A Zero-Delay Sequential Scheme for Lossy Coding of Individual Sequences *

Tamás Linder Gábor Lugosi

February 26, 2000

Abstract

We consider adaptive sequential lossy coding of bounded individual sequences when the performance is measured by the sequentially accumulated mean squared distortion. The encoder and the decoder are connected via a noiseless channel of capacity R and both are assumed to have zero delay. No probabilistic assumptions are made on how the sequence to be encoded is generated. For any bounded sequence of length n, the distortion redundancy is defined as the normalized cumulative distortion of the sequential scheme minus the normalized cumulative distortion of the best scalar quantizer of rate R which is matched to this particular sequence. We demonstrate the existence of a zero-delay sequential scheme which uses common randomization in the encoder and the decoder such that the normalized maximum distortion redundancy converges to zero at a rate $n^{-1/5}\log n$ as the length of the encoded sequence n increases without bound.

Index Terms: Lossy source coding, scalar quantization, sequential prediction, individual sequences.

*T. Linder is with the Department of Mathematics and Statistics, Queen's University, Kingston, Ontario, Canada K7L 3N6 (email: `linder@mast.queensu.ca`). G. Lugosi is with the Department of Economics, Pompeu Fabra University, Ramon Trias Fargas 25-27, 08005 Barcelona, Spain (email: `lugosi@upf.es`).
This research was supported in part by the Natural Sciences and Engineering Research Council (NSERC) of Canada and DGES grant PB96-0300.

1 Introduction

In a widely used model of lossy source coding, an infinite sequence of real-valued source symbols $x_1, x_2, \ldots$ is transformed into a sequence of channel symbols $y_1, y_2, \ldots$ (assumed to take values from a finite alphabet) which are transmitted through a noiseless channel. The received sequence of channel symbols are then used to produce the reproduction sequence $\widehat{x}_1, \widehat{x}_2, \ldots$. Such a system is called *causal* if the reproduction of the current source symbol depends on the present and past source symbols, but not on the future ones. In general, very little is known about the optimum performance theoretically attainable (OPTA) for causal coding of probabilistic sources. For the special case of a stationary and memoryless source Neuhoff and Gilbert [1] showed that the OPTA function of causal codes is achieved by time-sharing of entropy coded scalar quantizers.

A requirement more restrictive than causality is that of zero delay. A lossy coding scheme is said to have zero delay if each channel symbol y_n depends only on the past and present source symbols $x_1, \ldots, x_n$ and the reproduction $\widehat{x}_n$ for the present source symbol x_n depends only on the channel symbols $y_1, \ldots, y_n$ received so far. Zero-delay schemes have an obvious advantage over other coding methods (such as block codes) in applications where decoding delay is a crucial factor. For memoryless sources it has been shown by Ericson [2] and Gaarder and Slepian [3] (see also [4]) that the zero-delay OPTA function is achieved by the optimal (Lloyd-Max) scalar quantizer for the source.

In this work, we consider the problem of zero-delay source coding in a deterministic setting. Inspired by recent work on prediction of individual sequences, we study zero-delay sequential quantization of individual sequences. In this setting, it is not assumed that the source is generated by an underlying probabilistic mechanism. In general, the goal is to construct a *single* zero-delay scheme whose cumulative distortion on every bounded sequence is very close to that of the best scheme for the given sequence within a family of fixed zero-delay schemes. The probabilistic analogue of this problem is the problem of zero-delay universal coding with respect to a given class of sources.

Although lossless sequential source coding has been extensively studied (see Merhav and Feder [5] for am extensive survey), there seem to be no results available concerning its lossy counterpart we consider here. One main difficulty with the lossy case is that, unlike in the lossless case, the decoder does not have access to the past source outputs. Therefore the well-developed arsenal of universal lossless coding and sequential prediction cannot be directly applied.

In this paper we investigate the possibility of zero-delay lossy coding of individual sequences. Our main result in Section 2 describes a zero-delay sequential adaptive coding scheme which, asymptotically, achieves a cumulative mean squared distortion achieved by the best scalar quantizer of a given rate matched to the actual bounded source sequence. In other words, the proposed method has to compete sequentially with an "anticipating" scheme that sees the entire sequence in advance and chooses the best scalar quantizer for this sequence. The construction builds on techniques developed in the theory of prediction of individual sequences, namely, it uses an appropriately modified version of the exponential weighting method of Vovk [6]. The proposed method requires common randomization in the encoder and the decoder. Some aspects of common randomization are discussed in Section 3.

Admittedly, the special class of reference methods (i.e., the family of all fixed-rate scalar quantizers) limits the scope of this result, but it is still of interest, especially in view of the previously cited results of Ericson [2] and Gaarder and Slepian [3]. To our knowledge this is the first result concerning zero-delay sequential lossy coding of individual sequences.

2 Problem Formulation and Results

A (randomized) zero-delay sequential source code of rate $R = \log M$ (where M is a positive integer and log denotes base 2 logarithm) is described by an encoder-decoder pair which are connected via a noiseless channel of capacity R. It is assumed that both the encoder and the decoder have access to a common sequence of random variables $\{U_i\}_{i=1}^{\infty}$, where each U_i is uniformly distributed on the interval $[0,1]$. (Note that the U_i need not be independent.) The input to the encoder is a sequence of real numbers $x_1, x_2, \ldots$ assumed to be bounded such that $x_i \in [0,1]$ for all $i \geq 1$. (One could more generally assume that each x_i is in a fixed interval of length B, but since squared error distortion will be considered, the choice $x_i \in [0,1]$ does not limit generality.) At each time instant $i = 1, 2, \ldots$, the encoder observes x_i and the random number U_i. Based on x_i, U_i, and the past input values $x^{i-1} = (x_1, \ldots, x_{i-1})$, the encoder produces a channel symbol $y_i \in \{1, 2, \ldots, M\}$ which is then transmitted to the decoder. After receiving y_i, the decoder outputs the reconstruction value $\widehat{x}_i$ based on U_i and the channel symbols $y^n = (y_1, \ldots, y_n)$ received so far.

More formally, the code is given by a sequence of encoder-decoder functions $\{f_i, g_i\}_{i=1}^{\infty}$, where

$$f_i : [0,1]^i \times [0,1] \to \{1, 2, \ldots, M\}$$

and

$$g_i : \{1, 2, \ldots, M\}^i \times [0,1] \to [0,1].$$

so that $y_i = f_i(x^i, U_i)$ and $\widehat{x}_i = g_i(y^i, U_i)$, $i = 1, 2, \ldots$. Note that there is no delay in the encoding and decoding process. The *normalized cumulative squared distortion* of the sequential scheme at time instant n is given by

$$D_n(x^n) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \widehat{x}_i)^2$$

where the dependence of D_n on the randomizing sequence is suppressed in the notation. The expected cumulative distortion is

$$\overline{D}_n(x^n) = E\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \widehat{x}_i)^2\right]$$

where the expectation is taken with respect to the randomizing sequence $U^n = (U_1, \ldots, U_n)$.

An M-level scalar quantizer Q is measurable mapping $\mathbb{R} \to \mathcal{C}$, where the *codebook* $\mathcal{C}$ is a finite subset of $\mathbb{R}$ with cardinality $|\mathcal{C}| = M$. The elements of $\mathcal{C}$ are called the *codepoints*. The instantaneous squared distortion of Q for input x is $(x - Q(x))^2$. A quantizer Q is called a nearest neighbor quantizer if it satisfies

$$(Q(x) - x)^2 = \min_{y \in \mathcal{C}}(x - y)^2$$

for all x. It is well known that if Q is a nearest neighbor quantizer and $\widehat{Q}$ has the same codebook as Q, then $(Q(x) - x)^2 \leq (\widehat{Q}(x) - x)^2$ for all x. For this reason, we will only consider nearest-neighbor quantizers. Also, since we consider sequences with components in $[0, 1]$, we can assume without loss of generality that the domain of definition of Q is $[0, 1]$ and that all its codepoints are in $[0, 1]$.

Let $\mathcal{Q}$ denote the collection of all M-level nearest neighbor quantizers. For any sequence x^n, let $D_n^*(x^n)$ denote the minimum normalized cumulative distortion in quantizing x^n with an M-level scalar quantizer, that is, let

$$D_n^*(x^n) = \min_{Q \in \mathcal{Q}} \frac{1}{n}\sum_{i=1}^{n}(x_i - Q(x_i))^2.$$

Note that to find a $Q \in \mathcal{Q}$ achieving $D_n^*(x^n)$ one has to know the entire sequence x^n in advance. The next theorem asserts that there exists a zero-delay sequential source code of rate R which, for any bounded input sequence, performs asymptotically as well as the best scalar quantizer of rate R matched to the entire sequence.

Theorem 1 *For any $R = \log M$ there exists a randomized zero-delay sequential source code $\{f_i, g_i\}_{i=1}^{\infty}$ of rate R whose expected normalized cumulative distortion $\overline{D}_n(x^n)$ satisfies, for all $x^n \in [0,1]^n$,*

$$\overline{D}_n(x^n) - D_n^*(x^n) \leq Cn^{-1/5} \log n, \tag{1}$$

where C is a constant independent of n and x_1^n. In particular,

$$\limsup_{n\to\infty} \max_{x^n \in [0,1]^n} \left(\overline{D}_n(x^n) - D_n^*(x^n) \right) \leq 0.$$

To prove the theorem we first consider the case of sequential coding of sequences of a fixed finite length.

Proposition 1 *For any $\tilde{n} \geq 1$ and $R = \log M$ there exists a randomized zero-delay sequential source code $\{f_i^{(\tilde{n})}, g_i^{(\tilde{n})}\}_{i=1}^{\tilde{n}}$ of rate R for coding sequences of length $\tilde{n}$ such that for all $n \leq \tilde{n}$ and for all $x^n \in [0,1]^n$,*

$$n\overline{D}_n(x^n) \leq nD_n^*(x^n) + c\tilde{n}^{4/5} \log \tilde{n} \tag{2}$$

where c is a positive constant which does not depend on $\tilde{n}$.

Proposition 1 demonstrates the existence of a zero-delay scheme for sequentially coding sequences of length n which are asymptotically (for large n) efficient. To see this, let $n = \tilde{n}$ in Proposition 1. Then for any $n \geq 1$ there exists a sequential code for sequences of length n such that for any $x^n \in [0,1]^n$,

$$\overline{D}_n(x^n) \leq D_n^*(x^n) + cn^{-1/5} \log n \ . \tag{3}$$

Note that codes achieving (3) depend on the length n of the sequence to be encoded and therefore Proposition 1 does not directly imply the existence of a single sequential code $\{f_i, g_i\}_{i=1}^{\infty}$ capable of coding sequences of arbitrary length and achieving (1). The following proof exhibits a simple construction of a sequential code $\{f_i, g_i\}_{i=1}^{\infty}$ which satisfies Theorem 1 using the finite-length codes of Proposition 1. The proof is inspired by a similar trick in [7].

Proof of Theorem 1 For any $i = 1, 2, \dots,$ let l be a nonnegative integer such that $2^l \leq i < 2^{l+1}$. Now use the codes $\{f_i^{(\tilde{n})}, g_i^{(\tilde{n})}\}_{i=1}^{\tilde{n}}$ of Proposition 1 with $\tilde{n} = 2^l$ to define f_i and g_i by

$$y_i = f_i(x^i, U_i) = f_{i-2^l+1}^{(2^l)}(x_{2^l}^i, U_i)$$

and

$$\widehat{x}_i = g_i(y^i, U_i) = g^{(2^l)}_{i-2^l+1}(y^i_{2^l}, U_i).$$

That is, for $l = 0, 1, \ldots$, the segments $x_{2^l}^{2^{l+1}-1} = (x_{2^l}, \ldots, x_{2^{l+1}-1})$ are independently encoded using the code $\{f_i^{(2^l)}, g_i^{(2^l)}\}_{i=1}^{2^l}$. The resulting sequence $\{f_i, g_i\}_{i=1}^{\infty}$ is clearly a sequential code in the sense of Theorem 1.

To bound the cumulative distortion, for any $1 \le j \le k$, we let $\overline{D}(x_j^k) = E\left[\sum_{i=j}^k (x_i - \widehat{x}_i)^2\right]$ and also define $D^*(x_j^k) = \min_{Q\in\mathcal{Q}} \sum_{i=j}^k (x_i - Q(x_i))^2$. Note that for all $j < k' \le k$, one has $D^*(x_j^{k'-1}) + D^*(x_{k'}^k) \le D^*(x_j^k)$. Therefore Proposition 1 implies that for all n such that $2^m \le n < 2^{m+1}$,

$$\begin{aligned}
n\overline{D}_n(x^n) &= \sum_{l=0}^{m-1} \overline{D}(x_{2^l}^{2^{l+1}-1}) + \overline{D}(x_{2^m}^n) \\
&\le \sum_{l=0}^{m-1} \left(D^*(x_{2^l}^{2^{l+1}-1}) + c(2^l)^{4/5} \log 2^l \right) \\
&\quad + D^*(x_{2^m}^n) + c(2^m)^{4/5} \log 2^m \\
&\le D^*(x^n) + \sum_{l=0}^{m} c(2^l)^{4/5} \log 2^l \\
&\le D^*(x^n) + c \log 2^m \sum_{l=0}^{m} (2^l)^{4/5} \\
&= nD_n^*(x^n) + c \log 2^m \frac{2^{(m+1)\frac{4}{5}} - 1}{2^{\frac{4}{5}} - 1} \\
&\le nD_n^*(x^n) + Cn^{4/5} \log n
\end{aligned}$$

where $C = c2^{4/5}/(2^{4/5} - 1) \le 2.35c$. We conclude that for all $n \ge 1$ and $x^n \in [0,1]^n$, the normalized cumulative distortion of the sequential scheme is upper bounded as

$$\overline{D}_n(x^n) \le D_n^*(x^n) + Cn^{-1/5} \log n$$

which proves Theorem 1. □

Proof of Proposition 1 Let $\mathcal{Q}_N = \{Q_1, \ldots, Q_N\}$ be a fixed but arbitrary collection of M-level nearest neighbor scalar quantizers such that the codepoints of each Q_j are inside $[0,1]$. Fix $x^{\tilde{n}} \in [0,1]^{\tilde{n}}$, and for $n = 1, \ldots, \tilde{n}$, let

$$L_j^n = \sum_{i=1}^{n} (x_i - Q_j(x_i))^2$$

and set $L_j^0 = 0$. To simplify the description of the code, let us first construct a hypothetical coding scheme in which both the encoder and decoder have access to the values L_j^n, $j = 1, \ldots, N$, at each time instant n. The hypothetical scheme uses the well-known exponential weighting method of sequential prediction, (see, e.g., Vovk [6], Littlestone and Warmuth [8], and Cesa-Bianchi et al. [7]). Let $\eta > 0$ be fixed and for $n = 1, \ldots, \tilde{n}$, define the weights

$$\lambda_j^n = \frac{e^{-\eta L_j^{n-1}}}{\sum_{m=1}^N e^{-\eta L_m^{n-1}}}$$

(note that $\sum_{j=1}^N \lambda_j^n = 1$). At time n, the encoder uses the random number U_n and the weights λ_j^n to generate the random index $J_n \in \{1, \ldots, N\}$ with distribution

$$\Pr\{J_n = j\} = \lambda_j^n, \qquad j = 1, \ldots, N.$$

Then the encoder picks the quantizer Q_{J_n} to encode x_n and transmits the channel symbol representing the quantizer output $Q_{J_n}(x_n)$. After receiving this channel symbol, the decoder outputs $Q_{J_n}(x_n)$ (note that since the decoder has access to U_n and the λ_j^n, it can also generate J_n).

The expected normalized cumulative distortion of the hypothetical scheme (denoted by $\overline{d}_n(x^n)$) is given by

$$\begin{aligned} \overline{d}_n(x^n) = \frac{1}{n}\sum_{i=1}^n E(x_i - Q_{J_i}(x_i))^2 &= \frac{1}{n}\sum_{i=1}^n \sum_{j=1}^N (x_i - Q_{J_i}(x_i))^2 \Pr\{J_i = j\} \\ &= \frac{1}{n}\sum_{i=1}^n \sum_{j=1}^N \lambda_j^i (x_i - Q_j(x_i))^2. \end{aligned} \tag{4}$$

In the Appendix we show that for all $n \geq 1$,

$$n\overline{d}_n(x^n) \leq \min_{1\leq j\leq N} L_j^n + \frac{\ln N}{\eta} + \frac{n\eta}{8}. \tag{5}$$

Moreover, a simple argument presented in Lemma 2 in the Appendix shows that for each $N \geq 2$, there exists a collection of M-level nearest-neighbor quantizers $\mathcal{Q}_N = \{Q_1, \ldots, Q_N\}$, supported in $[0, 1]$, such that for all $x^n \in [0, 1]^n$,

$$\min_{Q\in\mathcal{Q}_N} \sum_{i=1}^n (x_i - Q(x_i))^2 \leq \min_{Q\in\mathcal{Q}} \sum_{i=1}^n (x_i - Q(x_i))^2 + n\frac{1}{N^{1/M} - 1}. \tag{6}$$

Using this $\mathcal{Q}_N$ in the definition of the hypothetical scheme, we can rewrite (5) as

$$n\overline{d}_n(x^n) \leq nD_n^*(x^n) + n\frac{1}{N^{1/M} - 1} + \frac{\ln N}{\eta} + \frac{n\eta}{8}. \tag{7}$$

Construction of sequential scheme

The bound (7) implies that η and N can be chosen (as functions of $\tilde{n}$) such that the cumulative distortion of the hypothetical scheme satisfies

$$n\overline{d}_n(x^n) \le nD_n^*(x^n) + O(\tilde{n}^{1/2}\log\tilde{n}).$$

To achieve this, however, the hypothetical scheme has to transmit the values of L_j^i, $j = 1,\ldots,N$ at all time instants i, which requires an additional channel of infinite capacity between the encoder and the decoder. The basic idea for constructing a sequential scheme of true rate $\log M$ is to periodically transmit approximate (quantized) versions of the cumulative quantizer losses L_j^i and to use these approximations to form the approximate weights $\widehat{\lambda}_j^i$ at the decoder. We show that using only a small fraction of the overall available rate to transmit the quantized cumulative distortions, the $\widehat{\lambda}_j^i$ will sufficiently well approximate the λ_j^i so that the difference between the distortion $\overline{D}_n(x^n)$ of the resulting sequential scheme and the distortion $\overline{d}_n(x^n)$ of the hypothetical scheme becomes negligible for large n.

To describe the scheme, let $\{n_l;\ l = 1,2,\ldots\}$ be a strictly increasing sequence of positive integers such that $n_1 = 1$. Let $K \ge 1$ be a fixed integer and let q_K denote the K-level uniform quantizer over $[0,1]$. Introduce [1]

$$b_l = \lceil K \log_M n_l \rceil .$$

For any $l \ge 1$, if $b_l < n_{l+1} - n_l$, then in the time interval $i = n_l,\ldots,n_{l+1}-1$, the encoder transmits losslessly the values of $\tilde{x}_r = q_K(x_r)$, $r = 1,\ldots,n_l - 1$, using the first b_l of the available $n_{l+1} - n_l$ channel uses. (Note that this is possible since the number of different ways of partitioning $n_l - 1$ points into K cells is not greater than n_l^K.) In these time instants (for $i = n_l,\ldots,n_l + b_l - 1$) the decoder's output is set to a constant value (say $\widehat{x}_i = 1/2$). In the remaining time instants $i = n_l + b_{l-1},\ldots,n_{l+1}-1$, the encoder forms the approximate weights

$$\widehat{\lambda}_j^i = \frac{e^{-\eta \widehat{L}_j^{i-1}}}{\sum_{m=1}^N e^{-\eta \widehat{L}_m^{i-1}}}, \qquad j = 1,\ldots,N \tag{8}$$

where

$$\widehat{L}_j^i = \sum_{r=1}^{n_l - 1} \left(\tilde{x}_r - Q_j(\tilde{x}_r)\right)^2, \quad j = 1,\ldots,N, \quad i = n_l,\ldots,n_{l+1}-1.$$

Using U_i, the encoder then generates the random index $\widehat{J}_i$ with distribution $\Pr\{\widehat{J}_i = j\} = \widehat{\lambda}_j^i$, picks $Q_{\widehat{J}_i}$ to quantize x_i, and transmits the channel symbol for the quantizer output

[1] $\lceil t \rceil$ denotes the smallest integer not less than t, and $\lfloor t \rfloor$ denotes the largest integer not greater than t.

$Q_{\widehat{J}_i}(x_i)$. Observe that at the same time instants, the decoder has already access to all $\widehat{x}_r$, $r = 1, \ldots, n_l - 1$, and thus it can simultaneously calculate $\widehat{\lambda}_j^i$ for $j = 1, \ldots, N$, and $i = n_l, \ldots, n_{l+1} - 1$. Using U_i and the received channel symbol, the decoder can output $\widehat{x}_i = Q_{\widehat{J}_i}(x_i)$.

If $n_{l+1} - n_l \le b_l$ for some l, then the encoder is defined to be in an idle state in the time segment $i = n_l, \ldots, n_{l+1} - 1$ and the decoder outputs some preset constant value (say $\widehat{x}_i = 1/2$).

Analysis of distortion

For $i = n_l, \ldots, n_{l+1} - 1$ such that $n_{l+1} - n_l \le b_l$, we have $(x_i - \widehat{x}_i)^2 \le 1$. If $n_{l+1} - n_l > b_l$, then $(x_i - \widehat{x}_i)^2 \le 1$ for $i = n_l, \ldots, n_l + b_l - 1$. On the other hand, for $i = n_l + b_l, \ldots, n_{l+1} - 1$, we have

$$E(x_i - \widehat{x}_i)^2 = \sum_{j=1}^{N} (x_i - Q_j(x_i))^2 \Pr\{\widehat{J}_i = j\} = \sum_{j=1}^{N} \widehat{\lambda}_j^i (x_i - Q_j(x_i))^2.$$

Fix $n \le \tilde{n}$ and consider the cumulative distortion at time n. Extending the definition (8) of the approximate weights to all $i = n_l, \ldots, n_{l+1} - 1$, we can now upper bound the expected cumulative distortion as

$$\begin{aligned} n\overline{D}_n(x^n) &= \sum_{i=1}^{n} E(x_i - \widehat{x}_i)^2 \\ &\le \sum_{i=1}^{n} \sum_{j=1}^{N} \widehat{\lambda}_j^i (x_i - Q_j(x_i))^2 + \sum_{l=1}^{S_n} b_l \end{aligned} \tag{9}$$

where $S_n = \max\{l : n_l \le n\}$. First we give an upper bound for the first term of the right side of (9) in terms of the cumulative distortion of the hypothetical scheme. Since $(Q_j(x) - x)^2 \le 1$ for j and $x \in [0, 1]$, equation (4) implies

$$\sum_{i=1}^{n} \sum_{j=1}^{N} \widehat{\lambda}_j^i (x_i - Q_j(x_i))^2 - n\overline{d}_n(x^n) \le \sum_{i=1}^{n} \sum_{j=1}^{N} |\widehat{\lambda}_j^i - \lambda_j^i| \tag{10}$$

Since $\widehat{\lambda}_j^i = \frac{e^{-\eta \widehat{L}_j^{i-1}}}{\sum_{m=1}^{N} e^{-\eta \widehat{L}_m^{i-1}}}$ and $\lambda_j^i = \frac{e^{-\eta L_j^{i-1}}}{\sum_{m=1}^{N} e^{-\eta L_m^{i-1}}}$, Lemma 3 in the Appendix implies that

$$\sum_{j=1}^{N} |\widehat{\lambda}_j^i - \lambda_j^i| \le 2\eta \max_{1 \le j \le N} \left|\widehat{L}_j^{i-1} - L_j^{i-1}\right|.$$

Now it is easy to see that for any nearest-neighbor quantizer Q supported in $[0, 1]$ and any $x, \tilde{x} \in [0, 1]$, we have

$$\left|(x - Q(x))^2 - (\tilde{x} - Q(\tilde{x}))^2\right| \le 2|x - \tilde{x}|.$$

This implies that for all $i = n_l, \dots, n_{l+1} - 1$ and all $j = 1, \dots, N$,

$$\begin{aligned}
\left|\widehat{L}_j^i - L_j^i\right| &= \left|\sum_{r=1}^{n_l - 1} (\tilde{x}_r - Q_j(\tilde{x}_r))^2 - \sum_{r=1}^{i} (x_r - Q_j(x_r))^2\right| \\
&\le 2\sum_{r=1}^{n_l-1} |\tilde{x}_r - x_r| + \sum_{r=n_l}^{i} (x_r - Q_j(x_r))^2 \\
&\le \frac{(n_l - 1)}{K} + (n_{l+1} - n_l)
\end{aligned}$$

where the second inequality follows since $|x_r - \tilde{x}_r| = |x_r - q_K(x_r)| \le 1/(2K)$. Summarizing these bounds, we obtain

$$n\overline{D}_n(x^n) - n\overline{d}_n(x^n) \le \sum_{l=1}^{S_n} b_l + 2\eta \sum_{l=1}^{S_n} (n_{l+1} - n_l) \left(\frac{n_l - 1}{K} + (n_{l+1} - n_l)\right). \qquad (11)$$

Combining this with the bound (7) on the cumulative distortion of the hypothetical scheme, for all $n \le \tilde{n}$ we obtain

$$\begin{aligned}
n\overline{D}_n(x^n) - nD_n^*(x^n) &= \left(n\overline{D}_n(x^n) - n\overline{d}_n(x^n)\right) + \left(n\overline{d}_n(x^n) - nD_n^*(x^n)\right) \\
&\le n\frac{1}{N^{1/M} - 1} + \frac{\ln N}{\eta} + \frac{n\eta}{8} \\
&\quad + \sum_{l=1}^{S_n} b_l + 2\eta \sum_{l=1}^{S_n} (n_{l+1} - n_l) \left(\frac{n_l - 1}{K} + (n_{l+1} - n_l)\right).
\end{aligned}$$

It only remains to choose the parameters η, N, K, and the sequence $\{n_l\}$ appropriately. We do this by setting $n_l = \lfloor l^a \rfloor$ for some $a > 1$ which allows us to approximately optimize the upper bound by an appropriate choice of the constant a. In this case we have $S_n \approx n^{1/a}$ and (ignoring the constants) the upper bound has the form

$$\frac{n}{N^{1/M}} + \frac{\ln N}{\eta} + n\eta + Kn^{1/a}\log n + \frac{\eta n^2}{K} + \eta n^{2-1/a}.$$

Straightforward calculation reveals that ignoring logarithmic and constant factors, the choice that approximately minimizes this upper bound is $a = 5/2$, $\eta \approx \tilde{n}^{-4/5}$, $K \approx \tilde{n}^{2/5}$. The number of reference quantizers N must be such that $N^{1/M}$ is bounded by a polynomial of $\tilde{n}$ and its order is at least $\tilde{n}^{1/5}$. Computationally it may be advantageous to choose N as small as possible (i.e., letting $N \approx n^{M/5}$). Resubstiting these values into the upper bound above gives the desired result. $\square$

3 Remarks on Common Randomization

The proposed quantization scheme has an obvious weakness: it requires that both the encoder and the decoder dispose of the same sequence of uniform random variables $U_1, U_2, \ldots$. This assumption is not uncommon in universal quantization of probabilistic sources (see the works of Ziv [9] and Zamir and Feder [10]) where the U_i sequence represents "subtractive dither". In practice, these may be replaced by a pseudorandom sequence generated at both the sender and receiver side.

Observe that the only requirement for $U_1, U_2, \ldots$ is that their distribution should be uniform. No assumption on the joint distribution of these variables is necessary for Theorem 1. In an extreme case, as in [9], one may even take $U_1 = U_2 = \cdots$, that is, use the *same* variable at each time instance. This has no effect on the expected behavior of the distortion. On the other hand, using the same randomizing variable at all time instants hides a danger of instability, as the true (random) distortion $D_n(x^n)$ may be far from its expected value $\overline{D}_n(x^n) = ED_n(x^n)$. The next fact shows that one may avoid instability by using an independent randomizing sequence.

Lemma 1 *If $U_1, U_2, \ldots$ are independent and uniformly distributed over $[0,1]$, then the distortion $D_n(x^n)$ of the quantization scheme of Theorem 1 satisfies, for all $t > 0$,*

$$\Pr\left\{\left|D_n(x^n) - \overline{D}_n(x^n)\right| > t\right\} \leq 2e^{-2nt^2}$$

In particular, by the Borel-Cantelli lemma, combining Lemma 1 with Theorem 1 yields:

Corollary 1 *Assume that a sequence of independent uniform random variables is available at both the encoder and the decoder. Then there exists a randomized zero-delay sequential source code $\{f_i, g_i\}_{i=1}^{\infty}$ of rate R whose normalized cumulative distortion $D_n(x^n)$ satisfies, for all $\{x_i\}_{i=1}^{\infty}$ such that $x_i \in [0,1]$ for all i,*

$$\limsup_{n\to\infty} \frac{D_n(x^n) - D_n^*(x^n)}{n^{-1/5}\log n} \leq C, \qquad \textit{almost surely} \tag{12}$$

where C is the same constant as in Theorem 1.

Proof of Lemma 1 Recall from the proof of Proposition 1 that for all n,

$$\left|D_n(x^n) - \overline{D}_n(x^n)\right| \leq \left|\frac{1}{n}\sum_{i=1}^{n}\left(\left(x_i - Q_{\widehat{J}_i}(x_i)\right)^2 - E\left(x_i - Q_{\widehat{J}_i}(x_i)\right)^2\right)\right|$$

where the random variable $\widehat{J}_i$ is a function of U_i and the approximate weights $\widehat{\lambda}^i_j$, $j = 1, \ldots, N$. Since the approximate weights are deterministic (i.e., their values do not depend on the sequence $U_1, U_2, \ldots$, see equation (8)), the expression on the right-hand side is an average of n independent random variables. Now recall Hoeffding's inequality [11] which states that if $S_n = \sum_{i=1}^n X_i$, where $X_1, \ldots, X_n$ are independent random variables such that $X_i \in [a,b]$ with probability one, then $\Pr\{n^{-1}|S_n - E(S_n)| > t\} \le 2e^{-2nt^2/(b-a)^2}$ for all $t > 0$. Letting $X_i = \left(x_i - Q_{\widehat{J}_i}(x_i)\right)^2$ and $[a,b] = [0,1]$ yields the claim of the lemma. □

Appendix

Proof of (5): distortion of hypothetical scheme

Using a standard technique (see, e.g., Cesa-Bianchi [12]) to upper bound $\overline{d}_n(x^n)$, we define $W_n = \sum_{j=1}^N e^{-\eta L_j^{n-1}}$. Then, since $W_1 = N$, on the one hand we have

$$\begin{aligned}
\ln \frac{W_{n+1}}{W_1} &= \ln \sum_{j=1}^N e^{-\eta L_j^n} - \ln N \\
&\ge \ln \left(\max_{1 \le j \le N} e^{-\eta L_j^n} \right) - \ln N \\
&= -\eta \min_{1 \le j \le N} L_j^n - \ln N. \qquad (13)
\end{aligned}$$

On the other hand, recalling that by Hoeffding's inequality [11], for any random variable $X \in [0,1]$ and $s \in \mathbb{R}$, $E\left(e^{sX}\right) \le e^{sE(X)+s^2/8}$,

$$\begin{aligned}
\ln \frac{W_{n+1}}{W_1} &= \sum_{i=1}^n \ln \frac{W_{i+1}}{W_i} \\
&= \sum_{i=1}^n \ln \frac{\sum_{j=1}^N e^{-\eta L_j^i}}{\sum_{j=1}^N e^{-\eta L_j^{i-1}}} \\
&= \sum_{i=1}^n \ln \frac{\sum_{j=1}^N e^{-\eta (x_i - Q_j(x_i))^2} e^{-\eta L_j^{i-1}}}{\sum_{j=1}^N e^{-\eta L_j^{i-1}}} \\
&= \sum_{i=1}^n \ln \left(\sum_{j=1}^N \lambda_j^i e^{-\eta (x_i - Q_j(x_i))^2} \right) \\
&\le \sum_{i=1}^n \ln \left(e^{-\eta \sum_{j=1}^N \lambda_j^i (x_i - Q_j(x_i))^2 + \eta^2/8} \right) \\
&\quad \text{(by Hoeffding's inequality)}
\end{aligned}$$

$$
\begin{aligned}
&= -\eta \sum_{i=1}^{n} \sum_{j=1}^{N} \lambda_j^i (x_i - Q_j(x_i))^2 + \frac{n\eta^2}{8} \\
&= -\eta\, n\overline{d}_n(x^n) + \frac{n\eta^2}{8}.
\end{aligned}
$$

Combining the preceding bound with (13) yields

$$
n\overline{d}_n(x^n) \le \min_{1\le j\le N} L_j^n + \frac{\ln N}{\eta} + \frac{n\eta}{8}.
$$

Lemma 2 *Let $\mathcal{Q}$ denote the family of all M-level nearest neighbor scalar quantizers whose codepoints are all inside $[0,1]$. Then for all $N \ge 2$ there exists a collection of N scalar quantizers $\mathcal{Q}_N = \{Q_1, \ldots, Q_N\} \subset \mathcal{Q}$ such that for all $x^n \in [0,1]$,*

$$
\min_{Q\in\mathcal{Q}_N} \sum_{i=1}^{n} (x_i - Q(x_i))^2 \le \min_{Q\in\mathcal{Q}} \sum_{i=1}^{n} (x_i - Q(x_i))^2 + n\frac{1}{N^{1/M} - 1}.
$$

Proof. The statement of the lemma will follow if we can construct a $\mathcal{Q}' \subset \mathcal{Q}$ with cardinality $|\mathcal{Q}'| \le N$ such that for any $Q \in \mathcal{Q}$ there is a $Q' \in \mathcal{Q}'$ satisfying

$$
\max_{x\in[0,1]} \left|(x - Q(x))^2 - (x - Q'(x))^2\right| \le \frac{1}{N^{1/M} - 1}.
$$

Toward this end, let $k = \lfloor N^{1/M} \rfloor$ and define $\mathcal{Q}^{(k)}$ as the family of nearest neighbor quantizers with M or less codepoints which all belong to the set

$$
\mathcal{C}^{(k)} = \{1/(2k), 3/(2k), \ldots, (2k-1)/(2k)\}.
$$

Since for any $y \in [0,1]$ there is a $y' \in \mathcal{C}^{(k)}$ with $|y - y'| \le 1/(2k)$, it is easy to see that for any M-level nearest neighbor quantizer Q with codepoints inside $[0,1]$ there is a $Q' \in \mathcal{Q}^{(k)}$ with

$$
\max_{x\in[0,1]} \left|(x - Q(x))^2 - (x - Q'(x))^2\right| \le \frac{1}{k}.
$$

Since $|\mathcal{Q}^{(k)}| \le k^M \le N$, the lemma follows. □

Lemma 3 *Let $\eta > 0$, and for any $v = (v_1, \ldots, v_N) \in \mathbb{R}^N$ and $j = 1, \ldots, N$, define*

$$
\lambda_j(v) = \frac{e^{-\eta v_j}}{\sum_{k=1}^{N} e^{-\eta v_k}}.
$$

Then for any $v, \widehat{v} \in \mathbb{R}^N$,

$$
\sum_{j=1}^{N} |\lambda_j(\widehat{v}) - \lambda_j(v)| \le 2\eta \max_{1\le j\le N} |\widehat{v}_j - v_j|.
$$

Proof. For $\alpha \in [0,1]$, let $h_j(\alpha) = \lambda_j(v + \alpha(\widehat{v} - v))$ and let $h'_j(\alpha)$ denote the derivative of $h_j(\alpha)$. Then by the mean value theorem of differentiation, for some $\tilde{\alpha} \in (0,1)$ we have $\lambda_j(\widehat{v}) - \lambda(v) = h_j(1) - h_j(0) = h'_j(\tilde{\alpha})$. Now,

$$h'_j(\tilde{\alpha}) = \sum_{i=1}^{N} \frac{\partial \lambda_j}{\partial v_i}(v + \tilde{\alpha}(\widehat{v} - v))(\widehat{v}_i - v_i)$$

where

$$\frac{\partial \lambda_j}{\partial v_i} = \begin{cases} -\eta\lambda_j(1-\lambda_j), & i = j \\ -\eta\lambda_j\lambda_i, & i \neq j. \end{cases}$$

Therefore, by letting $\tilde{\lambda}_j = \lambda_j(v + \tilde{\alpha}(\widehat{v} - v))$ and using the fact that $\sum_{j=1}^{N} \tilde{\lambda}_j = 1$, we obtain

$$\begin{aligned} |\lambda_j(\widehat{v}) - \lambda_j(v)| &= \eta\tilde{\lambda}_j \left| (1-\tilde{\lambda}_j)(\widehat{v}_j - v_j) + \sum_{i \neq j} \tilde{\lambda}_i(\widehat{v}_i - v_i) \right| \\ &\leq \eta\tilde{\lambda}_j 2(1-\tilde{\lambda}_j) \max_{1 \leq j \leq N} |\widehat{v}_j - v_j| \\ &\leq 2\eta\tilde{\lambda}_j \max_{1 \leq j \leq N} |\widehat{v}_j - v_j| \end{aligned}$$

which implies the lemma. □

References

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

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

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

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

[5] N. Merhav and M. Feder, "Quantization," *IEEE Trans. Inform. Theory,* (Special Commemorative Issue), vol. IT-44, pp. 2124–2147, Oct. 1998.

[6] V. Vovk, "Aggregating strategies," in *Proceedings of the 3rd Annual Workshop on Computational Learning Theory*, pp. 372–383, 1990.

[7] N. Cesa-Bianchi, Y. Freund, D. P. Helmbold, D. Haussler, R. Schapire, and M. K. Warmuth, “How to use expert advice,” *Journal of the ACM*, vol. 44, no. 3, pp. 427–485, 1997.

[8] N. Littlestone and M. K. Warmuth, “The weighted majority algorithm,” *Information and Computation*, vol. 108, pp. 212–261, 1994.

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

[10] R. Zamir and M. Feder, “On universal quantization by randomized uniform/lattice quantizers,” *IEEE Trans. Inform. Theory*, vol. IT-38, pp. 428–436, Mar. 1992.

[11] W. Hoeffding, “Probability inequalities for sums of bounded random variables,” *Journal of the American Statistical Association*, vol. 58, pp. 13–30, 1963.

[12] N. Cesa-Bianchi, “Analysis of two gradient-based algorithms for on-line regression,” in *Proceedings of the 10th Annual Conference on Computational Learning Theory*, pp. 163–170, ACM Press, 1997.