Arbitrary Source Models and Bayesian Codebooks in Rate-Distortion Theory

Ioannis Kontoyiannis Junshan Zhang

March 2002

Abstract — We characterize the best achievable performance of lossy compression algorithms operating on arbitrary random sources, and with respect to general distortion measures. Direct and converse coding theorems are given for variable-rate codes operating at a fixed distortion level, emphasizing: (a) non-asymptotic results, (b) optimal or near-optimal redundancy bounds, and (c) results with probability one. This development is based in part on the observation that there is a precise correspondence between compression algorithms and probability measures on the reproduction alphabet. This is analogous to the Kraft inequality in lossless data compression. In the case of stationary ergodic sources our results reduce to the classical coding theorems. As an application of these general results, we examine the performance of codes based on mixture codebooks for discrete memoryless sources. A mixture codebook (or Bayesian codebook) is a random codebook generated from a mixture over some class of reproduction distributions. We demonstrate the existence of universal mixture codebooks, and show that it is possible to *universally* encode memoryless sources with redundancy of approximately $(d/2)\log n$ bits, where d is the dimension of the simplex of probability distributions on the reproduction alphabet.

Index Terms — Rate-distortion theory, redundancy rate, mixture codebooks, data compression.

[1]Ioannis Kontoyiannis is with the Division of Applied Mathematics and the Department of Computer Science, Brown University, Box F, 182 George St., Providence, RI 02912, USA. Email: yiannis@dam.brown.edu Web: www.dam.brown.edu/people/yiannis/

[2]Junshan Zhang is with the Department of Electrical Engineering, Arizona State University, Tempe, AZ 85287-7206, USA. Email: junshan.zhang@asu.edu Web: www.eas.asu.edu/~junshan/

[3]Ioannis Kontoyiannis was supported in part by NSF grants #0073378-CCR, DMS-9615444, and by USDA-IFAFS grant #00-52100-9615. Junshan Zhang was supported in part by ASU Faculty Initiation grant DM11001.

1 Introduction

Suppose data are generated by a random process, or source, $\{X_n\ ;\ n \geq 1\}$. Roughly speaking, the main objective of data compression is to find efficient representations of data strings $x_1^n = (x_1, x_2, \ldots, x_n)$ by variable-length binary strings $\psi_n(x_1^n)$. If we let A denote the *source alphabet*, then the map $\psi_n : A^n \to \{0,1\}^*$ from A^n to the set of finite-length binary strings is a (variable-length) block code of length n. The compression performance of such a code is described by its length-function

$$L_n(x_1^n) = \text{ length of } [\psi_n(x_1^n)] \quad \text{bits, for } x_1^n \in A^n.$$

In *lossless* data compression the natural class of codes to consider is the class of uniquely decodable codes ψ_n. As is well-known, the Kraft inequality (see, e.g., [7, p. 90]) provides a correspondence between uniquely decodable codes ψ_n, and probability distributions Q_n on A^n:

> KRAFT INEQUALITY: ($\Leftarrow$) For any uniquely decodable code (ψ_n, L_n) there is a probability measure Q_n on A^n such that:
>
> $$L_n(x_1^n) \geq -\log Q_n(x_1^n) \quad \text{bits, for all } x_1^n.$$
>
> ($\Rightarrow$) Given any probability measure Q_n on A^n there is a uniquely decodable code (ψ_n, L_n) such that:
>
> $$L_n(x_1^n) \leq -\log Q_n(x_1^n) + 1 \quad \text{bits, for all } x_1^n.$$

[Above and throughout the paper, log denotes the logarithm taken to base 2.] In the first part above, for Q_n we can take the measure

$$Q_n(x_1^n) \stackrel{\triangle}{=} Z^{-1} 2^{-L_n(x_1^n)}$$

where Z is the normalizing constant

$$Z = \sum_{x_1^n \in A^n} 2^{-L_n(x_1^n)}.$$

Then the usual statement of the Kraft inequality says that $Z \leq 1$.

Turning to *lossy* compression, we consider the problem of variable-rate coding at a fixed distortion level. More precisely, for each data string $x_1^n = (x_1, x_2, \ldots, x_n) \in A^n$ produced by the source $\{X_n\}$, our goal is to find an "accurate" representation of x_1^n by a string $y_1^n = (y_1, y_2, \ldots, y_n)$ taking values in the *reproduction alphabet* $\hat{A}$. The accuracy or "distortion" between two such strings is measured by a family of arbitrary distortion measures $\rho_n : A^n \times \hat{A}^n \to [0, \infty)$, $n \geq 1$ (more precise definitions will be given below).

The class of codes we consider here is the collection of *variable-length 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 discrete (finite or countably infinite) subset of $\hat{A}^n$, called the *codebook*;

(*b*) $\phi_n : A^n \to B_n$ is the *encoder* or *quantizer*;

(c) $\psi_n : B_n \to \{0,1\}^*$ is a uniquely decodable representation of the elements of B_n by finite-length binary strings.

We say the code $C_n = (B_n, \phi_n, \psi_n)$ *operates at distortion level* D (for some $D \geq 0$), if it encodes each source string with distortion D or less:

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

The compression performance of a code C_n is described by its length function

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

For a code C_n with associated length function ℓ_n we will often write $C_n = (B_n, \phi_n, \psi_n, \ell_n)$ or simply (C_n, ℓ_n).

The main theoretical issue of interest here is to characterize the best achievable compression performance of such codes. For stationary ergodic sources, Shannon [25] gave the first such general characterization in terms of the rate-distortion function. In this paper we adopt a somewhat different point of view: We take as the starting point a lossy version of the Kraft inequality, and use that as the basis for the subsequent general development. This approach leads to a natural formulation of the rate-distortion question as a convex selection problem, and allows us to consider, at least for part of the way, completely arbitrary source distributions and distortion measures. This approach has its roots in the earlier work of Bell and Cover [3] and of Kieffer [18].

1.1 Outline

Our first main result (part of Theorem 1 in Section 2) is the following lossy version of the Kraft inequality. Given a source string $x_1^n \in A^n$ and a distortion level $D \geq 0$, let $B(x_1^n, D)$ denote the "distortion-ball" of radius D centered at x_1^n (see (7) for a precise definition).

> LOSSY KRAFT INEQUALITY: ($\Leftarrow$) For any code $C_n = (B_n, \phi_n, \psi_n)$ with associated length function ℓ_n, operating at distortion level D, there is a probability measure Q_n on $\hat{A}^n$ such that:
>
> $$\ell_n(x_1^n) \geq -\log Q_n(B(x_1^n, D)) \quad \text{bits, for all } x_1^n.$$
>
> ($\Rightarrow$) Given any "admissible" sequence of probability measures $\{Q_n\}$ on $\hat{A}^n$, there is a sequence of codes $\{C_n = (B_n, \phi_n, \psi_n)\}$ with associated length functions $\{\ell_n\}$, operating at distortion level D, such that:
>
> $$\ell_n(X_1^n) \leq -\log Q_n(B(X_1^n, D)) + (1+\epsilon)\log n \quad \text{bits, eventually, w.p.1}$$
>
> where 'w.p.1' above and throughout the paper means 'with probability one.'

As will become apparent below, the assumption of "admissibility" of the measures $\{Q_n\}$ is simply the natural requirement that random codebooks generated according to these measures do not yield codes with infinite rate.

In Theorems 1 and 2 in the next section it is also shown that the same code performance as in the second part of the lossy Kraft inequality can be achieved in expectation. Further, if for a given sequence of measures $\{Q_n\}$ more detailed information is available on the asymptotic behavior of the "code-lengths"

$$\ell_n(X_1^n) \;\approx\; -\log Q_n(B(X_1^n, D)), \tag{1}$$

then more precise statements can be made about the redundancy achieved by the corresponding codes. The converse part ($\Leftarrow$) of the lossy Kraft inequality is based on an extension of a simple argument that was implicitly used in [21], and the corresponding direct coding theorems (in contrast to the lossless case) are asymptotic, and they are based on random coding arguments. In order to obtain the precise form of the second-order terms in the description lengths (the terms of order $\log n$), extra care is needed in constructing efficient codes.

In view of the codes-measures correspondence implied by the lossy Kraft inequality, the problem of understanding the best achievable compression performance is reduced, at least conceptually, to identifying the "optimal" measures Q_n and understanding the exact behavior of the approximate code lengths (1). As we will see, this correspondence leads to a characterization of the achievable performance of compression algorithms not in terms of the rate-distortion function, but in terms of a related quantity $K_n(D)$, defined as

$$K_n(D) \stackrel{\triangle}{=} \inf_{Q_n} E[-\log Q_n(B(X_1^n, D))] \tag{2}$$

where the infimum is over all probability measures Q_n on $\hat{A}^n$.

Let us assume for a moment that the above infimum is achieved for some $Q_n = \widetilde{Q}_n$. In Theorems 3 and 4 we give both asymptotic and finite-n results on the optimality of the measures $\widetilde{Q}_n$ and the codes they generate. First we show that for any code C_n with length function ℓ_n operating at distortion level D,

$$E[\ell_n(X_1^n)] \geq K_n(D) \geq R_n(D) \quad \text{bits} \tag{3}$$

where $R_n(D)$ is the nth order rate distortion function of the source. Then we show that the measures $\widetilde{Q}_n$ are "competitively optimal" in that, for any measure Q_n and any $k > 0$,

$$\Pr\{-\log Q_n(B(X_1^n, D)) \leq -\log \widetilde{Q}_n(B(X_1^n, D)) - k\} \leq 2^{-k} \tag{4}$$

(see also Remark 3 after Theorem 3). Moreover, we prove that the codes generated according to the measures $\{\widetilde{Q}_n\}$ are *asymptotically* optimal, up to about $(\log n)$ bits:

$$-\log Q_n(B(X_1^n, D)) \geq -\log \widetilde{Q}_n(B(X_1^n, D)) - (1+\epsilon)\log n \quad \text{bits, eventually, w.p.1.} \tag{5}$$

The statements in (3)–(5) are given in Theorems 3 and 4. Special cases of these results under much more restrictive assumptions (a finite reproduction alphabet $\hat{A}$ and a bounded, single-letter distortion measure) recently appeared in [21].

Note that, so far, *no assumptions* have been made on the source distribution or the distortion measures ρ_n.[1] As a sanity check we consider the case of stationary ergodic processes and subadditive distortion measures, and we show in Theorem 5 that in this case our general results reduce to Kieffer's pointwise coding theorems in [18], where the quantity $K_n(D)$ was defined and used extensively.

As an application of this general framework we consider the problem of universal coding for memoryless sources with respect to single letter distortion measures. Following the corresponding development in universal lossless compression (see [6] and the references therein), we examine the performance of random codes based on mixture codebooks. Let $\{X_n\}$ be an independent and identically distributed (i.i.d.) source over a finite alphabet A, and let the reproduction alphabet $\hat{A}$ also be finite. A *mixture codebook* (or *Bayesian codebook*) is a random codebook generated according to sequence of distributions $\{M_n\}$, where each M_n is a mixture of i.i.d. distributions Q^n on $\hat{A}^n$:

$$M_n(y_1^n) \triangleq \int_{\text{all } Q} Q^n(y_1^n)\, d\pi(Q) \quad \text{for } y_1^n \in \hat{A}^n. \tag{6}$$

In Theorem 6 sufficient conditions are given for the "prior" distribution π, guaranteeing that the codes generated according to the mixture distributions $\{M_n\}$ are *universal* over the class of all memoryless sources on A.

Under further regularity conditions on the prior π (assuming it has a continuous and everywhere positive density with respect to Lebesgue measure), it is shown in Theorem 7 that the redundancy of the mixture codebooks asymptotically does not exceed $\approx (d/2)\log n$ bits, where $d = |\hat{A}| - 1$ is the dimension of the simplex of distributions on the reproduction alphabet $\hat{A}$. Therefore, the price paid for universality is about $(1/2)\log n$ bits per degree of freedom, where, in contrast with the case of lossless compression, "freedom" is measured with respect to the class of possible *codebook distributions* we are allowed to use, and not with respect to the size of the class of sources considered.

In view of the recent results in [28] this rate appears to be optimal and it agrees with the corresponding results in lossless compression [23][6], as well as with those obtained in [5] within the framework of vector-quantization. Moreover, the codes generated by mixture codebooks appear to be the first examples of codes whose redundancy is shown to be near-optimal not only in expectation, but also with probability one.

1.2 Earlier Work

A number of relevant papers have already been pointed out and briefly discussed. We also mention that, implicitly, the codes-measures correspondence has been used in the literature by various authors over the past five years or so; relevant works include [22][27][29][19][21], among others.

A different approach for dealing with arbitrary sources has been introduced by Steinberg, Verdú, and Han [26][13][14], based on the 'information-spectrum' method. This leads to a different (asymptotic) characterization of the best achievable performance in lossy data compression. Unlike in those

[1] To be absolutely precise, we should mention that for the results discussed above we do need to make the trivial assumption that finite-rate coding is indeed possible at the distortion level we consider.

works, more emphasis here has been placed on obtaining non-asymptotic converses, tight redundancy bounds, and coding theorems with probability one rather than in expectation.

The problem of determining the best achievable (expected) redundancy rate in lossy compression was extensively treated in [35]; see also the references therein. Suboptimal universal redundancy rates in expectation were computed in [31][15], and asymptotically tight bounds were recently obtained in [30][28] where converses were also established. Taking a different point of view, [5] employs high-rate quantization theory to address the question of how close one can come to the OPTA (Optimum Performance Theoretically Achievable) function, as opposed to the rate-distortion function. Another related problem, that of characterizing the optimal *pointwise* redundancy (including the question of universality) has been treated in detail in [21].

All of the works mentioned so far exhibit universal codes based on "multiple codebooks" or "two-stage descriptions." That is, the source string x_1^n is examined, and based on its statistical properties one of several possible codes is chosen to encode x_1^n. First the index of the chosen code is communicated to the decoder, then the encoded version of x_1^n is sent. In contrast, the universal codes presented here are based on a single mixture codebook that works well for all memoryless sources. This construction, facilitated by the codes-measures correspondence, is developed in close analogy to the corresponding lossless compression results; see [8][23][6][2] and the references therein. Mixture codebooks for lossy compression are also briefly considered in [34][33], but with the mixture being over fixed-composition codebooks of a given type, rather than over distributions.

Finally we note that a different connection between rate-distortion theory and Baysian inference has been drawn in [32][16].

2 Statements of Results

2.1 The Setting

We introduce some definitions and notation that will remain in effect throughout the paper. Let $\{X_n \ ; \ n \geq 1\}$ be a random process, or source, taking values in the *source alphabet* A, where A is assumed to be a complete, separable metric space, equipped with its Borel σ-field $\mathcal{A}$. For $1 \leq i \leq \infty$ we write X_i^j for the vector of random variables $(X_i, X_{i+1}, \ldots, X_j)$, and similarly write $x_i^j = (x_i, x_{i+1}, \ldots, x_j) \in A^{j-i+1}$ for a realization of X_i^j. The distribution of X_1^n is denoted by P_n (more precisely, P_n is a Borel probability measure on $(A^n, \mathcal{A}^n)$), and the probability measure describing the distribution of the entire process is denoted by $\mathbb{P}$.

Similarly, for the *reproduction alphabet* we take $\hat{A}$ to be a complete, separable metric space together with its Borel σ-field $\hat{\mathcal{A}}$, where $\hat{A}$ may or may not be the same as A.[2] For each $n \geq 1$ we assume that we are given a distortion measure ρ_n, that is, a nonnegative function $\rho_n : A^n \times \hat{A}^n \to [0, \infty)$.[3] For each source string $x_1^n \in A^n$ and distortion level $D \geq 0$ we define the distortion-ball $B(x_1^n, D)$ as the

[2] To avoid uninteresting technicalities, we assume throughout that all singletons are measurable, i.e., $\{x\} \in \mathcal{A}$ and $\{y\} \in \hat{\mathcal{A}}$ for all $x \in A$, $y \in \hat{A}$.

[3] Assuming, of course, that each ρ_n is measurable with respect to the product σ-field $\mathcal{A}^n \times \hat{\mathcal{A}}^n$.

collection of all strings $y_1^n \in \hat{A}^n$ that have distortion D or less with respect to x_1^n:

$$B(x_1^n, D) \stackrel{\triangle}{=} \{y_1^n \in \hat{A}^n : \rho_n(x_1^n, y_1^n) \leq D\}. \tag{7}$$

Finally, throughout the paper log denotes the logarithm taken to base 2 and $\log_e$ denotes the natural logarithm. Unless explicitly stated otherwise, all familiar information-theoretic quantities (the relative entropy, rate-distortion function, and so on) are defined in terms of logarithms taken to base 2, and are hence expressed in bits.

2.2 Random Codebooks

Given an arbitrary (Borel) probability measure Q_n on $\hat{A}^n$, by a *random codebook generated according to* Q_n we mean a collection of independent random vectors

$$Y(i) = Y_1^n(i), \quad i \geq 1$$

taking values in $\hat{A}^n$, each generated according to the measure Q_n. These random vectors will serve as the codebook in various direct coding theorems presented below. Given a random source string X_1^n, a random codebook as above, and a distortion level $D \geq 0$, we define the *waiting time* W_n as the index i of the first element of the codebook that matches X_1^n with distortion D or less

$$W_n \stackrel{\triangle}{=} \inf\{i \geq 1 \ : \ \rho_n(X_1^n, Y_1^n(i)) \leq D\}$$

with the usual convention that the infimum of the empty set equals ∞.

In the coding scenario we assume that the random codebook is available to both the encoder and the decoder. The gist of the direct coding theorems below is to find efficient ways for the encoder to communicate to the decoder the value of the waiting time W_n; once this is available, the decoder can read off the codeword $Y_1^n(W_n)$ and obtain a D-close version of X_1^n.

It perhaps worth mentioning that the relationship of the nth codebook to the $(n+1)$st codebook will be irrelevant in all our subsequent direct coding arguments. For the sake of being specific it may be convenient to think of codebooks with different block-lengths as being independent. On the other hand, if Q_n and Q_{n+1} are consistent measures, i.e., if Q_n happens to be the n-dimensional measure induced by Q_{n+1} on $\hat{A}^n$, then the $(n+1)$st codebook can be generated from the nth one by extending each of its codewords by an additional letter generated according to the conditional distribution induced by Q_{n+1}.

2.3 Arbitrary Sources

Let $\{X_n\}$ be an arbitrary source and $\{\rho_n\}$ a given sequence of distortion measures as above. At various points below we will need to impose the following assumptions. They are variations of [18, Condition 2$'$)].

(WQC): For a distortion level $D \geq 0$ we say that the *weak quantization condition (WQC) holds at level D* if for each n there is a (measurable) quantizer $q^{(n)} : A^n \to B_n \subset \hat{A}^n$ such that B_n is a finite or countably infinite set, and

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

(QC): For a distortion level $D \geq 0$ we say that the *quantization condition (QC) holds at level D*, if (WQC) holds with respect to quantizers $q^{(n)}$ of finite rate,

$$M_1 \stackrel{\triangle}{=} \sup_{n \geq 1} \frac{1}{n} H(q^{(n)}(X_1^n)) < \infty,$$

where $H(q^{(n)}(X_1^n))$ denotes the entropy (in bits) of the discrete random variable $q^{(n)}(X_1^n)$.

(pSQC): For a distortion level $D \geq 0$ we say that the *p-strong quantization condition (pSQC) holds at level D* for some $1 \leq p \leq \infty$, if (WQC) holds with respect to quantizers $q^{(n)}$ also satisfying

$$M_p \stackrel{\triangle}{=} \sup_{n \geq 1} \frac{1}{n} \left\{ E[(-\log \mu_n(q^{(n)}(X_1^n)))^p] \right\}^{1/p} < \infty \tag{8}$$

where μ_n denotes the (discrete) distribution of $q^{(n)}(X_1^n)$ on $\hat{A}^n$. [For $p = \infty$ the above expression is interpreted, as usual, to be the corresponding L_∞ norm, i.e., $M_\infty = \sup_n (1/n) \| -\log \mu_n(q^{(n)}(X_1^n)) \|_\infty$.]

Note that clearly (pSQC) $\Rightarrow$ (QC) $\equiv$ (1SQC) $\Rightarrow$ (WQC). Also observe that, if $\boldsymbol{X}$ is a stationary source and the $\{\rho_n\}$ are single-letter (or subadditive) distortion measures, then each of the above three conditions reduces to the existence of a suitable scalar quantizer, that is, each quantization condition reduces to the corresponding requirement for $n = 1$.

For the statement of the next theorem we also need the following definitions. For each $n \geq 1$, let Q_n be a probability measure on $\hat{A}^n$. The sequence $\{Q_n\}$ is called *admissible* if the random codebooks generated according to these measures yield codes with finite rate, with probability one. Formally, $\{Q_n\}$ is admissible if there is a finite constant R such that:

$$\limsup_{n \to \infty} -\frac{1}{n} \log Q_n(B(X_1^n, D)) \leq R < \infty \quad \text{w.p.1.} \tag{9}$$

Similarly, the sequence $\{Q_n\}$ is called *admissible in expectation* if it yields random codes with finite average rate, that is,[4]

$$R = \limsup_{n \to \infty} \frac{1}{n} E[-\log Q_n(B(X_1^n, D))] < \infty. \tag{10}$$

The following result demonstrates the correspondence between sequences of codes C_n operating at distortion level D and sequences of measures Q_n on the reproduction spaces $\hat{A}^n$. It is proved in Section 3.1.

[4] Note that, for any probability measure Q_n on $(\hat{A}^n, \hat{\mathcal{A}}^n)$, the function $x_1^n \mapsto -\log Q_n(B(x_1^n, D))$ is measurable with respect to $\mathcal{A}^n$. To see this, simply observe that we can take $Q_n(B(x_1^n, D))$ to be the regular conditional probability of the event $\{\rho_n(X_1^n, Y_1^n) \leq D\}$ with respect to $P_n \times Q_n$, conditional on the σ-field $\mathcal{A}^n$.

Theorem 1. Codes-Measures Correspondence: Given a distortion level $D \geq 0$, assume that the weak quantization condition (WQC) holds at level D.

(i) For any code $C_n = (B_n, \phi_n, \psi_n, \ell_n)$ operating at distortion level D there is a probability measure Q_n on $\hat{A}^n$ such that

$$\ell_n(x_1^n) \geq -\log Q_n(B(x_1^n, D)) \quad \text{bits, for all } x_1^n \in A^n.$$

(ii) For any admissible sequence of probability measures $\{Q_n\}$ there is a sequence of codes $\{C_n = (B_n, \phi_n, \psi_n, \ell_n)\}$ operating at distortion level D such that

$$\ell_n(X_1^n) \leq -\log Q_n(B(X_1^n, D)) + \log n + 3\log\log n + \text{Const.} \quad \text{bits, eventually, w.p.1.}$$

(iii) For any sequence of probability measures $\{Q_n\}$ that are admissible in expectation, there is a sequence of codes $\{C_n = (B_n, \phi_n, \psi_n, \ell_n)\}$ operating at distortion level D such that

$$E[\ell_n(X_1^n)] \leq E[-\log Q_n(B(X_1^n, D))] + \log n + 2\log\log n + \text{Const.} \quad \text{bits, eventually.}$$

Remark 1: The constant terms in (ii) and (iii) above depend only on the constant R that bounds the asymptotic rate of the measures Q_n in (9) and (10), respectively.

Our next result shows that the redundancy rates of the codes in parts (ii) and (iii) of Theorem 1 can be improved when we have more information about the asymptotic behavior of the approximate code-lengths (1) corresponding to the measures Q_n. It is proved in Section 3.2.

Theorem 2. Improved Redundancy Rates:

(i) Given a distortion level $D \geq 0$, assume that the weak quantization condition (WQC) holds at level D. Let $\{Q_n\}$ be an admissible sequence of probability measures and assume that

$$Z_n \stackrel{\triangle}{=} |-\log Q_n(B(X_1^n, D)) - nR| \leq B\sqrt{n}\log n \quad \text{eventually, w.p.1} \tag{11}$$

for some finite constant B. Then there are codes $\{C_n = (B_n, \phi_n, \psi_n, \ell_n)\}$ operating at distortion level D such that

$$\ell_n(X_1^n) \leq -\log Q_n(B(X_1^n, D)) + \frac{1}{2}\log n + 2\log\log n + \text{Const.} \quad \text{bits, eventually, w.p.1.}$$

(ii) Given a distortion level $D \geq 0$, assume that the p-strong quantization condition (pSQC) holds at level D for some $p \geq 2$. Assume that the sequence $\{Q_n\}$ is admissible in expectation, and let Z_n be defined as above. If for n large enough and for some finite nonzero constants B and C

$$\Pr\{Z_n > B\sqrt{n}\log n\} \leq \frac{C}{n^q}$$

where $1/p + 1/q = 1$, then there are codes $\{C_n = (B_n, \phi_n, \psi_n, \ell_n)\}$ operating at distortion level D such that

$$E[\ell_n(X_1^n)] \leq E[-\log Q_n(B(X_1^n, D))] + \frac{1}{2}\log n + 2\log\log n + \text{Const.} \quad \text{bits, eventually.}$$

Remark 2: The constant term in the statement of part (i) depends only on the constant B in (11). The constant term in part (ii) depends only on the constants B, C, and the constant M_p in the (pSQC) condition (8).

Note that in many important cases of interest, the assumptions of Theorems 1 and 2 on the asymptotic behavior of the approximate code lengths (1) can be verified via the "generalized AEP" and its refinements; see [10] for an extensive discussion.

Theorems 1 and 2 indicate that the best achievable performance of codes operating at distortion level D can be understood almost entirely on the basis of understanding the precise behavior of the approximate code lengths (1). To this end, for each $n \geq 1$ and distortion level $D \geq 0$ we define (cf. (2)) the quantity

$$K_n(D) \triangleq \inf_{Q_n} E[-\log Q_n(B(X_1^n, D))]$$

where the infimum is over all probability measures Q_n on $\hat{A}^n$.

The next theorem gives finite-n bounds on the achievable compression performance of arbitrary codes operating at distortion level D. In particular, it identifies the measures $Q_n = \widetilde{Q}_n$ that are optimal in terms of compression performance, as those that achieve the infimum in the definition of $K_n(D)$.

Theorem 3. Non-Asymptotic Bounds:

(i) For any code (C_n, ℓ_n) operating at distortion level D:

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

where $R_n(D)$ is the usual nth order rate-distortion function of the source $\{X_n\}$ [4], defined by

$$R_n(D) = \inf_{(X_1^n, Y_1^n):\, X_1^n \sim P_n \text{ and } E[\rho_n(X_1^n, Y_1^n)] \leq D} I(X_1^n; Y_1^n)$$

where $I(X_1^n; Y_1^n)$ denotes the mutual information between X_1^n and Y_1^n, and the infimum is taken over all jointly distributed random vectors (X_1^n, Y_1^n) with X_1^n having the source distribution P_n, and $E[\rho_n(X_1^n, Y_1^n)] \leq D$.

(ii) Assume that the infimum in the definition of $K_n(D)$ is achieved by some probability measure $\widetilde{Q}_n$ and that $K_n(D) < \infty$. Then for *any* probability measure Q_n on $\hat{A}^n$ we have

$$E\left[\frac{Q_n(B(X_1^n, D))}{\widetilde{Q}_n(B(X_1^n, D))}\right] \leq 1$$

and for any $k > 0$,

$$\Pr\{-\log Q_n(B(X_1^n, D)) \leq -\log \widetilde{Q}_n(B(X_1^n, D)) - k\} \leq 2^{-k}.$$

Remark 3. Competitive Optimality: The second result in part (ii) above is somewhat striking. It states that, for any fixed n, there is an optimal code operating at distortion level D (the code corresponding to $\widetilde{Q}_n$) with the following property: The probability that any other code beats the optimal one by k or more bits, is at most 2^{-k}. For a detailed example see [20, Section IV].

Remark 4. Achievability: Although in general there may not exist measures $\widetilde{Q}_n$ achieving the infimum in the definition of $K_n(D)$, when $\hat{A}$ is finite and $\{\rho_n\}$ is a sequence of bounded, single-letter distortion measures such $\widetilde{Q}_n$ were shown to exist in [21]. When the infimum is not achieved (or the achievability is not easy to check), it is still possible to recover the result of Theorem 3 (ii); see Remark 7 after Theorem 4 below.

Remark 5. $K_n(D)$ vs. $R_n(D)$: The function $K_n(D)$ can be defined in more familiar information-theoretic terms, making it more easily comparable to the rate-distortion function $R_n(D)$ – see Lemma 1 below. In that form, $K_n(D)$ is reminiscent of Kolmogorov's definition of ϵ-entropy (which explains the choice of the letter K in the notation). It is obvious from Lemma 1 that $K_n(D)$ is generally larger than $R_n(D)$, but for stationary ergodic sources their limiting values $\lim_n (1/n)K_n(D)$ and $\lim_n (1/n)R_n(D)$ are equal (see Theorem 5). Lemma 1 is proved in Appendix I.

Lemma 1. Alternative Characterization of $K_n(D)$: For all $D \geq 0$ we have

$$K_n(D) = \inf_{(X_1^n, Y_1^n):\, \rho_n(X_1^n, Y_1^n) \leq D \text{ w.p.1}} I(X_1^n; Y_1^n)$$

where the infimum is taken over all jointly distributed random vectors (X_1^n, Y_1^n) with X_1^n having the source distribution P_n, and $\rho_n(X_1^n, Y_1^n) \leq D$ with probability one.

Next we deduce the following asymptotic result from Theorem 3: Up to approximately $\log n$ bits, the code-lengths

$$\ell_n(X_1^n) \approx -\log \widetilde{Q}_n(B(X_1^n, D))$$

are both achievable and impossible to beat, with probability one. Since the proofs of Theorems 3 and 4 follow very closely along the lines of the corresponding results in [21] we only include brief outlines of their proofs in Section 3.3.

Theorem 4. Asymptotic Bounds: Assume that for all n large enough the infimum in the definition of $K_n(D)$ is achieved by some probability measure $\widetilde{Q}_n$, and that $K_n(D) < \infty$.

(i) For any sequence of codes $\{C_n, \ell_n\}$ operating at distortion level D:

$$\ell_n(X_1^n) \geq -\log \widetilde{Q}_n(B(X_1^n, D)) - \log n - 2\log\log n \quad \text{bits, eventually, w.p.1.}$$

(ii) Moreover, if the sequence $\{\widetilde{Q}_n\}$ is admissible and the weak quantization condition (WQC) holds at level D, then there is a sequence of codes $\{C_n, \ell_n\}$ operating at distortion level D such that

$$\ell_n(X_1^n) \leq -\log \widetilde{Q}_n(B(X_1^n, D)) + \log n + 3\log\log n + \text{Const.} \quad \text{bits, eventually, w.p.1.} \tag{12}$$

Similarly, if the sequence $\{\widetilde{Q}_n\}$ is admissible in expectation then there exist codes so that (12) holds in expectation, and with $2\log\log n$ in place of $3\log\log n$.

Remark 6. Admissibility: The assumption that the optimal measures $\{\widetilde{Q}_n\}$ are admissible is typically not restrictive. For example, as we will see in the next section, they are always admissible (as well as admissible in expectation) when the source is stationary and ergodic.

Remark 7. More on Achievability: If $K_n(D) < \infty$ but the infimum in the definition of $K_n(D)$ is not achieved, there always exists a function $\tilde{g}_n : A^n \to [0, \infty)$ in the L^1-closure of the collection $\mathcal{G}_n$ of functions on A^n defined by

$$\mathcal{G}_n \stackrel{\triangle}{=} \{g \in L^1(P_n) \ : \ g(x_1^n) \geq -\log Q_n(B(x_1^n, D)) \ \text{ for a probability measure } Q_n \text{ on } \hat{A}^n\} \tag{13}$$

such that $K_n(D) = E[\tilde{g}_n(X_1^n)]$; see [17, Theorem 1]. Moreover, the conclusions of Theorem 3 (ii) and the converse in Theorem 4 (i) all still hold if we replace the term $-\log \widetilde{Q}_n(B(X_1^n, D))$ by $\tilde{g}_n(X_1^n)$.

2.4 Ergodic Sources

We briefly consider the case of stationary ergodic sources and demonstrate how the classical coding theorems follow from the general results above. Assume that the source $\{X_n\}$ is stationary and ergodic, and that the distortion measures $\{\rho_n\}$ are *subadditive*, i.e.,

$$(m+n)\rho_{m+n}(x_1^{m+n}, y_1^{m+n}) \leq m\rho_m(x_1^m, y_1^m) + n\rho_n(x_{m+1}^{m+n}, y_{m+1}^{m+n}),$$

for all $x_1^{m+n} \in A^{m+n}$, $y_1^{m+n} \in \hat{A}^{m+n}$, and all $m, n \geq 1$. We will also assume the existence of a *reference letter* $\hat{a} \in \hat{A}$ such that

$$E[\rho_1(X_1, \hat{a})] < \infty.$$

The following theorem is essentially due to Kieffer [18]. To connect it with our earlier results we include a brief outline of its proof in Appendix II.

Theorem 5. Stationary Ergodic Sources: [18] Let $\{X_n\}$ be a stationary ergodic source such that the quantization condition (QC) holds for all $D > 0$. Assume that the distortion measures $\{\rho_n\}$ are subadditive, and that a reference letter exists. Then $K_n(D)$ is finite for all $n \geq 1$, and the limit

$$K(D) \stackrel{\triangle}{=} \lim_{n\to\infty} \frac{1}{n} K_n(D)$$

exists and is equal to the rate-distortion function of the source $\{X_n\}$,

$$R(D) \stackrel{\triangle}{=} \lim_{n\to\infty} \frac{1}{n} R_n(D).$$

If, moreover, for all n large enough the infimum in the definition of $K_n(D)$ is achieved by some probability measure $\widetilde{Q}_n$, then we also have:

(i) As $n \to \infty$:

$$-\frac{1}{n} \log \widetilde{Q}_n(B(X_1^n, D)) \to K(D) = R(D) \quad \text{w.p.1. and in } L^1. \tag{14}$$

(ii) For any sequence of codes $\{C_n, \ell_n\}$ operating at distortion level D,

$$\liminf_{n\to\infty} \frac{1}{n}\ell_n(X_1^n) \geq K(D) = R(D) \quad \text{bits/symbol, w.p.1} \tag{15}$$

and there exist codes achieving this bound with equality.

Remark 8. Even More on Achievability: As described in Remark 7, even when the achieving measure $\widetilde{Q}_n$ does not exist we can always find a function $\tilde{g}_n \in L^1(P_n)$ so that $\tilde{g}_n(X_1^n)$ plays the role of $-\log \widetilde{Q}_n(B(X_1^n, D))$. In this context, Theorem 2 of [17] implies that Theorem 5 (i) always holds with $\tilde{g}_n(X_1^n)$ in place of $-\log \widetilde{Q}_n(B(X_1^n, D))$, and therefore the pointwise converse (15) is also always valid.

2.5 Mixture Codebooks and Memoryless Sources

As we saw in Theorems 3 and 4, the optimal measures $\widetilde{Q}_n$ completely characterize the best compression performance of codes operating at a fixed distortion level. But the $\widetilde{Q}_n$ themselves are typically hard to describe explicitly.

In this section we restrict our attention to memoryless sources. Although it is not hard to see that even here the measures $\widetilde{Q}_n$ do not have a particularly simple structure, we do know [35][29][21] that, when dealing with a single memoryless source, *asymptotically* optimal compression can still be achieved by codes based on product measures, i.e., measures Q_n of the form $Q_n = Q^n$ on $\hat{A}^n$: Taking Q^* to be the optimal reproduction distribution at distortion level D, the codes generated according to the product measures $(Q^*)^n$ achieve near-optimal redundancy both in expectation and with probability one (see Proposition 1 and the discussion following Theorem 7 for details).

Turning to the problem of universal coding, the above discussion motivates us to consider codes based on random codebooks that are generated according to *mixtures* over the class of all product distributions Q^n on $\hat{A}^n$. The existence of universal mixture codebooks will be established, and sufficient conditions will be given under which the redundancy rate they achieve is optimal.

For the rest of this section $\{X_n\}$ is assumed to be a stationary memoryless source, that is, the random variables X_n are independent and identically distributed (i.i.d), with common distribution P on A. We take both A and $\hat{A}$ to be finite sets and write $|\hat{A}| = d + 1$ for the cardinality of $\hat{A}$. Given a distortion measure $\rho : A \times \hat{A} \to [0, \infty)$, we consider the family of single-letter distortion measures $\{\rho_n\}$:

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

We also make the customary assumption that for each $a \in A$ there is an $\hat{a} \in \hat{A}$ such that $\rho(a, \hat{a}) = 0$.

Following [27][21], for each probability distribution Q on $\hat{A}$ and all $D \geq 0$ we define the rate function

$$R(P, Q, D) \triangleq \inf_{(X,Y)\,:\,X\sim P,\ E[\rho(X,Y)]\leq D} \{I(X;Y) + H(Q_Y \| Q)\}$$

where $H(\cdot\|\cdot)$ denotes the relative entropy, the infimum is taken over all jointly distributed random variables (X, Y) such that X has the source distribution P and $E[\rho(X,Y)] \leq D$, and Q_Y denotes the marginal distribution of Y. It is not hard to see that $R(P,Q,D)$ is related to the rate-distortion function $R(D)$ of the source P via

$$R(D) = \inf_Q R(P,Q,D) = R(P,Q^*,D) \tag{16}$$

where the infimum is taken over all probability distributions Q on $\hat{A}$; cf. [27][21]. We let Q_n^* denote the product measures $(Q^*)^n$ on $\hat{A}^n$, $n \geq 1$, and we call the measures Q_n^* the *optimal reproduction distributions at distortion level* D (even though the achieving Q^* in (16) may not be unique).

The following proposition shows that $R(P,Q,D)$ characterizes the first-order compression performance of the random codebooks generated according to the product measures Q^n. In particular (recall Theorem 1) it implies that the codebooks generated according to the optimal reproduction distributions $\{Q_n^*\}$ achieve first-order optimal compression performance.

Proposition 1. [27] For any probability measure Q on $\hat{A}$, and for all $D > 0$:

$$\lim_{n\to\infty} -\frac{1}{n}\log Q^n(B(X_1^n, D)) = R(P,Q,D) \quad \text{w.p.1.} \tag{17}$$

Next define, as usual,

$$D_{\max} = D_{\max}(P) = \min_{y\in\hat{A}} E_P[\rho(X,y)]. \tag{18}$$

Since $R(D) = 0$ for $D \geq D_{\max}$, to avoid the trivial case when $R(D)$ is identically equal to zero we assume that $D_{\max} > 0$. Also, from now on we restrict our attention to distortion levels D in the interesting range $D \in (0, D_{\max})$.

Let $\mathcal{P}$ denote the d-dimensional simplex of probability measures Q on $\hat{A}$, and let π be a probability measure on $\mathcal{P}$; we refer to π as the "prior" distribution on $\mathcal{P}$. For each $n \geq 1$, we define the *mixture distribution* M_n on $\hat{A}^n$ as in (6): $M_n(y_1^n) = \int_{Q\in\mathcal{P}} Q^n(y_1^n)\, d\pi(Q)$, for $y_1^n \in \hat{A}^n$.

The next theorem gives a simple sufficient condition on the prior, under which the first-order performance of codes based on the mixture distributions M_n is universally optimal. Theorem 6 is proved in Section 4.1 using an argument similar to that used in the proof of the corresponding lossless result in [1].

Theorem 6. Universality of I.I.D. Mixtures: Let $\{X_n\}$ be an i.i.d. source with distribution P on A. For $D \in (0, D_{\max})$ let Q^* denote the optimal reproduction distribution of P at distortion level D. If the prior π has a density p with respect to Lebesgue measure on the simplex, and p is strictly positive in a neighborhood of Q^*, then:

$$-\log M_n(B(X_1^n, D)) \;\leq\; -\log Q_n^*(B(X_1^n, D)) + o(n) \quad \text{w.p.1, as } n \to \infty. \tag{19}$$

More generally, (19) remains valid as long as the prior π assigns positive mass to all neighborhoods

$$N(Q^*, \epsilon) \stackrel{\triangle}{=} \{Q \;:\; R(P,Q,D) < R(P,Q^*,D) + \epsilon\}, \quad \text{for } \epsilon > 0. \tag{20}$$

The conditions of Theorem 6 are easily seen to be satisfied, e.g., when the prior π has an everywhere strictly positive density p, or when π assigns positive mass to the optimal reproduction distribution itself. In particular, this includes the special case of discrete mixtures of the form

$$M_n = \sum_{i=1}^{\infty} w_i Q_i^n,$$

where Q^* is one of the mixing distributions Q_i. More generally, the discrete mixtures M_n are universal if the $\{Q_i\}$ are a countable dense subset of the simplex $\mathcal{P}$ and $\{w_i\}$ are strictly positive weights summing to one.

Remark 9. Universal Codes: Suppose that the assumptions of Theorem 6 hold for all Q^* on the simplex – e.g., take π to be normalized Lebesgue measure on the simplex or a discrete probability measure supported on a countable dense set. Then, in view of Proposition 1 and (16), Theorem 6 implies that the mixtures measures $\{M_n\}$ have

$$\limsup_{n\to\infty} -\frac{1}{n}\log M_n(B(X_1^n, D)) \le R(D) \quad \text{w.p.1.}$$

Therefore they are admissible, and Theorem 1 implies the existence of universal codes over the class of all memoryless sources on A.

Next it is shown that when the prior π satisfies certain smoothness conditions the asymptotic redundancy rate of the mixture codebooks is approximately $(d/2)\log n$ bits, where $d = |\hat{A}| - 1$ is the dimensionality of $\mathcal{P}$.

For Theorem 7 we only consider sources P that have $P(a) > 0$ for all $a \in A$, and we assume that the optimal reproduction distribution Q^* is unique and is achieved in the interior of the simplex. Formally, for each $D \ge 0$ we consider sources P in the collection:

$$\mathcal{S}(D) \stackrel{\triangle}{=} \{P \;:\; D < D_{\max}(P),\; Q^* \text{ is unique},\; \text{support}(P) = A,\; \text{and support}(Q^*) = \hat{A}\;\}. \tag{21}$$

Theorem 7. Pointwise Redundancy of I.I.D. Mixtures: Let $D > 0$, and assume that the prior π has a strictly positive, continuous density p with respect to Lebesgue measure. Then for any source distribution P in $\mathcal{S}(D)$,

$$-\log M_n(B(X_1^n, D)) \;\le\; -\log Q_n^*(B(X_1^n, D)) + \frac{d}{2}\log n + O(\log\log n) \quad \text{w.p.1, as } n\to\infty \tag{22}$$

where Q_n^* are the optimal reproduction distributions for the source P at distortion level D.

The proof of Theorem 7 (given in Section 4.3) is based on an argument using Laplace's method of integration, and closely parallels the proof of the corresponding result of Clarke and Barron [6] in the lossless case. Using (22) and applying Theorem 2 we get a sequence of universal codes with near-optimal redundancy rate.

Corollary 1. Universal Pointwise Redundancy: Let $D > 0$. There is a sequence of codes $\{C_n = (B_n, \phi_n, \psi_n, \ell_n)\}$ operating at distortion level D such that, for all memoryless sources P in $\mathcal{S}(D)$,

$$\ell_n(X_1^n) \leq -\log Q_n^*(B(X_1^n, D)) + \left(\frac{d+1}{2}\right) \log n + O(\log \log n) \quad \text{bits, w.p.1, as } n \to \infty \tag{23}$$

where Q_n^* are the optimal reproduction distributions for the source P, at distortion level D.

Corollary 1 follows from combining Theorem 7 with Theorem 2 (i). To verify that assumptions of Theorem 2 (i) are satisfied, recall [21, Proposition 3].

Remarks:

(a) *Interpretation*: What Theorem 7 and Corollary 1 really tell us is that the codes corresponding to the mixture distributions $\{M_n\}$ have code-lengths that do not exceed the "optimum"

$$\ell_n^*(X_1^n) \stackrel{\triangle}{=} -\log Q_n^*(B(X_1^n, D)) + \frac{1}{2} \log n \quad \text{bits}$$

by more than $\approx (d/2) \log n$ bits. Why call the code-lengths $\ell_n^*(X_1^n)$ optimum? In view of Theorem 2, $\ell_n^*(X_1^n)$ correspond to random codes generated according to the measures $\{Q_n^*\}$. In [29][35] and in [21] it was shown that these random codes are essentially asymptotically optimal, both in expectation and with probability one. Specifically, in [29] it is proved that

$$E[\ell_n^*(X_1^n)] \approx nR(D) + \log n \quad \text{bits}$$

but we also know [35] that *any* code operating at distortion level D has expected code-lengths at least as large as $\approx nR(D) + (1/2) \log n$ bits. Similarly, in [21] it was shown that no code operating at distortion level D can outperform $\ell_n^*(X_1^n)$ by more than $2 \log n$ bits, eventually with probability one, and it was also shown that $\ell_n^*(X_1^n)$ is approximately competitively optimal (see Corollary 2 and Proposition 3 in [21]).

Therefore, we interpret (23) as saying that the pointwise price paid for universality by a mixture codebook is approximately "$(1/2) \log n$ bits per degree of freedom." The converse recently proved in [28] indicates that this rate is asymptotically optimal.

(b) *Conditions for Universality*: The results of Theorem 7 and Corollary 1 only hold for sources P in $\mathcal{S}(D)$. The only essential restriction implied by this assumption is that the optimal reproduction distribution at distortion level D is unique and has full support. Although this is, of course, not always the case, it is typically true for all low enough distortion levels, and it *is* true in several important special cases. For example, it is easily seen that this assumption is satisfied in the case of binary sources with respect to Hamming distortion; cf. [21, Ex. 2]. More generally, Yang and Zhang in [30][28] give the following sufficient condition:

$$\text{The matrix} \quad \left(e^{\lambda \rho(a, \hat{a})}\right)_{a \in A,\ \hat{a} \in \hat{A}} \quad \text{is of full column rank, for all } \lambda < 0. \tag{$*$}$$

In [30, Appendix A] it is shown that for any source P satisfying $(*)$, the optimal reproduction distribution Q^* is unique for all $D \in (0, D_{\max})$. [Note that the direct coding theorems in [30][28] are stated for sources P that satisfy $(*)$ with Q^* of full support; these conditions are apparently stronger than requiring $P \in \mathcal{S}(D)$.]

(c) *Pointwise Redundancy and Minimal Coding Variance*: As discussed in (a) above, the codes exhibited in Corollary 1 have code-lengths that do not exceed the optimum $\ell_n^*(X_1^n)$ bits by more than $O(\log n)$ bits; cf. [21, Corollary 1]. Therefore, the mixture codebooks not only achieve the best rate, but their second-order performance is also optimal, in that their fluctuations *universally* achieve the "minimal coding variance" [21, p.139] of the source that is being compressed.

3 Arbitrary Sources: The Codes-Measures Correspondence

Before giving the proofs of Theorems 1-4, we state some preliminary lemmas that will be needed in the proofs of the direct coding theorems. The first lemma (given here without proof) describes a specific prefix-free code for the positive integers. It is based on iterating a simple idea of Elias [11].

Lemma 2. A Code for the Positive Integers: There exists a prefix-free code C for the positive integers with associated length function L, such that, for all $k \geq 1$

$$L(k) \leq \log k + \log^+ \log k + 2 \log^+ \log^+ \log k + \gamma$$

where $\log^+ x$ denotes the function $\log \max\{x, 1\}$, and γ is some finite constant.

All our direct coding theorems will be proved using random coding arguments based on random codebooks as described in Section 2.2. In the notation of that section, the next lemma establishes a precise relationship between the waiting time W_n for a D-close match in the codebook, and the probability $q_n = Q_n(B(X_1^n, D))$ of finding such a match.

Lemma 3. Waiting Times Bounds: In the notation of Section 2.2:

(i) If for some $x_1^n \in A^n$ the probability $q_n \stackrel{\triangle}{=} Q_n(B(x_1^n, D))$ is nonzero, then for any $\epsilon > 0$:

$$\Pr\{\log[(W_n - 1)q_n] \geq \epsilon \,|\, X_1^n = x_1^n\} \leq e^{-2^\epsilon}.$$

(ii) There is a universal constant $K < \infty$ such that

$$E\{\log[W_n Q_n(B(X_1^n, D))]\} \leq K, \quad \text{for all } n \geq 1,$$

where for each n the expectation is taken over the message X_1^n as well as over the random codebook.

Proof of Lemma 3: Fix x_1^n such that $q_n = Q_n(B(x_1^n, D)) > 0$. Then, conditional on $X_1^n = x_1^n$, the distribution of W_n is geometric with parameter q_n and

$$\Pr\{\log[(W_n - 1)q_n] \geq \epsilon \,|\, X_1^n = x_1^n\} = \Pr\{W_n \geq 2^\epsilon/q_n + 1 \,|\, X_1^n = x_1^n\} \leq (1 - q_n)^{2^\epsilon/q_n} \leq e^{-2^\epsilon}$$

where the last step follows from the simple inequality $(1-x)^{1/x} \leq e^{-1}$, for $x \in (0,1]$. This proves (i). For (ii) let G_n denote the event

$$G_n = \{W_n \neq 1 \text{ and } Q_n(B(X_1^n, D)) > 0\}$$

and write q_n for the (random) probability $Q_n(B(X_1^n, D))$. Then we can bound

$$\begin{aligned}
E\{\log(W_n q_n)\} &\leq E\{[\log(W_n q_n)]\mathbb{I}_{G_n}\} \\
&\leq E\{[\log((W_n - 1)q_n)]\mathbb{I}_{G_n}\} + 1 \\
&\leq \sum_{j \geq 0} \Pr\{[\log((W_n - 1)q_n)]\mathbb{I}_{G_n} \geq j\} + 1 \\
&\overset{(a)}{=} E_{\mathbb{P}}\left[\mathbb{I}_{G_n} \sum_{j \geq 0} \Pr\{\log((W_n - 1)q_n) \geq j \,|\, X_1^\infty\}\right] + 1 \\
&\overset{(b)}{\leq} \sum_{j \geq 0} e^{-2^j} + 1 \stackrel{\triangle}{=} K
\end{aligned}$$

where (a) follows by Fubini's theorem, (b) follows from part (i) of the lemma, and where for any event G, by $\mathbb{I}_G$ we denote its indicator function. □

3.1 Proof of Theorem 1

For part (i), given a code $C_n = (B_n, \phi_n, \psi_n, \ell_n)$, let $L_n : B_n \to \mathbb{N}$ be the length function of the uniquely decodable map ψ_n so that $\ell_n(x_1^n) = L_n(\phi_n(x_1^n))$. Define

$$Q_n(y_1^n) \stackrel{\triangle}{=} \begin{cases} 2^{-L_n(y_1^n)} & \text{if } y_1^n \in B_n \\ 0 & \text{otherwise.} \end{cases}$$

Then for any $x_1^n \in A^n$

$$\ell_n(x_1^n) = L_n(\phi_n(x_1^n)) = -\log Q_n(\phi_n(x_1^n)) \geq -\log Q_n(B(x_1^n, D)) \quad \text{bits}$$

where the inequality follows from the fact that C_n operates at distortion level D. Since ψ_n is uniquely decodable, Kraft's inequality implies that Q_n is a sub-probability measure. If it is a probability measure we are done; otherwise the above argument can be repeated with the measure $Q_n' = Z^{-1} Q_n$ in place of Q_n, where Z is the normalizing constant

$$Z = \sum_{y_1^n \in B_n} 2^{-L_n(y_1^n)} \leq 1.$$

The direct coding theorems of parts (ii) and (iii) are based on a random coding argument. Given a sequence of measures $\{Q_n\}$, for each $n \geq 1$ generate a random codebook $\{Y_1^n(i)\ ;\ i \geq 1\}$ as described in Section 2.2. These codebooks are assumed to be available both to the encoder and decoder, and the following coding scheme is adopted: If the waiting time W_n is finite, then X_1^n is described (with distortion D or less) by describing W_n to the decoder, using the code from Lemma 2. If $W_n = \infty$, then we describe X_1^n using the quantizer $q^{(n)}$ provided by the condition (WQC). Writing μ_n for the distribution of $q^{(n)}(X_1^n)$ on B_n, this description can be given using at most

$$\lceil -\log \mu_n(q^{(n)}(X_1^n))\rceil \quad \text{bits.} \tag{24}$$

Finally a one-bit flag is added to tell the decoder which of the two cases ($W_n < \infty$ or $W_n = \infty$) occurred. This code clearly operates at distortion level D, and its overall description length ℓ_n has:

$$\ell_n(X_1^n) \;\leq\; \begin{cases} \log W_n + \log^+ \log W_n + 2\log^+ \log^+ \log W_n + \gamma + 1 & \text{if } W_n < \infty \\ \lceil -\log \mu_n(q^{(n)}(X_1^n))\rceil + 1 & \text{if } W_n = \infty. \end{cases}$$

For (ii), since the sequence $\{Q_n\}$ is admissible, for $\mathbb{P}$-almost every realization of the source the probability $Q_n(B(X_1^n, D)) \geq 2^{-nR} > 0$ eventually, and therefore the waiting times W_n will be finite eventually with probability 1 (with respect to both the source and the codebook distribution). Therefore,

$$\ell_n(X_1^n) \;\leq\; \log W_n + \log^+ \log W_n + 2\log^+ \log^+ \log W_n + \gamma + 1 \quad \text{eventually, w.p.1.} \tag{25}$$

Write, as before, q_n for the (random) probability $Q_n(B(X_1^n, D))$. Then for $\mathbb{P}$-almost every realization of the source, for n large enough we can apply Lemma 3 (i) with $\epsilon = \log[2\log n]$ to get

$$\Pr\{\log[(W_n - 1)q_n] \geq \log[2\log n] \,|\, X_1^n = x_1^n\} \leq e^{-2\log n} \leq n^{-2}.$$

The Borel-Cantelli lemma implies that

$$\log[(W_n - 1)q_n] \leq \log\log n + 1 \quad \text{eventually, w.p.1}$$

and hence

$$\log[W_n q_n] \leq \log\log n + 2 \quad \text{eventually, w.p.1.} \tag{26}$$

From the admissibility of $\{Q_n\}$ and (26) we also have,

$$\log W_n \leq 2nR \quad \text{eventually, w.p.1} \tag{27}$$

and combining (25) (26) and (27),

$$\begin{aligned} \ell_n(X_1^n) \;&\leq\; -\log q_n + \log\log n + \log(2nR) + 2\log\log(2nR) + \gamma + 3 \\ &\leq\; -\log q_n + \log n + 3\log\log n + [\gamma + 3\log R + 6] \quad \text{eventually, w.p.1.} \end{aligned}$$

This proves (ii) with the constant term equal to $(\gamma + 3\log R + 6)$.

Turning to part (iii) we note that the assumption that $\{Q_n\}$ is admissible in expectation implies that for all $\epsilon > 0$ sufficiently small there is a finite (nonrandom) N such that, for all $n \geq N$,

$$\begin{aligned} q_n = Q_n(B(X_1^n, D)) \;&>\; 0 \quad \text{w.p.1} \\ \text{and } \; E[-\log Q_n(B(X_1^n, D))] \;&\leq\; 2n(R - \epsilon). \end{aligned} \tag{28}$$

In particular, the bound (25) holds with probability one for all $n \geq N$, and therefore

$$E[\ell_n(X_1^n)] \;\leq\; E[\log W_n] + E[\log^+ \log W_n] + 2E[\log^+ \log^+ \log W_n] + \gamma + 1, \quad \text{for } n \geq N.$$

Replacing all the $(\log W_n)$ terms above by $[\log(W_n q_n) - \log q_n]$, using Lemma 3 (ii), and applying Jensen's inequality, implies that for all $n \geq N$:

$$E[\ell_n(X_1^n)] \;\leq\; E[-\log q_n] + \log[K + E(-\log q_n)] + 2\log\log[K + E(-\log q_n)] + \gamma + K + 1.$$

Finally using the bound (28) yields, for all $n \geq N$ large enough,

$$E[\ell_n(X_1^n)] \;\leq\; E[-\log q_n] + \log n + 2\log\log n + [\gamma + K + 3\log R + 4].$$

This proves (iii) with the constant term equal to $(\gamma + K + 3\log R + 4)$ and concludes the proof. □

Remark: The proof of the direct coding theorem in part (ii) above establishes that the bound claimed in the statement of the theorem holds with probability one with respect to both the source and the random codebook distribution. In other words, not only do there exist deterministic codes as claimed, but (almost) any realization of the random codes constructed will provide such a code.

3.2 Proof of Theorem 2

The coding scheme used here is different from the one in Theorem 1. Let

$$\Delta_n \stackrel{\triangle}{=} -\log Q_n(B(X_1^n, D)) - nR,$$

so that $Z_n = |\Delta_n|$, and define the events

$$\begin{aligned} H_n &\stackrel{\triangle}{=} \{Z_n \leq B\sqrt{n}\log n\} \\ \text{and}\quad J_n &\stackrel{\triangle}{=} \{\log W_n \leq nR + \Delta_n + \log\log n + 2\}. \end{aligned}$$

If H_n and J_n both occur, then W_n is described in two steps: First we describe $\lceil \Delta_n \rceil$ using no more than

$$\lceil \log(2B\sqrt{n}\log n + 1) \rceil \;\leq\; \frac{1}{2}\log n + \log\log n + \log B + 3 \quad \text{bits} \tag{29}$$

and then describe W_n itself using no more than

$$nR + \Delta_n + \log\log n + 3 = -\log Q_n(B(X_1^n, D)) + \log\log n + 3 \quad \text{bits}. \tag{30}$$

If either H_n or J_n fails, then X_1^n is described without coding, using the quantizer $q^{(n)}$ provided by (WQC). After adding a one-bit flag to indicate which of the two methods was used, from (24), (29) and (30) the overall description length ℓ_n of this code has:

$$\ell_n(X_1^n) \;\leq\; \begin{cases} -\log Q_n(B(X_1^n, D)) + \frac{1}{2}\log n + 2\log\log n + (\log B + 7) & \text{if } H_n \text{ and } J_n \text{ both hold} \\ \lceil -\log \mu_n(q^{(n)}(X_1^n)) \rceil + 1 & \text{otherwise.} \end{cases} \tag{31}$$

For part (i), assumption (11) and the admissibility of $\{Q_n\}$ imply that both H_n and J_n hold eventually with probability 1 (see the derivation of (26) above), thereby proving (i) with the constant term being equal to $(\log B + 7)$.

For part (ii) we first need to bound the probability that J_n fails. With $q_n = Q_n(B(X_1^n, D))$, since $\{Q_n\}$ is admissible in expectation, for all n large enough $q_n > 0$ with probability one, and

$$\begin{aligned}\Pr\{J_n \text{ fails}\} &= \Pr\{\log[W_n q_n] > \log(4\log n)\} \\ &\leq \Pr\{(W_n - 1)q_n > 2\log n\} \stackrel{(a)}{\leq} \frac{1}{n^2}\end{aligned}$$

where (a) follows by Lemma 3 (i). Therefore, letting F_n denote the event that either H_n or J_n fails, for n large enough we have

$$\Pr\{F_n\} \leq \frac{C+1}{n^q}. \tag{32}$$

Taking expectations of both sides of (31) we get

$$\begin{aligned}E[\ell_n(X_1^n)] \leq E[-\log Q_n(B(X_1^n, D))] &+ \frac{1}{2}\log n + 2\log\log n + (\log B + 7) \\ &+ E\{\mathbb{I}_{F_n}[-\log \mu_n(q^{(n)}(X_1^n))]\} + 2\Pr\{F_n\}.\end{aligned}$$

Applying Hölder's inequality to the last expectation above, using bound (32), and recalling the bound from the condition (pSQC), yields

$$E\{\mathbb{I}_{F_n}[-\log \mu_n(q^{(n)}(X_1^n))]\} + 2\Pr\{F_n\} \leq M_p(C+1)^{1/q} + 1, \quad \text{eventually.}$$

This completes the proof of part (ii) with the constant term being $M_p(C+1)^{1/q} + \log B + 8$. □

3.3 Proofs of Theorems 3 and 4

Lemma 1 immediately implies that $K_n(D) \geq R_n(D)$. Given a code (C_n, ℓ_n), define a probability measure Q_n' as in the proof of Theorem 1, and notice that $E[\ell_n(X_1^n)] \geq E[-\log Q_n'(B(X_1^n, D))] \geq K_n(D)$. Part (ii) of Theorem 3 and part (i) of Theorem 4 follow from the Kuhn-Tucker conditions given in [3] exactly as in the proof of Theorem 6 in [21]. Finally Theorem 4 (ii) is an immediate consequence of parts (ii) and (iii) of Theorem 1. □

4 Memoryless Sources: Mixture Codebooks and Redundancy

In the notation of Section 2.5, assume that $\{X_n\}$ is a stationary memoryless source with distribution P on A, and write $\mathbb{P}$ for the distribution of the entire process. Let π be a prior distribution on the simplex $\mathcal{P}$ of probability measures on $\hat{A}$. For each $n \geq 1$, the mixture distributions M_n on $\hat{A}^n$ are defined as in (6).

4.1 Proof of Theorem 6

It suffices to show that for every $\epsilon > 0$

$$\frac{1}{n}\log\frac{Q_n^*(B(X_1^n,D))}{M_n(B(X_1^n,D))} \leq \epsilon \quad \text{eventually w.p.1}$$

or, equivalently,

$$\frac{2^{n\epsilon}M_n(B(X_1^n,D))}{Q_n^*(B(X_1^n,D))} \geq 1, \quad \text{eventually w.p.1.} \tag{33}$$

For $\epsilon > 0$ define the neighborhoods $N(Q^*,\epsilon)$ as in (20), and assume that $\pi\{N(Q^*,\epsilon)\} > 0$ for all $\epsilon > 0$ (we deal with the case of π having a positive density at the end).

Observe that

$$\begin{aligned}
\frac{2^{n\epsilon}M_n(B(X_1^n,D))}{Q_n^*(B(X_1^n,D))} &= 2^{n\epsilon}\int_{Q\in\mathcal{P}}\frac{Q^n(B(X_1^n,D))}{Q_n^*(B(X_1^n,D))}\,d\pi(Q)\\
&\geq 2^{n\epsilon}\int_{N(Q^*,\epsilon)}\frac{Q^n(B(X_1^n,D))}{Q_n^*(B(X_1^n,D))}\,d\pi(Q)\\
&= \int_{N(Q^*,\epsilon)} 2^{n[\epsilon-\frac{1}{n}\log\frac{Q_n^*(B(X_1^n,D))}{Q^n(B(X_1^n,D))}]}\,d\pi(Q).
\end{aligned}\tag{34}$$

From (17) we know that for every Q

$$r_n(X_1^n) \stackrel{\triangle}{=} \frac{1}{n}\log\frac{Q_n^*(B(X_1^n,D))}{Q^n(B(X_1^n,D))} \to R(P,Q,D) - R(D) \quad \text{with } \mathbb{P}\text{-prob.1}$$

therefore, by Fubini's theorem, for $\mathbb{P}$-almost every realization x_1^∞ of the source and for π-almost every $Q \in N(Q^*,\epsilon)$,

$$r_n(x_1^n) \to R(P,Q,D) - R(D) < \epsilon. \tag{35}$$

Fix any one of the (almost all) x_1^∞ such that (35) holds. By Fatou's lemma and the fact that $\pi\{N(Q^*,\epsilon)\} > 0$ it follows that

$$\liminf_{n\to\infty}\int_{N(Q^*,\epsilon)} 2^{n(\epsilon-r_n(x_1^n))}\,d\pi(Q) \geq \int_{N(Q^*,\epsilon)}\liminf_{n\to\infty} 2^{n(\epsilon-r_n(x_1^n))}\,d\pi(Q) = \infty,$$

and combining this with (34) implies (33) as required.

Finally we claim that, if π has a density $p(Q)$ with respect to Lebesgue measure on $\mathcal{P}$ and $p(Q)$ is strictly positive in a neighborhood of Q^*, then $\pi\{N(Q^*,\epsilon)\} > 0$ for all $\epsilon > 0$ small enough. Note that, since $p(Q) > 0$ in a neighborhood of Q^*, it suffices to show that the neighborhoods $N(Q^*,\epsilon)$ have positive Lebesgue measure. Recall that by the representation of $R(P,Q,D)$ in [21, Proposition 2], $R(P,Q,D)$ is convex as a function of Q. Since $R(P,Q^*,D) = R(D) < \infty$, by the definition of $R(P,Q,D)$ it follows that it is finite for all Q with support at least as large as the support of Q^*. Let $\mathcal{M}(Q^*)$ denote the set of all such Q:

$$\mathcal{M}(Q^*) \stackrel{\triangle}{=} \{Q\in\mathcal{P} \;:\; \text{support}(Q) \supseteq \text{support}(Q^*)\}.$$

Since $\mathcal{M}(Q^*)$ is locally simplicial, $R(P,Q,D)$ is upper-semicontinuous on $\mathcal{M}(Q^*)$ (cf. [24, Theorem 10.2]). Therefore the neighborhoods $N(Q^*,\epsilon)$ contain nonempty open sets and hence have positive Lebesgue measure. This proves the claim and completes the proof of the theorem. $\square$

4.2 Technical Properties

Following the notation of Section 2.5, we let $\mathcal{P}$ denote the simplex of probability distributions on $\hat{A}$. Let $\hat{A} = \{\hat{a}_0, \hat{a}_1, \ldots, \hat{a}_d\}$. For the sake of rigor, we need to introduce a simple parametrization of $\mathcal{P}$. We identify $\mathcal{P}$ with the d-dimensional subset of $\mathbb{R}^d$

$$\Theta = \left\{ \theta = (\theta_1, \theta_2, \ldots, \theta_d) \in [0,1]^d \;:\; \sum_{i=1}^{d} \theta_i \leq 1 \right\}$$

via the 1-1 correspondence $Q \in \mathcal{P} \leftrightarrow \theta \in \Theta$:

$$\begin{aligned} Q(\hat{a}_i) &= \theta_i, \quad i = 1, 2, \ldots, d \\ Q(\hat{a}_0) &= 1 - \sum_{i=1}^{d} \theta_i. \end{aligned}$$

We often write Q_θ for the distribution in $\mathcal{P}$ corresponding to $\theta \in \Theta$.

Let π be a prior distribution on Θ (or, equivalently, on $\mathcal{P}$). From now on we also assume that π has a strictly positive, continuous density

$$p(\theta) = \frac{d\pi(\theta)}{d\theta}, \quad \theta \in \Theta$$

with respect to Lebesgue measure on Θ.

Throughout this section we fix a $D > 0$ and a source distribution $P \in \mathcal{S}(D)$ (recall (21)). Following [21] we define, for any $Q \in \mathcal{P}$:

$$\begin{aligned} D_{\min}^{P,Q} &\triangleq E_P\left[\min_{y \,\in\, \text{support}(Q)} \rho(X, y) \right] \\ D_{\max}^{P,Q} &\triangleq E_{P\times Q}\left[\rho(X,Y)\right] \end{aligned}$$

and for all $\lambda < 0$ and $x \in A$,

$$\begin{aligned} \Lambda_{P,Q}(\lambda) &\triangleq E_P\left\{ \log_e E_Q\left[e^{\lambda\rho(X,Y)} \right] \right\} \\ \Lambda_x(Q,\lambda) &\triangleq \log_e E_Q\left[e^{\lambda\rho(x,Y)} \right]. \end{aligned}$$

By [21, Lemma 1], for any $Q = Q_\theta \in \mathcal{P}$ and any $D \in (D_{\min}^{P,Q_\theta}, D_{\max}^{P,Q_\theta})$ there exists a unique $\lambda = \lambda_\theta < 0$ such that

$$\Lambda'_{P,Q}(\lambda_\theta) \triangleq \frac{\partial}{\partial\lambda}[\Lambda_{P,Q}(\lambda)]\Big|_{\lambda=\lambda_\theta} = D.$$

From [21, Lemma 1 and Proposition 1] we have

$$R_e(P, Q_\theta, D) = \lambda_\theta D - \Lambda_{P,Q}(\lambda_\theta) \tag{36}$$

where $R_e(P,Q,D)$ is the same as $R(P,Q,D)$ but in nats rather than bits,

$$R_e(P,Q,D) \triangleq (\log_e 2) R(P,Q,D).$$

Now let θ^* correspond to the optimal reproduction distribution $Q^* = Q_{\theta^*}$ for the source distribution P at distortion level D. Recall [21, Proposition 2] that whenever $D \in (0, D_{\max})$, we also have $D \in (D_{\min}^{P,Q^*}, D_{\max}^{P,Q^*})$. Therefore, writing $\lambda^* = \lambda_{\theta^*}$,

$$R_e(D) = R_e(P, Q^*, D) = \lambda^* D - \Lambda_{P,Q^*}(\lambda^*) \tag{37}$$

where $R_e(D) = (\log_e 2)R(D)$ denotes the rate-distortion function in nats.

Lemma 4. Differentiability Properties: Let $D > 0$ and $P \in \mathcal{S}(D)$. Write $Q^* = Q_{\theta^*}$ for the optimal reproduction distribution for P at distortion level D.

(i) For each $x \in A$, the function $\Lambda_x(Q_\theta, \lambda_\theta)$ is twice differentiable in θ on a neighborhood of θ^*.

(ii) $R_e(P, Q_\theta, D)$ is twice differentiable in θ on a neighborhood of θ^*.

Proof of Lemma 4: For $\lambda < 0$ and θ in the interior of Θ (corresponding to Q_θ in the interior of the simplex $\mathcal{P}$), define

$$\Psi_1(\theta, \lambda) = \frac{\partial}{\partial \lambda}[\Lambda_{P,Q_\theta}(\lambda)] - D.$$

By the definitions of λ_θ and λ^* we have $\Psi_1(\theta, \lambda_\theta) = \Psi_1(\theta^*, \lambda^*) = 0$. It is easy to see that Ψ_1 is twice differentiable in either of its arguments, and, moreover, the derivative

$$\frac{\partial \Psi_1(\theta, \lambda)}{\partial \lambda} = \Lambda''_{P,Q_\theta}(\lambda)$$

exists and is strictly positive as long as $D_{\min}^{P,Q_\theta}$ is strictly smaller than $D_{\max}^{P,Q_\theta}$ (see [21, Lemma 1]). But since P and Q_θ both have full support, our assumption that $D_{\max} > 0$ implies that this is always the case. Thus, by the implicit function theorem, λ_θ is differentiable in θ on a neighborhood of θ^*, and

$$\frac{\partial \lambda_\theta}{\partial \theta_i} = -\left[\Lambda''_{P,Q_\theta}(\lambda_\theta)\right]^{-1} \frac{\partial \Psi_1(\theta, \lambda)}{\partial \theta_i}\bigg|_{\lambda = \lambda_\theta}.$$

From this expression and from the differentiability of Ψ_1 it immediately follows that λ_θ is in fact twice differentiable in θ on a neighborhood of θ^*, and since

$$\Lambda_x(Q_\theta, \lambda_\theta) = \log_e \sum_{y \in \hat{A}} \left[Q_\theta(y) e^{\lambda_\theta \rho(x,Y)}\right]$$

it follows that, for any fixed $x \in A$, $\Lambda_x(Q_\theta, \lambda_\theta)$ is also twice differentiable in θ in a neighborhood of θ^*. This proves (i).

Recall the expression for $R_e(P, Q, D)$ in (36) and define, for $R > 0$ and θ in the interior of Θ,

$$\Psi_2(\theta, R) = R - \lambda_\theta D + \Lambda_{P,Q_\theta}(\lambda_\theta).$$

By the discussion preceding the statement of the lemma, $\Psi_2(\theta^*, R_e(D)) = \Psi_2(\theta, R_e(P, Q_\theta, D)) = 0$. Also, by the arguments above, Ψ_2 is twice differentiable in either of its arguments and

$$\frac{\partial \Psi_2(\theta, R)}{\partial R} = 1 \neq 0.$$

Therefore, by the implicit function theorem $R_e(P, Q_\theta, D)$ is differentiable in θ on a neighborhood of θ^* and

$$\frac{\partial R_e(P, Q_\theta, D)}{\partial \theta_i} = -\frac{\partial \Psi_2(\theta, R)}{\partial \theta_i}\Big|_{R=R(P,Q_\theta,D)}.$$

This, together with the definition $\Psi_2(\theta, R)$ and the differentiability of λ_θ proved above, imply that $R_e(P, Q_\theta, D)$ is twice differentiable, proving (ii). $\square$

4.3 Proof of Theorem 7

The outline of the proof is similar to that of the corresponding lossless result in [6] and it heavily relies on the precise asymptotics for $Q_n^*(B(X_1^n, D))$ developed in [9] and [29], so our notation follows closely the notation in [6][9][29].

Let $D > 0$ be given, and choose and fix a source distribution $P \in \mathcal{S}(D)$ with a corresponding optimal reproduction distribution $Q^* = Q_{\theta^*}$. According to Lemma 4 we can define

$$S_n \stackrel{\triangle}{=} S_n(\theta^*, \lambda^*; X_1^n) \stackrel{\triangle}{=} \frac{1}{\sqrt{n}} \sum_{i=1}^n \nabla_\theta [\Lambda_{X_i}(Q_\theta, \lambda_\theta) - \Lambda_{P,Q_\theta}(\lambda_\theta)]\Big|_{\theta=\theta^*}$$

where $\lambda^* = \lambda_{\theta^*}$ is chosen so that (37) holds. Similarly, Lemma 4 guarantees the existence of the matrix of partial derivatives

$$J \stackrel{\triangle}{=} \frac{\partial^2}{\partial \theta^2} R_e(P, Q_\theta, D)\Big|_{\theta=\theta^*}.$$

Note that, since $P \in \mathcal{S}(D)$, J is positive definite and hence invertible.

Since S_n is a (normalized) partial sum of zero-mean, independent random vectors, the law of the iterated logarithm implies that each of its components is of order $O(\sqrt{\log_e \log_e n})$ with probability one. Therefore, the quadratic form

$$S_n^t J^{-1} S_n \;=\; O(\log_e \log_e n) \quad \text{w.p.1.} \tag{38}$$

Choosing a $d_0 > d$, we define a "perturbed" version of θ^* as

$$\hat{\theta} = \theta^* + \frac{1}{\sqrt{n}} J^{-1} S_n \cdot \mathbb{I}_{\{S_n^t J^{-1} S_n \le d_0\}}.$$

Let $\delta_n \stackrel{\triangle}{=} \sqrt{\frac{d_0}{n}}$, $n \ge 1$, and define a sequence of neighborhoods centered at $\hat{\theta}$

$$\hat{N}_{\delta_n} \stackrel{\triangle}{=} \{\theta \in \Theta : \|\theta - \hat{\theta}\| \le \delta_n\},$$

where, throughout the proof, $\|\cdot\|$ denotes the L^2-norm with respect to the matrix J:

$$\|\xi\| \stackrel{\triangle}{=} \sqrt{\xi^t J \xi}, \quad \xi \in \mathbb{R}^d.$$

Similarly define the neighborhoods $N_{\delta_n} = \{\theta \in \Theta : \|\theta - \theta^*\| \le \delta_n\}$, and note that

$$\hat{N}_{\delta_n} \subset N_{2\delta_n}. \tag{39}$$

Next we obtain an upper bound on the quantity $-\frac{n}{2}\|\theta-\hat{\theta}\|^2$. Using the definition of $\hat{\theta}$, we expand $\|\theta-\hat{\theta}\|^2$ as:

$$\|\theta-\theta^*\|^2+\frac{1}{n}S_n^tJ^{-1}S_n-\frac{2}{\sqrt{n}}(\theta-\theta^*)^tS_n-\left\{\frac{1}{n}S_n^tJ^{-1}S_n-\frac{2}{\sqrt{n}}(\theta-\theta^*)^tS_n\right\}\mathbb{I}_{\{S_n^tJ^{-1}S_n>d_0\}}. \qquad (40)$$

Since on the event $\{S_n^tJ^{-1}S_n>d_0\}$ we have $\hat{\theta}=\theta^*$, the Cauchy-Schwarz inequality implies that

$$\sqrt{n}(\theta^*-\theta)^tS_n\mathbb{I}_{\{S_n^tJ^{-1}S_n>d_0\}}\leq S_n^tJ^{-1}S_n, \quad \text{for } \theta\in\hat{N}_{\delta_n}. \qquad (41)$$

Combining (40) and (41) we have that, for all $\theta\in\hat{N}_{\delta_n}$,

$$-\frac{n}{2}\|\theta-\hat{\theta}\|^2 \quad\leq\quad -\frac{n}{2}\|\theta-\theta^*\|^2+\sqrt{n}(\theta-\theta^*)^tS_n\;+S_n^tJ^{-1}S_n. \qquad (42)$$

Now we are ready to examine the term $\log_e\left[\frac{Q_n^*(B(X_1^n,D))}{M_n(B(X_1^n,D))}\right]$. Let $\phi_n(\theta)$, $n\geq 1$, denote the truncated normal density functions

$$\phi_n(\theta)=\frac{1}{c_n}e^{-(n/2)\|\theta-\hat{\theta}\|^2}\mathbb{I}_{\{\theta\in\hat{N}_{\delta_n}\}}, \quad \theta\in\Theta$$

where c_n is the normalizing constant

$$c_n=\int_{\hat{N}_{\delta_n}}e^{-(n/2)\|\theta-\hat{\theta}\|^2}\,d\theta.$$

Writing $Q_n^*=(Q^*)^n$ and $Q_\theta^n=(Q_\theta)^n$, with probability one we have

$$\begin{aligned}
&\log_e\frac{Q_n^*(B(X_1^n,D))p(\theta^*)}{M_n(B(X_1^n,D))}\\
&=\quad -\log_e\int_\Theta\frac{Q_\theta^n(B(X_1^n,D))p(\theta)}{Q_n^*(B(X_1^n,D))p(\theta^*)}\,d\theta\\
&\leq\quad -\log_e\int_{\hat{N}_{\delta_n}}\frac{Q_\theta^n(B(X_1^n,D))p(\theta)}{Q_n^*(B(X_1^n,D))p(\theta^*)}\,d\theta\\
&=\quad -\log_e\int_{\hat{N}_{\delta_n}}\frac{Q_\theta^n(B(X_1^n,D))}{Q_n^*(B(X_1^n,D))}\frac{p(\theta)}{p(\theta^*)}c_ne^{(n/2)\|\theta-\hat{\theta}\|^2}\phi_n(\theta)\,d\theta\\
&\overset{(a)}{\leq}\quad \int_{\hat{N}_{\delta_n}}\left[-\log_e\frac{Q_\theta^n(B(X_1^n,D))}{Q_n^*(B(X_1^n,D))}-\frac{n}{2}\|\theta-\hat{\theta}\|^2-\log_e\frac{p(\theta)}{p(\theta^*)}\right]\phi_n(\theta)\,d\theta-\log_e c_n\\
&\overset{(b)}{\leq}\quad \int_{\hat{N}_{\delta_n}}\left[-\log_e\frac{Q_\theta^n(B(X_1^n,D))}{Q_n^*(B(X_1^n,D))}-\frac{n}{2}\|\theta-\theta^*\|^2+\sqrt{n}(\theta-\theta^*)^tS_n\right]\phi_n(\theta)\,d\theta\\
&\qquad\qquad -\log_e c_n+O(\log_e\log_e n)\\
&\overset{(c)}{\leq}\quad \int_{N_{2\delta_n}}\left|-\log_e\frac{Q_\theta^n(B(X_1^n,D))}{Q_n^*(B(X_1^n,D))}-\frac{n}{2}\|\theta-\theta^*\|^2+\sqrt{n}(\theta-\theta^*)^tS_n\right|e_0\;\phi_n^*(\theta)\,d\theta\\
&\qquad\qquad -\log_e c_n+O(\log_e\log_e n),
\end{aligned} \qquad (43)$$

where $\phi_n^*(\theta)$ is the d-dimensional normal density with mean θ^* and covariance matrix $(nJ)^{-1}$. Above, (a) follows by Jensen's inequality, (b) follows from the continuity of $p(\theta)$, (42) and (38), and (c) follows from (39) and the fact that

$$\phi_n(\theta)\quad\leq\quad e_0\phi_n^*(\theta), \quad \text{for some constant } \; e_0>0,\; \theta\in\hat{N}_{\delta_n}.$$

This inequality is easily derived from the bound after eq. (5.7) in [6, p.467].

Next we consider the integrand in (43). Appealing to [29, Corollary 1], for θ close enough to θ^*,

$$\begin{aligned} -\log_e Q_n^*(B(X_1^n, D)) &= nR_e(\hat{P}_n, Q^*, D) + \frac{1}{2}\log_e n + O(1) \quad \text{w.p.1} \\ \text{and} \quad -\log_e Q_\theta^n(B(X_1^n, D)) &= nR_e(\hat{P}_n, Q_\theta, D) + \frac{1}{2}\log_e n + O(1) \quad \text{w.p.1} \qquad (44) \end{aligned}$$

where $\hat{P}_n$ denotes the empirical distribution induced by X_1^n on A. Moreover, a close examination of the proof of Corollary 1 in [29] reveals that, in our setting (where A and $\hat{A}$ are finite and ρ is bounded), the convergence in (44) is uniform for θ in a small enough compact neighborhood of θ^*. Therefore, uniformly for $\theta \in N_{2\delta_n}$,

$$\log_e \frac{Q_n^*(B(X_1^n, D))}{Q_\theta^n(B(X_1^n, D))} = n\left[R_e(\hat{P}_n, Q_\theta, D) - R_e(\hat{P}_n, Q^*, D)\right] + O(1) \quad \text{w.p.1.} \qquad (45)$$

To expand the above right-hand-side further, following [21] we define

$$f_\theta(x) \quad = \quad -\Lambda_x(Q_\theta, \lambda_\theta) + \Lambda_{P,Q_\theta}(\lambda_\theta), \quad x \in A.$$

Writing the difference $n[R_e(\hat{P}_n, Q_\theta, D) - R_e(\hat{P}_n, Q^*, D)]$ as

$$n[R_e(\hat{P}_n, Q_\theta, D) - R_e(P, Q_\theta, D)] - n[R_e(\hat{P}_n, Q^*, D) - R_e(P, Q^*, D)] + n[R_e(P, Q_\theta, D) - R_e(P, Q^*, D)],$$

we can apply [9, Theorem 3] to get that $n[R_e(\hat{P}_n, Q_\theta, D) - R_e(\hat{P}_n, Q^*, D)]$ is equal to

$$\sum_{i=1}^n [f_\theta(X_i) - f_{\theta^*}(X_i)] + n[R_e(P, Q_\theta, D) - R_e(P, Q^*, D)] + O(\log_e \log_e n) \quad \text{w.p.1.} \qquad (46)$$

[Note that we have actually used the sharper remainder term of order $(\log_e \log_e n)$, as identified in [21, Appendix VI].] Further, a careful look at the proof of [9, Theorem 3] (see also the discussion in the proof of Theorem 18 in [10]) shows the convergence in [9, Theorem 3] is uniform for θ in a small compact neighborhood in θ^* (provided that A and $\hat{A}$ are finite and ρ is bounded). Therefore, convergence in (46) holds uniformly for $\theta \in N_{2\delta_n}$.

Notice that in the above notation S_n can be rewritten as

$$S_n = \frac{1}{\sqrt{n}} \sum_{i=1}^n \nabla_\theta [-f_\theta(X_i)]\Big|_{\theta=\theta^*}.$$

From the Taylor expansion of f_θ around θ^*, and taking the maximum over all possible realizations for $X_1^n = x_1^n \in A^n$, we get

$$\sum_{i=1}^n (f_\theta(X_i) - f_{\theta^*}(X_i)) + \sqrt{n}(\theta - \theta^*)^t S_n \;=\; n\, O(\|\theta - \theta^*\|^2) \quad \text{w.p.1.}$$

Therefore, uniformly for $\theta \in N_{2\delta_n}$,

$$\sum_{i=1}^n (f_\theta(X_i) - f_{\theta^*}(X_i)) + \sqrt{n}(\theta - \theta^*)^t S_n = O(1) \quad \text{w.p.1.} \qquad (47)$$

Combining (46) and (47), uniformly for $\theta \in N_{2\delta_n}$ we can bound

$$\begin{aligned}
&\left| n[R_e(\hat{P}_n, Q_\theta, D) - R_e(\hat{P}_n, Q^*, D)] - \frac{n}{2}\|\theta - \theta^*\|^2 + \sqrt{n}(\theta - \theta^*)^t S_n \right| \\
&\qquad \leq \left| \sum_{i=1}^{n} (f_\theta(X_i) - f_{\theta^*}(X_i)) + \sqrt{n}(\theta - \theta^*)^t S_n \right| \\
&\qquad\qquad + n \left| R_e(P, Q_\theta, D) - R_e(P, Q^*, D) - \frac{1}{2}\|\theta - \theta^*\|^2 \right| + O(\log_e \log_e n) \\
&\qquad = O(\log_e \log_e n) \quad \text{w.p.1} \qquad (48)
\end{aligned}$$

where in the last step we have used the Taylor expansion of $R_e(P, Q_\theta, D)$ around θ^*:

$$R_e(P, Q_\theta, D) - R_e(P, Q^*, D) = \frac{1}{2}\|\theta - \theta^*\|^2 + o\left(\|\theta - \theta^*\|^2\right)$$

(recall Lemma 4 and the definition of the norm $\|\cdot\|$). Substituting the bounds (45) and (48) in (43) we conclude that

$$\log_e \frac{Q_n^*(B(X_1^n, D))}{M_n(B(X_1^n, D))} \leq -\log_e c_n + O(\log_e \log_e n) \quad \text{w.p.1.}$$

Finally (recall that J is positive definite), repeating the exact same argument that leads to [6, eq.(5.3)], by a comparison with a multivariate normal integral we can estimate

$$-\log_e c_n \leq \frac{d}{2} \log_e n + O(1)$$

thereby completing the proof. □

Acknowledgment

We wish to thank Matthew Harrison for his comments on an earlier draft of this paper.

Appendix I: Proof of Lemma 1

The proof that $K_n(D)$ is no smaller than the right-hand-side follows exactly as in the proof of [21, Lemma 4], with the only difference that, if the infimum in the definition of $K_n(D)$ is not achieved, we can assume without loss of generality that $K_n(D) < \infty$ and choose $\widetilde{Q}_n$ with $E[-\log \widetilde{Q}_n(B(X_1^n, D))] \leq K_n(D) + \epsilon$. Repeating the argument in [21, Appendix 5] up to step (d), yields that $K_n(D) \geq \inf I(X_1^n; Y_1^n) - \epsilon$ and letting $\epsilon \downarrow 0$ gives the desired inequality.

To prove the reverse inequality let (X_1^n, Y_1^n) be an arbitrary pair of random vectors as in the infimum in the statement of the lemma, write μ_n for the joint distribution of (X_1^n, Y_1^n), let Q_n denote the Y_1^n-marginal, and assume that $I(X_1^n; Y_1^n) < \infty$. [Note that if no such $(X_1^n, Y_1^n) \sim \mu_n$ exists then the infimum equals $+\infty$ and we are done.] This guarantees the existence of all the Radon-Nikodym derivatives below (in particular recall [12, eq. (5.2.8)]), so that for P_n-almost every x_1^n,

$$\begin{aligned}\int d\mu_n(y_1^n|x_1^n)\log\frac{d\mu_n(y_1^n|x_1^n)}{dQ_n} &\overset{(a)}{=} -\int_{B(x_1^n,D)} d\mu_n(y_1^n|x_1^n)\log\frac{dQ_n}{d\mu_n}(y_1^n|x_1^n)\\ &\overset{(b)}{\geq} -\log Q_n(B(x_1^n,D))\end{aligned}$$

where (a) follows from Fubini's theorem and the assumption that $\rho(X_1^n,Y_1^n)\leq D$ with μ_n-probability one, and (b) follows from Jensen's inequality. Integrating both sides above with respect to P_n yields

$$I(X_1^n;Y_1^n)=H(\mu_n\|P_n\times Q_n)\geq E[-\log Q_n(B(X_1^n,D))]\geq K_n(D)$$

and completes the proof. □

Appendix II: Proof Outline for Theorem 5

We assume that (for all n) the infimum in the definition of $K_n(D)$ is achieved by some probability measure $\widetilde{Q}_n$ on $\hat{A}^n$; as explained in the remarks following Theorems 4 and 5 the general case is similar. The existence of $R(D)$ is well-known [12]. For each $n\geq 1$, let $\mathcal{G}_n$ be the family of functions defined by (13). The quantization condition (QC) implies that $K_n(D)<\infty$ and therefore each $\mathcal{G}_n$ is nonempty. Moreover, in the terminology of [17], it is straightforward to check that each $\mathcal{G}_n$ is log-convex and that the sequence $\{\mathcal{G}_n\}$ is additive. Then Kieffer's generalized ergodic theorem [17] implies (14) with $K(D)$ on the right-hand-side, and also establishes the existence of

$$K(D)=\lim_{n\to\infty}K_n(D)=\inf_{n\geq 1}K_n(D).$$

Before proving the equality $K(D)=R(D)$ we note that Theorem 4 (i) and (14) imply (15), and similarly Theorem 1 (ii) and (14) show the existence of codes achieving (15) with equality. This proves part (ii).

Finally we argue that $K(D)=R(D)$. From their definitions and Lemma 1 it immediately follows that $R(D)\leq K(D)<\infty$. Fatou's lemma applied to (15) implies that for any sequence of codes $\{(C_n,\ell_n)\}$ operating at distortion level D,

$$\liminf_{n\to\infty}\frac{1}{n}E[\ell_n(X_1^n)]\geq K(D). \tag{49}$$

But there are codes operating at distortion level D that achieve the rate-distortion function in expectation: A close examination of the proofs of Theorems 11.4.1 and 11.5.1 in [12] shows that for any $\epsilon>0$ and $\delta>0$ there are fixed-rate codes with asymptotic rate bounded above by $R(D-\delta)+\epsilon$ and with vanishing probability of encoding a source string with distortion greater than D. Therefore, using the quantizers $\{q^{(n)}\}$ provided by the quantization condition (QC) we can modify these codes to operate at distortion level D, with an additional cost ϵ in the rate. This and (49) imply that for all $\epsilon>0$ and $\delta>0$

$$R(D-\delta)+2\epsilon\geq K(D).$$

Since $R(D)$ is continuous when finite, we can let both ϵ and δ go to zero to get that indeed $R(D)\geq K(D)$. □

References

[1] A.R. Barron. *Logically Smooth Density Estimation.* PhD thesis, Dept. of Electrical Engineering, Stanford University, 1985.

[2] A.R. Barron, J. Rissanen, and B. Yu. The minimum description length principle in coding and modeling. (Information theory: 1948–1998). *IEEE Trans. Inform. Theory*, 44(6):2743–2760, 1998.

[3] R. Bell and T.M. Cover. Game-theoretic optimal portfolios. *Management Sci.*, 34(6):724–733, 1988.

[4] T. Berger. *Rate Distortion Theory: A Mathematical Basis for Data Compression.* Prentice-Hall Inc., Englewood Cliffs, NJ, 1971.

[5] P.A. Chou, M. Effros, and R.M. Gray. A vector quantization approach to universal noiseless coding and quantizations. *IEEE Trans. Inform. Theory*, 42(4):1109–1138, 1996.

[6] B.S. Clarke and A.R. Barron. Information-theoretic asymptotics of Bayes methods. *IEEE Trans. Inform. Theory*, 36(3):453–471, 1990.

[7] T.M. Cover and J.A. Thomas. *Elements of Information Theory.* J. Wiley, New York, 1991.

[8] L.D. Davisson. Universal noiseless coding. *IEEE Trans. Inform. Theory*, 19(6):783–795, 1973.

[9] A. Dembo and I. Kontoyiannis. The asymptotics of waiting times between stationary processes, allowing distortion. *Ann. Appl. Probab.*, 9:413–429, 1999.

[10] A. Dembo and I. Kontoyiannis. Source coding, large deviations, and approximate pattern matching. *To appear, IEEE Trans. Inform. Theory*, June 2002. [Available from www.dam.brown.edu/people/yiannis].

[11] P. Elias. Universal codeword sets and representations of the integers. *IEEE Trans. Inform. Theory*, 21:194–203, 1975.

[12] R.M. Gray. *Source Coding Theory.* Kluwer Academic Publishers, Boston, 1990.

[13] T.S. Han. An information-spectrum approach to source coding theorems with a fidelity criterion. *IEEE Trans. Inform. Theory*, 43(4):1145–1164, 1997.

[14] T.S. Han. *Information-Spectrum Methods in Information Theory.* Baifukan-Press, Tokyo, Japan, 1998. (in Japanese).

[15] D. Ishii and H. Yamamoto. The redundancy of universal coding with a fidelity criterion. *IEICE Trans. Fundamentals*, E80-A:2225–2231, 1997.

[16] F. Kanaya and K. Nakagawa. On the practical implication of mutual information for statistical decisionmaking. *IEEE Trans. Inform. Theory*, 37(4):1151–1156, 1991.

[17] J.C. Kieffer. An almost sure convergence theorem for sequences of random variables selected from log-convex sets. In *Almost everywhere convergence, II (Evanston, IL, 1989)*, pages 151–166. Academic Press, Boston, MA, 1991.

[18] J.C. Kieffer. Sample converses in source coding theory. *IEEE Trans. Inform. Theory*, 37(2):263–268, 1991.

[19] I. Kontoyiannis. An implementable lossy version of the Lempel-Ziv algorithm – Part I: Optimality for memoryless sources. *IEEE Trans. Inform. Theory*, 45(7):2293–2305, November 1999.

[20] I. Kontoyiannis. Model selection via rate-distortion theory. In *Proceedings of the 34th Annual Conference on Information Sciences and Systems, CISS 2000*, pages WP8–7 – WP8–11, Princeton, NJ, March 2000.

[21] I. Kontoyiannis. Pointwise redundancy in lossy data compression and universal lossy data compression. *IEEE Trans. Inform. Theory*, 46(1):136–152, January 2000.

[22] T. Łuczak and W. Szpankowski. A suboptimal lossy data compression algorithm based on approximate pattern matching. *IEEE Trans. Inform. Theory*, 43(5):1439–1451, 1997.

[23] J. Rissanen. Stochastic complexity. *J. Roy. Statist. Soc. Ser. B*, 49(3):223–239, 253–265, 1987. With discussion.

[24] R.T. Rockafellar. *Convex Analysis.* Princeton University Press, Princeton, NJ, 1997. Reprint of the 1970 original, Princeton Paperbacks.

[25] C.E. Shannon. Coding theorems for a discrete source with a fidelity criterion. *IRE Nat. Conv. Rec.*, part 4:142–163, 1959. Reprinted in D. Slepian (ed.), *Key Papers in the Development of Information Theory*, IEEE Press, 1974.

[26] Y. Steinberg and S. Verdú. Simulation of random processes and rate-distortion theory. *IEEE Trans. Inform. Theory*, 42(1):63–86, 1996.

[27] E.-h. Yang and J.C. Kieffer. On the performance of data compression algorithms based upon string matching. *IEEE Trans. Inform. Theory*, 44(1):47–65, 1998.

[28] 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.*

[29] E.-h. Yang and Z. Zhang. On the redundancy of lossy source coding with abstract alphabets. *IEEE Trans. Inform. Theory*, 45(4):1092–1110, 1999.

[30] E.-h. Yang and Z. Zhang. The redundancy of source coding with a fidelity criterion – Part II: Coding at a fixed rate level with unknown statistics. *IEEE Trans. Inform. Theory*, 47(1):126–145, 2001.

[31] B. Yu and T.P. Speed. A rate of convergence result for a universal D-semifaithful code. *IEEE Trans. Inform. Theory*, 39(3):813–820, 1993.

[32] A. Yuan and B.S. Clarke. An information criterion for likelihood selection. *IEEE Trans. Inform. Theory*, 45(2):562–571, 1999.

[33] R. Zamir and K. Rose. A type generator model for adaptive lossy compression. In *Proc. of the IEEE International Symposium on Inform. Theory*, page 186, Ulm, Germany, June/July 1997.

[34] R. Zamir and K. Rose. Natural type selection in adaptive lossy compression. *IEEE Trans. Inform. Theory*, 47(1):99–111, 2001.

[35] 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*, 43(1):71–91, 1997.