Pointwise Redundancy in Lossy Data Compression and Universal Lossy Data Compression

Ioannis Kontoyiannis, *Member, IEEE*

***Abstract*—We characterize the achievable pointwise redundancy rates for lossy data compression at a fixed distortion level. "Pointwise redundancy" refers to the difference between the description length achieved by an nth-order block code and the optimal $nR(D)$ bits. For memoryless sources, we show that the best achievable redundancy rate is of order $O(\sqrt{n})$ in probability. This follows from a second-order refinement to the classical source coding theorem, in the form of a "one-sided central limit theorem." Moreover, we show that, along (almost) any source realization, the description lengths of any sequence of block codes operating at distortion level D exceed $nR(D)$ by at least as much as $C\sqrt{n \log \log n}$, infinitely often. Corresponding direct coding theorems are also given, showing that these rates are essentially achievable. The above rates are in sharp contrast with the *expected* redundancy rates of order $O(\log n)$ recently reported by various authors. Our approach is based on showing that the compression performance of an arbitrary sequence of codes is essentially bounded below by the performance of Shannon's random code. We obtain partial generalizations of the above results for arbitrary sources with memory, and we prove lossy analogs of "Barron's Lemma."**

***Index Terms*—Large deviations, lossy data compression, rate-distortion, redundancy, universal coding.**

I. Introduction

BROADLY speaking, the objective of lossy data compression is to find efficient approximate representations for relatively large amounts of data. Let $x_1^n \stackrel{\Delta}{=} (x_1, x_2, \cdots, x_n)$ denote a data string generated by a random source $\boldsymbol{X} = \{X_n;\ n \geq 1\}$ taking values in the source alphabet A. We wish to represent each x_1^n by a corresponding string $y_1^n \stackrel{\Delta}{=} (y_1, y_2, \cdots, y_n)$ taking values in the reproduction alphabet $\hat{A}$ (where $\hat{A}$ may or may not be the same as A), so that the distortion between each data string and its representation lies within some fixed allowable range. For our purposes, distortion is measured by a family of single-letter distortion measures

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

where ρ: $A \times \hat{A} \to [0, \infty)$ is a fixed nonnegative function.

Manuscript received June 14, 1999; revised September 18, 1999. This work was supported in part by a grant from the Purdue Research Foundation. Preliminary results from this work were reported at the 1999 Canadian Workshop on Information Theory, Kingston, Ont., June 1999.

The author is with the Department of Statistics, Purdue University, 1399 Mathematical Sciences Building, W. Lafayette, IN 47907-1399 USA (e-mail: yiannis@stat.purdue.edu).

Communicated by N. Merhav, Associate Editor for Source Coding.
Publisher Item Identifier S 0018-9448(00)00074-2.

To be specific, we consider "variable-length block codes operating at a fixed distortion level," that is, codes C_n defined by triplets (B_n, ϕ_n, ψ_n) where

a) B_n is a subset of $\hat{A}^n$ called the *codebook*;
b) ϕ_n: $A^n \to B_n$ is the *encoder* or *quantizer*;
c) ψ_n: $B_n \to \{0, 1\}^*$ is an invertible (and prefix-free) representation of the elements of B_n by finite-length binary strings.

For $D \geq 0$, the block code $C_n = (B_n, \phi_n, \psi_n)$ is said to *operate at distortion level* D [14] (or to be D-*semifaithful* [23]), if it encodes each source string with distortion D or less

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

From the point of view of data compression, the main quantity of interest is the description length of a block code C_n, expressed in terms of its associated length function $\ell_n : A^n \to \mathbb{N}$. Here, $\ell_n(x_1^n)$ denotes the description length, in bits, assigned by C_n to the string x_1^n. Formally

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

Roughly speaking, the smaller the description length, the better the code.

Shannon in 1959 characterized the best achievable compression performance of block codes. Suppose, for example, that the data are generated by a memoryless source $\boldsymbol{X} = \{X_n; n \geq 1\}$, that is, the X_n are independent and identically distributed (i.i.d.) random variables with common distribution P on A. Suppose also that $\{C_n = (B_n, \phi_n, \psi_n); n \geq 1\}$ is an arbitrary sequence of block codes operating at distortion level D. In [28] Shannon identified the minimal *expected* description length that can be achieved by any such sequence $\{C_n\}$. He showed that the expected compression ratio $E[\ell_n(X_1^n)]/n$ is asymptotically bounded below by the *rate-distortion function* $R(D)$

$$\liminf_{n \to \infty} \frac{E[\ell_n(X_1^n)]}{n} \geq R(D) \quad \text{bits per symbol} \qquad (1)$$

where $R(D) = R(P, D)$ is defined by the well-known formula

$$R(D) = R(P, D) = \inf_{(X,Y):\ X \sim P,\ E[\rho(X,Y)] \leq D} I(X; Y).$$

(Precise definitions are given in the next section.) Moreover, Shannon demonstrated the existence of codes achieving the above lower bound with equality; see [28] or Berger's classic text [4].

A stronger version of Shannon's "converse" (1) was proved by Kieffer in 1991 [14], who showed that the rate-distortion

function is an asymptotic lower bound for $\ell_n(X_1^n)$ not just in expectation but also in the pointwise sense

$$\liminf_{n\to\infty} \frac{\ell_n(X_1^n)}{n} \geq R(D) \quad \text{a.s.} \tag{2}$$

(Here and throughout this paper the terms "pointwise," "almost surely" (denoted "a.s."), and "with probability one" are used interchangeably.) Kieffer's result says that, asymptotically, it is impossible to beat the rate-distortion function even on a small fraction of the messages generated by the source. In [14] it is also demonstrated that the bound in (2) can be achieved with equality.

Our main aim in this paper is to characterize the achievable *pointwise redundancy rates* for block codes applied to memoryless sources, where the pointwise redundancy is defined as the difference between the description length $\ell_n(X_1^n)$ of an nth-order block code C_n and the optimum description length given by $nR(D)$. Mathematically, this problem translates to describing the possible rates of convergence in (2), and, in particular, finding the fastest such rate. The main gist of our approach will be to show that the performance of any sequence of block codes operating at a fixed distortion level is bounded below by the performance of a (simple variant of) Shannon's random code.

In terms of data compression, knowing the possible convergence rates that can be achieved in(2) tells us how big blocks of data we need to take in order to come reasonably close to optimal compression performance. Clearly, these questions are of significant practical relevance.

A. Outline

For simplicity, assume for now that A and $\hat{A}$ are both finite sets, and let $\boldsymbol{X} = \{X_n; n \geq 1\}$ be a memoryless source with rate-distortion function $R(D)$. Our main results (Theorems 4 and 5, summarized in Corollaries 1 and 2) state that the performance of an arbitrary sequence of codes $\{C_n, \ell_n\}$ is essentially dominated by the performance of a random code, to which we refer as the "Shannon code."

(MAIN RESULT): Let Q_n^* be the optimal reproduction distribution at distortion level D, and write $B(x_1^n, D)$ for the distortion-ball of radius D around x_1^n (precise definitions are given in the next section). For any sequence of block codes $\{C_n\}$ operating at distortion level D, with associated length functions $\{\ell_n\}$, we have

$$\ell_n(X_1^n) \geq \log[1/Q_n^*(B(X_1^n, D))] + O(\log n) \quad \text{a.s.}$$

Moreover, the Shannon code asymptotically achieves this lower bound with equality.

(Throughout the paper, "log" denotes the logarithm taken to base 2 and "$\log_e$" denotes the natural logarithm.) Next, motivated by corresponding results in the case of lossless data compression [15], we interpret Kieffer's result (2) as a "one-sided" law of large numbers, and we state and prove corresponding second-order refinements to (2). In Theorem 1 we give a "one-sided" central limit theorem (CLT) corresponding to the pointwise lower bound in (2).

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

$$\ell_n(X_1^n) \geq nR(D) + \sqrt{n}G_n + O(\log n) \quad \text{a.s.} \tag{3}$$

where the G_n converge in distribution (as $n \to \infty$) to a Gaussian random variable. Moreover, there exist codes $\{C_n\}$ achieving the lower bound in (3) (see Theorem 2).

This means that for *any* sequence of codes, about half the time, the description length $\ell_n(X_1^n)$ will deviate from the optimum $nR(D)$ bits by $O(\sqrt{n})$ bits.

A further refinement to the pointwise converse (2) is also given in Theorem 1, in the form of a "one-sided" law of the iterated logarithm (LIL) (under some mild conditions). This provides a complete characterization of the pointwise redundancy of block codes at a fixed distortion level.

(LIL): For any sequence of codes $\{C_n, \ell_n\}$ operating at distortion level D, the pointwise redundancy exceeds $C\sqrt{n \log\log n}$ for infinitely many values of n (for some $C > 0$)

$$\ell_n(X_1^n) - nR(D) \geq C\sqrt{n \log\log n} \quad \text{infinitely often, a.s.} \tag{4}$$

Moreover, there exist codes $\{C_n\}$ asymptotically achieving this lower bound (Theorem 2).

The pointwise redundancy rates in (3) and (4) are in sharp contrast with the corresponding *expected* redundancy results recently reported by Zhang, Yang, and Wei in [39]. There, it shown that the best possible expected redundancy

$$E[\ell_n(X_1^n)] - nR(D)$$

achievable by block codes is of order $O(\log n)$. For practical purposes, this difference suggests the following interpretation: Since any compression algorithm used in practice is bound to have fluctuations in the description length of order at least as large as $O(\sqrt{n})$, for big enough block lengths n it may or may not be worth putting a lot of effort into optimizing the algorithm's *expected* performance. Instead, it might be more useful to either: a) try to control the variance of the description lengths $\ell_n(X_1^n)$ or b) optimize the algorithm's implementation. Indeed, it seems to often be the case in practice that "implementation complexity might be the dominating issue" [5].

Our next result says, perhaps somewhat surprisingly, that there is no cost for universality in pointwise redundancy. That is, essentially the same performance can be achieved, even when the source distribution is not known in advance. For the class of all memoryless sources P over the alphabet A, Theorem 3 demonstrates the existence of a sequence of universal codes $\{C_n^*\}$ with length functions $\{\ell_n^*\}$ such that, for every source P (and for some $C' > 0$)

a) $\ell_n^*(X_1^n)/n \to R(P, D)$ a.s.
b) $\ell_n^*(X_1^n) = nR(P, D) + \sqrt{n}G_n + O(\log n)$ a.s.
c) $\ell_n^*(X_1^n) - nR(P, D) \leq C'\sqrt{n \log\log n}$ eventually, a.s.

A natural next question to ask is whether these results remain true when sources with memory are considered. The fundamental coding theorems in (1) and (2) are, of course, still valid

(with the rate-distortion function now defined in terms of the distribution of the whole process $\boldsymbol{X}$), but redundancy questions appear to be much more delicate. For arbitrary sources with memory (not even necessarily stationary or ergodic), Theorem 6 gives a general pointwise lower bound for the performance of block codes at a fixed distortion level. This result can be thought of as the natural analog to the case of lossy compression of a well-known result from lossless compression, sometimes [29] referred to as "Barron's Lemma" [1], [2]. A more detailed discussion of this connection is given in Section II-D. Finally Theorem 8 is a direct coding theorem demonstrating a pointwise achievability result which complements the lower bound of Theorem 6.

B. History

Despite its obvious practical relevance, the redundancy problem for lossy data compression at a fixed distortion level seems to have only been considered relatively recently, and, with few exceptions, attention has been restricted to questions regarding expected redundancy.

In 1993, Yu and Speed [38] demonstrated the existence of a sequence of universal codes with expected redundancy rate of order $O(\log n)$ over the class of memoryless sources with finite source and reproduction alphabets. In the case of the Hamming distortion measure, Merhav in 1995 [21] proved a corresponding lower bound showing that the expected redundancy (even when the source distribution is known in advance) is bounded below by $(1/2)\log n$. The question was essentially settled by the work of Zhang, Yang, and Wei [39] in 1997, where it is demonstrated that Merhav's lower bound is true quite generally, and corresponding direct coding theorems are given, exhibiting codes with redundancy bounded above by $[\log n + o(\log n)]$. A similar direct coding theorem for sources with abstract alphabets was recently proved by Yang and Zhang [35]. For universal coding at a fixed distortion level, Chou, Effros, and Gray [8] showed that the price paid for being universal over k-dimensional parametric classes of sources is essentially $(k/2)\log n$. A universal direct coding theorem for memoryless sources over finite alphabets was recently reported by Yang and Zhang in [37].

With only a couple of notable exceptions from 1968 (Pilc [24] and Wyner [32]), the dual problem of lossy compression at a fixed-rate level appears to also have been considered rather recently. Linder, Lugosi, and Zeger [19], [20] studied various aspects of the *distortion redundancy* problem and exhibited universal codes with distortion redundancy of order $O(\log n)$. Zhang, Yang, and Wei [39] proved a lower bound of order $O(\log n)$, and they constructed codes achieving this lower bound (to first order). Coding for sources with abstract alphabets is considered in [35], and questions of universality are treated in [8] and [36], among many others.

The rest of the paper is organized as follows. In the next section we state and discuss our main results. Section III contains the proofs of the pointwise converses for memoryless sources (Theorems 1 and 4), and Section IV contains the proofs of the corresponding direct coding theorems (Theorems 2, 3, and 5). In Section V, we prove our results for arbitrary sources (Theorems 6–8), and the Appendices contain proofs of various technical steps needed along the way.

II. Results

Let $\boldsymbol{X} = \{X_n;\, n \geq 1\}$ be a random source taking values in the *source alphabet* A, where A is assumed to be a Polish space (i.e., a complete, separable metric space); let $\mathcal{A}$ denote its associated Borel σ-field. Although all our results will be stated for the general case, there is no essential "loss of ideas" in thinking of A as being finite. For $1 \leq i \leq j \leq \infty$, write X_i^j for the vector of random variables $(X_i, X_{i+1}, \cdots, X_j)$ and similarly write $x_i^j = (x_i, x_{i+1}, \cdots, x_j) \in A^{j-i+1}$ for a realization of X_i^j.

Let $\hat{A}$ denote the *reproduction alphabet*. Given a nonnegative measurable function $\rho : A \times \hat{A} \to [0, \infty)$, we define a sequence of single-letter distortion measures $\rho_n : A^n \times \hat{A}^n \to [0, \infty)$, $n \geq 1$, by

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

Throughout the paper we will assume that the set $\hat{A}$ is finite, and that the function ρ is bounded, i.e., $\rho(x, y) \leq M < \infty$ for some fixed constant M, for all $x \in A$, $y \in \hat{A}$. Although these assumptions are not necessary for the validity of all of our results, they are made here for the sake of simplicity of the exposition. We also make the customary assumption that

$$\sup_{x \in A} \min_{y \in \hat{A}} \rho(x, y) = 0. \tag{5}$$

We are interested in *variable-length block codes C_n operating at a fixed distortion level*, where $C_n = (B_n, \phi_n, \psi_n)$ is defined in terms of a subset B_n of $\hat{A}^n$ called the *codebook*, an *encoder* $\phi_n : A^n \to B_n$, and a lossless (prefix-free) binary code $\psi_n : B_n \to \{0, 1\}^*$ for B_n. For $D \geq 0$, we say that the code C_n *operates at distortion level* D, if $\rho_n(x_1^n, \phi_n(x_1^n)) \leq D$ for all source strings $x_1^n \in A^n$. The *length function* $\ell_n : A^n \to \mathbb{N}$ induced by C_n is defined by

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

so that $\ell_n(x_1^n)$ is the length (in bits) of the description of x_1^n by C_n.

For $D \geq 0$ and $n \geq 1$, the *nth-order rate-distortion function* of $\boldsymbol{X}$ (see, e.g., [4]) is defined by

$$R_n(D) = \inf_{(X_1^n, Y_1^n)} I(X_1^n; Y_1^n)$$

where $I(X_1^n; Y_1^n)$ denotes the mutual information (in bits) between X_1^n and Y_1^n, and the infimum is over all jointly distributed random vectors (X_1^n, Y_1^n) with values in $A^n \times \hat{A}^n$, such that X_1^n has the source distribution and $E[\rho_n(X_1^n, Y_1^n)] \leq D$; if there are no such (X_1^n, Y_1^n), we let $R_n(D) = \infty$. (Similarly, throughout the paper, the infimum of an empty set is taken to be $+\infty$.) The *rate-distortion function* $R(D)$ of $\boldsymbol{X}$ is defined as the limit of $(1/n)R_n(D)$ as $n \to \infty$, provided the limit exists.

A. Second-Order Coding Theorems for Memoryless Sources

In this section we assume that $\boldsymbol{X}$ is a memoryless source with *fixed* distribution P. That is, the random variables $\{X_n\}$ are i.i.d. according to P, where, strictly speaking, P is a probability

measure on $(A, \mathcal{A})$. As is well known [4], the rate-distortion function of a memoryless source reduces to its first-order rate-distortion function

$$R(D) = R(P, D) = \inf_{(X,Y)} I(X; Y) \tag{6}$$

where the infimum is over all jointly distributed random variables (X, Y) such that X has distribution P and $E[\rho(X, Y)] \le D$. Let

$$D_{\max} = D_{\max}(P) = \min_{y \in \hat{A}} E_P[\rho(X, y)] \tag{7}$$

and note that $R(D) = 0$ for $D \ge D_{\max}$ (see, e.g., Proposition 1-iv) in Section III). In order to avoid the trivial case when $R(D)$ is identically zero, we assume that $D_{\max} > 0$.

In our first result, Theorem 1, we give lower bounds on the pointwise deviations of the description lengths $\ell_n(X_1^n)$ of any code C_n from the optimum $nR(D)$ bits. It is proved in Section III-B by an application of the general lower bound in Theorem 4.

Theorem 1. Second-Order Converses: Let $\boldsymbol{X}$ be a memoryless source with rate-distortion function $R(D)$, and let $D \in (0, D_{\max})$.

i) 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

$$\ell_n(X_1^n) - nR(D) \ge \sqrt{n} G_n - 2 \log n \quad \text{eventually, a.s.} \tag{8}$$

and the G_n converge in distribution to a Gaussian random variable

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

with variance σ^2 explicitly identified.

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

$$\limsup_{n \to \infty} \frac{\ell_n(X_1^n) - nR(D)}{\sqrt{2n \log_e \log_e n}} \ge \sigma \quad \text{a.s.}$$
$$\liminf_{n \to \infty} \frac{\ell_n(X_1^n) - nR(D)}{\sqrt{2n \log_e \log_e n}} \ge -\sigma \quad \text{a.s.}$$

(Recall that $\log_e$ denotes the natural logarithm and $\log \equiv \log_2$.) Our next result, Theorem 2, shows that these lower bounds are tight. It is proved in Section IV using a random coding argument. Although the construction is essentially identical to Shannon's classical argument, determining its pointwise asymptotic behavior is significantly more delicate, and it relies heavily on the recent results of Dembo and Kontoyiannis [10] and Yang and Zhang [35] on the asymptotics of the probability of "distortion balls." See the discussion after Theorem 4.

Theorem 2. Direct Coding Theorem: Let $\boldsymbol{X}$ be a memoryless source with rate-distortion function $R(D)$, and let $D \in (0, D_{\max})$. There is a sequence of codes $\{C_n, \ell_n\}$ operating at distortion level D, which achieve asymptotic equality (to first order) in all the almost-sure statements of Theorem 1

a) $$\lim_{n \to \infty} \left[\frac{\ell_n(X_1^n) - nR(D)}{\sqrt{n}} - G_n \right] = 0 \quad \text{a.s.}$$

b) $$\limsup_{n \to \infty} \frac{\ell_n(X_1^n) - nR(D)}{\sqrt{2n \log_e \log_e n}} = \sigma \quad \text{a.s.}$$

c) $$\liminf_{n \to \infty} \frac{\ell_n(X_1^n) - nR(D)}{\sqrt{2n \log_e \log_e n}} = -\sigma \quad \text{a.s.}$$

Remarks:

1) Variance: The variance σ^2 in Theorems 1 and 2 is a quantity characteristic of the source, which tells us that, when the source is encoded in the most efficient way, the deviations of the codeword lengths $\ell_n(X_1^n)$ from the optimum $nR(D)$ bits will have a variance roughly equal to $n\sigma^2$. If any other code is used, these deviations will be asymptotically bounded below by a Gaussian random variable of variance $n\sigma^2$. In view of this, we think of $\sigma^2 = \sigma^2(P, D)$ as the *minimal coding variance* of the source P at distortion level D. The precise definition of σ^2 is given in the next section and its properties are discussed in some detail in Section II-C. In particular, σ^2 is always nonnegative (typically it is strictly positive), and it can be expressed as

$$\sigma^2 = \mathrm{Var}(-\log F(X_1)) \tag{9}$$

for some function F: $A \to (0, \infty)$.

2) Pointwise Redundancy: Let $\{C_n, \ell_n\}$ be arbitrary codes operating at distortion level D. If $\sigma^2 > 0$, part ii) of Theorem 1 says that when the codes $\{C_n\}$ are applied to almost any realization of the source $\boldsymbol{X}$, then for infinitely many n

$$\ell_n(X_1^n) - nR(D) \ge C\sqrt{n \log_e \log_e n} \tag{10}$$

where for C we can take any constant $C \in (0, \sqrt{2}\sigma)$. Moreover, the amount by which we can "beat" the rate-distortion function satisfies

$$\ell_n(X_1^n) - nR(D) \ge -C\sqrt{n \log_e \log_e n} \quad \text{eventually, a.s.}$$

The $O(\sqrt{n \log_e \log_e n})$ rate in (10) is in sharp contrast with the *expected* redundancy rates of order $O(\log n)$ reported in [39].

3) Expected Versus Pointwise Redundancy: The difference between the two types of redundancy is reminiscent of the classical bias/variance tradeoff in statistics. Here, if the goal is to design a lossy compression algorithm that will be used repeatedly and on large data sets, then it is probably a good idea to optimize the expected performance. On the other hand, if it is important to guarantee compression performance within certain bounds, it might be possible to give up some rate in order to reduce the variance.

4) Lossless Compression: The results in Theorem 1 are close parallels of the corresponding lossless compression results in [15, Theorems 1 and 2]. There, the coding variance takes the simple form

$$\sigma^2 = \mathrm{Var}(-\log P(X_1)) \tag{11}$$

(cf. (9) above), which can be viewed as the natural second-order analog of the entropy $H = E(-\log P(X_1))$. In the lossless

case, the pointwise lower bounds are easily achieved, for example, by the Huffman code or the Shannon code [9]. In fact, it is well known [18], [30] that we can come within $O(\log n)$ of the Shannon code universally over all memoryless sources, for *all* message strings x_1^n. Therefore, in the lossless case, the same pointwise behavior can be achieved universally at no extra cost [15].

Next we show that the pointwise redundancy rates of Theorem 2 can be achieved universally over all memoryless sources on A. The proof of Theorem 3 (Section IV) is similar in spirit to that of Theorem 2, with the difference that here, in order to be universal, we generate multiple random codebooks and we allow the encoder to choose the best one. The additional cost of transmitting the index of the codebook that was used turns out to be negligible, and the pointwise behavior obtained is identical (up to terms of order $O(\log n)$) to that achieved with knowledge of the source distribution. The idea of multiple random codebooks is well known in information theory, dating at least as far back as Ziv's 1972 paper [40] and the work of Neuhoff, Gray, and Davisson in 1975 [22]. Nevertheless, to determine the exact pointwise behavior of this random code is more delicate, and our analysis relies on recent results from [10] and [35].

Theorem 3. Universal Coding: There is a sequence of universal codes $\{C_n^*, \ell_n^*\}$ operating at distortion level D, such that, if the data $(X_1, X_2, \cdots)$ are generated by *any* memoryless source P on A, and if $D \in (0, D_{\max}(P))$, then

$$\text{a}') \quad \lim_{n\to\infty} \left[\frac{\ell_n^*(X_1^n) - nR(P, D)}{\sqrt{n}} - G_n(P, D) \right] = 0, \quad P\text{-a.s.}$$

$$\text{b}') \quad \limsup_{n\to\infty} \frac{\ell_n^*(X_1^n) - nR(P, D)}{\sqrt{2n \log_e \log_e n}} = \sigma(P, D), \quad P\text{-a.s.}$$

$$\text{c}') \quad \liminf_{n\to\infty} \frac{\ell_n^*(X_1^n) - nR(P, D)}{\sqrt{2n \log_e \log_e n}} = -\sigma(P, D), \quad P\text{-a.s.}$$

where the random variables $G_n = G_n(P, D)$ and the variance $\sigma^2 = \sigma^2(P, D)$ are as in Theorem 1.

B. Main Results: Pointwise Optimality of the Shannon Code

In this section we state our main results, from which Theorems 1–3 of the previous section will follow.

Assume that $\boldsymbol{X}$ is a memoryless source with distribution P on A, and let Q be an arbitrary measure on $\hat{A}$; since $\hat{A}$ is a finite set, we think of Q simply as a discrete probability mass function (p.m.f.). For each Q, define

$$R(P, Q, D) = \inf_{(X,Y)} [I(X; Y) + H(Q_Y || Q)] \tag{12}$$

where $H(R||Q)$ denotes the relative entropy (in bits) between two distributions R and Q, Q_Y denotes the distribution of 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 and $E[\rho(X, Y)] \le D$. It is easy to see that the rate-distortion function of $\boldsymbol{X}$ can be expressed as

$$R(D) = R(P, D) = \inf_Q R(P, Q, D) \tag{13}$$

where the infimum is over all p.m.f.'s Q on $\hat{A}$ (simply interchange the two infima).

For each source P on A and distortion level $D \ge 0$, let $Q^* = Q^*(P, D)$ denote a p.m.f. achieving the infimum in (13)

$$R(D) = R(P, Q^*, D). \tag{14}$$

(See Proposition 2 part ii) in Section III-A for the existence of Q^*.) We call this Q^* the *optimal reproduction distribution* for P at distortion level D.

For a fixed source P, a distortion level $D \in (0, D_{\max})$, and a corresponding Q^* as in (14), we let $\Lambda_x(\lambda)$, $x \in A$, $\lambda \le 0$, be the log-moment generating function of the random variable $\rho(x, Y)$ when $Y \sim Q^*$

$$\Lambda_x(\lambda) = \log_e E_{Q^*} \left(e^{\lambda \rho(x, Y)} \right), \qquad x \in A, \lambda \le 0.$$

Then there exists a unique $\lambda = \lambda^* < 0$ such that

$$\frac{d}{d\lambda} [E_P(\Lambda_X(\lambda))] = D$$

(see Lemma 1 in Section III-A).

Our next result, Theorem 4, shows that the pointwise redundancy of any sequence of block codes is essentially bounded below by a sum of i.i.d. random variables. As we discuss in the remarks following Theorem 4, this lower bound can be interpreted as saying that the performance of any sequence of block codes is dominated by the performance of Shannon's random code. Theorem 4 is proved in Section III-B.

Theorem 4. Pointwise Lower Bound: Let $\boldsymbol{X}$ be a memoryless source with rate-distortion function $R(D)$, and let $D \in (0, D_{\max})$. For any sequence of codes $\{C_n, \ell_n\}$ operating at distortion level D and any sequence $\{b_n\}$ of positive constants such that $\sum_n 2^{-b_n} < \infty$, we have

$$\ell_n(X_1^n) - nR(D) \ge \sum_{i=1}^n f(X_i) - b_n \quad \text{eventually, a.s.} \tag{15}$$

where

$$f(x) \triangleq (\log e)(-\Lambda_x(\lambda^*) - E_P[-\Lambda_X(\lambda^*)]). \tag{16}$$

Remarks:

1) Consequences: It is easy to see that Theorem 1 is an immediate consequence of the lower bound (15). In particular, the coding variance σ^2 in Theorems 1 and 2 is simply the variance of the random variable $f(X_1)$.

2) Intuition: Suppose we generate a random (Shannon) codebook according to Q^*, that is, we generate i.i.d. codewords

$$Y(i) = (Y_{i,1}, Y_{i,2}, \cdots, Y_{i,n}), \qquad i = 1, 2, \cdots$$

each drawn from the distribution $Q_n^* = (Q^*)^n$. We can encode each source sequence X_1^n by specifying the index $i = W_n$ of the first codeword $Y(i)$ such that $\rho_n(X_1^n, Y(i)) \le D$. This description takes approximately $(\log W_n)$ bits. But W_n,

the "waiting time" until the first D-close match for X_1^n, is approximately equal to the reciprocal of the probability of finding such a match, so

$$\log W_n \approx \log [1/Q_n^*(B(X_1^n, D))]$$

where the "distortion balls" $B(x_1^n, D)$ are defined by

$$B(x_1^n, D) = \left\{ y_1^n \in \hat{A}^n \colon \rho_n(x_1^n, y_1^n) \le D \right\}, \qquad x_1^n \in A^n. \tag{17}$$

From the recent work of Dembo and Kontoyiannis [10] and Yang and Zhang [35] we know that these probabilities behave like

$$\log [1/Q_n^*(B(X_1^n, D))] = nR(D) + \sum_{i=1}^{n} f(X_i) + \tfrac{1}{2} \log n + O(\log \log n) \quad \text{a.s.} \tag{18}$$

(See Proposition 3 in Section IV.) Therefore, the pointwise description length of the Shannon code is, approximately,

$$\log W_n \approx nR(D) + \sum_{i=1}^{n} f(X_i) \quad \text{bits a.s.}$$

In view of this, we can rephrase Theorem 4 by saying that, in a strong sense, *the performance of any code is bounded below by the performance of the Shannon code.* Indeed, the proof of Theorem 4 is based on first showing that any code can be thought of as a *random* code (according to a measure on $\hat{A}^n$ different from Q_n^*), and then proving that the pointwise performance of any random code is dominated by the performance of the Shannon code.

The following result, Theorem 5, formalizes the above random coding argument, and also shows that essentially the same performance can be achieved universally over all memoryless sources.

Theorem 5. The Shannon Code—Random Coding:

i) Let $\boldsymbol{X}$ be a memoryless source with rate-distortion function $R(D)$, and let $D \in (0, D_{\max})$. There is a sequence of codes $\{C_n, \ell_n\}$ operating at distortion level D, such that

$$\ell_n(X_1^n) \le \log [1/Q_n^*(B(X_1^n, D))] + 4 \log n + \text{Const.} \quad \text{eventually, a.s.}$$

where Q^* is the optimal reproduction distribution at distortion level D, and $Q_n^* = (Q^*)^n$.

ii) There is a sequence of universal codes $\{C_n^*, \ell_n^*\}$ operating at distortion level D, such that, if the data $(X_1, X_2, \cdots)$ are generated by *any* memoryless source P on A, and if $D \in (0, D_{\max}(P))$, then

$$\ell_n^*(X_1^n) \le \log [1/Q_n^*(B(X_1^n, D))] + (4+k) \log n + \text{Const.} \quad \text{eventually, } P\text{-a.s.}$$

where k is the number of elements in $\hat{A}$, $Q^* = Q^*(P, D)$ is the optimal reproduction distribution corresponding to the true source P at distortion level D, and $Q_n^* = (Q^*)^n$.

Next, in Corollary 1 we combine Theorem 4 (with $b_n = (1+\delta) \log n$) with (18) and Theorem 5 part i), to rewrite the above results in an intuitively more appealing form. And in Corollary 2 we point out that from the proof of Theorem 4 we can read off a lower bound on the performance of an arbitrary sequence of codes, which holds for any finite block length n. Although the two corollaries are simple consequences of Theorems 4 and 5, conceptually they contain the main contributions of this paper.

Corollary 1. Pointwise Optimality of Shannon Code: Under the assumptions of Theorem 4, for any sequence of codes $\{C_n, \ell_n\}$ operating at distortion level D, we have

$$\ell_n(X_1^n) \ge \log [1/Q_n^*(B(X_1^n, D))] - 2 \log n \quad \text{eventually, a.s.} \tag{19}$$

where the Shannon code achieves

$$\ell_n(X_1^n) \le \log [1/Q_n^*(B(X_1^n, D))] + 4 \log n + \text{Const.} \quad \text{eventually, a.s.}$$

Corollary 2. (Nonasymptotic Lower Bound): Under the assumptions of Theorem 4, for any block length n and any constant $c > 0$, if the code (C_n, ℓ_n) operates at distortion level D, then

$$\Pr \left\{ \ell_n(X_1^n) \le -\log E_{Q_n^*} \left(e^{n\lambda^*(\rho_n(X_1^n, Y_1^n) - D)} \right) - c \right\} \le 2^{-c}.$$

In the case of lossless compression, this reduces to Barron's lower bound (see [2, eq. (3.5)])

$$\Pr\{\ell_n(X_1^n) \le -\log P(X_1^n) - c\} \le 2^{-c}.$$

C. Minimal Coding Variance

Suppose $\boldsymbol{X}$ is a memoryless source with distribution P, let $D \in (0, D_{\max})$, and let Q^* be the corresponding optimal reproduction distribution. In the notation of the previous section, the minimal coding variance $\sigma^2 = \text{Var}[f(X_1)]$ can be written as

$$\sigma^2 = \text{Var}_P \left[-\log E_{Q^*} \left(e^{\lambda^*[\rho(X,Y) - D]} \right) \right]$$

(here X and Y are independent random variables with distributions P and Q^*, respectively). As we will see in Section III-A, the rate-distortion function $R(D)$ can similarly be expressed as

$$R(D) = E_P \left[-\log E_{Q^*} \left(e^{\lambda^*[\rho(X,Y) - D]} \right) \right].$$

Comparing the last two expressions suggests that we may think of σ^2 as a second-order version of $R(D)$, and further justifies the term minimal coding variance, by analogy to the minimal coding rate $R(D)$.

It is obvious that σ^2 is always nonnegative, and it is typically strictly positive, since the only way it can be zero is if the expectation $E_{Q^*}(e^{\lambda^* \rho(x,Y)})$ is constant for P-almost all $x \in A$. We give three simple examples illustrating this.

Example 1. Lossless Compression: As mentioned above, in the case of lossless compression the minimal coding variance reduces to $\sigma^2 = \text{Var}[-\log P(X_1)]$ (cf. (11)), from which it is

immediate that $\sigma^2 = 0$ if and only if P is the uniform distribution over the finite alphabet A. (See [15] for more details and a corresponding characterization for Markov sources.)

Example 2. Binary Source, Hamming Distortion: This is the simplest nontrivial lossy example. Suppose $\boldsymbol{X}$ has Bernoulli(p) distribution for some $p \in (0, 1/2]$. Let $A = \hat{A} = \{0, 1\}$ and let ρ be the Hamming distortion measure, $\rho(x, y) = |x-y|$. For $D \in (0, p)$, easy but tedious calculations (cf. [4, Example 2.7.1] and [9, Theorem 13.3.1]) show that Q^* is a Bernoulli(q) distribution with $q=(p-D)/(1-2D)$, $\lambda^* = \log_e (D/(1-D))$, and

$$E_{Q^*}\left(e^{\lambda^* \rho(x, Y)}\right) = \frac{P(x)}{1-D}.$$

As we already noted, $\sigma^2 = 0$ if and only if the above expression is constant in x, so that, here, $\sigma^2 = 0$ if and only if $p = 1/2$, i.e., if and only if P is the uniform distribution on $A = \{0, 1\}$.

Example 3. A Quaternary Example: This is a standard example from Berger's text [4, Example 2.7.2]. Suppose $\boldsymbol{X}$ takes values in the alphabet $A = \{1, 2, 3, 4\}$ and let $\hat{A} = A$. Suppose the distribution of $\boldsymbol{X}$ is given by $P = (p/2, (1-p)/2, (1-p)/2, p/2)$, for some $p \in (0, 1/2]$, and let the distortion measure ρ be specified by the matrix $(\rho_{ij}) = (\rho(i, j))$, $i, j \in A$, where

$$(\rho_{ij}) = \begin{pmatrix} 0 & 1/2 & 1/2 & 1 \\ 1/2 & 0 & 1 & 1/2 \\ 1/2 & 1 & 0 & 1/2 \\ 1 & 1/2 & 1/2 & 0 \end{pmatrix}.$$

For $D \in (0, (1-\sqrt{1-2p})/2)$ it is possible (although tedious) to calculate Q^* and λ^* explicitly, and to obtain that

$$E_{Q^*}\left(e^{\lambda^* \rho(x, Y)}\right) = \frac{P(x)}{(1-D)^2}.$$

Once again, this is generally not constant in x, with the exception of the case $p = 1/2$. So, σ^2 will be strictly positive, unless P is the uniform distribution on $A = \{1, 2, 3, 4\}$.

There is an obvious trend in all three examples above: *The variance σ^2 is strictly positive, unless P is uniformly distributed over A.* It is an interesting problem to determine how generally this pattern persists.

D. Sources with Memory

Here we present analogs of Theorems 4 and 5 for arbitrary sources. Of course, at this level of generality, the results we get are not as strong as the ones in the memoryless case. Still, adopting a different approach, we are able to get interesting partial generalizations.

Let $\boldsymbol{X}$ be an arbitrary source with values in A, and let P_n denote the distribution of X_1^n. By a *subprobability measure* Q_n on $\hat{A}^n$ we mean a positive measure with total mass $0 < Q_n(\hat{A}^n) \leq 1$. For each $D \geq 0$ we define (recall the notation in (17))

$$K_n(D) = K_n(P, D) = \inf_{Q_n} E_{P_n}\{-\log Q_n(B(X_1^n, D))\}$$

where the infimum is over all subprobability measures Q_n.

Suppose $\tilde{Q}_n$ achieves the above infimum (the existence of such a Q_n is established by Lemma 3 in Section V). Our next result gives an analog of Theorem 4 for the case of sources with memory. It is proved in Section V using an argument similar to the one used by Kieffer in the proof of Theorem 2 in [14].

Theorem 6. A Lossy "Barron's Lemma": Suppose $\boldsymbol{X}$ is an arbitrary source, and let $D \geq 0$ be such that $K_n(D) < \infty$. For any sequence of codes $\{C_n, \ell_n\}$ operating at distortion level D, we have the following.

i) For all n

$$E[\ell_n(X_1^n)] \geq K_n(D) \geq R_n(D).$$

ii) For any sequence $\{b_n\}$ of positive constants such that $\sum_n 2^{-b_n} < \infty$

$$\ell_n(X_1^n) \geq \log\left[1/\tilde{Q}_n(B(X_1^n, D))\right] - b_n \quad \text{eventually, a.s.} \tag{20}$$

The lower bound in (20) is a natural "lossy analog" of a well-known result from lossless compression, often called "Barron's Lemma" [1], [2]. Barron's Lemma states that for any sequence of lossless codes $\{C_n, \ell_n\}$

$$\ell_n(X_1^n) \geq \log[1/P_n(X_1^n)] - b_n \quad \text{eventually, a.s.}$$

Similarly, we can interpret the lower bound of Corollary 1 (19) as a different generalization of Barron's Lemma, valid only for memoryless sources. The reason why (19) is preferable over (20) is because the Q_n^* are product measures, whereas it is apparently hard to characterize the measures $\tilde{Q}_n$ in general. For example, in the case of memoryless sources one would expect that, for large n, the measures $\tilde{Q}_n$ "converge" to the measures Q_n^* in some sense. The only way in which we have been able to make this intuition precise is by proving the following result asserting the asymptotic equivalence between the compression performance of $\tilde{Q}_n$ and Q_n^*.

Theorem 7. Equivalence of Measures: Let $\boldsymbol{X}$ be a memoryless source with distribution P, and let $D \in (0, D_{\max})$. Then

$$\log\left[1/\tilde{Q}_n(B(X_1^n, D))\right] = \log[1/Q_n^*(B(X_1^n, D))] + O(\log n) \quad \text{a.s.}$$

Clearly, Theorem 7 can be combined with the recent results on the probabilities of D-balls mentioned in (18), to give alternative proofs of the pointwise converses in Theorems 1 and 4. Finally, we state a direct coding theorem, demonstrating that the lower bound in Theorem 6 is asymptotically tight. It is proved in Section V using a random coding argument.

Theorem 8. Partial Achievability: Suppose $\boldsymbol{X}$ is an arbitrary source, and let $D \geq 0$ be such that $K_n(D) < \infty$. Then there is a sequence of codes $\{C_n, \ell_n\}$ operating at distortion level D, such that

$$\ell_n(X_1^n) \leq \log\left[1/\tilde{Q}_n(B(X_1^n, D))\right] + 2\log n + 2\log\log\left[\frac{2n^2}{\tilde{Q}_n(B(X_1^n, D))}\right] + \text{Const.} \quad \text{eventually, a.s.}$$

Theorem 8, combined with Theorem 7 and (18), provides alternative proofs for Theorem 2 and Theorem 5 part i).

Finally, we remark that, for "nice" measures $\tilde{Q}_n$, it is natural to expect that the probabilities $\tilde{Q}_n(B(X_1^n, D))$ will decay to zero exponentially fast. This would be true, for example, if the $\tilde{Q}_n$ were the finite-dimensional marginals of a "nice" process, like a Markov chain [34] or a process with "rapid mixing" properties [7]. In that case, the "$\log\log$" term in Theorem 8 would grow like $\log n$, implying that the lower bound of Theorem 7 is tight up to terms of order $O(\log n)$.

III. Converses for Memoryless Sources

In Section III-A we collect some useful technical facts, and in Section III-B we prove Theorem 4 and use it to deduce Theorem 1.

A. Representations and Properties of $R(P, Q, D)$

For $n \geq 1$, let μ and ν be arbitrary probability measures on A^n and $\hat{A}^n$, respectively (of course, since $\hat{A}$ is a finite set, ν is a discrete p.m.f. on $\hat{A}^n$). Write X_1^n for a random vector with distribution μ on A^n, and Y_1^n for an independent random vector with distribution ν on $\hat{A}^n$. Let $S_n = \{y_1^n \in \hat{A}^n \colon \nu(y_1^n) > 0\} \subseteq \hat{A}^n$ denote the support of ν, and define

$$D_{\min}^{\mu,\nu} = E_\mu \left[\min_{y_1^n \in S_n} \rho_n(X_1^n, y_1^n) \right]$$

$$D_{\max}^{\mu,\nu} = E_{\mu\times\nu}\left[\rho_n(X_1^n, Y_1^n)\right].$$

Clearly, $0 \leq D_{\min}^{\mu,\nu} \leq D_{\max}^{\mu,\nu} < \infty$. For $\lambda \leq 0$, we define

$$\Lambda_{\mu,\nu}(\lambda) = E_\mu \left[\log_e E_\nu \left(e^{\lambda \rho_n(X_1^n, Y_1^n)} \right) \right]$$

and, for $D \geq 0$, we write $\Lambda_{\mu,\nu}^*$ for the Fenchel–Legendre transform of $\Lambda_{\mu,\nu}$

$$\Lambda_{\mu,\nu}^*(D) = \sup_{\lambda \leq 0} \left[\lambda D - \Lambda_{\mu,\nu}(\lambda)\right].$$

In analogy with (12), we also define

$$R(\mu, \nu, D) = \inf_{(X_1^n, Z_1^n)} \left[I(X_1^n; Z_1^n) + H(Q_{Z_1^n} || \nu) \right]$$

where $H(R||Q)$ denotes the relative entropy (in bits) between two distributions R and Q, $Q_{Z_1^n}$ denotes the distribution of Z_1^n, and the infimum is over all jointly distributed random vectors (X_1^n, Z_1^n) with values in $A^n \times \hat{A}^n$ such that X_1^n has distribution μ and $E[\rho_n(X_1^n, Z_1^n)] \leq D$. In the next lemma we collect various standard properties of $\Lambda_{\mu,\nu}$ and $\Lambda_{\mu,\nu}^*$. Parts i)–iii) can be found, e.g., in [10] or [17]; part iv) is proved in Appendix I.

Lemma 1:

i) $\Lambda_{\mu,\nu}$ is infinitely differentiable on $(-\infty, 0)$, $\Lambda'_{\mu,\nu}(0) = D_{\max}^{\mu,\nu}$, and $\Lambda'_{\mu,\nu}(\lambda) \to D_{\min}^{\mu,\nu}$ as $\lambda \to -\infty$.

ii) $\Lambda''_{\mu,\nu}(\lambda) \geq 0$ for all $\lambda \leq 0$; if, moreover, $D_{\min}^{\mu,\nu} < D_{\max}^{\mu,\nu}$, then $\Lambda''_{\mu,\nu}(\lambda) > 0$ for all $\lambda \leq 0$.

iii) If $D_{\min}^{\mu,\nu} < D_{\max}^{\mu,\nu}$ and $D \in (D_{\min}^{\mu,\nu}, D_{\max}^{\mu,\nu})$, then there exists a unique $\lambda < 0$ such that $\Lambda'_{\mu,\nu}(\lambda) = D$ and $\Lambda_{\mu,\nu}^*(D) = \lambda D - \Lambda_{\mu,\nu}(\lambda)$.

iv) For every $\lambda \leq 0$ and every probability measure μ on A, $\Lambda_{\mu,\nu}(\lambda)$ is upper semicontinuous as a function of ν.

In the following propositions we give two alternative representations of the function $R(\mu, \nu, D)$, and state several of its properties. Propositions 1 and 2 are proved in Appendices II and III, respectively.

Proposition 1. Representations of $R(\mu, \nu, D)$:

i) For all $D \geq 0$

$$R(\mu, \nu, D) = \inf_{\Theta} \; E_\mu \left[H(\Theta(\cdot|X_1^n) || \nu(\cdot)) \right]$$

where the infimum is taken over all probability measures Θ on $A^n \times \hat{A}^n$ such that the A^n-marginal of Θ equals μ and $E_\Theta[\rho_n(X_1^n, Y_1^n)] \leq D$.

ii) For all $D \geq 0$ we have

$$R(\mu, \nu, D) = (\log e)\Lambda_{\mu,\nu}^*(D).$$

Proposition 2. Properties of $R(\mu, \nu, D)$:

i) For every $D \geq 0$ and every probability measure μ on A^n, $R(\mu, \nu, D)$ is lower semicontinuous as a function of ν.

ii) For every $D \geq 0$, there exists a p.m.f. $Q = Q^*$ on $\hat{A}$ achieving the infimum in (14).

iii) For $D < D_{\min}^{\mu,\nu}$, $R(\mu, \nu, D) = \infty$; for $D_{\min}^{\mu,\nu} < D < D_{\max}^{\mu,\nu}$, $0 < R(\mu, \nu, D) < \infty$; and for $D \geq D_{\max}^{\mu,\nu}$, $R(\mu, \nu, D) = 0$.

iv) For $0 < D < D_{\max}$ we have $0 < R(D) < \infty$, whereas for $D \geq D_{\max}$, $R(D) = 0$.

v) If $D \in (0, D_{\max})$, then $D_{\min}^{P,Q^*} < D_{\max}^{P,Q^*}$ and $D \in (D_{\min}^{P,Q^*}, D_{\max}^{P,Q^*})$.

B. Proofs of Converses

Proof of Theorem 1 from Theorem 4: Taking $b_n = 2\log n$ in Theorem 4, yields

$$\ell_n(X_1^n) - nR(D) = \sum_{i=1}^n f(X_i) - 2\log n \quad \text{eventually, a.s.} \tag{21}$$

Writing $G_n = (1/\sqrt{n}) \sum_{i=1}^n f(X_i)$ we get (8) since the random variables $\{f(X_n)\}$ are zero-mean, bounded, i.i.d. random variables, so the ordinary CLT implies that

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

where $\sigma^2 = \mathrm{Var}\,(f(X_1))$. This proves part i).

Next, dividing both sides of (21) by $\sqrt{2n \log_e \log_e n}$, letting $n \to \infty$, and invoking the classical LIL (see, e.g., [6, Theorem 13.25] or [12, p. 437]), immediately gives the two statements in part ii). □

Proof of Theorem 4: Let $\{C_n = (B_n, \phi_n, \psi_n)\}$ be an arbitrary sequence of block codes operating at distortion level D, and let $\{\ell_n\}$ be the sequence of corresponding length functions. By Proposition 2 part ii) we can choose a p.m.f. Q^* on $\hat{A}$ so that $R(D) = R(D, Q^*, D)$. Since we assume $D \in (0, D_{\max})$,

Proposition 2 part v) implies that $D \in (D_{\min}^{P,Q^*}, D_{\max}^{P,Q^*})$, so by Lemma 1 we can pick a $\lambda^* < 0$ with

$$\begin{aligned}\lambda^* D - \Lambda_{P,Q^*}(\lambda^*) &= \Lambda^*_{P,Q^*}(D) \\ &= (\log_e 2) R(P, Q^*, D) \\ &= (\log_e 2) R(D) \end{aligned} \tag{22}$$

where the second equality comes from Proposition 1 part ii).

Since ψ_n is a prefix-free lossless code (for each n), it induces a length function L_n on B_n given by

$$L_n(y_1^n) = \text{length of } [\psi_n(y_1^n)], \qquad y_1^n \in B_n.$$

The functions L_n and ℓ_n are clearly related by $\ell_n(x_1^n) = L_n(\phi_n(x_1^n))$. The key idea of the proof is to consider the following subprobability measure on $\hat{A}^n$:

$$Q_{C_n}(F) \triangleq \sum_{y_1^n \in F \cap B_n} 2^{-L_n(y_1^n)}, \qquad \text{for all } F \subseteq \hat{A}^n.$$

Note that Q_{C_n} is supported entirely on B_n; the fact that it is a subprobability measure follows by the Kraft inequality. Our main use for Q_{C_n} will be to bound the description lengths $\ell_n(x_1^n)$ in terms of an expectation over Q_{C_n}. For any $x_1^n \in A^n$

$$\begin{aligned} 2^{-\ell_n(x_1^n)} &= 2^{-L_n(\phi_n(x_1^n))} \\ &\overset{\text{a)}}{\le} 2^{-L_n(\phi_n(x_1^n))} e^{n\lambda^*[\rho_n(x_1^n, \phi_n(x_1^n)) - D]} \\ &\le \sum_{y_1^n \in B_n} 2^{-L_n(y_1^n)} e^{n\lambda^*[\rho_n(x_1^n, y_1^n) - D]} \\ &= E_{Q_{C_n}}\left(e^{n\lambda^*[\rho_n(x_1^n, Y_1^n) - D]}\right) \end{aligned}$$

where a) follows from the fact that C_n operates at distortion level D. This gives the lower bound

$$\ell_n(x_1^n) \ge -\log E_{Q_{C_n}}\left(e^{n\lambda^*[\rho_n(x_1^n, Y_1^n) - D]}\right). \tag{23}$$

Now consider the following family of functions on A^n:

$$\begin{aligned}\mathcal{F}_n \triangleq \Big\{ & g\colon g(x_1^n) = E_{Q_n}\left(e^{n\lambda^*[\rho_n(x_1^n, Y_1^n) - D]}\right) \\ & \text{for a subprobability measure } Q_n \text{ on } \hat{A}^n \Big\}\end{aligned}$$

and notice that $\mathcal{F}_n$ is a convex family. We are interested in the infimum

$$\begin{aligned}&\inf_{g \in \mathcal{F}_n} E_{P^n}\{-\log_e g(X_1^n)\} \\ &\quad = \inf_{Q_n} E_{P^n}\left\{-\log_e E_{Q_n}\left(e^{n\lambda^*[\rho_n(X_1^n, Y_1^n) - D]}\right)\right\} \end{aligned} \tag{24}$$

where P^n denotes the distribution X_1^n. According to Lemma 2 below, this infimum is achieved by the function $g^* \in \mathcal{F}_n$ defined in terms of the measure $Q_n^* = (Q^*)^n$

$$\begin{aligned} g^*(x_1^n) &= E_{Q_n^*}\left(e^{n\lambda^*[\rho_n(x_1^n, Y_1^n) - D]}\right) \\ &= \prod_{i=1}^n E_{Q^*}\left(e^{\lambda^*[\rho(x_i, Y) - D]}\right) \end{aligned} \tag{25}$$

i.e.,

$$E_{P^n}\left\{\log_e\left(\frac{g(X_1^n)}{g^*(X_1^n)}\right)\right\} \le 0, \qquad \text{for all } g \in \mathcal{F}_n.$$

But these are exactly the Kuhn–Tucker conditions for the optimality of g^* in (24); therefore, by [3, Theorem 2] we have that

$$E_{P^n}\left\{\frac{g(X_1^n)}{g^*(X_1^n)}\right\} \le 1, \qquad \text{for all } g \in \mathcal{F}_n. \tag{26}$$

The result of Theorem 4 can now be proved as follows. Define $g_n \in \mathcal{F}_n$ by

$$g_n(x_1^n) = E_{Q_{C_n}}\left(e^{n\lambda^*[\rho_n(x_1^n, Y_1^n) - D]}\right). \tag{27}$$

Recall the function f on A defined in (16), and observe that, using (22), it can be rewritten as

$$f(x) = -R(D) - \log E_{Q^*}\left(e^{\lambda^*[\rho(x, Y) - D]}\right). \tag{28}$$

Then, the probability that the assertion of the theorem fails can be bounded above as

$$\begin{aligned} &\Pr\left\{\ell_n(X_1^n) - nR(D) \le \sum_{i=1}^n f(X_i) - b_n\right\} \\ &\overset{\text{a)}}{\le} \Pr\left\{ -\log E_{Q_{C_n}}\left(e^{n\lambda^*[\rho_n(X_1^n, Y_1^n) - D]}\right) \right. \\ &\qquad \left. \le \sum_{i=1}^n [f(X_i) + R(D)] - b_n \right\} \\ &\overset{\text{b)}}{=} \Pr\{-\log g_n(X_1^n) \ge -\log g^*(X_1^n) b_n\} \\ &= \Pr\left\{\frac{g_n(X_1^n)}{g^*(X_1^n)} \ge 2^{b_n}\right\} \\ &\overset{\text{c)}}{\le} 2^{-b_n} E_{P^n}\left\{\frac{g_n(X_1^n)}{g^*(X_1^n)}\right\} \\ &\overset{\text{d)}}{\le} 2^{-b_n} \end{aligned} \tag{29}$$

where a) follows from the bound (23), b) follows from the definitions of f, g_n, and g^* in (28), (27), and (25), c) is simply Markov's inequality, and d) follows from the Kuhn–Tucker conditions (26) with $g = g_n$. Now since the sequence 2^{-b_n} is summable by assumption, an application of the Borel–Cantelli lemma to the bound in (29) completes the proof. □

Lemma 2: The infimum in (24) is achieved by $Q_n^* = (Q^*)^n$.

Proof of Lemma 2: Write $\mathcal{P}_n$ (or $\mathcal{SP}_n$) for the collection of all probability (respectively, subprobability) measures on $\hat{A}^n$. For $\lambda \leq 0$ and Q_n a p.m.f. on $\hat{A}^n$, let

$$h(\lambda, Q_n) = n\lambda D - \Lambda_{P^n, Q_n}(n\lambda).$$

Taking $Q_n = Q_n^* = (Q^*)^n$ in the infimum in (24), the result of the lemma follows from the following series of relations:

$$\begin{aligned}
&E_{P^n}\left\{-\log_e E_{Q_n^*}\left(e^{n\lambda^*[\rho_n(X_1^n, Y_1^n)-D]}\right)\right\} \\
&\quad \stackrel{\text{a)}}{=} n\lambda^* D - \Lambda_{P^n, Q_n^*}(n\lambda^*) \\
&\quad \stackrel{\text{b)}}{=} n\left[\lambda^* D - \Lambda_{P, Q^*}(\lambda^*)\right] \\
&\quad \stackrel{\text{c)}}{=} (\log_e 2) nR(P, Q^*, D) \\
&\quad \stackrel{\text{d)}}{=} (\log_e 2) nR(D) \\
&\quad \stackrel{\text{e)}}{=} (\log_e 2) R_n(D) \\
&\quad \stackrel{\text{f)}}{=} (\log_e 2) \inf_{Q_n \in \mathcal{P}_n} R(P^n, Q_n, D) \\
&\quad \stackrel{\text{g)}}{=} \inf_{Q_n \in \mathcal{P}_n} \sup_{\lambda \leq 0} h(\lambda/n, Q_n) \\
&\quad \stackrel{\text{h)}}{=} \inf_{Q_n \in \mathcal{P}_n} \sup_{\lambda \leq 0} h(\lambda, Q_n) \\
&\quad \stackrel{\text{i)}}{=} \sup_{\lambda \leq 0} \inf_{Q_n \in \mathcal{P}_n} h(\lambda, Q_n) \\
&\quad \stackrel{\text{j)}}{=} \inf_{Q_n \in \mathcal{P}_n} h(\lambda^*, Q_n) \\
&\quad \stackrel{\text{k)}}{=} \inf_{Q_n \in \mathcal{P}_n} E_{P^n}\left\{-\log_e E_{Q_n}\left(e^{n\lambda^*[\rho_n(X_1^n, Y_1^n)-D]}\right)\right\} \\
&\quad \stackrel{(\ell)}{=} \inf_{Q_n \in \mathcal{SP}_n} E_{P^n}\left\{-\log_e E_{Q_n}\left(e^{n\lambda^*[\rho_n(X_1^n, Y_1^n)-D]}\right)\right\}
\end{aligned}$$

where a) and b) follow from the definitions of $\Lambda_{P^n, (Q^*)^n}$ and Λ_{P, Q^*}, respectively; c) follows from the choice of λ^* and d) follows from the choice of Q^* in (14); e) follows from the well-known fact that the rate-distortion function of a vector of n i.i.d. random variables equals n times the rate-distortion function of one of them; f) follows in the same way as we noted in the Introduction that (6) is the same as (13); g) follows from Proposition 1 part ii); h) follows simply by replacing λ by $n\lambda$; i) follows by an application of the minimax theorem (see below for an explanation); j) follows from a simple continuity argument given below; k) follows from the definitions of the functions h and Λ_{P^n, Q_n}; and (ℓ) follows from noting that if Q_n is strictly a subprobability measure with $Q_n(\hat{A}^n) = Z < 1$, then using the probability measure $Q_n'(\cdot) = Z^{-1} Q_n(\cdot)$ we can make the expression that is being minimized on line (ℓ) smaller by $\log Z < 0$.

To justify the application of the minimax theorem in step i), first we note that, by Lemma 1 part iv), h is lower semicontinuous as a function of Q_n, and by Lemma 1 part i) it is a continuous function of λ. Also, since by Lemma 1 part ii) $\Lambda''_{P^n, Q_n}(\lambda) \geq 0$, h is concave in λ, and by Jensen's inequality (and the concavity of the logarithm), h is convex in Q_n. And since the space of all p.m.f.'s Q_n on $\hat{A}^n$ is compact, we can invoke Sion's minimax theorem [31, Corollary 3.3] to justify the exchange of the infimum and the supremum in step i).

Finally, we need to justify step j). For $\lambda \leq 0$ define the functions $M(\lambda) = h(\lambda, Q_n^*)$, and

$$m(\lambda) = \inf_{Q_n \in \mathcal{P}_n} h(\lambda, Q_n).$$

Our goal is to show

$$\sup_{\lambda \leq 0} m(\lambda) = m(\lambda^*). \tag{30}$$

From the above argument a)–i), we have that

$$M(\lambda^*) = \sup_{\lambda \leq 0} M(\lambda) = \sup_{\lambda \leq 0} m(\lambda).$$

As noted in the beginning of the proof of Theorem 4, $M(\lambda)$ is strictly concave and its supremum is achieved uniquely by $\lambda^* < 0$, so we may restrict our attention to an interval $I = [\lambda^* - \delta, \lambda^* + \delta] \subseteq (-\infty, 0)$, and since $m(\lambda) \leq M(\lambda)$ for all λ

$$M(\lambda^*) = \sup_{\lambda \in I} M(\lambda) = \sup_{\lambda \in I} m(\lambda).$$

By the definition of m as the infimum of continuous functions it follows that it is upper semicontinuous (see, e.g., [27, p. 38]), so it achieves its supremum on the compact set I. Since $m(\lambda) \leq M(\lambda)$ for all λ, and $M(\lambda) = M(\lambda^*)$ only at λ^*, this implies that the supremum of m over I must also be achieved at λ^*, giving (30). □

IV. Shannon's Random Codes and D-Ball Probabilities

In this section we prove Theorem 5, and we deduce Theorems 2 and 3 from it.

We continue in the notation of the previous section, and recall the definition of a distortion ball in (17). Let $\boldsymbol{X}$ be a memoryless source with distribution P, fix $D \in (0, D_{\max})$, let Q^* be the optimal reproduction distribution for P at distortion level D, and write $Q_n^* = (Q^*)^n$. The following proposition will be used repeatedly in the proofs. If follows easily from some recent results in [10] and [35]; see Appendix VI.

Proposition 3: If $D \in (0, D_{\max})$, we have

$$\begin{aligned}
&-\log Q_n^*(B(X_1^n, D)) \\
&\qquad = nR(D) + \sum_{i=1}^n f(X_i) + \frac{1}{2}\log n + O(\log\log n) \quad \text{a.s.}
\end{aligned}$$

Proofs of Theorems 2 and 3 from Theorem 5: Combining Theorem 5 part i) with Proposition 3 gives

$$\ell_n(X_1^n) - nR(D) \leq \sum_{i=1}^n f(X_i) + O(\log n) \quad \text{a.s.}$$

In view of the corresponding lower bound in Theorem 4 (with $b_n = 2\log n$), this, together with the classical CLT and the LIL

applied to the sum of the bounded, zero-mean, i.i.d. random variables $\{f(X_i)\}$, yield the three statements of Theorem 2. Theorem 3 follows from Theorem 5 part ii) in exactly the same way. □

Proof of Theorem 5 Part i): For each $n \geq 1$ we generate a random codebook according to $Q_n^* = (Q^*)^n$. Let $Y(i) = (Y_{i,1}, Y_{i,2}, \cdots, Y_{i,n})$, $i = 1, 2, \cdots$, be i.i.d. random vectors in $\hat{A}^n$, each drawn according to Q_n^*. Given a source string X_1^n to be encoded, let $W_n = W_n(X_1^n)$ denote the first index i such that X_1^n matches the ith codeword with distortion D or less

$$W_n = \inf\{i \geq 1 \colon \rho_n(X_1^n, Y(i)) \leq D\}.$$

If such a match exists, we describe X_1^n (with distortion no more than D) by describing the integer W_n to the decoder; this can be done using (cf. [13] and [33])

$$\lceil \log(W_n + 1) \rceil + 2\lceil \log(\lceil \log(W_n + 1) \rceil + 1) \rceil \leq \log W_n + 2\log\log(2W_n) + 10 \quad \text{bits.}$$

Otherwise, we describe X_1^n exactly, using $\lceil n \log |\hat{A}| \rceil$ bits (recall (5)). Our code C_n consists of combining these two descriptions, together with a one-bit flag to specify which one was chosen.

Next we show that, for large n, the "waiting time" W_n until a D-match is found can be approximated by the reciprocal of the probability of finding such a match, $Q_n^*(B(X_1^n, D))$. Specifically, we claim that the difference

$$\epsilon_n \stackrel{\Delta}{=} \log W_n - \log[1/Q_n^*(B(X_1^n, D))]$$

satisfies

$$\epsilon_n \leq 2\log n \quad \text{eventually, a.s.} \tag{31}$$

where the almost-sure statement above (and also in all subsequent statements) is with respect to P-almost any source realization, and almost any sequence of codebooks generated according to the above procedure. We prove (31) by an argument along the lines of the "strong approximation" results in [10], [16], and [17]. Let

$$G_m = \{x_1^\infty \colon 0 < Q_n^*(B(x_1^n, D)) < 1/2 \text{ for all } n \geq m\}.$$

Proposition 2 implies that, eventually, almost every string x_1^∞ generated by $\boldsymbol{X}$ will belong to a G_m, i.e.,

$$P\left(\bigcup_{m \geq 1} G_m\right) = 1 \tag{32}$$

where, with a slight abuse of notation, we write P for the one-dimensional marginal of the distribution of $\boldsymbol{X}$ as well as the infinite-dimensional product distribution it induces. Now let $n \geq m \geq 1$. Conditional on $X_1^\infty = x_1^\infty \in G_m$, the waiting time W_n has a geometric distribution with parameter $p_n \stackrel{\Delta}{=} Q_n^*(B(X_1^n, D))$, so that

$$\begin{aligned}\Pr\{\epsilon_n > 2\log n | X_1^n = x_1^n\} &= \Pr\{W_n > n^2/p_n | X_1^n = x_1^n\} \\ &\leq (1-p_n)^{(n^2/p_n)-1} \\ &\leq 2(1-p_n)^{n^2/p_n} \\ &\leq 2/n^2\end{aligned}$$

where the last step follows from the inequality $(1-p)^M \leq 1/(Mp)$, for $p \in (0, 1)$ and $M > 0$. Since this bound is uniform over $x_1^\infty \in G_m$, and $(2/n^2)$ is summable, the Borel–Cantelli lemma implies that $\epsilon_n \leq 2\log n$, eventually, for almost every codebook sequence, and P-almost all $x_1^\infty \in G_m$. This together with (32) establishes (31).

In particular, (31) implies that with probability one, $W_n < \infty$ eventually. Therefore, the description length of our code

$$\ell_n(X_1^n) \leq \log W_n + 2\log\log(2W_n) + 11 \quad \text{bits, eventually, a.s.}$$

and this can be bounded above as

$$\begin{aligned}\ell_n(X_1^n) &\leq \log[1/Q_n^*(B(X_1^n, D))] + \epsilon_n \\ &\quad + 2\log[1 + \epsilon_n - \log Q_n^*(B(X_1^n, D))] + 11 \\ &\stackrel{\text{a)}}{\leq} \log[1/Q_n^*(B(X_1^n, D))] + 2\log n \\ &\quad + 2\log[1 + 2\log n + 2nR(D)] + 11 \\ &\leq \log[1/Q_n^*(B(X_1^n, D))] + 4\log n + \text{Const.} \\ &\qquad \text{eventually, a.s.}\end{aligned}$$

where a) follows from (31) and Proposition 3. This proves part i). □

Note that the above proof not only demonstrates the existence of a good sequence of codes $\{C_n, \ell_n\}$, but it also shows that almost every sequence of random codes generated as above will satisfy the statement of the theorem.

Proof of Theorem 5 Part ii): Here we assume that the source $\boldsymbol{X} = \{X_n; n \geq 1\}$ has a distribution P, where P is unknown to the encoder and decoder, but such that $D \in (0, D_{\max}(P))$. For each $n \geq 1$ we generate a family of codebooks, one for each n-type on $\hat{A}$. Recall [9] that a p.m.f. Q on $\hat{A}$ is called an n-type if, for each $y \in \hat{A}$, $Q(y) = m/n$ for an integer m. The number $T(n)$ of n-types grows polynomially in n, and it is bounded above as

$$T(n) \leq (n+1)^k \tag{33}$$

where $k = |\hat{A}|$ denotes the cardinality of $\hat{A}$; see [9, Ch. 13].

For $1 \leq j \leq T(n)$, let $Q^{(j)}$ denote the jth n-type. The jth codebook consists of i.i.d. random vectors

$$Y^{(j)}(i) = (Y_{i,1}^{(j)}, Y_{i,2}^{(j)}, \cdots, Y_{i,n}^{(j)}), \qquad i = 1, 2, \cdots$$

where each $Y^{(j)}(i)$ is drawn according to $(Q^{(j)})^n$. Given a source string X_1^n, we let $W_n^{(j)} = W_n^{(j)}(X_1^n)$ be the waiting time until a D-close match for X_1^n is found in the jth codebook

$$W_n^{(j)} = \inf\left\{i \geq 1 \colon \rho_n(X_1^n, Y^{(j)}(i)) \leq D\right\}$$

and we define

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

It is not hard to see that, for large enough n, there will (almost) always be a D-match for X_1^n in one of the codebooks, so that

$$W_n^* < \infty \quad \text{eventually, a.s.}$$

where the almost-sure statement here (and also in all subsequent statements) is with respect to P-almost any source realization, and almost any sequence of codebooks generated as above. (This is so because, for $n \ge k$, at least one of the n-types has positive probability on all elements of $\hat{A}$, so with probability one every possible $\hat{A}^n$-string will appear infinitely often. Assumption (5) then guarantees the existence of a D-match.) Therefore, we can describe X_1^n (with distortion no more than D) to the decoder by specifying the waiting time W_n^*, and the codebook in which W_n^* is achieved. As in part i), and using the bound in (33), this can be done using

$$\begin{aligned}\ell_n^*(X_1^n) &= \lceil \log(W_n^*+1)\rceil + 2\lceil \log(\lceil \log(W_n^*+1)\rceil + 1)\rceil \\ &\quad + \lceil k\log(n+1)\rceil \\ &\le \log W_n^* + 2\log\log(2W_n^*) + k\log n + (11+k) \\ &\quad \text{bits, eventually, a.s.}\end{aligned}$$

Now, following [17], we pick a sequence of n-types that are close to Q^*. We let q_n be an n-type such that $q_n(y) > 0$ and $|q_n(y) - Q^*(y)| \le k/n$, for all $y \in \hat{A}$. This can be done for all $n \ge N$, for some fixed integer N (see [17] for the details). Let V_n denote the waiting time associated with the codebook corresponding to q_n, and write $Q_n = (q_n)^n$. The same argument as the one used to prove (31) in part i) can be used here to show that

$$\begin{aligned}\epsilon_n' &= \triangleq \log V_n - \log[1/Q_n(B(X_1^n, D))] \\ &\le 2\log n \quad \text{eventually, a.s.}\end{aligned} \tag{34}$$

Using the obvious fact that W_n^* is never greater than V_n, we can bound $\ell_n^*(X_1^n)$ above by

$$\begin{aligned}\ell_n^*(X_1^n) &\le \log V_n + 2\log\log(2V_n) + k\log n + (11+k) \\ &\le \log(1/p_n) + \delta_n + \epsilon_n' \\ &\quad + 2\log[1+\delta_n+\epsilon_n' + \log(1/p_n)] \\ &\quad + k\log n + (11+k)\end{aligned} \tag{35}$$

eventually, almost surely, where $p_n = Q_n^*(B(X_1^n, D))$ as before, and

$$\delta_n = \delta_n(X_1^n) = \log\left[\frac{Q_n^*(B(X_1^n, D))}{Q_n(B(X_1^n, D))}\right].$$

Next we claim that there exist absolute constants C and N_0 such that

$$\delta_n(x_1^n) \le C \quad \text{for all } n \ge N_0, \text{ and all } x_1^n \in A^n. \tag{36}$$

Before proving this, let us see how it allows us to complete the proof. Recalling Proposition 3 and substituting the bounds (34) and (36) into (35) gives

$$\begin{aligned}\ell_n^*(X_1^n) &\le \log(1/p_n) + 2\log[1 + C + 2\log n + 3nR(D)] \\ &\quad + (2+k)\log n + (C+11+k) \\ &\le \log(1/p_n) + 2\log[4nR(D)] + (2+k)\log n \\ &\quad + (C+11+k) \\ &\le \log[1/Q_n^*(B(X_1^n, D))] + (4+k)\log n + \text{Const.} \\ &\quad \text{eventually, } P-\text{a.s.}\end{aligned}$$

Finally, we need to prove (36). Pick $N_0 \ge N$ large enough, so that for all $n \ge N_0$ and all $y \in \hat{A}$, $Q^*(y)$ is either equal to zero or $Q^*(y) > k/n$. Let $n \ge N_0$ and $x_1^n \in A^n$ be arbitrary. Then

$$\begin{aligned}\frac{Q_n^*(B(x_1^n, D))}{Q_n(B(x_1^n, D))} &= \frac{\sum_{y_1^n \in B(x_1^n, D)} \left[Q_n(y_1^n)\frac{Q_n^*(y_1^n)}{Q_n(y_1^n)}\right]}{\sum_{y_1^n \in B(x_1^n, D)} Q_n(y_1^n)} \\ &\le \max_{y_1^n \in B(x_1^n, D)} \frac{Q_n^*(y_1^n)}{Q_n(y_1^n)} \\ &= \max_{y_1^n \in B(x_1^n, D)} \prod_{i=1}^n \frac{Q^*(y_i)}{q_n(y_i)} \\ &\le \left(\max_{y\in\hat{A}:\, Q^*(y)>0} \frac{Q^*(y)}{q_n(y)}\right)^n \\ &\overset{\text{a)}}{\le} \left(\max_{y\in\hat{A}:\, Q^*(y)>0} \frac{Q^*(y)}{Q^*(y) - k/n}\right)^n \\ &\le \left(1 - \frac{k}{nQ^*(y^*)}\right)^{-n}\end{aligned}$$

where a) is by the choice of q_n, and y^* is the $y \in \hat{A}$ with the smallest nonzero Q^* probability. So

$$\delta_n(x_1^n) \le -n\log(1 - C'/n)$$

with $C' = k/Q^*(y^*)$, and this is a convergent sequence so it must be bounded. □

As in part i), this proof actually shows that almost every sequence of random codes generated as above will satisfy the statement of the theorem.

V. Arbitrary Sources

Let $\boldsymbol{X}$ be an A-valued source, and write P_n for the distribution of X_1^n. In this section, we prove Theorems 6–8. We begin with two useful lemmas; they are proved in Appendices IV and V, respectively.

Lemma 3: The infimum

$$\inf_{Q_n} E_{P_n}\{-\log Q_n(B(X_1^n, D))\} \tag{37}$$

over all subprobability measures Q_n on $\hat{A}^n$ is the same as the infimum over all probability measures, and it is achieved by some probability measure $\tilde{Q}_n$.

Lemma 4:

$$K_n(D) = E_{P_n}\left\{-\log \tilde{Q}_n(B(X_1^n, D))\right\} \geq R_n(D).$$

Proof of Theorem 6: Let $\{C_n, \ell_n\}$ be an arbitrary sequence of codes operating at distortion level D, where each C_n consists of a triple (B_n, ϕ_n, ψ_n). Let L_n be the length function induced by ψ_n on B_n. As in the proof of Theorem 4, the key idea is to consider the subprobability measure Q_{C_n} on $\hat{A}^n$ defined by

$$Q_{C_n}(F) \stackrel{\Delta}{=} \sum_{y_1^n \in F \cap B_n} 2^{-L_n(y_1^n)}, \qquad \text{for all } F \subseteq \hat{A}^n.$$

Since C_n operates at distortion level D, for any $x_1^n \in A^n$ we have

$$\begin{aligned}\ell_n(x_1^n) &= L_n(\phi_n(x_1^n)) \\ &= -\log Q_{C_n}(\phi_n(x_1^n)) \\ &\geq -\log Q_{C_n}(B(x_1^n, D)).\end{aligned} \tag{38}$$

From (38) and the definition of $K_n(D)$ we immediately get that

$$E_{P_n}[\ell_n(X_1^n)] \geq K_n(D)$$

and, in view of Lemma 4, this proves part i).

For part ii), we define a family of functions on A^n

$$\mathcal{G}_n \stackrel{\Delta}{=} \left\{g\colon g(x_1^n) = Q_n(B(x_1^n, D)) \text{ for a subprobability measure } Q_n \text{ on } \hat{A}^n\right\}$$

and note that $\mathcal{G}_n$ is a convex family. By Lemma 3 we know that

$$\inf_{g \in \mathcal{G}_n} E_{P_n}\{-\log g(X_1^n)\} = E_{P_n}\{-\log \tilde{g}(X_1^n)\} \tag{39}$$

where $\tilde{g}$ is the function $\tilde{g}(x_1^n) = \tilde{Q}_n(B(x_1^n, D))$. But for each $n \geq 1$, (39) are exactly the Kuhn–Tucker conditions for the optimality of $\tilde{g}$ in $\mathcal{G}_n$, so [3, Theorem 2] implies that

$$E_{P_n}\left\{\frac{g(X_1^n)}{\tilde{g}(X_1^n)}\right\} \leq 1, \qquad \text{for all } g \in \mathcal{G}_n. \tag{40}$$

Therefore, letting

$$g_n(x_1^n) = Q_{C_n}(B(x_1^n, D))$$

the probability that the assertion of part ii) fails can be bounded above as

$$\begin{aligned}&\Pr\left\{\ell_n(X_1^n) \leq \log\left[1/\tilde{Q}_n(B(X_1^n, D))\right] - b_n\right\} \\ &\stackrel{\text{a)}}{\leq} \Pr\left\{\log\left[1/Q_{C_n}(B(X_1^n, D))\right] \right. \\ &\qquad \left. \leq \log\left[1/\tilde{Q}_n(B(X_1^n, D))\right] - b_n\right\} \\ &= \Pr\left\{\frac{Q_{C_n}(B(X_1^n, D))}{\tilde{Q}_n(B(X_1^n, D))} \geq b_n\right\} \\ &\stackrel{\text{b)}}{\leq} 2^{-b_n} E_{P_n}\left\{\frac{g_n(X_1^n)}{\tilde{g}(X_1^n)}\right\} \\ &\stackrel{\text{c)}}{\leq} 2^{-b_n}\end{aligned}$$

where a) follows from the bound (38), b) is simply Markov's inequality, and c) follows from the Kuhn–Tucker conditions (40) with $g = g_n$. Since the sequence 2^{-b_n} is summable by assumption, the Borel–Cantelli lemma completes the proof. □

Proof of Theorem 7: Suppose $\boldsymbol{X}$ is a memoryless source with distribution P, let $P_n = P^n$ denote the distribution of X_1^n, and write $Q_n^* = (Q^*)^n$, where Q^* is the optimal reproduction distribution at distortion level D.

Replacing g_n by $g_n'(x_1^n) = Q_n^*(B(x_1^n, D))$ in the last part of the argument of the proof of Theorem 6, and taking $b_n \stackrel{\Delta}{=} 2\log n$, we get

$$\begin{aligned}&\log[1/Q_n^*(B(X_1^n, D))] \\ &\geq \log\left[1/\tilde{Q}_n(B(X_1^n, D))\right] - 2\log n \quad \text{eventually, a.s.}\end{aligned} \tag{41}$$

Similarly, taking

$$g_n''(x_1^n) = E_{\tilde{Q}_n}\left(e^{n\lambda^*[\rho_n(x_1^n, Y_1^n) - D]}\right)$$

in place of g_n in the proof of Theorem 4, and choosing $b_n \stackrel{\Delta}{=} 2\log n$, we get

$$\begin{aligned}&-\log E_{\tilde{Q}_n}\left(e^{n\lambda^*[\rho_n(X_1^n, Y_1^n) - D]}\right) \\ &\qquad \geq -\log E_{Q_n^*}\left(e^{n\lambda^*[\rho_n(X_1^n, Y_1^n) - D]}\right) \\ &\qquad\quad - 2\log n \quad \text{eventually, a.s.}\end{aligned} \tag{42}$$

But by Proposition 3 we know that (in the notation of the proof of Theorem 4)

$$\begin{aligned}&\log[1/Q_n^*(B(X_1^n, D))] \\ &= nR(D) + \sum_{i=1}^n f(X_i) + \tfrac{1}{2}\log n + O(\log\log n) \quad \text{a.s.} \\ &= \log E_{Q_n^*}\left(e^{n\lambda^*[\rho_n(X_1^n, Y_1^n) - D]}\right) + \frac{1}{2}\log n \\ &\quad + O(\log\log n) \quad \text{a.s.}\end{aligned} \tag{43}$$

and also, by a simple Chernoff-type bound

$$\begin{aligned}\tilde{Q}_n(B(X_1^n, D)) &= E_{\tilde{Q}_n}\left\{\mathbb{I}_{\{\rho_n(X_1^n, Y_1^n) \leq D\}}\right\} \\ &\leq E_{\tilde{Q}_n}\left\{e^{n\lambda^*[\rho_n(X_1^n, Y_1^n) - D]}\right\}.\end{aligned} \tag{44}$$

From (42)–(44) we have

$$\log[1/\tilde{Q}_n(B(X_1^n, D))] \geq \log[1/Q_n^*(B(X_1^n, D))] - \frac{5}{2}\log n + O(\log\log n) \quad \text{a.s.}$$

Combining this with the corresponding lower bound in (41) completes the proof. □

Proof of Theorem 8: We use a random coding argument, very similar to the ones used in the proofs of Theorem 5 parts i) and ii). For each $n \geq 1$ we generate a random codebook according to $\tilde{Q}_n$: Let

$$Y(i) = (Y_{i,1}, Y_{i,2}, \cdots, Y_{i,n}), \qquad i = 1, 2, \cdots$$

be i.i.d. random vectors in $\hat{A}^n$, each drawn according to $\tilde{Q}_n$. Given a source string X_1^n, let $W_n = W_n(X_1^n)$ denote the first index i such that X_1^n matches the ith codeword with distortion D or less

$$W_n = \inf\{i \geq 1 \colon \rho_n(X_1^n, Y(i)) \leq D\}.$$

If such a match exists, we describe X_1^n to the decoder (with distortion no more than D) by describing W_n, using, as before, no more than

$$\log W_n + 2\log\log(2W_n) + 10 \quad \text{bits.}$$

Otherwise, we describe X_1^n exactly, using $\lceil n \log|\hat{A}| \rceil$ bits; this is possible because of our initial assumption (5). Our code C_n consists of combining these two descriptions, together with a one-bit flag to specify which one was chosen.

Next we claim that the waiting times W_n can be approximated by the quantities $1/\tilde{Q}_n(B(X_1^n, D))$, in that their difference satisfies

$$\epsilon_n \stackrel{\Delta}{=} \log W_n - \log[1/\tilde{Q}_n(B(X_1^n, D))] \leq 2\log n \quad \text{eventually, a.s.} \qquad (45)$$

The assumption that $K_n(D) < \infty$ implies that

$$\tilde{Q}_n(B(x_1^n, D)) > 0$$

for P_n-almost all x_1^n, so the strong approximation argument from the proof of Theorem 5 goes through essentially *verbatim* to prove (45). In particular, (45) implies that $W_n < \infty$ eventually, almost surely, so the description length of our code can be bounded above as

$$\begin{aligned}\ell_n(X_1^n) &\leq \log W_n + 2\log\log(2W_n) + 11 \\ &\leq \log\left[1/\tilde{Q}_n(B(X_1^n, D))\right] + \epsilon_n \\ &\quad + 2\log\log\left[\frac{2^{\epsilon_n+1}}{\tilde{Q}_n(B(X_1^n, D))}\right] + 11 \\ &\leq \log\left[1/\tilde{Q}_n(B(X_1^n, D))\right] + 2\log n \\ &\quad + 2\log\log\left[\frac{2n^2}{\tilde{Q}_n(B(X_1^n, D))}\right] + \text{Const.} \\ &\qquad \text{eventually, a.s.}\end{aligned}$$

and we are done. □

APPENDIX I

Proof of Lemma 1 Part iv): Fix a $\lambda \leq 0$ and a probability measure μ on A^n. Let $\{\nu^{(k)}\}$ be a sequence of p.m.f.'s on $\hat{A}^n$, such that the $\nu^{(k)}$ converge, as $k \to \infty$, to some p.m.f. ν on $\hat{A}^n$. Then

$$\begin{aligned}&\limsup_{k\to\infty} \Lambda_{\mu,\nu^{(k)}}(\lambda) \\ &\quad = -\liminf_{k\to\infty} E_\mu\left\{-\log_e E_{\nu^{(k)}}\left(e^{\lambda\rho_n(X_1^n, Y_1^n)}\right)\right\} \\ &\quad \stackrel{a)}{\leq} -E_\mu\left\{\liminf_{k\to\infty} -\log_e E_{\nu^{(k)}}\left(e^{\lambda\rho_n(X_1^n, Y_1^n)}\right)\right\} \\ &\quad \stackrel{b)}{=} -E_\mu\left\{-\log_e E_\nu\left(e^{\lambda\rho_n(X_1^n, Y_1^n)}\right)\right\} \\ &\quad = \Lambda_{\mu,\nu}(\lambda)\end{aligned}$$

where a) follows from Fatou's Lemma and b) follows from the assumption that $\nu^{(k)} \to \nu$. Therefore, $\Lambda_{\mu,\nu}(\lambda)$ is upper semicontinuous in ν. □

APPENDIX II

Proof of Proposition 1: The alternative representation of $R(\mu, \nu, D)$ in part i) can be obtained from its definition by a simple application of the chain rule for relative entropy (see, e.g., [25, eq. (3.11.5)] or [11, Theorem D.13]).

For part ii) we will use the representation in part i). Fix $D \geq 0$ arbitrary, and recall (see [11, Lemma 6.2.13]) that for any bounded measurable function ϕ: $\hat{A}^n \to \mathbb{R}$, any $x_1^n \in A^n$, and any candidate measure Θ on $A^n \times \hat{A}^n$ with A^n-marginal equal μ and $E_\Theta[\rho_n(X_1^n, Y_1^n)] \leq D$, we have

$$(\log_e 2)H(\Theta(\cdot|x_1^n)||\nu(\cdot)) \geq \int \phi(y_1^n)\, d\Theta(y_1^n|x_1^n) - \log_e E_\nu\left(e^{\phi(Y_1^n)}\right).$$

Choosing $\phi(y_1^n) = \lambda\rho_n(x_1^n, y_1^n)$ and taking expectations of both sides with respect to μ, yields that

$$(\log_e 2)E_\mu\left[H(\Theta(\cdot|X_1^n)||\nu(\cdot))\right] \geq \lambda D - \Lambda_{\mu,\nu}(\lambda).$$

Taking the infimum over all candidate measures Θ and the supremum over all $\lambda \leq 0$, implies (from part i))

$$R(\mu, \nu, D) \geq (\log e)\Lambda^*_{\mu,\nu}(D). \qquad (46)$$

To prove the reverse inequality we consider four cases. (Note that we only need to consider cases when $\Lambda^*_{\mu,\nu}(D) < \infty$.)

Case I: $D^{\mu,\nu}_{\min} > 0$ and $D \in (0, D^{\mu,\nu}_{\min})$. By Lemma 1, for all $\lambda \leq 0$

$$\frac{d}{d\lambda}\left[\lambda D - \Lambda_{\mu,\nu}(\lambda)\right] = D - \Lambda'_{\mu,\nu}(\lambda) \leq D - D^{\mu,\nu}_{\min} < 0.$$

Therefore,

$$\Lambda^*_{\mu,\nu}(D) \geq \limsup_{\lambda\to-\infty}\left[\lambda D - \Lambda_{\mu,\nu}(\lambda)\right] = \infty$$

so in this case

$$R(\mu, \nu, D) = (\log e)\Lambda^*_{\mu,\nu}(D) = \infty. \qquad (47)$$

Case II: $D \geq D_{\max}^{\mu,\nu}$. Here, by Lemma 1, for all $\lambda \leq 0$

$$\frac{d}{d\lambda}\left[\lambda D - \Lambda_{\mu,\nu}(\lambda)\right] = D - \Lambda'_{\mu,\nu}(\lambda) \geq D_{\max}^{\mu,\nu} - D_{\max}^{\mu,\nu} = 0$$

so $\Lambda^*_{\mu,\nu}(D)$ is achieved at $\lambda = 0$, giving

$$\Lambda^*_{\mu,\nu}(D) = \Lambda_{\mu,\nu}(0) = 0.$$

On the other hand, taking $\Theta = \mu \times \nu$, noting that

$$E_\Theta[\rho_n(X_1^n, Y_1^n)] = D_{\max}^{\mu,\nu}$$

and recalling that relative entropy is nonnegative, implies that $R(\mu, \nu, D) = 0$. Hence, here

$$R(\mu, \nu, D) = (\log e)\Lambda^*_{\mu,\nu}(D) = 0. \tag{48}$$

Case III: $D_{\min}^{\mu,\nu} < D_{\max}^{\mu,\nu}$ and $D \in (D_{\min}^{\mu,\nu}, D_{\max}^{\mu,\nu})$. By Lemma 1, there is a unique $\lambda^* < 0$ such that $\Lambda'_{\mu,\nu}(\lambda^*) = D$ and

$$\Lambda^*_{\mu,\nu}(D) = \lambda^* D - \Lambda_{\mu,\nu}(\lambda^*) > 0 - \Lambda_{\mu,\nu}(0) = 0. \tag{49}$$

Let

$$\frac{d\Theta}{d\mu \times d\nu}(x_1^n, y_1^n) = \frac{e^{\lambda^* \rho_n(x_1^n, y_1^n)}}{E_\nu\left(e^{\lambda^* \rho_n(x_1^n, Y_1^n)}\right)}$$

and observe that $E_\Theta[\rho_n(X_1^n, Y_1^n)] = \Lambda'_{\mu,\nu}(\lambda^*) = D$. Then

$$\begin{aligned} R(\mu, \nu, D) &\leq E_\mu\left[H(\Theta(\cdot|X_1^n)\|\nu(\cdot))\right] \\ &= \int \log\left[\frac{d\Theta}{d\mu \times d\nu}\right] d\mu \times d\nu \\ &= (\log e)\left[\lambda^* D - \Lambda_{\mu,\nu}(\lambda^*)\right] \\ &= (\log e)\Lambda^*_{\mu,\nu}(D). \end{aligned}$$

This, together with (46) and (49) imply that here

$$0 < R(\mu, \nu, D) = (\log e)\Lambda^*_{\mu,\nu}(D) < \infty. \tag{50}$$

Case IV: $D_{\min}^{\mu,\nu} < D_{\max}^{\mu,\nu}$ and $D = D_{\min}^{\mu,\nu}$. Write $S_n \subseteq \hat{A}^n$ for the support of ν, and for $x_1^n \in A^n$ let

$$\rho_n(x_1^n) = \min_{y_1^n \in S_n} \rho_n(x_1^n, y_1^n)$$

so that $D_{\min}^{\mu,\nu} = E_\mu[\rho_n(X_1^n)]$ and

$$\begin{aligned} &\lambda D_{\min}^{\mu,\nu} - \Lambda_{\mu,\nu}(\lambda) \\ &\qquad = E_\mu\left[-\log_e E_\nu\left(e^{\lambda[\rho_n(X_1^n, Y_1^n) - \rho_n(X_1^n)]}\right)\right]. \end{aligned}$$

Also, by Lemma 1

$$\frac{d}{d\lambda}\left[\lambda D_{\min}^{\mu,\nu} - \Lambda_{\mu,\nu}(\lambda)\right] = D_{\min}^{\mu,\nu} - \Lambda'_{\mu,\nu}(\lambda) < 0$$

so $\Lambda^*_{\mu,\nu}(D_{\min}^{\mu,\nu})$ is the increasing limit of $[\lambda D_{\min}^{\mu,\nu} - \Lambda_{\mu,\nu}(\lambda)]$ as $\lambda \to -\infty$. Therefore, letting $Z(x_1^n)$ denote the event

$$\{y_1^n : \rho_n(x_1^n, y_1^n) = \rho_n(x_1^n)\} \subseteq \hat{A}^n$$

and $\overline{Z}(x_1^n)$ denote its complement

$$\begin{aligned} &\Lambda^*_{\mu,\nu}(D_{\min}^{\mu,\nu}) \\ &\quad = \lim_{\lambda \to -\infty} E_\mu\left[-\log_e E_\nu\left(e^{\lambda[\rho_n(X_1^n, Y_1^n) - \rho_n(X_1^n)]}\right)\right] \\ &\quad = \lim_{\lambda \to -\infty} E_\mu\left[-\log_e \left\{\nu(Z(X_1^n)) \right.\right. \\ &\qquad \left.\left. + E_\nu\left(e^{\lambda[\rho_n(X_1^n, Y_1^n) - \rho_n(X_1^n)]}\mathbb{I}_{\overline{Z}(X_1^n)}\right)\right\}\right] \\ &\quad = E_\mu\left[-\log_e \nu(Z(X_1^n))\right] \end{aligned}$$

where the last equality follows from the monotone convergence theorem. Since we are only interested in the case $\Lambda^*_{\mu,\nu}(D_{\min}^{\mu,\nu}) < \infty$, the above calculation implies that we may assume, without loss of generality, that $\nu(Z(x_1^n)) > 0$ for μ-almost all $x_1^n \in A^n$. We can then define a measure Θ by

$$\frac{d\Theta}{d\mu \times d\nu}(x_1^n, y_1^n) = \frac{1}{\nu(Z(x_1^n))}\mathbb{I}_{\{y_1^n \in Z(x_1^n)\}}$$

which has $E_\Theta[\rho_n(X_1^n, Y_1^n)] = D_{\min}^{\mu,\nu}$, and

$$\begin{aligned} R(\mu, \nu, D) &\leq E_\mu\left[H\left(\Theta(\cdot|X_1^n)\|\nu(\cdot)\right)\right] \\ &= \int \log\left[\frac{d\Theta}{d\mu \times d\nu}\right] d\mu \times d\nu \\ &= (\log e)\int -\log_e \nu(Z(x_1^n))\, d\mu(x_1^n) \\ &= (\log e)\Lambda^*_{\mu,\nu}(D_{\min}^{\mu,\nu}). \end{aligned}$$

This together with (46) complete the proof. □

Appendix III

Proof of Proposition 2: By Lemma 1 part iv), $\Lambda_{\mu,\nu}(\lambda)$ is upper semicontinuous as a function of ν, so $[\lambda D - \Lambda_{\mu,\nu}(\lambda)]$ is lower semicontinuous. Therefore, by the representation of $R(\mu, \nu, D)$ in Proposition 1 part ii) and the fact that the supremum of lower semicontinuous functions is itself lower semicontinuous (see, e.g., [27, p. 38]) we get that $R(\mu, \nu, D)$ is lower semicontinuous as a function of ν, proving part i).

Part ii) follows immediately from part i): since $\hat{A}$ is finite, the set of all p.m.f.'s Q on $\hat{A}$ is compact, and therefore the lower semicontinuous function $R(P, Q, D)$ must achieve its infimum over that compact set (see, e.g., [26, p. 195]), proving the existence of the required Q^*.

For part iii): it is easy to check that the stated properties of $R(\mu, \nu, D)$ are actually proved in the course of proving Proposition 1 part ii); see (47), (48), and (50).

Part iv): First, if $D \in (0, D_{\max})$, then letting U denote the uniform distribution on $\hat{A}$ and recalling our basic assumption (5), we have

$$D_{\min}^{P,U} = E_\mu\left[\min_{y \in \hat{A}} \rho(X, y)\right] \leq \sup_{x \in A} \min_{y \in \hat{A}} \rho(x, y) = 0$$

i.e., $D_{\min}^{P,U} = 0$. Therefore, $D > 0$ means that $D > D_{\min}^{P,U}$, so by part iii) above $R(P, U, D) < \infty$ and hence $R(D) < \infty$. Also, for any distribution Q on $\hat{A}$

$$D_{\max} = \min_{y \in \hat{A}} E_P[\rho(X, y)] \leq E_{P \times Q}[\rho(X, Y)] = D_{\max}^{P,Q}$$

so, in particular, $D < D_{\max}^{P,Q^*}$. Part iii) then implies that $R(D) = R(P, Q^*, D) > 0$.

On the other hand, if $D \geq D_{\max}$, then by the definition of $D_{\max}$ in (7) there exists a $z \in \hat{A}$ such that

$$D_{\max} = E_P[\rho(X, z)] = E_{P\times\delta_z}[\rho(X, Y)] = D_{\max}^{P,\delta_z}$$

where δ_z is the measure attaching unit mass at z. This means that $D \geq D_{\max}^{P,\delta_z}$, so by part iii) above $R(P, \delta_z, D) = 0$, and hence $R(D) = 0$.

Part v): Since

$$R(P, Q^*, D) = R(D) \in (0, \infty)$$

from part iii) we have that $D \in [D_{\min}^{P,Q^*}, D_{\max}^{P,Q^*})$, and, in particular that $D_{\min}^{P,Q^*} < D_{\max}^{P,Q^*}$. So we only have to rule out the case $D = D_{\min}^{P,Q^*}$, but this was done in [17, Appendix II]. □

Appendix IV

Proof of Lemma 3: First observe that if Q_n is strictly a subprobability measure with $Q_n(\hat{A}^n) = Z < 1$, then using the probability measure $Q_n'(\cdot) = Z^{-1}Q_n(\cdot)$ we can make the expectation in (37) smaller by $\log Z < 0$. Therefore, it is enough to consider probability measures Q_n.

As for the achievability of the infimum, it suffices to note that it is taken over a compact set (the set over all p.m.f.'s Q_n on $\hat{A}^n$), and that the map $Q_n \mapsto E_{P_n}\{-\log Q_n(B(X_1^n, D))\}$ is lower semicontinuous. This follows from Fatou's Lemma in exactly the same way as it was shown in Appendix I that $\Lambda_{\mu,\nu}(\lambda)$ is upper semicontinuous in ν. □

Appendix V

Proof of Lemma 4: Define a joint probability measure Θ on the product space $A^n \times \hat{A}^n$ by restricting the product measure $P_n \times \tilde{Q}_n$ to be supported on $\{(x_1^n, y_1^n)\colon \rho_n(x_1^n, y_1^n) \leq D\}$

$$\frac{d\Theta}{dP_n \times \tilde{Q}_n}(x_1^n, y_1^n) = \frac{1}{\tilde{Q}_n(B(x_1^n, D))} \mathbb{I}_{\{y_1^n \in B(x_1^n, D)\}}.$$

Observe that the A^n-marginal of Θ is P_n, and let Θ_2 denote its $\hat{A}^n$-marginal. Then, with (X_1^n, Y_1^n) distributed according to Θ

$$\begin{aligned} K_n(D) &\overset{\text{a)}}{=} E_{P_n}\left\{-\log \tilde{Q}_n(B(X_1^n, D))\right\} \\ &= H(\Theta \| P_n \times \tilde{Q}_n) \\ &\overset{\text{b)}}{=} E_{\Theta_2}\{H(\Theta(\cdot|Y_1^n) \| P_n(\cdot))\} + H(\Theta_2 \| \tilde{Q}_n) \\ &\overset{\text{c)}}{\geq} H(\Theta \| P_n \times \Theta_2) \\ &\overset{\text{d)}}{=} I(X_1^n; Y_1^n) \\ &\overset{\text{e)}}{\geq} R_n(D) \end{aligned}$$

where a) follows from the definition of $K_n(D)$, b) follows from the chain rule for relative entropy (see, e.g., [25, eq. (3.11.5)] or [11, Theorem D.13]), c) follows from the nonnegativity of relative entropy, d) is just the definition of mutual information, and e) comes from the definition of $R(D)$, since $E_\Theta[\rho_n(X_1^n, Y_1^n)] \leq D$. □

Appendix VI

Proof of Proposition 3: Since $D \in (0, D_{\max})$, Proposition 2 part v) implies that $D \in (D_{\min}^{P,Q^*}, D_{\max}^{P,Q^*})$, so we may invoke [35, Corollary 1] to obtain

$$\begin{aligned} &-\log Q_n^*(B(X_1^n, D)) \\ &\qquad = nR(\hat{P}_n, Q^*, D) + \frac{1}{2}\log n + O(1) \quad \text{a.s.} \end{aligned}$$

where $\hat{P}_n$ is the empirical measure induced by X_1^n on A, i.e., the measure that assigns mass $1/n$ to each one of the values X_i, $i = 1, 2, \cdots, n$. Also, [10, Theorem 3] says that

$$nR(\hat{P}_n, Q^*, D) = nR(D) + \sum_{i=1}^{n} f(X_i) + o(\sqrt{n}) \quad \text{a.s.}$$

but a simple examination of the proof in [10] shows that we may replace the term $o(\sqrt{n})$ above by $O(\log\log n)$, without any changes in the proof. Combining these two results completes the proof of the proposition. □

Acknowledgment

The author wishes to thank A. Dembo, D. Gatzouras, and H. Rubin for various enlightening technical discussions, and A. Barron for his useful comments on an earlier version of this paper.

References

[1] P. H. Algoet, "Log-optimal investment," Ph.D. dissertation, Dept. Elec. Eng., Stanford Univ., Stanford, CA, 1985.
[2] A. R. Barron, "Logically smooth density estimation," Ph.D. dissertation, Dept. Elec. Eng., Stanford Univ., Stanford, CA, 1985.
[3] R. Bell and T. M. Cover, "Game-theoretic optimal portfolios," *Management Sci.*, vol. 34, no. 6, pp. 724–733, 1988.
[4] T. Berger, *Rate Distortion Theory: A Mathematical Basis for Data Compression*. Englewood Cliffs, NJ: Prentice-Hall, 1971.
[5] T. Berger and J. D. Gibson, "Lossy source coding," *IEEE Trans. Inform. Theory*, vol. 44, pp. 2693–2723, Oct. 1998.
[6] L. Breiman, *Probability* , ser. SIAM Classics in Applied Mathematics, vol. 7. Philadelphia, PA: SIAM, 1992.
[7] Z. Chi, "The first order asymptotics of waiting times with distortion between stationary processes," preprint, 1999.
[8] P.A. Chou, M. Effros, and R. M. Gray, "A vector quantization approach to universal noiseless coding and quantizations," *IEEE Trans. Inform. Theory*, vol. 42, pp. 1109–1138, July 1996.
[9] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York, NY: J. Wiley, 1991.
[10] A. Dembo and I. Kontoyiannis, "The asymptotics of waiting times between stationary processes, allowing distortion," *Ann. Appl. Probab.*, vol. 9, pp. 413–429, 1999.
[11] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*, 2nd ed. New York, NY: Springer-Verlag, 1998.
[12] R. Durrett, *Probability: Theory and Examples*, 2nd ed. Belmont, CA: Duxbury, 1996.
[13] P. Elias, "Universal codeword sets and representations of the integers," *IEEE Trans. Inform. Theory*, vol. IT-21, pp. 194–203, 1975.
[14] J. C. Kieffer, "Sample converses in source coding theory," *IEEE Trans. Inform. Theory*, vol. 37, pp. 263–268, Mar. 1991.
[15] I. Kontoyiannis, "Second-order noiseless source coding theorems," *IEEE Trans. Inform. Theory*, vol. 43, pp. 1339–1341, July 1997.
[16] ——, "Asymptotic recurrence and waiting times for stationary processes," *J. Theor. Probab.*, vol. 11, pp. 795–811, 1998.
[17] ——, "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.
[18] R. E. Krichevsky and V. K. Trofimov, "The performance of universal coding," *IEEE Trans. Inform. Theory*, vol. IT-27, pp. 199–207, Mar. 1981.

[19] T. Linder, G. Lugosi, and K. Zeger, "Rates of convergence in the source coding theorem, in empirical quantizer design, and in universal lossy source coding," *IEEE Trans. Inform. Theory*, vol. 40, pp. 1728–1740, Nov. 1994.

[20] ——, "Fixed-rate universal lossy source coding and rates of convergence for memoryless sources," *IEEE Trans. Inform. Theory*, vol. 41, pp. 665–676, May 1995.

[21] N. Merhav, "A comment on 'A rate of convergence result for a universal D-semifaithful code'," *IEEE Trans. Inform. Theory*, vol. 41, pp. 1200–1202, July 1995.

[22] R. M. Neuhoff, D. L. Gray, and L. D. Davisson, "Fixed rate universal block source coding with a fidelity criterion," *IEEE Trans. Inform. Theory*, vol. IT-21, pp. 511–523, Sept. 1975.

[23] D. Ornstein and P. C. Shields, "Universal almost sure data compression," *Ann. Probab.*, vol. 18, pp. 441–452, 1990.

[24] R. J. Pilc, "The transmission distortion of a source as a function of the encoding block length," *Bell Syst. Tech. J.*, vol. 47, pp. 827–885, 1968.

[25] M. S. Pinsker, *Information and Information Stability of Random Variables and Processes*. San Francisco, CA: Holden-Day, 1964.

[26] H. L. Royden, *Real Analysis*. New York, NY: Macmillan, 1988.

[27] W. Rudin, *Real and Complex Analysis*. New York, NY: McGraw-Hill, 1987.

[28] C. E. Shannon, "Coding theorems for a discrete source with a fidelity criterion," in *IRE Nat. Conv. Rec.*, vol. 4, 1959, pp. 142–163.

[29] P. C. Shields, "String matching bounds via coding," *Ann. Probab.*, vol. 25, pp. 329–336, 1997.

[30] J. Shtarkov, "Coding of discrete sources with unknown statistics," in *Topics in Information Theory*, I. Csiszár and P. Elias, Eds. Amsterdam, The Netherlands: North Holland (Coll. Math. Soc. J. Bolyai), 1977, vol. 16, pp. 559–574.

[31] M. Sion, "On general minimax theorems," *Pac. J. Math.*, vol. 8, pp. 171–176, 1958.

[32] A. D. Wyner, "Communication of analog data from a Gaussian source over a noisy channel," *Bell Syst. Tech. J.*, vol. 47, pp. 801–812, 1968.

[33] A. D. Wyner and J. Ziv, "The sliding-window Lempel–Ziv algorithm is asymptotically optimal," *Proc. IEEE*, vol. 82, pp. 872–877, June 1994.

[34] 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.

[35] E.-H. Yang and Z. Zhang, "On the redundancy of lossy source coding with abstract alphabets," *IEEE Trans. Inform. Theory*, vol. 45, pp. 1092–1110, May 1999.

[36] ——, "The redundancy of source coding with a fidelity criterion—Part II: Coding at a fixed rate level with unknown statistics," preprint, 1999.

[37] ——, "The redundancy of source coding with a fidelity criterion—Part III: Coding at a fixed distortion level with unknown statistics," preprint, 1999.

[38] B. Yu and T. P. Speed, "A rate of convergence result for a universal D-semifaithful code," *IEEE Trans. Inform. Theory*, vol. 39, pp. 813–820, May 1993.

[39] 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.

[40] J. Ziv, "Coding of sources with unknown statistics—Part II: Distortion relative to a fidelity criterion," *IEEE Trans. Inform. Theory*, vol. IT-18, pp. 389–394, May 1972.