A "Follow the Perturbed Leader"-type Algorithm for Zero-Delay Quantization of Individual Sequences

András György Tamás Linder Gábor Lugosi

Abstract

Zero-delay lossy source coding schemes are considered for individual sequences. 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. Recently, Weissman and Merhav constructed a randomized scheme which, for any bounded individual sequence of length n, achieves a distortion redundancy $O(n^{-1/3} \log n)$. However, this scheme has prohibitive complexity (both space and time) which makes practical implementation infeasible. In this paper, we present an efficiently computable algorithm based on a "follow the perturbed leader"-type prediction method by Kalai and Vempala. Our algorithm achieves distortion redundancy $O(n^{-1/4} \log n)$, which is somewhat worse than that of the scheme by Merhav and Weissman, but it has computational complexity that is *linear* in the sequence length n, and requires $O(n^{1/4})$ storage capacity.

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 zero delay if each channel symbol y_n depends only on the source symbols $x_1, \ldots, x_n$ and the reproduction $\hat{x}_n$ for the source symbol x_n depends only on the channel symbols $y_1, \ldots, y_n$. Thus the

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, the NATO Science Fellowship of Canada, and by DGES grant PB96-0300 of the Spanish Ministry of Science and Technology.

Proceedings of the Data Compression Conference (DCC'04)
1068-0314/04 $ 20.00 © 2004 IEEE

encoder produces y_n as soon as x_n is available, and the decoder can produce $\hat{x}_n$ when y_n is received.

In this work, we concentrate on zero-delay (sequential) 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.

The study of zero-delay coding for individual sequences was initiated in [1]. There a zero-delay scheme was constructed whose normalized accumulated mean squared distortion for any bounded sequence of n source symbols is not larger than that of the best scalar quantizer that is matched to the sequence plus an error term (called the distortion redundancy) of order $O(n^{-1/5}\log n)$. The scheme was based on a generalization of exponentially weighted average prediction of individual sequences (see Vovk [2, 3], Littlestone and Warmuth [4]) and required common randomization at the encoder and the decoder. This result was improved by Weissman and Merhav [5] who constructed a zero-delay scheme which uses randomization only at the encoder and has distortion redundancy $O(n^{-1/3}\log n)$. (This is currently the best known redundancy bound for this problem.)

Although both schemes have the attractive property of performing uniformly well on individual sequences, they are computationally inefficient. 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]. Clearly, even for moderate values of the encoding rate R, these complexities make the implementation infeasible. A low complexity algorithm for implementing the scheme of [5] was recently developed in [6]. The method reduces the computational complexity to the tractable $O(2^R n^{4/3})$ without increasing the order of the distortion redundancy. In effect, the algorithm efficiently generates randomly chosen quantizers according to an exponential weighting scheme without calculating and storing the cumulative distortions of $n^{c\,2^R}$ reference quantizers as was done in [1] and [5]. By adjusting the parameters of the algorithm, the complexity of the scheme can be reduced to $O(2^R n)$ (i.e., linear in the length of the sequence) at the cost of increasing the distortion redundancy to $O(n^{-1/4}\sqrt{\log n})$.

In this paper we construct a new, low complexity algorithm for zero-delay lossy coding of individual sequences. Our approach combines the elegant "follow the perturbed leader" prediction scheme of Kalai and Vempala [7] and Hannan [8] with the coding method of [5]. The new algorithm has linear-time computational complexity $O(2^R n)$ and distortion redundancy $O(n^{-1/4}\log n)$ (almost the same as the linear-time version of the algorithm in [6]), but it can be implemented with $O(2^R n^{1/4})$ storage capacity while the method of [6] has $O(2^R n^{1/2})$ storage requirement. In addition, the new algorithm has the advantage of being conceptually simpler as it essentially manages to reduce the problem to the well-understood off-line design of empirically optimal scalar quantizers [9].

2 Zero-delay universal quantization of individual sequences

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 $U_1, U_2, \dots$ of independent random variables distributed uniformly over the interval $[0, 1]$. The input to the encoder is a sequence of real numbers $x_1, x_2, \dots$ taking values in the interval $[0, 1]$. (All results may be extended trivially for arbitrary bounded sequences of input symbols.) At each time instant $i = 1, 2, \dots$, the encoder observes x_i and the random number U_i. Based on x_i, U_i, the past input values $x^{i-1} = (x_1, \dots, x_{i-1})$, and the past values of the randomization sequence $U^{i-1} = (U_1, \dots, U_{i-1})$, the encoder produces a channel symbol $y_i \in \{1, 2, \dots, M\}$ 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 $y^i = (y_1, \dots, y_i)$ received so far.

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

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

and

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

so that $y_i = f_i(x^i, U^i)$ and $\hat{x}_i = g_i(y^i)$, $i = 1, 2, \dots$. 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 $U^n = (U_1, \dots, U_n)$.

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*. Without loss of generality, we only consider nearest neighbor quantizers Q such that $(x - Q(x))^2 = \min_{y \in \mathcal{C}}(x - y)^2$ for all x. Also, since we consider sequences with components in $[0, 1]$, we will assume without loss of generality that all quantizers Q are defined on and take values in $[0, 1]$.

Let $\mathcal{Q}$ denote the collection of all M-level nearest neighbor quantizers taking values in $[0, 1]$. 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 taken 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

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

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), and at the end of the kth block, that is, at time instants $i=kl$, $k=0,\dots,n/l-1$, a quantizer Q_k is drawn randomly (using exponential weighting) from the class $\mathcal{Q}_K$ of all M-level nearest-neighbor quantizers whose code points all belong to the finite grid

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

according to the probabilities

$$\mathbb{P}\{Q_k=Q\}=\frac{e^{-\eta\sum_{t=1}^{kl}(x_t-Q(x_t))^2}}{\sum_{\widehat{Q}\in\mathcal{Q}_K}e^{-\eta\sum_{t=1}^{kl}(x_t-\widehat{Q}(x_t))^2}} \tag{2}$$

where $\eta>0$ is a parameter used to optimize the algorithm. 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, by transmitting an index identifying Q_k (note that $|\mathcal{Q}_K|=\binom{K}{M}$), and in the rest of the block 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 instants 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 $\hat{x}_i=Q_k(x_i)$. Optimizing the values of η, K and l, the upper bound (1) is shown to hold in [5] for the expected distortion redundancy of the scheme.

3 An efficient “follow the perturbed leader”-type algorithm

In the straightforward implementation of Weissman and Merhav's 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 $O(n^{M/3})$ when K is chosen optimally to be proportional to $n^{1/3}$. Thus, the overall computational complexity of encoding a sequence of length n becomes $O(n^{1+M/3})$, and the storage requirement 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. (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.) Clearly, this complexity is prohibitive for all except very low coding rates.

Recently, an efficient implementation of the algorithm was given in [6], achieving $O(n^{-1/3}\log n)$ distortion redundancy using $O(Mn^{4/3})$ computations. The algorithm can also be implemented with $O(Mn)$ time and $O(Mn^{1/2})$ storage complexity, resulting in a slightly worse $O(n^{-1/4}\sqrt{\log n})$ distortion redundancy. This rather substantial reduction in complexity is achieved by a nontrivial sequential algorithm for drawing a quantizer according to the distribution in (2), without having to compute the cumulative distortions for all $Q \in \mathcal{Q}_K$.

In the following we combine Weissman and Merhav's [5] coding scheme with an efficient, novel prediction algorithm, due originally to Hannan [8], and recently rediscovered and simplified by Kalai and Vempala [7]. In the prediction context, the algorithm forfeits choosing predictors according to the (essentially optimal) exponential weighting method, and instead it chooses the predictor that is optimal in hindsight for a randomly perturbed version of the past data (thus the name "follow the perturbed leader"). Although sequential prediction schemes cannot directly be applied in sequential lossy coding problems (where the loss incurred at every step is *not* available at the decoder), we show how to use this idea in the context of zero-delay lossy coding. The resulting algorithm reduces the computational complexity of the online problem by solving its off-line version, that is, the problem of finding an empirically optimal quantizer for a given source sequence, which can be solved in linear time [9]. Indeed, the algorithm to be presented here has computational complexity of order of $O(Mn)$ and requires storage capacity of order of $O(Mn^{1/4})$, at the expense of a slightly increased $O(n^{-1/4}\log n)$ expected distortion redundancy.

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

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{x}_i)^2\right] - \min_{Q\in\mathcal{Q}} \frac{1}{n}\sum_{i=1}^{n}(x_i - Q(x_i))^2 \leq Cn^{-1/4}\log n \qquad (3)$$

for some constant $C > 0$ depending only on M, and the coding procedure has computational complexity $O(Mn)$ and requires $O(Mn^{1/4})$ storage capacity.

Remark. The algorithm of the theorem is conceptually simpler and uses less storage space than the linear-time version of the algorithm in [6]. However, it is less flexible in terms of a trading off complexity for performance; it appears that the distortion redundancy cannot be further reduced at the cost of slightly increasing the time complexity, as was the case for the algorithm in [6].

Proof. Fix the positive integers $K > M$ and $l > \log \binom{K}{M} / \log M$ (for simplicity assume that l divides n) to be specified later. Let q_K denote a K-level uniform quantizer on $[0,1]$ with code points $\{1/(2K), 3/(2K), \ldots, (2K-1)/(2K)\}$. Notice that we do not lose too much in terms of distortion, if instead of the sequence $x_1, \ldots, x_n$, we encode its finely quantized version $\bar{x}_1, \ldots, \bar{x}_n$, where

$$\bar{x}_i = q_K(x_i), \qquad i = 1, \ldots, n.$$

It is easy to check that for any nearest neighbor quantizer Q with code points in $[0,1]$ we have

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

Thus for any sequence $Q_0, Q_1, \ldots, Q_{n/l-1}$ of quantizers in $\mathcal{Q}$,

$$\begin{aligned} &\sum_{k=0}^{n/l-1} \sum_{i=kl+1}^{(k+1)l} (x_i - Q_k(x_i))^2 - \min_{Q \in \mathcal{Q}} \sum_{i=1}^{n} (x_i - Q(x_i))^2 \\ &\qquad \le \sum_{k=0}^{n/l-1} \sum_{i=kl+1}^{(k+1)l} (\bar{x}_i - Q_k(\bar{x}_i))^2 - \min_{Q \in \mathcal{Q}} \sum_{i=1}^{n} (\bar{x}_i - Q(\bar{x}_i))^2 + \frac{2n}{K}. \end{aligned} \tag{4}$$

Now let I_A denote the indicator function of the event A, and for $i = 1, \ldots, n$ and $j = 1, \ldots, K$, let

$$h_i(j) = \sum_{t=1}^{i} I_{\{\bar{x}_t = \frac{2j-1}{2K}\}}.$$

Hence for any i, the vector of integers $\mathbf{h}_i = (h_i(1), \ldots, h_i(K))$ describes the histogram of $\bar{x}^i$. For any vector $\mathbf{a} = (a_1, \ldots, a_K)$ with $a_j \ge 0$, $j = 1, \ldots, K$, let $Q_{\mathbf{a}} \in \mathcal{Q}$ denote a K-level quantizer that is optimal for the discrete distribution that assigns probability $a_j / |\mathbf{a}|_1$ to each point $\frac{2j-1}{2K}$, $j = 1, \ldots, K$, where $|\mathbf{a}|_1 = \sum_{j=1}^{K} |a_j|$. Thus, for example, $Q_{\mathbf{h}_n}$ is the M-level quantizer that quantizes the entire sequence $\bar{x}^n$ with minimum distortion.

For any $Q \in \mathcal{Q}$ with codebook $\{y_1, \ldots, y_K\} \subset [0,1]$, let $q_K(Q) \in \mathcal{Q}_K$ denote a nearest neighbor quantizer with codebook $\{q_K(y_1), \ldots, q_K(y_K)\}$. Notice that

$$\sup_{x \in [0,1]} |(x - Q(x))^2 - (x - q_K(Q)(x))^2| \le \frac{1}{K}. \tag{5}$$

Our coding scheme works as follows: the quantized source sequence $\bar{x}^n$ is divided into non-overlapping blocks of length l, and at the end of the kth block, that is, at time instants $i = kl$, $k = 0, \ldots, n/l - 1$, a quantizer $Q_k \in \mathcal{Q}_K$ is chosen such that

$$Q_k = q_K(Q_{\mathbf{h}_{kl} + \mathbf{V}_k})$$

where $\mathbf{V}_k$ is a random variable uniformly distributed in the K-dimensional cube $[0, 1/\epsilon]^K$ and $\epsilon > 0$ is a parameter to be specified later. 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

```
Input: $n, M, K, l, \epsilon, x_1, \ldots, x_n$.
$k := 0$ and $h_0(j) := 0$ for all $j$.
For $i := 1$ to $n$
  if $i - 1 = kl$ then
    draw $\mathbf{V}_k$ uniformly from $[0, 1/\epsilon]^K$;
    $Q_k := q_K(Q_{(h_{kl}(1),\ldots,h_{kl}(K))+\mathbf{V}_k})$;
  $\bar{x}_i := q_K(x_i)$;
  $h_i(j) := h_{i-1}(j) + I_{\{\bar{x}_i = \frac{2j-1}{2K}\}}$ for all $j$;
  if $i - kl \le \lceil \frac{1}{\log M} \log \binom{K}{M} \rceil$
    then transmit the corresponding index symbol for $Q_k$;
    else transmit $Q_k(x_i)$;
  if $i = (k+1)l$ then $k := k + 1$.
```

Figure 1: Universal low complexity zero-delay source coding scheme

selected quantizer Q_k to the receiver, (note that $|\mathcal{Q}_K| = \binom{K}{M}$)). In the rest of the block 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 instants 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 $\hat{x}_i = Q_k(x_i)$. The algorithm is summarized in Figure 1.

Upper bound on the expected distortion redundancy: Since the algorithm does not code the first $\lceil \log \binom{K}{M} / \log M \rceil$ source symbols in each block, the distortion redundancy of the coding scheme can be bounded as

$$\begin{aligned}
&\sum_{i=1}^{n} (x_i - \hat{x}_i)^2 - \min_{Q \in \mathcal{Q}} \sum_{i=1}^{n} (x_i - Q(x_i))^2 \\
&\quad\le \sum_{k=0}^{n/l-1} \sum_{i=kl+1}^{(k+1)l} (x_i - Q_k(x_i))^2 - \min_{Q \in \mathcal{Q}} \sum_{i=1}^{n} (x_i - Q(x_i))^2 + \frac{n}{l} \left\lceil \frac{\log \binom{K}{M}}{\log M} \right\rceil \\
&\quad\le \sum_{k=0}^{n/l-1} \sum_{i=kl+1}^{(k+1)l} (\bar{x}_i - Q_k(\bar{x}_i))^2 - \min_{Q \in \mathcal{Q}} \sum_{i=1}^{n} (\bar{x}_i - Q(\bar{x}_i))^2 + \frac{n}{l} \left\lceil \frac{\log \binom{K}{M}}{\log M} \right\rceil + \frac{2n}{K} \qquad (6)
\end{aligned}$$

where the second inequality holds by (4). In what follows, we bound the distortion redundancy for encoding the sequence $\bar{x}^n$. First notice that by (5)

$$\sum_{k=0}^{n/l-1} \sum_{i=kl+1}^{(k+1)l} (\bar{x}_i - Q_k(\bar{x}_i))^2 \le \sum_{k=0}^{n/l-1} \sum_{i=kl+1}^{(k+1)l} (\bar{x}_i - Q_{\mathbf{h}_{kl}+\mathbf{V}_k}(\bar{x}_i))^2 + \frac{n}{K}. \qquad (7)$$

Next we show that for any $\epsilon > 0$, the expectation of the first term on the right hand side of (7) can be bounded as

$$\mathbb{E}\left[\sum_{k=0}^{n/l-1}\sum_{i=kl+1}^{(k+1)l}(\bar{x}_i - Q_{\mathbf{h}_{kl}+\mathbf{V}_k}(\bar{x}_i))^2\right] \le \min_{Q\in\mathcal{Q}}\sum_{i=1}^{n}(\bar{x}_i - Q(\bar{x}_i))^2 + \frac{K}{\epsilon} + \epsilon nl. \qquad (8)$$

The proof of (8) is an appropriately adapted version of the proof of Theorem 1 in [7] (given in a prediction context). For any quantizer $Q \in \mathcal{Q}$ let

$$d_j(Q) = \left(\frac{2j-1}{2K} - Q\left(\frac{2j-1}{2K}\right)\right)^2 \qquad \text{for } j = 1,\ldots,K$$

and set $\mathbf{d}(Q) = (d_1(Q),\ldots,d_K(Q))$. Since $\mathbf{h}_{(k+1)l}(j) - \mathbf{h}_{kl}(j)$ is the number of times $(2j-1)/(2K)$ occurs in the sequence $\bar{x}_{kl+1},\ldots,\bar{x}_{(k+1)l}$, we have

$$\sum_{k=0}^{n/l-1}\sum_{i=kl+1}^{(k+1)l}(\bar{x}_i - Q_{\mathbf{h}_{kl}+\mathbf{V}_k}(\bar{x}_i))^2 = \sum_{k=0}^{n/l-1}\mathbf{d}(Q_{\mathbf{h}_{kl}+\mathbf{V}_k})\cdot(\mathbf{h}_{(k+1)l} - \mathbf{h}_{kl})$$

where $\mathbf{a}\cdot\mathbf{b} = \sum_{j=1}^{K} a_j b_j$ for $\mathbf{a} = (a_1,\ldots,a_K)$ and $\mathbf{b} = (b_1,\ldots,b_K)$.

As we will show later, for large values of k the distributions induced by $(\mathbf{h}_{kl} + \mathbf{V}_k)$ and $(\mathbf{h}_{(k+1)l} + \mathbf{V}_k)$ are close, so first we consider the more tractable expectation

$$\mathbb{E}\left[\sum_{k=0}^{n/l-1}\mathbf{d}(Q_{\mathbf{h}_{(k+1)l}+\mathbf{V}_k})\cdot(\mathbf{h}_{(k+1)l} - \mathbf{h}_{kl})\right].$$

Defining $\mathbf{V}_{-1} = (0,\ldots,0)$, for any $n/l \ge m \ge 1$, we have

$$\begin{aligned}&\sum_{k=0}^{m-1}\mathbf{d}(Q_{\mathbf{h}_{(k+1)l}+\mathbf{V}_k})\cdot(\mathbf{h}_{(k+1)l} + \mathbf{V}_k - \mathbf{h}_{kl} - \mathbf{V}_{k-1})\\ &\quad\le\ \mathbf{d}(Q_{\mathbf{h}_{ml}+\mathbf{V}_{m-1}})\cdot(\mathbf{h}_{ml} + \mathbf{V}_{m-1}) \le \mathbf{d}(Q_{\mathbf{h}_{ml}})\cdot(\mathbf{h}_{ml} + \mathbf{V}_{m-1}). \qquad (9)\end{aligned}$$

Here the second inequality follows since $Q_{\mathbf{h}_{ml}+\mathbf{V}_{m-1}}$ is optimal for $\mathbf{h}_{ml} + \mathbf{V}_{m-1}$, and the first inequality follows by induction: for $m = 1$, the inequality holds trivially; the induction step from m to $m+1$ follows from

$$\mathbf{d}(Q_{\mathbf{h}_{ml}+\mathbf{V}_{m-1}})\cdot(\mathbf{h}_{ml} + \mathbf{V}_{m-1}) \le \mathbf{d}(Q_{\mathbf{h}_{(m+1)l}+\mathbf{V}_m})\cdot(\mathbf{h}_{ml} + \mathbf{V}_{m-1})$$

which holds again by the optimality of $Q_{\mathbf{h}_{ml}+\mathbf{V}_{m-1}}$ for $\mathbf{h}_{ml} + \mathbf{V}_{m-1}$. Since $\mathbf{V}_{m-1} = \sum_{k=0}^{m-1}(\mathbf{V}_k - \mathbf{V}_{k-1})$, from (9) we obtain

$$\begin{aligned}&\sum_{k=0}^{m-1}\mathbf{d}(Q_{\mathbf{h}_{(k+1)l}+\mathbf{V}_k})\cdot(\mathbf{h}_{(k+1)l} - \mathbf{h}_{kl})\\ &\quad\le\ \mathbf{d}(Q_{\mathbf{h}_{ml}})\cdot\mathbf{h}_{ml} + \sum_{k=0}^{m-1}(\mathbf{d}(Q_{\mathbf{h}_{ml}}) - \mathbf{d}(Q_{\mathbf{h}_{(k+1)l}+\mathbf{V}_k}))\cdot(\mathbf{V}_k - \mathbf{V}_{k-1})\end{aligned}$$

$$\leq \min_{Q\in\mathcal{Q}} \sum_{i=1}^{ml} (\bar{x}_i - Q(\bar{x}_i))^2 + K \sum_{k=0}^{m-1} |\mathbf{V}_k - \mathbf{V}_{k-1}|_\infty \tag{10}$$

where $|\mathbf{a}|_\infty = \max_j |a_j|$, and the second inequality follows since $Q_{\mathbf{h}_{ml}}$ is optimal for $\mathbf{h}_{ml}$, $\mathbf{a}\cdot\mathbf{b} \leq |\mathbf{a}|_1 |\mathbf{b}|_\infty$ for any $\mathbf{a},\mathbf{b} \in \mathbb{R}^K$, and $0 \leq d_j(Q) \leq 1$ for all $Q \in \mathcal{Q}$ and $j = 1,\ldots,K$.

Since the expectation $\mathbb{E}\left[\sum_{k=0}^{n/l-1} \mathbf{d}(Q_{\mathbf{h}_{(k+1)l}+\mathbf{V}_k}) \cdot (\mathbf{h}_{(k+1)l} - \mathbf{h}_{kl})\right]$ does not change if $\mathbf{V}_k$ is replaced by $\mathbf{V}_0$ for all $k = 0,\ldots,n/l-1$, from (10) we get

$$\begin{aligned}\mathbb{E}\left[\sum_{k=0}^{n/l-1} \mathbf{d}(Q_{\mathbf{h}_{(k+1)l}+\mathbf{V}_k}) \cdot (\mathbf{h}_{(k+1)l} - \mathbf{h}_{kl})\right] &\leq \min_{Q\in\mathcal{Q}} \sum_{i=1}^{n} (\bar{x}_i - Q(\bar{x}_i))^2 + K\mathbb{E}\big(|\mathbf{V}_0|_\infty\big) \\ &\leq \min_{Q\in\mathcal{Q}} \sum_{i=1}^{n} (\bar{x}_i - Q(\bar{x}_i))^2 + \frac{K}{\epsilon}. \qquad (11)\end{aligned}$$

In order to prove (8) from (11), we need to give an upper bound on the expected difference in the distortion between using the quantizer $Q_{\mathbf{h}_{(k+1)l}+\mathbf{V}_k}$ instead of $Q_{\mathbf{h}_{kl}+\mathbf{V}_k}$. Notice that $\mathbf{h}_{(k+1)l} + \mathbf{V}_k$ and $\mathbf{h}_{kl} + \mathbf{V}_k$ are both uniformly distributed over cubes. Assuming that both $\mathbf{h}_{(k+1)l} + \mathbf{V}_k$ and $\mathbf{h}_{kl} + \mathbf{V}_k$ fall into the intersection of the two cubes, their conditional distributions are the same (both being uniform), and hence the corresponding conditional expectations of $\mathbf{d}(Q_{\mathbf{h}_{(k+1)l}+\mathbf{V}_k})\cdot(\mathbf{h}_{(k+1)l} - \mathbf{h}_{kl})$ and $\mathbf{d}(Q_{\mathbf{h}_{kl}+\mathbf{V}_k}) \cdot (\mathbf{h}_{(k+1)l} - \mathbf{h}_{kl})$ are the same. Therefore, since for any quantizer $Q \in \mathcal{Q}$, $\mathbf{d}(Q) \cdot (\mathbf{h}_{(k+1)l} - \mathbf{h}_{kl}) \leq |\mathbf{h}_{(k+1)l} - \mathbf{h}_{kl}|_1 = l$, if the two cubes overlap on a fraction δ of their volume, then

$$\mathbb{E}\left[\mathbf{d}(Q_{\mathbf{h}_{kl}+\mathbf{V}_k}) \cdot (\mathbf{h}_{(k+1)l} - \mathbf{h}_{kl})\right] - \mathbb{E}\left[\mathbf{d}(Q_{\mathbf{h}_{(k+1)l}+\mathbf{V}_k}) \cdot (\mathbf{h}_{(k+1)l} - \mathbf{h}_{kl})\right] \leq (1-\delta)l. \tag{12}$$

Clearly, (12) holds for $k = 0,\ldots,n/l-1$. It is easy to see that for any $\mathbf{a} \in \mathbb{R}^K$, the cubes $[0,1/\epsilon]^K$ and $\mathbf{a} + [0,1/\epsilon]^K$ overlap in at least a $(1 - \epsilon|\mathbf{a}|_1)$ fraction of their volume if $\epsilon|\mathbf{a}|_\infty \leq 1$. Therefore, $\delta \geq 1 - \epsilon|\mathbf{h}_{(k+1)l} - \mathbf{h}_{kl}|_1 = 1 - \epsilon l$; hence summing (12) for all k, by (11) we obtain (8).

Combining (6), (7) and (8) we have

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

where we used the fact that $\log\binom{K}{M}/\log M + 1 \leq M\log K/\log M$. Letting $\epsilon = \sqrt{\frac{K}{ln}}$, $K = c_1 n^{1/4}$, and $l = c_2 n^{1/4}$ for some constants $c_1, c_2 > 0$ satisfying $K > M$ and $l > \log\binom{K}{M}/\log M$ gives (3).

Finally, it is easy to see that in case l does not divide n, the distortion on the last, truncated block can be accounted in the bound by slightly increasing the constant C.

Complexity analysis: By the scalar quantizer design algorithm of Wu and Zhang [9], for any nonnegative vector $\mathbf{a} \in \mathbb{R}^K$, the mean-square optimal M-level scalar quantizer $Q_{\mathbf{a}}$ can be found in $O(MK)$ time, thus the computational complexity of the algorithm is $O(MKn/l)+O(n) = O(Mn)$. Since the design of each quantizer requires $O(MK)$ storage capacity, the storage requirement of the algorithm is $O(Mn^{1/4})$. □

Note that when the quantizer Q_k is drawn, we implicitly assumed that we are able to perform $O(Mn^{1/4})$ operations in 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 $O(n^{1/4})$, 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.

References

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

[2] V. Vovk, "Aggregating strategies," in *Proceedings of the Third Annual Workshop on Computational Learning Theory*, (New York), pp. 372–383, Association of Computing Machinery, 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 Trans. Inform. Theory*, vol. 48, pp. 721–733, Mar. 2002.

[6] A. György, T. Linder, and G. Lugosi, "Efficient adaptive algorithms and minimax bounds for zero-delay lossy source coding," *submitted to IEEE Transactions on Signal Processing*, 2003. available at `www.szit.bme.hu/~gya/publications/GyLiLu03.ps`.

[7] 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`.

[8] 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.

[9] X. Wu and K. Zhang, "Quantizer monotonicities and globally optimal scalar quantizer design," *IEEE Trans. Inform. Theory*, vol. 39, pp. 1049–1053, 1993.