is exactly the rate-distortion function of the i.i.d. $N(0, \sigma^2)$ source, we conclude that the rate achieved is the same as what we would have obtained on the Gaussian source we originally expected. This offers yet another justification of the folk theorem that the Gaussian source is the hardest one to compress, among sources with a fixed variance. In fact, the above result is a natural fixed-distortion analog of [58, Theorem 3].

Example 3—Gaussian Codebook With Mismatched Variance: Here we consider a different type of mismatch. As before, we are prepared to encode an i.i.d. Gaussian source, but we have an incorrect estimate of its variance, say $\hat{\sigma}^2$ instead of the true variance σ^2. So we are using a random codebook with respect to the optimal reproduction distribution $Q_n^* = (Q^*)^n$, where Q^* is the $N(0, \hat{\sigma}^2 - D)$ distribution, but the actual source is i.i.d. $N(0, \sigma^2)$. In this case, the rate achieved by the random codebooks according to the distributions Q_n^* is given by the expression in Theorem 12 b), with τ^2 replaced by $\hat{\sigma}^2 - D$. Although the resulting expression is somewhat long and not easy to manipulate analytically, it is straightforward to evaluate numerically. For example, Fig. 1 shows the asymptotic rate achieved, as a function of the error $e = \sigma^2 - \hat{\sigma}^2$ in the estimate of the true variance. As expected, the best rate is achieved when the codebook distribution is matched to the source (corresponding to $e = 0$), and it is equal to the rate- distortion function of the source. Moreover, as one might expect, it is more harmful to underestimate the variance than to overestimate it.

C. Waiting Times and Idealized Lempel–Ziv Coding

Given $D \ge 0$ and two independent realizations from the stationary-ergodic processes $\boldsymbol{X}$ and $\boldsymbol{Y}$, our main quantity of interest here is the *waiting time* $W_n = W_n(D)$ until a D-close version of the initial string X_1^n first appears in Y_1^∞. Formally

$$W_n = \inf\{i \ge 1\colon \rho_n(X_1^n, Y_i^{i+n-1}) \le D\} \tag{27}$$

with the convention, as before, that the infimum of the empty set equals $+\infty$.

The motivation for studying the asymptotic behavior of W_n for large n is twofold.

Idealized Lempel–Ziv Coding: The natural extension of the idealized scenario described in the Introduction is to consider a message X_1^n that is to be encoded with the help of a database Y_1^∞. The source and the database are assumed to be independent, and the database distribution may or may not be the same as that of the source. In order to communicate X_1^n to the decoder with distortion D or less, the encoder simply describes W_n, using no more than

$$\log_2 W_n + O(\log_2 \log_2 W_n) \quad \text{bits.}$$

Therefore, the asymptotic performance of this idealized scheme can be completely understood in terms of the asymptotic of $\log W_n$, for large n.

DNA Pattern Matching: Here we imagine that X_1^n represents a DNA or protein "template," and we want to see whether it appears, either exactly or approximately, as a contiguous substring of a database DNA sequence Y_1^∞. We are interested in quantifying the "degree of surprise" in the fact that a D-close match was found at position W_n. Specifically, was the match found "atypically" early, or is the value of W_n consistent with the hypothesis that the template and the database are independent? For a detailed discussion, see, e.g., [29, Sec. 3.2], [45], [3], [2], [4], and the references therein.

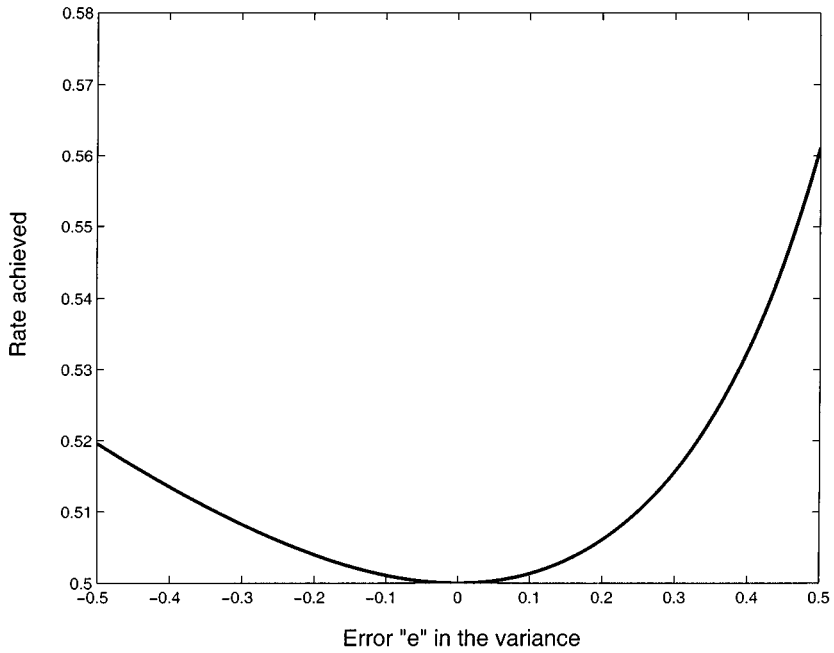

Fig. 1. This graph shows the rate achieved by an i.i.d. Gaussian codebook of variance $\hat{\sigma}^2 - D$ when applied to i.i.d. $N(0,\ \sigma^2)$ data. The rate is shown as a function of the error $e = \sigma^2 - \hat{\sigma}^2$ in the variance estimate. In this particular example, $\sigma^2 = 2$, $D = 1$, the error e ranges from $-1/2$ to $1/2$, and the rate-distortion function of the source equals 0.5 bit/symbol.

If for a moment we consider the case when both $\boldsymbol{X}$ and $\boldsymbol{Y}$ are i.i.d., we see that the waiting time W_n is, at least intuitively, closely related to the index i_n of Section III-A. As the following result shows, although the distribution of W_n is not exactly geometric, W_n behaves very much like i_n, at least in the exponent. That is, the difference

$$\log W_n - [-\log Q_n(B(X_1^n,\ D))]$$

is "small," eventually with probability one.

Recall the definition of ψ-mixing from Section II-C, and also the definition of the ϕ-mixing coefficients of $\boldsymbol{Y}$

$$\phi(k) = \sup\{|\mathbb{Q}(B|A) - \mathbb{Q}(B)|: \\ B \in \sigma(Y_k^\infty),\ A \in \sigma(Y_{-\infty}^0),\ \mathbb{Q}(A) > 0\}$$

where, as before, $\sigma(Y_i^j)$ denotes the σ-field generated by Y_i^j. The process $\boldsymbol{Y}$ is called *ϕ-mixing* if $\phi(k) \to 0$ as $k \to \infty$; see [12] for an extensive discussion of ϕ-mixing and related mixing conditions.

Theorem 13—Strong Approximation [51], [27]: Let $\boldsymbol{X}$ and $\boldsymbol{Y}$ be stationary-ergodic processes, and assume that $\boldsymbol{Y}$ is either ψ-mixing or ϕ-mixing with summable ϕ-mixing coefficients, $\sum_{k\geq 1}\phi(k) < \infty$. If $Q_n(B(X_1^n,\ D)) > 0$ eventually with probability one, then for any $\epsilon > 0$

$$-(1+\epsilon)\log n \leq \log[W_n Q_n(B(X_1^n,\ D))] \\ \leq (2+\epsilon)\log n \quad \text{eventually, w.p. 1.}$$

Theorem 13, of course, implies that

$$\log W_n = -\log Q_n(B(X_1^n,\ D)) + O(\log n) \quad \text{w.p. 1} \quad (28)$$

and combining this with the generalized AEP statements of Theorems 1 and 4 we immediately obtain the first-order (or strong-law-of-large-numbers, SLLN) asymptotic behavior of the waiting times W_n:

Theorem 14—SLLN for Waiting Times: Let $\boldsymbol{X}$ and $\boldsymbol{Y}$ be stationary-ergodic processes.

a) If $\boldsymbol{Y}$ is i.i.d. and the average distortion D_{av} is finite, then for any $D \in (D_{\min},\ D_{\text{av}})$

$$\frac{1}{n}\log W_n \to R_1(P_1,\ Q_1,\ D) \quad \text{w.p. 1.} \quad (29)$$

b) If $\boldsymbol{Y}$ is $\psi^{\pm}$-mixing and the distortion measure ρ is essentially bounded, i.e., $D_{\max} < \infty$, then for any $D \in (D_{\min},\ D_{\text{av}})$

$$\frac{1}{n}\log W_n \to R(\mathbb{P},\ \mathbb{Q},\ D) \quad \text{w.p. 1.} \quad (30)$$

Note that similar results can be obtained under different assumptions on the process $\boldsymbol{Y}$, using Theorems 3 and 5 in place of Theorems 1 and 4 as done above. When $\boldsymbol{X}$ is taken to be an arbitrary stationary-ergodic process, it is natural to expect that the mixing conditions for $\boldsymbol{Y}$ in Theorem 14 b) cannot be substantially relaxed. In fact, even in the case of exact matching between finite-alphabet processes, Shields [76] has produced a counterexample demonstrating that the analog of Theorem 13 does not hold for arbitrary stationary-ergodic $\boldsymbol{Y}$.

Historical Remarks: Waiting times in the context of lossy data compression were studied by Steinberg and Gutman [77] and Łuczak and Szpankowski [60]. Yang and Kieffer [88] identified the limiting rate-function for a wide range of finite alphabet sources, and Dembo and Kontoyiannis [27] and Chi [20] generalized these results to processes with general alphabets.

The strong approximation idea was introduced in [51] in the case of exact matching. For processes $\boldsymbol{Y}$ with summable ϕ-mixing coefficients, Theorem 13 was proved in [27], and when $\boldsymbol{Y}$ is ψ-mixing it was proved, for the case of no distortion, in [51]. Examining the latter proof, Chi [20] observed that it immediately generalizes to the statement of Theorem 13.

Related results were obtained by Kanaya and Muramatsu [42], who extended some of the results of [77] to processes with general alphabets, and by Koga and Arimoto [49] who considered *nonoverlapping* waiting times between finite-alphabet processes and Gaussian processes. Finally, Shields [76] and Marton and Shields [61] considered waiting times with respect to Hamming distortion and for $\boldsymbol{X}$ and $\boldsymbol{Y}$ having the same distribution over a finite alphabet. For the case of small distortion they showed, under some conditions, that approximate matching results like (29) and (30) reduce to their natural exact matching analogs as $D \to 0$.

D. Match Lengths and Practical Lempel–Ziv Coding

In the idealized coding scenario of the preceding subsection, we considered the case where a fixed-length message X_1^n is to be compressed using an infinitely long database Y_1^∞. But, in practice, the reverse situation is much more common. We typically have a "long" message $(X_1, X_2, \cdots)$ to be compressed, and only a finite-length database Y_1^m is available to the encoder and decoder. It is therefore natural (following the corresponding development in the case of lossless compression) to try and match "as much as possible" from the message $(X_1, X_2, \cdots)$ into the database Y_1^m. With this in mind we define the *match-length* L_m as the length ℓ of the longest prefix X_1^ℓ that matches somewhere in the database with distortion D or less

$$L_m = \sup\{\ell \geq 1\colon \rho_\ell(X_1^\ell, Y_j^{j+\ell-1}) \leq D, \quad \text{for some } j = 1, 2, \ldots, m\}. \tag{31}$$

Intuitively, there is a connection between match lengths and waiting times. Long matches should mean short waiting times, and *vice versa*. In the case of exact matching, this connection was precisely formalized by Wyner and Ziv [84], who observed that the following "duality" relationship always holds:

$$W_n \leq m \Leftrightarrow L_m \geq n. \tag{32}$$

This is almost identical to the standard relationship in renewal theory between the number of events by a certain time and the time of the nth event (see, e.g., [35]). Wyner and Ziv [84] utilized (32) to translate their first-order asymptotic results about W_n to corresponding results about L_m.

Unfortunately, this simple relationship no longer holds in the case of *approximate* matching, when a distortion measure is introduced. Instead, the following modified duality was observed in [60] and employed in [27] to obtain corresponding results in approximate matching and lossy data compression:

$$W_n \leq m \Rightarrow L_m \geq n \text{ and } L_m \geq n \Rightarrow \inf_{k \geq n} W_k \leq m. \tag{33}$$

In [27], it is shown that (33) can be used to deduce the asymptotic behavior of L_m from that of W_n, but this translation is not straightforward anymore. In fact, as we discuss in Section V-B, a more delicate analysis is needed in this case. Nevertheless, once the behavior of the waiting times is understood, the first implication in (33) immediately yields asymptotic *lower bounds* on the behavior of the match lengths. This is significant for data compression since long match lengths usually mean good compression performance. Indeed, this observation allowed Kontoyiannis [53] to introduce a new lossy version of the Lempel–Ziv algorithm that achieves asymptotically optimal compression performance for memoryless sources. The key characteristics of the algorithm are that it has polynomial implementation complexity, while at the same time it achieves redundancy comparable to that of its lossless counterpart, the FDLZ [85]. Note that the main issue of practical interest here is not simply the encoding complexity, but rather the tradeoff between complexity and redundancy. For example, the encoding complexity can be made arbitrarily small by using a very slowly growing (yet asymptotically dense) set of database distributions, but in that case, the redundancy rate of the algorithm would also be extremely slow.

In terms of practical algorithms, the utility of pattern-matching-base methods has been extensively studied; see [5], [1], and the references therein. A different approach to using pattern matching for adaptive lossy compression was introduced in [95], [96].

We also mention that, before [53], several practical (yet suboptimal) lossy versions of the Lempel–Ziv algorithm were introduced, perhaps most notably by Steinberg and Gutman [77] and Łuczak and Szpankowski [60]. Roughly speaking, the reason for their suboptimal compression performance was that the coding was done with respect to a database that had the same distribution as the source. In view of the discussion in the previous section, it is clear that the asymptotic code-rate of these algorithms is $R_1(P, P, D)$, which is typically significantly larger than the optimal $R(D) = \inf_Q R_1(P, Q, D)$; see [88] or [53] for more detailed discussions.

E. Sphere-Covering and Weighted Codebooks

Finally, we briefly describe a related question that was recently considered in [55]. In the classical rate-distortion problem, one is interested in finding "efficient" codebooks for describing the output X_1^n of some random source $\boldsymbol{X} = \{X_n\}$ to within some tolerable distortion level. In terms of data compression, a codebook is efficient when it contains relatively few codewords. Here, we are interested in the more general problem of finding codebooks with small "mass." Let $M\colon A \to (0, \infty)$ be an arbitrary nonnegative function assigning mass $M^n(C_n)$ to subsets C_n of A^n

$$M^n(C_n) \triangleq \sum_{y_1^n \in C_n} M^n(y_1^n) \triangleq \sum_{y_1^n \in C_n} \prod_{i=1}^n M(y_i).$$

The question of interest can be stated as follows. Let C_n be a subset A^n (we think of C_n as the codebook) that nearly D-covers all of A^n, i.e., with high probability, every string X_1^n generated by the source will match at least one element of C_n with distortion D or less

$$\Pr\{\text{there is an } y_1^n \in C_n \text{ such that } \rho_n(X_1^n, y_1^n) \leq D\} \approx 1. \tag{34}$$

If (34) holds, how small can the mass of C_n be?

This question is motivated, in part, by the fact that a number of important statistical problems can be restated in this framework. For example, taking M identically equal to one, this problem reduces to the rate-distortion question. Taking M to be a different probability measure Q, it reduces to the classical hypothesis testing question, whereas $M = P$ (the source distribution) yields "converses" to some measure-concentration inequalities.

A precise answer to this question is offered in [55], where a single-letter characterization is given for the best achievable exponential rate at which $M^n(C_n)$ can grow, among all codebooks C_n satisfying (34). With different choices for M and the distortion measure on A, this result gives various corollaries as special cases, including the classical rate-distortion theorem, Stein's lemma in hypothesis testing, and a new converse to some measure-concentration inequalities on discrete spaces.

Once again, the main ingredient in the proof of the corresponding direct coding theorem in [55] is provided by yet another version of the generalized AEP.

IV. Refinements of the Generalized AEP

As we saw in Section III, the generalized AEP can be used to determine the first-order asymptotic behavior of a number of interesting objects arising in applications. For example, the generalized AEP of Theorem 1

$$-\frac{1}{n}\log Q^n(B(X_1^n, D)) \to R_1(P, Q, D) \quad \text{w.p. 1}$$

immediately translated (via the strong approximation of Theorem 13) to an SLLN result for the waiting times

$$\frac{1}{n}\log W_n \to R_1(P, Q, D) \quad \text{w.p. 1.}$$

In this section, we will prove refinements to the generalized AEP of Section II-B, and in Section V we will revisit the applications of the previous section and use these refinements to prove corresponding second-order asymptotic results.

To get some motivation, let us consider for a moment the simplest version of the classical AEP, for an i.i.d. process $\boldsymbol{X}$ with distribution P on the finite alphabet A. The AEP here follows by a simple application of the law of large numbers

$$-\frac{1}{n}\log P^n(X_1^n) = \frac{1}{n}\sum_{i=1}^{n}[-\log P(X_i)] \to H \tag{35}$$

where H is the entropy of P. But (35) contains more information than that: it says that $-\log P^n(X_1^n)$ is, in fact, equal to the partial sum $S_n = \sum_{i=1}^n Z_i$ of the i.i.d. random variables $Z_i = -\log P(X_i)$. Therefore, we can apply the central limit theorem (CLT) or the law of the iterated logarithm (LIL) to get more precise information on the convergence of the AEP.

The same strategy can be carried out for non-i.i.d. processes. Initially, Ibragimov [40] and then Philipp and Stout [71] showed that even when $\boldsymbol{X}$ is a Markov chain, or, more generally, a weakly dependent random process, the quantities $-\log P^n(X_1^n)$ can be approximated by the partial sums of an associated weakly dependent process. These results have found a number of applications in lossless data compression and related areas [51], [50].

In this and the following sections, we will carry out a similar program in the lossy case. The main idea will be to show that, in analogy with the lossless case, the quantities $-\log Q^n(B(X_1^n, D))$ are asymptotically close to the partial sums of a function of the X_i, i.e.,

$$-\frac{1}{n}\log Q^n(B(X_1^n, D)) \approx R_1(P, Q, D) + \frac{1}{n}\sum_{i=1}^{n} g(X_i).$$

See Corollary 17 for the precise statement.

Throughout this section, we will adopt the notation and assumptions of Section II-B. Let $\boldsymbol{X}$ be a stationary-ergodic process with first-order marginal $P_1 = P$ on A, and let Q be an arbitrary probability measure on $\hat{A}$. Define $D_{\min}$ and D_{av}, as before (as in (7) and (8)), and assume that $D_{\min} < D_{\text{av}}$ so that the distortion measure $\rho(X, Y)$ is not essentially constant in Y with positive probability. We also impose here the additional assumption that ρ has a finite third moment

$$D_3 \triangleq E_{P\times Q}[\rho^3(X, Y)] < \infty. \tag{36}$$

The first result of this section refines Theorem 1 by giving a more precise asymptotic estimate of the quantity

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

in terms of the rate-function $R_1(P, Q, D)$ and the empirical measure $\hat{P}_n$ induced by X_1^n on A^n

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

where δ_x denotes the measure assigning unit mass to $x \in A$.

Theorem 15 [90]: Let $\boldsymbol{X}$ be a stationary-ergodic process with marginal P on A, and let Q be an arbitrary probability measure on $\hat{A}$. Assume that $D_3 = E_{P\times Q}[\rho^3(X, Y)]$ is finite. Then, for any $D \in (D_{\min}, D_{\text{av}})$

$$-\log Q^n(B(X_1^n, D)) = nR_1(\hat{P}_n, Q, D) + \frac{1}{2}\log n + O(1) \quad \text{w.p. 1.} \tag{37}$$

Next we show that the most significant term in (37) can be approximated by the partial sum of a weakly dependent random process. Recall the definition of the α-mixing coefficients of $\boldsymbol{X}$

$$\alpha(k) = \sup\{|\mathbb{P}(A\cap B) - \mathbb{P}(A)\mathbb{P}(B)|: A \in \sigma(X_{-\infty}^0),\ B \in \sigma(X_k^\infty)\}$$

where $\sigma(X_i^j)$ is the σ-field generated by X_i^j. The process $\boldsymbol{X}$ is called *α-mixing* if $\alpha(k) \to 0$ as $k \to \infty$; see [12] for more details.

We also need to recall some of the notation from the proof of Theorem 1 in Section II-B. For $x \in A$ and $\lambda \in \mathbb{R}$, let $\Lambda_x(\lambda)$ denote the log-moment generating function of the random variable $\rho(x, Y)$

$$\Lambda_x(\lambda) \triangleq \log E_Q\left(e^{\lambda\rho(x, Y)}\right)$$

and note that the function $\Lambda(\lambda)$ defined in (10) can be written as $\Lambda(\lambda) = E_P[\Lambda_X(\lambda)]$. Also, recall that for any

$D \in (D_{\min}, D_{\text{av}})$ there exists a unique $\lambda^* < 0$ such that $\Lambda'(\lambda^*) = D$.

Theorem 16 [27]: Let $\boldsymbol{X}$ be a stationary α-mixing process with marginal P on A, and let Q be an arbitrary probability measure on $\hat{A}$. Assume that the α-mixing coefficients of $\boldsymbol{X}$ satisfy

$$\sum_{k=1}^{\infty} \alpha^t(k) < \infty, \qquad \text{for some } t \in (0, 1/3) \tag{38}$$

and that

$$D_3 = E_{P\times Q}[\rho^3(X, Y)]$$

is finite. Then for any $D \in (D_{\min}, D_{\text{av}})$

$$nR_1(\hat{P}_n, Q, D) = nR_1(P, Q, D) + \sum_{i=1}^{n} g(X_i) + O(\log\log n) \quad \text{w.p. 1}$$

where

$$g(x) \stackrel{\Delta}{=} \Lambda(\lambda^*) - \Lambda_x(\lambda^*), \qquad x \in A. \tag{39}$$

Theorem 16 is a small generalization of [27, Theorem 3]. Before giving its proof outline, we combine Theorems 15 and 16 to show that, as promised, $-\log Q^n(B(X_1^n, D))$ can be accurately approximated as the partial sum of the weakly dependent random process $\{g(X_n)\}$.

Corollary 17—Second-Order Generalized AEP: Let $\boldsymbol{X}$ be a stationary α-mixing process with marginal P on A, and let Q be an arbitrary probability measure on $\hat{A}$. Assume that the α-mixing coefficients of $\boldsymbol{X}$ satisfy (38) and that

$$D_3 = E_{P\times Q}[\rho^3(X, Y)]$$

is finite. Then for any $D \in (D_{\min}, D_{\text{av}})$, and with $g(x)$ defined as in (39)

$$-\log Q^n(B(X_1^n, D)) = nR_1(P, Q, D) + \sum_{i=1}^{n} g(X_i) + \frac{1}{2}\log n + O(\log\log n) \quad \text{w.p. 1.}$$

Proof Outline for Theorem 16: Adapting the argument leading from [27, eqs. (22)–(24)], one easily checks that the result of Theorem 16 holds as soon as

$$\liminf_{n\to\infty} \inf_{|\theta|<\delta} B_n(\theta) > 0 \quad \text{w.p. 1} \tag{40}$$

and

$$\limsup_{n\to\infty} \frac{nA_n^2}{\log\log n} < \infty \quad \text{w.p. 1} \tag{41}$$

where $A_n = n^{-1}\sum_{k=1}^{n} \zeta_k$ is the empirical mean of the centered random variables $\zeta_k = \Lambda'_{X_k}(\lambda^*) - D$, and $B_n(\theta)$ is the empirical mean of the nonnegative random variables $\Lambda''_{X_k}(\lambda^* + \theta)$. By the ergodic theorem we have, with probability one

$$\liminf_{n\to\infty} \inf_{|\theta|<\delta} B_n(\theta) \geq \liminf_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} \inf_{|\theta|<\delta} \Lambda''_{X_k}(\lambda^* + \theta) = E_P\left[\inf_{|\theta|<\delta} \Lambda''_X(\lambda^* + \theta)\right]$$

and by Fatou's lemma and the continuity of the map $\theta \mapsto \Lambda''_x(\lambda^* + \theta)$ it follows that

$$\liminf_{\delta\downarrow 0} E_P\left[\inf_{|\theta|<\delta} \Lambda''_X(\lambda^*+\theta)\right] \geq E_P[\Lambda''_X(\lambda^*)] = \Lambda''(\lambda^*) > 0.$$

This implies that (40) holds once $\delta > 0$ is made small enough. (Note that the above argument also avoids an incorrect—but also unnecessary—application of the uniform ergodic theorem in the derivation of [27, eq. (26)].)

Turning to (41), since $\lambda^* < 0$, it follows by the convexity of $\Lambda_x(\lambda)$ that for any $x \in A$

$$0 \leq \Lambda'_x(\lambda^*) \leq \Lambda'_x(0) = E_Q[\rho(x, Y)].$$

Consequently, Hölder's inequality and assumption (36) imply that the random variable

$$|\zeta_k| \leq E_Q[\rho(X_k, Y)|X_k] + D$$

has a finite third moment. Recall [64] that the LIL holds for the partial sum A_n of a zero-mean, stationary process $\{\zeta_k\}$ with a finite third moment, as soon as the α-mixing coefficients of $\{\zeta_k\}$ satisfy (38). The observation that ζ_k is a deterministic function of X_k for all k completes the proof. □

V. Applications: Second-Order Results

Here we revisit the applications considered in Section III, and using the "second-order generalized AEP" of Corollary 17 we prove second-order refinements for many of the results from Section III. In Section V-A, we consider the problem of lossy data compression in the same setting as in Section III-A. We use the second-order AEP to determine the precise asymptotic behavior of the Shannon random codebooks, and show that, with probability one, they achieve optimal compression performance up to terms of order $(\log n)$ bits. Moreover, essentially the same compression performance can be achieved universally. For arbitrary variable-length codes operating at a fixed rate level, we show that the rate at which they can achieve the optimal rate of $n\mathcal{R}(D)$ bits is at best of order $O(\sqrt{n})$ bits. This is the best possible redundancy rate as long as the "minimal coding variance" of the source is strictly positive. For discrete i.i.d. sources, a characterization is given of when this variance can be zero.

In Section V-B, we look at waiting times, and we prove a second-order refinement to Theorem 14, and in Section V-C, we consider the problem of determining the asymptotic behavior of longest match lengths. As discussed briefly in Section III-D, their asymptotic can be deduced from the corresponding waiting-times results via duality.

A. Lossy Data Compression

1) Random Codes and Second-Order Converses: Here we consider the exact same setup as in Section III-A. An i.i.d. source $\boldsymbol{X}$ with distribution P on A is to be compressed with distortion D or less with respect to a bounded distortion measure ρ, satisfying, as before, the usual assumption (23)—see the remark at the end of this section for its implications. We take the reproduction alphabet $\hat{A}$ to be finite, define $\overline{D}$ as in (24), and assume that $\overline{D} > 0$.

For $D \in (0, \overline{D})$, let Q_n^*, $n \geq 1$, denote the optimal reproduction distributions at distortion level D. Combining the strong approximation Theorem 9 with the second-order generalized AEP of Corollary 17 and the discussion in Section III-A yields the following.

Theorem 18—Pointwise Redundancy for i.i.d. Sources [54]: Suppose $\boldsymbol{X}$ is an i.i.d. source with distribution P on A, and with rate-distortion function $\mathcal{R}(D)$ (in bits). Let Q_n^* denote the optimal reproduction distributions at distortion level $D \in (0, \overline{D})$, and define the function $h(x) = (\log_2 e) g(x)$, $x \in A$, with g defined as in (39). Then we get the following.

a) The codes based on almost any realization of the Shannon random codebooks according to the measures $\{Q_n^*\}$ have code lengths $\ell_n(X_1^n)$ satisfying

$$\ell_n(X_1^n) \leq n\mathcal{R}(D) + \sum_{i=1}^{n} h(X_i) + 4 \log n \text{ bits,}$$

eventually, w.p. 1.

b) The codes based on almost any realization of the universal Shannon random codebooks have code-lengths $\ell_n(X_1^n)$ satisfying

$$\ell_n(X_1^n) \leq n\mathcal{R}(D) + \sum_{i=1}^{n} h(X_i) + (4 + |\hat{A}|) \log n, \text{ bits}$$

eventually, w.p. 1.

We remark that the coefficients of the $(\log n)$ terms in a) and b) are not the best possible, and can be significantly improved; see [56] for more details.

Perhaps somewhat surprisingly, it turns out that the performance of the above random codes is optimal up to terms of order $(\log n)$ bits. Recall that a *code C_n operating at distortion level* $D \geq 0$ is defined by a triplet (B_n, ϕ_n, ψ_n) where

a) B_n is a subset of $\hat{A}^n$, called the *codebook*;

b) $\phi_n \colon A^n \to B_n$ is the *encoder*;

c) $\psi_n \colon B_n \to \{0, 1\}^*$ is a uniquely decodable map;

such that

$$\rho_n(x_1^n, \phi_n(x_1^n)) \leq D, \qquad \text{for all } x_1^n \in A^n.$$

The code lengths $\ell_n(X_1^n)$ achieved by such a code are simply

$$\ell_n(x_1^n) = \text{length of } [\psi_n(\phi_n(x_1^n))] \qquad \text{bits.}$$

Theorem 19—Pointwise Converse for i.i.d. Sources [54]: Let $\boldsymbol{X}$ be an i.i.d. source with distribution P on A, and let $\{C_n\}$ be an arbitrary sequence of codes operating at distortion level $D \in (0, \overline{D})$, with associated code lengths $\{\ell_n\}$. Then

$$\ell_n(X_1^n) \geq n\mathcal{R}(D) + \sum_{i=1}^{n} h(X_i) - \log n \text{ bits,}$$

eventually, w.p. 1

where $h(x)$ is defined as in Theorem 18.

The proof of Theorem 19 in [54] uses techniques quite different to those developed in this paper. In particular, the key step in the proof is established by an application of the generalized Kuhn–Tucker conditions of Bell and Cover [8].

Theorems 18 and 19 are next combined to yield "second-order" refinements to Shannon's classical source-coding theorem. For a source $\boldsymbol{X}$ as in Theorem 19 and a $D \in (0, \overline{D})$, the *minimal coding variance* $\sigma^2 = \sigma^2(P, D)$ *of source* P *at distortion level* D is

$$\sigma^2 = \sigma^2(P, D) \triangleq \text{Var}[h(X_1)] \tag{42}$$

with $h(x)$ as in Theorem 18.

Theorem 20—Second-Order Source-Coding Theorems [54]: Let $\boldsymbol{X}$ be an i.i.d. source with distribution P on A and with rate-distortion function $\mathcal{R}(D)$ (in bits). For $D \in (0, \overline{D})$

(CLT) There is a sequence of random variables $G_n = G_n(P, D)$ such that, for any sequence of codes $\{C_n, \ell_n\}$ operating at distortion level D, we have

$$\ell_n(X_1^n) - n\mathcal{R}(D) \geq \sqrt{n} G_n \text{ bits,}$$

eventually, w.p. 1 (43)

and the G_n converge in distribution to a Gaussian random variable

$$G_n \xrightarrow{\mathcal{D}} N(0, \sigma^2)$$

where $\sigma^2 = \sigma^2(P, D)$ is the minimal coding variance.

(LIL) With σ^2 as above, for any sequence of codes $\{C_n, \ell_n\}$ operating at distortion level D

$$\limsup_{n \to \infty} \frac{\ell_n(X_1^n) - n\mathcal{R}(D)}{\sqrt{2n \log \log n}} \geq \sigma \quad \text{w.p. 1}$$

$$\liminf_{n \to \infty} \frac{\ell_n(X_1^n) - n\mathcal{R}(D)}{\sqrt{2n \log \log n}} \geq -\sigma \quad \text{w.p. 1.}$$

$(\Rightarrow)$ Moreover, there exist codes $\{C_n, \ell_n\}$ operating at distortion level D, that asymptotically achieve equality *universally* in all these lower bounds.

Remark on Assumption (23): When the distortion measure does not satisfy assumption (23) [as, for example, when $\rho(x, y) = (x - y)^2$ with $A = \mathbb{R}$ and $\hat{A}$ a finite subset of $\mathbb{R}$], we can modify ρ to

$$\rho'(x, y) = \rho(x, y) - f(x)$$

with $f(x) = \min_{y \in \hat{A}} \rho(x, y)$, so that ρ' satisfies (23). Then, to generate codes operating at distortion level D with respect to ρ, we can construct random codebooks for as before but do the encoding with respect to $\rho'(x, y)$ at the *random* distortion level $D_n = D - E_{\hat{P}_n}(f(X))$. It is not hard to check that [27, Theorem 2] can be extended to apply when D is replaced by the sequence $\{D_n\}$. Since $D_n \to D - E_P(f(X))$ as $n \to \infty$, this results with the first-order approximation

$$-\frac{1}{n} \log Q_n^*(B(X_1^n, D_n)) \approx R_1^{\rho'}(\hat{P}_n, Q^*, D_n).$$

Simple algebra then shows that

$$R_1^{\rho'}(\hat{P}_n, Q^*, D_n) = R_1^{\rho}(\hat{P}_n, Q^*, D)$$

implying that all the results of Section V-A1 remain valid [despite the fact that ρ does not satisfy (23)], with the function $h(\cdot)$ taken in terms of the log-moment generating function $\Lambda_x(\lambda)$

of the *original* distortion measure ρ (and not that of the modified ρ').

2) Critical Behavior: In view of Theorems 18 and 19 above, the code lengths $\ell_n^*(X_1^n)$ of the best code operating at distortion level D have

$$\ell_n^*(X_1^n) \approx n\mathcal{R}(D) + \sum_{i=1}^{n} h(X_i) + O(\log n) \quad \text{bits}.$$

This reveals an interesting dichotomy in the behavior of the "pointwise" redundancy of the best code:

- either the minimal coding variance σ^2 (recall (42)) is nonzero, in which case the best rate at which optimality can be achieved is of order $\sqrt{n}$ bits by the CLT;
- or $\sigma^2 = 0$, and the best redundancy rate is of order $(\log n)$ bits (cf. [97]).

Under certain conditions, in this section we give a precise characterization of when each of these two cases can occur. Before stating it, we briefly discuss two examples to gain some intuition.

Example 4—Lossless Compression: Lossless data compression can be considered as an extreme case of lossy compression, where $\boldsymbol{X}$ is an i.i.d. source with distribution P on a finite set $A = \hat{A}$, and the distortion level D is set to zero. Here it is well known that (ignoring the integer length constraints) the best code is given by the idealized Shannon code $\ell_n(X_1^n) = -\log_2 P^n(X_1^n)$. In agreement with the upper and lower bounds of Theorems 19 and 20, here it is trivial to see that the code lengths of the Shannon code in fact satisfy

$$\ell_n(X_1^n) = n\mathcal{H}(P) + \sum_{i=1}^{n} h(X_i)$$

where $\mathcal{H}(P)$ is the entropy of P in bits, and with

$$h(x) \stackrel{\Delta}{=} -\log_2 P(x) - \mathcal{H}(P), \qquad x \in A.$$

When is $\sigma^2 = 0$? By its definition (42), σ^2 is zero if and only if the function $h(x)$ is constant over x, which, in this case, can only happen if $P(x)$ is constant over $x \in A$. Therefore, here $\sigma^2 = 0$ *if and only if the source has a uniform distribution over* A.

Example 5—Binary Source With Hamming Distortion: Consider the simplest nontrivial lossy example. Let $\boldsymbol{X}$ be an i.i.d. source with Bernoulli(p) distribution (for some $p \in (0, 1/2]$), let $A = \hat{A} = \{0, 1\}$, and take ρ to be Hamming distortion: $\rho(x, y) = |x - y|$. For $D \in (0, p)$, it is not hard to evaluate all the relevant quantities explicitly (see, e.g., [9, Example 2.7.1] or [24, Theorem 13.3.1]). In particular, the optimal reproduction distribution Q^* is Bernoulli(q), with $q = (p - D)/(1 - 2D)$, and our function of interest is

$$h(x) = -\log_2\left(\frac{P(x)}{1-D}\right) - E_P\left[-\log_2\left(\frac{P(X_1)}{1-D}\right)\right].$$

Recalling that the minimal coding variance is zero if and only if $h(x)$ is constant, from the above expression we see that, similarly to the previous example, also here $\sigma^2 = 0$ *if and only if the source has a uniform distribution*.

For discrete sources, the next result gives conditions under which the characterization suggested by these two examples remains valid. Suppose $A = \hat{A} = \{a_1, a_2, \ldots, a_k\}$ is a finite set, write ρ_{ij} for $\rho(a_i, a_j)$, and assume that ρ is symmetric and that $\rho_{ij} = 0$ if and only if $i = j$. We call ρ a *permutation distortion measure*, if all rows of the matrix $(\rho_{ij})_{i,\,j=1,\ldots,k}$ are permutations of one another.

Theorem 21—Variance Characterization [28]: Let $\boldsymbol{X}$ be a discrete source with distribution P and rate-distortion function $R(D)$. Assume that $R(D)$ is strictly convex over $(0, \overline{D})$. There are exactly two possibilities:

a) either $\sigma^2 = \sigma^2(P, D)$ is only zero for finitely many $D \in (0, \overline{D})$;
b) or $\sigma^2 = \sigma^2(P, D) \equiv 0$ for *all* $D \in (0, \overline{D})$, in which case P is the uniform distribution on A and ρ is a permutation distortion measure.

A general discussion of this problem, including the case of continuous sources, is given in [28]. Also, in the lossless case, the problem of characterizing when $\sigma^2 = 0$ for sources with memory is dealt with in [50].

Before moving on to waiting times and match lengths we mention that, in a somewhat similar vain, the problem of understanding the best *expected* redundancy rate in lossy data compression has also been recently considered in [97], [92], [89], [41].

B. Waiting Times

Next we turn to waiting times. Recall that, given $D \geq 0$ and two independent realizations of the stationary ergodic processes $\boldsymbol{X}$ and $\boldsymbol{Y}$, the waiting time W_n was defined as the time of the first appearance of X_1^n in $\boldsymbol{Y}$ with distortion D or less (see (27) for the precise definition). In Theorem 14, we gave conditions that identified the first-order limiting behavior of W_n. In particular, when $\boldsymbol{Y}$ is i.i.d., it was shown in Theorem 14 a) that, for $D \in (D_{\min}, D_{\text{av}})$

$$\frac{\log W_n}{n} \to R_1(P, Q, D) \quad \text{w.p. 1} \tag{44}$$

where P and Q are the first-order marginals of $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively.

The next result gives conditions under which the SLLN-type statement of (44) can be refined to a CLT and a LIL.

Theorem 22—CLT and LIL for Waiting Times: Let $\boldsymbol{X}$ be a stationary α-mixing process and $\boldsymbol{Y}$ be an i.i.d. process, with marginal distributions P and Q, on A and $\hat{A}$, respectively. Assume that the α-mixing coefficients of $\boldsymbol{X}$ satisfy (38) and that $D_3 = E_{P\times Q}[\rho^3(X, Y)]$ is finite. Then for any $D \in (D_{\min}, D_{\text{av}})$ the following series converges:

$$\sigma^2 \stackrel{\Delta}{=} E_P[g^2(X_1)] + 2\sum_{k=2}^{\infty} E_P[g(X_1)g(X_k)] \tag{45}$$

with $g(x)$ defined as in (39), and, moreover

(CLT) With $R_1 = R_1(P, Q, D)$

$$\frac{\log W_n - nR_1}{\sqrt{n}} \xrightarrow{\mathcal{D}} N(0, \sigma^2).$$

(LIL) The set of limit points of the sequence

$$\left\{\frac{\log W_n - nR_1}{\sqrt{2n\log\log n}}\right\}, \qquad n \geq 3$$

coincides with $[-\sigma, \sigma]$, with probability one.

Proof Outline: For a bounded distortion measure ρ, Theorem 22 was proved in [27]. To obtain the more general statement of the theorem, combine the strong approximation of Theorem 13 with the second-order AEP in Corollary 17 to get

$$\log W_n = nR_1(P, Q, D) + \sum_{i=1}^{n} g(X_i) + O(\log n) \quad \text{w.p. 1.} \tag{46}$$

Since $\boldsymbol{X}$ satisfies the mixing assumption (38), so does the process $\{g(X_n)\}$. Also, since $\lambda^* < 0$, the function $\Lambda_x(\lambda^*)$ is bounded above by zero, and by Jensen's inequality it is bounded below by $\lambda^* E_Q[\rho(x, Y)]$. Therefore,

$$|\Lambda_x(\lambda^*)| \leq |\lambda^*| E_Q[\rho(x, Y)]$$

and this, together with Hölder's inequality and the definition of $g(x)$, imply that $E_P[|g(X_1)|^3] < \infty$. Therefore, we can apply the CLT of [70, Theorem 1.7] to the process $\{g(X_n)\}$ in order to deduce the CLT part of the theorem from (46). Similarly, applying the LIL of [64] to $\{g(X_n)\}$, from (46) we get the LIL part of the theorem. □

Remark 5: When the variance σ^2 in (45) is positive, then the *functional* versions of the above CLT and LIL given in [27] still hold, under exactly the conditions of Theorem 22. (This follows by applying the functional CLT of [70, Theorem 1.7] and the functional LIL of [65, Theorem 1 (IV)].)

C. Match Lengths and Duality

We turn to our last application, match lengths. Recall that, given a distortion level $D \geq 0$ and two independent realizations of the processes $\boldsymbol{X}$ and $\boldsymbol{Y}$, the match length L_m is defined as the length ℓ of the longest prefix X_1^ℓ that appears (with distortion D or less) starting somewhere in the "database" Y_1^m. See (31) for the precise definition. As we briefly mentioned in Section III-D, there is a duality relationship between match lengths and waiting times. Roughly speaking, long matches mean short waiting times, and *vice versa*; see (33).

Although the relation (33) is not as simple as the duality (32) for exact matching, it is still possible to use (33) to translate the asymptotic results for W_n to corresponding results for L_m. These are given in Theorem 23 below. This translation, carried out in [27], is more delicate than in the case of exact matching. For example, in order to prove the CLT for the match lengths L_m one invokes the functional CLT for the waiting times (see Remark 5 and [27, proof of Theorem 4]).

Theorem 23—Match Lengths Asymptotic: Let $\boldsymbol{X}$ be a stationary process and $\boldsymbol{Y}$ be an i.i.d. process, with marginal distributions P and Q, on A and $\hat{A}$, respectively. Assume that $D_3 = E_{P\times Q}[\rho^3(X, Y)]$ is finite. Then for any $D \in (D_{\min}, D_{\text{av}})$ we have

$$\textbf{(LLN)} \qquad \frac{L_m}{\log m} \to \frac{1}{R_1} \quad \text{w.p. 1}$$

where $R_1 = R_1(P, Q, D)$. If, moreover, the α-mixing coefficients of $\boldsymbol{X}$ satisfy (38) and the variance σ^2 in (45) is nonzero, then, with $\tau^2 \triangleq \sigma^2 R_1^{-3}$, we have

$$\textbf{(CLT)} \qquad \frac{L_m - \frac{\log m}{R_1}}{\sqrt{\log m}} \xrightarrow{\mathcal{D}} N(0, \tau^2)$$

$$\textbf{(LIL)} \qquad \limsup_{m\to\infty} \frac{L_m - \frac{\log m}{R_1}}{\sqrt{2\log m \log\log\log m}} = \tau \quad \text{w.p. 1.}$$

The results of Theorem 23 were proved in [27] for any bounded distortion measure ρ. The slightly more general version stated above is proved in exactly the same way, using the results of Section IV in place of Theorems 2 and 3 of [27].

VI. Random Fields: First-Order Results

This and the following sections are devoted to generalizations of the results of Sections II–V to the case of random fields. Specifically, the role of the processes $\boldsymbol{X}$ and $\boldsymbol{Y}$ will now be played by stationary ergodic random fields $\boldsymbol{X} = \{X_u;\, u \in \mathbb{Z}^d\}$ and $\boldsymbol{Y} = \{Y_u;\, u \in \mathbb{Z}^d\}$. As we will see, many of the problems that we considered have natural analogs in this case, and the overall theme carries over. The generalized AEP and its refinement can be extended to random fields, and the corresponding questions in data compression and pattern matching can be answered following the same path as before.

A. Notation and Definitions

The following definitions and notation will remain in effect throughout Sections VI and VII.

We consider two random fields

$$\boldsymbol{X} = \{X_u;\, u \in \mathbb{Z}^d\} \quad \text{and} \quad \boldsymbol{Y} = \{Y_u;\, u \in \mathbb{Z}^d\}, \qquad d \geq 2$$

taking values in A and $\hat{A}$, respectively, and indexed by points $u = (u_1, u_2, \ldots, u_d)$ on the integer lattice $\mathbb{Z}^d$. As before, A and $\hat{A}$ are complete, separable metric spaces, equipped with their Borel σ-fields $\mathcal{A}$ and $\hat{\mathcal{A}}$, respectively. Let $\mathbb{P}$ and $\mathbb{Q}$ denote the (infinite-dimensional) measures of the entire random fields $\boldsymbol{X}$ and $\boldsymbol{Y}$. Unless explicitly stated otherwise, we always assume that $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent of each other.

Throughout the rest of the paper we will assume that $\boldsymbol{X}$ and $\boldsymbol{Y}$ are stationary and ergodic. To be precise, by that we mean that the Abelian group of translations $\{T_u\colon u \in \mathbb{Z}^d\}$ acts on both $(A^{\mathbb{Z}^d}, \mathcal{A}^{\mathbb{Z}^d}, \mathbb{P})$ and $(\hat{A}^{\mathbb{Z}^d}, \hat{\mathcal{A}}^{\mathbb{Z}^d}, \mathbb{Q})$ in a measure-preserving, ergodic manner; see [57] for a detailed exposition.

For $v, w \in \mathbb{Z}^d$, the distance between v and w is defined by

$$d(v, w) \triangleq \max_{1\leq i\leq d} |v_i - w_i|$$

and the distance between two subsets $V, W \subset \mathbb{Z}^d$ is

$$d(V, W) \triangleq \inf_{v\in V,\, w\in W} d(v, w).$$

Given $v, w \in \mathbb{Z}^d$, we let

$$[v, w] = \{u \in \mathbb{Z}^d\colon v_j \leq u_j \leq w_j \text{ for all } j\}$$

where $[v, w]$ is empty in case $v_j > w_j$ for some j.

We write $C(n)$ for the d-dimensional cube of side $n \geq 1$

$$C(n) = \{u \in \mathbb{Z}^d \colon 1 \leq u_j \leq n \text{ for all } j\}$$

and $[0, \infty)$ for the "infinite cube"

$$[0, \infty) = \{u \in \mathbb{Z}^d \colon u_j \geq 0 \text{ for all } j\}.$$

For an arbitrary subset $U \subset \mathbb{Z}^d$ we let $|U|$ denote its size; for example, $|C(n)| = n^d$. Also, for $U \subset \mathbb{Z}^d$ we write

$$X_U \triangleq \{X_u; u \in U\}$$

so that, in particular

$$X_{[0,\infty)} = \{X_u; u_j \geq 0 \text{ for all } j\}.$$

For $V \subset \mathbb{Z}^d$ and $u \in \mathbb{Z}^d$ we let $u + V$ denote the translate

$$u + V = \{u + v \colon v \in V\}.$$

For each $n \geq 1$, let P_n denote the marginal distribution of $X_{C(n)}$ on A^{n^d}, and similarly write Q_n for the distribution of $Y_{C(n)}$. Let $\rho \colon A \times \hat{A} \to [0, \infty)$ be an arbitrary nonnegative (measurable) function, and define a sequence of single-letter distortion measures $\rho_n \colon A^{n^d} \times \hat{A}^{n^d} \to [0, \infty)$, $n \geq 1$ by

$$\rho_n(x_{C(n)}, y_{C(n)}) \triangleq \frac{1}{n^d} \sum_{u \in C(n)} \rho(x_u, y_u), \quad x_{C(n)} \in A^{n^d}, \quad y_{C(n)} \in \hat{A}^{n^d}.$$

Given $D \geq 0$ and $x_{C(n)} \in A^{n^d}$, we write $B(x_{C(n)}, D)$ for the distortion-ball of radius D

$$B(x_{C(n)}, D) = \left\{ y_{C(n)} \in \hat{A}^{n^d} \colon \rho_n(x_{C(n)}, y_{C(n)}) \leq D \right\}.$$

B. Generalized AEP

It is well known that the classical AEP

$$-\frac{1}{n} \log P_n(X_1^n) \to H(\mathbb{P}) \quad \text{w.p. 1}$$

generalizes to the case of finite-alphabet random fields on $\mathbb{Z}^d$, as well as to other amenable group actions [68]. In this subsection, we give two versions of the generalized AEP of Theorems 1 and 4 to the case of random fields on $\mathbb{Z}^d$.

$\boldsymbol{Y}$ is i.i.d. In the notation of Section VI-A, we take $\boldsymbol{X}$ to be a stationary ergodic random field with first-order marginal $P_1 = P$, and $\boldsymbol{Y}$ to be i.i.d. with first-order marginal $Q_1 = Q$. We define $D_{\min}$ and D_{av} as in the one-dimensional case (recall (7) and (8)), and assume that $\rho(x, y)$ is not essentially constant for ($\mathbb{P}$-almost) all $x \in A$, that is, $D_{\min} < D_{\text{av}}$.

A simple examination of the proof of Theorem 1 shows that it extends *verbatim* to the case of random fields, with the only difference that instead of the usual ergodic theorem we now need to invoke the ergodic theorem for $\mathbb{Z}^d$ actions; see [57, Ch. 6]. We thus obtain the following.

Theorem 24—Generalized AEP When $\boldsymbol{Y}$ is i.i.d.: Let $\boldsymbol{X}$ be a stationary ergodic random field on $\mathbb{Z}^d$ and $\boldsymbol{Y}$ be i.i.d., with marginal distributions P and Q on A and $\hat{A}$, respectively. Assume that $D_{\text{av}} = E_{P \times Q}[\rho(X, Y)]$ is finite. Then for any $D \in (D_{\min}, D_{\text{av}})$

$$-\frac{1}{n^d} \log Q^{n^d}\left(B\left(X_{C(n)}, D\right)\right) \to R_1(P, Q, D) \quad \text{w.p. 1}$$

with the (one-dimensional) rate-function $R_1(P, Q, D)$ defined as in Theorem 1.

$\boldsymbol{Y}$ is not i.i.d. Let $\boldsymbol{X}$ and $\boldsymbol{Y}$ be stationary random fields and define D_{av} and $D_{\max}$ exactly as in the one-dimensional case (recall (8) and (14)). We assume that the distortion measure ρ is essentially bounded, i.e., $D_{\max} < \infty$, and define

$$D_{\min} \triangleq \sup_{n \geq 1} D_{\min}^{(n)} = \lim_{n \to \infty} D_{\min}^{(n)} \tag{47}$$

where

$$D_{\min}^{(n)} \triangleq E_{P_n}\left[\operatorname*{ess\,inf}_{Y_{C(n)} \sim Q_n} \rho_n\left(X_{C(n)}, Y_{C(n)}\right)\right]. \tag{48}$$

To see that the limit in (47) exists and equals the supremum, first note that $\{n^d D_{\min}^{(n)}\}$ is an increasing sequence, and that $D_{\min}^{(nk)} \geq D_{\min}^{(k)}$ for all $n, k \geq 1$. Now fix $k \geq 1$ arbitrary. Given $n \geq k$ we write $n = mk + r$ for some $0 \leq r \leq k - 1$, so that

$$n^d D_{\min}^{(n)} \geq (mk)^d D_{\min}^{(mk)} \geq (mk)^d D_{\min}^{(k)}.$$

Since $n/mk \to 1$ as $n \to \infty$, this implies that

$$\liminf_{n \to \infty} D_{\min}^{(n)} \geq D_{\min}^{(k)}.$$

Since k was arbitrary we are done.

Finally, we assume once again that the distortion measure ρ is not essentially constant, that is, $D_{\min} < D_{\text{av}}$. Our next result is the random fields analog of Theorem 4; it is proved in Appendix C.

Theorem 25—Generalized AEP Rate Function: Let $\boldsymbol{X}$ and $\boldsymbol{Y}$ be stationary random fields. Assume that ρ is essentially bounded, i.e., $D_{\max} < \infty$, and that with $\mathbb{P}$-probability one, conditional on $X_{[0,\infty)} = x_{[0,\infty)}$, the random variables $\{\rho_n(x_{C(n)}, Y_{C(n)})\}$ satisfy a large deviations principle with some deterministic, convex rate-function. Then for all $D \in (D_{\min}, D_{\text{av}})$, except possibly at $D = D_{\min}^{(\infty)}$

$$\lim_{n \to \infty} -\frac{1}{n^d} \log Q_n\left(B\left(X_{C(n)}, D\right)\right) = R(\mathbb{P}, \mathbb{Q}, D) \quad \text{w.p. 1} \tag{49}$$

where $D_{\min}^{(\infty)}$ and the rate-function $R(\mathbb{P}, \mathbb{Q}, D)$ are defined as in the one-dimensional case, by (17) and (16), respectively, and the rate-functions $R_n(P_n, Q_n, D)$ are now defined as

$$R_n(P_n, Q_n, D) = \inf_{V_n} \frac{1}{n^d} H(V_n \| P_n \times Q_n) \tag{50}$$

with the infimum taken over all joint distributions V_n on $A^{n^d} \times \hat{A}^{n^d}$ such that the A^{n^d}-marginal of V_n is P_n and

$$E_{V_n}[\rho_n(X_{C(n)}, Y_{C(n)})] \leq D.$$

Note Added in Proof: After this work was submitted, we received an interesting preprint from Chi [19] written in response to some questions raised in the earlier version of this paper. In

[19], Chi verifies the assumptions of Theorem 25 for the case when $\boldsymbol{Y}$ is a Gibbs field. In [19, Theorem 1], it is shown that if $\boldsymbol{X}$ is a stationary-ergodic random field with a finite alphabet and $\boldsymbol{Y}$ is a stationary Gibbs field also with a finite alphabet, then the LDP assumption of Theorem 25 is satisfied. Therefore, the generalized AEP also holds in this case with the rate function $R(\mathbb{P}, \mathbb{Q}, D)$ defined as in (49) and (50). We will discuss the further implications of this result for data compression on random fields in subsequent work.

Remark 6: Suppose that $(\boldsymbol{X}, \boldsymbol{Y})$ is a stationary random field satisfying a "process-level LDP" with a convex, good rate-function. To be precise, given $x_{C(n)} \in A^{n^d}$, write $x^{(n)}$ for the periodic extension of $x_{C(n)}$ to an infinite realization in $A^{[0,\infty)}$ and let $X^{(n)}$ and $Y^{(n)}$ denote the periodic extensions of $X_{C(n)}$ and $Y_{C(n)}$, respectively. The process-level empirical measure $\mathcal{L}_n$ induced by $\boldsymbol{X}$ and $\boldsymbol{Y}$ on $(A^{[0,\infty)} \times \hat{A}^{[0,\infty)})$ is defined by

$$\mathcal{L}_n \triangleq \frac{1}{n^d} \sum_{u \in C(n)} \delta_{(X^{(n)}_{u+[0,\infty)}, Y^{(n)}_{u+[0,\infty)})}$$

where $\delta_{s,s'}$ denotes the measure assigning unit mass to the joint realization

$$(s, s') \in A^{[0,\infty)} \times \hat{A}^{[0,\infty)}$$

and $X^{(n)}_{u+[0,\infty)}$ (or $Y^{(n)}_{u+[0,\infty)}$) denotes $X^{(n)}$ (respectively, $Y^{(n)}$) shifted by u [i.e., the value of $X^{(n)}_{u+[0,\infty)}$ at position v is the same as the value of $X^{(n)}$ at position $u+v$; similarly, for $Y^{(n)}_{u+[0,\infty)}$]. By assuming that $(\boldsymbol{X}, \boldsymbol{Y})$ satisfy a "process-level LDP" we mean that the sequence of measures $\{\mathcal{L}_n\}$ satisfies the LDP in the space of stationary probability measures on $(A^{[0,\infty)} \times \hat{A}^{[0,\infty)})$ equipped with the topology of weak convergence, with some convex, good rate-function $I(\cdot)$. These assumptions are satisfied by many of the random field models used in applications, and in particular by a large class of Gibbs fields (see, e.g., [22], [37], [63] for general theory and [39], [82] for examples in the areas of image processing and image analysis).

As in the one-dimensional case, suppose that the process-level LDP condition holds, and that the distortion measure ρ is bounded and continuous on $A \times \hat{A}$. Then with $\mathbb{P}$-probability one, conditional on $X_{[0,\infty)} = x_{[0,\infty)}$, the sequence $\{\rho_n(x_{C(n)}, Y_{C(n)})\}$ satisfies the LDP upper bound with respect to the deterministic, convex rate-function $J(\cdot)$ as in Remark 3. Moreover, assuming sufficiently strong mixing properties for $\boldsymbol{Y}$ one may also verify the corresponding lower bound (for example, by adapting the stochastic subadditivity approach of [21]).

C. Applications

In Sections VI-C1 and VI-C2 we consider the random field analogs of the problems discussed in Section III in the context of one-dimensional processes. In the instances when our analysis was restricted to i.i.d. processes, the extension to random fields is trivial—an i.i.d. random field is no different from an i.i.d. process. For that reason, we only give the full statements of corresponding random fields results when the generalization from $d = 1$ to $d \geq 2$ does involve some modifications. Otherwise, only a brief description of the corresponding results is mentioned.

1) Lossy Data Compression: Here we very briefly discuss the problem of data compression, when the data is in the form of a two- or more generally a d-dimensional array. In this case, the underlying data source is naturally modeled as a d-dimensional random field. Extensive discussions of the general information-theoretic problems on random fields are given in [10] and the recent monograph [93]; see also [36].

First we discuss the results given in Section III-A. The construction of the random codebooks described there generalizes to random fields in an obvious fashion, and the statement as well as the proof of Theorem 9 remain unchanged. Following the notation exactly as developed for i.i.d. sources, the strengthened coding theorems given in Theorems 10 and 11 follow by combining (the obvious generalization of) Theorem 9 with the generalized AEP of Theorem 24.

Similarly, the mismatched-codebook results of Section III-B only rely on Theorem 9 and the generalized AEP of Theorem 1, and therefore immediately generalize to the random field case.

2) Waiting Times: Here we consider the natural d-dimensional analogs of the waiting times questions considered in Section III-C. Given two independent realizations of the random fields $\boldsymbol{X}$ and $\boldsymbol{Y}$, our main quantity of interest here is how "far" we have to look in $\boldsymbol{Y}$ until we find a match for the pattern $X_{C(n)}$ with distortion D or less. Given $n \geq 1$ and a distortion level $D \geq 0$, we define the *waiting time* W_n as the smallest length i such that a copy of the pattern $X_{C(n)}$ appears somewhere in $Y_{C(i+n-1)}$, with distortion D or less. Formally,

$$W_n = \inf\{i \geq 1 \colon \rho_n(X_{C(n)}, Y_{u+C(n)}) \leq D \text{ for some } u \in [0, i-1]^d\}$$

with the convention that the infimum of the empty set equals $+\infty$.

In the one-dimensional case, our main tool in investigating the asymptotic behavior of the waiting times was the strong approximation in Theorem 13. Roughly speaking, Theorem 13 stated that the waiting time W_n for a D-close match of X_1^n in $\boldsymbol{Y}$ is inversely proportional to the probability $Q_n(B(X_1^n, D))$ of such a match. In Theorem 26 we generalize this result to the d-dimensional case by showing that the d-dimensional volume $(W_n)^d$ we have to search in $\boldsymbol{Y}$ in order to find a D-close match for $X_{C(n)}$ is, roughly, inversely proportional to the probability $Q_n(B(X_{C(n)}, D))$ of finding such a match.

Before stating Theorem 26, we need to recall the following definition. Dobrushin's *nonuniform ϕ-mixing coefficients* of a stationary random field $\boldsymbol{Y}$ are

$$\phi_\ell(k) = \sup\{|\mathbb{Q}(B|A) - \mathbb{Q}(B)| \colon B \in \sigma(Y_U),\ A \in \sigma(Y_V),\ \mathbb{Q}(A) > 0\ |U| \leq \ell,\ |V| < \infty,\ d(U,V) \geq k\}$$

where $\sigma(Y_U)$ denotes the σ-field generated by the random variables Y_U, $U \subset \mathbb{Z}^d$. See [59, Ch. 6] or [31] for detailed discussions of the coefficients $\{\phi_\ell(k)\}$ and their properties.

Theorem 26—Strong Approximation: Let $\boldsymbol{X}$ and $\boldsymbol{Y}$ be stationary ergodic random fields, and assume that the nonuniform ϕ-mixing coefficients of $\boldsymbol{Y}$ satisfy

$$\limsup_{n\to\infty} \sum_{j=1}^{\infty} (j+1)^{d-1} \phi_{n^d}(jn) < \infty. \tag{51}$$

If $Q_n(B(X_{C(n)}, D)) > 0$ eventually with probability one, then for any $\epsilon > 0$

$$-(1+\epsilon)\log n \le \log[W_n^d Q_n(B(X_{C(n)}, D))] \le (d+1+\epsilon)\log n \text{ eventually, w.p. 1.}$$

The proof of Theorem 26 is a straightforward modification of the corresponding one-dimensional argument in [27]; it is given in Appendix D.

Remark 7: The mixing condition (51) is satisfied by a rather large class of stationary random fields. For example, in the case of Markov random fields, it is easy to check that under Dobrushin's uniqueness condition (D) the limit in (51) is finite; see [38, Sec. 8.2] or [31] for more details.

Next we combine the above strong approximation result with the generalized AEPs of Theorems 24 and 25, to read off the first-order asymptotic behavior of the waiting times. Theorem 27 generalizes Theorem 14 to the random field case.

Theorem 27—SLLN for Waiting Times: Let $\boldsymbol{X}$ and $\boldsymbol{Y}$ be stationary ergodic random fields.

a) If $\boldsymbol{Y}$ is i.i.d. and the average distortion D_{av} is finite, then for any $D \in (D_{\min}, D_{\text{av}})$

$$\frac{1}{n^d}\log W_n^d \to R_1(P_1, Q_1, D) \quad \text{w.p. 1.}$$

b) Suppose that the conditions of Theorem 25 are satisfied, and that $\boldsymbol{Y}$ also satisfies the mixing assumption (51). Then, for any $D \in (D_{\min}^{(\infty)}, D_{\text{av}})$

$$\frac{1}{n^d}\log W_n^d \to R(\mathbb{P}, \mathbb{Q}, D) \quad \text{w.p. 1.}$$

VII. Random Fields: Second-Order Results

We turn to the random field extensions of the second-order results of Sections IV and V. In Section VII-A, we state the random field analog of the second-order generalized AEP, and in Section VII-B we discuss its application to the problems of lossy data compression and pattern matching.

A. Refinements of Generalized AEP

Let $\boldsymbol{X}$ be a stationary ergodic random field with marginal distribution P on A, and let Q be a fixed probability measure on $\hat{A}$. We will assume throughout that the distortion measure ρ has a finite third moment

$$D_3 \stackrel{\Delta}{=} E_{P\times Q}[\rho^3(X, Y)] < \infty \tag{52}$$

and that it is not essentially constant, i.e., $D_{\min} < D_{\text{av}}$, with $D_{\min}$ and D_{av} defined as before (cf. (7) and (8)).

The goal of this section is to give the random field analogs of Theorems 15 and 16 and of Corollary 17 from the one-dimensional case.

An examination of the proof of Theorem 15 in [90] shows that its proof only depends on the ergodicity of $\boldsymbol{X}$ and the i.i.d. structure of the product measures Q^n. Simply replacing the application of the ergodic theorem by the ergodic theorem for $\mathbb{Z}^d$ actions [57, Ch. 6] immediately yields the following generalization. As long as condition (52) is satisfied, for all $D \in (D_{\min}, D_{\text{av}})$ we have

$$-\log Q^{n^d}(B(X_{C(n)}, D)) = n^d R_1(\hat{P}_n, Q, D) + \frac{d}{2}\log n + O(1) \quad \text{w.p. 1} \tag{53}$$

where $\hat{P}_n$ is now the empirical measure induced by $X_{C(n)}$ on A.

In order to generalize Theorem 16 to $\mathbb{Z}^d$ we need to introduce a measure of dependence analogous to α-mixing in the one-dimensional case. For a stationary random field $\boldsymbol{X}$ on $\mathbb{Z}^d$ we define the *uniform α-mixing coefficients* of $\boldsymbol{X}$ by

$$\alpha(k) = \sup\{|\mathbb{P}(A\cap B) - \mathbb{P}(A)\mathbb{P}(B)|\colon A \in \sigma(X_U), B \in \sigma(X_V), d(U, V) \ge k\}$$

where, as before, $\sigma(X_U)$ denotes the σ-field generated by the random variables Y_U. See [59], [31] for more details.

Apart from ergodicity, the main technical ingredient in the proof of Theorem 16 (see also the proof of [27, Theorem 3]) is the LIL for $\boldsymbol{X}$. Similarly to the one-dimensional case, the LIL for a random field $\boldsymbol{X}$ holds as soon as the following mixing condition is satisfied:

$$\alpha(k) \le Ck^{-3d(1+\epsilon)}, \qquad \text{for some } \epsilon > 0 \text{ and } C < \infty. \tag{54}$$

[This follows from the almost sure invariance principle in [11, Theorem 1].]

Assuming that (54) and the third-moment condition (52) both hold, we get the following generalization of Theorem 16. For all $D \in (D_{\min}, D_{\text{av}})$

$$n^d R_1(\hat{P}_n, Q, D) = n^d R_1(P, Q, D) + \sum_{u\in C(n)} g(X_u) + O(\log\log n) \quad \text{w.p. 1} \tag{55}$$

with $g(x)$ defined exactly as in the one-dimensional case (39).

Combining (53) and (55) gives the following generalization of Corollary 17.

Theorem 28—Second-Order Generalized AEP: Let $\boldsymbol{X}$ be a stationary ergodic random field with marginal distribution P on A, and let Q be an arbitrary probability measure on $\hat{A}$. Assume that the uniform α-mixing coefficients of $\boldsymbol{X}$ satisfy (54) and that $D_3 = E_{P\times Q}[\rho^3(X, Y)]$ is finite. Then for any $D \in (D_{\min}, D_{\text{av}})$, and with $g(x)$ defined as in (39)

$$-\log Q^{n^d}(B(X_1^n, D)) = n^d R_1(P, Q, D) + \sum_{u\in C(n)} g(X_u) + \frac{d}{2}\log n + O(\log\log n) \quad \text{w.p. 1.}$$

B. Applications

Next we discuss applications of the second-order generalized AEP to the d-dimensional analogs of the data compression and pattern matching problems of Section IV. As in Section VI-C, the only results stated explicitly are those whose extensions to $\mathbb{Z}^d$ require modifications.

As mentioned in Section VI-C1, the one-dimensional construction of the random codes, as well as the main tool used in their analysis, Theorem 9, immediately generalize to the random

field case. And since all the second-order results of Section V-A (Theorems 18–21) are stated for i.i.d. sources, their statements as well as proofs carry over *verbatim* to this case.

For the problem of waiting times, we can use the second-order generalized AEP of Theorem 28 to refine the SLLN of Theorem 27

$$\frac{1}{n^d}\log W_n^d \to R_1(P, Q, D) \quad \text{w.p. } 1$$

to a corresponding CLT and LIL as in the one-dimensional case. These refinements are stated in Theorem 29. Its proof is identical to that of Theorem 22 in the one-dimensional case. The only difference here is that we need to invoke the CLT and LIL for the partial sums of the random field $\{g(X_u);\, u \in \mathbb{Z}^d\}$. Under the conditions of the theorem, these follow from the almost sure invariance principle of [11, Theorem 1].

Theorem 29: Let $\boldsymbol{X}$ be a stationary ergodic random field and $\boldsymbol{Y}$ be i.i.d., with marginal distributions P and Q on A and $\hat{A}$, respectively. Assume that the uniform α-mixing coefficients of $\boldsymbol{X}$ satisfy (54) and that $D_3 = E_{P\times Q}[\rho^3(X, Y)]$ is finite. Then for any $D \in (D_{\min}, D_{\text{av}})$ the following series is absolutely convergent:

$$\sigma^2 \triangleq \sum_{u\in\mathbb{Z}^d} E_P[g(X_{\mathbf{0}})g(X_u)] \tag{56}$$

with $g(x)$ defined as in (39), and, moreover

(CLT) With $R_1 = R_1(P, Q, D)$

$$\frac{\log W_n^d - n^d R_1}{n^{d/2}} \xrightarrow{\mathcal{D}} N(0, \sigma^2).$$

(LIL) The set of limit points of the sequence

$$\left\{\frac{\log W_n^d - n^d R_1}{\sqrt{2n^d \log\log n}}\right\}, \qquad n \geq 3$$

coincides with $[-\sigma, \sigma]$, with probability one.

Appendix A
Proof of Theorem 7

We prove the upper and lower bounds separately. For the upper bound, recalling the definition of $r_n(X_1^n, D)$ in (20) we observe that

$$r_n(X_1^n, D) \leq \frac{1}{n}\log P_n(B(X_1^n, D)) - \frac{1}{n}\log Q^n(X_1^n)$$

where the second term converges to $H(P) + H(P\|Q)$ as $n \to \infty$, by the ergodic theorem. Since the first term is increasing in D, for any fixed $D > 0$ we have with $\mathbb{P}$-probability one

$$\limsup_{\substack{n\to\infty\\ D\downarrow 0}} r_n(X_1^n, D) \leq H(P) + H(P\|Q) + \limsup_{n\to\infty} \frac{1}{n}\log P_n(B(X_1^n, D)). \tag{57}$$

Now the pointwise source-coding theorem (see [56, Theorems 1 and 5]) implies that

$$\liminf_{n\to\infty} -\frac{1}{n}\log P_n(B(X_1^n, D)) \geq R(D) \quad \text{w.p. } 1 \tag{58}$$

where $R(D)$ is the rate-distortion function of the source $\boldsymbol{X}$ (in nats). From (57) and (58) we get

$$\begin{aligned}\limsup_{\substack{n\to\infty\\ D\downarrow 0}} r_n(X_1^n, D) &\leq H(P) + H(P\|Q) - R(D)\\ &\leq H(P) + H(P\|Q) - H(\mathbb{P}) + H(P) - R_1(D) \quad \text{w.p. } 1\end{aligned}$$

where $R_1(D)$ denotes the first-order rate-distortion function of $\boldsymbol{X}$, $H(\mathbb{P})$ is the entropy rate of $\boldsymbol{X}$ (both in nats), and the second inequality follows from the Wyner–Ziv bound; see [83, Remark 4]. The assumption that $\rho(x, y) = 0$ if and only if $x = y$ implies that $\lim_{D\to 0} R_1(D) = H(P)$, so letting $D \downarrow 0$ the above right-hand side becomes $H(P) + H(P\|Q) - H(\mathbb{P})$ and it is an easy calculation to verify that this is indeed the same as $H(\mathbb{P}\|\mathbb{Q})$. This gives the required upped bound.

For the lower bound, we proceed similarly by noting that

$$r_n(X_1^n, D) \geq \frac{1}{n}\log P_n(X_1^n) - \frac{1}{n}\log Q^n(B(X_1^n, D))$$

where the first term converges to $-H(\mathbb{P})$ by the classical AEP (as $n \to \infty$). Since the second term is decreasing in D, for any fixed $D \in (0, D_{\text{av}})$ we have with probability one

$$\begin{aligned}\liminf_{\substack{n\to\infty\\ D\downarrow 0}} r_n(X_1^n, D) &\geq -H(\mathbb{P}) - \limsup_{n\to\infty}\frac{1}{n}\log Q^n(B(X_1^n, D))\\ &= -H(\mathbb{P}) + R_1(P, Q, D)\end{aligned}$$

where the last step follows from the generalized AEP in Theorem 1 (note that $D_{\min} = 0$ here). By the characterization of the rate-function in Theorem 2 we know that

$$\begin{aligned}R_1(P, Q, D) &= \sup_{\lambda'\leq 0}[\lambda' D - \Lambda(\lambda')] \geq [\lambda D - \Lambda(\lambda)]\\ &= -E_P\left[\log E_Q\left(e^{\lambda(\rho(X,Y)-D)}\right)\right]\end{aligned}$$

for any fixed $\lambda < 0$. Therefore, for any $D \in (0, D_{\text{av}})$ and $\lambda < 0$, we have

$$\liminf_{\substack{n\to\infty\\ D\downarrow 0}} r_n(X_1^n, D) \geq -H(\mathbb{P}) - E_P\left[\log E_Q\left(e^{\lambda(\rho(X,Y)-D)}\right)\right] \quad \text{w.p. } 1.$$

Letting $D \to 0$ and then $\lambda \to -\infty$, by the dominated convergence theorem (and the assumption $\rho(x, y) = 0$ iff $x = y$) the right-hand side above converges to

$$-H(\mathbb{P}) + H(P\|Q) + H(P) = H(\mathbb{P}\|\mathbb{Q})$$

proving the lower bound.

Finally, since for each fixed n the limit as $D \downarrow 0$ of $r_n(X_1^n, D)$ exists, it follows that the repeated limit $\lim_n \lim_D$, also exists and is equal to the double limit $H(\mathbb{P}\|\mathbb{Q})$. □

Appendix B
Proof of Theorem 8

Part a): Fixing n, let $f_n = dP_n/dQ_n$ and consider the set

$$A_n \triangleq \left\{x_1^n\colon Q_n(B(x_1^n, D)) > 0\ \forall D > 0,\ f_n(x_1^n) = \limsup_{D\downarrow 0}\frac{P_n(B(x_1^n, D))}{Q_n(B(x_1^n, D))} = \liminf_{D\downarrow 0}\frac{P_n(B(x_1^n, D))}{Q_n(B(x_1^n, D))}\right\}.$$

By the Lebesgue–Besicovitch differentiation theorem (cf. [33, Theorems 1.6.1, 1.6.2]), we know that $Q_n(A_n) = 1$, hence also $P_n(A_n) = 1$. With $\mathbb{P}(\cup_n A_n^c) = 0$, we conclude the proof of part a) by applying Theorem 6 for $M_n = Q^n$ (in which case $H_n \geq 0$).

Part b): As $Q(A_1) = 1$, in particular $Q(B(x, D)) > 0$ for all $D > 0$ and Q-almost every $x \in \mathbb{R}^d$ (hence also for $P = P_1$-almost every $x \in \mathbb{R}^d$), implying that $D_{\min}$ of (7) is zero. The same argument yields also that $P(B(x, D)) > 0$ for all $D > 0$ and P-almost every x, hence $D_{\min}$ is still zero if we replace Q by P. Thus, for all

$$D < \min\{E_{P\times Q}[\rho(X, Y)], E_{P\times P}[\rho(X, Y)]\}$$

applying Theorem 1 twice we get

$$\lim_{n\to\infty} r_n(X_1^n, D) = R_1(P, Q, D) - R_1(P, P, D) \quad \text{w.p. 1.}$$

For any probability measure μ and any $\lambda \leq 0$, let

$$\Lambda(\lambda; \mu) = \int \left[\log \int e^{\lambda\rho(x, y)}\, d\mu(y)\right] dP(x).$$

Fixing $D > 0$ small enough, we have by Theorem 2 that

$$R_1(P, P, D) = \lambda D - \Lambda(\lambda; P)$$

for the unique $\lambda = \lambda(D) < 0$ such that $\Lambda'(\lambda; P) = D$, whereas

$$R_1(P, Q, D) \geq \lambda D - \Lambda(\lambda; Q).$$

Since $E_{P\times P}[\rho(X, Y)] > 0$, we have also that $\lambda(D) \downarrow -\infty$ as $D \downarrow 0$ (see (11)). Consequently,

$$\liminf_{D\downarrow 0} \{R_1(P, Q, D) - R_1(P, P, D)\} \geq \liminf_{\lambda\downarrow -\infty} \{\Lambda(\lambda; P) - \Lambda(\lambda; Q)\}.$$

Similarly, by Theorem 2 we have

$$R_1(P, Q, D) = \tilde{\lambda} D - \Lambda(\tilde{\lambda}; Q), \qquad \text{for } \tilde{\lambda} < 0$$

such that $\Lambda'(\tilde{\lambda}; Q) = D$, $R_1(P, P, D) \geq \tilde{\lambda} D - \Lambda(\tilde{\lambda}; P)$, and with $E_{P\times Q}[\rho(X, Y)] > 0$, also $\tilde{\lambda} \downarrow -\infty$ when $D \downarrow 0$. Therefore, it suffices to show that

$$\lim_{\lambda\downarrow -\infty} \{\Lambda(\lambda; P) - \Lambda(\lambda; Q)\} = H(P\|Q). \tag{59}$$

To this end, for any $\lambda < 0$ and $x \in \mathbb{R}^d$, let

$$h_\lambda(x) \triangleq \frac{E_P\left(e^{\lambda\rho(x, Y)}\right)}{E_Q\left(e^{\lambda\rho(x, Y)}\right)}$$

noting that

$$\Lambda(\lambda; P) - \Lambda(\lambda; Q) = \int \log h_\lambda(x)\, dP(x).$$

Using the change of variable $U = \rho(x, Y) \geq 0$ followed by integration by parts, we see that

$$h_\lambda(x) = \frac{\int_0^\infty e^{\lambda u} g_x(u)\, du}{\int_0^\infty e^{\lambda u} k_x(u)\, du}$$

where $g_x(r) = P(B(x, r))$ and $k_x(r) = Q(B(x, r))$ are nonnegative, nondecreasing and bounded above by 1. Considering separately $u \leq 2\eta$ and $u > 2\eta$, it is easy to check that for any $\eta > 0$

$$\sup_{0<r\leq 2\eta} \frac{g_x(r)}{k_x(r)} + \psi_{\lambda, x} \geq h_\lambda(x) \geq \inf_{0<r\leq 2\eta} \frac{g_x(r)}{k_x(r)} \frac{1}{1 + \psi_{\lambda, x}} \tag{60}$$

where

$$\psi_{\lambda, x} \triangleq \frac{\int_{2\eta}^\infty e^{\lambda u}\, du}{\int_0^{2\eta} e^{\lambda u} k_x(u)\, du} \leq \frac{1}{\eta|\lambda| k_x(\eta)}. \tag{61}$$

Fix $x \in A_1$ of part a), in which case $k_x(r) > 0$ for all $r > 0$ and $g_x(r)/k_x(r) \to f_1(x)$ as $r \to 0$. Letting $\lambda \downarrow -\infty$ and then $\eta \to 0$, it follows by (60) and (61) that

$$\lim_{\lambda\downarrow -\infty} h_\lambda(x) = f_1(x).$$

Recall that $P(A_1) = 1$ and our assumption that

$$\int \log k_x(\eta) dP(x) > -\infty$$

for any $\eta > 0$. By our integrability conditions, the function $\min\{0, \inf_{\lambda\geq 1} \log h_\lambda(x)\}$ is P-integrable, hence, by Fatou's lemma

$$\liminf_{\lambda\downarrow -\infty} \int \log h_\lambda(x)\, dP(x) \geq \int \log f_1(x)\, dP(x) = H(P\|Q).$$

Moreover, in case $H(P\|Q) < \infty$, our assumptions imply that $\sup_{\lambda\geq 1} |\log h_\lambda(x)|$ is P-integrable, hence by dominated convergence

$$\int \log h_\lambda(x)\, dP(x) \to \int \log f_1(x)\, dP(x)$$

for $\lambda \downarrow -\infty$, as required to complete the proof of (59). □

Appendix C
Proof of Theorem 25

Recall our assumption that, for $\mathbb{P}$-a.e. $x_{[0,\infty)}$, conditional on $X_{[0,\infty)} = x_{[0,\infty)}$ the random variables $\{\rho_n(x_{C(n)}, Y_{C(n)})\}$ satisfy the LDP with a *deterministic* convex good rate-function denoted hereafter $R(\mathbb{P}, \mathbb{Q}, \cdot)$. Since ρ is bounded, by Varadhan's lemma and convex duality, this implies that

$$R(\mathbb{P}, \mathbb{Q}, D) = \sup_{\lambda\in\mathbb{R}} [\lambda D - \Lambda_\infty(\lambda)] \triangleq \Lambda_\infty^*(D) \tag{62}$$

where for any $\lambda \in \mathbb{R}$, the finite, deterministic limit

$$\Lambda_\infty(\lambda) \triangleq \lim_{n\to\infty} \frac{1}{n^d} \log \int e^{\lambda \sum_{u\in C(n)} \rho(x_u, y_u)}\, dQ_n(y_{C(n)})$$

exists for $\mathbb{P}$-a.e. $x_{[0,\infty)}$ (cf. [29, Theorem 4.5.10]). By bounded convergence, $\Lambda_\infty(\lambda)$ is also the limit of

$$\Lambda_n(\lambda) \triangleq \frac{1}{n^d} \int \left[\log \int e^{\lambda \sum_{u\in C(n)} \rho(x_u, y_u)} \cdot dQ_n(y_{C(n)})\right] dP_n(x_{C(n)}).$$

By stationarity

$$D_{\text{av}} = E_{P_n \times Q_n}\left(\rho_n(X_{C(n)}, Y_{C(n)})\right), \qquad \forall\, n \geq 1 \quad (63)$$

so replacing P_1, Q_1, and $\rho(x, y)$ of Theorem 2 by P_n, Q_n, and $n^d \rho_n(x_{C(n)}, y_{C(n)})$, respectively, we see that

$$R_n(P_n, Q_n, D) = \sup_{\lambda \in \mathbb{R}} [\lambda D - \Lambda_n(\lambda)] \triangleq \Lambda_n^*(D). \quad (64)$$

Note that $|\Lambda_n(\lambda) - \Lambda_n(\lambda')| \leq c|\lambda - \lambda'|$ for some $c < \infty$ and all n, λ, $\lambda' \in \mathbb{R}$, hence, the convergence of $\Lambda_n(\cdot)$ to $\Lambda_\infty(\cdot)$ is uniform on compact subsets of $\mathbb{R}$. In particular, the convex, continuous functions $\Lambda_n(\cdot)$ converge infimally to $\Lambda_\infty(\cdot)$, and, consequently, by [80, Theorem 5], the convex functions $\Lambda_n^*(\cdot)$ converge infimally to $\Lambda_\infty^*(\cdot)$, that is,

$$\begin{aligned}\Lambda_\infty^*(D) &= \lim_{\delta \to 0} \limsup_{n \to \infty} \inf_{|\hat{D} - D| < \delta} \Lambda_n^*(\hat{D}) \\ &= \lim_{\delta \to 0} \liminf_{n \to \infty} \inf_{|\hat{D} - D| < \delta} \Lambda_n^*(\hat{D}). \end{aligned} \quad (65)$$

It follows from (63) and Jensen's inequality that $\Lambda_n(\lambda) \geq \lambda D_{\text{av}}$ for all n and λ, hence, for $D \leq D_{\text{av}}$ suffices to consider $\lambda \leq 0$ in (62) and in (64). Thus, for $1 \leq n \leq \infty$, Λ_n^* are nonnegative, convex, and monotone nonincreasing on $[0, D_{\text{av}}]$, with $\Lambda_n^*(D_{\text{av}}) = 0$. For $1 \leq n \leq \infty$, let

$$D_{\min}^{(n)} \triangleq \lim_{\lambda \downarrow -\infty} \frac{\Lambda_n(\lambda)}{\lambda}$$

so that $\Lambda_n^*(D) = \infty$ for $D < D_{\min}^{(n)}$, while $\Lambda_n^*(D) < \infty$ for $D > D_{\min}^{(n)}$. Note that for $n < \infty$ this coincides with the definition of $D_{\min}^{(n)}$ given in (48). It is easy to check then that (65) implies the pointwise convergence of $\Lambda_n^*(\cdot) = R_n(\mathbb{P}, \mathbb{Q}, \cdot)$ to $\Lambda_\infty^*(\cdot) = R(\mathbb{P}, \mathbb{Q}, \cdot)$ at any D for which $\Lambda_\infty^*(D-\delta) \downarrow \Lambda_\infty^*(D)$, that is, for all $D \neq D_{\min}^{(\infty)}$. In particular, necessarily $D_{\min}^{(\infty)} \in [D_{\min}, D_{\text{av}}]$, and $D_{\min}^{(\infty)}$ may also be defined via (17). The continuity of $R(\mathbb{P}, \mathbb{Q}, D)$ at $D \in (D_{\min}, D_{\text{av}})$, $D \neq D_{\min}^{(\infty)}$ implies the equality in (49) for such D, thus, completing the proof of the theorem. □

Appendix D
Proof of Theorem 26

For each $m \geq 1$, let G_m be the collection of "good" realizations $x_{\mathbb{Z}^d} \in A^{\mathbb{Z}^d}$

$$G_m = \left\{x_{\mathbb{Z}^d} \in A^{\mathbb{Z}^d} : Q_n(B(x_{C(n)}, D)) > 0 \text{ for all } n \geq m\right\}$$

so that the assumption that $Q_n(B(X_{C(n)}, D)) > 0$ eventually, with probability one translates to

$$\mathbb{P}\left\{\bigcup_{m \geq 1} G_m\right\} = 1. \quad (66)$$

To prove the lower bound we choose and fix an $m \geq 1$ and a realization $x_{\mathbb{Z}^d} \in G_m$. Then for any $K > 1$

$$\begin{aligned}&\Pr\left\{W_n^d < K | X_{C(n)} = x_{C(n)}\right\} \\ &\quad \leq \sum_{u \in [0, \lfloor K^{1/d} \rfloor - 1]^d} Q_n\left\{Y_{u + C(n)} \in B(x_{C(n)}, D)\right\} \\ &\quad \leq K Q_n(B(x_{C(n)}, D)).\end{aligned}$$

Since, by its definition, W_n is always greater than or equal to one, this inequality trivially holds also for $K \in (0, 1]$. Setting $K = [n^{1+\epsilon} Q_n(B(x_{C(n)}, D))]^{-1}$ above gives, for all $n \geq m$,

$$\begin{aligned}&\Pr\left\{\log\left[W_n^d Q_n\left(B\left(X_{C(n)}, D\right)\right)\right]\right. \\ &\qquad \left. < -(1+\epsilon)\log n | X_{C(n)} = x_{C(n)}\right\} \leq \frac{1}{n^{1+\epsilon}}.\end{aligned}$$

Since this bound is uniform over $x_{\mathbb{Z}^d} \in G_m$ and summable, the Borel–Cantelli lemma and assumption (66) imply that

$$\begin{aligned}&\log\left[W_n^d Q_n\left(B\left(X_{C(n)}, D\right)\right)\right] \\ &\qquad \geq -(1+\epsilon)\log n \text{ eventually, w.p. 1.}\end{aligned} \quad (67)$$

For the upper bound we choose and fix an $m \geq 1$ and a realization $x_{\mathbb{Z}^d} \in G_m$, and take $K \geq (n+1)^d$. Note that

$$\begin{aligned}&\Pr\left\{W_n^d > K | X_{C(n)} = x_{C(n)}\right\} \\ &\qquad \leq \Pr\left\{\sum_{u \in [0, M]^d} \mathbb{I}_{\{Y_{nu + C(n)} \in B(x_{C(n)}, D)\}} = 0\right\}\end{aligned}$$

where the sum is over the $(M+1)^d$ integer positions $u \in [0, M]^d \subset \mathbb{Z}^d$, nu denotes the point $(nu_1, nu_2, \ldots, nu_d) \in \mathbb{Z}^d$, and

$$M = M(K, n) \triangleq \left\lfloor \frac{K^{1/d} - 1}{n} \right\rfloor.$$

Let Σ_n denote the sum in the above probability

$$\Sigma_n = \sum_{u \in [0, M]^d} I_n(u)$$

where $I_n(u)$ is the indicator function of the event

$$\{Y_{nu + C(n)} \in B(x_{C(n)}, D)\}.$$

In this notation

$$\Pr\left\{W_n^d > K | X_{C(n)} = x_{C(n)}\right\} \leq \mathbb{Q}\{\Sigma_n = 0\} \leq \frac{\text{Var}_{\mathbb{Q}}(\Sigma_n)}{[E_{\mathbb{Q}}(\Sigma_n)]^2}. \quad (68)$$

By stationarity

$$E_{\mathbb{Q}}(\Sigma_n) = [M+1]^d Q_n(B(x_{C(n)}, D)) \quad (69)$$

and by the definition of the ϕ-mixing coefficients, if $u \neq v$

$$\begin{aligned}E_{\mathbb{Q}}\{I_n(u) I_n(v)\} &\leq Q_n(B(x_{C(n)}, D)) \\ &\quad \cdot [\phi_{n^d}(nd(u, v) - n + 1) + Q_n(B(x_{C(n)}, D))].\end{aligned}$$

Using the last two estimates we can bound the variance as

$$\begin{aligned}\text{Var}_{\mathbb{Q}}\{\Sigma_n\} &= \sum_{u, v \in [0, M]^d} \text{Cov}_{\mathbb{Q}}(I_n(u), I_n(v)) \\ &\leq [M+1]^d Q_n(B(x_{C(n)}, D)) + \sum_{u, v \in [0, M]^d, u \neq v} \\ &\quad \cdot \left[Q_n\left(B\left(x_{C(n)}, D\right)\right) \phi_{n^d}(nd(u, v) - n + 1)\right] \\ &\leq [M+1]^d Q_n(B(x_{C(n)}, D)) \\ &\quad \cdot \left[1 + \sum_{j=1}^{M} c_d j^{d-1} \phi_{n^d}(nj - n + 1)\right]\end{aligned} \quad (70)$$

where $c_d j^{d-1}$ bounds the number of possible points u that can be at a distance exactly j from a given point v (for some constant c_d). By assumption (51) we can find a finite constant Φ such that the expression in square brackets in (70) is bounded above by Φ, uniformly in n. Substituting this bound, together with (69) and (70), in (68), gives

$$\Pr\{W_n > K | X_{C(n)} = x_{C(n)}\} \leq \frac{\Phi}{[M+1]^d Q_n\left(B\left(x_{C(n)}, D\right)\right)}. \quad (71)$$

Let $\epsilon > 0$ arbitrary, take n large enough so that $n^{(1+\epsilon)/d} \geq 2$, and let $K = n^{d+1+\epsilon}/Q_n(B(x_{C(n)}, D))$. Simple algebra shows that with this choice of K we have

$$[M+1]^d Q_n\left(B\left(x_{C(n)}, D\right)\right) \geq \frac{1}{2} n^{1+\epsilon}$$

and substituting this in (71) yields

$$\Pr\left\{\log\left[W_n^d Q_n\left(B\left(X_{C(n)}, D\right)\right)\right] > (d+1+\epsilon)\log n \middle| X_{C(n)} = x_{C(n)}\right\} \leq \frac{2\Phi}{n^{1+\epsilon}}.$$

This bound is uniform over $x_{\mathbb{Z}^d} \in G_m$ and summable, so the Borel–Cantelli lemma and (66) imply that

$$\log\left[W_n^d Q_n\left(B\left(X_{C(n)}, D\right)\right)\right] \leq (d+1+\epsilon)\log n \quad \text{eventually, w.p. 1.} \quad (72)$$

Combining (72) and (67) completes the proof. □

ACKNOWLEDGMENT

The authors would like to thank Tamás Linder and Yuval Peres for useful discussions regarding Theorems 7 and 8, and Wojciech Szpankowski for a number of bibliographical comments. They would also like to thank the reviewers for their very careful reading of the original manuscript and their valuable comments.

REFERENCES

[1] M. Alzina, W. Szpankowski, and A. Grama, "2D-pattern matching image and video compression," *IEEE Trans. Image Processing*, vol. 11, pp. 318–331, Mar. 2002.
[2] R. Arratia, L. Gordon, and M. S. Waterman, "The Erdös–Rényi law in distribution for coin tossing and sequence matching," *Ann. Statist.*, vol. 18, pp. 539–570, 1990.
[3] R. Arratia and M. S. Waterman, "The Erdös–Rényi strong law for pattern matching with a given proportion of mismatches," *Ann. Probab.*, vol. 17, no. 3, pp. 1152–1169, 1989.
[4] ——, "A phase transition for the score in matching random sequences allowing deletions," *Ann. Appl. Probab.*, vol. 4, pp. 200–225, 1994.
[5] M. Atallah, Y. Génin, and W. Szpankowski, "Pattern matching image compression: Algorithmic and empirical results," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 21, pp. 614–627, July 1999.
[6] A. R. Barron, "The strong ergodic theorem for densities: Generalized Shannon–McMillan–Breiman theorem," *Ann. Probab.*, vol. 13, pp. 1292–1303, 1985.
[7] J. G. Bell, T. C. Cleary, and I. H. Witten, *Text Compression*. Englewood Cliffs, NJ: Prentice-Hall, 1990.
[8] R. Bell and T. M. Cover, "Game-theoretic optimal portfolios," *Manag. Sci.*, vol. 34, no. 6, pp. 724–733, 1988.
[9] T. Berger, *Rate Distortion Theory: A Mathematical Basis for Data Compression*. Englewood Cliffs, NJ: Prentice-Hall, 1971.
[10] T. Berger, S. Y. Shen, and Z. X. Ye, "Some communication problems of random fields," *Int. J. Math. Statist. Sci.*, vol. 1, no. 1, pp. 47–77, 1992.
[11] I. Berkes and G. J. Morrow, "Strong invariance principles for mixing random fields," *Z. Wahrsch. Verw. Gebiete*, vol. 57, no. 1, pp. 15–37, 1981.
[12] B. C. Bradley, "Basic properties of strong mixing conditions," in *Dependence in Probability and Statistics*, E. Eberlein and M. S. Taqqu, Eds. Boston, MA: Birkhauser, 1986, pp. 165–192.
[13] L. Breiman, "The individual ergodic theorem for information theory," *Ann. Math. Statist.*, vol. 28, pp. 809–811, 1957.
[14] ——, "Correction to 'The individual ergodic theorem for information theory'," *Ann. Math. Statist.*, vol. 31, pp. 809–810, 1960.
[15] W. Bryc and A. Dembo, "Large deviations and strong mixing," *Ann. Inst. H. Poincaré Probab. Statist.*, vol. 32, no. 4, pp. 549–569, 1996.
[16] J. A. Bucklew, "The source coding theorem via Sanov's theorem," *IEEE Trans. Inform. Theory*, vol. IT-33, pp. 907–909, Nov. 1987.
[17] ——, "A large deviation theory proof of the abstract alphabet source coding theorem," *IEEE Trans. Inform. Theory*, vol. 34, pp. 1081–1083, Sept. 1988.
[18] J.-R. Chazottes, E. Floriani, and R. Lima, "Relative entropy and identification of Gibbs measures in dynamical systems," *J. Statist. Phys.*, vol. 90, pp. 697–725, 1998.
[19] Z. Chi, "Conditional large deviation principles for Gibbs random fields," preprint, Oct. 2001.
[20] ——, "The first-order asymptotics of waiting times with distortion between stationary processes," *IEEE Trans. Inform. Theory*, vol. 47, pp. 338–347, Jan. 2001.
[21] ——, "Stochastic sub-additivity approach to conditional large deviation principle," *Ann. Probab.*, vol. 23, pp. 1303–1328, 2001.
[22] F. Comets, "Grandes déviations pour des champs de Gibbs sur $\mathbf{Z}^d$," *C. R. Acad. Sci. Paris Sér. I Math.*, vol. 303, no. 11, pp. 511–513, 1986.
[23] ——, "Large deviation estimates for a conditional probability distribution. Applications to random interaction Gibbs measures," *Probab. Theory Related Fields*, vol. 80, pp. 407–432, 1989.
[24] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.
[25] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. New York: Academic, 1981.
[26] D. A. Dawson and J. Gärtner, "Large deviations from the McKean–Vlasov limit for weakly interacting diffusions," *Stochastics*, vol. 20, no. 4, pp. 247–308, 1987.
[27] A. Dembo and I. Kontoyiannis, "The asymptotics of waiting times between stationary processes, allowing distortion," *Ann. Appl. Probab.*, vol. 9, pp. 413–429, 1999.
[28] ——, "Critical behavior in lossy source coding," *IEEE Trans. Inform. Theory*, vol. 47, pp. 1230–1236, Mar. 2001.
[29] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*, 2nd ed. New York: Springer-Verlag, 1998.
[30] J. D. Deuschel and D. W. Stroock, *Large Deviations*. Boston, MA: Academic, 1989.
[31] P. Doukhan, *Mixing: Properties and Examples*. New York: Springer-Verlag, 1994.
[32] P. Elias, "Universal codeword sets and representations of the integers," *IEEE Trans. Inform. Theory*, vol. IT-21, pp. 194–203, Mar. 1975.
[33] L. C. Evans and R. F. Gariepy, *Measure Theory and Fine Properties of Functions*. Boca Raton, FL: CRC, 1992.
[34] J. Feldman, "r-entropy, equipartition, and Ornstein's isomorphism theorem in $\mathbf{R}^n$," *Israel J. Math.*, vol. 36, no. 3/4, pp. 321–345, 1980.
[35] W. Feller, *An Introduction to Probability Theory and Its Applications*, 2nd ed. New York: Wiley, 1971, vol. II.
[36] H. Föllmer, "On entropy and information gain in random fields," *Z. Wahrsch. Verw. Gabiete*, vol. 26, pp. 207–217, 1973.
[37] H. Föllmer and S. Orey, "Large deviations for the empirical field of a Gibbs measure," *Ann. Probab.*, vol. 16, no. 3, pp. 961–977, 1988.
[38] H.-O. Georgii, *Gibbs Measures and Phase Transitions*. Berlin, Germany: W. de Gruyter, 1989.
[39] X. Guyon, *Random Fields on a Network: Modeling, Statistics, and Applications*. New York: Springer-Verlag, 1995.
[40] I. A. Ibragimov, "Some limit theorems for stationary processes," *Theory Probab. Appl.*, vol. 7, pp. 349–382, 1962.
[41] D. Ishii and H. Yamamoto, "The redundancy of universal coding with a fidelity criterion," *IEICE Trans. Fundamentals*, vol. E80-A, pp. 2225–2231, 1997.
[42] F. Kanaya and J. Muramatsu, "An almost sure recurrence theorem with distortion for stationary ergodic sources," *IEICE Trans. Fundamentals*, vol. E80-A, pp. 2264–2267, 1997.
[43] A. Kanlis, "Compression and transmission of information at multiple resolutions," Ph.D. dissertation, Dept. Elec. Comput. Eng., Univ. Maryland, College Park, 1998.

[44] A. Kanlis, P. Narayan, and B. Rimoldi, "On three topics for a course in information theory," in *Statistical Methods in Imaging, Medicine, Optics, and Communication*, J. A. O'Sullivan, Ed. New York: Springer-Verlag, 2001.

[45] S. Karlin and F. Ost, "Maximal length of common words among random letter sequences," *Ann. Probab.*, vol. 16, pp. 535–563, 1988.

[46] J. C. Kieffer, "A counterexample to Perez's generalization of the Shannon–McMillan theorem," *Ann. Probab.*, vol. 1, pp. 362–364, 1973.

[47] ——, "Correction to: 'A counterexample to Perez's generalization of the Shannon–McMillan theorem' ," *Ann. Probab.*, vol. 4, pp. 153–154, 1976. The paper being corrected in *Ann. Probab.*, vol.1, pp. 362–364, 1973.

[48] ——, "Sample converses in source coding theory," *IEEE Trans. Inform. Theory*, vol. 37, pp. 263–268, Mar. 1991.

[49] H. Koga and S. Arimoto, "On the asymptotic behaviors of the recurrence time with a fidelity criterion for discrete memoryless sources and memoryless Gaussian sources," *IEICE Trans. Fundamentals*, vol. E81-A, pp. 981–986, 1998.

[50] I. Kontoyiannis, "Second-order noiseless source coding theorems," *IEEE Trans. Inform. Theory*, vol. 43, pp. 1339–1341, July 1997.

[51] ——, "Asymptotic recurrence and waiting times for stationary processes," *J. Theor. Probab.*, vol. 11, pp. 795–811, 1998.

[52] ——, "Recurrence and waiting times in stationary processes, and their applications in data compression," Ph.D. dissertation, Dept. Elec. Eng., Stanford Univ.Stanford, CA, May 1998.

[53] ——, "An implementable lossy version of the Lempel–Ziv algorithm—Part I: Optimality for memoryless sources," *IEEE Trans. Inform. Theory*, vol. 45, pp. 2293–2305, Nov. 1999.

[54] ——, "Pointwise redundancy in lossy data compression and universal lossy data compression," *IEEE Trans. Inform. Theory*, vol. 46, pp. 136–152, Jan. 2000.

[55] ——, "Sphere-covering, measure concentration, and source coding," *IEEE Trans. Inform. Theory*, vol. 47, pp. 1544–1552, May 2001.

[56] I. Kontoyiannis and J. Zhang, "Arbitrary source models and Bayesian codebooks in rate-distortion theory," *IEEE Trans. Inform. Theory*, to be published.

[57] U. Krengel, *Ergodic Theorems*. Berlin, Germany: W. de Gruyte, 1985.

[58] A. Lapidoth, "On the role of mismatch in rate distortion theory," *IEEE Trans. Inform. Theory*, vol. 43, pp. 38–47, Jan. 1997.

[59] Z. Lin and C. Lu, *Limit Theory for Mixing Dependent Random Variables*. Dordrecht, The Netherlands: Kluwer, 1996.

[60] T. Łuczak and W. Szpankowski, "A suboptimal lossy data compression algorithm based on approximate pattern matching," *IEEE Trans. Inform. Theory*, vol. 43, pp. 1439–1451, Sept. 1997.

[61] K. Marton and P. C. Shields, "Almost sure waiting time results for weak and very weak Bernoulli processes," *Ergod. Theory Dynam. Syst.*, vol. 15, pp. 951–960, 1995.

[62] B. McMillan, "The basic theorems of information theory," *Ann. Math. Statist.*, vol. 24, pp. 196–219, 1953.

[63] S. Olla, "Large deviations for Gibbs random fields," *Probab. Theory Related Fields*, vol. 77, no. 3, pp. 343–357, 1988.

[64] H. Oodaira and K. I. Yoshihara, "The law of the iterated logarithm for stationary processes satisfying mixing conditions," *Kōdai Math. Sem. Rept.*, vol. 23, pp. 311–334, 1971.

[65] ——, "Note on the law of the iterated logarithm for stationary processes satisfying mixing conditions," *Kōdai Math. Sem. Rep.*, vol. 23, pp. 335–342, 1971.

[66] S. Orey, "On the Shannon–Perez–Moy theorem," in *Particle Systems, Random Media and Large Deviations (Brunswick, ME, 1984)*. Providence, RI: Amer. Math. Soc., 1985, pp. 319–327.

[67] ——, "Large deviations in ergodic theory," in *Seminar on Stochastic Processes, 1984 (Evanston, IL 1984)*. Boston, MA: Birkhäuser, 1986, pp. 195–249.

[68] D. Ornstein and B. Weiss, "The Shannon–McMillan–Breiman theorem for a class of amenable groups," *Israel J. Math.*, vol. 44, pp. 53–60, 1983.

[69] ——, "Entropy and data compression schemes," *IEEE Trans. Inform. Theory*, vol. 39, pp. 78–83, Jan. 1993.

[70] M. Peligrad, "Recent advances in the central limit theorem and its weak invariance principle for mixing sequences of random variables (a survey)," in *Dependence in Probability and Statistics*, E. Eberlein and M. S. Taqqu, Eds. Boston, MA: Birkhauser, 1986, pp. 193–223.

[71] W. Philipp and W. Stout, *Almost Sure Invariance Principles for Partial Sums of Weakly Dependent Random Variables*, ser. Memoirs of the AMS, 1975, vol. 2, iss. 2, no. 161.

[72] D. J. Sakrison, "The rate distortion function for a class of sources," *Inform. Contr.*, vol. 15, pp. 165–195, 1969.

[73] ——, "The rate of a class of random processes," *IEEE Trans. Inform. Theory*, vol. IT-16, pp. 10–16, Jan. 1970.

[74] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 379–423, 623–656, 1948.

[75] ——, "Coding theorems for a discrete source with a fidelity criterion," in *IRE Nat. Conv. Rec.*, 1959, pp. 142–163. Reprinted in D. Slepian, Ed., *Key Papers in the Development of Information Theory* New York: IEEE Press, 1974.

[76] P. C. Shields, "Waiting times: Positive and negative results on the Wyner–Ziv problem," *J. Theoret. Probab.*, vol. 6, no. 3, pp. 499–519, 1993.

[77] Y. Steinberg and M. Gutman, "An algorithm for source coding subject to a fidelity criterion, based upon string matching," *IEEE Trans. Inform. Theory*, vol. 39, pp. 877–886, May 1993.

[78] W. Szpankowski, "Asymptotic properties of data compression and suffix trees," *IEEE Trans. Inform. Theory*, vol. 39, pp. 1647–1659, Sept. 1993.

[79] ——, *Average Case Analysis of Algorithms on Sequences*. New York: Wiley, 2001.

[80] R. A. Wijsman, "Convergence of sequences of convex sets, cones and functions," *Bull. Amer. Math. Soc.*, vol. 70, pp. 186–188, 1964.

[81] F. M. J. Willems, "Universal data compression and repetition times," *IEEE Trans. Inform. Theory*, vol. 35, pp. 54–58, Jan. 1989.

[82] G. Winkler, *Image Analysis, Random Fields and Dynamic Monte Carlo Methods: A Mathematical Introduction*. Berlin, Germany: Springer-Verlag, 1995.

[83] A. D. Wyner and J. Ziv, "Bounds on the rate-distortion function for stationary sources with memory," *IEEE Trans. Inform. Theory*, vol. IT-17, pp. 508–513, Sept. 1971.

[84] ——, "Some asymptotic properties of the entropy of a stationary ergodic data source with applications to data compression," *IEEE Trans. Inform. Theory*, vol. 35, pp. 1250–1258, Sept. 1989.

[85] ——, "Fixed data base version of the Lempel–Ziv data compression algorithm," *IEEE Trans. Inform. Theory*, vol. 37, pp. 878–880, May 1991.

[86] ——, "The sliding-window Lempel–Ziv algorithm is asymptotically optimal," *Proc. IEEE*, vol. 82, pp. 872–877, June 1994.

[87] A. D. Wyner, J. Ziv, and A. J. Wyner, "On the role of pattern matching in information theory. (Information theory: 1948–1998)," *IEEE Trans. Inform. Theory*, vol. 44, pp. 2045–2056, Sept. 1998.

[88] E.-H. Yang and J. C. Kieffer, "On the performance of data compression algorithms based upon string matching," *IEEE Trans. Inform. Theory*, vol. 44, pp. 47–65, Jan. 1998.

[89] E.-H. Yang and Z. Zhang, "The redundancy of source coding with a fidelity criterion—Part III: Coding at a fixed distortion level with unknown statistics," preprint.

[90] ——, "On the redundancy of lossy source coding with abstract alphabets," *IEEE Trans. Inform. Theory*, vol. 45, pp. 1092–1110, May 1999.

[91] ——, "The shortest common superstring problem: Average case analysis for both exact and approximate matching," *IEEE Trans. Inform. Theory*, vol. 45, pp. 1867–1886, Sept. 1999.

[92] ——, "The redundancy of source coding with a fidelity criterion—Part II: Coding at a fixed rate level with unknown statistics," *IEEE Trans. Inform. Theory*, vol. 47, pp. 126–145, Jan. 2001.

[93] Z. Ye and T. Berger, *Information Measures for Discrete Random Fields*. Beijing, China: Science, 1998.

[94] R. Zamir, "The index entropy of mismatched lossy source coding," *IEEE Trans. Inform. Theory*, vol. 48, pp. 523–528, Feb. 2002.

[95] R. Zamir and K. Rose, "A type generator model for adaptive lossy compression," in *Proc. IEEE Int. Symp. Information Theory*, Ulm, Germany, June/July 1997, p. 186.

[96] ——, "Natural type selection in adaptive lossy compression," *IEEE Trans. Inform. Theory*, vol. 47, pp. 99–111, Jan. 2001.

[97] Z. Zhang, E.-H. Yang, and V. K. Wei, "The redundancy of source coding with a fidelity criterion—Part I: Known statistics," *IEEE Trans. Inform. Theory*, vol. 43, pp. 71–91, Jan. 1997.

[98] J. Ziv and A. Lempel, "A universal algorithm for sequential data compression," *IEEE Trans. Inform. Theory*, vol. IT-23, pp. 337–343, May 1977.

[99] ——, "Compression of individual sequences by variable rate coding," *IEEE Trans. Inform. Theory*, vol. IT-24, pp. 530–536, Sept. 1978.