Mismatched Codebooks and the Role of Entropy-Coding in Lossy Data Compression

I. Kontoyiannis R. Zamir

November 2, 2005

Abstract — We introduce a universal quantization scheme based on random coding, and we analyze its performance. This scheme consists of a source-independent random codebook (typically *mismatched* to the source distribution), followed by optimal entropy-coding that is *matched* to the quantized codeword distribution. A single-letter formula is derived for the rate achieved by this scheme at a given distortion, in the limit of large codebook dimension. The rate reduction due to entropy-coding is quantified, and it is shown that it can be arbitrarily large. In the special case of "almost uniform" codebooks (e.g., an i.i.d. Gaussian codebook with large variance) and difference distortion measures, a novel connection is drawn between the compression achieved by the present scheme and the performance of "universal" entropy-coded dithered lattice quantizers. This connection generalizes the "half-a-bit" bound on the redundancy of dithered lattice quantizers. Moreover, it demonstrates a strong notion of universality where a single "almost uniform" codebook is near-optimal for *any* source and *any* difference distortion measure. The proofs are based on the fact that the limiting empirical distribution of the first matching codeword in a random codebook can be precisely identified. This is done using elaborate large deviations techniques, that allow the derivation of a new "almost sure" version of the conditional limit theorem.

Index Terms — Rate-distortion theory, random coding, mismatch, universal quantization, universal Gaussian codebook, pattern-matching, large deviations, data compression, robustness.

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

[2]Ram Zamir is with the Department of Electrical Engineering - Systems, Tel-Aviv University, Ramat Aviv 69978, Israel. Email: `zamir@eng.tau.ac.il` Web: `www.eng.tau.ac.il/~zamir/`

[3]I. Kontoyiannis was supported in part by NSF grant #0073378-CCR, and by USDA-IFAFS grant #00-52100-9615. R. Zamir was supported in part by the US-Israel Bi-National Science Foundation grant 1998-309.

1 Introduction

1.1 Mismatched Quantization and Compression

Variable-rate lossless compression – or *entropy-coding* – is an efficient method for enhancing the compression performance of quantizers [12, 4]. This paper investigates the role of entropy-coding when the quantizer codebook is *mismatched* with respect to the source distribution. Our motivation mainly comes from Ziv's concept of *universal quantization* for lossy compression of real-valued sources with unknown statistics [37]. Ziv's scheme uses a randomized ("dithered") lattice quantizer, which is scaled to meet the target distortion level, and the quantizer is followed by a universal lossless encoder which reduces the coding rate to the true entropy of the quantized sequence. Neuhoff [23] suggested that the universal quantizer could be viewed as an efficient combination of a "simple" robust quantizer and a "complex" lossless encoder. Variations on the problem of entropy-coded dithered quantization (ECDQ) can be found in [14, 32, 33].

Intuitively, the quantizer mismatch leaves much room for rate savings using entropy-coding. Moreover, unlike *optimum* entropy-constrained vector quantization (ECVQ) [11, 22], the entropy-coding gain in the mismatched case does not vanish even in the limit of large vector dimension. For a rather trivial example, note that the un-coded rate of an unbounded lattice quantizer is infinite, but it becomes finite after entropy-coding if the source has finite variance. The advantage of entropy-coding a mismatched quantizer is particularly prominent at high resolution quantization conditions. Gray and Linder show that *optimum* high-rate performance for mean-squared distortion can be achieved even if the quantizer codebook is mismatched with respect to the source (specifically, if it is designed for a source with uniform density), as long as the quantizer output is entropy-coded according to the *true* quantizer output distribution [13, sec. VII]. As we shall see here, similar behavior occurs at *any* resolution, only with a slight rate loss due to the codebook mismatch.

One of the central results of universal quantization theory is that, after entropy-coding, the rate loss of the universal quantizer with respect to the optimum ECVQ is bounded for *all* sources and *all* distortion levels by a universal constant [37, 32]. For example, for squared error distortion, the rate loss of a k-dimensional lattice ECDQ is bounded by $(1/2)\log(4\pi e G_k)$ bits, where G_k is the normalized second moment of the lattice; this bound is ≈ 0.754 bits for $k = 1$, and it converges to $1/2$ bit as $k \to \infty$ (where $\log = \log_2$). These results are limited, however, to lattice structured quantizers, and more specifically to those lattice dimensions and distortion measures which are covered by lattice coding theory.

The central goal of this paper is to develop a structure-free framework for mismatched, entropy-coded quantization at an arbitrary distortion level, based on random coding ideas and techniques. The random coding framework, although not constructive, allows us to precisely quantify two important operational quantities: (a) The potential rate gain due to entropy-coding when using a mismatched random codebook; equivalently, this can be thought of as the rate loss of the straightforward scheme which uses a mismatched codebook without entropy-coding. (b) The rate loss due to quantizer mismatch, over the optimal rate-distortion function: We will derive a universal upper bound for this rate

loss, analogous to the half-a-bit bound for lattice ECDQs and quadratic distortion described above.

Mismatched random codebooks for *fixed-rate* lossy source coding have been investigated by Sakrison [26, 27], Zhang and Wei [36], Lapidoth [20], Zamir and Rose [34], and others. See [15, 16] and the references therein. Specifically, source coding with a mismatched random codebook (or string matching with a mismatched database) has been considered by Steinberg and Gutman [28], Yang and Kieffer [29], and Dembo and Kontoyiannis [8], among others. These works (and the references therein) develop an extensive theory of mismatched random lossy coding in the limit of large codebook dimension, with an emphasis on precisely characterizing the asymptotic rate and the redundancy of these schemes. Here we continue that investigation, but we introduce the additional step of entropy-coding the index of the codebook before transmitting it to the decoder.

Entropy coding the codeword index in a source-matched random lossy codebook has been considered in the early work by Pinkston [24]. Mismatched high resolution quantization has been considered by Bucklew [3] and by Gray and Linder [13], where several results, some of which parallel those derived here, are presented. Preliminary results on the entropy rates achieved by mismatched random codebooks for general (non vanishing) distortions and discrete memoryless sources appear in [31]. Here we strengthen these results, and extend them to richer classes of sources and codebook distributions. In particular, we establish a formal connection between ECDQ and entropy-coded random codebooks.

1.2 Discussion of Main Results

We begin in Section 2, where we derive asymptotic single-letter characterizations for the compression rate achieved by two different coding schemes, both based on a random codebook $\boldsymbol{C}_n = \{Y_1^n(i),\ i = 1, 2, \ldots\}$ consisting of i.i.d. n-dimensional words Y_1^n, each having i.i.d. components generated by an arbitrary distribution Q. We shall discuss later specific interesting choices for the codebook distribution Q. A natural motivation for the use of a mismatched Q is the observation that, in many important applications, the source statistics are generally unknown *a priori* or they change with time – or both.

Given a source string X_1^n to be compressed with distortion D or less, we consider the index N_n of the first codeword in $\boldsymbol{C}_n$ that matches X_1^n within distortion D. Our first result says that as $n \to \infty$, the empirical distribution of this first matching word converges to a distribution Q^*_{PQD} which can be identified as the solution of an single-letter minimization problem. [Here $P = P_1$ denotes the first-order marginal of the source distribution.] The proof is based on large deviations techniques, and generalizes the "favorite type theorem" of [34].

Using this result we establish an upper bound on the rate achieved when such a random codebook is used in conjunction with entropy-coding: Suppose that the encoder first finds the first D-close match at position N_n, and then entropy-codes the index N_n conditional on the codebook $\boldsymbol{C}_n$. The rate achieved for this D-accurate description of X_1^n is

$$H(N|\boldsymbol{C}) \stackrel{\triangle}{=} \lim_{n\to\infty} \frac{1}{n} H(N_n|\boldsymbol{C}_n) \quad \text{bits/symbol.}$$

We then compare our bound with the limiting rate $R(P, Q, D)$ achieved in the "naive coding" scenario,

where the encoder simply transmits the index N_n using Elias' code for the integers,

$$R(P, Q, D) = \lim_{n\to\infty} \frac{1}{n} \log(N_n) \quad \text{bits/symbol.}$$

We show that the rate gain of entropy-coding over the naive coding scheme (or, equivalently, the rate loss of naive coding) satisfies

$$\text{rate gain} = R(P, Q, D) - H(N|\boldsymbol{C}) \geq H(Q^*_{PQD}\|Q) \quad \text{bits/symbol,}$$

i.e., it is at least as large as the relative entropy between the limiting empirical distribution of $Y_1^n(N_n)$ and the codebook-generating distribution Q. For example, it is approximately $H(P\|Q)$ for small mean squared distortion D. This lower bound is strictly positive, unless Q is the optimal reproduction distribution (i.e., the optimal output distribution of the rate-distortion function).

This expression resembles the rate loss due to mismatch in the *lossless* component of the code at high resolution quantization (see, e.g., [13]). Indeed, for small mean-squared distortion we have $H(N|\boldsymbol{C}) \approx h(P) - \frac{1}{2}\log(2\pi eD)$, and $R(P, Q, D) \approx h(P) - \frac{1}{2}\log(2\pi eD) + H(P\|Q)$, hence the latter amounts to encoding a source $\sim P$ using a code designed for a source $\sim Q$. At *non*-high resolution conditions, however, the rate loss remains positive even if the lossless component of the code is matched, i.e., $H(N|\boldsymbol{C})$ is in general strictly above the rate-distortion function.

Of particular interest is the case of universal Gaussian codebooks: Suppose we encode a real-valued *memoryless* source using a white Gaussian codebook $\sim N(0, \tau^2)$, with respect to squared error distortion. If we simply use this codebook in the "naive" sense described above, robust source coding theory implies that taking $\tau^2 = \sigma^2 - D$, where σ^2 denotes the source variance, guarantees achieving the Gaussian rate-distortion function $R(D) = \frac{1}{2}\log(\sigma^2/D)$ for *any* source [26, 20]. Since the Gaussian source is the hardest to compress in this class, this implies high redundancy when the source is far from Gaussian.

On the other hand, if we also allow the encoder to entropy code the index, the results are fundamentally different. If we take the codebook variance τ^2 to be large, the codebook distribution becomes flat and it is tempting to think that the codebook itself looks approximately like the codebook of a lattice quantizer. Indeed, we show that as $\tau^2 \to \infty$ the rate $H(N|\boldsymbol{C})$ achieved by entropy-coding this Gaussian codebook is no greater than the rate of a dithered lattice quantizer with large lattice dimension, given by

$$\limsup_{\tau^2\to\infty} H(N|\boldsymbol{C}) \leq I(X; X + Z_D) \quad \text{bits/symbol,}$$

where $I(X; X + Z_D)$ denotes the mutual information between the first source symbol X and $X + Z_D$, where Z_D is an independent $N(0, D)$ random variable. Combining this with well-known facts about universal quantizers [37, 32], it follows that the naive coding rate is going to infinity as $\tau^2 \to \infty$, whereas the limiting rate achieved by entropy-coding, $I(X; X + Z_D)$, is at most 1/2 bit above the rate-distortion function of X, and it coincides with the rate-distortion function of X in the limit of small D. This new derivation provides an interesting bridge between universal quantization theory and mismatched random coding.

As observed by Yang and Kieffer [29], the naive coding rate $R(P, Q, D)$ depends only on the first-order marginal of the source distribution, hence it does not benefit from memory in the source. Moreover, as shown in detail in the sequel, the entire dependence of the naive coding rate on the source distribution P may be very weak (in fact, sometimes it is *entirely independent* of P), thus it is far from being optimal. As we shall see, these disadvantages are eliminated by the use of entropy coding.

In particular, for the case of memoryless Gaussian codebooks with large variance, we argue that the entropy-coding scheme achieves a rate no greater than the mutual information rate $I(\boldsymbol{X}, \boldsymbol{X} + \boldsymbol{Z}_D)$, where $\boldsymbol{Z}_D$ is an independent white Gaussian process with variance D. As before, this in turn implies that the rate of the entropy-coded scheme is no greater than $R(D) + 1/2$ bits/symbol, where $R(D)$ is the rate distortion function of the entire source (not just the first-order rate-distortion function).

We also show that these results generalize beyond the Gaussian codebook case to a much wider class of codebooks, namely, "approximately flat" codebooks with distributions of exponential type, and to general difference distortion measures. We quantify the entropy-coding gain in this case, and show that the resulting compression rate is bounded above by $R(D) + C^*$ bits/symbol, where $R(D)$ is the rate-distortion function of the source, and C^* is an upper bound for the "min-max capacity" defined in [35]. This is a new generalization of the well-known half-a-bit bound derived for dithered lattice quantizers and squared distortion [32] to the case of general difference distortion measures. Moreover, *it implies the existence of a single ensemble of codebooks which is universal with respect to both the source distribution and the distortion criterion*, resembling the results of Yang and Kieffer in [30].

The paper is organized as follows. In Section 2 we describe in detail the entropy-coding scenario and the naive coding scheme based on a mismatched random codebook, and we state the main result on the index entropy in Theorem 1. Section 3 contains two important examples illustrating the entropy-coding gain, including the case of universal Gaussian codebooks mentioned above. In Section 4 we state and prove an almost sure conditional limit theorem (Theorem 3), which forms the basis for the favorite type theorem (Theorem 2) and for the proof of Theorem 1 which is given in Section 5. Finally, in Section 6 we give tighter bounds on the index entropy and the entropy-coding gain for sources with memory.

2 The Performance of Mismatched Codebooks

In this section we characterize the compression performance achieved by memoryless random codebooks when used to compress data generated from a stationary ergodic source. Two coding scenarios are considered: The "naive coding" scenario where data is simply described by the index of the first match in the codebook, and the "entropy-coded" case where this index is entropy-coded.

In Section 3 we compare the performance of these two schemes, and explicitly evaluate the entropy-coding gain in two important special cases.

2.1 Notation and Definitions

We begin by introducing some basic definitions and notation that will remain in effect for the rest of the paper.

Consider a stationary ergodic process (or source) $\boldsymbol{X} = \{X_n \,;\, n \geq 1\}$ taking values in the source alphabet A. We will assume throughout that A is a complete, separable metric space (often called a Polish space), equipped with its associated Borel σ-field $\mathcal{A}$. For the sake of simplicity we also make the (rather harmless) assumption that all singletons are measurable, i.e., $\{x\} \in \mathcal{A}$ for all $x \in A$. Similarly, for the reproduction alphabet $\hat{A}$ we take $(\hat{A}, \hat{\mathcal{A}})$ to be the Borel measurable space corresponding to a complete, separable metric (or Polish) space $\hat{A}$ and assume that $\{y\} \in \hat{\mathcal{A}}$ for all $y \in \hat{A}$. We write X_i^j for the vector of random variables $X_i^j = (X_i, X_{i+1}, \ldots, X_j)$, and similarly $x_i^j = (x_i, x_{i+1}, \ldots, x_j) \in A^{j-i+1}$ for a realization of these random variables, $-\infty \leq i \leq j \leq \infty$. We let P_n denote the marginal distribution of X_1^n on A^n ($n \geq 1$), and write $\mathbb{P}$ for the distribution of the whole process. We use P for the first-order marginal P_1.

Given an arbitrary nonnegative (measurable) function $\rho : A \times \hat{A} \to [0, \infty)$, define a sequence of single-letter (or "additive") distortion measures $\rho_n : A^n \times \hat{A}^n \to [0, \infty)$ by

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

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

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

Throughout the paper, log denotes the logarithm to base 2 and ln denotes the natural logarithm. Unless otherwise mentioned, all familiar information-theoretic quantities (entropy, mutual information, and so on) are defined in terms of logarithms taken to base 2, and are therefore expressed in bits.

2.2 Random Codebooks

Given a probability measure Q on the reproduction alphabet $\hat{A}$, a *memoryless random codebook* $\boldsymbol{C}_n$ *with distribution* Q is an infinite sequence of i.i.d. random vectors $Y_1^n(i)$, $i \geq 1$, with each $Y_1^n(i)$ being distributed according to the product measure Q^n on $\hat{A}^n$. In other words, the components of $Y_1^n(i)$ are i.i.d. with distribution Q. We write

$$\boldsymbol{C}_n \stackrel{\triangle}{=} \{Y_1^n(i) \,;\, i \geq 1\}$$

for the entire codebook, and we call Q the *codebook distribution.*

Suppose that, for a fixed n, this codebook is available to both the encoder and decoder. Given a distortion level D and a source string X_1^n to be described with distortion D or less, the encoder looks for a D-close match of X_1^n into the codebook $\boldsymbol{C}_n$. Let N_n be the position of the first such match,

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

with the convention that the infimum of the empty set equals $+\infty$. Roughly speaking, the way the encoder describes X_1^n is by describing the position N_n of this first match.

Given a codebook distribution Q on $\hat{A}$, we define

$$\begin{aligned} D_{\min} &\stackrel{\triangle}{=} E_P[\operatorname*{ess\,inf}_{Y\sim Q} \rho(X,Y)] \\ D_{\mathrm{av}} &\stackrel{\triangle}{=} E_{P\times Q}[\rho(X,Y)], \end{aligned}$$

where $P = P_1$ denotes the first-order marginal of $\boldsymbol{X}$.[1] We will assume throughout that D_{av} is finite. Clearly $0 \le D_{\min} \le D_{\mathrm{av}}$. To avoid the trivial case when $\rho(x,y)$ is constant for (P-almost) all $x \in A$, we assume that with positive P-probability $\rho(x,y)$ is not essentially constant in y, that is:

$$D_{\min} < D_{\mathrm{av}}.$$

Note also that for D greater than D_{av} the rate-distortion function $R(D)$ of $\boldsymbol{X}$ is zero, and that for D below $D_{\min}$ no match can ever be found. Therefore, from now on we restrict our attention to the interesting range of distortion levels $D \in (D_{\min}, D_{\mathrm{av}})$.

We consider two possible ways in which the encoder can transmit N_n: The simplest thing to do is describe N_n directly, using some predetermined code for the positive integers; see, e.g., [10]. This can be done with approximately $\log(N_n)$ bits. Alternatively, once the codebook $\boldsymbol{C}_n$ has been fixed, the encoder may choose to "entropy-code" N_n, giving it an average description length of roughly $H(N_n|\boldsymbol{C}_n)$ bits. This is equivalent to re-ordering the codewords according to decreasing order of probabilities, and then describing the new index $\pi(N_n)$ using approximately $\log(\pi(N_n))$ bits like above. When the statistics of the source are a-priori unknown, we use a universal algorithm to entropy-code (or re-order) the codewords.

2.3 Naive Coding

First we consider the case when the encoder describes the index N_n without entropy-coding; we refer to this scenario as "naive coding." As mentioned in the Introduction, this coding scheme (and many variations on it) has been analyzed extensively in [28][29][8] and several other works cited therein. To avoid potentially infinite searches in the codebook, we make the simplifying assumption that the encoder only describes N_n when it is smaller than 2^{nb}, where b is some positive constant to be chosen later. Accordingly, we define the truncated index N_n':

$$N_n' \stackrel{\triangle}{=} \begin{cases} N_n, & \text{if } N_n \le \lfloor 2^{nb} \rfloor, \\ \lfloor 2^{nb} \rfloor + 1, & \text{otherwise.} \end{cases}$$

When N_n exceeds $\lfloor 2^{nb} \rfloor$, the encoder uses an alternative description for X_1^n. In order to ensure that such a description can be given with finite rate, we introduce the following simple conditions; cf. [17, 19, 8].

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

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

$$\rho(x, q(x)) \leq D \quad \text{for all } x \in A.$$

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

$$M_p \stackrel{\triangle}{=} \{E_P[(-\log \mu(q(X_1)))^p]\}^{1/p} < \infty,$$

where μ denotes the (discrete) distribution of the quantized random variable $q(X)$.

Note that for all $p' \geq p \geq 1$ we clearly have (p'SQC) $\Rightarrow$ (pSQC) $\Rightarrow$ (WQC), and that if the quantizer q of (WQC) has finite range then (pSQC) automatically holds for all $p \geq 1$. In particular, (1SQC) amounts simply to the requirement that there exists an appropriate scalar quantizer q with $H(q(X_1)) < \infty$.

The encoder describes X_1^n with distortion D or less in two steps. First, a description of N_n' is given using Elias' code for the integers [10]. This takes

$$\log N_n' + 2 \log \log N_n' + O(1) \quad \text{bits.} \tag{1}$$

If $N_n' \leq \lfloor 2^{nb} \rfloor$, then $N_n = N_n'$ and the above description is sufficient for the decoder to recover a D-close version of X_1^n from the codebook, so the second step is omitted. And if $N_n' = \lfloor 2^{nb} \rfloor + 1$, then the encoder also gives a representation of X_1^n with distortion D or less using the quantizer q provided by (WQC). This can be given in

$$\sum_{i=1}^{n} \lceil -\log \mu(q(X_i)) \rceil \quad \text{bits.}$$

Let $\ell_n(X_1^n)$ denote the overall description length of the algorithm just described. As we will see the constant b can be chosen in such a way that N_n will eventually be small enough so that the encoder will never need to resort to the alternative coding method. Therefore, in view of (1), to understand this code's compression performance (i.e., to understand the asymptotic behavior of $\ell_n(X_1^n)$) it suffices to understand the behavior of $(\log N_n)$ for large n.

Suppose that a source string X_1^n is given; the probability that any particular codeword $Y_1^n(i)$ matches X_1^n with distortion D or less is $Q^n(B(X_1^n, D))$. If this probability is nonzero, then, conditional on X_1^n, the distribution of N_n is geometric with parameter $Q^n(B(X_1^n, D))$. From this observation it is easy to deduce that N_n is close to its mean, namely $1/Q^n(B(X_1^n, D))$, when n is large. The following result is an easy consequence of this fact and of Theorem B below.

Theorem A. Naive Coding Performance. [8]: Suppose that $\boldsymbol{X}$ is a stationary ergodic source with first-order marginal $P_1 = P$, and that Q is an arbitrary codebook distribution on $\hat{A}$ with $D_{\text{av}} < \infty$. If $D \in (D_{\min}, D_{\text{av}})$ and $\boldsymbol{X}$ satisfies (WQC) at D, then for almost every sequence of memoryless random codebooks $\boldsymbol{C}_n$ with distribution Q:

$$\lim_{n\to\infty} \frac{1}{n} \ell_n(X_1^n) \;=\; \lim_{n\to\infty} \frac{1}{n} \log N_n \;=\; R(P, Q, D) \quad \text{bits/symbol, w.p.1.}$$

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

$$R(P, Q, D) = \inf_W H(W\|P \times Q)\,, \tag{2}$$

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

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

and the infimum in (2) is taken over all joint distributions W on $A \times \hat{A}$ such that the first marginal of W is P and $E_W[\rho(X, Y)] \le D$. This result holds as long as the constant b is chosen $b > R(P, Q, D)$.

See Section 3 for specific examples where the asymptotic rate $R(P, Q, D)$ can be explicitly evaluated.

Theorem B. [8]: Let $\boldsymbol{X}$ be a stationary ergodic source with first-order marginal distribution P, and let Q be an arbitrary codebook distribution on $\hat{A}$ with $D_{\text{av}} < \infty$. Then for all $D \in (D_{\min}, D_{\text{av}})$:

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

2.4 Entropy-Coding the Index

Next we consider the case of entropy-coding, where, after the codebook $\boldsymbol{C}_n$ has been fixed, the encoder uses the conditional distribution (given $\boldsymbol{C}_n$) of the position of the first D-close match to optimally describe this position to the decoder: The truncated index N_n' is first described using $H(N_n'|\boldsymbol{C}_n)+O(1)$ bits, on the average. As before, if $N_n \le \lfloor 2^{nb} \rfloor$ this offers a complete D-close representation of X_1^n. Otherwise, the encoder adds to this an alternative representation of X_1^n using the quantizer q provided by (WQC). On the average, this takes

$$\sum_{i=1}^{n} E_P\Big(\lceil -\log \mu(q(X_i))\rceil\Big) \quad \text{bits.}$$

Let $\mathcal{L}_n(X_1^n)$ denote the overall description length of the above coding scheme. Next we give an upper bound on the asymptotic rate it achieves. Given an arbitrary codebook distribution Q on $\hat{A}$, define the *output-constrained rate-distortion function* (or lower mutual information (LMI)) [31, 36] by

$$I_m(P\|Q, D) \stackrel{\triangle}{=} \inf_{X\sim P,\, Y\sim Q\ E\rho(X,Y)\le D} I(X; Y), \tag{3}$$

where $I(X; Y)$ denotes the mutual information between X and Y, and the infimum is taken over all jointly distributed random variables (X, Y) such that $X \sim P$, $Y \sim Q$, and $E\rho(X, Y) \le D$. Using the chain rule for relative entropy it is easy to verify that the earlier rate-function $R(P, Q, D)$ can be expressed as

$$R(P, Q, D) = \inf_{\widetilde{Q}} [I_m(P\|\widetilde{Q}, D) + H(\widetilde{Q}\|Q)], \tag{4}$$

where the infimum is over all probability measures $\widetilde{Q}$ on $\hat{A}$. As we show in Section 4.1, the minimizer of (4) exists and is unique, and we denote it by $Q^*_{P,Q,D}$:

$$Q^*_{P,Q,D} = \arg\min_{\widetilde{Q}} [I_m(P\|\widetilde{Q}, D) + H(\widetilde{Q}\|Q)].$$

Theorem 1. Entropy-Coding Performance: Suppose that $\boldsymbol{X}$ is a stationary ergodic source with first-order marginal distribution P, and that Q is an arbitrary (i.i.d.) codebook distribution with $D_{\rm av} < \infty$. Assume that $D \in (D_{\min}, D_{\rm av})$ and that $\boldsymbol{X}$ satisfies (pSQC) at D for some $p > 1$. Then the rate of the entropy-coded scheme with memoryless codebooks $\boldsymbol{C}_n$ with distribution Q, satisfies,

$$\limsup_{n\to\infty} \frac{1}{n} E[\mathcal{L}_n(X_1^n)] \;=\; \limsup_{n\to\infty} \frac{1}{n} H(N_n'|\boldsymbol{C}_n) \;\le\; I_m(P\|Q^*_{P,Q,D}, D) \quad \text{bits/symbol}, \tag{5}$$

where the expectation is taken over both the message X_1^n and the random codebook $\boldsymbol{C}_n$. This result holds as long as $b > R(P, Q, D)$.

We immediately obtain from this and Theorem A above:

Corollary 1. Entropy-Coding Gain: Under the assumptions of Theorem 1, the rate gain of entropy-coding over the naive coding scheme is at least

$$R(P, Q, D) - I_m(P\|Q^*_{P,Q,D}, D) = H(Q^*_{P,Q,D}\|Q) \quad \text{bits/symbol}. \tag{6}$$

As shown in [31], entropy coding *without* conditioning on the codebook yields a rate equal to the naive coding rate. Thus, the quantity in the right hand side of (6) can also be thought of as the rate-gain due to matching the entropy coder to the specific realization of the codebook.

In Section 6 we generalize and refine the bounds (5) and (6). The discussion there suggests that, as in the case of discrete memoryless sources [31], these bounds are tight also for general memoryless sources; for sources with memory the inequality (5) is strict and hence the entropy-coding gain (6) is larger.

The measure $Q^*_{P,Q,D}$ has an interesting coding interpretation that will be clarified further in Theorem 2: When n is large, the empirical distribution of the first matching codeword $Y_1^n(N_n)$ in the codebook is close to $Q^*_{P,Q,D}$ with high probability. In the case of discrete memoryless sources this phenomenon can be explained using the method of types as in [34]. The lower mutual information $I_m(P\|\widetilde{Q}, D)$ represents the rate achieved by a fixed-composition codebook, namely a codebook consisting exclusively of codewords with type $\widetilde{Q}$. Equivalently, $I_m(P\|\widetilde{Q}, D)$ is the exponent in the probability that a source string will match a type-Q string with distortion D or less. In this light, a memoryless random codebook with distribution Q can be thought of as a union of polynomially many fixed-composition codebooks, where the proportion of words of type $\widetilde{Q}$ is $\exp[-nH(\widetilde{Q}\|Q)]$. Now, codewords of type $\widetilde{Q} = Q$ are very frequent in the codebook but their lower mutual information is very high (i.e., low matching probability), whereas codewords of type $\widetilde{Q} = Q^*_{P,D}$, where $Q^*_{P,D}$ achieves the rate-distortion function (11), have the lowest lower mutual information (i.e., high matching probability), but they are too rare in the codebook. Therefore, we can think of the achieving measure

$\widetilde{Q} = Q^*_{P,Q,D}$ in (4) as corresponding to the codeword type that strikes the optimum balance between the competing requirements of high matching probability and high frequency in the codebook.

Note that it follows from (3) and (4) that

$$R(P,Q,D) \geq I_m(P\|Q^*_{P,Q,D},D) \geq R(D)$$

with equality in both inequalities if and only if Q achieves the rate-distortion function in (11).

Proof outline of Theorem 1. The equality in (5) follows simply from the observation that $E[\mathcal{L}_n(X_1^n)] \geq H(N_n'|\boldsymbol{C}_n)$ combined with

$$\begin{aligned} E[\mathcal{L}_n(X_1^n)] &\leq H(N_n'|\boldsymbol{C}_n) + E\left[\sum_{i=1}^n \Big\{[-\log\mu(q(X_i))+1]\mathbb{I}_{\{N_n'=\lfloor 2^{nb}\rfloor+1\}}\Big\}\right] \;+\; O(1) \\ &\overset{(a)}{\leq} H(N_n'|\boldsymbol{C}_n) + n(M_p+1)\Pr\{N_n \geq \lfloor 2^{nb}\rfloor+1\} \;+\; O(1) \\ &\overset{(b)}{\leq} H(N_n'|\boldsymbol{C}_n) + o(n), \end{aligned} \tag{7}$$

where $\mathbb{I}_E$ denotes the indicator function of an arbitrary event E, (a) follows by Hölder's inequality, and (b) follows from Theorem B.

The existence and uniqueness of Q^*_{PQD} is established in Section 4. The inequality in (5) is the main technical content of the theorem; its proof is given in Section 5. □

3 The Entropy-Coding Gain

In this section we illustrate the gain of entropy-coding over the naive coding scheme in two particular instances where it can be explicitly evaluated. As mentioned in the introduction, we first consider the case of Gaussian codebooks with large variance. Since such a codebook distribution is approximately uniform over the whole real line, it is tempting to think of the entropy-coded scheme as a "randomized" version of an entropy-coded, uniform lattice quantizer. Indeed, we show that, as the codebook variance grows to infinity, the rate achieved by the entropy-coded scheme is at least as good as the asymptotic rate of entropy-coded dithered quantization (ECDQ)

Then in Section 3.2 we consider a more general class of approximately uniform, or "asymptotically flat," codebook distributions, corresponding to appropriately defined exponential families. In this case we argue that the resulting compression performance can be determined in a way analogous to the analysis given for the Gaussian case.

3.1 Universal Gaussian Codebooks

Let $\boldsymbol{X}$ be a stationary and ergodic, real-valued source to be compressed, and suppose $\boldsymbol{X}$ has zero mean $E(X_1) = 0$ and finite variance $\sigma^2 = \mathrm{Var}(X_1) < \infty$. We consider memoryless random codebooks generated according to the Gaussian distribution $Q \sim N(0,\tau^2)$, and we take ρ be the squared-error

distortion measure $\rho(x,y) = (x-y)^2$. Under these assumptions, the rate achieved by the naive coding scheme is [8],

$$R(P,Q,D) = \begin{cases} \infty\,, & D = 0 \\ \frac{1}{2}\log\left(\frac{v}{D}\right) - (\log e)\frac{(v-D)(v-\sigma^2)}{2v\tau^2}\,, & 0 < D < \sigma^2+\tau^2 \\ 0\,, & D \geq \sigma^2+\tau^2, \end{cases} \tag{8}$$

where

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

[Note that here $D_{\min} = 0$, $D_{\rm av} = \sigma^2 + \tau^2$.] We observe that in this case $R(P,Q,D)$ depends only on the first and second moments of the source distribution, and that asymptotically for large codebook variance it takes the form

$$R(P,Q,D) = \frac{1}{2}\log\left(\frac{\tau^2}{eD}\right) + o(1) \tag{9}$$

where $o(1) \to 0$ as $\tau^2 \to \infty$.

Remark: In more familiar information-theoretic terms, the rate-function $R(P,Q,D)$ can equivalently be expressed as

$$R(P,Q,D) = \inf_{(X,Y)} \left[I(X;Y) + H(Q_Y \| Q)\right] \tag{10}$$

where the infimum is over all jointly distributed random variables (X,Y) with values in $A \times \hat{A}$, such that X has distribution P, $E[\rho(X,Y)] \leq D$, and Q_Y denotes the distribution of Y; cf. [29].

This expression shows that, typically, the rate achieved by the naive coding scheme is strictly suboptimal, unless of course the source itself is memoryless and Q is chosen to minimize $R(P,Q,D)$. In fact from (10) it is immediate that for a memoryless source with rate-distortion function $R(D)$ we indeed have

$$R(D) = \inf_{\widetilde{Q}} R(P,\widetilde{Q},D), \tag{11}$$

where the infimum is over all probability distributions $\widetilde{Q}$ on $\hat{A}$.

Example. Known Variance. Now suppose that the source $\boldsymbol{X}$ is believed to be i.i.d. Gaussian with $N(0,\sigma^2)$ distribution. As is well-known [1, 5], for any $D \in (0,\sigma^2)$ the optimal coding distribution is $Q^* \sim N(0,\sigma^2 - D)$, therefore we construct memoryless random codebooks according to Q^*. But instead of the Gaussian source we expected, we are faced with data from some arbitrary stationary ergodic $\boldsymbol{X}$ with zero mean and variance σ^2. From the previous example it follows that the asymptotic rate achieved by the naive coding scheme will be [substituting $\tau^2 = \sigma^2 - D$ in (8)]

$$\frac{1}{2}\log\left(\frac{\sigma^2}{D}\right) \quad \text{bits/symbol.}$$

This is exactly the rate-distortion function of the i.i.d. $N(0,\sigma^2)$ source, so the rate achieved is the same as what we would have obtained on the Gaussian source we originally expected. This coincides with Sakrison's robust fixed-rate for a class of sources [26]. It is yet another version of the folk theorem that the Gaussian source is the hardest one to compress, among all real-valued sources with a fixed variance; cf. [20].

Turning back to the general case, suppose $\boldsymbol{X}$ is a zero-mean, stationary and ergodic, real-valued source, with variance $\sigma^2 = \mathrm{Var}(X_1) < \infty$, and let the codebook distribution be $Q \sim N(0,\tau^2)$. Choose and fix a distortion level $D > 0$. From (9) we have that, for τ^2 large, the rate achieved by the naive coding scheme is

$$\log(\tau) + O(1) \quad \text{bits/symbol}$$

which of course grows to infinity as $\tau^2 \to \infty$. On the other hand, as we show in the following proposition the rate achieved by the entropy-coding scheme stays bounded, and for memoryless sources it coincides with the asymptotic (large vector dimension) rate of entropy-coded dithered lattice quantization (ECDQ). This confirms the natural intuition that the behavior of a random codebook with an approximately flat distribution should mimic the behavior of an entropy-coded uniform quantizer.

Proposition 1. Entropy-Coding Gain for Universal Gaussian Codebooks: Let $\boldsymbol{X}$ be a real-valued, zero-mean, stationary ergodic source, with finite variance $\sigma^2 = \mathrm{Var}(X_1)$, let $D > 0$ be a fixed distortion level, and take $Q \sim N(0,\tau^2)$. Let Q^*_{PQD} be as in Theorem 1. We have:

(i) The measure Q^*_{PQD} converges to $P * N(0,D)$ as $\tau \to \infty$, in that Q^*_{PQD} has a density $f_{Q^*_{PQD}}(y)$ (with respect to Lebesgue measure) for all τ large enough and

$$f_{Q^*_{PQD}}(y) \to E_P[\phi_D(y - X)] \quad \text{as } \tau^2 \to \infty, \text{ for all } y \in \mathbb{R},$$

where ϕ_D denotes the density of the $N(0,D)$ distribution.

(ii) The upper bound $I_m(P\|Q^*_{PQD}, D)$ to the rate achieved by the entropy-coding scheme satisfies

$$\lim_{\tau^2\to\infty} I_m(P\|Q^*_{PQD}, D) = I(X; X + Z_D) \quad \text{bits/symbol},$$

where $X \sim P$, and Z_D denotes a $N(0,D)$ random variable independent of X.

(iii) If the source $\boldsymbol{X}$ is memoryless then as $\tau \to \infty$ the rate achieved by the entropy-coding scheme is no greater than

$$R(D) + \frac{1}{2} \quad \text{bits/symbol},$$

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

A proof outline for Proposition 1 is given in Appendix A. As we will discuss in Section 6, for sources with memory the entropy-coding gain is generally significantly larger. In fact when $\boldsymbol{X}$ is not memoryless:

1. the rate achieved by the entropy-coding scheme is actually equal to the mutual information rate $I(\boldsymbol{X}; \boldsymbol{X} + \boldsymbol{Z}_D)$, where $\boldsymbol{Z}_D$ is a white Gaussian process with variance D;

2. the result in part (iii) of the proposition is valid for all stationary and ergodic sources.

3.2 Approximately Flat Codebooks and Difference Distortion Measures

We now extend the asymptotic result above to more general codebook distributions and to arbitrary difference distortion measures (not necessarily squared error).

Suppose that $\hat{A} = A = \mathbb{R}$ and that ρ is a difference distortion measure of the form $\rho(x, y) = \rho(y-x)$ for some $\rho : \mathbb{R} \to [0, \infty)$. Here we consider real-valued, stationary ergodic sources $\boldsymbol{X}$, and codebook distributions Q that have a density f_Q (with respect to Lebesgue measure).

We begin by deriving a lower bound for the rate-function $R(P, Q, D)$, in the spirit of the Shannon Lower Bound [21].

Lemma SLB. A "Shannon Lower Bound" for Difference Distortion Measures: Assume that the codebook distribution Q has a density f_Q (with respect to Lebesgue measure), and let $Q_{\max} = \sup_y f_Q(y)$. Then,

$$R(P, Q, D) \geq \log(1/Q_{\max}) - h_{\max}(D), \tag{12}$$

where $h_{\max}(D)$ is the maximum entropy associated with ρ and D,

$$h_{\max}(D) = \max_{f:\ E_f[\rho(Z)] \leq D} h(f)$$

and where $h(f) = -\int f(x) \log f(x) dx = h(Z)$ denotes the differential entropy of a random variable Z with density f.

Note that the lower bound (12) is independent of the source distribution P.

Proof. Consider the infimum in (10). For any jointly distributed (X, Y) such that $E[\rho(Y-X)] \leq D$, let Q_Y and f_Y denote the measure and the density describing the distribution of Y, respectively. We can then write [5],

$$I(X;Y) = h(Y) - h(Y|X) = h(Y) - h(Y - X|X) \geq h(Y) - h(Y - X) \geq h(f_Y) - h_{\max}(D)$$

where the first inequality holds since conditioning reduces the entropy. On the other hand, we can expand

$$H(Q_Y \| Q) = -h(f_Y) + E_{Q_Y}[-\log f_Q(Y)] \geq -h(f_Y) + \log(1/Q_{\max}).$$

Combining, we get

$$I(X;Y) + H(Q_Y \| Q) \geq \log(1/Q_{\max}) - h_{\max}(D). \tag{13}$$

Now note that we have implicitly assumed that Y has a conditional density given x for P-almost all x, but if it did not then the relative entropy $H(W \| P \times Q) = E_P[H(W(\cdot|X) \| Q)] = I(X;Y) + H(Q_Y \| Q)$

between the joint distribution W of (X, Y) and $(P \times Q)$ would be infinite, so the above bound would still hold. Therefore, in view of (10), the right-hand side of (13) is also a bound for $R(P, Q, D)$. □

Similarly to the Shannon lower bound for the rate-distortion function [21], the lower bound above turns out to be tight for several interesting special cases. To see this we first derive an upper bound for $R(P, Q, D)$. In (10) we can always pick

$$Y = X + Z_D, \tag{14}$$

where $Z_D \sim f_D$ is independent of X and it achieves the maximum entropy associated with ρ and D, i.e., $h(Z_D) = h(f_D) = h_{\max}(D)$. For this choice $I(X;Y) = h(Y) - h(Z_D) = h(f_Y) - h_{\max}(D)$, where f_Y is the density of $Y = X + Z_D$. Therefore,

$$\begin{aligned} R(P, Q, D) &\leq h(f_Y) - h_{\max}(D) + H(Q_Y \| Q) && (15) \\ &= E_{f_Y}[-\log f_Q(Y)] - h_{\max}(D), && (16) \end{aligned}$$

where $Y = X + Z_D$.

Now if $f_Q(y)$ is continuous near $y_{\max} = \arg\max_y f_Q(y)$, and f_Y is concentrated around $y_{\max}$, then $E_{f_Y}[-\log f_Q(Y)] \approx \log(1/Q_{\max})$, and the two bounds (12) and (16) are close. Since the lower bound (12) is independent of P, closeness of the bounds would imply that $R(P, Q, D)$ is only weakly dependent on the source distribution. For example, for a uniform codebook distribution $Q \sim U[-K, K]$ we have $Q_{\max} = 1/2K$ so the lower bound is $R(P, Q, D) \geq \log(2K) - h_{\max}(D)$. On the other hand, if K is large enough so that $f_Y(y) = 0$ for $|y| > K$, then $E_{f_Y}[-\log f_Q(Y)] = \log(2K)$, and the lower bound is tight. See also [31, Lemma 1]

More generally, suppose that the codebook distribution $Q = Q_s$ has an exponential density of the form

$$f_s(y) = B_s \ \exp(-sg(y)) \ , \ s > 0, \tag{17}$$

where g is any suitable (nonnegative) function with $g(0) = 0$. Gaussian codebooks correspond to the case $g(y) = y^2$, while uniform codebooks correspond to a "well-shaped" g. Moreover, for any "nice" g (as stated rigorously in the next lemma), as $s \to 0$ the exponential density $f_s(y)$ tends to be *locally uniform* relative to the Y of (14). This explains the following asymptotic characterization of $R(P, Q, D)$.

Lemma TIGHT. Asymptotically Flat Codebooks: For a difference distortion measure and an exponential codebook distribution $Q = Q_s$ of the form (17), if $E[g(X + Z_D)]$ is finite, then the lower bound (12) becomes tight as $s \to 0$:

$$R(P, Q_s, D) = \log(1/B_s) - h_{\max}(D) + o(1).$$

Proof. For $Q = Q_s$ we have $Q_{\max} = B_s$ and $y_{\max} = 0$, so the lower bound (12) is equal to $\log(1/B_s) - h_{\max}(D)$. On the other hand, $-\log f_Q(y) = \log(1/B_s) + sg(y)$, so the upper bound (16) is equal to $\log(1/B_s) + sE[g(X + Z_D)] - h_{\max}(D)$, which approaches the lower bound as $s \to 0$. □

An interesting consequence of Lemma TIGHT is that, like for uniform codebooks, for very flat codebook distributions Q_s the rate-function $R(P, Q_s, D)$ is *almost independent* of the source distribution P. In particular, for a Gaussian $Q \sim N(0, \tau^2)$ and any source P,

$$R(P, Q, D) = \frac{1}{2}\log(2\pi\tau^2) - h_{\max}(D) + o(1)$$

as $\tau \to \infty$. If ρ = squared error, then $h_{\max}(D) = (1/2)\log(2\pi eD)$, and we obtain $R(P, Q, D) = (1/2)\log(\tau^2/eD) + o(1)$ as in (8).

Another consequence of the asymptotic tightness of the upper bound (16) is that an additive maximum entropy noise channel of the form (14) asymptotically achieves the minimizations (2) and (10). This observation extends the asymptotic additive Gaussian noise channel characterization of $I_m(P\|Q^*_{PQD}, D)$ in Proposition 1. We state this result in the following proposition and prove it in Appendix D.

Proposition 2. Entropy-Coding Gain for Approximately Flat Codebooks: Let ρ be a difference distortion measure such that the maximum entropy $h_{\max}(D)$ defined in (13) exists and is strictly monotonically increasing with D. Let Q_s be any exponential codebook distribution of the form (17) such that $E[g(X + Z_D)]$ is finite. Then:

(i) The measures $Q^*_{PQ_sD}$ converge to $P * f_D$ as $s \to 0$, where $P * f_D$ is the distribution of $Y = X + Z_D$, and where Z_D is the random variable achieving $h_{\max}(D)$. This convergence is in the sense that the density of $Q^*_{PQ_sD}$ converges to the density of $Y = X + Z_D$.

(ii) The upper bound $I_m(P\|Q^*_{PQ_sD}, D)$ to the rate achieved by the entropy-coding scheme satisfies

$$\lim_{s\to 0} I_m(P\|Q^*_{PQ_sD}, D) = I(X; X + Z_D) \quad \text{bits/symbol},$$

where $X \sim P$, and Z_D is the maximum entropy random variable achieving $h_{\max}(D)$.

(iii) If the source $\boldsymbol{X}$ is memoryless, then as $s \to 0$ the rate achieved by the entropy-coding scheme is no greater than

$$R(D) + C^* \quad \text{bits/symbol}, \tag{18}$$

where the universal constant C^* is defined as

$$C^* = C^*(\rho, D) = \sup_{U:\, E\rho(U)\le D} I(U; U + Z_D).$$

Note that C^* is an upper bound for the "min-max capacity" defined in [35] in connection with the Wyner-Ziv problem. In particular, for an rth-power distortion measure $\rho(\hat{x} - x) = |\hat{x} - x|^r$, we have $0.5 \le C^* \le 1$ bit, and for $r = 2$ we have $C^* = 1/2$ bit in accordance with Proposition 1 (iii); see [32].

We note that the codebook distribution Q_s in Proposition 2 is independent of the source, of the distortion measure, *and* of the distortion level. This implies a strong robustness property for a memoryless codebook drawn from Q_s: For sufficiently small s, such a codebook is universal (in

the sense of the bounded loss in (18)) for any source and any distortion criterion admissible by the proposition. Thus, we may imagine that we first fix the codebook; then we select the desired distortion criterion to generate the D-balls; and finally we let the source induce the codeword distribution which determines the entropy code.

4 An Almost Sure Conditional Limit Theorem

4.1 Preliminaries

As before, we assume throughout this section that $\boldsymbol{X}$ is a stationary ergodic source and Q is an arbitrary codebook distribution with $0 \leq D_{\min} < D_{\text{av}} < \infty$. Also we fix a distortion level $D \in (D_{\min}, D_{\text{av}})$. Under these conditions, from [8, Theorem 2] we know that $R(P, Q, D)$ is finite and strictly positive, and that the infimum in its definition in (2) is always achieved by some joint distribution W^* with $E_{W^*}[\rho(X, Y)] = D$. Moreover, since the set of W over which the infimum is taken is convex, from [6] we know that this W^* is the unique minimizer.

Alternatively, $R(P, Q, D)$ can be expressed as

$$(\ln 2)R(P, Q, D) = \sup_{\lambda \leq 0}[\lambda D - \Lambda(\lambda)] = \lambda^* D - \Lambda(\lambda^*), \tag{19}$$

where

$$\Lambda(\lambda) \stackrel{\triangle}{=} E_P\left[\ln E_Q\left(e^{\lambda\rho(X,Y)}\right)\right],$$

and λ^* is the unique negative real number with

$$\Lambda'(\lambda^*) = D; \tag{20}$$

where prime denotes derivative, cf. [8, Theorem 2].

Let $Q' = W_Y^*$ denote the Y-marginal of W^*. From the above discussion and from equations (10), (3) and (4) it follows that the infimum in (4) is uniquely achieved by Q'. That is

$$Q' = W_Y^* = Q_{P,Q,D}^*$$

where as before $P = P_1$ denotes the first-order marginal of the source distribution. Now let $\widehat{Q}_n$ denote the empirical distribution induced by the matching codeword $Z_1^n \stackrel{\triangle}{=} Y_1^n(N_n)$ on $\hat{A}$:

$$\widehat{Q}_n \stackrel{\triangle}{=} \hat{P}_{Y_1^n(N_n)} \stackrel{\triangle}{=} \frac{1}{n}\sum_{i=1}^{n} \delta_{Z_i}.$$

4.2 Results

Our first result says that, when n is large, $\widehat{Q}_n \approx Q'$ with high probability. This generalizes and strengthens the "favorite-type theorem" of [34].

Theorem 2. Empirical Distribution of the Matching Codeword ("Favorite Type"):

(i) For every (measurable) $E \subset \hat{A}$, any $\delta > 0$, and $\mathbb{P}$-almost every source realization x_1^∞, as $n \to \infty$ we have:

$$\Pr\left\{|\widehat{Q}_n(E) - Q'(E)| > \delta \,\middle|\, X_1^n = x_1^n\right\} \to 0 \text{ exponentially fast.}$$

(ii) With probability one,

$$\widehat{Q}_n \Rightarrow Q', \quad \text{as } n \to \infty,$$

where '$\Rightarrow$' denotes weak convergence of probability measures.

As we will see below, Theorem 2 is a consequence of the following generalization of the conditional limit theorem (see, e.g., [5, Ch.12] for the standard form of the conditional limit theorem).

Theorem 3. Almost Sure Conditional Limit Theorem: Let X_1^n and Y_1^n be two independent random vectors with distributions P_n and Q^n, respectively, and write $\hat{P}_n = \hat{P}_{Y_1^n}$ for the empirical distribution of Y_1^n. For every (measurable) $E \subset \hat{A}$, any $\delta > 0$, and for $\mathbb{P}$-almost every realization x_1^∞, as $n \to \infty$ we have:

$$\Pr\left\{|\hat{P}_n(E) - Q'(E)| > \delta \,\middle|\, \rho_n(X_1^n, Y_1^n) \le D \text{ and } X_1^n = x_1^n\right\} \to 0 \text{ exponentially fast.}$$

Note that since $\rho_n(X_1^n, Y_1^n) \le D$ is a rare event, the conditional probability in Theorem 3 would have been different without conditioning on a $\mathbb{P}$-almost sure realization x_1^∞. We first deduce Theorem 2 from Theorem 3 and then we give the proof of Theorem 3.

Proof of Theorem 2. Observe that

$$\begin{aligned}
&\Pr\left\{|\widehat{Q}_n(E) - Q'(E)| > \delta \,\middle|\, X_1^n\right\} \\
&\quad \stackrel{(a)}{=} \sum_{k\ge 1} \Pr\left\{|\widehat{Q}_n(E) - Q'(E)| > \delta \,\middle|\, N_n = k \text{ and } X_1^n\right\} \Pr\{N_n = k \,|\, X_1^n\} \\
&\quad \stackrel{(b)}{=} \sum_{k\ge 1} \Pr\left\{|\hat{P}_n(E) - Q'(E)| > \delta \,\middle|\, \rho_n(X_1^n, Y_1^n) \le D \text{ and } X_1^n\right\} \Pr\{N_n = k \,|\, X_1^n\} \\
&\quad \stackrel{(c)}{=} \Pr\left\{|\hat{P}_n(E) - Q'(E)| > \delta \,\middle|\, \rho_n(X_1^n, Y_1^n) \le D \text{ and } X_1^n\right\},
\end{aligned}$$

where (a) and (c) follow from the fact that $N_n < \infty$, eventually with probability one (by Theorem A); and (b) follows from the observation that, due to the codewords' independence, the random variables $\rho_n(X_1^n, Y_1^n(k)),\ k = 1, 2, \ldots$ are conditionally independent given X_1^n. This implies that, given N_n, the distribution of the matching codeword is exactly the same as the distribution of Y_1^n conditioned on the event $\{\rho_n(X_1^n, Y_1^n) \le D\}$.

This together with Theorem 3 proves (i). From (i) and the Borel-Cantelli lemma we conclude that, for any measurable $E \subset \hat{A}$,

$$\widehat{Q}_n(E) \to Q'(E), \quad \text{as } n \to \infty, \text{ w.p.1.}$$

Since $\hat{A}$ is a Polish space, there exists a countable convergence-determining class $\mathcal{E} = \{E_i\} \subset \hat{\mathcal{A}}$.[2] Therefore, with probability one we have that

$$\widehat{Q}_n(E_i) \to Q'(E_i), \quad \text{as } n \to \infty, \text{ for all } i,$$

and this implies (ii). □

Proof of Theorem 3. The probability in Theorem 3 can be expanded as

$$\begin{aligned}
&\Pr\left\{|\hat{P}_n(E) - Q'(E)| > \delta \text{ and } \rho_n(X_1^n, Y_1^n) \le D \,\middle|\, X_1^n\right\} \Big/ \Pr\{\rho_n(X_1^n, Y_1^n) \le D \,|\, X_1^n\} \\
&\quad = \Pr\left\{\hat{P}_n(E) < Q'(E) - \delta \text{ and } \rho_n(X_1^n, Y_1^n) \le D \,\middle|\, X_1^n\right\} \Big/ Q^n(B(X_1^n, D)) \\
&\qquad + \Pr\left\{\hat{P}_n(E) > Q'(E) + \delta \text{ and } \rho_n(X_1^n, Y_1^n) \le D \,\middle|\, X_1^n\right\} \Big/ Q^n(B(X_1^n, D)).
\end{aligned}$$

We only treat the first of the two terms above; the second one can be dealt with similarly.

If $Q'(E) \le \delta$ there is nothing to prove, so let us assume that $Q'(E) > \delta$. In view of Theorem B, it suffices to show that

$$\limsup_{n\to\infty} \frac{1}{n} \log \Pr\left\{\hat{P}_n(E) < Q'(E) - \delta \text{ and } \rho_n(X_1^n, Y_1^n) \le D \,\middle|\, X_1^n\right\} < -R(P, Q, D) \quad \text{w.p.1.} \tag{21}$$

This will be proved by an application of the Gärtner-Ellis theorem. Toward that end, choose and fix an arbitrary realization x_1^∞ of $\boldsymbol{X}$, and define a sequence of random vectors $\{\xi_n\}$ in $\mathbb{R}^2$ as

$$\xi_n \stackrel{\triangle}{=} \left(\rho_n(x_1^n, Y_1^n), \hat{P}_n(E)\right) = \frac{1}{n}\sum_{i=1}^n \left(\rho(x_i, Y_i), \mathbb{I}_E(Y_i)\right),$$

where the random variables $\{Y_i\}$ are as in the statement of the theorem. Let $\Theta_n(\lambda)$ denote the log-moment generating function of ξ_n,

$$\Theta_n(\lambda) \stackrel{\triangle}{=} \ln E_{Q^n}\left[\exp\{\lambda_1 \rho_n(x_1^n, Y_1^n) + \lambda_2 \hat{P}_n(E)\}\right], \qquad \lambda = (\lambda_1, \lambda_2) \in (-\infty, 0]^2.$$

By the ergodic theorem we have, for $\mathbb{P}$-almost every realization x_1^∞,

$$\begin{aligned}
\lim_{n\to\infty} \frac{1}{n}\Theta_n(n\lambda) &= \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^n \ln E_Q\left[\exp\{\lambda_1 \rho(x_i, Y) + \lambda_2 \mathbb{I}_E(Y)\}\right] \\
&= \Theta(\lambda) \stackrel{\triangle}{=} E_P\left\{\ln E_Q\left[\exp\{\lambda_1 \rho(X, Y) + \lambda_2 \mathbb{I}_E(Y)\}\right]\right\},
\end{aligned} \tag{22}$$

where, by Jensen's inequality, the limiting log-moment generating function $\Theta(\lambda)$ satisfies $-\infty < (\lambda_1 D_{\text{av}} + \lambda_2) \le \Theta(\lambda) \le 0$. It is easy to check (using the dominated convergence theorem) that $\Theta(\lambda)$ is differentiable, with partial derivatives

$$\begin{aligned}
\frac{\partial\Theta}{\partial\lambda_1} &= E_{W^{(\lambda)}}[\rho(X, Y)], \\
\frac{\partial\Theta}{\partial\lambda_2} &= W_Y^{(\lambda)}(E),
\end{aligned}$$

[2]For example, let B be a countable dense subset of $\hat{A}$, and take $\mathcal{E}$ to be the collection of all open balls with rational radii centered at the points of B, together with all finite intersections of such balls. Then $\mathcal{E}$ is countable (by construction) and it is easy to check that [2, Theorem 2.3] applies, verifying that $\mathcal{E}$ is convergence-determining.

where $W^{(\lambda)}$ is the probability measure defined on $A \times \hat{A}$ by

$$\frac{dW^{(\lambda)}(x,y)}{d(P \times Q)} \stackrel{\triangle}{=} \frac{\exp\{\lambda_1 \rho(x,y) + \lambda_2 \mathbb{I}_E(y)\}}{E_Q\left[\exp\{\lambda_1 \rho(x,Y) + \lambda_2 \mathbb{I}_E(Y)\}\right]},$$

and $W_Y^{(\lambda)}$ is the Y-marginal of $W^{(\lambda)}$.

Note that $\Theta(\lambda)$ is a convex function, and define its convex dual,

$$\Theta^*(z) = \sup_{\lambda_1 \le 0,\, \lambda_2 \le 0} [(z,\lambda) - \Theta(\lambda)], \quad z = (z_1, z_2) \in \mathbb{R}^2,$$

where (z,λ) denotes the usual Euclidean inner product $(z,\lambda) = z_1\lambda_1 + z_2\lambda_2$.

In view of (22) and of the above discussion, we can apply the Gärtner-Ellis theorem [9, Theorem 2.3.6] to conclude that, with probability one, the probabilities in (21) satisfy

$$\limsup_{n\to\infty} \frac{1}{n} \ln \Pr\left\{\hat{P}_n(E) < Q'(E) - \delta \ \text{ and } \ \rho_n(X_1^n, Y_1^n) \le D \,\middle|\, X_1^n\right\} \le - \inf_{z_1 \in [0,D],\, z_2 \in [0,q]} \Theta^*(z_1, z_2),$$

with probability one, where we write $q = Q'(E) - \delta > 0$. From its definition, it is obvious that $\Theta^*(z_1, z_2)$ is nonincreasing in each of its coordinates. Therefore, to prove (21) and conclude the proof of the theorem it suffices to show (recall (19)) that:

$$\Theta^*(D, q) > (\ln 2) R(P, Q, D) = \lambda^* D - \Lambda(\lambda^*). \tag{23}$$

To prove (23) we consider

$$g(\lambda_1, \lambda_2) \stackrel{\triangle}{=} \lambda_1 D + \lambda_2 q - \Theta(\lambda_1, \lambda_2).$$

Using the dominated convergence theorem as before we can differentiate g with respect to λ_2 to get that for all $(\lambda_1, \lambda_2) \in (-\infty, 0]^2$,

$$\frac{\partial g(\lambda_1, \lambda_2)}{\partial \lambda_2} = q - E_{P\times Q}\left[\mathbb{I}_E(Y) \frac{\exp\{\lambda_1 \rho(X,Y) + \lambda_2 \mathbb{I}_E(Y)\}}{E_Q[\exp\{\lambda_1 \rho(X,Y') + \lambda_2 \mathbb{I}_E(Y')\}]}\right] = q - W_Y^{(\lambda)}(E),$$

where at the endpoint $\lambda_2 = 0$ this is understood as the corresponding right-derivative. Now, if this derivative evaluated at $(\lambda_1, \lambda_2) = (\lambda^*, 0)$ is nonnegative, i.e., if

$$W_Y^{(\lambda^*, 0)}(E) \le q < Q'(E), \tag{24}$$

then, after some simple algebra, $W^{(\lambda^*,0)}$ is easily seen to satisfy

$$(\ln 2) H(W^{(\lambda^*,0)} \| P \times Q) = \lambda^* D - \Lambda(\lambda^*) = (\ln 2) R(P, Q, D),$$

and also

$$E_{W^{(\lambda^*,0)}}[\rho(X,Y)] = \Lambda'(\lambda^*) = D$$

(see (20)). Since as we remarked above $R(P,Q,D)$ is uniquely achieved by W^*, we must have that $W^{(\lambda^*,0)} = W^*$, and, in particular, $W_Y^{(\lambda^*,0)} = W_Y^* = Q'$. But this contradicts (24).

Therefore, it must be the case that

$$\left.\frac{\partial g(\lambda_1,\lambda_2)}{\partial \lambda_2}\right|_{(\lambda^*,0)} < 0,$$

which means that by taking $\lambda_2 = \lambda'$ slightly negative, we can make $g(\lambda^*, \lambda')$ strictly larger than $g(\lambda^*, 0) = (\ln 2)R(P, Q, D)$. Hence

$$\Theta^*(D, q) = \sup_{\lambda_1,\,\lambda_2} g(\lambda_1, \lambda_2) \geq g(\lambda^*, \lambda') > (\ln 2)R(P, Q, D),$$

establishing (23) and thereby completing the proof. □

5 Entropy-Coding Performance

Before giving the proof of Theorem 1 we need to state four simple Lemmas that establish some of the technical properties we need in the proof.

5.1 Four Lemmas

Recall the notation and assumptions of Section 4.1. We begin with some preliminary lemmas.

Lemma 1:

$$\lim_{\delta \downarrow 0} \sup_{\{F_i\}} \inf_{\widetilde{Q}:|\widetilde{Q}-Q'|<\delta} H(\widetilde{Q}\|Q) = H(Q'\|Q),$$

where the supremum is taken over all finite partitions $\{F_1, F_2, \ldots, F_k\}$ of $\hat{A}$, and for any such partition the infimum is over all probability measures $\widetilde{Q}$ on $\hat{A}$ such that $|\widetilde{Q}(F_i) - Q'(F_i)| < \delta$ for all i.

Proof. Since Q' is always among the measures over which the infimum is taken, we obviously have that the above left-hand side is no larger than the right-hand side.

To prove the corresponding lower bound let $\epsilon > 0$ arbitrary, and choose a finite partition $\mathcal{P} = \{F_1, \ldots, F_k\}$ such that $H(Q'_{\mathcal{P}}\|Q_{\mathcal{P}}) > H(Q'\|Q) - \epsilon$, where for any measure μ and any partition $\mathcal{P}$ we write $\mu_{\mathcal{P}}$ for the corresponding discrete measure which assigns probability $\mu(F_i)$ to each i in the alphabet $\{1, 2, \ldots, k\}$. The fact that this is possible follows from [25, Chapter 2] and the fact that $H(Q'\|Q) \leq R(P, Q, D) < \infty$. Without loss of generality we assume that $Q(F_i) \neq 0$ for all $F_i \in \mathcal{P}$.

By the uniform continuity of relative entropy on a finite alphabet, we can choose $\delta_0 > 0$ small enough so that

$$H(\widetilde{Q}_{\mathcal{P}}\|Q_{\mathcal{P}}) > H(Q'_{\mathcal{P}}\|Q_{\mathcal{P}}) - \epsilon$$

for all probability measures $\widetilde{Q}$ on $\{1, \ldots, k\}$ with $|\widetilde{Q}_{\mathcal{P}}(F_i) - Q_{\mathcal{P}}(F_i)| < \delta_0$ for all i. Then by the data processing inequality for relative entropy, for all $\delta < \delta_0$,

$$\inf_{\widetilde{Q}:|\widetilde{Q}-Q'|<\delta_0} H(\widetilde{Q}_{\mathcal{P}}\|Q_{\mathcal{P}}) \geq H(Q'_{\mathcal{P}}\|Q_{\mathcal{P}}) - \epsilon \geq H(Q'\|Q) - 2\epsilon,$$

and since $\epsilon > 0$ was arbitrary, we are done. □

The following lemma contains a simple observation based on Lemma 1; it is stated here without proof.

Lemma 2: For any $\epsilon > 0$ there is a $\delta > 0$ and a finite partition $\mathcal{P} = \{F_1, \ldots, F_k\}$ of $\hat{A}$ such that

$$R(\delta) \stackrel{\triangle}{=} \inf_{\widetilde{Q}:|\widetilde{Q}-Q'|<\delta} H(\widetilde{Q}\|Q) > H(Q'\|Q) - \epsilon.$$

Writing $I(\delta) \stackrel{\triangle}{=} R(P, Q, D) - R(\delta)$, we have

$$I_m(P\|Q', D) \leq I(\delta) \leq I_m(P\|Q', D) + \epsilon.$$

Given a $\delta > 0$ and a finite partition $\mathcal{P} = \{F_1, \ldots, F_k\}$, let B_δ denote the set of probability measures on $\hat{A}$ that are δ-close to Q' on the sets F_i:

$$B_\delta \stackrel{\triangle}{=} \{\widetilde{Q} \,:\, |\widetilde{Q}(F_i) - Q'(F_i)| < \delta \ \text{ for all } i\}. \tag{25}$$

Lemma 3. Let X_1^n and Y_1^n be two independent random vectors with distributions P_n and Q^n, respectively, and write $\hat{P}_n$ for the empirical distribution of Y_1^n. Suppose a $\delta > 0$ and a finite partition $\mathcal{P} = \{F_1, \ldots, F_k\}$ of $\hat{A}$ are given, and write

$$p_n \stackrel{\triangle}{=} p_n(x_1^n, \delta) \stackrel{\triangle}{=} \Pr\{\rho_n(x_1^n, Y_1^n) \leq D \,|\, \hat{P}_n \in B_\delta\}.$$

Then we have:

$$\lim_{n\to\infty} -\frac{1}{n} \log p_n(X_1^n, \delta) = I(\delta) \quad \text{w.p.1.}$$

Proof. We expand

$$p_n(X_1^n, \delta) \;=\; \Pr\{\rho_n(X_1^n, Y_1^n) \leq D \text{ and } \hat{P}_n \in B_\delta\} \Big/ \Pr\{\hat{P}_n \in B_\delta\}$$

and evaluate the exponential behavior of the numerator and denominator separately. First, by Sanov's theorem [7],

$$\lim_{n\to\infty} \frac{1}{n} \log \Pr\{\hat{P}_n \in B_\delta\} = -R(\delta). \tag{26}$$

We will also show that

$$\lim_{n\to\infty} \frac{1}{n} \log \Pr\{\rho_n(X_1^n, Y_1^n) \leq D \text{ and } \hat{P}_n \in B_\delta\} = -R(P, Q, D) \quad \text{w.p.1,} \tag{27}$$

and, recalling that $R(P, Q, D) - R(\delta) = I(\delta)$, this will complete the proof.

First note that, since

$$\Pr\{\rho_n(X_1^n, Y_1^n) \leq D \text{ and } \hat{P}_n \in B_\delta\} \leq Q^n(B(X_1^n, D)),$$

Theorem B implies

$$\limsup_{n\to\infty} \frac{1}{n} \log \Pr\{\rho_n(X_1^n, Y_1^n) \le D \text{ and } \hat{P}_n \in B_\delta\} \le -R(P, Q, D) \quad \text{w.p.1.} \tag{28}$$

For the corresponding lower bound we employ the Gärtner-Ellis theorem, much as in the proof of Theorem 3. Let x_1^∞ be some fixed realization of $\boldsymbol{X}$, and define a sequence of random vectors $\{\zeta_n\}$ in $\mathbb{R}^{k+1}$ by

$$\zeta_n \stackrel{\triangle}{=} \Big(\rho_n(x_1^n, Y_1^n), \hat{P}_n(F_1), \ldots, \hat{P}_n(F_k)\Big) = \frac{1}{n}\sum_{i=1}^{n} \left(\rho(x_i, Y_i), \mathbb{I}_{F_1}(Y_i), \ldots, \mathbb{I}_{F_k}(Y_i)\right).$$

Let $\Gamma_n(\lambda)$ denote the log-moment generating function of ζ_n,

$$\Gamma_n(\lambda) \stackrel{\triangle}{=} \ln E_{Q^n}\left[\exp\left\{\lambda_0 \rho_n(x_1^n, Y_1^n) + \sum_{i=1}^{k} \lambda_i \hat{P}_n(F_i)\right\}\right], \qquad \lambda = (\lambda_0, \ldots, \lambda_k) \in (-\infty, 0]^{k+1}.$$

As before, the ergodic theorem says that for $\mathbb{P}$-almost every realization x_1^∞,

$$\lim_{n\to\infty} \frac{1}{n}\Gamma_n(n\lambda) \;=\; \Gamma(\lambda) \;\stackrel{\triangle}{=}\; E_P\left\{\ln E_Q\left[\exp\{\lambda_0 \rho(X, Y) + \sum_{i=1}^{k} \lambda_i \mathbb{I}_{F_i}(Y)\}\right]\right\},$$

where, by Jensen's inequality, the limiting moment generating function $\Gamma(\lambda)$ satisfies $-\infty < (\lambda_0 D_{\text{av}} + \sum_{i=1}^{k} \lambda_i) \le \Gamma(\lambda) \le 0$. Once again, a routine application of the dominated convergence theorem verifies that $\Gamma(\lambda)$ is differentiable, so the Gärtner-Ellis theorem [9, Theorem 2.3.6] yields that

$$\liminf_{n\to\infty} \frac{1}{n} \ln \Pr\{\rho_n(X_1^n, Y_1^n) \le D \text{ and } \hat{P}_n \in B_\delta\} \ge -\inf_z \Gamma^*(z) \quad \text{w.p.1,}$$

where the infimum is taken over all $z \in \mathbb{R}^{k+1}$ with $z_0 \in [0, D)$ and $|z_i - Q'(F_i)| < \delta$, $i = 1, \ldots, k$, and Γ^* denotes the convex dual of Γ,

$$\Gamma^*(z) \stackrel{\triangle}{=} = \sup_{\lambda \in (-\infty, 0]^{k+1}} [(z, \lambda) - \Gamma(\lambda)], \quad z \in \mathbb{R}^{k+1},$$

where (z, λ) is the Euclidean inner product in $\mathbb{R}^{k+1}$, $(z, \lambda) = \sum_{i=0}^{k} z_i \lambda_i$. Therefore, with probability one,

$$\liminf_{n\to\infty} \frac{1}{n} \log \Pr\{\rho_n(X_1^n, Y_1^n) \le D \text{ and } \hat{P}_n \in B_\delta\} \ge -(\log e)\Gamma^*(D, Q'(F_1), \ldots, Q'(F_k)), \tag{29}$$

where we used the (easily verifiable) fact that Γ^* is continuous in $z_0 \in (D_{\min}, D_{\text{av}})$. Finally we claim that

$$(\log e)\Gamma^*(D, Q'(F_1), \ldots, Q'(F_k)) \le R(P, Q, D). \tag{30}$$

Combining (30) with (29) and with the upper bound (28) proves (27) and the Lemma.

So it only remains to establish (30). Note that for any $x \in A$, any measurable function $\phi : \hat{A} \to \mathbb{R}$ which is bounded above, and any measure W on $A \times \hat{A}$,

$$(\ln 2) H(W(\cdot|x) \| Q(\cdot)) \geq \int \phi(y) W(dy|x) - \ln E_Q[e^{\phi(Y)}]. \tag{31}$$

[This can be proved in exactly the same way as the corresponding statement in the proof of [8, Theorem 2].] Take $W = W^*$ to be the achieving measure in the definition of $R(P, Q, D)$, and let $\phi(y) = \lambda_0 \rho(x, y) + \sum_{i=1}^k \lambda_i \mathbb{I}_{F_i}(y)$ for some $\lambda \in (-\infty, 0]^{k+1}$. Applying (31) and integrating both sides with respect to P we get that

$$\begin{aligned} R(P, Q, D) &\geq (\log e) \left[\lambda_0 E_{W^*}[\rho(X, Y)] + \sum_{i=1}^k \lambda_i Q'(F_i) - \Gamma(\lambda) \right] \\ &\geq (\log e) \left[(\lambda, (D, Q'(F_1), \ldots, Q'(F_k))) - \Gamma(\lambda) \right], \end{aligned}$$

and since this holds for any $\lambda \in (-\infty, 0]^{k+1}$ we have established (30), as required. □

Finally we give a simple general result on the asymptotic behavior of the entropy of sequences of random variables. Its proof is in Appendix B.

Lemma 4: Let $\xi_1, \xi_2, \ldots$ be a sequence of random variables, and $A_1, A_2, \ldots$be a sequence of events with $\Pr\{A_n\} \to 1$, as $n \to \infty$. Assume that $\xi_n \in \{1, 2, 3, \ldots, 2^{n\beta}\}$, for all n and some $\beta < \infty$. Then,

$$\lim_{n \to \infty} \frac{1}{n} \Big[H(\xi_n) - H(\xi_n | \mathbb{I}_{A_n} = 1) \Big] = 0.$$

5.2 Proof of Theorem 1

Let $\epsilon > 0$ be arbitrary, and choose a $\delta > 0$ and a finite partition $\mathcal{P} = \{F_1, \ldots, F_k\}$ of $\hat{A}$ as in Lemma 2. With B_δ as in (25) and with $\hat{P}_{Y_1^n(k)}$ denoting the empirical distribution of the kth codeword, for $k = 1, 2, \ldots$ we define:

$$J_k = \begin{cases} \mathbb{I}_{\{\hat{P}_{Y_1^n(k)} \in B_\delta\}} & \text{if } 1 \leq k \leq \lfloor 2^{nb} \rfloor, \\ 1 & \text{if } k \geq \lfloor 2^{nb} \rfloor + 1. \end{cases}$$

Now we consider two sub-codebooks of $\boldsymbol{C}_n$,

$$\begin{aligned} \boldsymbol{C}_n^{(0)} &\stackrel{\triangle}{=} \Big\{ Y_1^n(k) \ : \ J_k = 0, \ 1 \leq k \leq \lfloor 2^{nb} \rfloor \Big\} \\ \boldsymbol{C}_n^{(1)} &\stackrel{\triangle}{=} \Big\{ Y_1^n(k) \ : \ J_k = 1, \ 1 \leq k \leq \lfloor 2^{nb} \rfloor \Big\}. \end{aligned}$$

Also, for $j = 0, 1$, let $N_n^{(j)}$ be the index of the first codeword in $\boldsymbol{C}_n^{(j)}$ that matches X_1^n with distortion D or less, and let $M_n^{(j)}$ be the index of the position of $Y_1^n(N_n^{(j)})$ in $\boldsymbol{C}_n^{(j)}$. If no match is found in $\boldsymbol{C}_n^{(j)}$, then let $N_n^{(j)} = M_n^{(j)} = \lfloor 2^{nb} \rfloor + 1$. From these definitions it immediately follows that, given $\boldsymbol{C}_n$, the value of N_n' and the values of $(M_n^{(J_{N_n})}, J_{N_n})$ are in a one-to-one correspondence.

To bound $E[\mathcal{L}_n(X_1^n)]$ we begin by expanding

$$\begin{aligned} H(N_n'|\boldsymbol{C}_n) &= H(M_n^{(J_{N_n})}, J_{N_n}|\boldsymbol{C}_n) \\ &\leq 1 + H(M_n^{(J_{N_n})}|J_{N_n}) \\ &= 1 + \Pr\{J_{N_n}=0\}H(M_n^{(J_{N_n})}|J_{N_n}=0) + \Pr\{J_{N_n}=1\}H(M_n^{(J_{N_n})}|J_{N_n}=1) \\ &\leq 1 + \Pr\{J_{N_n}=0\}\log(\lfloor 2^{nb}\rfloor+1) + H(M_n^{(1)}|J_{N_n}=1), \end{aligned}$$

therefore, in view of (7)

$$\begin{aligned} \limsup_{n\to\infty}\frac{1}{n}E[\mathcal{L}_n(X_1^n)] &\leq \limsup_{n\to\infty}\frac{1}{n}H(N_n'|\boldsymbol{C}_n) \\ &\leq \limsup_{n\to\infty}\left[\frac{1}{n}H(M_n^{(1)}|J_{N_n}=1) + \frac{1}{n}\log(2^{nb}+1)\Pr\{\widehat{Q}_n\notin B_\delta\}\right] \\ &\overset{(a)}{=} \limsup_{n\to\infty}\frac{1}{n}H(M_n^{(1)}|J_{N_n}=1) \\ &\overset{(b)}{=} \limsup_{n\to\infty}\frac{1}{n}H(M_n^{(1)}), \end{aligned} \tag{32}$$

where (a) follows from Theorem 2, and (b) follows from Lemma 4. Now recall the definition of the (conditional) probability $p_n = p_n(X_1^n,\delta)$ from Lemma 3, and for arbitrary $\Delta>0$ let $E_n^{(\Delta)}$ denote a quantized version of the exponent of p_n:

$$E_n^{(\Delta)} \stackrel{\triangle}{=} \Delta\left\lceil -\frac{1}{n\Delta}\log p_n\right\rceil.$$

Note that, given a source string x_1^n, the random variable $M_n^{(1)}$ has a "truncated Geometric" distribution, which we denote by $\mathrm{Geom}^*(p_n)$; formally, for a parameter $q\in(0,1)$,

$$\Pr\{\mathrm{Geom}^*(q)=k\} = \begin{cases} q(1-q)^{k-1} & \text{if } 1\leq k\leq\lfloor 2^{nb}\rfloor \\ (1-q)^{k-1} & \text{if } k=1+\lfloor 2^{nb}\rfloor \\ 0 & \text{otherwise.}\end{cases}$$

A useful bound on the entropy of a mixture of $\mathrm{Geom}^*(q)$ distributions is given in the following lemma; its proof is given in Appendix C.

Lemma 5: If the distribution of the random variable Z is a mixture of $\mathrm{Geom}^*(q)$ distributions with $q\in[\alpha,\beta]$, then $H(Z)\leq\log(e/\alpha)$.

Now observe that, given $E_n^{(\Delta)}$ is equal to some e_n, the conditional distribution of $M_n^{(1)}$ is a mixture of $\mathrm{Geom}^*(q)$ distributions, for parameter values $q\geq 2^{-ne_n}$. Therefore, by Lemma 5,

$$H(M_n^{(1)}|E_n^{(\Delta)}=e_n)\leq ne_n+\log(e),$$

and hence, with probability one,

$$\limsup_{n\to\infty}\frac{1}{n}H(M_n^{(1)}|E_n^{(\Delta)}=e_n) \quad\leq\quad \limsup_{n\to\infty}E_n^{(\Delta)}$$

$$
\begin{aligned}
&\leq \limsup_{n\to\infty} -\frac{1}{n}\log p_n(X_1^n,\delta)+\Delta\\
&\overset{(a)}{\leq} I(\delta)+\Delta\\
&\overset{(b)}{\leq} I_m(P\|Q',D)+\epsilon+\Delta,
\end{aligned}
$$

where (a) follows by Lemma 2 and (b) by Lemma 3. Since for all n large enough $(1/n)H(M_n^{(1)}|E_n^{(\Delta)}=e_n)\leq(b+1/n)\leq(b+1)$ with probability one, we can apply Fatou's lemma to get,

$$
\begin{aligned}
\limsup_{n\to\infty}\frac{1}{n}H(M_n^{(1)}|E_n^{(\Delta)}) &= \limsup_{n\to\infty} E\Big\{\frac{1}{n}H(M_n^{(1)}|E_n^{(\Delta)}=e_n)\Big\}\\
&\leq E\Big\{\limsup_{n\to\infty}\frac{1}{n}H(M_n^{(1)}|E_n^{(\Delta)}=e_n)\Big\}\\
&\leq I_m(P\|Q',D)+\epsilon+\Delta. \qquad (33)
\end{aligned}
$$

Next we will show that

$$
\limsup_{n\to\infty}\frac{1}{n}H(M_n^{(1)})\leq\limsup_{n\to\infty}\frac{1}{n}H(M_n^{(1)}|E_n^{(\Delta)}). \qquad (34)
$$

Since $\epsilon>o$ and $\Delta>0$ were arbitrary, combining this with (32) and (33) will complete the proof of the theorem.

Turning to the proof of (34), we take $\epsilon'>0$ arbitrary, and define

$$
I_n=\mathbb{I}_{\{E_n^{(\Delta)}\leq I_m(P\|Q',D)+\Delta+\epsilon+\epsilon'\}}, \qquad (35)
$$

and observe that, by Lemmas 2 and 3,

$$
\Pr\{I_n=1\}\geq\Pr\{-\tfrac{1}{n}\log p_n(X_1^n,\delta)\leq I(\delta)+\epsilon'\}\to 1 \quad \text{as } n\to\infty. \qquad (36)
$$

We can expand

$$
\begin{aligned}
H(M_n^{(1)}|E_n^{(\Delta)}) &\geq H(M_n^{(1)}|I_n,E_n^{(\Delta)})\\
&\geq \Pr\{I_n=1\}\,H(M_n^{(1)}|I_n=1,E_n^{(\Delta)})\\
&= \Pr\{I_n=1\}\Big[H(M_n^{(1)},E_n^{(\Delta)}|I_n=1)-H(E_n^{(\Delta)}|I_n=1)\Big]\\
&\geq \Pr\{I_n=1\}\Big[H(M_n^{(1)},E_n^{(\Delta)}|I_n=1)-K\Big]\\
&\overset{(a)}{\geq} H(M_n^{(1)}|I_n=1)+(\Pr\{I_n=1\}-1)\log(\lceil 2^{nb}\rceil+1)+K\Pr\{I_n=1\}
\end{aligned}
$$

where

$$
K=\log\left(\frac{I_m(P\|Q',D)+\Delta+\epsilon+\epsilon'}{\Delta}+1\right)+1
$$

and where (a) follows since by (35) the number of values of $E_n^{(\Delta)}$ is at most 2^K. Therefore, from this and (36) we have

$$
\limsup_{n\to\infty}\frac{1}{n}H(M_n^{(1)}|I_n=1)\leq\limsup_{n\to\infty}\frac{1}{n}H(M_n^{(1)}|E_n^{(\Delta)}).
$$

Finally, (36) allows us to apply Lemma 4 and thus conclude that (34) holds, completing the proof of the theorem. □

6 Tighter Bounds for Sources with Memory

In this section we discuss the following questions: (1) Under what conditions is the bound in Theorem 1 tight? (2) When it is not tight, what is the actual performance of the entropy-coded scheme? *Only heuristic arguments and proof outlines are given.*

To gain some intuition, we first consider the extreme case of *lossless* compression of a finite-alphabet, stationary ergodic source $\boldsymbol{X}$, that is $D = 0$ relative to Hamming distortion. Let Q be a codebook distribution on $\hat{A} = A$ with $Q(a) > 0$ for all $a \in A$, and let $\boldsymbol{C}_n$ be a *memoryless* random codebook with distribution Q. Then all possible n-strings from A^n will appear infinitely often in $\boldsymbol{C}_n$, and the matching codeword will always be identical to the source string. Moreover, this also implies that Q^*_{PQD}, the limiting first-order empirical distribution of the matching codeword (Theorem 2), will simply be the first-order marginal P of the source. It is therefore an immediate consequence of the AEP (Asymptotic Equipartition Property [5]) that the asymptotic rate achieved by this scheme will be exactly equal to the entropy rate $H(\boldsymbol{X})$ of the source $\boldsymbol{X}$. In this case it is easy to calculate the bound given in Theorem 1 explicitly to get that, at $D = 0$,

$$I_m(P\|Q^*_{PQD}, D) = I_m(P\|P, D) = H(X_1).$$

The above argument indicates that the bound in Theorem 1 will be tight *if and only if* the source $\boldsymbol{X}$ is memoryless. Indeed, for finite-alphabet memoryless sources this was shown to be the case in [31].

Now let us turn to general alphabet sources and positive distortions $D > 0$. For general stationary sources, it is well-known that the rate-distortion function decreases as the memory increases, so it is natural to expect that the rate achieved by any "good" coding scheme will also take advantage of such dependencies.

For the *naive coding scheme* Theorem A immediately shows that, if the codebook distribution is memoryless, then memory in the source does *not* affect the rate achieved. Formally, this observation is reflected in the identity [29],

$$R(P_k, Q^k, kD) = kR(P_1, Q, D), \quad \text{for all } k. \tag{37}$$

In contrast, in the *entropy-coded* case we expect that memory in the source *does* affect the rate. For example, the above heuristic argument shows that for $D = 0$ entropy-coding achieves the entropy-rate of the source, and not just $H(X_1)$. But since the bound $I_m(P_1\|Q^*_{P_1QD}, D)$ in Theorem 1 only depends on the first-order marginal P_1, memory in the source does not affect it and therefore it cannot be tight in this case. As we discuss next we can establish tighter bounds showing that, in fact, entropy-coding the index *does* take advantage of memory. This more desirable behavior is reflected in the multi-dimensional behavior of the lower mutual information (LMI) function: In contrast to (37), whenever $P_k \neq P^k$,

$$I_m(P_k\|Q^*_{P_k,Q^k,kD}, kD) < kI_m(P_1\|Q^*_{P_1,Q,D}, D),$$

where $Q^*_{P_k,Q^k,kD}$ is the (unique) k-dimensional distribution $\tilde{Q}_k$ that achieves $R(P_k, Q^k, kD)$ in (4). Therefore, the LMI decreases due to memory in the source even if the codebook is memoryless.

Example. Universal Gaussian Codebooks. To appreciate this decrease in LMI due to memory in the source, consider a memoryless Gaussian codebook with large variance τ^2 and squared error distortion measure, as in Section 3.1. For a real-valued source $\boldsymbol{X}$ with zero mean and finite variance σ^2, a straightforward k-dimensional extension of Proposition 1 gives,

$$\lim_{\tau^2\to\infty} I_m(P_k\|Q^*_{P_k,Q^k,D}, kD) = I(X_1^k; X_1^k + Z_D^{(k)}) \tag{38}$$

where $X_1^k \sim P_k$, and $Z_D^{(k)}$ denotes an i.i.d. $N(0, D)$ random vector independent of X_1^k. Since $Z_D^{(k)}$ has a density we can write

$$I(X_1^k; X_1^k + Z_D^{(k)}) = h(X_1^k + Z_D^{(k)}) - kh(Z_D^{(1)}).$$

If X_1^k also has a density, then for small D this expression becomes $h(X_1^k) - kh(Z_D^{(1)}) + o(1)$, where $o(1) \to 0$ as $D \to 0$; see [21]. It follows that, for small D,

$$I_m(P_1\|Q^*_{P_1,Q,D}, D) - \frac{1}{k} I_m(P_k\|Q^*_{P_k,Q^k,kD}, kD) = h(P_1) - \frac{1}{k}h(P^k) + o(1)$$

where $\lim_{D\to 0}\lim_{\tau^2\to\infty} o(1) = 0$. That is, for small D the LMI rate reduction relative to the marginal case is asymptotically

$$h(P_1) - \frac{1}{k}h(P^k) \to I(X_1; X_{-\infty}^0)$$

as $k \to \infty$. This is the information the past has about the present, which for some sources can be very large.

In general, the tighter bounds on the rate of the entropy-coded scheme follow from natural k-dimensional extensions of the results in Theorems 1 and 2. As before, we restrict attention to memoryless random codebooks with arbitrary distribution Q, single-letter distortion measures, and stationary ergodic sources. As in [18], the extension to the case with memory follows by considering k-blocks of super-symbols in the source and the codebook, but the technicalities, although not particularly insightful, are very involved. The reader will have probably been convinced of this by seeing the proofs in the simpler memoryless case. Under the same assumptions as in Theorems 1 and 2 (and perhaps under mild additional regularity conditions on the source as in [18]), we obtain the following analogs.

Theorem 1-k.: For any k we have:

$$\limsup_{n\to\infty} \frac{1}{n} H(N_n'|\boldsymbol{C}_n) \;\le\; \frac{1}{k} I_m(P_k\|Q^*_{P_k,Q^k,kD}, kD) \quad \text{bits/symbol.}$$

Theorem 2-k.: Let $\widehat{Q}_n^{(k)}$ denote the kth order empirical distribution induced by the matching codeword $Z_1^n \stackrel{\triangle}{=} Y_1^n(N_n)$ on $\hat{A}$. With probability one, for any k we have:

$$\widehat{Q}_n^{(k)} \Rightarrow Q^*_{P_k,Q^k,kD}, \quad \text{as } n \to \infty.$$

Following standard arguments used in the analysis of the rate-distortion function [1], we can define the *LMI rate* as

$$I_m(\mathbb{P}\|Q^*_{\mathbb{P},Q^\infty,D}, D) \stackrel{\triangle}{=} \inf_k \frac{1}{k} I_m(P_k\|Q^*_{P_k,Q^k,kD}, kD) = \lim_{k\to\infty} \frac{1}{k} I_m(P_k\|Q^*_{P_k,Q^k,kD}, kD).$$

It follows that the best upper bound on the index entropy is $I_m(\mathbb{P}\|Q^*_{\mathbb{P},Q^\infty,D}, D)$, and we conjecture that this bound is in fact tight, i.e.,

$$\lim_{n\to\infty} \frac{1}{n} H(N_n'|\boldsymbol{C}_n) = I_m(\mathbb{P}\|Q^*_{\mathbb{P},Q^\infty,D}, D).$$

Example. Universal Gaussian Codebooks. Returning to the special case considered in the last example, if $Q \sim N(0, \tau^2)$, then as the codebook variance $\tau^2 \to \infty$ the rate $I_m(\mathbb{P}\|Q^*_{\mathbb{P},Q^\infty,D}, D)$ achieved by the entropy-coded scheme satisfies

$$\lim_{\tau^2\to\infty} I_m(\mathbb{P}\|Q^*_{\mathbb{P},Q^\infty,D}, D) = I(\boldsymbol{X}; \boldsymbol{X} + \boldsymbol{Z}_D), \tag{39}$$

where $\boldsymbol{Z}_D$ is a white Gaussian process with variance D and $I(\boldsymbol{X}; \boldsymbol{X} + \boldsymbol{Z}_D)$ is the mutual information rate between $\boldsymbol{X}$ and $\boldsymbol{Z}_D$. (A simple heuristic calculation indicating that (39) holds is to divide (38) by k and take k to infinity.) Combining this with the fact that $I(\boldsymbol{X}; \boldsymbol{X} + \boldsymbol{Z}_D) \le R(D) + 1/2$, [37, 32], where $R(D)$ is the rate-distortion function of the entire process (not just the first-order rate-distortion function), we get that, as $\tau^2 \to \infty$, the rate achieved by the entropy-coded scheme is no worse than $R(D) + 1/2$ bits/symbol.

Acknowledgments

We thank Zhiyi Chi for some interesting technical conversations.

Appendix

A Proof Outline of Proposition 1

Although the result of the proposition can be obtained by little more than elementary calculus, the calculations are rather lengthy so we only give an outline of the proof here.

First observe that, in the notation of Section 4.1, the log-moment generating $\Lambda(\lambda)$ can be evaluated explicitly,

$$\Lambda(\lambda) = -\frac{1}{2}\ln(1 - 2\lambda\tau^2) + \frac{\lambda\sigma^2}{1 - 2\lambda\tau^2},$$

and (20) can be solved to show that the optimizing value of $\lambda = \lambda^*$ is given by

$$\lambda = \frac{2D - \tau^2 - \Delta}{4\Delta\tau^2},$$

where

$$\Delta = \sqrt{\tau^2 + 4\sigma^2 D},$$

so that, as $\tau^2 \to \infty$ we have

$$\lambda = -\frac{1}{2D} + \frac{1}{2\tau^2} + O(\tau^{-4}) \tag{40}$$

$$\lambda^2 = \frac{1}{4D^2} + O(\tau^{-4}). \tag{41}$$

From the proof of [8, Theorem 2] it follows that the joint distribution W^* that achieves the infimum in the definition of $R(P, Q, D)$ is given by

$$\frac{dW^*}{d(P \times Q)}(x, y) = \frac{e^{\lambda(x-y)^2}}{E_P\left[e^{\lambda(X-y)^2}\right]},$$

and, as discussed in Section 4.1, Q' is the Y-marginal W^*_Y of W^*. Therefore, writing $\phi_a(y)$ for the $N(0, a)$ density, the density $f_{Q'}(y)$ of Q' with respect to Lebesgue measure $m(dy)$ can be expressed as

$$f_{Q'}(y) = \frac{dQ'}{dm}(y) = E_P\left[\frac{e^{\lambda(X'-y)^2}}{E_P\left[e^{\lambda(X-y)^2}\right]}\right] \phi_{\tau^2}(y),$$

where X and X' denote two independent random variables with the same distribution P. Evaluating the denominator explicitly and rearranging terms, the above expression becomes

$$f_{Q'}(y) = \sqrt{\frac{\frac{1}{2\tau^2} - \lambda}{\pi}}\, E_P\left[\exp\left\{-\left(y\sqrt{\frac{1}{2\tau^2} - \lambda} - X\sqrt{\frac{\lambda^2}{\frac{1}{2\tau^2} - \lambda}}\right)^2\right\}\right], \tag{42}$$

and recalling (40) and (41), we can let $\tau^2 \to \infty$ to get that

$$\begin{aligned} \frac{dW^*}{d(P \times Q)}(x, y)\, \frac{dQ}{dm}(y) &= \sqrt{\frac{\frac{1}{2\tau^2} - \lambda}{\pi}}\, \exp\left\{-\left(y\sqrt{\frac{1}{2\tau^2} - \lambda} - x\sqrt{\frac{\lambda^2}{\frac{1}{2\tau^2} - \lambda}}\right)^2\right\} \\ &\to \phi_D(y - x) \quad \text{as } \tau^2 \to \infty. \end{aligned} \tag{43}$$

Invoking the dominated convergence theorem we can conclude that

$$f_{Q'}(y) \to E_P[\phi_D(y - X)] \quad \text{as } \tau^2 \to \infty, \tag{44}$$

as claimed. This proves (a).

For part (b) note that, from (4) and the above discussion it follows that

$$\begin{aligned} I_m(P\|Q, D) &= R(P, Q', D) - H(Q'\|Q) \\ &= H(W^*\|P \times Q') - H(Q'\|Q) \\ &= H(W^*\|P \times Q) \\ &= \int \left[\frac{dW^*}{d(P \times Q)}(x, y)\frac{dQ}{dm}(y) \ln\left\{\frac{dW^*}{d(P \times Q)}(x, y)\frac{dQ}{dm}(y)\left(\frac{dQ'}{dm}(y)\right)^{-1}\right\}\right] dP(x)dm(y). \end{aligned}$$

Using the expressions for the densities in (42) and (43), recalling that $dQ/dm(y) = \phi_{\tau^2}(y)$, and applying the convergence bounds in (40), (41) and (44), it is straightforward to show that the last integrand in $[\cdots]$ above converges to

$$\phi_D(y-x)\ln\left(\frac{\phi_D(y-x)}{E_P[\phi_D(y-X)]}\right).$$

Writing V_D for the joint distribution of the random variables $(X, X+Z_D)$ as in the statement of the proposition, and Q_D for the distribution of $(X+Z_D)$, the above expression can be rewritten as

$$\frac{dV_D}{d(P\times m)}(x,y)\ln\left\{\frac{dV_D}{d(P\times Q_D)}(x,y)\right\}.$$

Finally, using (40) and (41) to justify the use of the dominated convergence theorem we get that also the integrals converge, i.e., as $\tau^2\to\infty$,

$$\begin{aligned}
I_m(P\|Q,D) &= \int\left[\frac{dW^*}{d(P\times Q)}(x,y)\frac{dQ}{dm}(y)\ln\left\{\frac{dW^*}{d(P\times Q)}(x,y)\frac{dQ}{dm}(y)\left(\frac{dQ'}{dm}(y)\right)^{-1}\right\}\right]dP(x)dm(y)\\
&\to \int\frac{dV_D}{d(P\times m)}(x,y)\ln\left\{\frac{dV_D}{d(P\times Q_D)}(x,y)\right\}dP(x)dm(y)\\
&= H(V_D\|P\times Q_D)\\
&= I(X;X+Z_D),
\end{aligned}$$

proving (b).

Finally part (c) follows from the well-known fact [37, 32, 35] that the rate-distortion function $R(D)$ of a real-valued memoryless source (with respect to squared error distortion) is bounded below by $I(X;X+Z_D)-1/2$. □

B Proof of Lemma 4

First observe that

$$H(\xi_n)\le H(\xi_n,\mathbb{I}_{A_n}) = H(\xi_n)+H(\mathbb{I}_{A_n}|\xi_n)\le H(\xi_n)+1,$$

so that

$$\lim_{n\to\infty}\frac{1}{n}\Big[H(\xi_n)-H(\xi_n,\mathbb{I}_{A_n})\Big]=0. \tag{45}$$

Also we can expand

$$\begin{aligned}
\frac{1}{n}H(\xi_n,\mathbb{I}_{A_n}) &= \frac{1}{n}H(\mathbb{I}_{A_n})+\frac{1}{n}H(\xi_n|\mathbb{I}_{A_n}=1)\Pr\{A_n\}+\frac{1}{n}H(\xi_n|\mathbb{I}_{A_n}=0)(1-\Pr\{A_n\})\\
&= O(\tfrac{1}{n})+\frac{1}{n}H(\xi_n|\mathbb{I}_{A_n}=1)+(1-\Pr\{A_n\})\frac{1}{n}\Big[H(\xi_n|\mathbb{I}_{A_n}=0)-H(\xi_n|\mathbb{I}_{A_n}=1)\Big]\\
&= \frac{1}{n}H(\xi_n|\mathbb{I}_{A_n}=1)+O(\tfrac{1}{n})+(1-\Pr\{A_n\})\frac{1}{n}\log(2^{n\beta}),
\end{aligned}$$

i.e.,

$$\lim_{n\to\infty}\frac{1}{n}\Big[H(\xi_n,\mathbb{I}_{A_n})-H(\xi_n|\mathbb{I}_{A_n}=1)\Big]=0. \tag{46}$$

Combining (45) with (46) proves the lemma. □

C Proof of Lemma 5

It is well-known that the Geometric (non-truncated) distribution has the largest entropy among all nonnegative variables with a given mean. Now, it is easy to verify that if Z_q is a Geom(q) random variable [i.e., if $Z_q = k$ with probability $q(1-q)^{k-1}$, for $k = 1, 2, \ldots$], then $E[Z_q] = 1/q$, and

$$H(Z_q) = \log(1/q) - \frac{1-q}{q}\log(1-q) \leq \log(e/q).$$

Thus, since the mean of a mixture of truncated Geometric distributions is smaller than or equal to the mean of the Geometric distribution with the smallest parameter q (in our case, α), we obtain that $H(Z) \leq H(Z_\alpha) \leq \log(e/\alpha)$. □

D Proof of Proposition 2

We first introduce some convenient notation. Let W_s^* denote the joint distribution minimizing $H(W\|P \times Q_s)$ in (2) (i.e., achieving $R(P, Q_s, D)$), and let Q_s^* denote its induced Y-marginal. In our previous notation, $Q_s^* = Q_{P,Q_s,D}$ and $I(W_s^*) = I_m(P||Q_{P,Q_s,D}, D)$, where $I(W)$ is the mutual information associated with a joint distribution W. Let $W_{\rm add}$ denote the joint input-output distribution associated with the additive noise channel $Y = X + Z_D$ in (14). In this notation, part (ii) of the proposition amounts to

$$I_m(P\|Q_{P,Q_s,D}, D) = I(W_s^*) \to I(W_{\rm add})$$

as $s \to \infty$.

Clearly $W_{\rm add}$ is in the set of admissible distributions W in (2). Moreover, by Lemma TIGHT we know that $H(W_{\rm add}\|P \times Q_s) - R(P, Q_s, D) \to 0$ as $s \to \infty$. That is, $W_{\rm add}$ asymptotically achieves the minimum of $H(W||P \times Q_s)$. Since, by (2) and the Pythagorean theorem for divergence [5] for any admissible W

$$H(W\|P \times Q_s) \geq R(P, Q_s, D) + H(W_s^*\|W),$$

we conclude that

$$H(W_s^*\|W_{\rm add}) \to 0. \tag{47}$$

Since relative entropy dominates L_1 distance, this implies that the density of W_s^* converges to that of $W_{\rm add}$, *a fortiori* proving part (i).

Part (i) and the semi-continuity of the divergence [25] imply that

$$\liminf_{s\to 0} I(W_s^*) \geq I(W_{\rm add}). \tag{48}$$

On the other hand, it follows from (47) and the chain rule for relative entropy, [5], that the conditional relative entropy $H(W_s^*\|W_{\rm add}|P) \to 0$ as $s \to \infty$. Alternatively, if we expand $H(W_s^*\|W_{\rm add}|P)$ in terms of differential entropy, this becomes

$$\lim_{s\to 0} [h(Z_D) - h(Y_s^*|X)] = 0$$

where (X, Y_s^*) are jointly distributed as W_s^*, Z_D is a maximum entropy random variable independent of X, and $E\rho(Y_s^* - X) = D$ (equality here is due to the strict monotonicity of $h_{\max}(D)$ as a function of D). Therefore,

$$h(Y_s^*|X) \to h(Z_D) = h_{\max}(D). \tag{49}$$

We can also conclude from (47) that the relative entropy between the outputs vanishes

$$\lim_{s\to 0} H(Q_s^* \| Q_Y) = 0,$$

where Q_Y denotes the distribution of $Y = X + Z_D$. Again by the semi-continuity of the divergence this implies that

$$\limsup_{s\to 0} h(Y_s^*) \le h(X + Z_D); \tag{50}$$

see [21]. Combining (49) and (50) we thus have

$$\begin{aligned} I(W_s^*) &= I(X; Y_s^*) && (51)\\ &= h(Y_s^*) - h(Y_s^*|X) && (52)\\ &\le h(X + Z_D) - h(Z_D) + o(1) && (53)\\ &= I(X; X + Z_D) + o(1) && (54) \end{aligned}$$

where $o(1) \to 0$ as $s \to \infty$. This, together with (48), proves part (ii).

Finally, part (iii) follows from [35]. □

References

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

[2] P. Billingsley. *Convergence of Probability Measures*. John Wiley & Sons Inc., New York, second edition, 1999.

[3] J. A. Bucklew. Two results on the asymptotic performance of quantizers. *IEEE Trans. Inform. Theory*, 30:341–348, March 1984.

[4] P.A. Chou, T. Lookabaugh, and R.M. Gray. Entropy constrained vector quantization. *IEEE Trans. Acoustics, Speech and Signal Processing*, 37:31–42, 1989.

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

[6] I. Csiszár. I-divergence geometry of probability distributions and minimization problems. *Ann. Probab.*, 3:146–158, 1975.

[7] I. Csiszár. Sanov property, generalized I-projection and a conditional limit theorem. *Ann. Probab.*, 12(3):768–793, 1984.

[8] A. Dembo and I. Kontoyiannis. Source coding, large deviations, and approximate pattern matching. *IEEE Trans. Inform. Theory*, 48:1590–1615, June 2002.

[9] A. Dembo and O. Zeitouni. *Large Deviations Techniques And Applications*. Springer-Verlag, New York, second edition, 1998.

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

[11] A. Gersho. Asymptotically optimal block quantization. *IEEE Trans. Inform. Theory*, 25:373–380, 1979.

[12] H. Gish and N.J. Pierce. Asymptotically efficient quantization. *IEEE Trans. Inform. Theory*, 14:676–683, 1968.

[13] R.M. Gray and T. Linder. Mismatch in high-rate entropy-constrained vector quantization. *IEEE Trans. Inform. Theory*, 49:1204–1218, 2003.

[14] M. Gutman. On universal quantization with various distortion measures. *IEEE Trans. Inform. Theory*, 33: Jan. 1987.

[15] A. Kanlis. *Compression and Transmission of Information at Multiple Resolutions*. PhD thesis, Dept. of Electrical and Computer Engineering, University of Maryland at College Park, 1998.

[16] A. Kanlis, P. Narayan, and B. Rimoldi. On three topics for a course in information theory. In *Statistical Methods in Imaging, Medicine, Optics, and Communication, Edt. J.A. O'Sullivan.* Springer Verlag, 2001.

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

[18] I. Kontoyiannis. Sphere-covering, measure concentration, and source coding. *IEEE Trans. Inform. Theory*, 47:1544–1552, May 2001.

[19] I. Kontoyiannis and J. Zhang. Arbitrary source models and Bayesian codebooks in rate-distortion theory. *IEEE Trans. Inform. Theory*, 48:2276–2290, 2002.

[20] A. Lapidoth. On the role of mismatch in rate distortion theory. *IEEE Trans. Inform. Theory*, 43(1):38–47, 1997.

[21] T. Linder and R. Zamir. On the asymptotic tightness of the Shannon lower bound. *IEEE Trans. Inform. Theory*, 40(6):2026–2031, 1994.

[22] T. Lookabaugh and R.M. Gray. High resolution quantization theory and the vector quantizer advantage. *IEEE Trans. Inform. Theory*, 35:1020–1033, 1989.

[23] D.L. Neuhoff. Source coding strategies: Simple quantizers vs. simple noiseless codes. *Proceedings 1986 Conf. on Information Sciences and Systems*, 1:267–271, 1986.

[24] J.T. Pinkston. *Encoding Independent Sample Information Sources.* PhD Thesis, MIT, 1967.

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

[26] D.J. Sakrison. The rate distortion function for a class of sources. *Information and Control*, 15:165–195, 1969.

[27] D.J. Sakrison. The rate of a class of random processes. *IEEE Trans. Inform. Theory*, 16:10–16, 1970.

[28] Y. Steinberg and M. Gutman. An algorithm for source coding subject to a fidelity criterion, based upon string matching. *IEEE Trans. Inform. Theory*, 39(3):877–886, 1993.

[29] 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.

[30] E.-h. Yang and J.C. Kieffer. Simple universal lossy data compression schemes derived from the Lempel-Ziv algorithm. *IEEE Trans. Inform. Theory*, 42:239–245, Jan. 1996.

[31] R. Zamir. The index entropy of a mismatched codebook. *IEEE Trans. Inform. Theory*, 48(2):523–528, 2002.

[32] R. Zamir and M. Feder. On universal quantization by randomized uniform/lattice quantizers. *IEEE Trans. Inform. Theory*, 38:428–436, 1992.

[33] R. Zamir and M. Feder. Information rates for pre/post filtered dithered quantizers. *IEEE Trans. Inform. Theory*, 42:1340–1353, 1996.

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

[35] R. Zamir. The rate loss in the Wyner-Ziv problem. *IEEE Trans. Inform. Theory*, 42:2073-2084, Nov. 1996.

[36] Z. Zhang and V.K. Wei. An on-line universal lossy data compression algorithm by continuous codebook refinement – Part I: Basic results. *IEEE Trans. Inform. Theory*, 42(3):803–821, 1996.

[37] J. Ziv. On universal quantization. *IEEE Trans. Inform. Theory*, 31(3):344–347, 1985.