Efficient Adaptive Algorithms and Minimax Bounds for Zero-Delay Lossy Source Coding

András György Tamás Linder Gábor Lugosi

IEEE Trans. on Signal Processing
Submitted: July 23, 2003
Revised: December 22, 2003

Abstract

Zero-delay lossy source coding schemes are considered for both individual sequences and random sources. Performance is measured by the distortion redundancy, defined as the difference between the normalized cumulative mean squared distortion of the scheme and the normalized cumulative distortion of the best scalar quantizer of the same rate which is matched to the entire sequence to be encoded. By improving and generalizing a scheme of Linder and Lugosi, Weissman and Merhav showed the existence of a randomized scheme which, for any bounded individual sequence of length n, achieves a distortion redundancy $O(n^{-1/3} \log n)$. However, both schemes have prohibitive complexity (both space and time) which makes practical implementation infeasible. In this paper, we present an algorithm that computes Weissman and Merhav's scheme efficiently. In particular, we introduce an algorithm with encoding complexity $O(n^{4/3})$ and distortion redundancy $O(n^{-1/3} \log n)$. The complexity can be made linear in the sequence length n at the price of increasing the distortion redundancy to $O(n^{-1/4}\sqrt{\log n})$. We also consider the problem of minimax distortion redundancy in zero-delay lossy coding of random sources. By introducing a simplistic scheme and proving a lower bound, we show that for the class of bounded memoryless sources, the minimax expected distortion redundancy is upper and lower bounded by (constant multiples of) $n^{-1/2}$.

Index Terms: Algorithmic efficiency, individual sequences, lossy source coding, minimax redundancy, sequential coding, scalar quantization.

A. György and T. Linder are with the Department of Mathematics and Statistics, Queen's University, Kingston, Ontario, Canada K7L 3N6 (email: `{gyorgy}{linder}@mast.queensu.ca`). A. György is on leave from the Computer and Automation Research Institute of the Hungarian Academy of Sciences, Lágymányosi u. 11, Budapest, Hungary, H-1111. 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 by DGES grant PB96-0300 of the Spanish Ministry of Science and Technology.

1 Introduction

Consider the widely used model for fixed-rate lossy source coding at rate R where an infinite sequence of real-valued source symbols $x_1, x_2, \ldots$ is transformed into a sequence of channel symbols $y_1, y_2, \ldots$ taking values from the finite channel alphabet $\{1, 2, \ldots, M\}$, $M = 2^R$, and these channel symbols are then used to produce the reproduction sequence $\hat{x}_1, \hat{x}_2, \ldots$. The scheme is said to have overall delay at most δ if there exist nonnegative integers d_1 and d_2 with $d_1 + d_2 \leq \delta$ such that each channel symbol y_n depends only on the source symbols $x_1, \ldots, x_{n+d_1}$ and the reproduction $\hat{x}_n$ for the source symbol x_n depends only on the channel symbols $y_1, \ldots, y_{n+d_2}$. When $\delta = 0$, the scheme is said to have zero delay. In this case, y_n depends only on $x_1, \ldots, x_n$, and $\hat{x}_n$ on $y_1, \ldots, y_n$, so that the encoder produces y_n as soon as x_n is available, and the decoder can produce $\hat{x}_n$ when y_n is received.

Lossy source coding schemes with limited delay (in particular with zero delay) are of obvious practical interest in all applications where small delay is a crucial requirement. In this paper we investigate the construction of provably efficient and computationally feasible methods for zero-delay lossy source coding. We mainly concentrate on methods that perform uniformly well with respect to a given reference coder class on every individual (deterministic) sequence. In this individual-sequence setting no probabilistic assumptions are made on the source sequence, which provides a natural model for situations where very little is known about the source to be encoded. We also investigate the best performance of zero-delay schemes for probabilistic sources and determine tight performance bounds for the class of memoryless sources.

The study of zero-delay coding for individual sequences was initiated in [1]. There a zero-delay scheme was constructed that, uniformly over all individual sequences, performs essentially as well as the best scalar quantizer that is matched to the particular sequence to be encoded. More precisely, it was shown that for any bounded sequence of n source symbols, the scheme's normalized accumulated mean squared distortion is not larger than the normalized cumulative distortion of the best scalar quantizer of the same rate plus an error term (called the distortion redundancy) of order $n^{-1/5} \log n$. The scheme was based on a generalization of exponentially weighted average prediction of individual sequences (see Vovk [2, 3], Littlestone and Warmuth [4]) and it required that both the encoder and the decoder have access to a common randomization sequence.

The results in [1] were improved and generalized by Weissman and Merhav [5]. They considered the construction of schemes that can compete with any finite set of limited-delay and finite-memory coding schemes without requiring that the decoder have access to the randomization sequence. In the special case dealt with in [1] where the reference class is the (zero-delay) family of scalar quantizers of a given rate, the resulting scheme has distortion redundancy of order $n^{-1/3}\log n$. Similarly to the method of [1], the basic idea is to assign a weight to each of a finite collection of quantizers approximating all possible quantizers of rate R such that the weight is an exponentially decreasing function of the accumulated distortion of the quantizer. Then a quantizer is chosen randomly with probabilities proportional to the assigned weights and used in transmitting symbols for a certain period.

Although both schemes have the attractive property of performing uniformly well on individual sequences, they are computationally inefficient in that the number of weights they have to maintain is polynomial in n with degree that is proportional to $M = 2^R$, where R is the rate of the scheme. In particular, in their straightforward implementation, they require a computational time of order $n^{c\,2^R}$, where $c = 1/5$ for the scheme in [1] and $c = 1/3$ for the scheme in [5]. This prohibitive complexity comes from the fact that in order to well approximate the performance of the best scalar quantizer by the performance of the best quantizer from a finite set of quantizers, these methods have to calculate and store the cumulative distortion of about $n^{c\,2^R}$ quantizers. Clearly, even for moderate values of the encoding rate, this complexity makes the implementation of both methods infeasible. It was identified as an important open problem in both [1] and [5] to find an algorithm with similar performance properties but significantly lower complexity.

The main result of this paper is an algorithm for implementing the scheme of Weissman and Merhav whose computational complexity is of order $2^R n^{4/3}$. The key idea is to use the special structure of scalar quantizers to efficiently generate randomly chosen quantizers according to the exponential weighting scheme without having to calculate and store the cumulative losses of all $n^{c\,2^R}$ reference quantizers. The complexity of the scheme can be reduced to be of order $2^R n$ (i.e., linear in the length of the sequence) by increasing the distortion redundancy to $O(n^{-1/4}\sqrt{\log n})$.

In the second part of the paper we investigate the distortion redundancy problem for zero-delay coding schemes in the probabilistic setting. In particular, we provide lower and upper bounds for stationary and memoryless random sources. These bounds are based on learning-theoretic analyses of the minimax distortion redundancy in the design of empirically optimal quantizers [6, 7]. We show that there exists a simple (not randomized) zero-delay scheme whose expected distortion redundancy is bounded by a constant times $n^{-1/2}$. In the other direction, we show an $n^{-1/2}$-type lower bound on the maximum distortion redundancy over the class of memoryless sources for any zero-delay scheme. This proves that for memoryless sources the minimax distortion redundancy of zero-delay lossy coding is essentially proportional to $n^{-1/2}$. Note that this is in contrast to the best known $O(n^{-1/3}\log n)$ convergence rate for zero-delay coding of individual sequences given by Weissman and Merhav's scheme. Whether this $O(n^{-1/3}\log n)$ rate can be improved for individual sequences remains an open problem.

The rest of the paper is organized as follows. In Section 2, after giving formal definitions, we construct an algorithm efficiently implementing the scheme of Weissman and Merhav, and analyze its performance and complexity. In Section 3 we show that the minimax distortion redundancy of zero-delay schemes for memoryless sources is at least of order $n^{-1/2}$, and we also describe and analyze a simplistic scheme which provides a matching $n^{-1/2}$-type upper bound. Conclusions are drawn in Section 4.

2 A fast algorithm for individual sequences

In this section, first we formally define the model of fixed-rate zero-delay sequential lossy source coding and describe the coding scheme of Weissman and Merhav. The main result of this section is an efficiently computable algorithm to implement their method.

A fixed-rate zero-delay sequential source code of rate $R = \log M$ (M is a positive integer and log denotes base-2 logarithm) is defined by an encoder-decoder pair connected via a discrete noiseless channel of capacity R. We assume that the encoder has access to a sequence $U_1, U_2, \ldots$ of independent random variables distributed uniformly over the interval $[0,1]$. The input to the encoder is a sequence of real numbers $x_1, x_2, \ldots$ taking values in the interval $[0,1]$. (All results may be extended trivially for arbitrary bounded sequences of input symbols.) 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, the past input values $x^{i-1} = (x_1, \ldots, x_{i-1})$, and the past values of the randomization sequence $U^{i-1} = \{U_1, \ldots, U_{i-1}\}$, 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 $\hat{x}_i$ based on the channel symbols $y^i = (y_1, \ldots, y_i)$ received so far.

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

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

and

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

so that $y_i = f_i(x^i, U^i)$ and $\hat{x}_i = g_i(y^i)$, $i = 1, 2, \ldots$. 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

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

The expected cumulative distortion is

$$\mathbb{E}\left[\frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{x}_i)^2\right]$$

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 a 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 *code points*. The instantaneous squared distortion of Q for input x is $(x - Q(x))^2$. A quantizer

Q is called a nearest neighbor quantizer if for all x it satisfies

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

It is immediate from the definition that if Q is a nearest neighbor quantizer and $\widehat{Q}$ has the same codebook as Q, then $(Q(x)-x)^2 \le (\widehat{Q}(x)-x)^2$ 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 code points are in $[0,1]$.

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

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

Note that to find a $Q \in \mathcal{Q}$ achieving this minimum one has to know the entire sequence x^n in advance.

The expected *distortion redundancy* of a scheme (with respect to the class of scalar quantizers) is the quantity

$$\sup_{x^n} \left(\mathbb{E}\left[\frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{x}_i)^2 \right] - \min_{Q \in \mathcal{Q}} \frac{1}{n} \sum_{i=1}^{n} (x_i - Q(x_i))^2 \right).$$

where the supremum is over all individual sequences of length n with components in $[0,1]$ (recall that the expectation is taken over the randomizing sequence). In [1] a zero-delay sequential scheme was constructed whose distortion redundancy converges to zero as n increases without bound. In other words, for any bounded input sequence the scheme performs asymptotically as well as the best scalar quantizer that is matched to the entire sequence. The main result of Weissman and Merhav [5], specialized to the zero-delay case, improves the construction in [1] and yields the best distortion redundancy known to date given by

$$\sup_{x^n}\left(\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(x_i-\hat{x}_i)^2\right]-\min_{Q\in\mathcal{Q}}\frac{1}{n}\sum_{i=1}^{n}(x_i-Q(x_i))^2\right)\le cn^{-1/3}\log n$$

where c is a constant depending only on M.

The coding scheme of [5] works as follows: the source sequence x^n is divided into non-overlapping blocks of length l (for simplicity assume that l divides n). At the end of the kth block, that is, at time instances $t = kl$, $k = 0, 1, \ldots, n/l - 1$, a quantizer Q_k is chosen randomly from the class $\mathcal{Q}_K$ of all M-level nearest-neighbor quantizers whose code points all belong to the finite grid

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

according to the probabilities

$$\mathbb{P}\{Q_k = Q\} = p_k(Q) = \frac{e^{-\eta D_{kl}(Q)}}{\sum_{\widehat{Q}\in\mathcal{Q}_K} e^{-\eta D_{kl}(\widehat{Q})}} \tag{1}$$

where $\eta > 0$ is a parameter to be specified later,

$$D_t(Q) = \frac{1}{t}\sum_{i=1}^{t}(x_i - Q(x_i))^2 \qquad \text{for all } t = 1, \ldots, n$$

and $D_0(Q) = 0$ for all $Q \in \mathcal{Q}_K$. At the beginning of the $(k+1)$st block the encoder uses the first $\lceil \frac{1}{R}\log\binom{K}{M}\rceil$ time instants to describe the selected quantizer Q_k to the receiver ($\lceil x\rceil$ denotes the smallest integer not less than x), that is, for time instants

$$i = kl + 1, \ldots, kl + \left\lceil \frac{1}{R}\log\binom{K}{M}\right\rceil$$

an index identifying Q_k is transmitted (note that $|\mathcal{Q}_K| = \binom{K}{M}$). In the rest of the block, that is, for time instants

$$i = kl + \left\lceil \frac{1}{R}\log\binom{K}{M}\right\rceil + 1, \ldots, (k+1)l$$

the encoder uses Q_k to encode the source symbol x_i and transmits $Q_k(x_i)$ to the receiver. In the first $\lceil \frac{1}{R}\log\binom{K}{M}\rceil$ time instances of the $(k+1)$st block, that is, while the index of the quantizer Q_k is communicated, the decoder emits an arbitrary symbol $\hat{x}_i$. In the remainder of the block, the decoder uses Q_k to decode the transmitted $\hat{x}_i = Q_k(x_i)$.

Choosing $\eta = c_1\sqrt{\log n/(nl)}$ one obtains, as it is implicitly proven in Theorem 1 and Corollary 2 in [5], that for all $x^n \in [0,1]^n$, the expected cumulative distortion of this scheme is bounded as

$$\frac{1}{n}\mathbb{E}\left[\sum_{i=1}^{n}(x_i-\hat{x}_i)^2\right] - \min_{Q\in\mathcal{Q}}\frac{1}{n}\sum_{i=1}^{n}(x_i-Q(x_i))^2 \;\leq\; C_1\sqrt{\frac{l\log K}{n}} + \frac{C_2\log K}{l} + \frac{1}{K} \quad (2)$$

where c_1, C_1, and C_2 are positive constants depending only on M. The right-hand side of (2) is asymptotically minimized by setting $l = c_2 n^{1/3}$ and $K = c_3 n^{1/3}$ for positive constants c_2 and c_3; in this case one obtains that

$$\frac{1}{n}\mathbb{E}\left[\sum_{i=1}^{n}(x_i-\hat{x}_i)^2\right] - \min_{Q\in\mathcal{Q}}\frac{1}{n}\sum_{i=1}^{n}(x_i-Q(x_i))^2 = O(n^{-1/3}\log n).$$

To be able to set l and K this way, the encoder and the decoder need to know the sequence length n in advance. However, using the well-known method of exponentially increasing block lengths (see, e.g., [8]), the algorithm can be modified so that it performs essentially just as well without the prior knowledge of n (only the constants will slightly increase).

In the straightforward implementation of this algorithm, one has to compute the distortion for all the $\binom{K}{M}$ quantizers in $\mathcal{Q}_K$ in parallel. This method is computationally inefficient since it has to perform $O(K^M)$ computations for each input symbol, which becomes $O(n^{M/3})$ with the optimal choice $K = c_3 n^{1/3}$. Thus, the overall computational complexity of encoding a sequence of length n becomes $O(n^{1+M/3})$, and the space complexity[1] of the algorithm is $O(K^M) = O(n^{M/3})$, since the cumulative distortion for each quantizer in $\mathcal{Q}_K$ has to be stored. Clearly, this complexity is prohibitive for all except very low coding rates.

[1]Throughout this paper we do not consider specific models for storing real numbers; for simplicity we assume that a real number can be stored in a memory space of fixed size.

In the following we describe an efficient way to implement the above algorithm. The main point is that one can draw a quantizer according to the distribution in (1) without computing the cumulative distortions $D_t(Q)$ for all $Q \in \mathcal{Q}_K$.

Theorem 1 *For any $n \geq 1$, $M \geq 2$, $K > M$, and $l > \log \binom{K}{M} / \log M$, there exists a zero-delay source coding scheme of rate $R = \log M$ for coding sequences of length n such that for all $x^n \in [0,1]^n$,*

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{x}_i)^2\right] - \min_{Q\in\mathcal{Q}} \frac{1}{n}\sum_{i=1}^{n}(x_i - Q(x_i))^2 \leq C_1\sqrt{\frac{l \log K}{n}} + C_2\frac{\log K}{l} + \frac{3}{K}$$

where C_1, C_2 are positive constants that depend only on M, and the coding procedure has $O(MK^2 n/l)$ computational complexity and $O(MK^2)$ space complexity.

Remarks. It is easy to check that to minimize the above upper bound one has to choose $l = c_2' n^{1/3}$ and $K = c_3' n^{1/3}$ for positive constants c_2' and c_3'. This way a distortion redundancy of $O(n^{-1/3}\log n)$ is achieved. As a result, the computational complexity becomes $O(Mn^{4/3})$ and the memory need of the algorithm is $O(Mn^{2/3})$. The algorithm can also be implemented with computational complexity $O(Mn)$ (that is, linear both in n and M). In this case, to minimize the distortion we have to set $l = c_2'' n^{1/2}$ and $K = c_3'' n^{1/4}$, implying a distortion redundancy of order $n^{-1/4}\sqrt{\log n}$ and $O(Mn^{1/2})$ space complexity.

It can be shown that the actual distortion of the scheme (for the current realization of the randomizing sequence $U_1, \ldots, U_n$) is, with high probability, close to the expected performance given in the theorem. In particular, by a straightforward application of the Azuma-Hoeffding inequality (see [5] for details), for any $\epsilon > 0$,

$$\mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{x}_i)^2 - \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{x}_i)^2\right] > \epsilon\right\} \leq e^{-n\epsilon^2/(2l)}.$$

Recently, in [9] another low complexity algorithm was developed for the same problem. This algorithm uses the "follow the perturbed leader"-type prediction method of Hannan [10] and Kalai and Vempala [11], instead of the exponentially weighted average prediction. This algorithm, which is conceptually somewhat simpler than the one in the theorem, can be implemented in linear $O(Mn)$ time, and it achieves a slightly worse distortion redundancy of order $n^{-1/4}\log n$, while having only $O(Mn^{1/4})$ space complexity. However, unlike the algorithm in the theorem, the performance of the algorithm of [9] cannot be improved at the price of increasing its complexity. In other words, that algorithm cannot achieve the best known $O(n^{-1/3}\log n)$ distortion redundancy.

Proof of Theorem 1 In the proof we use the algorithm of [5] but we draw the random quantizers Q_k in a computationally efficient way.

Let I_B denote the indicator function of the event B. For any fixed k and $z < \hat{z}$ such that $z, \hat{z} \in \widehat{C}^{(K)} = C^{(K)} \cup \{0, 1\}$, let

$$\Delta_k(z, \hat{z}) = \begin{cases} \sum_{i=1}^{kl} I_{\{x_i \le \hat{z}\}}(x_i - \hat{z})^2 & \text{if } z = 0; \\ \sum_{i=1}^{kl} I_{\{x_i \in (z, \frac{z+\hat{z}}{2}]\}}(x_i - z)^2 + I_{\{x_i \in (\frac{z+\hat{z}}{2}, \hat{z}]\}}(x_i - \hat{z})^2 & \text{if } 0 < z < \hat{z} < 1; \\ \sum_{i=1}^{kl} I_{\{x_i \ge z\}}(x_i - z)^2 & \text{if } 0 < z \text{ and } \hat{z} = 1. \end{cases} \tag{3}$$

Define $z_0 = 0$, $z_{M+1} = 1$, and denote the code points of $Q \in \mathcal{Q}_K$ by $z_1 < \ldots < z_M$. Then for $j = 1, \ldots, M+1$, $\Delta_k(z_{j-1}, z_j)$ denotes the partial distortion of Q in the interval (z_{j-1}, z_j) when quantizing the sequence $x^{kl} = (x_1, \ldots, x_{kl})$, and the distortion $D_{kl}(Q)$ of Q can be decomposed as

$$D_{kl}(Q) = \sum_{j=1}^{M+1} \Delta_k(z_{j-1}, z_j).$$

Next we provide an algorithm that for any fixed k chooses a quantizer randomly according to the distribution $\{p_k(Q)\}$ given in (1). This algorithm assumes that the partial distortions $\Delta_k(z, \hat{z})$ are known for all $z < \hat{z}$, $z, \hat{z} \in \widehat{C}^{(K)}$. The efficient computation of the $\Delta_k(z, \hat{z})$ will be treated later.

We construct Q_k by choosing its code points sequentially, in an increasing order: first we compute the distribution of the smallest code point and draw the code point randomly according to this distribution; having chosen the smallest $m-1$ code points, we compute the conditional distribution of the mth smallest code point, and draw the code point according to this distribution. After having chosen all the M code points, the resulting quantizer Q_k (a random object) will satisfy $\mathbb{P}(Q_k = Q) = p_k(Q)$ for all $Q \in \mathcal{Q}_K$.

For any $1 \le m \le M$ and $z_1 < \ldots < z_m$, $z_i \in C^{(K)}$, let $\mathcal{Q}(z_1, \ldots, z_m) \subset \mathcal{Q}_K$ denote the set of M-level quantizers in $\mathcal{Q}_k$ with m smallest code points $z_1 < \ldots < z_m$. For $m = 0$ define formally $\mathcal{Q}(z_1, \ldots, z_m) = \mathcal{Q}_K$. Let $p_k(z_m|z_{m-1}, \ldots, z_1)$ denote the probability that the mth code point of Q_K is z_m given the smallest $m-1$ code points are $z_1 < \ldots < z_{m-1}$. Clearly, for $m = 1$ we have

$$p_k(z_1) = \sum_{Q \in \mathcal{Q}(z_1)} p_k(Q) \tag{4}$$

and for $m \ge 2$,

$$p_k(z_m|z_{m-1}, \ldots, z_1) = \frac{\sum_{Q \in \mathcal{Q}(z_1, \ldots, z_m)} p_k(Q)}{\sum_{Q \in \mathcal{Q}(z_1, \ldots, z_{m-1})} p_k(Q)}. \tag{5}$$

To compute these probabilities efficiently, for any $z \in C_0^{(K)} = C^{(K)} \cup \{0\}$ define

$$G_k(1, z) = e^{-\eta \Delta(z,1)}$$

and for $2 \le m \le M+1$ and $z \in C_0^{(K)}$ define

$$G_k(m, z) = \sum_{z_2 > z} \sum_{z_3 > z_2} \cdots \sum_{z_m > z_{m-1}} e^{-\eta(\Delta_t(z,z_2) + \Delta_t(z_m,1))} \prod_{j=2}^{m-1} e^{-\eta \Delta_t(z_j, z_{j+1})}.$$

where $z_i \in C^{(K)}$ for all i. Setting $z_1 = z$ and $z_{m+1} = 1$, we can simplify the notation as

$$G_k(m, z) = G_k(m, z_1) = \sum_{z_2 > z_1} \cdots \sum_{z_m > z_{m-1}} \prod_{j=1}^{m} e^{-\eta \Delta_t(z_j, z_{j+1})}.$$

Expressions (4) and (5) can be rewritten in terms of $G_k(\cdot, \cdot)$. Introducing the notation $z_0 = \hat{z}_0 = 0$ and $z_{M+1} = \hat{z}_{M+1} = 1$, for $m = 1$ we have

$$\begin{aligned} p_k(z_1) &= \frac{\sum_{z_2 > z_1} \cdots \sum_{z_M > z_{M-1}} \prod_{j=0}^{M} e^{-\eta \Delta_k(z_j, z_{j+1})}}{\sum_{\hat{z}_1 > \hat{z}_0} \cdots \sum_{\hat{z}_M > \hat{z}_{M-1}} \prod_{j=0}^{M} e^{-\eta \Delta_k(\hat{z}_j, \hat{z}_{j+1})}} \\ &= e^{-\eta \Delta_k(z_0, z_1)} \frac{G_k(M, z_1)}{G_k(M+1, 0)}. \end{aligned} \tag{6}$$

For $2 \le m \le M$, letting $\hat{z}_j = z_j$ for $j = 1, \ldots, m-1$, we have

$$\begin{aligned}
& p_k(z_m | z_{m-1}, \ldots, z_1) \\
&= \frac{\sum_{z_{m+1} > z_m} \cdots \sum_{z_M > z_{M-1}} \prod_{j=0}^{M} e^{-\eta \Delta_k(z_j, z_{j+1})}}{\sum_{\hat{z}_m > \hat{z}_{m-1}} \cdots \sum_{\hat{z}_M > \hat{z}_{M-1}} \prod_{j=0}^{M} e^{-\eta \Delta_k(\hat{z}_j, \hat{z}_{j+1})}} \\
&= \frac{e^{-\eta(\Delta_k(0,z_1) + \cdots + \Delta_k(z_{m-1}, z_m))} \sum_{z_{m+1} > z_m} \cdots \sum_{z_M > z_{M-1}} \prod_{j=m}^{M} e^{-\eta \Delta_k(z_j, z_{j+1})}}{e^{-\eta(\Delta_k(0,z_1) + \cdots + \Delta_k(z_{m-2}, z_{m-1}))} \sum_{\hat{z}_m > \hat{z}_{m-1}} \cdots \sum_{\hat{z}_M > \hat{z}_{M-1}} \prod_{j=m-1}^{M} e^{-\eta \Delta_k(\hat{z}_j, \hat{z}_{j+1})}} \\
&= e^{-\eta \Delta_k(z_{m-1}, z_m)} \frac{G_k(M - m + 1, z_m)}{G_k(M - m + 2, z_{m-1})}. \qquad (7)
\end{aligned}$$

Note that (7) reduces to (6) for $m = 1$.

The values of $G_k(m, z)$ can be computed for all $z \in C_0^{(K)}$ and $m = 2, \ldots, M+1$ via the following recursion.

$$\begin{aligned}
& G_k(m, z) \\
&= \sum_{z_2 > z} \sum_{z_3 > z_2} \cdots \sum_{z_m > z_{m-1}} e^{-\eta(\Delta_k(z, z_2) + \Delta_k(z_m, 1))} \prod_{j=2}^{m-1} e^{-\eta \Delta_k(z_j, z_{j+1})} \\
&= \sum_{z_2 > z} e^{-\eta \Delta_k(z, z_2)} \sum_{z_3 > z_2} \cdots \sum_{z_m > z_{m-1}} e^{-\eta \Delta_k(z_m, 1)} \prod_{j=2}^{m-1} e^{-\eta \Delta_k(z_j, z_{j+1})} \\
&= \sum_{\hat{z} > z} e^{-\eta \Delta_k(z, \hat{z})} G_k(m-1, \hat{z}). \qquad (8)
\end{aligned}$$

Note that the case $z = 0$ has to be considered only when $m = M + 1$.

In summary, we have the following algorithm.

Algorithm 1 (Drawing a random quantizer according to (1))

Input: $\mathtt{M}, \mathtt{K}, \Delta_{\mathtt{k}}(\cdot,\cdot)$.

$\mathtt{G_k(1,z)} := \mathtt{e}^{-\eta\Delta(\mathtt{z},1)}$ `for all` $\mathtt{z} \in \mathtt{C}^{(\mathtt{K})}$, $\mathtt{z_0} := 0$.

`For` $\mathtt{m} := 2$ `to` $\mathtt{M}+1$

  `compute` $\mathtt{G_k(m,z)}$ `using (8) for all` $\mathtt{z} \in \mathtt{C}^{(\mathtt{K})}$ `(also for` $\mathtt{z}=0$ `if` $\mathtt{m}=\mathtt{M}+1$`).`

`For` $\mathtt{m} := 1$ `to` $\mathtt{M}$

  `compute` $\mathtt{p_k}(\mathtt{z_m}|\mathtt{z_{m-1}},\ldots,\mathtt{z_1})$ `for all` $\mathtt{z_m} > \mathtt{z_{m-1}}$, $\mathtt{z_m} \in \mathtt{C}^{(\mathtt{K})}$ `according to (7);`

  `choose` $\mathtt{z_m}$ `randomly according to the computed conditional probability distribution.`

`Let` $\mathtt{Q_k}$ `be a nearest-neighbor quantizer with code points` $\mathtt{z_1} < \ldots < \mathtt{z_M}$.

From the derivation of the algorithm the following lemma is straightforward:

Lemma 1 *The quantizer Q_k generated by Algorithm 1 satisfies (1).*

Since $|C^{(K)}| = K$, the complexity to compute $G_k(m,z)$ from the function $G_k(m-1,\cdot)$ is proportional to K, and since z can be chosen in K ways, the computation of $G_k(m,\cdot)$ from $G_k(m-1,\cdot)$ has complexity $O(K^2)$. Thus the computation of G_k for all possible values has complexity $O(MK^2)$, which in turn implies that the computational complexity of Algorithm 1 is also $O(MK^2)$, provided the partial distortions $\Delta_k(z,\hat{z})$ are known.

To maintain these distortion values, for each input symbol x_i we have to update the distortion of each interval $(z,\hat{z})$ containing x_i. Since the number of such intervals can vary from approximately K to $K^2/4$, this implies extra computations of the order of $O(nK^2)$ for the whole sequence, making the overall computational complexity $O(nK^2) + O(MK^2n/l)$, which becomes $O(Mn^{5/3})$ in the minimum distortion case when both l and K are proportional to $n^{1/3}$.

The amount of necessary computations can be reduced by storing only approximate distortion values, at the price of only slightly increasing the normalized cumulative distortion. The idea is that instead of the original sequence x^n, we use its finely quantized version $\bar{x}^n = (\bar{x}_1,\ldots,\bar{x}_n)$ to compute the approximate distortion values which are then used to determine the distribution for generating the random quantizers. The $\bar{x}_i$ are obtained via a K-level uniform scalar quantizer, that is,

$$\bar{x}_i = \begin{cases} \frac{\lfloor K x_i \rfloor}{K} + \frac{1}{2K} & \text{if } x_i < 1; \\ \frac{2K-1}{2K} & \text{if } x_i = 1 \end{cases}$$

(here $\lfloor x \rfloor$ denotes the largest integer not greater than x). It is easy to check that for any nearest neighbor quantizer Q with code points in $[0,1]$, we have

$$\max_{x \in [0,1]} |(x - Q(x))^2 - (\bar{x} - Q(\bar{x}))^2| \le 1/K$$

where $\bar{x}$ is the K-level uniform scalar quantized version of x. Thus for any sequence $Q_0, Q_1, \ldots, Q_{n/l-1}$ of quantizers in $\mathcal{Q}$,

$$\begin{aligned} &\mathbb{E}\left[\frac{1}{n}\sum_{k=0}^{n/l-1}\sum_{i=kl+1}^{(k+1)l}(x_i - Q_k(x_i))^2\right] - \min_{Q' \in \mathcal{Q}} \frac{1}{n}\sum_{i=1}^{n}(x_i - Q'(x_i))^2 \\ &\quad \le \mathbb{E}\left[\frac{1}{n}\sum_{k=0}^{n/l-1}\sum_{i=kl+1}^{(k+1)l}(\bar{x}_i - Q_k(\bar{x}_i))^2\right] - \min_{Q' \in \mathcal{Q}} \frac{1}{n}\sum_{i=1}^{n}(\bar{x}_i - Q'(\bar{x}_i))^2 + \frac{2}{K}. \qquad (9) \end{aligned}$$

Define $\widehat{\Delta}_k(z, \hat{z})$ for all $z < \hat{z}$ and k as $\Delta_k(z, \hat{z})$ was defined in (3), but with $\bar{x}_i$ in place of x_i. That is,

$$\widehat{\Delta}_k(z, \hat{z}) = \begin{cases} \sum_{i=1}^{kl} I_{\{\bar{x}_i \le \hat{z}\}}(\bar{x}_i - \hat{z})^2 & \text{if } z = 0; \\ \sum_{i=1}^{kl} I_{\{\bar{x}_i \in (z, \frac{z+\hat{z}}{2}]\}}(\bar{x}_i - z)^2 + I_{\{\bar{x}_i \in (\frac{z+\hat{z}}{2}, \hat{z}]\}}(\bar{x}_i - \hat{z})^2 & \text{if } 0 < z < \hat{z} < 1; \\ \sum_{i=1}^{kl} I_{\{\bar{x}_i \ge z\}}(\bar{x}_i - z)^2 & \text{if } 0 < z \text{ and } \hat{z} = 1. \end{cases} \qquad (10)$$

Then for $j = 1, \ldots, M+1$, $\widehat{\Delta}_k(z_{j-1}, z_j)$ denotes the partial distortion of the quantizer Q with code points $z_1 < \ldots < z_M$ in the interval (z_{j-1}, z_j) when applied to the sequence $\bar{x}_1, \ldots, \bar{x}_{kl}$. Unlike Δ_k, $\widehat{\Delta}_k$ can be computed efficiently for all k.

For each time instant t, define the histogram

$$h_t(j) = \sum_{i=1}^{t} I_{\{\bar{x}_i = \frac{2j-1}{2K}\}} \quad j = 1, \ldots, K$$

counting the number of input symbols falling in the jth cell of the K-level uniform quantizer. Clearly, $h_t(j)$ can easily be computed using constant computational capacity in each time instant. (The index j satisfying $\bar{x}_i = (2j-1)/(2K)$ can be identified in constant time; then $h_t(j)$ is increased by one.) This way the $h_{kl}(i)$ are immediately available at the end of the kth block. The next lemma, which is proved in the Appendix (Algorithms 3–5) shows that using h_{kl}, $\widehat{\Delta}_k(\cdot,\cdot)$ can be computed efficiently.

Lemma 2 *Given K and $h_{kl}(i), i = 1, \ldots, K$, the values of $\widehat{\Delta}_k(z, \hat{z})$ for all $z < \hat{z}$ ($z, \hat{z} \in \widehat{C}^{(K)}$) can be computed in $O(K^2)$ time.*

Using this lemma we obtain the following zero-delay source coding scheme:

Algorithm 2 (Universal low-complexity zero-delay source coding scheme)

Input: $\mathtt{n, M, K, l, x_1, \ldots, x_n}$.

```
k := 0 and h_0(j) := 0 for all j.
For i := 1 to n
  if i − 1 = kl then
    compute Δ̂_k(z, ẑ) for all z < ẑ (using Algorithms 3-5 with input K, h_kl(·));
    choose randomly Q_k using Algorithm 1 with input M, K, Δ̂_k(·,·);
  h_i(j) := h_{i−1}(j) + I_{x_i = (2j−1)/(2K)} for all j
  if i − kl ≤ ⌈(1/R) log (K choose M)⌉
    then transmit the corresponding index symbol for Q_k;
    else transmit Q_k(x_i);
  if i = (k + 1)l then k := k + 1.
```

By (2) and (9), the above coding scheme can be decoded with expected distortion redundancy

$$C_1\sqrt{\frac{l\log K}{n}}+\frac{C_2\log K}{l}+\frac{3}{K}$$

and the encoding procedure has a computational complexity $O(MK^2n/l)$ and $O(MK^2)$ space complexity (decoding can obviously be performed in linear time with $O(M)+O(K)$ space complexity). $\square$

Remarks. Algorithm 2 may be difficult to implement on-line since in order to choose a quantizer randomly at the end of each block, $O(MK^2)$ computations have to be performed during a single time slot. With the choice of parameters $l = c_2''n^{1/2}$ and $K = c_3''n^{1/4}$ yielding linear complexity in n, this amounts to $O(Mn^{1/2})$ computations during one time slot. To alleviate this problem, one can modify the algorithm so that Q_k is determined during the $(k+1)$st block which is of length $O(n^{1/2})$, and then Q_k can be applied in the $(k+2)$nd block instead of the $(k+1)$st block. This way at each time instant only a constant number of computations is carried out. It is not difficult to see that this modification results in essentially the same distortion redundancy, and only the constants will slightly increase.

Although, in principle, only one random number is needed to generate the code points $z_1,\ldots,z_M$ in Algorithm 1, in practice one may want to use M random numbers (one for each code point). In this case, the additional condition $l \geq M$ should be satisfied (this always holds for large enough n if either $l = c_2'n^{1/3}$ or $l = c_2''n^{1/2}$).

Even though here we only consider squared distortion, most of the arguments presented above generalize in a quite straightforward way to more general distortion measures. In particular, it is easy to see that for difference distortion measures of the form $\rho(|x-\hat{x}|)$ where ρ is nondecreasing and Lipschitz on $[0,1]$, Algorithm 1 can be modified in a natural manner so that Lemma 1 remains true. The modified algorithm preserves the computational complexity of order nK^2+MK^2n/l. Moreover, a bound similar to Theorem 1 holds with modified constants. To construct an algorithm with a reduced complexity similar to Algorithm 2, additional assumptions on the distortion measure may be needed. If, for example, $\rho(|x-\hat{x}|) = |x-\hat{x}|^r$ for a positive integer r, then Algorithm 2 may be modified by straightforward adjustments in Algorithms 3–5.

3 Minimax distortion redundancy for memoryless sources

The purpose of this section is to show that if the source is a stationary and memoryless random sequence, then the rate of convergence may be speeded up so that the distortion redundancy is of order $n^{-1/2}$, as opposed to the $O(n^{-1/3}\log n)$ proved by Weissman and Merhav [5] for individual sequences. We first prove a lower bound of order $n^{-1/2}$ that holds not only for the reference class of all scalar quantizers, but also for the entire reference class of all zero-delay coding schemes.

We assume that the source is a sequence of independent and identically distributed (i.i.d.) random variables $\{X_i\}_{i=1}^{\infty}$, the randomizing sequence $\{U_i\}_{i=1}^{\infty}$ is independent of the source, and both the source and the randomizing sequence take values in the interval $[0,1]$. Consider any zero-delay encoder-decoder sequence $\{f_i, g_i\}_{i=1}^{\infty}$, where, as before

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

and

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

so that the channel input at time i is $Y_i = f_i(X^i, U^i)$ and the reconstruction is $\hat{X}_i = g_i(Y^i)$, $i = 1,2,\ldots$.

The following lemma was proved (in different forms) by Ericson [12] and Gaarder and Slepian [13] (see also [14]). It states that, for memoryless sources, the best performance over the class of zero-delay codes is achieved by a (memoryless) scalar quantizer. We give the short proof for completeness.

Lemma 3 *If $\{X_i\}_{i=1}^{\infty}$ is a sequence of independent random variables, then for any sequence $\{f_i, g_i\}_{i=1}^{\infty}$ we have for all $i \geq 1$,*

$$\mathbb{E}\left[(X_i - \hat{X}_i)^2\right] \geq \min_{Q\in\mathcal{Q}} \mathbb{E}\left[(X_i - Q(X_i))^2\right]$$

where $\mathcal{Q}$ denotes the class of scalar nearest-neighbor quantizers with M reconstruction *levels.*

Proof. Define the "reproduction coder" $\hat{g}_i : [0,1]^i \times [0,1]^i \to [0,1]$ by

$$\hat{X}_i = \hat{g}_i(X^i, U^i) = g_i(f_1(X_1, U_1), \ldots, f_i(X^i, U_i)).$$

Denote the distribution of (X^{i-1}, U^i) by μ and recall that (X^{i-1}, U^i) and X_i are independent. Thus

$$\begin{aligned}
\mathbb{E}\left[(X_i - \hat{X}_i)^2\right] &= \mathbb{E}\left[(X_i - \hat{g}_i(X^i, U^i))^2\right] \\
&= \int \mathbb{E}\left[(X_i - \hat{g}_i(X^i, U^i))^2 \middle| X^{i-1} = x^{i-1}, U^i = u^i\right] d\mu(x^{i-1}, u^i) \\
&= \int \mathbb{E}\left[(X_i - \hat{g}_i(X_i, x^{i-1}, u^i))^2\right] d\mu(x^{i-1}, u^i).
\end{aligned}$$

Since among $f_1, \ldots, f_i$ only f_i depends on x_i and it can take at most M values, the function $\hat{g}_i(\,\cdot\,, x^{i-1}, u^i)$ can take at most M values for each fixed (x^{i-1}, u^i). Hence, if $\mathcal{G}$ denotes the class of measurable real functions of a real variable with at most M distinct values, then for μ almost all (x^{i-1}, u^i),

$$\mathbb{E}\left[(X_i - \hat{g}_i(X_i, x^{i-1}, u^i))^2\right] \geq \inf_{g \in \mathcal{G}} \mathbb{E}\left[(X_i - g(X_i))^2\right].$$

Since the class of M-level scalar nearest-neighbor quantizers achieves the infimum on the right-hand side,

$$\inf_{g \in \mathcal{G}} \mathbb{E}\left[(X_i - g(X_i))^2\right] = \min_{Q \in \mathcal{Q}} \mathbb{E}\left[(X_i - Q(X_i))^2\right]$$

and the lemma is proved. □

It was shown in [7, Theorem 1] that for any $M \geq 3$ there exists a bounded i.i.d. sequence $\{X_i\}_{i=1}^{\infty}$ such that for some $c > 0$ and all $n \geq \frac{2}{3}M$,

$$\min_{Q \in \mathcal{Q}} \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - Q(X_i))^2\right] \geq \mathbb{E}\left[\min_{Q \in \mathcal{Q}} \frac{1}{n}\sum_{i=1}^{n}(X_i - Q(X_i))^2\right] + \frac{c}{\sqrt{n}}.$$

Combining this with Lemma 3 gives the following lower bound for bounded memoryless sequences of length n on the normalized distortion redundancy of any zero-delay scheme with respect to the best scalar quantizer matched to the entire sequence.

Theorem 2 *For any $M \geq 3$ there exist a stationary and memoryless source $\{X_i\}_{i=1}^{\infty}$ taking values in $[0,1]$ and a constant $c > 0$ such that for any randomizing sequence $\{U_i\}_{i=1}^{\infty}$, zero-delay encoder-decoder sequence $\{f_i, g_i\}_{i=1}^{\infty}$ of rate $R = \log M$, and all $n \geq \frac{2}{3}M$,*

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{X}_i)^2 - \min_{Q\in\mathcal{Q}}\frac{1}{n}\sum_{i=1}^{n}(X_i - Q(X_i))^2\right] \geq \frac{c}{\sqrt{n}}.$$

Remark. The theorem immediately implies that the minimax distortion redundancy for individual sequences is lower bounded as

$$\inf_{\{f_i,g_i\}_{i=1}^{n}}\sup_{x^n}\left(\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{x}_i)^2\right] - \min_{Q\in\mathcal{Q}}\frac{1}{n}\sum_{i=1}^{n}(x_i - Q(x_i))^2\right) \geq \frac{c}{\sqrt{n}}.$$

Note that there is a gap between this lower bound and the best known $n^{-1/3}\log n$-type upper bound given in [5].

Next we show that the $n^{-1/2}$ convergence rate is in fact achievable by a simplistic zero-delay scheme described as follows. Time is divided into exponentially increasing blocks of length $1, 2, 2^2, 2^3, \ldots$. At the end of the kth block, the encoder selects an M-level nearest neighbor quantizer Q_k, minimizing the *empirical distortion*, that is,

$$D_m(Q_k) = \operatorname*{arg\,min}_{Q\in\mathcal{Q}_{K_k}} D_m(Q)$$

where $m = 1 + 2 + \cdots + 2^{k-1} = 2^k - 1$,

$$D_m(Q) = \frac{1}{m}\sum_{i=1}^{m}(X_i - Q(X_i))^2$$

and the minimum is taken over the class $\mathcal{Q}_{K_k}$ of all M-level nearest neighbor quantizers whose code points all belong to the finite grid

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

where we choose $K_k = \lfloor 2^{k/2} \rfloor$. At the beginning of the $(k+1)$st block, the encoder describes the selected quantizer Q_k to the receiver. This may be done using $\lceil Mk/2 \rceil$ bits, that is, in at most $\lceil Mk/(2\log M) \rceil$ time periods. In the rest of the $(k+1)$st block, the encoder uses the quantizer Q_k to transmit $Q_k(X_i)$ at each time instant i.

Remark. Wu and Zhang [15] gave an algorithm with computational complexity $O(Mn)$ which finds an M-level empirically optimal quantizer for an ordered input sequence of length n. Using this algorithm it is easy to see that the zero-delay scheme defined above may be implemented at a total computational cost of $O(Mn) + O(n \log n)$, where the second term is the time needed to sort the input sequence in each block.

The performance of this zero-delay scheme may be bounded as follows.

Theorem 3 *Consider the scheme described above and assume that $X_1, X_2, \ldots$ are independent and identically distributed random variables taking values in $[0,1]$. Then there exists a constant c, depending on M only, such that*

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \widehat{X}_i)^2 - \min_{Q\in\mathcal{Q}}\frac{1}{n}\sum_{i=1}^{n}(X_i - Q(X_i))^2\right] \le \frac{c}{\sqrt{n}} .$$

Moreover, almost surely, for n sufficiently large,

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \widehat{X}_i)^2 - \min_{Q\in\mathcal{Q}}\frac{1}{n}\sum_{i=1}^{n}(X_i - Q(X_i))^2 \le \sqrt{\frac{c\log\log n}{n}} .$$

Remarks. It follows from Lemma 3 that the upper bound for the expectation also holds if the minimum is taken over *all* rate-R zero-delay schemes instead of the class of M-level scalar quantizers. Thus Theorems 2 and 3 also imply that the minimax expected distortion redundancy over the class of memoryless sources and for the reference class of all zero-delay schemes is sandwiched between constant multiples of $n^{-1/2}$.

It is easy to see that the above described simplistic scheme fails in the individual sequence setting. This can be shown by constructing a sequence for which the scheme performs poorly (we use a construction from [5] where the Hamming distortion measure was considered). For simplicity consider the case $M = 2$ and assume that $x_i \in \{0, 1/2, 1\}$

for all $i = 1, \ldots, 2^k - 1$. Since the empirically optimal quantizer Q_k has only two code points, it is always possible in the $(k+1)$st block to choose $x_{k+1}^{\max} \in \{0, 1/2, 1\}$ such that $|x_{k+1}^{\max} - Q_k(x_{k+1}^{\max})| \geq \frac{1}{4}$. We let all x_i in the $(k+1)$st block be equal to $x_{k+1}^{\max}$, so that $(x_i - Q_k(x_i))^2 \geq \frac{1}{16}$ for all $2^k \leq i < 2^{k+1}$. Thus the normalized cumulative distortion for this sequence is at least $\frac{1}{16}$ for all n. On the other hand, for any $2^k \leq i < 2^{k+1}$, let Q_i^* denote a quantizer with two code points that is empirically optimal for x^i. Let p_0, p, and p_1 denote the empirical frequencies in the sequence x^i of 0, 1/2, and 1, respectively, and assume without loss of generality that $p_0 < p_1$ (i.e., $p_0 < (1-p)/2$). Then the Lloyd conditions for quantizer optimality [16] imply that 1 must be a code point of Q_i^*, and the other code point of Q_i^* lies in the interval $[0, 1/2]$. The distortion of Q_i^* on x_i is easily seen to equal $\frac{p_0 p}{4(p_0+p)}$, an expression whose maximum in p_0 under the constraint $p_0 \leq (1-p)/2$ is $\frac{3}{4} - \frac{1}{\sqrt{2}}$. Thus the empirical distortion of Q_i^* on x^i is at most $\frac{3}{4} - \frac{1}{\sqrt{2}}$, so the distortion redundancy of the simplistic scheme is at least $\frac{1}{16} - \left(\frac{3}{4} - \frac{1}{\sqrt{2}}\right) > 0$ for all n.

Proof of Theorem 3. Denote the "expected" distortion of the empirically selected quantizer Q_k by

$$D(Q_k) = \mathbb{E}\left[(X - Q_k(X))^2 | X_1, \ldots, X_m\right]$$

where X has the same distribution as the X_i and is independent of them. Also, let the distortion of the optimal quantizer be denoted by

$$D^* = \min_{Q \in \mathcal{Q}} \mathbb{E}(X - Q(X))^2.$$

It was shown by Linder, Lugosi, and Zeger [6] (see also Linder [17]) that

$$D(Q_k) - \min_{Q \in \mathcal{Q}_{K_k}} D(Q) \leq 2 \max_{Q \in \mathcal{Q}_{K_k}} |D(Q) - D_m(Q)| \leq 2 \sup_{Q \in \mathcal{Q}} |D(Q) - D_m(Q)| \qquad (11)$$

and also that

$$\mathbb{E} \sup_{Q \in \mathcal{Q}} |D(Q) - D_m(Q)| \leq \frac{c}{\sqrt{m}} \qquad (12)$$

where the constant c only depends on M. (In the rest of the proof c denotes a constant depending on M only whose value may change from line to line.) Combining these results with the fact that, by Lemma 2 in [1],

$$\min_{Q \in \mathcal{Q}_{K_k}} D(Q) - D^* \le \frac{1}{K_k} \le \frac{2}{\sqrt{m}} \, ,$$

we conclude that $\mathbb{E}D(Q_k) - D^* \le cm^{-1/2}$ for a constant c depending on M. To analyze the expected distortion of the zero-delay scheme, recall that in the $(k+1)$st block the first at most $\lceil Mk/(2\log M)\rceil$ time instances are used to transmit the quantizer Q_k and the contribution of this part to the cumulative distortion is at most $\lceil Mk/(2\log M)\rceil$. In the rest of the $(k+1)$st block, the cumulative distortion

$$\sum_{i=m+\lceil Mk/(2\log M)\rceil}^{2m-1} (X_i - \widehat{X}_i)^2$$

conditionally, given $X_1, \ldots, X_m$, is a sum of i.i.d. random variables, with expected value $D(Q_k)$.

To bound the expected cumulative distortion, let n be arbitrary such that n falls in the $(k+1)$st block, that is, $2^k \le n \le 2^{k+1} - 1$. By the argument above,

$$\begin{aligned}
\mathbb{E}\sum_{i=1}^{n}(X_i - \widehat{X}_i)^2 &= \sum_{j=1}^{k} \mathbb{E} \sum_{i=2^{j-1}}^{2^j-1} (X_i - \widehat{X}_i)^2 + \mathbb{E}\sum_{i=2^k}^{n}(X_i - \widehat{X}_i)^2 \\
&\le \sum_{j=1}^{k}\left(\left\lceil \frac{M(j-1)}{2\log M} \right\rceil + 2^{j-1}\mathbb{E}D(Q_{j-1}) \right) \\
&\quad + \left\lceil \frac{Mk}{2\log M} \right\rceil + (n - 2^k)\mathbb{E}D(Q_k) \\
&\le \sum_{j=1}^{k}\left(\frac{M(j-1)}{2\log M} + 2^{j-1}\left(D^* + \frac{c}{\sqrt{2^{j-1}}} \right) \right) \\
&\quad + \frac{Mk}{2\log M} + (n - 2^k)\left(D^* + \frac{c}{\sqrt{2^k}} \right) + k + 1 \\
&\le \frac{M}{2\log M}\frac{k(k+1)}{2} + nD^* + c\sum_{j=1}^{k+1} 2^{(j-1)/2} + k + 1.
\end{aligned}$$

Since $k \le \log n$, we obtain that

$$\mathbb{E}\sum_{i=1}^{n}(X_i - \widehat{X}_i)^2 - nD^* \le c\left(\log^2 n + \sqrt{n}\right) .$$

Finally, since $nD^* - n\min_{Q\in\mathcal{Q}} D_n(Q) \le n\sup_{Q\in\mathcal{Q}}|D(Q) - D_n(Q)|$ whose expected value is bounded by a constant times $n^{1/2}$, the proof of the first statement is complete.

In the proof of the second statement, we use the following version of Kolmogorov's inequality (see, e.g., Rényi [18]):

Lemma 4 *If $Y_1, \ldots, Y_n$ are zero-mean i.i.d. random variables with variance σ^2 then for all $t > 0$,*

$$\mathbb{P}\left\{\max_{i\le n}\sum_{\ell=1}^{i} Y_\ell > t\right\} \le \frac{4}{3}\mathbb{P}\left\{\sum_{i=1}^{n} Y_i > t - 2\sigma\sqrt{n}\right\} .$$

In particular, if the Y_i take their values in the interval $[-1, 1]$ then $\sigma \le 1$, and by Hoeffding's inequality [19], for any $t > 0$,

$$\mathbb{P}\left\{\max_{i\le n}\sum_{\ell=1}^{i} Y_\ell > 2\sqrt{n} + t\right\} \le \frac{4}{3}e^{-t^2/2n} .$$

To prove the almost sure statement of the theorem, first note that it follows by the bounded differences inequality of McDiarmid [20] that for any $\epsilon > 0$,

$$\mathbb{P}\left\{\sup_{Q\in\mathcal{Q}}|D(Q) - D_m(Q)| > \mathbb{E}\sup_{Q\in\mathcal{Q}}|D(Q) - D_m(Q)| + \epsilon\right\} \le e^{-2m\epsilon^2} . \tag{13}$$

Thus, the total distortion over the jth block may be bounded as

$$\begin{aligned}
&\sum_{i=2^{j-1}}^{2^j-1}(X_i - \widehat{X}_i)^2 \\
&\le \left\lceil\frac{M(j-1)}{2\log M}\right\rceil + \sum_{i=2^{j-1}}^{2^j-1}\left((X_i - \widehat{X}_i)^2 - \mathbb{E}\left[(X_i - \widehat{X}_i)^2 | X_1, \ldots, X_{2^{j-1}-1}\right]\right) + 2^{j-1}D(Q_{j-1}) \\
&= \frac{M(j-1)}{2\log M} + 1 + \sum_{i=2^{j-1}}^{2^j-1} Y_i + 2^{j-1}D^* + 2^{j-1}\left(D(Q_{j-1}) - D^*\right)
\end{aligned}$$

$$\leq \frac{M(j-1)}{2\log M} + 1 + \sum_{i=2^{j-1}}^{2^j-1} Y_i + 2^{j-1}D^* + 2^{j-1}\left(\frac{c}{\sqrt{2^{j-1}}} + Z_j\right)$$

where we denote

$$Y_i = (X_i - \widehat{X}_i)^2 - \mathbb{E}\left[(X_i - \widehat{X}_i)^2 | X_1, \ldots, X_{2^{j-1}-1}\right] , \quad i = 2^{j-1}, \ldots, 2^j - 1$$

and

$$Z_j = 2\left(\sup_{Q\in\mathcal{Q}} |D(Q) - D_{2^{j-1}}(Q)| - \mathbb{E}\sup_{Q\in\mathcal{Q}} |D(Q) - D_{2^{j-1}}(Q)|\right)$$

and the inequality follows from (11) and (12) since

$$\begin{aligned}
& D(Q_{j-1}) - D^* \leq 2\sup_{Q\in\mathcal{Q}} |D(Q) - D_{2^{j-1}}(Q)| \\
= \ & 2\mathbb{E}\left(\sup_{Q\in\mathcal{Q}} |D(Q) - D_{2^{j-1}}(Q)|\right) \\
& + 2\left(\sup_{Q\in\mathcal{Q}} |D(Q) - D_{2^{j-1}}(Q)| - \mathbb{E}\sup_{Q\in\mathcal{Q}} |D(Q) - D_{2^{j-1}}(Q)|\right) \\
\leq \ & \frac{c}{\sqrt{2^{j-1}}} + Z_j.
\end{aligned}$$

Note that conditioned on $X_1, \ldots, X_{2^{j-1}-1}$, the random variables $Y_{2^{j-1}}, \ldots, Y_{2^j-1}$ are i.i.d. with zero mean taking values in $[-1, 1]$ and by (13), Z_j is a zero-mean random variable with $\mathbb{P}\{Z_j > t/\sqrt{2^{j-1}}\} \leq e^{-t^2/2}$. Thus, by Hoeffding's inequality, and the union bound, for any $j = 1, \ldots, k$ and $t_j > 0$,

$$\begin{aligned}
& \mathbb{P}\left\{\sum_{i=2^{j-1}}^{2^j-1} (X_i - \widehat{X}_i)^2 - 2^{j-1}D^* > \frac{M(j-1)}{2\log M} + 1 + 2^{j-1}\frac{c}{\sqrt{2^{j-1}}} + t_j\sqrt{2^{j-1}}\right\} \\
& \qquad \leq \ \mathbb{P}\left\{\sum_{i=2^{j-1}}^{2^j-1} Y_i + 2^{j-1}Z_j > t_j\sqrt{2^{j-1}}\right\} \\
& \qquad \leq \ \mathbb{P}\left\{\sum_{i=2^{j-1}}^{2^j-1} Y_i > \frac{t_j\sqrt{2^{j-1}}}{2}\right\} + \mathbb{P}\left\{Z_j > \frac{t_j}{2\sqrt{2^{j-1}}}\right\} \\
& \qquad \leq \ 2e^{-t_j^2/8} .
\end{aligned}$$

The distortion accumulated during the $(k+1)$st period may be bounded similarly, though here we use Lemma 4 instead of Hoeffding's inequality. We obtain, for any $t_{k+1} > 0$,

$$
\begin{aligned}
& \mathbb{P}\left\{ \exists n \in \{2^k, \ldots, 2^{k+1}-1\} : \sum_{i=2^k}^{n} (X_i - \widehat{X}_i)^2 - (n-2^k)D^* \right. \\
& \qquad \left. > \frac{Mk}{2\log M} + 1 + (n-2^k)\frac{c}{\sqrt{2^k}} + 2\sqrt{2^k} + t_{k+1}\sqrt{2^k} \right\} \\
& \le \mathbb{P}\left\{ \max_{n\in\{2^k,\ldots,2^{k+1}-1\}} \sum_{i=2^k}^{n} Y_i + 2^k Z_{k+1} > 2\sqrt{2^k} + t_{k+1}\sqrt{2^k} \right\} \\
& \le \mathbb{P}\left\{ \max_{n\in\{2^k,\ldots,2^{k+1}-1\}} \sum_{i=2^k}^{n} Y_i > 2\sqrt{2^k} + \frac{t_{k+1}\sqrt{2^k}}{2} \right\} + \mathbb{P}\left\{ Z_{k+1} > \frac{t_{k+1}}{2\sqrt{2^k}} \right\} \\
& \le \frac{7}{3} e^{-t_{k+1}^2/8}.
\end{aligned}
$$

Choosing $t_j = \sqrt{8\log(7k^2j^2/3)}$ and using the union bound, we obtain that, for all $k \ge 1$ the probability that there exists an $n \in \{2^k, \ldots, 2^{k+1}-1\}$ such that

$$
\sum_{i=1}^{n} (X_i - \widehat{X}_i)^2 - nD^* > \frac{Mk(k+1)}{4\log M} + k + 1 + c\sum_{j=1}^{k} 2^{j/2} + 2\sqrt{n} + \sum_{j=1}^{k} \sqrt{8\log(7k^2j^2/3)}2^{j/2}
$$

is at most $\sum_{j=1}^{k} k^{-2}j^{-2} < 2k^{-2}$. Since $k \le \log n$, we obtain that there exists a constant (depending on M) such that for all $k \ge 1$ the probability that there exists an $n \in \{2^k, \ldots, 2^{k+1}-1\}$ such that

$$
\sum_{i=1}^{n} (X_i - \widehat{X}_i)^2 - nD^* > c\sqrt{n\log\log n}
$$

is at most $\sum_{j=1}^{k} k^{-2}j^{-2} < 2k^{-2}$. Applying the Borel-Cantelli lemma concludes the proof of the almost-sure statement of the theorem. □

4 Concluding remarks

We presented an efficiently computable algorithm for zero-delay lossy source coding whose normalized cumulative distortion is guaranteed to be almost as small as that of the best scalar quantizer. We have also determined the best possible convergence rate for the distortion redundancy in zero-delay lossy coding of memoryless sources.

Since our algorithm depends on the special structure of the class of all M-level nearest-neighbor scalar quantizers, it is not clear whether it can be generalized to other, richer reference classes of encoders. Such an extension would be of both practical and theoretical interest since the special reference class of all scalar quantizers somewhat limits the scope of our results. The results of Weissman and Merhav [5] on which we have built our algorithm cover all finite classes of limited-delay finite-memory coding schemes. Of special practical importance would be to extend our efficient method to the classes of sliding block codes, trellis source codes, and codes based on differential pulse code modulation (DPCM). For these classes an additional difficulty is the efficient approximation of the full reference class by a finite set of encoders from the class.

On the theoretical side, an interesting open problem is to determine whether the $n^{-1/3}\log n$ convergence rate obtained in [5] for the distortion redundancy in the case of individual sequences can be improved.

Appendix

Proof of Lemma 2 To compute $\widehat{\Delta}_k(z,\hat{z})$ we have to consider three cases.

Case 1: $z=0$ and $\hat{z}\in C^{(K)}$. Obviously we have $\widehat{\Delta}_k(0,1/(2K))=0$. Since

$$\begin{aligned}
\widehat{\Delta}_k\left(0,\frac{2j-1}{2K}\right) &= \sum_{i=1}^{j-1} h_{kl}(i)\left(\frac{2i-1}{2K}-\frac{2j-1}{2K}\right)^2 \\
&= \frac{1}{K^2}\sum_{i=1}^{j-1} h_{kl}(i)(i-j)^2 \\
&= \frac{1}{K^2}\left(j^2\sum_{i=1}^{j-1} h_{kl}(i) - 2j\sum_{i=1}^{j-1} i h_{kl}(i) + \sum_{i=1}^{j-1} i^2 h_{kl}(i)\right) \\
&\triangleq (j^2 s_1(j) - 2j s_2(j) + s_3(j))/K^2
\end{aligned}$$

we can compute $\widehat{\Delta}_k(0,\frac{2j-1}{2K})$ for increasing $j=2,\ldots,K$ by storing and computing s_1,s_2,s_3 recursively as follows.

Algorithm 3 (Computing $\widehat{\Delta}_k(0, \frac{2j-1}{2K})$)

Input: $\mathtt{K}, \mathtt{h_{kl}}(\cdot)$.

$\mathtt{s_1 := s_2 := s_3 := 0}$.

$\mathtt{For\ j := 1\ to\ K}$

$\widehat{\Delta}_\mathtt{k}(0, \frac{\mathtt{2j-1}}{\mathtt{2K}}) := (\mathtt{j^2 s_1 - 2js_2 + s_3})/\mathtt{K}^2$;

$\mathtt{s_1 := s_1 + h_{kl}(j)}$;

$\mathtt{s_2 := s_2 + jh_{kl}(j)}$;

$\mathtt{s_3 := s_3 + j^2 h_{kl}(j)}$.

Case 2: $z \in C^{(K)}$, $\hat{z} = 1$. Here, similarly to Case 1, we obtain

$$\begin{aligned}\widehat{\Delta}_k\left(\frac{2j-1}{2K}, 1\right) &= \frac{1}{K^2}\left(j^2 \sum_{i=j+1}^{K} h_{kl}(i) - 2j \sum_{i=j+1}^{K} i h_{kl}(i) + \sum_{i=j+1}^{K} i^2 h_{kl}(i)\right) \\ &\triangleq (j^2 r_1(j) - 2j r_2(j) + r_3(j))/K^2.\end{aligned}$$

Thus $\widehat{\Delta}_k(\frac{2j-1}{2K}, 1)$ can be computed recursively as follows.

Algorithm 4 (Computing $\widehat{\Delta}_k(\frac{2j-1}{2K}, 1)$)

Input: $\mathtt{K}, \mathtt{h_{kl}}(\cdot)$.

$\mathtt{r_1 := r_2 := r_3 := 0}$.

$\mathtt{For\ j := K\ to\ 1}$

$\widehat{\Delta}_\mathtt{k}(\frac{\mathtt{2j-1}}{\mathtt{2K}}, 1) := (\mathtt{j^2 r_1 - 2jr_2 + r_3})/\mathtt{K}^2$;

$\mathtt{r_1 := r_1 + h_{kl}(j)}$;

$\mathtt{r_2 := r_2 + jh_{kl}(j)}$;

$\mathtt{r_3 := r_3 + j^2 h_{kl}(j)}$.

Case 3: $z, \hat{z} \in C^{(K)}$. In this case $z = (2u-1)/(2K)$ and $\hat{z} = (2v-1)/(2K)$ for some integers $1 \le u < v \le K$. For $v = u+1$ we have $\widehat{\Delta}_k(z, \hat{z}) = 0$; otherwise $\widehat{\Delta}_k(z, \hat{z})$ can be computed recursively for increasing v, since

$$
\begin{aligned}
&\widehat{\Delta}_k\left(\frac{2u-1}{2K},\frac{2v+1}{2K}\right)-\widehat{\Delta}_k\left(\frac{2u-1}{2K},\frac{2v-1}{2K}\right)\\
=\;& \sum_{i=u+1}^{\left\lfloor\frac{u+v+1}{2}\right\rfloor} h_{kl}(i)\left(\frac{2i-1}{2K}-\frac{2u-1}{2K}\right)^2+\sum_{i=\left\lfloor\frac{u+v+1}{2}\right\rfloor+1}^{v} h_{kl}(i)\left(\frac{2i-1}{2K}-\frac{2v+1}{2K}\right)^2\\
&-\sum_{i=u+1}^{\left\lfloor\frac{u+v}{2}\right\rfloor} h_{kl}(i)\left(\frac{2i-1}{2K}-\frac{2u-1}{2K}\right)^2-\sum_{i=\left\lfloor\frac{u+v}{2}\right\rfloor+1}^{v-1} h_{kl}(i)\left(\frac{2i-1}{2K}-\frac{2v-1}{2K}\right)^2\\
=\;& \frac{1}{K^2}\Bigg(h_{kl}\left(\frac{u+v+1}{2}\right)\left(\left(\frac{v-u+1}{2}\right)^2-\left(\frac{v-u-1}{2}\right)^2\right)\\
&+h_{kl}(v)+\sum_{i=\left\lfloor\frac{u+v+1}{2}\right\rfloor+1}^{v-1} h_{kl}(i)\left((v-i+1)^2-(v-i)^2\right)\Bigg)\\
=\;& \frac{1}{K^2}\left(h_{kl}\left(\frac{u+v+1}{2}\right)(v-u)+h_{kl}(v)+\sum_{i=\left\lfloor\frac{u+v+1}{2}\right\rfloor+1}^{v-1} h_{kl}(i)(2v-2i+1)\right)\\
=\;& \frac{1}{K^2}\left(h_{kl}\left(\frac{u+v+1}{2}\right)(v-u)+h_{kl}(v)+(2v+1)\sum_{i=\left\lfloor\frac{u+v+1}{2}\right\rfloor+1}^{v-1} h_{kl}(i)-2\sum_{i=\left\lfloor\frac{u+v+1}{2}\right\rfloor+1}^{v-1} ih_{kl}(i)\right)\\
\triangleq\;& \frac{1}{K^2}\left(h_{kl}\left(\frac{u+v+1}{2}\right)(v-u)+h_{kl}(v)+(2v+1)s_1(u,v)-2s_2(u,v)\right)
\end{aligned}
$$

where $h_{kl}(a)=0$ if a is not an integer. Thus, $\widehat{\Delta}_k(z,\hat{z})$ can be computed in this case by the following algorithm.

Algorithm 5 (Computing $\widehat{\Delta}_k(\frac{2u-1}{2K}, \frac{2v-1}{2K})$)

Input: $\mathtt{K}, \mathtt{h_{kl}}(\cdot)$.

For $\mathtt{u} := 1$ to $\mathtt{K}-1$

  for $\mathtt{v} := \mathtt{u}+1$ to $\mathtt{K}$

    if $\mathtt{v} = \mathtt{u}+1$ then

      $\widehat{\Delta}_\mathtt{k}(\frac{2\mathtt{u}-1}{2\mathtt{K}}, \frac{2\mathtt{v}-1}{2\mathtt{K}}) := 0$;

      $\mathtt{s_1} := \mathtt{s_2} := 0$;

    else

      $\mathtt{s_1} := \mathtt{s_1} + \mathtt{h_{kl}}(\mathtt{v}-1) - \mathtt{h_{kl}}\left(\frac{\mathtt{u}+\mathtt{v}}{2}+1\right)$;

      $\mathtt{s_2} := \mathtt{s_2} + (\mathtt{v}-1)\mathtt{h_{kl}}(\mathtt{v}-1) - \left(\frac{\mathtt{u}+\mathtt{v}}{2}+1\right)\mathtt{h_{kl}}\left(\frac{\mathtt{u}+\mathtt{v}}{2}+1\right)$;

      $\widehat{\Delta}_\mathtt{k}(\frac{2\mathtt{u}-1}{2\mathtt{K}}, \frac{2\mathtt{v}-1}{2\mathtt{K}}) := \left(\mathtt{h_{kl}}\left(\frac{\mathtt{u}+\mathtt{v}+1}{2}\right)(\mathtt{v}-\mathtt{u}) + \mathtt{h_{kl}}(\mathtt{v}) + (2\mathtt{v}+1)\mathtt{s_1} - 2\mathtt{s_2}\right)/\mathtt{K}^2$.

Clearly, the computational complexity of Algorithm 3 and Algorithm 4 is $O(K)$ while to perform Algorithm 5 we need $O(K^2)$ operations. Thus, at the end of the kth block, determining $\widehat{\Delta}_k(z, \hat{z})$ for all $z < \hat{z}$ has computational complexity $O(K^2)$. $\square$

References

[1] T. Linder and G. Lugosi, "A zero-delay sequential scheme for lossy coding of individual sequences," *IEEE Transactions on Information Theory*, vol. 47, pp. 2533–2538, 2001.

[2] V. Vovk, "Aggregating strategies," in *Proceedings of the Third Annual Workshop on Computational Learning Theory*, pp. 372–383, Association of Computing Machinery, New York, 1990.

[3] V. Vovk, "A game of prediction with expert advice," *Journal of Computer and System Sciences*, vol. 56, pp. 153–173, 1998.

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

[5] T. Weissman and N. Merhav, "On limited-delay lossy coding and filtering of individual sequences," *IEEE Transactions on Information Theory*, vol. 48, pp. 721–733, 2002.

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

[7] T. Linder, "On the training distortion of vector quantizers," *IEEE Transactions on Information Theory*, vol. 46, pp. 1617–1623, 2000.

[8] 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.

[9] A. György, T. Linder, and G. Lugosi, "A "follow the perturbed leader"-type algorithm for zero-delay quantization of individual sequences." accepted to DCC 2004.

[10] J. Hannan, "Approximation to bayes risk in repeated plays," in *Contributions to the Theory of Games* (M. Dresher, A. Tucker, and P. Wolfe, eds.), vol. 3, pp. 97–139, Princeton University Press, 1957.

[11] A. Kalai and S. Vempala, "Efficient algorithms for the online decision problem," in *Proc. 16th Conf. on Computational Learning Theory*, (Washington, D. C., USA), 2003. available at `http://www-math.mit.edu/~vempala/papers/online.ps`.

[12] T. Ericson, "A result on delayless information transmission," in *IEEE International Symposium on Informatiuon Theory*, (Grignano, Italy), 1979.

[13] N. T. Gaarder and D. Slepian, "On optimal finite-state digital transmission systems," in *IEEE International Symposium on Information Theory*, (Grignano, Italy), 1979.

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

[15] X. Wu and K. Zhang, "Quantizer monotonicities and globally optimal scalar quantizer design," *IEEE Transactions on Information Theory*, vol. 39, pp. 1049–1053, 1993.

[16] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression.* Boston: Kluwer, 1992.

[17] T. Linder, "Learning-theoretic methods in vector quantization," in *Principles of nonparametric learning* (L. Györfi, ed.), no. 434 in CISM Courses and Lecture Notes, New York: Springer-Verlag, 2002.

[18] A. Rényi, *Probability theory.* Amsterdam: North-Holland, 1970.

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

[20] C. McDiarmid, "On the method of bounded differences," in *Surveys in Combinatorics 1989*, pp. 148–188, Cambridge University Press, Cambridge, 1989.