Causal Coding of Stationary Sources and Individual Sequences with High Resolution

Tamás Linder Ram Zamir

September 15, 2003
Submitted to: *IEEE Transactions on Information Theory*

Abstract

In a causal source coding system the reconstruction of the present source sample is restricted to be a function of the present and past source samples, while the code stream itself may be non-causal and have variable rate. Neuhoff and Gilbert showed that for memoryless sources, optimum performance among all causal source codes is achieved by time-sharing at most two memoryless codes (quantizers) followed by entropy coding. In this work, we extend Neuhoff and Gilbert's result in the limit of small distortion (high resolution) to two new settings. First we show that at high resolution an optimal causal code for a stationary source with finite differential entropy rate consists of a uniform quantizer followed by a (sequence) entropy coder. This implies that the price of causality at high resolution is approximately 0.254 bits, i.e., the space-filling loss of the uniform quantizer. Then we consider individual sequences and introduce a deterministic analogue of differential entropy, which we call "Lempel-Ziv differential entropy." We show that for any bounded individual sequence with finite Lempel-Ziv differential entropy rate, optimum high-resolution performance among all finite-memory variable-rate causal codes is achieved by dithered scalar uniform quantization followed by Lempel-Ziv coding. As a by-product, we also prove an individual-sequence version of the Shannon lower bound.

Index Terms: Causal source codes, differential entropy, finite-memory codes, individual sequences, Lempel-Ziv complexity, stationary sources, uniform quantizer.

T. Linder is with the Department of Mathematics and Statistics, Queen's University, Kingston, Ontario, Canada K7L 3N6 (email: `linder@mast.queensu.ca`). R. Zamir is with the Department of Electrical Engineering - Systems, Tel-Aviv University, Tel-Aviv, 69978, Israel (email: `zamir@eng.tau.ac.il`).

This research was supported in part by the Natural Sciences and Engineering Research Council (NSERC) of Canada.

1 Introduction

The performance gap between vector and scalar quantization is a basic figure of interest in lossy data compression. On the one extreme, scalar quantizers are the most easy-to-implement and commonly used source coding devices. On the other extreme, vector quantizers of unbounded dimension yield the rate-distortion function, $R(D)$, the minimum rate theoretically attainable by coding the source with distortion D [1]. The performance gain resulting from going to higher quantization dimensions is attributed to three factors in the quantization literature: ability to exploit memory in the source, ability to shape the quantizer codebook, and existence of better space-filling quantization cells (see the illuminating paper by Lookabaugh and Gray [2].) If the quantizer output sequence is "entropy coded," (encoded with a variable-rate lossless code), then most of the gain due to the first two factors can be achieved even with *scalar* quantization. In fact, in the limit of small distortion ($D \to 0$), known as "high resolution conditions," the rate loss of an optimum entropy-coded quantizer (ECQ) with respect to the rate-distortion function is due solely to the quantizer's space-filling (in)efficiency. By a classic result of Gish and Pierce [3], a uniform quantizer is approximately an optimum scalar ECQ at high resolution, and hence the rate loss of scalar quantization is asymptotically the space-filling loss of a cubic cell; i.e., $(1/2)\log_2\big(2\pi e/12\big) \approx 0.254$ bits per sample (assuming the squared error distortion measure).

The popularity of scalar quantizers is due not only to their very simple structure, but also to the fact that (fixed-rate) scalar quantizers have no encoding delay. However, scalar (memoryless) quantizers form only a special subclass of codes having zero delay, which, in general, can also have memory. It is an interesting and challenging problem to determine how much (if any) of the advantage offered by vector quantization can be realized with codes that introduce no additional delay, but allow the encoder output to depend also on the *past* samples of the source. For memoryless sources, Ericson [4] and Gaarder and Slepian [5, 6] showed that optimal performance among (fixed-rate) zero-delay codes is achieved by optimal scalar quantization, and thus zero-delay coding of memoryless sources does not offer any of the advantages of vector quantization. For sources with memory, the problem in general is still unresolved and only partial results are known (see, e.g., [6, 7]). Zero-delay codes [8] and limited-delay codes [9] have also been investigated in the individual-sequence setting. Recently, source coding exponents for zero-delay, finite-memory coding of memoryless sources have been derived by Merhav and Kontoyiannis [10].

In the context of entropy-coded quantization, the problem is also complicated by the fact that with entropy coding the overall system delay cannot be strictly zero. Neuhoff and Gilbert [11] proposed an alternative model, called "causal source coding," which ignores delays created by the variable-rate coding of the quantizer output. In a causal source code the reconstruction of the present source sample depends only on the present and the past source samples, but the decoder can generate the reconstruction with arbitrary delay. The minimum coding rate achievable with distortion D by such systems is denoted $r_c(D)$. With this definition, Neuhoff and Gilbert were able to show that for *memoryless* sources causal source coding cannot achieve any of the vector quantization advantages. Specifically, as described in detail in the Section 3, the optimum causal source coder times-shares at most two entropy-coded scalar quantizers. In essence, this result implies that by looking into the source's past one cannot create multidimensional cells that have better space-filling properties than the cubic cell. In the limit of high resolution, the loss of causality $r_c(D) - R(D)$ is therefore the same as the space-filling loss of the scalar ECQ; i.e., approximately 0.254 bits per sample.

When trying to extend Neuhoff and Gilbert's result to sources with memory, one encounters a substantial difficulty: due to the dependence between consecutive source samples, the quantized current and past samples become the "context" for quantizing the next sample. The optimization of such a system requires the little-understood optimal design of the quantization function over the entire (correlated) sequence.

In this paper we extend Neuhoff and Gilbert's result for two new settings under high resolution conditions. Intuitively, the high resolution assumption allows us to circumvent the difficulty outlined above because the finely-quantized past samples effectively provide an *unquantized* context for entropy coding. The first setting we consider is that of probabilistic stationary sources. Assuming the squared error distortion measure, we prove an asymptotic lower bound on the performance of causal coding of stationary sources with finite differential entropy rate, and show that an entropy-coded uniform scalar quantizer asymptotically achieves this bound. Hence, just as in the memoryless case, the rate loss in causal coding is asymptotically the space-filling loss of the cubic cell.

The second setting is inspired by Ziv and Lempel's model of coding an "individual sequence" using a finite-state machine [12, 13]. We consider encoding a deterministic bounded sequence of real numbers using a finite-resolution, finite-memory causal coder followed by a finite-state lossless encoder. We prove an asymptotic converse theorem for the performance of such systems. The resulting lower bound is given in terms of a new quantity, called the "Lempel-Ziv differential entropy rate," which, in the context of deterministic sequences and complexity-constrained encoders, plays a role similar to Shannon's differential entropy rate. We show via a direct coding theorem that a *dithered* uniform scalar quantizer ([14, 15]) combined with a finite-state lossless coder achieves the lower bound of the converse theorem. We also derive an individual-sequence version of the Shannon lower bound [1] to the rate-distortion function in which the Lempel-Ziv differential entropy rate replaces the Shannon differential entropy rate. This bound implies that the loss of causality for individual sequences is the same as in the probabilistic setting.

The paper is organized as follows. After reviewing some notation and definitions in Section 2, we derive the converse and direct coding theorems for causal coding of probabilistic stationary sources in Section 3. In Section 4, causal coding of deterministic sequences is studied. In Section 4.1, we introduce the notion of Lempel-Ziv differential entropy rate and present a result which characterizes individual sequences for which this quantity is finite. The converse and direct coding theorems for causal coding of individual sequences are given in Section 4.2. We prove the Shannon lower bound for individual sequences in Section 4.3. Section 5 concludes the paper. Some of the more technical proofs are relegated to the Appendix.

2 Preliminaries

For any sequence of random variables $\{X_n\}_{n\in I}$, where I is either the set of integers or the set of positive integers, and for any $n \geq m$, the segment (vector) $(X_m, X_{m+1}, \dots, X_n)$ will be denoted by X_m^n. We allow m and n to be infinite; for example, we write $X_{-\infty}^{\infty}$ for the entire sequence $\{X_n\}_{n=-\infty}^{\infty}$. A similar convention applies to deterministic sequences which are usually denoted by lower case letters.

The entropy of an n-dimensional discrete random vector X_1^n with values in the countable set A is defined by

$$H(X_1^n) \triangleq -\sum_{x\in A} \Pr(X_1^n = x)\log \Pr(X_1^n = x)$$

where log denotes base-2 logarithm. If the distribution of the real random vector X_1^n is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^n$, having probability density function (pdf) f, the *differential entropy* of X_1^n is

$$h(X_1^n) \triangleq -\int_{\mathbb{R}^n} f(x)\log f(x)\,dx$$

provided the integral exists. The normalized versions of $H(X_1^n)$ and $h(X_1^n)$ are denoted by $\bar{H}(X_1^n)$ and $\bar{h}(X_1^n)$, respectively; i.e.,

$$\bar{H}(X_1^n) \triangleq \frac{1}{n}H(X_1^n) \qquad \text{and} \qquad \bar{h}(X_1^n) \triangleq \frac{1}{n}h(X_1^n).$$

The *entropy rate* of a stationary sequence of discrete random variables X_1^∞ is

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

where the limit exists and is finite if $H(X_1)$ is finite [16].

If X_1^∞ is stationary and X_1^n has a pdf and finite differential entropy $h(X_1^n)$ for all $n \geq 1$, then the *differential entropy rate* of X_1^∞ is defined by

$$h(X_1^\infty) \triangleq \lim_{n\to\infty} \frac{1}{n} h(X_1^n).$$

By stationarity, the above limit is either finite or equal to $-\infty$. Entropy rates and differential entropy rates for double-sided stationary sequences are defined in a similar way. For example, $h(X_{-\infty}^\infty) \triangleq \lim_n \frac{1}{2n} h(X_{-n}^n)$.

Entropy rates will also be expressed via conditional entropies [16, 17]. For any discrete stationary $X_{-\infty}^\infty$,

$$H(X_{-\infty}^\infty) = \lim_{n\to\infty} H(X_1|X_{-n}^0) \triangleq H(X_1|X_{-\infty}^0)$$

while if $X_{-\infty}^\infty$ is stationary and has finite differential entropy rate,

$$h(X_{-\infty}^\infty) = \lim_{n\to\infty} h(X_1|X_{-n}^0) \triangleq h(X_1|X_{-\infty}^0).$$

A scalar quantizer is a measurable function $q : \mathbb{R} \to \mathbb{R}$ with a countable range. A scalar quantizer of particular interest is the uniform quantizer with step size $\Delta > 0$: Let Q_Δ denote the quantizer defined by $Q_\Delta(x) = k\Delta + \Delta/2$ if $k\Delta \leq x < (k+1)\Delta$, $k = 0, \pm 1, \pm 2\ldots.$ When Q_Δ is applied componentwise to X_1^n, we write $Q_\Delta(X_1^n)$ to denote the resulting (discrete) random vector $(Q_\Delta(X_1), \ldots, Q_\Delta(X_n))$. A similar convention holds for infinite sequences of random variables; e.g., $Q_\Delta(X_{-\infty}^\infty)$ denotes the sequence $\{Q_\Delta(X_n)\}_{n=-\infty}^\infty$.

The following result by Csiszár [18] shows a fundamental connection between the differential entropy of a random vector and the the asymptotic entropy of its uniformly quantized version.

Lemma 1 *Assume X_1^n is an n-vector of real random variables such that $H(Q_1(X_1^n)) < \infty$. If X_1^n has finite differential entropy, then*

$$\lim_{\Delta\to 0} [\bar{H}(Q_\Delta(X_1^n)) + \log \Delta] = \bar{h}(X_1^n). \tag{1}$$

It is also shown in [18] that the limit is equal to $-\infty$ if $h(X_1^n) = -\infty$ or X_1^n does not have a pdf. Since $h(X_1^n) \leq H(Q_1(X_1^n))$ by Jensen's inequality, we obtain that in case $H(Q_1(X_1^n))$ is finite,[1] X_1^n possesses a pdf and finite differential entropy if and only if the limit on the left-hand side of (1) is finite.

The following extension of Csiszár's result to stationary processes will play an important role in this paper.

[1] It is straightforward to show that $H(Q_1(X_1^n))$ is finite if and only if $H(Q_\Delta(X_1^n))$ is finite for all $\Delta > 0$.

Lemma 2 *If $X_{-\infty}^{\infty}$ is stationary, has finite differential entropy rate, and $H(Q_1(X_1)) < \infty$, then*

$$\lim_{\Delta \to 0} [H(Q_\Delta(X_{-\infty}^{\infty})) + \log \Delta] = h(X_{-\infty}^{\infty}). \tag{2}$$

The proof is given in Appendix A. Combined with the previous remark, the proof also implies that whenever $H(Q_1(X_1))$ is finite, the process has a finite differential entropy rate if and only is the limit on the left-hand side is finite.

3 Causal coding of stationary sources

Consider the following model for causal (non-anticipating) encoding a discrete-time real random process $X_{-\infty}^{\infty}$.[2] The encoder accepts the source sequence $\ldots, X_{-1}, X_0, X_1, X_2, \ldots$ and applies to it a sequence of *reproduction functions* $\{g_n\}_{n=1}^{\infty}$, where g_n maps $X_{-\infty}^{n}$ into the real-valued reproduction symbol

$$\hat{X}_n = g_n(X_{-\infty}^{n}), \quad n = 1, 2, \ldots$$

Each g_n is assumed to be a measurable function of one-sided infinite real sequences $x_{-\infty}^{n}$ and have a countable range (thus each $\hat{X}_n$ is a discrete random variable). The encoder losslessly encodes the reproduction sequence $\hat{X}_1, \hat{X}_2, \hat{X}_3, \ldots$ and thereby creates the variable-rate binary representation $Z_1, Z_2, Z_3 \ldots$. The decoder receives $Z_1, Z_2, Z_3, \ldots$ and losslessly decodes the reproduction sequence $\hat{X}_1, \hat{X}_2, \hat{X}_3, \ldots$. The code is called *causal* because the reproduction $\hat{X}_n$ depends only on the present and past source symbols $X_{-\infty}^{n}$. This means that all delays are due to the lossless coding part of the code. Note that although the encoder has access to the entire source sequence $X_{-\infty}^{\infty}$, only X_1^{∞} is to be represented and reproduced by the code.

The collection $\{g_n\}_{n=1}^{\infty}$ is called a *casual reproduction coder*. The distortion of the system is defined by the accumulated expected mean-squared error

$$d(\{g_n\}) \triangleq \limsup_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} E(X_i - \hat{X}_i)^2.$$

Note that the distortion is determined solely by the reproduction coder.

The rate of the code is measured by

$$\limsup_{n \to \infty} \frac{1}{n} E\left[L_n(X_{-\infty}^{\infty})\right]$$

[2] We follow the model introduced by Neuhoff and Gilbert [11]. They allowed general source and reproduction alphabets and an arbitrary single-letter distortion measure; we only consider the case of real sources and squared error distortion measure which is amenable to high-resolution analysis.

where $L_n(X_{-\infty}^{\infty})$ is the cumulative number of bits received by the encoder when it produces $\hat{X}_n$. Neuhoff and Gilbert [11] showed that the infimum of rates for all causal codes with a given reproduction coder $\{g_n\}$ is the *limsup entropy rate* of the reproduction process, defined by

$$\limsup_{n\to\infty} \frac{1}{n} H(\hat{X}_1^n)$$

where $\hat{X}_n = g_n(X_{-\infty}^n)$ for all $n \geq 1$. We follow [11] to define the rate of the system to be

$$r(\{g_n\}) \triangleq \limsup_{n\to\infty} \frac{1}{n} H(\hat{X}_1^n)$$

which makes the rate definition independent of the particular choice of the lossless code used in the scheme.

An important class of reproduction coders is the class of *sliding-block coders* (also called stationary or time-invariant coders). A causal sliding-block coder is characterized and denoted by a real function g of one-sided infinite sequences such that $\hat{X}_n = g(X_{-\infty}^n)$ for all $n \geq 1$. In this case, the distortion and rate are denoted, respectively, by $d(g)$ and $r(g)$. Note that if $X_{-\infty}^{\infty}$ is stationary, then $\hat{X}_1^{\infty}$ and $\{X_n, \hat{X}_n\}_{n=1}^{\infty}$ are both stationary. Thus $r(g)$ is equal to the (ordinary) entropy-rate of $\hat{X}_1^{\infty}$ and $d(g) = E(X_1 - \hat{X}_1)^2$. If $\hat{X}_n = q(X_n)$, $n = 1, 2, \ldots$, for a scalar quantizer q, then $g = q$ is called a *memoryless* reproduction coder, and $r(g)$ is given by the entropy rate of the stationary sequence $\{q(X_n)\}_{n=-\infty}^{\infty}$.

The optimal performance theoretically attainable (OPTA) with causal source codes is the minimum rate achievable when encoding the source $X_{-\infty}^{\infty}$ by any causal code with distortion D or less. Formally, for all $D > 0$ the causal OPTA function is defined by

$$r_c(D) \triangleq \inf_{\{g_n\}:\, d(\{g_n\}) \leq D} r(\{g_n\})$$

where the infimum is over all causal reproduction coders with distortion not exceeding D.

The main result of [11] shows that if $X_{-\infty}^{\infty}$ is *stationary and memoryless,* then

$$r_c(D) = \bar{r}_m(D)$$

where $\bar{r}_m(D)$ is the lower convex hull of the OPTA function, $r_m(D)$, for memoryless reproduction coders (scalar quantizers), given by

$$r_m(D) = \inf_{q:\, E(X-q(X))^2 \leq D} H(q(X)). \tag{3}$$

Here X is a random variable having the common distribution of the X_n and the infimum is over all scalar quantizers having squared distortion $E(X - q(X))^2 \leq D$. $r_m(D)$ is called the OPTA function for scalar entropy-constrained quantization of the memoryless source $X_{-\infty}^{\infty}$

[19, 20]. Any quantizer q such that $H(q(X)) = r_m(D)$ and $E(X - q(X))^2 \leq D$ is called an optimal quantizer.

Since any point on the graph of $\bar{r}_m(D)$ can be obtained as the convex combination of at most two points on the graph of $r_m(D)$, Neuhoff and Gilbert's result is equivalent to the statement that for memoryless sources, optimum performance among all causal source codes is achieved by time-sharing at most two optimal entropy-constrained scalar quantizers. The following shows that this result continues to hold for sources with memory in the limit of small distortion, in which case uniform quantizers are known to be (asymptotically) optimal in the entropy-constrained sense.

Theorem 1 *Assume the real stationary source $X_{-\infty}^{\infty}$ has finite differential entropy rate and suppose $H(Q_1(X_1)) < \infty$. Then*

$$\lim_{D\to 0}\left(r_c(D) + \frac{1}{2}\log(12D)\right) = h(X_{-\infty}^{\infty}). \tag{4}$$

Furthermore, $r_c(D)$ is asymptotically achieved by a uniform scalar quantizer Q_Δ with step size $\Delta = \sqrt{12D}$ in the sense that $\lim_{D\to 0} d(Q_{\sqrt{12D}})/D = 1$ *and*

$$\lim_{D\to 0}\left(r(Q_{\sqrt{12D}}) + \frac{1}{2}\log(12d(Q_{\sqrt{12D}}))\right) = h(X_{-\infty}^{\infty}). \tag{5}$$

Remarks.

1. Let $r(D)$ denote the rate-distortion function (with respect to the squared error distortion) of the stationary source $X_{-\infty}^{\infty}$. The rate loss of causal coding is the difference

   $$\delta(D) \triangleq r_c(D) - r(D).$$

   Since $r(D)$ is the OPTA function of all unrestricted coding schemes, the rate loss is always nonnegative. We have the Shannon lower bound [1] on $r(D)$,

   $$r(D) \geq r_{\mathrm{SLB}}(D) \triangleq h(X_{-\infty}^{\infty}) - \frac{1}{2}\log(2\pi e D) \tag{6}$$

   which is known to be asymptotically tight [21, 22] under the present conditions in the sense that $\lim_{D\to 0}(r(D) - r_{\mathrm{SLB}}(D)) = 0$. Combining this with Theorem 1 shows that the "price of causality" at high rates is

   $$\lim_{D\to 0}\delta(D) = \lim_{D\to 0}\left(r_c(D) - r_{\mathrm{SLB}}(D)\right) = \frac{1}{2}\log\left(\frac{\pi e}{6}\right) = 0.254 \text{ bits/sample.}$$

   This is the "space-filling loss" of the uniform quantizer; i.e., the high-resolution rate loss of a uniform scalar quantizer with respect to an optimal vector quantizer with asymptotically large dimension [3, 2].

2. The requirement of causality can be relaxed by allowing finite anticipation $K \geq 0$ for the reproduction coder. In this case $X_n = g(X_{-\infty}^{n+K})$, and casual codes correspond to the $K = 0$ case. In view of the (high-resolution) causal solution, it is tempting to replace the scalar uniform quantizer by a $(K+1)$-dimensional lattice quantizer [23] as a candidate for source coding with anticipation K. Indeed, by quantizing the source in blocks of size $K+1$ and applying sequence entropy coding, one obtains, for small distortion, the achievable rate-distortion curve $h(X_{-\infty}^{\infty}) - \frac{1}{2}\log(D/G_{K+1})$, where G_{K+1} is the normalized second moment of the $(K+1)$-dimensional lattice. Denoting the OPTA for anticipation K by $r^{(K)}(D)$ the rate loss with respect to unlimited anticipation is upper bounded as
$$\lim_{D\to 0}\big(r^{(K)}(D) - r(D)\big) \leq \frac{1}{2}\log(2\pi e G_{K+1}).$$
The lattice scheme and the bound are asymptotically optimal for $K = 0$ by Theorem 1, and also for large anticipation since for "good" lattices $G_K \longrightarrow 1/(2\pi e)$ as $K \to \infty$ [24]. However, it is not at all clear whether this scheme is optimal and hence this bound is tight for any finite positive K.

Proof of Theorem 1 We start with proving the second statement (5). Recall that the (common) marginal distribution of the X_n is absolutely continuous (i.e., has a pdf). From high-resolution quantization theory [25, Lemma 1], this implies without any further conditions that
$$\lim_{\Delta\to 0}\frac{E(X_1 - Q_\Delta(X_1))^2}{\Delta^2/12} = 1.$$
Since Q_Δ is a memoryless reproduction coder, $d(Q_\Delta) = E(X_1 - Q_\Delta(X_1))^2$, and hence we obtain
$$\lim_{D\to 0}\frac{d(Q_{\sqrt{12D}})}{D} = 1. \tag{7}$$
The rate of the memoryless reproduction coder Q_Δ is the entropy rate of $Q_\Delta(X_{-\infty}^{\infty})$. Using Lemma 2 with $\Delta = \sqrt{12D}$ we obtain
$$\lim_{D\to 0}\left(H(Q_{\sqrt{12D}}(X_{-\infty}^{\infty})) + \frac{1}{2}\log(12D)\right) = h(X_{-\infty}^{\infty}). \tag{8}$$
This proves the second statement of the theorem on the asymptotic optimality of Q_Δ.

Since $d(Q_\Delta)$ is clearly continuous in Δ, it is easy to see that (7) and (8) also imply the following asymptotic upper bound on $r_c(D)$:
$$\limsup_{D\to 0}\left(r_c(D) + \frac{1}{2}\log(12D)\right) \leq h(X_{-\infty}^{\infty}). \tag{9}$$

The rest of the proof is devoted to showing the reverse inequality

$$\liminf_{D\to 0}\left(r_c(D)+\frac{1}{2}\log(12D)\right)\geq h(X_{-\infty}^{\infty}). \tag{10}$$

We use the proof technique of [11] (proof of Theorem 3, steps 1 and 2) which needs to be adapted to sources with memory in the limit of small distortion. The key to this is the following "conditional" version of a classic result on high-rate entropy-constrained quantization by Zador [26, 27] and Gish and Pierce [3]. The lemma is proved in Appendix A.

Lemma 3 *Assume $X_{-\infty}^{\infty}$ is stationary, has finite differential entropy rate, and suppose $H(Q_1(X_1))<\infty$. For any $D>0$ define*

$$r_L(D)=\inf_{g:\,E(X_1-g(X_{-\infty}^1))^2\leq D} H(g(X_{-\infty}^1)|X_{-\infty}^0)$$

where the infimum is over all measurable real functions g of $X_{-\infty}^1$ that have countable range and satisfy $E\big(X_1-g(X_{-\infty}^1)\big)^2\leq D$. Then

$$\liminf_{D\to 0}\big(r_L(D)+\frac{1}{2}\log(12D)\big)\geq h(X_1|X_{-\infty}^0).$$

The inequality (10) follows once we show that

$$\liminf_{D\to 0}\big(r(\{g_n^{(D)}\})+\frac{1}{2}\log(12D)\big)\geq h(X_1|X_{-\infty}^0) \tag{11}$$

for an arbitrary family of causal reproduction coders $\big\{\{g_n^{(D)}\}:\ D>0\big\}$ such that $d(\{g_n^{(D)}\})\leq D$ for all $D>0$. In the proof of Theorem 3 in [11] the following lower bound on the the rate of any causal reproduction coder $\{g_n^{(D)}\}$ was shown to hold:

$$r(\{g_n^{(D)}\})\geq\limsup_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}H(\hat{X}_i^{(D)}|X_{-\infty}^{i-1}) \tag{12}$$

where $\hat{X}_i^{(D)}=g_n^{(D)}(X_{-\infty}^i)$. Define $d_n(D)=E(X_n-\hat{X}_n^{(D)})^2$. Then from the definition of r_L,

$$H(\hat{X}_n^{(D)}|X_{-\infty}^{n-1})\geq r_L(d_n(D)).$$

Now let $\bar{r}_L$ denote the the lower convex hull of r_L. Since $r_L(d_n(D))\geq\bar{r}_L(d_n(D))$, and $\bar{r}_L(d)$ is nonincreasing and convex (and therefore continuous at any $d>0$) we obtain

$$\limsup_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}H(\hat{X}_i^{(D)}|X_{-\infty}^{i-1}) \quad\geq\quad \limsup_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\bar{r}_L(d_i(D))$$

$$
\begin{aligned}
&\geq \limsup_{n\to\infty} \bar{r}_L\left(\frac{1}{n}\sum_{i=1}^{n} d_i(D)\right) \\
&\geq \bar{r}_L\left(\limsup_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} d_i(D)\right) \\
&\geq \bar{r}_L(D).
\end{aligned}
$$

From Lemma 3,

$$\liminf_{D\to 0}\left(r_L(D) + \frac{1}{2}\log(12D) - h(X_1|X_{-\infty}^{0})\right) \geq 0$$

As we show in Appendix A, this implies

$$\liminf_{D\to 0}\left(\bar{r}_L(D) + \frac{1}{2}\log(12D) - h(X_1|X_{-\infty}^{0})\right) \geq 0. \tag{13}$$

Hence

$$
\begin{aligned}
&\liminf_{D\to 0}\left(\limsup_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} H(\hat{X}_i^{(D)}|X_{-\infty}^{i-1}) - \tau(D)\right) \\
&\qquad \geq \liminf_{D\to 0}\left(\bar{r}_L(D) + \frac{1}{2}\log(12D) - h(X_1|X_{-\infty}^{0})\right) \geq 0.
\end{aligned}
$$

Combined with (12), this proves (11) and completes the proof of the theorem. □

4 Causal coding of individual sequences

In this section, our aim is to investigate the high-resolution behavior of causal codes when no probabilistic assumption is made on the sequence to be encoded. We introduce the "Lempel-Ziv differential entropy rate" of a bounded deterministic sequence, a concept that will prove crucial in characterizing the OPTA function of finite-memory causal codes for individual sequences. As well, it will provide an individual sequence version of the Shannon lower bound.

We begin with introducing some new notation and definitions. Let $\mathcal{P}^l$ denote the set of all probability measures on the Borel subsets of $\mathbb{R}^l$, and let $\mathcal{P}_a^l \subset \mathcal{P}^l$ be the collection of all P in $\mathcal{P}^l$ that are absolutely continuous with respect to the l-dimensional Lebesgue measure (i.e., each $P \in \mathcal{P}_a^l$ has a pdf). For any P in $\mathcal{P}_a^l$, $h(P)$ denotes the differential entropy of P, and if $P \in \mathcal{P}^l$ is discrete, $H(P)$ denotes its entropy. The normalized versions of $h(P)$ and $H(P)$ are denoted by $\bar{h}(P)$ and $\bar{H}(P)$, respectively, i.e., $\bar{h}(P) \triangleq \frac{1}{l}h(P)$ and $\bar{H}(P) \triangleq \frac{1}{l}H(P)$. We write $X \sim P$ if a random vector X has distribution P, so that $h(P) = h(X)$ or $H(P) = H(X)$ (whichever is appropriate) if $X \sim P$.

Given a sequence of real numbers $x_1^\infty = x_1, x_2, \ldots$, and positive integers $n \geq l$, let $\hat{P}^l_{x_1^n}$ denote the "sliding-window" empirical distribution of l-blocks in the initial segment x_1^n. That is, for any Borel set $B \subset \mathbb{R}^l$,

$$\hat{P}^l_{x_1^n}(B) = \frac{1}{n-l+1} \sum_{i=1}^{n-l+1} 1_B(x_i^{i+l-1})$$

where $1_B(y) = 1$ if $y \in B$ and $1_B(y) = 0$ otherwise.

For simplicity we always assume in the sequel that x_1^∞ (the sequence to be encoded) is bounded so that each x_n is from the interval $[0,1]$. All results can be trivially extended for arbitrary bounded sequences of real numbers.

4.1 Lempel-Ziv differential entropy

To define the individual-sequence analogue of differential entropy, we use the concept of *finite-state* compressibility of an individual sequence y_1^∞ over a finite alphabet $\mathcal{Y}$ introduced by Ziv and Lempel [12]. A variable-length finite-state lossless coder $E = (g, e)$ is characterized by a next state function $g : S \times \mathcal{Y} \to S$, where S is a finite set of states, and an encoder function $e : S \times \mathcal{Y} \to \{0,1\}^*$, where $\{0,1\}^*$ denotes the set of finite-length binary strings, including the empty string. The sequence y_1^∞ is encoded into the bit stream $b_1 b_2 b_3 \ldots$, (a concatenation of finite-length binary strings) while going through an infinite sequence of states $s_1, s_2, s_3 \ldots$, according to

$$\begin{aligned} b_i &= e(s_i, y_i) \\ s_{i+1} &= g(s_i, y_i) \quad i = 1, 2, \ldots \end{aligned}$$

It is assumed that the initial state s_1 is a prescribed element of S. The coder (g, e) is assumed to be *information lossless* [12] so that y_1^∞ can be losslessly recovered from s_1 and $b_1 b_2 b_3 \ldots$. Let $l(b_i)$ denote the length of the binary string b_i (the empty string has length zero), and let $L(y_1^n, E) \triangleq \sum_{i=1}^n l(b_i)$. The finite-state compressibility of y_1^∞ is defined by

$$\rho_{LZ}(y_1^\infty) \triangleq \lim_{s\to\infty} \limsup_{n\to\infty} \min_{E \in \mathcal{E}(s)} \frac{L(y_1^n, E)}{n} \tag{14}$$

where $\mathcal{E}(s)$ is the set of all finite-state coders with the number of states bounded as $|S| \leq s$. Clearly, $\rho_{LZ}(y_1^\infty) \leq \log |\mathcal{Y}|$ and $\rho_{LZ}(y_1^\infty)$ is an ultimate lower bound on the rate of any finite-state binary lossless code for y_1^∞.

A fundamental result [12, Thm. 3] states that the finite-state compressibility of y_1^∞ is the limit of the lth order normalized "empirical entropies" of y_1^∞, i.e,

$$\rho_{\mathrm{LZ}}(y_1^\infty) = \lim_{l\to\infty} \bar{H}_l(y_1^\infty) \tag{15}$$

where

$$\bar{H}_l(y_1^\infty) \triangleq \limsup_{n\to\infty} \bar{H}(\hat{P}^l_{y_1^n}). \tag{16}$$

The limit in (15) exists since $l\bar{H}_l(y_1^\infty)$ is subadditive in l [12, Lemma 1].

Another fundamental characterization of $\rho_{\mathrm{LZ}}(y_1^\infty)$, given in [12], is that

$$\rho_{\mathrm{LZ}}(y_1^\infty) = \limsup_{n\to\infty} \frac{1}{n} c(y_1^n) \log c(y_1^n)$$

where $c(y_1^n)$ denotes the number of phrases obtained via the incremental parsing of y_1^n; i.e., when y_1^n is sequentially parsed into shortest strings that have not appeared so far. It follows that $\rho_{\mathrm{LZ}}(y_1^\infty)$ can be achieved by the universal Lempel-Ziv algorithm based on incremental parsing.

The following definition provides an individual-sequence analogue of differential entropy. We adapt Csiszár's operational characterization (Lemmas 1 and 2) of differential entropy via the asymptotic entropy of a uniform quantizer, but replace the process entropy with the finite-state compressibility of the sequence. Recall that $Q_\Delta(x_1^\infty)$ denotes the uniformly quantized sequence $\{Q_\Delta(x_n)\}_{n=1}^\infty$.

Definition 1 *The* Lempel-Ziv differential entropy rate *of a sequence of real numbers x_1^∞, with $x_n \in [0,1]$ for all n, is defined by*

$$h_{\mathrm{LZ}}(x_1^\infty) \triangleq \limsup_{\Delta\to 0} \big[\rho_{\mathrm{LZ}}(Q_\Delta(x_1^\infty)) + \log\Delta\big]. \tag{17}$$

Remarks.

1. Note that $Q_\Delta(x)$ can take at most $\lceil\frac{1}{\Delta}\rceil$ values as x varies in $[0,1]$, where $\lceil a\rceil$ denotes the smallest integer that is greater than a.[3] Thus $\rho_{\mathrm{LZ}}(Q_\Delta(x_1^\infty)) \le \log\lceil\frac{1}{\Delta}\rceil$ and $\rho_{\mathrm{LZ}}(Q_\Delta(x_1^\infty)) + \log\Delta \le \log(\Delta+1)$, implying $h_{\mathrm{LZ}}(x_1^\infty) \le 0$. Consequently, $h_{LZ}(x_1^\infty)$ is either finite or $h_{LZ}(x_1^\infty) = -\infty$.

2. If each x_n belongs to the same finite set $\mathcal{X} \subset [0,1]$, one always has $h_{\mathrm{LZ}}(x_1^\infty) = -\infty$ since in this case $\rho_{LZ}(Q_\Delta(x_1^\infty))$ is bounded from above by the logarithm of the size of $\mathcal{X}$.

3. Examples where $h_{LZ}(x_1^\infty)$ is finite can be generated by letting x_1^∞ be a typical sample path of a stationary and ergodic process $\{X_n\}_{n=1}^\infty$ with finite differential entropy rate $h(X_1^\infty)$. From the ergodic theorem, with probability one, we have for all l,

[3] Note that this definition slightly differs from the usual definition of the ceiling function.

$\lim_{n\to\infty} \hat{P}^l_{X_1^n}(B) = \Pr(X_1^l \in B)$ for any Borel set $B \subset \mathbb{R}^l$. Thus, for almost all realizations x_1^∞, for all Δ,

$$\lim_{n\to\infty} \bar{H}(\hat{P}^l_{Q_\Delta(x_1^n)}) = \bar{H}(Q_\Delta(X_1^l))$$

and hence, from (15),

$$\rho_{\mathrm{LZ}}(Q_\Delta(x_1^\infty)) = \lim_{l\to\infty} \bar{H}(Q_\Delta(X_1^l)).$$

From Lemma 2,

$$\lim_{\Delta\to 0}\big[\lim_{l\to\infty} \bar{H}(Q_\Delta(X_1^l)) + \log\Delta\big] = h(X_1^\infty)$$

so, for almost all realizations x_1^∞,

$$h_{\mathrm{LZ}}(x_1^\infty) = h(X_1^\infty). \tag{18}$$

4. For nonstationary processes the Lempel-Ziv differential entropy rate of a typical sample path may not coincide with the ordinary differential entropy rate of the process (or the latter may not exist at all.) For example, typical sample paths of a discrete process can also have finite Lempel-Ziv differential entropy rate. Let X_1^∞ be a sequence of independent random variables such that X_n is uniformly distributed in $\big\{0, 1/2^n, \ldots, 1-1/2^n\big\}$. Then for any positive integer m, $Q_{1/2^m}(X_m^\infty)$ is a sequence of independent and identically distributed (i.i.d.) random variables that are uniformly distributed on a set with 2^m elements. Since the effect of the initial segment $Q_{1/2^m}(X_1^{m-1})$ on $\hat{P}^l_{Q_\Delta(X_1^n)}$ vanishes as $n \to \infty$, with probability one, we have for all $l \geq 1$,

   $$\lim_{n\to\infty} \bar{H}(\hat{P}^l_{Q_{1/2^m}(X_1^n)}) = m$$

   so by (15), $\rho_{\mathrm{LZ}}(Q_{1/2^m}(X_1^\infty)) = m$. Thus,

   $$\limsup_{\Delta\to 0}\big(\rho_{\mathrm{LZ}}(Q_\Delta(X_1^\infty)) + \log\Delta\big) \geq 0.$$

   Since the left-hand side is always nonpositive (see Remark 1), we obtain that $h_{\mathrm{LZ}}(X_1^\infty) = 0$ with probability one, while $h(X_1^\infty)$ does not exist.

We need some new definitions to state some important facts about $h_{\mathrm{LZ}}(x_1^\infty)$. If a sequence of probability measures $P_n \in \mathcal{P}^l$, $n = 1, 2, \ldots$, converges weakly to some $P \in \mathcal{P}^l$, we write $P_n \Rightarrow P$. Let $\mathcal{P}^l(x_1^\infty)$ denote the collection of all $P \in \mathcal{P}^l$ for which there is an infinite subsequence $\{n_k\}$ of the positive integers such that $\hat{P}^l_{x_1^{n_k}} \Rightarrow P$. Thus $\mathcal{P}^l(x_1^\infty)$ is the set of *subsequential limits* (with respect to weak convergence) of the sequence $\hat{P}^l_{x_1^n}$, $n = l, l+1, \ldots$.

Also, define $\mathcal{P}_a^l(x_1^\infty) \triangleq \mathcal{P}^l(x_1^\infty) \cap \mathcal{P}_a^l$, the set of probability measures in $\mathcal{P}^l(x_1^\infty)$ that possess a density. Note that for an arbitrary x_1^∞, both $\mathcal{P}^l(x_1^\infty)$ and $\mathcal{P}_a^l(x_1^\infty)$ may be empty.

We have seen that an individual sequence with finite Lempel-Ziv differential entropy rate might or might not be a typical sample path of a stationary and ergodic process. Nevertheless, the next result shows that any such individual sequence can be characterized via an associated stationary random source having finite differential entropy rate.

Theorem 2 *Assume x_1^∞ is a sequence with $x_n \in [0,1]$ for all n such that $h_{\mathrm{LZ}}(x_1^\infty)$ is finite. Then there exists a real-valued stationary process X_1^∞ with finite-dimensional distributions $X_1^l \sim P_l$ such that $P_l \in \mathcal{P}_a^l(x_1^\infty)$ for all $l \geq 1$. Furthermore, X_1^∞ has finite differential entropy rate $h(X_1^\infty)$, and*

$$h_{\mathrm{LZ}}(x_1^\infty) = h(X_1^\infty) = \lim_{l\to\infty} \sup_{P\in\mathcal{P}_a^l(x_1^\infty)} \bar{h}(P). \tag{19}$$

The proof is given in Appendix B. The theorem states that for all x_1^∞ with finite $h_{\mathrm{LZ}}(x_1^\infty)$, there is a stationary process X_1^∞ with finite differential entropy rate whose finite-dimensional distributions are the subsequential limits of empirical distributions of overlapping blocks of x_1^∞. (In particular, $\mathcal{P}_a^l(x_1^\infty)$ is nonempty for all l.) Furthermore, the differential entropy rate of the process coincides with $h_{\mathrm{LZ}}(x_1^\infty)$, and for asymptotically large l, the blocks X_1^l have maximum differential entropy. Thus, in a sense, X_1^∞ represents the dominant empirical behavior of x_1^∞. This characterization of x_1^∞ will prove crucial in our development of casual coding of individual sequences.

4.2 Finite-memory causal coding of individual sequences

Consider an infinite bounded sequence of real numbers $x_1^\infty = x_1, x_2, \ldots$, such that $x_n \in [0,1]$ for all n. A *causal, finite-resolution, finite-memory (CFRFM) encoder* with memory of size $M \geq 0$ is described by a reproduction coder f which, for each $i \geq 1$, maps the source string x_{i-M}^i into a reproduction letter $\hat{x}_i$, and by a finite-state coder which losslessly encodes $\hat{x}_1^\infty = \hat{x}_1, \hat{x}_2, \ldots$ into a sequence of variable-length binary strings. To unambiguously specify $\hat{x}_i = f(x_{i-M}^i)$ for $i = 1, \ldots, M$, we formally define $x_{-M+1} = \cdots = x_0 = 0$, but only $x_1, x_2, \ldots$ are reproduced. The reproduction coder is said to have finite resolution because it is assumed that it only sees a finely quantized version of the input. Formally, $f : \mathbb{R}^{M+1} \to \mathbb{R}$ is called a reproduction coder with input resolution $\delta > 0$ if for all $(z_1, \ldots, z_{M+1}) \in \mathbb{R}^{M+1}$

$$f(z_1, \ldots, z_{M+1}) = f(Q_\delta(z_1), \ldots, Q_\delta(z_{M+1})) \tag{20}$$

where, as before, Q_δ is the uniform quantizer with step size δ.

Since we assume that each x_i is in $[0,1]$, the finite input resolution property implies that there are only finitely many possible values of $\hat{x}_i = f(x_{i-M}^i)$, the collection of which we denote by $\hat{\mathcal{X}}_f$. The reproduction sequence $\hat{x}_1^\infty$ is encoded by a finite-state, variable-length lossless coder $E = (g, e)$ which emits the bit stream $b_1 b_2 \ldots$, where the binary string b_i has length $l(b_i)$. Analogously to causal codes for random sources, we define the rate of the system, measured in bits per source letter, by

$$r(x_1^\infty, f, E) \triangleq \limsup_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} l(b_i).$$

We eliminate the dependence of the system performance on the particular choice of the lossless coder by considering the minimum rate achievable by finite-state, variable-rate lossless coding of the reproduction sequence. Hence the rate of the CFRFM code with reproduction coder f is

$$r(x_1^\infty, f) \triangleq \inf_E r(x_1^\infty, f, E) \tag{21}$$

where the infimum is taken over all codes E with an arbitrary (but finite) number of states.

Comparing definitions (14) and (21), we clearly have $\rho_{\mathrm{LZ}}(\hat{x}_1^\infty) \le r(x_1^\infty, f)$. On the other hand, the fact that $\rho_{LZ}(\hat{x}_1^\infty)$ is achievable by *universal* finite-state schemes [12, Thm. 2] implies the reverse inequality, so we have

$$r(x_1^\infty, f) = \rho_{\mathrm{LZ}}(\hat{x}_1^\infty).$$

The distortion of the CFRFM coder (which only depends on the reproduction coder f) is given by the average cumulative squared error

$$d(x_1^\infty, f) \triangleq \limsup_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{x}_i)^2.$$

For $\delta > 0$ and $M \ge 0$, let $\mathcal{F}_\delta^M$ denote the family of all reproduction coders with input resolution δ and memory M; then $\mathcal{F} \triangleq \bigcup_{M\ge 0} \bigcup_{\delta > 0} \mathcal{F}_\delta^M$ is the collection of all finite input resolution reproduction coders having finite memory. The OPTA function for CFRFM codes with respect to x_1^∞ is defined by

$$r_{\mathcal{F}}(D, x_1^\infty) \triangleq \inf_{f\in\mathcal{F}:\, d(x_1^\infty, f)\le D} r(x_1^\infty, f). \tag{22}$$

Thus $r_{\mathcal{F}}(D, x_1^\infty)$ is the minimum rate achievable at distortion level D by any CFRFM code with reproduction coder having arbitrarily fine input resolution and arbitrarily large (but finite) memory size, and lossless coder having arbitrarily large (but finite) number of states. The following is an individual-sequence analogue of the converse part of Theorem 1.

Theorem 3 (Converse) *Assume x_1^∞ is a sequence with $x_n \in [0,1]$ for all n for which $h_{\rm LZ}(x_1^\infty)$ is finite. Then the OPTA function for CFRFM codes with respect to x_1^∞ satisfies*

$$\liminf_{D\to 0}\bigl(r_{\mathcal{F}}(D,x_1^\infty) + \frac{1}{2}\log(12D)\bigr) \ge h_{\rm LZ}(x_1^\infty).$$

Remark. Note that CFRFM coders form a subclass of the set of all causal reproduction coders we considered in the probabilistic setting. The condition that every coder in $\mathcal{F}$ is time-invariant is a natural restriction in the individual sequence setting. The other two conditions are imposed for technical reasons (and are quite heavily relied on in the proof). The finite-memory requirement, also assumed in [10] when studying the large deviations performance of special classes of causal codes, does restrict generality, although codes with long enough (but finite) memory may well approximate codes with infinite, but rapidly fading memory. On the other hand, the finite-resolution condition is non-restrictive from a practical viewpoint since any coder implemented on a digital computer must have finite input resolution.

Proof. Let X_1^∞ be the stationary process associated with x_1^∞ via Theorem 2. For convenience, we extend X_1^∞ into a two-sided process $X_{-\infty}^\infty$ by specifying that the n-blocks X_{-n+i+1}^i, $i = n-1, n-2, \ldots$ have the same distribution as X_1^n for all $n \ge 1$. We show that $r_{\mathcal{F}}(D,x_1^\infty)$ is lower bounded by the causal OPTA function of $X_{-\infty}^\infty$, from which the result will follow.

Consider any reproduction coder f with arbitrary input resolution $\delta > 0$ and memory $M \ge 0$. Since $r(x_1^\infty, f) = \rho_{\rm LZ}(\hat{x}_1^\infty)$, where $\hat{x}_i = f(x_{i-M}^i)$ for all i, from (15) and (16),

$$r(x_1^\infty, f) = \lim_{l\to\infty} \bar{H}_l(\hat{x}_1^\infty)$$

where

$$\bar{H}_l(\hat{x}_1^\infty) = \limsup_{n\to\infty} \bar{H}(\hat{P}^l_{\hat{x}_1^n}).$$

Fix $l \ge 1$ and define $\hat{f} : \mathbb{R}^{M+l} \to \mathbb{R}^l$ by

$$\hat{f}(z_1^{M+l}) = (f(z_1^{M+1}), f(z_2^{M+2}), \ldots, f(z_l^{M+l}))$$

for any $z_1^{M+1} \in \mathbb{R}^{M+l}$. Since f has input resolution δ, the range of the $\hat{x}_i$, $\hat{\mathcal{X}}_f = f([0,1]^{M+1})$, is finite. Thus for all $n \ge l$, $\hat{P}^l_{\hat{x}_1^n}$ is a discrete distribution such that for any $w \in \hat{f}([0,1]^{M+l}) \subset \hat{\mathcal{X}}_f^l$,

$$\begin{aligned}\hat{P}^l_{\hat{x}_1^n}(w) &= \frac{1}{n-l+1}\sum_{i=1}^{n-l+1} 1_{\{w\}}(\hat{x}_i^{i+l-1})\\ &= \frac{1}{n-l+1}\sum_{i=1}^{n-l+1} 1_{\hat{f}^{-1}(w)}(x_{i-M}^{i+l-1})\end{aligned}$$

$$= \hat{P}^{M+l}_{x^n_{-M+1}}(\hat{f}^{-1}(w)).$$

(Recall that $x_{-M+1} = \cdots = x_0 = 0$ by definition.) Hence $\hat{P}^l_{\hat{x}^n_1} = \hat{P}^{M+l}_{x^n_{-M+1}} \circ \hat{f}^{-1}$, where for any probability measure $P \in \mathcal{P}^l$ and measurable function $g : \mathbb{R}^l \to \mathbb{R}^m$, $P \circ g^{-1}$ denotes the probability measure in $\mathcal{P}^m$ induced by P and g; i.e., for any Borel set $B \subset \mathbb{R}^m$,

$$P \circ g^{-1}(B) \triangleq P(g^{-1}(B)) = P(\{x : g(x) \in B\}). \tag{23}$$

Since $X_1^{M+l} \sim P_{M+l} \in \mathcal{P}_a^{M+l}(x_1^\infty)$ by Theorem 2, there is a subsequence $\{n_j\}$ such that $P^{M+l}_{x_1^{n_j}} \Rightarrow P_{M+l}$. Clearly, we also have $\hat{P}^{M+l}_{x^{n_j}_{-M+1}} \Rightarrow P_{M+l}$ since the effect of the initial segment x^0_{-M+1} vanishes asymptotically. By the δ input resolution property, $\hat{f}(z_1^{M+l}) = \hat{f}(Q_\delta^{M+l}(z_1^{M+l}))$, so $\hat{f}$ is constant on the interior of each of the cells of Q_δ^{M+l}, which are $(M+l)$-dimensional hypercubes, and the discontinuities of $\hat{f}$ occur on the faces of the hypercubes. Thus the set of discontinuities of $\hat{f}$ have zero P_{M+l} probability (recall that P_{M+1} has a pdf), and so $\hat{P}^{M+l}_{x^{n_j}_{-M+1}} \circ \hat{f}^{-1} \Rightarrow P_{M+l} \circ \hat{f}^{-1}$ by [28, Thm. 5.1]. Since $P_{M+l} \circ \hat{f}^{-1}$ is discrete with finite support, this implies

$$\begin{aligned}
\bar{H}_l(\hat{x}_1^\infty) &= \limsup_{n\to\infty} \bar{H}(\hat{P}^l_{\hat{x}_1^n}) \\
&\geq \lim_{j\to\infty} \bar{H}(\hat{P}^l_{\hat{x}_1^{n_j}}) \\
&= \lim_{j\to\infty} \bar{H}(P^{M+l}_{x^n_{-M+1}} \circ \hat{f}^{-1}) \\
&= \bar{H}(P_{M+l} \circ \hat{f}^{-1}) \\
&= \bar{H}(f(X_1^{M+1}), \ldots, f(X_l^{l+M})).
\end{aligned}$$

We obtain

$$r(x_1^\infty, f) = \lim_{l\to\infty} \bar{H}_l(\hat{x}_1^\infty) \geq \lim_{l\to\infty} \bar{H}(\hat{X}_1^{M+l}) = H(\hat{X}_1^\infty)$$

where $\hat{X}_i = \hat{f}(X^i_{i-M})$.

Similarly, let $\{n_i\}$ be a subsequence such that $\hat{P}^{M+1}_{x^{n_i}_{-M+1}} \Rightarrow P_{M+1}$. Since the set of discontinuities of the bounded function $(z_{M+1} - f(z_1^{M+1}))^2$, $z_1^{M+1} \in [0,1]^{M+1}$, has P_{M+1} probability zero, we obtain

$$\begin{aligned}
d(x_1^\infty, f) &= \limsup_{n\to\infty} \frac{1}{n} \sum_{i=1}^n (x_i - \hat{x}_i)^2 \\
&\geq \lim_{i\to\infty} \frac{1}{n_i} \sum_{i=1}^{n_i} (x_i - f(x^i_{i-M}))^2 \\
&= \lim_{i\to\infty} \int (z_{M+1} - f(z_1^{M+1}))^2 \, dP^{M+1}_{x^{n_i}_{-M+1}}(z_1^{M+1})
\end{aligned}$$

$$
\begin{aligned}
&= \int (z_{M+1} - f(z_1^{M+1}))^2 \, dP_{M+1}(z_1^{M+1}) \\
&= E(X_1 - f(X_{-M+1}^1))^2 \triangleq d(f).
\end{aligned}
$$

Thus the rate and distortion of any CFRFM code for x_1^∞ are lower bounded by the rate and distortion, respectively, of its stationary causal reproduction coder f encoding $X_{-\infty}^\infty$. Hence

$$r(x_1^\infty, f) \geq r_c(d(f)) \geq r_c(d(x_1^\infty, f))$$

where r_c denotes the casual OPTA function of the stationary source $X_{-\infty}^\infty$. Since the above holds for any reproduction coder f having arbitrary input resolution and memory size, by definition of the OPTA function for CFRFM codes (22), for all $D > 0$,

$$r_{\mathcal{F}}(D, x_1^\infty) \geq r_c(D).$$

Therefore we obtain

$$
\begin{aligned}
\liminf_{D\to 0} \bigl(r_{\mathcal{F}}(D, x_1^\infty) + \frac{1}{2}\log(12D) \bigr) &\geq \lim_{D\to 0} \bigl(r_c(D) + \frac{1}{2}\log(12D) \bigr) \\
&= h(X_{-\infty}^\infty) = h_{\mathrm{LZ}}(x_1^\infty)
\end{aligned}
$$

where the first equality follows from Theorem 1 (whose conditions are clearly satisfied by $X_{-\infty}^\infty$) and the second from Theorem 2. This completes the proof. □

Theorem 2 and the preceding proof suggest that similarly to the random source case, the asymptotic lower bound for $r_{\mathcal{F}}(D, x_1^\infty)$ in Theorem 3 is achievable by a simple scheme in which the output of a memoryless uniform scalar quantizer Q_Δ is encoded using a finite-state lossless coder. Indeed, all one needs to show is that the finiteness of $h_{\mathrm{LZ}}(x_1^\infty)$ implies $d(x_1^\infty, Q_\Delta) \approx \Delta^2/12$ as $\Delta \to 0$ (which is according to high-rate quantization theory the typical asymptotic behavior of uniform quantizers for sources with a density). Then one could conclude directly from the definition of $h_{LZ}(x_1^\infty)$ that

$$\limsup_{\Delta\to 0} \bigl(r(x_1^\infty, Q_\Delta) + \frac{1}{2}\log(12 d(x_1^\infty, Q_\Delta)) \bigr) = h_{\mathrm{LZ}}(x_1^\infty)$$

since $r(x_1^\infty, Q_\Delta) = \rho_{LZ}(Q_\Delta(x_1^\infty))$. However, as the next example shows, it is not hard to construct sequences x_1^∞ with finite $h_{\mathrm{LZ}}(x_1^\infty)$ that do not exhibit this behavior.

(Counter) Example. Let $\{Y_n\}_{n=1}^\infty$ be a sequence of i.i.d. random variables with each Y_n being uniformly distributed on $[0, 1]$, and let y_1^∞ by a typical sample path of $\{Y_n\}$ such that $\hat{P}_{y_1^n}^l \Rightarrow u^l$ as $n \to \infty$ for all $l \geq 1$, where u^l denotes the uniform distribution on $[0, 1]^l$. Let $n_0 = 0$ and $\{n_j\}_{j=1}^\infty$ be an increasing sequence of positive integers such that

$$\lim_{j\to\infty} \frac{n_j - n_{j-1}}{n_j} = 1. \tag{24}$$

For each $m = 1, 2, 3, \dots$ and $j = m(m+1)/2$, let

$$x_{n_{j-1}+1}^{n_j} = y_1^{n_j - n_{j-1}}$$

and for $j = m(m+1)/2 + k$, $k = 1, \dots, m$, let $x_{n_{j-1}+1}^{n_j}$ be any sequence with components

$$x_i \in \left\{0, \frac{1}{k}, \dots, \frac{(k-1)}{k}, 1\right\} \quad \text{for } i = n_{j-1}+1, \dots, n_j. \tag{25}$$

The condition (24) clearly implies that along the subsequence n_j, $j = m(m+1)/2$, $m = 1, 2, \dots$ the empirical distribution of y_1^∞ dominates in the sense that $\hat{P}^l_{x_1^{n_{m(m+1)/2}}} \Rightarrow u^l$ as $m \to \infty$ for any $l \geq 1$. Letting Q_Δ^l denote the l-fold product of Q_Δ, we thus have

$$\bar{H}_l(Q_\Delta(x_1^\infty)) \geq \bar{H}(u^l \circ (Q_\Delta^l)^{-1}) = H(u^1 \circ Q_\Delta^{-1})$$

and, since $H(u^1 \circ Q_\Delta^{-1}) + \log \Delta \to h(u^1) = 0$ as $\Delta \to 0$ by Lemma 1, we obtain

$$\liminf_{\Delta \to 0} \bigl(\rho_{\mathrm{LZ}}(Q_\Delta(x_1^\infty) + \log \Delta\bigr) \geq 0.$$

Since $\rho_{\mathrm{LZ}}(Q_\Delta(x_1^\infty) \leq \log(1/\Delta + 1)$, this yields

$$\lim_{\Delta \to 0} \bigl(\rho_{\mathrm{LZ}}(Q_\Delta(x_1^\infty) + \log \Delta\bigr) = 0$$

so we conclude that x_1^∞ has (maximum) Lempel-Ziv differential entropy rate $h_{\mathrm{LZ}}(x_1^\infty) = 0$.

On the other hand, for any fixed integer $k \geq 1$, if $\Delta = 1/k$ and $j = m(m+1)/2 + k$ for $m = k, k+1, k+2, \dots$, then from (25) we have

$$(Q_\Delta(x_i) - x_i)^2 = \Delta^2/4 \quad \text{for all } i = n_{j-1}+1, \dots, n_j. \tag{26}$$

Since $(Q_\Delta(x) - x)^2 \leq \Delta^2/4$ for all x, (26) and (24) imply

$$d(x_1^\infty, Q_\Delta) = \limsup_{n \to \infty} \frac{1}{n} \sum_{i=1}^n (Q_\Delta(x_i) - x_i)^2 = \frac{\Delta^2}{4}$$

for all $\Delta = 1/k$, $k = 1, 2, \dots$. Thus

$$\lim_{\Delta \to 0} \bigl(r(x_1^\infty, Q_\Delta) + \frac{1}{2}\log(12 d(x_1^\infty, Q_\Delta))\bigr) = \frac{1}{2}\log 3 > h_{\mathrm{LZ}}(x_1^\infty).$$

so the asymptotic lower bound of Theorem 3 is not achieved by memoryless uniform scalar quantization.

Next we show that the asymptotic lower bound of Theorem 3 can be achieved by scalar uniform quantization and *subtractive dithering* [14, 15] (followed by Lempel-Ziv coding). Let

$\{Z_n\}_{n=1}^{\infty}$ be a sequence of i.i.d. random variables with each Z_n uniformly distributed on $(-1/2, 1/2]$. It is assumed that $\{Z_n\}$ is available to both the encoder and the decoder. The dithered uniform quantizer maps each x_i into

$$\hat{x}_i = Q_\Delta(x_i + Z_{\Delta,i})$$

where $Z_{\Delta,i} = \Delta Z_i$ (so that $Z_{\Delta,i}$ is uniformly distributed on $(-\Delta/2, \Delta/2]$), and the sequence $\hat{x}_1^\infty$ is encoded using a finite-state, variable-length coder. We measure the rate of the system, $r(x_1^\infty, Z_1^\infty, Q_\Delta)$, by the minimum rate achievable by finite-state variable-length coding of $\hat{x}_1^\infty$:

$$r(x_1^\infty, Z_1^\infty, Q_\Delta) \triangleq \rho_{\mathrm{LZ}}(\hat{x}_1^\infty). \tag{27}$$

Note that for any bounded sequence x_1^∞ and fixed $\Delta > 0$, $\hat{x}_i$ is a sequence from a finite alphabet, so $\rho_{\mathrm{LZ}}(\hat{x}_1^\infty)$ is well defined. Moreover, the Lempel-Ziv coding of $\hat{x}_1^\infty$ achieves this rate [12].

At the decoder (where Z_1^∞ is also available) x_i is reproduced as

$$\tilde{x}_i \triangleq \hat{x}_i - Z_{\Delta,i} = Q_\Delta(x_i + Z_{\Delta,i}) - Z_{\Delta,i}$$

and, accordingly, the distortion of the system is measured by

$$d(x_1^\infty, Z_1^\infty, Q_\Delta) \triangleq \limsup_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} (x_i - \tilde{x}_i)^2. \tag{28}$$

Both the rate and the distortion of the dithered scheme are random quantities which depend on the dither sequence Z_1^∞. The next result states that for any fixed input sequence x_1^∞, with probability one the asymptotic lower bound of Theorem 3 can be achieved by dithered schemes that use the same realization of the dither sequence for all Δ.

Theorem 4 (Achievability) *Assume x_1^∞ is a sequence with $x_n \in [0,1]$ for all n such that $h_{\mathrm{LZ}}(x_1^\infty)$ is finite. Then for almost all realizations z_1^∞ of Z_1^∞, the dithered scalar uniform quantizer has asymptotic performance*

$$\limsup_{\Delta\to 0} \Big(r(x_1^\infty, z_1^\infty, Q_\Delta) + \frac{1}{2}\log(12 d(x_1^\infty, z_1^\infty, Q_\Delta)) \Big) \le h_{\mathrm{LZ}}(x_1^\infty).$$

Proof. First we consider the distortion. It is well known [29] that if Z is uniformly distributed on $(-\Delta/2, \Delta/2]$, then for any $x \in \mathbb{R}$ the random variable

$$Q_\Delta(x + Z) - Z - x$$

is also uniformly distributed on $(-\Delta/2, \Delta/2]$, and therefore

$$E(Q_\Delta(x+Z) - Z - x)^2 = \frac{\Delta^2}{12}$$

(see also [15] for a generalization to dithered lattice quantizers). Thus for any $\Delta > 0$, $\{Q_\Delta(x_i + Z_{\Delta,i}) - Z_{\Delta,i} - x_i\}_{i=1}^\infty$ is a sequence of i.i.d. random variables with common distribution that is uniform on $(-\Delta/2, \Delta/2]$. Hence, by the strong law of large numbers, with probability one,

$$d(x_1^\infty, Z_1^\infty, Q_\Delta) = \limsup_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} (Q_\Delta(x_i + Z_{\Delta,i}) - Z_{\Delta,i} - x_i)^2 = \frac{\Delta^2}{12}.$$

which implies

$$\Pr\left(d(x_1^\infty, Z_1^\infty, Q_\Delta) = \frac{\Delta^2}{12} \text{ for all rational } \Delta > 0 \right) = 1.$$

It is easy to check that for any $x \in [0,1]$, $z \in (-1/2, 1/2]$, and $\Delta, \Delta' > 0$,

$$\Big| |Q_\Delta(x + \Delta z) - x - \Delta z| - |Q_{\Delta'}(x + \Delta' z) - x - \Delta' z| \Big| \le |\Delta - \Delta'| + 2\frac{|\Delta - \Delta'|}{\min\{\Delta, \Delta', 1\}}\ .$$

It follows that $d(x_1^\infty, Z_1^\infty, Q_\Delta)$ is a continuous function of Δ with probability one, so we obtain

$$\Pr\left(d(x_1^\infty, Z_1^\infty, Q_\Delta) = \frac{\Delta^2}{12} \text{ for all } \Delta > 0 \right) = 1. \tag{29}$$

Next recall that by (15),

$$r(x_1^\infty, Z_1^\infty, Q_\Delta) = \rho_{\mathrm{LZ}}(\hat{x}_1^\infty) = \lim_{l\to\infty} \bar{H}_l(\hat{x}_1^\infty) \tag{30}$$

where $\hat{x}_n = Q_\Delta(x_n + Z_{\Delta,n})$ for all n. Fix $l \ge 1$ and $n \ge l$, let $(z_{\Delta,1}, \ldots, z_{\Delta,n})$ be any length n sequence such that $z_{\Delta,i} \in (-\Delta/2, \Delta/2]$ for all i, and let $y_i = x_i + z_{\Delta,i}$ for $i = 1, \ldots, n$. Consider the joint empirical probability of overlapping l-blocks, $\hat{P}^l_{(x_1^n, y_1^n)}$, given for all Borel sets $B \subset \mathbb{R}^{2l}$ by

$$\hat{P}^l_{(x_1^n, y_1^n)}(B) = \frac{1}{n-l+1} \sum_{i=1}^{n-l+1} 1_B(x_i^{i+l-1}, y_i^{i+l-1}).$$

Furthermore, let $X_1^l = (X_1, \ldots, X_l)$ and $Y_1^l = (Y_1, \ldots, Y_l)$ be random vectors such that the pair (X_1^l, Y_1^l) has joint distribution $\hat{P}^l_{(x_1^n, y_1^n)}$. Then $|X_i - Y_i| \le \Delta/2$ for all $i = 1, \ldots, l$ with probability one. Hence for any x and $\Delta' > 0$, we have

$$\Pr(Y_i \in [x - \Delta/2, x + \Delta' + \Delta/2] \,|\, X_i \in [x, x + \Delta']) = 1$$

so that, conditioned on the event that $Q_{\Delta'}(X_i)$ is a given constant, $Q_\Delta(Y_i)$ can take at most $\lceil \frac{\Delta'}{\Delta} \rceil + 2$ different values. Consequently,

$$\begin{aligned}
H(Q_\Delta(Y_1^l) \,|\, Q_{\Delta'}(X_1^l)) &\le \sum_{i=1}^{l} H(Q_\Delta(Y_i) \,|\, Q_{\Delta'}(X_1^l)) \\
&\le \sum_{i=1}^{l} H(Q_\Delta(Y_i) \,|\, Q_{\Delta'}(X_i)) \\
&\le l \log \left(\frac{\Delta'}{\Delta} + 3 \right).
\end{aligned}$$

Therefore

$$\begin{aligned}
H(Q_\Delta(Y_1^l)) &\le H(Q_\Delta(Y_1^l), Q_{\Delta'}(X_1^l)) \\
&= H(Q_{\Delta'}(X_1^l)) + H(Q_\Delta(Y_1^l) \,|\, Q_{\Delta'}(X_1^l)) \\
&\le H(Q_{\Delta'}(X_1^l)) + l \log \left(\frac{\Delta'}{\Delta} + 3 \right).
\end{aligned}$$

Note that $H(Q_{\Delta'}(X_1^l)) = H(\hat{P}^l_{Q_{\Delta'}(x_1^n)})$ and, if $(z_{\Delta,1}, \ldots, z_{\Delta,n}) = (Z_{\Delta,1}, \ldots, Z_{\Delta,n})$, then $H(Q_\Delta(Y_1^l)) = H(\hat{P}^l_{\hat{x}_1^n})$. Thus we obtain that with probability one, for any $\Delta, \Delta' > 0$, $l \ge 1$, and $n \ge l$,

$$\bar{H}(\hat{P}^l_{\hat{x}_1^n}) \le \bar{H}(\hat{P}^l_{Q_{\Delta'}(x_1^n)}) + \log \left(\frac{\Delta'}{\Delta} + 3 \right).$$

Taking the limit superior of both sides as $n \to \infty$ and then the limit as $l \to \infty$ we obtain from (15) and (30) that with probability one, for all $\Delta, \Delta' > 0$

$$r(x_1^\infty, Z_1^\infty, Q_\Delta) = \rho_{\mathrm{LZ}}(\hat{x}_1^\infty) \le \rho_{\mathrm{LZ}}(Q_{\Delta'}(x_1^\infty)) + \log \left(\frac{\Delta'}{\Delta} + 3 \right).$$

Thus for every fixed $\Delta' > 0$,

$$\begin{aligned}
\limsup_{\Delta \to 0} \big(r(x_1^\infty, Z_1^\infty, Q_\Delta) + \log \Delta \big) &\le \rho_{\mathrm{LZ}}(Q_{\Delta'}(x_1^\infty)) + \limsup_{\Delta \to 0} \log\left(\Delta' + 3\Delta\right) \\
&= \rho_{\mathrm{LZ}}(Q_{\Delta'}(x_1^\infty)) + \log \Delta'
\end{aligned}$$

which, combined with the definition of $h_{\mathrm{LZ}}(x_1^\infty)$, implies, with probability one,

$$\limsup_{\Delta \to 0} \big(r(x_1^\infty, Z_1^\infty, Q_\Delta) + \log \Delta \big) \le h_{\mathrm{LZ}}(x_1^\infty).$$

Combined with (29), this proves the theorem. □

4.3 A Shannon lower bound for individual sequences

For two sequences of real numbers x_1^∞ and $\hat{x}_1^\infty$, let

$$d(x_1^\infty, \hat{x}_1^\infty) \triangleq \limsup_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{x}_i)^2.$$

If x_1^∞ is a bounded sequence of real numbers and $D \geq 0$, let

$$\rho(x_1^\infty, D) \triangleq \inf_{\hat{x}_1^\infty:\, d(x_1^\infty,\hat{x}_1^\infty)\leq D} \rho_{\mathrm{LZ}}(\hat{x}_1^\infty)$$

where the infimum is over all sequences $\hat{x}_1^\infty$ from some finite set of reals (so that $\rho_{\mathrm{LZ}}(\hat{x}_1^\infty)$ is well defined) that satisfy $d(x_1^\infty, \hat{x}_1^\infty) \leq D$. In analogy to [13], where a similar quantity was defined with the finite-state fixed-rate complexity of $\hat{x}_1^\infty$ replacing the finite-state variable-rate complexity $\rho_{\mathrm{LZ}}(\hat{x}_1^\infty)$, we call $\rho(x_1^\infty, D)$ the variable-rate rate-distortion function of x_1^∞. Intuitively, $\rho(x_1^\infty, D)$ expresses the minimum achievable rate in encoding the individual sequence x_1^∞ with *unbounded delay* using a variable-rate finite-state encoders.

The following lower bound on $\rho(x_1^\infty, D)$ gives an individual-sequence version of the Shannon lower bound for stationary sources with finite entropy rate.

Theorem 5 *Assume x_1^∞ is a sequence with $x_n \in [0,1]$ for all n, and suppose $h_{\mathrm{LZ}}(x_1^\infty)$ is finite. Then for any $D > 0$,*

$$\rho(x_1^\infty, D) \geq h_{\mathrm{LZ}}(x_1^\infty) - \frac{1}{2}\log(2\pi e D).$$

Remarks.

1. Although we do not have a coding theorem showing the exact operational significance of $\rho(x_1^\infty, D)$ for individual sequences, it can be proved using results of Yang and Kieffer [30] that with probability one, $\rho(X_1^\infty, D) = R(D)$ for any bounded stationary and ergodic source X_1^∞ with rate-distortion function $R(D)$. Thus the theorem gives back the Shannon lower bound for sample paths of bounded stationary and ergodic sources with finite differential entropy.

2. Theorems 4 and 5 imply that for systems that allow subtractive dithering, the price of causality for small distortion is upper bounded by $(1/2)\log(2\pi e/12)$ bits per sample. It can also be shown that the lower bound of Theorem 5 is asymptotically tight in the sense that it can be asymptotically achieved with schemes using multidimensional dithered lattice quantization followed by Lempel-Ziv coding. Thus in the limit of small distortion, the price of causality is the same as in the probabilistic case; i.e., the rate loss of the cubic quantizer cell.

Proof of Theorem 5 First we note that finite-state compressibility preserves some important properties of (Shannon) entropy. In particular, if $\mathcal{Y}$ and $\mathcal{Z}$ are finite sets, $T : \mathcal{Y} \to \mathcal{Z}$ is an arbitrary function, y_1^∞ is sequence from $\mathcal{Y}$, and $T(y_1^\infty) \triangleq T(y_1), T(y_2), T(y_3), \ldots$, then

$$\rho_{\mathrm{LZ}}(T(y_1^\infty)) \le \rho_{\mathrm{LZ}}(y_1^\infty). \tag{31}$$

Note that equality must hold if T has an inverse. Furthermore, if u_1^∞ and y_1^∞ are sequences from the finite alphabets $\mathcal{U}$ and $\mathcal{Y}$, respectively, and $(u_1^\infty, y_1^\infty) \triangleq (u_1, y_1), (u_2, y_2), (u_3, y_3), \ldots$ (a sequence from the finite alphabet $\mathcal{U} \times \mathcal{Y}$), then we have

$$\rho_{\mathrm{LZ}}(u_1^\infty) \le \rho_{\mathrm{LZ}}(u_1^\infty, y_1^\infty) \le \rho_{\mathrm{LZ}}(u_1^\infty) + \rho_{\mathrm{LZ}}(y_1^\infty). \tag{32}$$

These inequalities follow directly from the characterization of finite-state compressibility of a sequence in terms of the empirical entropies of overlapping blocks, but for completeness they are proved in Appendix D.

Let $D > 0$ and $\hat{x}_1^\infty$ any sequence over a finite subset of reals such that $d(x_1^\infty, \hat{x}_1^\infty) \le D$. We will show that

$$\rho_{\mathrm{LZ}}(\hat{x}_1^\infty) \ge h_{\mathrm{LZ}}(x_1^\infty) - \frac{1}{2}\log(2\pi e D) \tag{33}$$

which clearly implies the theorem.

Note that we can assume that $\hat{x}_n \in [0, 1]$ for all n, since otherwise we can define

$$\tilde{x}_n = \begin{cases} 0 & \text{if } \hat{x}_n < 0 \\ \hat{x}_n & \text{if } 0 \le \hat{x}_n \le 1 \\ 1 & \text{if } \hat{x}_n > 1 \end{cases}$$

and replace $\hat{x}_1^\infty$ by $\tilde{x}_1^\infty$. The new sequence will satisfy $\rho_{\mathrm{LZ}}(\tilde{x}_1^\infty) \le \rho_{\mathrm{LZ}}(\hat{x}_1^\infty)$ by (31), and $d(x_1^\infty, \tilde{x}_1^\infty) \le d(x_1^\infty, \hat{x}_1^\infty)$ since $x_n \in [0, 1]$ for all n.

For $\delta > 0$ let $Q_\delta(x_1^\infty) - Q_\delta(\hat{x}_1^\infty)$ denote the sequence $\{Q_\delta(x_n) - Q_\delta(\hat{x}_n)\}_{n=1}^\infty$. First using (32), then applying (31) with the invertible mapping $T(u, v) = (u - v, v)$, and then using (32) again, we obtain

$$\begin{aligned} \rho_{\mathrm{LZ}}(Q_\delta(x_1^\infty)) &\le \rho_{\mathrm{LZ}}(Q_\delta(x_1^\infty), Q_\delta(\hat{x}_1^\infty)) \\ &= \rho_{\mathrm{LZ}}(Q_\delta(x_1^\infty) - Q_\delta(\hat{x}_1^\infty), Q_\delta(\hat{x}_1^\infty)) \\ &\le \rho_{\mathrm{LZ}}(Q_\delta(x_1^\infty) - Q_\delta(\hat{x}_1^\infty)) + \rho_{\mathrm{LZ}}(Q_\delta(\hat{x}_1^\infty)). \end{aligned}$$

Note also that by (31), $\rho_{\mathrm{LZ}}(Q_\delta(\hat{x}_1^\infty)) \le \rho_{\mathrm{LZ}}(\hat{x}_1^\infty)$. Hence

$$\rho_{\mathrm{LZ}}(\hat{x}_1^\infty) \ge \rho_{\mathrm{LZ}}(Q_\delta(x_1^\infty)) - \rho_{\mathrm{LZ}}(Q_\delta(x_1^\infty) - Q_\delta(\hat{x}_1^\infty)). \tag{34}$$

Let $H_{\max}(D,\delta)$ denote the maximum entropy of any discrete random variable with values in $A_\delta \triangleq \{0, \pm\delta, \pm 2\delta, \ldots\}$ having second moment at most D, i.e.,

$$H_{\max}(D,\delta) \triangleq \max\{H(Z) : \ \Pr(Z \in A_\delta) = 1 \text{ and } E(Z^2) \le D\}.$$

We show in Appendix D that $d(Q_\delta(x_1^\infty), Q_\delta(\hat{x}_1^\infty)) \le D + 3\delta$ for all $0 < \delta < 1$, and also that this implies

$$\rho_{\mathrm{LZ}}(Q_\delta(x_1^\infty) - Q_\delta(\hat{x}_1^\infty)) \le H_{\max}(D + 3\delta, \delta). \tag{35}$$

Hence by (34),

$$\rho_{\mathrm{LZ}}(\hat{x}_1^\infty) \ge \rho_{\mathrm{LZ}}(Q_\delta(x_1^\infty)) - H_{\max}(D + 3\delta, \delta).$$

We also show in Appendix D that for all $D > 0$,

$$\limsup_{\delta\to 0} H_{\max}(D,\delta) + \log\delta \le \frac{1}{2}\log(2\pi e D). \tag{36}$$

Since $H_{\max}(D,\delta)$ is monotone increasing in D for any fixed δ, the preceding implies

$$\limsup_{\delta\to 0} H_{\max}(D + 3\delta,\delta) + \log\delta \le \frac{1}{2}\log(2\pi e D).$$

Thus

$$\begin{aligned}
\rho_{\mathrm{LZ}}(\hat{x}_1^\infty) &\ge \limsup_{\delta\to 0}\big(\rho_{\mathrm{LZ}}(Q_\delta(x_1^\infty)) - H_{\max}(D+3\delta,\delta)\big) \\
&\ge \limsup_{\delta\to 0}\big(\rho_{\mathrm{LZ}}(Q_\delta(x_1^\infty)) + \log\delta\big) - \limsup_{\delta\to 0}\big(H_{\max}(D+3\delta,\delta)\big) + \log\delta\big) \\
&\ge h_{\mathrm{LZ}}(x_1^\infty) - \frac{1}{2}\log(2\pi e D)
\end{aligned}$$

where the last inequality follows from the definition of $h_{\mathrm{LZ}}(x_1^\infty)$. □

5 Concluding remarks

We extended results on causal coding by Neuhoff and Gilbert to (stationary) sources with memory, and to individual sequences encoded by complexity-limited systems, under high resolution conditions. The price of causality was identified in both cases as the space-filling loss of the cubic lattice cell; i.e., approximately 0.254 bits.

For the individual sequence setting we also derived a lower bound on the performance of non-causal encoding systems. The bound, which parallels the Shannon lower bound on the rate-distortion function, is based on the notion of Lempel-Ziv (finite-state) complexity of a discrete individual sequence. We note that similar results can be obtained using other sequence complexity measures (e.g., Kolmogorov complexity), provided they satisfy the two very intuitive properties used in the proof.

Our analyses focused on the high-resolution limit, which, in effect, allowed the decoupling of the quantizer's rate-distortion behavior from its ability to form contexts for entropy coding. It is worth noting that at the other extreme (that of high distortion) the price of causality is expected to be smaller. For example, at the maximum distortion the loss is zero since a memoryless scalar quantizer with one level (placed at the mean of the source) achieves optimum rate-distortion performance. At intermediate distortion values one can always bound the price of causality by the rate loss of an entropy-coded dithered scalar quantizer, which is at most (approximately) 0.754 bits [14, 15] at all distortion values.

In light of these results, one could use similar intuition and tools to analyze fixed-rate zero-delay encoding with high resolution. The corresponding asymptotic performance limit in this case should be given in terms of Bennett's integral (e.g., [20]). We conjecture that for stationary sources possessing a conditional pdf given the infinite past, a conditional version of Bennett's integral, calculated with respect to the conditional pdf and averaged over the condition, gives the minimum distortion in zero-delay coding with high resolution.

Appendix A

Proof of Lemma 2 Since $h(X_{-\infty}^{\infty})$ exists and is finite, the mutual information between X_1 and the past $X_{-\infty}^{0}$ is finite:

$$I(X_1; X_{-\infty}^0) = h(X_1) - h(X_1|X_{-\infty}^0) < \infty.$$

Also, the condition $H(Q_1(X_n)) < \infty$ implies that for all $\Delta > 0$, $H(Q_\Delta(X_{-\infty}^{\infty})) < \infty$ and

$$I(Q_\Delta(X_1); Q_\Delta(X_{-\infty}^0)) = H(Q_\Delta(X_1)) - H(Q_\Delta(X_1)|Q_\Delta(X_{-\infty}^0)) < \infty.$$

Since for any decreasing sequence $\{\Delta_m\}$ with $\lim_m \Delta_m = 0$, the partitions (quantizer cells) of $\{Q_{\Delta_m}\}$ asymptotically generate the Borel sigma field on the real line, by [17, Lemma 5.5.5] we have

$$\lim_{\Delta\to 0} I(Q_\Delta(X_1); X_{-\infty}^0) = I(X_1; X_{-\infty}^0).$$

Therefore

$$\begin{aligned} h(X_1) - h(X_1|X_{-\infty}^0) &= \lim_{\Delta\to 0} I(Q_\Delta(X_1); X_{-\infty}^0) \\ &= \lim_{\Delta\to 0} \big[\, H(Q_\Delta(X_1)) - H(Q_\Delta(X_1)|X_{-\infty}^0) \,\big] \\ &= \lim_{\Delta\to 0} \big[\, H(Q_\Delta(X_1)) + \log\Delta - H(Q_\Delta(X_1)|X_{-\infty}^0) - \log\Delta \,\big] \end{aligned}$$

$$= \quad h(X_1) - \lim_{\Delta \to 0} \big[H(Q_\Delta(X_1)|X_{-\infty}^0) - \log \Delta \big]$$

where the last equality follows from Lemma 1. Hence we obtain

$$\begin{aligned} &\liminf_{\Delta \to 0} \big[H(Q_\Delta(X_{-\infty}^\infty)) + \log \Delta \big] \\ &\qquad = \quad \liminf_{\Delta \to 0} \big[H(Q_\Delta(X_1)|Q_\Delta(X_{-\infty}^0)) + \log \Delta \big] \\ &\qquad \geq \quad \liminf_{\Delta \to 0} \big[H(Q_\Delta(X_1)|X_{-\infty}^0) + \log \Delta \big] \\ &\qquad = \quad h(X_1|X_{-\infty}^0) = h(X_{-\infty}^\infty). \end{aligned}$$

To prove a reverse inequality, note that by stationarity,

$$H(Q_\Delta(X_{-\infty}^\infty)) \leq \bar{H}(Q_\Delta(X_1^n))$$

for any $n \geq 1$. Thus by Lemma 1,

$$\limsup_{\Delta \to 0} \big[(H(Q_\Delta(X_{-\infty}^\infty)) + \log \Delta \big] \leq \bar{h}(X_1^n).$$

As $n \to \infty$, the right-hand-side converges to $h(X_{-\infty}^\infty)$. Thus

$$\limsup_{\Delta \to 0} \big[H(Q_\Delta(X_{-\infty}^\infty)) + \log \Delta \big] \leq h(X_{-\infty}^\infty)$$

which completes the proof. □

Proof of Lemma 3 We need the following fact characterizing $r_m(D)$ in the limit of low distortion. The proposition is essentially due to Zador [26, 27] and Gish and Pierce [3]; it was proved with the present general conditions in [31].

Proposition 1 *If X is a real random variable with a pdf such that $h(X)$ and $H(Q_1(X))$ are finite, then*

$$\lim_{D \to 0} \big(r_m(D) + \frac{1}{2} \log(12D) \big) = h(X).$$

To prove the lemma, it suffices to show that if the family of functions $\{g_D; D > 0\}$ satisfies $E(X_1 - g_D(X_{-\infty}^1))^2 \leq D$ for all $D > 0$, then

$$\liminf_{D \to 0} \big(H(g_D(X_{-\infty}^1)|X_{-\infty}^0) + \frac{1}{2} \log(12D) \big) \geq h(X_1|X_{-\infty}^0).$$

To simplify the notation, let Y denote $X_{-\infty}^0$, and let y denote a particular realization $x_{-\infty}^0$. Let $P_{X_1|Y=y}$ denote the conditional distribution of X_1 given the infinite past $Y = y$, and note that $P_{X_1|Y=y}$ exist as a regular conditional probability [32]. Define

$$d_D(y) = E[(X_1 - g_D(X_1, Y))^2 | Y = y].$$

Since $E[d_D(Y)] \le D$ and by the concavity of the logarithm, we have

$$\begin{aligned} H(g_D(X_{-\infty}^1)|X_{-\infty}^0) + \frac{1}{2}\log(12D) \\ \ge \int \left[H(g_D(X_1,Y)|Y=y) + \frac{1}{2}\log(12 d_D(y))\right] d\mu(y) \end{aligned}$$

where μ denotes the distribution of $Y = X_{-\infty}^0$. Thus it suffices to show that

$$\liminf_{D\to 0} \int \left[H(g_D(X_1,Y)|Y=y) + \frac{1}{2}\log(12 d_D(y))\right] d\mu(y) \ge h(X_1|Y). \tag{A.1}$$

The finiteness of $h(X_1|Y)$ implies that $P_{X_1|Y=y}$ is absolutely continuous with pdf $f_{X_1|Y}(x_1|y)$ for μ-almost all y. For any y and $d > 0$ let $r_m(d, P_{X_1|Y=y})$ denote the OPTA of entropy-constrained scalar quantizers for a random variable X with distribution $P_{X_1|Y=y}$ and differential entropy $h(P_{X_1|Y=y}) = h(X_1|Y=y)$ (see definition (3)). Furthermore, define

$$F(y,d) = r_m(d, P_{X_1|Y=y}) + \frac{1}{2}\log(2\pi e d) - h(X_1|Y=y).$$

By definition, $H(g_D(X_1,Y)|Y=y) \ge r_m(d_D(y), P_{X_1|Y=y})$. Thus (A.1) holds if

$$\liminf_{D\to 0} \int F(y, d_D(y))\, d\mu(y) \ge \frac{1}{2}\log\left(\frac{2\pi e}{12}\right) \triangleq c. \tag{A.2}$$

The rest of the proof is devoted to showing that (A.2) holds.

Observe that by the conditions of the lemma, both $h(X_1|Y=y)$ and $H(Q_1(X_1)|Y=y]$ are finite for μ-almost all y. Therefore Proposition 1 implies that for μ-almost all y,

$$\liminf_{d\to 0} F(y,d) \ge c. \tag{A.3}$$

Also, by the Shannon lower bound (6), for μ-almost all y

$$F(y,d) \ge 0 \quad \text{for all } d > 0. \tag{A.4}$$

For any positive integer k and $D, \delta > 0$, define the sets

$$A_{D,k} \triangleq \{y : d_D(y) < 1/k\}$$

and

$$B_{\delta,k} \triangleq \{y :\ F(y,d) > c - \delta \ \text{ for all } \ d \in (0, 1/k)\}.$$

Then, using (A.4),

$$\int F(y, d_D(y))\, d\mu(y) \ \ge \ \int_{A_{D,k}\cap B_{\delta,k}} F(y, d_D(y))\, d\mu(y)$$

$$\geq \mu(A_{D,k} \cap B_{\delta,k})(c - \delta).$$

Since $d_D(y)$ is nonnegative and $E[d_D(Y)] \leq D$, Markov's inequality implies that $\lim_{D\to 0} \mu(A_{D,k}) = 1$ for all $k \geq 1$. Hence,

$$\begin{aligned} \liminf_{D\to 0} \int F(y, d_D(y))\, d\mu(y) &\geq \liminf_{D\to 0} \mu(A_{D,k} \cap B_{\delta,k})(c - \delta) \\ &= \mu(B_{\delta,k})(c - \delta). \qquad \text{(A.5)} \end{aligned}$$

Since

$$\{y : \liminf_{d\to 0} F(y, d) \geq c\} \subset \bigcup_{k\geq 1} B_{\delta,k}$$

we have $\mu(\bigcup_{k\geq 1} B_{\delta,k}) = 1$ for all $\delta > 0$ by (A.3). Since $B_{\delta,k} \subset B_{\delta,k'}$ if $k < k'$, the continuity of μ as a set function implies that $\lim_{k\to\infty} \mu(B_{\delta,k}) = 1$. Thus letting $k \to \infty$ in (A.5), we obtain

$$\liminf_{D\to 0} \int F(y, d_D(y))\, d\mu(y) \geq c - \delta$$

which completes the proof since $\delta > 0$ was arbitrary. □

Proof of (13): By appropriate shifting, normalization, and scaling, it suffices to show that if $r(t)$, $t > 0$ is a positive nonincreasing function such that

$$\liminf_{t\to 0} \big(r(t) + \ln t\big) \geq 0$$

then its lower convex hull $\bar{r}(t)$ satisfies

$$\liminf_{t\to 0} \big(\bar{r}(t) + \ln t\big) \geq 0. \qquad \text{(A.6)}$$

We prove (A.6) by contradiction. If (A.6) does not hold, then there is an $\epsilon > 0$ and a sequence of decreasing positive numbers t_n, $n = 1, 2 \ldots$, with $\lim_n t_n = 0$ such that

$$\bar{r}(t_n) \leq -\ln t_n - \epsilon \qquad \text{(A.7)}$$

for all n. Now consider the affine functions

$$g_{t_n,\epsilon}(t) = 1 - \frac{t}{t_n} - \ln t_n - \epsilon/2$$

that represent the lines supporting the convex function $-\ln t - \epsilon/2$ at the points $t = t_n$ (i.e., $g_{t_n,\epsilon}(t_n) = -\ln t_n - \epsilon/2$ and $g_{t_n,\epsilon}(t) \leq -\ln t - \epsilon/2$ for all $t > 0$). Let $t^* > 0$ be such that $r(t) \geq -\ln t - \epsilon/2$ if $0 < t \leq t^*$. Since $g_{t_n,\epsilon}(t)$ is strictly decreasing and $g_{t_n,\epsilon}(t) = 0$ at $t = t_n(1 - \epsilon/2) - t_n \ln t_n$, by choosing n large enough (so that t_n is small enough) we have $g_{t_n,\epsilon}(t) \leq 0$ for all $t > t^*$. Hence, we have for $0 < t \leq t^*$

$$g_{t_n,\epsilon}(t) \leq -\ln t - \epsilon/2 \leq r(t)$$

and for $t > t^*$

$$g_{t_n,\epsilon}(t) \le 0 \le r(t).$$

Thus $g_{t_n,\epsilon}(t) \le r(t)$ for all $t > 0$. Since $\bar{r}(t)$ is the pointwise supremum of all affine functions that are majorized by $r(t)$, it follows that $\bar{r}(t) \ge g_{t_n,\epsilon}(t)$ for all $t > 0$. But from (A.7) we have

$$\bar{r}(t_n) \le -\ln t_n - \epsilon < -\ln t_n - \epsilon/2 = g_{t_n,\epsilon}(t_n)$$

a contradiction. □

Appendix B

Proof of Theorem 2 First we construct the desired stationary process. Let $\{\Delta_k\}_{k=1}^{\infty}$ be a decreasing sequence of positive numbers converging to zero such that

$$h_{\mathrm{LZ}}(x_1^\infty) = \lim_{k\to\infty} \big[\rho_{\mathrm{LZ}}(Q_{\Delta_k}(x_1^\infty)) + \log \Delta_k\big].$$

From (15) and (16) we have

$$\begin{aligned} h_{\mathrm{LZ}}(x_1^\infty) &= \lim_{k\to\infty} \big(\lim_{l\to\infty} \bar{H}_l(Q_{\Delta_k}(x_1^\infty)) + \log \Delta_k \big) \\ &= \lim_{k\to\infty} \big(\lim_{l\to\infty} \limsup_{n\to\infty} \bar{H}(\hat{P}^l_{Q_{\Delta_k}(x_1^n)}) + \log \Delta_k \big). \end{aligned}$$

Recall that $l\bar{H}_l(y_1^\infty)$ is subadditive in l. Thus $\lim_l \bar{H}_l(y_1^\infty) = \inf_l \bar{H}_l(y_1^\infty)$, so we have for all l,

$$h_{\mathrm{LZ}}(x_1^\infty) \le \limsup_{k\to\infty} \big(\limsup_{n\to\infty} \bar{H}(\hat{P}^l_{Q_{\Delta_k}(x_1^n)}) + \log \Delta_k \big) \triangleq L(l). \tag{B.1}$$

Now note that $\hat{P}^l_{x_1^n}$ is supported in the hypercube $[0,1]^l$, so the family of probability measures $\{\hat{P}^l_{x_1^n};\, n = l, l+1, \ldots\}$ is uniformly tight. Therefore Prokhorov's theorem [32] implies that every subsequence of $\hat{P}^l_{x_1^n}$, $n = 1, 2, \ldots$ has a sub-subsequence, say $\hat{P}^l_{x_1^{n_k}}$, $k = 1, 2, \ldots$ converging weakly to some $P \in \mathcal{P}^l$. In particular, $\mathcal{P}^l(x_1^\infty)$ is nonempty for all l. It follows that for each l there exists a $P^l \in \mathcal{P}^l(x_1^\infty)$ and a subsequence $\{n_k\}$ (which depends on l) such that

$$\hat{P}^l_{x_1^{n_k}} \Rightarrow P^l \quad \text{and} \quad \limsup_{k\to\infty} \big(\bar{H}(\hat{P}^l_{Q_{\Delta_k}(x_1^{n_k})}) + \log \Delta_k \big) = L(l). \tag{B.2}$$

The collection $\{P^l;\, l \ge 1\}$ thus obtained will play an important role in the subsequent proof.

Let $a_k \triangleq \Delta_k \lceil \frac{1}{\Delta_k} \rceil$ and u_k^l denote the uniform distribution (Lebesgue measure) on $[0, a_k]^l$. Also, let u^l denote the uniform distribution on $[0,1]^l$ and Q^l_Δ the l-fold product of the uniform quantizer Q_Δ. Then the induced distribution $u_k^l \circ (Q^l_{\Delta_k})^{-1}$ (recall definition (23)), is the uniform distribution on $Q^l_{\Delta_k}([0,1]^l)$, a set of cardinality $\lceil \frac{1}{\Delta_k} \rceil^l$. Thus we have

$$\bar{H}(\hat{P}^l_{Q_{\Delta_k}(x_1^{n_k})}) + \log\left(\frac{\Delta_k}{a_k}\right) = -\frac{1}{l} D(\hat{P}^l_{Q_{\Delta_k}(x_1^{n_k})} \| u_k^l \circ (Q^l_{\Delta_k})^{-1})$$

where $D(P\|P')$ denotes the relative entropy (Kullback-Leibler divergence) [16, 17] between two probability measures P and P'. In Appendix C we show that $\hat{P}^l_{x_1^{n_k}} \Rightarrow P^l$ implies

$$\hat{P}^l_{Q_{\Delta_k}(x_1^{n_k})} \Rightarrow P^l. \tag{B.3}$$

Since $a_k \to 1$ as $k \to \infty$, it follows similarly that $u_k^l \circ (Q^l_{\Delta_k})^{-1} \Rightarrow u^l$. Thus from (B.1) and (B.2),

$$\begin{aligned} h_{\mathrm{LZ}}(x_1^\infty) \le L(l) &= \limsup_{k\to\infty} \big(\bar{H}(\hat{P}^l_{Q_{\Delta_k}(x_1^{n_k})}) + \log \Delta_k\big) \\ &= \limsup_{k\to\infty} \left(-\frac{1}{l} D(\hat{P}^l_{Q_{\Delta_k}(x_1^{n_k})} \| u_k^l \circ (Q^l_{\Delta_k})^{-1}) - \log a_k\right) \\ &= -\liminf_{k\to\infty} \frac{1}{l} D(\hat{P}^l_{Q_{\Delta_k}(x_1^{n_k})} \| u_k^l \circ (Q^l_{\Delta_k})^{-1}) \\ &\le -\frac{1}{l} D(P^l \| u^l) \end{aligned} \tag{B.4}$$

where P^l is defined in (B.2), and the inequality follows from the lower semicontinuity (with respect to weak convergence) of the relative entropy [33]. For any $P \in \mathcal{P}^l$ supported on $[0,1]^l$, the relative entropy $D(P\|u^l)$ is finite if and only if P has a pdf and finite differential entropy, in which case $h(P) = -D(P\|u^l)$. Hence (B.4) implies that $P^l \in P^l_a(x_1^\infty)$ (thus $\mathcal{P}^l_a(x_1^\infty)$ is nonempty) and it has finite differential entropy which is bounded as

$$\bar{h}(P^l) \ge L(l) \ge h_{\mathrm{LZ}}(x_1^\infty). \tag{B.5}$$

Now let $P \in \mathcal{P}^l_a(x_1^\infty)$ be arbitrary and $\{n_j\}$ a subsequence such that $\hat{P}^l_{x_1^{n_j}} \Rightarrow P$. Then

$$\begin{aligned} L(l) = \limsup_{k\to\infty} \big(\limsup_{n\to\infty} \bar{H}(\hat{P}^l_{Q_{\Delta_k}(x_1^n)}) + \log \Delta_k\big) & \\ \ge\ \limsup_{k\to\infty} \big(\lim_{j\to\infty} \bar{H}(\hat{P}^l_{Q_{\Delta_k}(x_1^{n_j})}) + \log \Delta_k\big) & \\ =\ \limsup_{k\to\infty} \big(\bar{H}(P \circ (Q^l_{\Delta_k})^{-1}) + \log \Delta_k\big) & \\ =\ \bar{h}(P) & \end{aligned}$$

where the second equality follows from Lemma 1 and the first equality holds since P has a pdf and the discontinuities of Q^l_Δ form a set of Lebesgue measure zero, and so from [28, Thm. 5.1], $\hat{P}^l_{x_1^{n_j}} \Rightarrow P$ implies that as $j \to \infty$,

$$\hat{P}^l_{Q_{\Delta_k}(x_1^{n_j})} = \hat{P}^l_{x_1^{n_j}} \circ (Q^l_{\Delta_k})^{-1} \Rightarrow P \circ (Q^l_{\Delta_k})^{-1}. \tag{B.6}$$

Thus $\bar{h}(P) \leq L(l)$ for all $P \in \mathcal{P}_a^l(x_1^\infty)$. Since $\bar{h}(P^l) \geq L(l)$ by (B.5), this implies

$$\bar{h}(P^l) = \sup_{P \in \mathcal{P}_a^l(x_1^\infty)} \bar{h}(P). \tag{B.7}$$

Next, using the collection $\{P^l;\, l \geq 1\}$, we construct the desired stationary process $\{X_n\}$ with marginal distributions in $\mathcal{P}_a^l(x_1^\infty)$. For $m > l$, let P_l^m denote the l-dimensional marginal of P^m corresponding to the first l coordinates; i.e., $P_l^m(B) = P^m(B \times \mathbb{R}^{m-l})$ for any measurable $B \subset \mathbb{R}^l$. Since each P_l^m is supported in $[0,1]^l$, for each l the family $\{P_l^m; m = l+1, l+2, \ldots\}$ is uniformly tight. Thus we can use Cantor's diagonal method to pick a subsequence $\{m_j\}$ of the positive integers such that for all $l \geq 1$,

$$P_l^{m_j} \Rightarrow P_l \ \text{ for some } P_l \in \mathcal{P}^l. \tag{B.8}$$

We show that the marginals $\{P_l;\, l = 1, 2, \ldots\}$ define a stationary process that satisfies the theorem statement. Recall that $\mathcal{P}^l(x_1^\infty)$ is the set of subsequential limits (with respect to weak convergence) of the sequence $\hat{P}^l_{x_1^n};\, n = l, l+1, \ldots$. Since the weak convergence of probability measures on a Euclidean space is metrizable [32], it follows that $\mathcal{P}^l(x_1^\infty)$ is closed under weak convergence. As shown in Appendix C,

$$P_l^m \in \mathcal{P}^l(x_1^\infty) \ \text{ for all } l \geq 1 \text{ and } m > l \tag{B.9}$$

and hence $P_l^{m_j} \Rightarrow P_l$ implies that $P_l \in \mathcal{P}^l(x_1^\infty)$. By construction, the family of finite-dimensional distributions $\{P_l; l = 1, 2, \ldots\}$ is consistent in the usual sense: for all $l \geq 1$ and $l' > l$, $P_l(B) = P_{l'}(B \times \mathbb{R}^{l'-l})$ for all measurable $B \subset \mathbb{R}^l$. Thus by the Kolmogorov extension theorem there exists a stochastic process $\{X_n\}_{n=1}^\infty$ with marginals $X_1^l \sim P_l$. Furthermore, note that $P_l \in \mathcal{P}^l(x_1^\infty)$ means that each P_l is the limit of sliding-block empirical distributions, and as such is stationary in the sense that if $X_1^l \sim P_l$, then for any $l' < l$, the l'-blocks $X_1^{l'}, X_2^{l'+1}, \ldots, X_{l-l'+1}^l$ have identical distribution. Hence $\{X_n\}_{n=1}^\infty$ is a stationary process.

We prove the first equality in (19) via matching upper and lower bounds. Fix $m \geq 1$ and let $Z_1^m = (Z_1, \ldots, Z_m)$ be jointly distributed according to P^m (defined in equation (B.2)). Since $P^m \in \mathcal{P}^m(x_1^\infty)$, we have $h(Z_1^l) = h(Z_{1+j}^{l+j})$ for all $1 \leq j \leq m - l$. Thus, writing m as $m = Nl + i$ for integers $N \geq 1$ and $0 \leq i < l$, we have

$$\begin{aligned}
\bar{h}(P^m) &= \frac{1}{m} h(Z_1^m) \\
&\leq \frac{1}{m}\Big(h(Z_1^l) + h(Z_{l+1}^{2l}) + \cdots h(Z_{(N-1)l+1}^{Nl}) + h(Z_{Nl+1}^m) \Big) \\
&\leq \frac{N}{m} h(Z_1^l)
\end{aligned}$$

$$\leq \frac{1-l/m}{l} h(Z_1^l)$$

where the second and third inequalities hold since the differential entropies are nonpositive since each Z_i is supported in $[0,1]$. Since $Z_1^l \sim P_l^m$, we have $h(Z_1^l) = h(P_l^m)$ for all $l < m$. Hence (B.5) implies that for all $m > l$,

$$h_{\mathrm{LZ}}(x_1^\infty) \leq \bar{h}(P^m) \leq \left(1 - \frac{l}{m}\right) \bar{h}(P_l^m). \tag{B.10}$$

Thus for the subsequence $\{m_j\}$ associated with the P_l in (B.8), similarly to (B.4), we obtain

$$\begin{aligned} h_{\mathrm{LZ}}(x_1^\infty) &\leq \limsup_{j\to\infty} \bar{h}(P_l^{m_j}) \\ &= \limsup_{j\to\infty} -\frac{1}{l} D(P_l^{m_j} \| u^l) \\ &= -\liminf_{j\to\infty} \frac{1}{l} D(P_l^{m_j} \| u^l) \\ &\leq -\frac{1}{l} D(P_l \| u^l) \\ &= \bar{h}(P_l) \end{aligned} \tag{B.11}$$

where the last equality holds since the preceding inequalities show that $D(P_l \| u^l)$ is finite, so $P_l \in \mathcal{P}_a^l(x_1^\infty)$. Since $h(P_l) = h(X_1^l)$, the above implies

$$h_{\mathrm{LZ}}(x_1^\infty) \leq \lim_{l\to\infty} \bar{h}(X_1^l). \tag{B.12}$$

To show the reverse inequality, recall that $P_l^{m_j} \in \mathcal{P}_a^l(x_1^\infty)$, so there is a subsequence $\{n_i\}$ such that $P_{x_1^{n_i}}^l \Rightarrow P_l^{m_j}$. Since $P_l^{m_j}$ has a pdf, similarly to (B.6), we have $P_{x_1^{n_i}}^{m_j} \circ (Q_\Delta^l)^{-1} \Rightarrow P_l^{m_j} \circ (Q_\Delta^l)^{-1}$. Hence

$$\begin{aligned} \limsup_{n\to\infty} \bar{H}_l(Q_\Delta(x_1^n)) &= \limsup_{n\to\infty} \bar{H}(\hat{P}_{x_1^n}^l \circ (Q_\Delta^l)^{-1}) \\ &\geq \lim_{i\to\infty} \bar{H}(\hat{P}_{x_1^{n_i}}^l \circ (Q_\Delta^l)^{-1}) \\ &= \bar{H}(P_l^{m_j} \circ (Q_\Delta^l)^{-1}) \end{aligned}$$

Since $P_l^{m_j} \Rightarrow P_l \sim X_1^l$, this implies

$$\begin{aligned} \limsup_{n\to\infty} \bar{H}_l(Q_\Delta(x_1^n)) &\geq \lim_{j\to\infty} \bar{H}(P_l^{m_j} \circ (Q_\Delta^l)^{-1}) \\ &= \bar{H}(P_l \circ (Q_\Delta^l)^{-1}) \\ &= \bar{H}(Q_\Delta(X_1^l)). \end{aligned}$$

Thus we obtain

$$\begin{aligned} h_{\mathrm{LZ}}(x_1^\infty) &= \limsup_{\Delta\to 0}\big(\lim_{l\to\infty}\limsup_{n\to\infty}\bar{H}_l(Q_\Delta(x_1^n)) + \log\Delta\big) \\ &\geq \limsup_{\Delta\to 0}\big(\lim_{l\to\infty}\bar{H}(Q_\Delta(X_1^l)) + \log\Delta\big) \\ &= \lim_{l\to\infty}\bar{h}(X_1^l) \end{aligned} \tag{B.13}$$

where the last equality holds by Lemma 2. Combined with (B.12), this proves the first equality in (19).

To show the second equality in (19), note that by (B.5) and (B.7) we have for all $l \geq 1$

$$h_{\mathrm{LZ}}(x_1^\infty) \leq \sup_{P\in\mathcal{P}_a^l(x_1^\infty)} \bar{h}(P).$$

Conversely, since $P_l^m \in \mathcal{P}_a^l(x_1^\infty)$ and $\bar{h}(P^l) = \sup_{P\in\mathcal{P}_a^l(x_1^\infty)}\bar{h}(P)$, (B.10) implies

$$\bar{h}(P^m) \leq \left(1 - \frac{l}{m}\right)\bar{h}(P^l)$$

for all $m > l$. Thus the limit $\lim_m \bar{h}(P^m)$ exists, and from (B.10) and (B.11) we obtain

$$\bar{h}(X_1^l) \geq \lim_{m\to\infty}\bar{h}(P^m) = \lim_{m\to\infty}\sup_{P\in\mathcal{P}_a^m(x_1^\infty)}\bar{h}(P)$$

Combining these bounds with (B.12) and (B.13) proves the second equality in (19). □

Appendix C

Proof of (B.3): Recall that $P_n \Rightarrow P$ if and only if $\int g\,dP_n \to \int g\,dP$ for any bounded and continuous real function g. Pick such a g and note that we can also assume that g has a compact support since a large enough hypercube contains the support of all $\hat{P}^l_{Q_{\Delta_k}(x_1^{n_k})}$. We have

$$\left|\int g\,d\hat{P}^l_{Q_{\Delta_k}(x_1^{n_k})} - \int g\,d\hat{P}^l_{x_1^{n_k}}\right| \leq \frac{1}{n_k - l + 1}\sum_{i=1}^{n_k-l+1}\left|g(Q_{\Delta_k}(x_i^{i+l-1})) - g(x_i^{i+l-1})\right|$$

Since g is uniformly continuous and $\|Q_{\Delta_k}(x_i^{i+l-1}) - x_i^{i+l-1}\| \leq \sqrt{l}\Delta_k/2$, the right-hand side converges to zero as $k\to\infty$. Thus $\hat{P}^l_{Q_{\Delta_k}(x_1^{n_k})} \Rightarrow P^l$ if and only if $\hat{P}^l_{x_1^{n_k}} \Rightarrow P^l$. □

Proof of (B.9): We show that $P_l^m \in \mathcal{P}^l(x_1^\infty)$ for all $m > l$. Let $g_1 : \mathbb{R}^l \to \mathbb{R}$ be bounded and continuous and define $g : \mathbb{R}^m \to \mathbb{R}$ by $g(x_1^m) = g_1(x_1^l)$ for all $x_1^m \in \mathbb{R}$. Then $g : \mathbb{R}^m \to \mathbb{R}$ is bounded and continuous. Suppose $P^m_{x_1^{n_i}} \Rightarrow P^m$. Then,

$$\begin{aligned}
\int g\, d\hat{P}^m_{x_1^{n_i}} &= \frac{1}{n_i - m + 1} \sum_{i=1}^{n_i - m + 1} g(x_i^{i+m-1}) \\
&= \frac{1}{n_i - m + 1} \sum_{i=1}^{n_i - m + 1} g_1(x_i^{i-l+1}) \\
&= \frac{1}{n_i - l + 1} \frac{n_i - l + 1}{n_i - m + 1} \left(\sum_{i=1}^{n_i - l + 1} g_1(x_i^{i+l-1}) - \sum_{i=n_i - m + 2}^{n_i - l + 1} g_1(x_i^{i+l-1}) \right) \\
&= (1 + a_i) \int g_1\, d\hat{P}^l_{x_1^{n_i}} + b_i
\end{aligned}$$

where $a_i \to 0$ and $b_i \to 0$ as $i \to \infty$. Also,

$$\lim_{i\to\infty} \int g\, d\hat{P}^m_{x_1^{n_i}} = \int g\, dP^m = \int g_1\, dP^m_l.$$

Thus if $\hat{P}^m_{x_1^{n_i}} \Rightarrow P^m$, then $\hat{P}^l_{x_1^{n_i}} \Rightarrow P^m_l$, and so $P^m_l \in \mathcal{P}^l_a(x_1^\infty)$. □

Appendix D

Proof of (31) and (32): To show the first inequality, let $T(a) \triangleq (T(a_1), \ldots, T(a_l))$ for any $a = (a_1, \ldots, a_l) \in \mathcal{Y}^l$ and $l \geq 1$. Fix $n \geq l$ and let Y be any $\mathcal{Y}^l$-valued random variable with distribution $\hat{P}^l_{y_1^n}$. It is easy to check that $T(Y)$ has distribution $\hat{P}^l_{T(y_1^n)}$, so the well known inequality $H(T(Y)) \leq H(Y)$ gives

$$H(\hat{P}^l_{T(y_1^n)}) \leq H(\hat{P}^l_{y_1^n})$$

implying

$$\begin{aligned}
\rho_{\mathrm{LZ}}(T(y_1^\infty)) &= \lim_{l\to\infty} \limsup_{n\to\infty} \bar{H}(\hat{P}^l_{T(y_1^n)}) \\
&\leq \lim_{l\to\infty} \limsup_{n\to\infty} \bar{H}(\hat{P}^l_{y_1^n}) = \rho_{\mathrm{LZ}}(y_1^\infty). \qquad \text{(D.1)}
\end{aligned}$$

To show (32), let (U, Y) be a $\mathcal{U}^l \times \mathcal{Y}^l$-valued pair of random variables with distribution $\hat{P}^l_{(u_1^n, y_1^n)}$. Then U and Y have distributions $\hat{P}^l_{u_1^n}$ and $\hat{P}^l_{y_1^n}$, respectively, so from the corresponding inequality for the entropy of random variables, for all $n \geq l$,

$$H(\hat{P}^l_{u_1^n}) \leq H(\hat{P}^l_{(u_1^n, y_1^n)}) \leq H(\hat{P}^l_{u_1^n}) + H(\hat{P}^l_{y_1^n})$$

from which (32) follows similarly to (D.1). □

Proof of (35): Since $x_n, \hat{x}_n \in [0,1]$ for all n, we have for arbitrary $\delta \in (0,1)$,

$$|Q_\delta(x_n) - Q_\delta(\hat{x}_n)|^2 \leq (\delta + |x_n - \hat{x}_n|)^2 \leq 3\delta + |x_n - \hat{x}_n|^2$$

so that

$$d(Q_\delta(x_1^\infty), Q_\delta(\hat{x}_1^\infty)) \leq D + 3\delta.$$

Let $z_n \triangleq Q_\delta(x_n) - Q_\delta(\hat{x}_n)$ for all n, and $\{n_k\}$ be a subsequence such that

$$\limsup_{n\to\infty} H(P^1_{z_1^n}) = \lim_{k\to\infty} H(P^1_{z_1^{n_k}})$$

and

$$P^1_{z_1^{n_k}} \Rightarrow P \quad \text{for some } P \in \mathcal{P}^1(z_1^\infty).$$

Since all elements of z_1^∞ are from the finite set $A = A_\delta \cap [0,1]$, the $P^1_{z_1^n}$, as well as P, are concentrated on A. Thus, recalling that $l\bar{H}_l(z_1^\infty)$ is subadditive in l, we have

$$H(P) = \lim_{k\to\infty} H(P^1_{z_1^{n_k}}) = \bar{H}_1(z_1^\infty) \geq \lim_{l\to\infty} \bar{H}_l(z_1^\infty) = \rho_{\mathrm{LZ}}(z_1^\infty) \tag{D.2}$$

and furthermore

$$\begin{aligned}\int t^2 dP(t) = \lim_{k\to\infty} \int t^2 dP^1_{z_1^{n_k}}(t) &= \lim_{k\to\infty} \frac{1}{n_k}\sum_{i=1}^{n_k} z_i^2 \\ &\leq \limsup_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} z_i^2 \\ &= d(Q_\delta(x_n), Q_\delta(\hat{x}_n)) \\ &\leq D + 3\delta.\end{aligned}$$

Since $P(A_\delta) = 1$, it follows that

$$H(P) \leq H_{\max}(D + 3\delta, \delta).$$

Combining this with (D.2) proves (35). □

Proof of (36): We use differential entropy to bound discrete entropy as in [16, Thm. 9.7.1]. Let $Z_{D,\delta}$ be an A_δ-valued discrete random variable achieving $H_{\max}(D,\delta)$. (Although we will not need the specific form of the distribution, it can be shown that $\Pr(Z_{D,\delta} = i\delta) = ae^{-b(i\delta)^2}$ with constants a and b such that $E[Z_{D,\delta}^2] = D$.) Let U_δ be independent of $Z_{D,\delta}$ and uniformly distributed on the interval $(-\delta/2, \delta/2]$. Then, since in each interval of length δ centered at $i\delta$, the pdf of $Z_{D,\delta} + U_\delta$ is constant with magnitude $\frac{1}{\delta}\Pr(Z_{D,\delta} = i\delta)$, we have

$$h(Z_{D,\delta} + U_\delta) = H(Z_{\Delta,\delta}) + \log \delta.$$

Also, by independence,

$$E(Z_{D,\delta} + U_\delta)^2 = E[Z_{D,\delta}^2] + E[U_\delta^2] \le D + \frac{\delta^2}{12}$$

which implies

$$h(Z_{D,\delta} + U_\delta) \le \frac{1}{2}\log(2\pi e(D + \delta/12))$$

since the Gaussian maximizes differential entropy over all pdf's satisfying a second moment constraint [16]. Combining these we obtain

$$\begin{aligned} \limsup_{\delta\to 0}\big(H_{\max}(D,\delta) + \log\delta\big) &= \limsup_{\delta\to 0} h(Z_{D,\delta} + U_\delta) \\ &\le \limsup_{\delta\to 0} \frac{1}{2}\log(2\pi e(D + \delta/12)) \\ &= \frac{1}{2}\log(2\pi e D) \end{aligned}$$

which completes the proof. □

Acknowledgement

The authors wish to thank András György for his many helpful comments.

References

[1] T. Berger, *Rate Distortion Theory.* Englewood Cliffs, New Jersey: Prentice–Hall, 1971.

[2] T. D. Lookabaugh and R. M. Gray, "High-resolution quantization theory and the vector quantizer advantage," *IEEE Trans. Inform. Theory*, vol. IT-35, pp. 1020–1033, Sep. 1989.

[3] H. Gish and J. N. Pierce, "Asymptotically efficient quantizing," *IEEE Trans. Inform. Theory*, vol. IT-14, pp. 676–683, Sep. 1968.

[4] T. Ericson, "A result on delayless information transmission." IEEE Int. Symp. Inform. Theory, Grignano, Italy, 1979.

[5] N. T. Gaarder and D. Slepian, "On optimal finite-state digital transmission systems." IEEE Int. Symp. Inform. Theory, Grignano, Italy, 1979.

[6] N. T. Gaarder and D. Slepian, "On optimal finite-state digital transmission systems," *IEEE Trans. Information Theory*, vol. 28, pp. 167–186, Mar. 1982.

[7] G. Gábor and Z. Györfi, *Recursive Source Coding.* New York: Springer-Verlag, 1986.

[8] T. Linder and G. Lugosi, "A zero-delay sequential scheme for lossy coding of individual sequences," *IEEE Trans. Inform. Theory*, vol. IT-47, pp. pp. 2533–2538, Sep. 2001.

[9] T. Weissman and N. Merhav, "On delay-limited lossy coding and filtering of individual sequences," *IEEE Trans. Inform. Theory*, vol. 48, pp. 5721–733, Mar. 2002.

[10] N. Merhav and I. Kontoyiannis, "Source coding exponents for zero-delay coding with finite memory," *IEEE Trans. Inform. Theory*, vol. IT-49, pp. 609–625, Mar. 2003.

[11] D. L. Neuhoff and R. K. Gilbert, "Causal source codes," *IEEE Trans. Inform. Theory*, vol. IT-28, pp. 701–713, Sep. 1982.

[12] J. Ziv and A. Lempel, "Compression of individual sequences via variable-rate coding," *IEEE Trans. Inform. Theory*, vol. IT-24, pp. 530–536, Sep. 1978.

[13] J. Ziv, "Distortion-rate theory for individual sequences," *IEEE Trans. Inform. Theory*, vol. IT-26, pp. 137–143, Mar. 1980.

[14] J. Ziv, "On universal quantization," *IEEE Trans. Inform. Theory*, vol. IT-31, pp. 344–347, May 1985.

[15] R. Zamir and M. Feder, "Rate distortion performance in coding band-limited sources by sampling and dithered quantization," *IEEE Trans. Inform. Theory*, vol. IT-42, pp. 141–154, Jan. 1995.

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

[17] R. M. Gray, *Entropy and Information Theory.* New York: Springer-Verlag, 1990.

[18] I. Csiszár, "Generalized entropy and quantization problems," in *Transactions of the Sixth Prague Conference on Information Theory, Statistical Decision Functions, Random Processes*, (Prague), pp. 29–35, Akademia, 1973.

[19] P. A. Chou, T. Lookabaugh, and R. M. Gray, "Entropy-constrained vector quantization," *IEEE Trans. Acoust. Speech, Signal Processing*, vol. ASSP-37, pp. 31–42, Jan. 1989.

[20] R. M. Gray and D. L. Neuhoff, "Quantization," *IEEE Trans. Inform. Theory,* (Special Commemorative Issue), vol. IT-44, pp. 2325–2383, Oct. 1998.

[21] Y. N. Linkov, "Evaluation of epsilon entropy of random variables for small epsilon," *Problems of Information Transmission*, vol. 1, pp. 12–18, 1965. Translated from Problemy Peredachi Informatsii, Vol. 1, 18-28.

[22] T. Linder and R. Zamir, "On the asymptotic tightness of the Shannon lower bound," *IEEE Trans. Inform. Theory*, vol. IT-40, pp. 2026–2031, Nov. 1994.

[23] J. H. Conway and N. J. A. Sloane, *Sphere Packings, Lattices and Groups*. New York: Springer-Verlag, 1988.

[24] R. Zamir and M. Feder, "On lattice quantization noise," *IEEE Trans. Inform. Theory*, vol. IT-42, pp. 1152–1159, July 1996.

[25] T. Linder and K. Zeger, "Asymptotic entropy constrained performance of tessellating and universal randomized lattice quantization," *IEEE Trans. Inform. Theory*, vol. 40, pp. 575–579, Mar. 1994.

[26] P. Zador, "Topics in the asymptotic quantization of continuous random variables." Technical Memorandum, Bell Laboratories, Murray Hill, NJ, Feb. 1966.

[27] P. Zador, "Asymptotic quantization error of continuous signals and the quantization dimension," *IEEE Trans. Inform. Theory*, vol. IT-28, pp. 139–149, Mar. 1982.

[28] P. Billingsley, *Convergence of Probability Measures*. New York: Wiley, 1968.

[29] L. Schuchman, "Dither signals and their effect on quantization noise," *IEEE Trans. on Communication Technology*, vol. 12, pp. 162–165, Dec. 1964.

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

[31] R. M. Gray, T. Linder, and J. Li, "A Lagrangian formulation of Zador's entropy-constrained quantization theorem," *IEEE Trans. Inform. Theory*, vol. IT-48, pp. 695–707, Mar. 2002.

[32] R. M. Dudley, *Real Analysis and Probability*. New York: Chapman & Hall, 1989.

[33] I. Csiszár, "On an extremum problem of information theory," *Studia Scientiarum Mathematicarum Hungarica*, pp. 57–70, 1974.