Near-Optimal Rates for Limited-Delay Universal Lossy Source Coding

András György[1]
[1]Machine Learning Research Group
Computer and Automation Research Institute
Hungarian Academy of Sciences
Budapest, Hungary
(email: gya@szit.bme.hu)

Gergely Neu[1,2]
[2]Department of Computer Science and
Information Theory
Budapest University of Technology and Economics
Budapest, Hungary
(email: neu.gergely@gmail.com)

***Abstract*—We consider the problem of limited-delay lossy coding of individual sequences. Here the goal is to design (fixed-rate) compression schemes to minimize the normalized expected distortion redundancy relative to a reference class of coding schemes, measured as the difference between the average distortion of the algorithm and that of the best coding scheme in the reference class. In compressing a sequence of length T, the best schemes available in the literature achieve an $O(T^{-1/3})$ normalized distortion redundancy relative to finite reference classes of limited delay and limited memory. It has also been shown that the distortion redundancy is at least of order $1/\sqrt{T}$ in certain cases. In this paper we narrow the gap between the upper and lower bounds, and give a compression scheme whose distortion redundancy is $O(\sqrt{\ln(T)/T})$, only a logarithmic factor larger than the lower bound. The method is based on the recently introduced Shrinking Dartboard prediction algorithm, a variant of the exponentially weighted average prediction. Our method is also applied to the problem of zero-delay scalar quantization, where $O(\ln(T)/\sqrt{T})$ distortion redundancy is achieved relative to the (infinite) class of scalar quantizers of a given rate, almost achieving the known lower bound of order $1/\sqrt{T}$.**

I. Introduction

In this paper we consider the problem of fixed-rate sequential lossy source coding of individual sequences with limited delay. Here a source sequence $x_1, x_2, \ldots$ taking values from the source alphabet $\mathcal{X}$ has to be transformed into a sequence $y_1, y_2, \ldots$ of channel symbols taking values in the finite channel alphabet $\{1, \ldots, M\}$, and these channel symbols are then used to produce the reproduction sequence $\hat{x}_1, \hat{x}_2, \ldots$. The rate of the scheme is defined as $\ln M$ nats (where $\ln$ denotes the natural logarithm), and the scheme is said to have δ_1 encoding and δ_2 decoding delay if, for any $t = 1, 2, \ldots$, the channel symbol y_t depends on $x^{t+\delta_1} = (x_1, x_2, \ldots, x_{t+\delta_1})$ and $\hat{x}_t$ depends on $y^{t+\delta_2} = (y_1, \ldots, y_{t+\delta_2})$. The goal of the coding scheme is to minimize the distortion between the source sequence and the reproduction sequence. In this work we concentrate on the individual sequence setting and aim to find methods that work uniformly well with respect to a reference coder class on every individual (deterministic) sequence. Thus, no probabilistic assumption is made on the source sequence, and the performance of a scheme is measured by the distortion redundancy defined as the maximal difference, over all source sequences of a given length, between the normalized distortion of the given coding scheme and that of the best reference coding scheme matched to the underlying source sequence.

The study of limited-delay (zero-delay) lossy source coding in the individual sequence setting was initiated by Linder and Lugosi [1], who showed the existence of randomized coding schemes that perform, on any bounded source sequence, essentially as well as the best scalar quantizer matched to the underlying sequence. More precisely, it was shown that the normalized squared error distortion of their scheme on any source sequence x^T of length T is at most $O(T^{-1/5} \ln T)$ larger than the normalized distortion of the best scalar quantizer matched to the source sequence in hindsight. The method of [1] is based on the exponentially weighted average (EWA) prediction method [2], [3], [4]: at each time instant a coding scheme (a scalar quantizer) is selected based on its "estimated" performance.

The coding scheme of [1] was improved and generalized by Weissman and Merhav [5]. They considered the more general case when the reference class $\mathcal{F}$ is a finite set of limited-delay and limited-memory coding schemes. To reduce the communication about the actual decoder to be used at the receiver, Weissman and Merhav introduced a coding scheme where the source sequence is split into blocks of equal length, and in each block a fixed encoder-decoder pair is used with its identity communicated at the beginning of each block. Similarly to [1], the code for each block is chosen using the EWA prediction method. The resulting scheme achieves an $O(T^{-1/3} \ln^{2/3} |\mathcal{F}|)$ distortion redundancy, or, in the case of the infinite class of scalar quantizers, the distortion redundancy

978-1-4577-0595-3/11/$26.00 ©2011 IEEE

becomes $O(T^{-1/3}\ln T)$.

The results of [5] have been extended in various ways, but all of these works are based on the block-coding procedure described above. A stochastic noisy channel setting is considered in [6], network quantization in [7], and a Wyner-Ziv setting (with side information at the decoder) is considered in [8]. Complexity issues have also been addressed in several papers: efficient solutions are given for the zero-delay case in [9], [10] (and, based on these papers, in [6], [7], [8] as well), and for the problem of competing with large classes of time-varying source codes in [7].

Since the above coding schemes are based on the block-coding scheme of [5], they cannot achieve better distortion redundancy than $O(T^{-1/3})$ up to some logarithmic factors. On the other hand, the distortion redundancy is known to be bounded from below by a constant multiple of $T^{-1/2}$ in the zero-delay case [9], leaving a gap between the best known upper and lower bounds. Thus, to improve upon the existing coding schemes, the communication overhead (describing the actually used coding schemes) between the encoder and the decoder has to be reduced, which is achievable by controlling the number of times the coding scheme changes in a better way then blockwise coding. This goal can be achieved by the recent Shrinking Dartboard (SD) algorithm of Geulen, Voecking, and Winkler [11], a modified version of the EWA prediction method that is designed to control the number of expert switches.

In this paper we construct a randomized coding strategy, which uses a slightly modified version of the SD algorithm as the prediction component, that achieves an $O(\sqrt{\ln T/T})$ average distortion redundancy with respect to a finite reference class of limited-delay and limited-memory source codes. The method can also be applied to compete with the (infinite) reference class of scalar quantizers, where it achieves an $O(\ln T/\sqrt{T})$ distortion redundancy. Note that these bounds are only logarithmic factors larger than the corresponding lower bound.

In Section II we revisit the SD algorithm of [11] with slight improvements relative to its original version. Our randomized coding strategy, based on the SD prediction method, is introduced and analyzed in Section IV. The strategy is applied to the problem of adaptive zero-delay lossy source coding in Section V. Conclusions are drawn and extensions are described in Section VI.

II. The Shrinking Dartboard algorithm revisited

In this section we define the problem of sequential decision making (prediction) with expert advice, and present the Shrinking Dartboard algorithm of [11]. Suppose we want to perform a sequence of decisions from a finite set $\mathcal{F}$ of size $N = |\mathcal{F}|$ without the knowledge of the future. At each time step $t = 1, 2, \ldots$ the decision maker chooses an action $\mathbf{i}_t \in \mathcal{F}$ and suffers a loss $d_{t,\mathbf{i}_t}$. At the end of each time step t the loss $d_{t,i} \in [0,1]$ for all $i \in \mathcal{F}$ is also revealed to the decision maker, whose goal is to minimize, for some $T > 0$, the average regret

$$\mathbf{R}_T = \max_{i\in\mathcal{F}} \frac{1}{T}\left(\sum_{t=1}^{T} d_{t,\mathbf{i}_t} - D_{T,i}\right)$$

with respect to the constant actions $i \in \mathcal{F}$, where $D_{T,i} = \sum_{t=1}^{T} d_{t,i}$ is the cumulative loss of action i up to time T. It is assumed that the decision maker has access to a sequence $\mathbf{U}_1, \mathbf{U}_2, \ldots$ of independent random variables with uniform distribution over the interval $[0,1]$, and its decision $\mathbf{i}_t$ depends only on $\mathbf{U}^t = (\mathbf{U}_1, \ldots, \mathbf{U}_t)$ and $d_{\tau,i}, \tau = 1, \ldots, t-1, i \in \mathcal{F}$. It is also assumed that the sequence $\{d_{t,i}\}$ is fixed in advance for all $i \in \mathcal{F}$ and $t = 1, 2, \ldots$, and, in particular, it is not affected by the (random) choices $\mathbf{i}_t$ of the decision maker.

A well-known solution to this problem (which is optimal under various conditions) is the EWA prediction method that, at time t, chooses action i with probability proportional to $e^{-\eta_t D_{t-1,i}}$ for some sequence of positive step size parameters $\{\eta_t\}_{t=1}^{T}$. It can be shown (using techniques developed in [12]) that if $\eta_{t+1} \le \eta_t$ for all t then the average expected regret of this algorithm satisfies $\mathbb{E}[\mathbf{R}_T] \le \sum_{t=1}^{T} \eta_t/(8T) + \ln N/(\eta_T T)$, hence setting the step sizes $\eta_t = 2\sqrt{\ln N/t}$ one obtains $\mathbb{E}[\mathbf{R}_T] \le \sqrt{\ln N/T}$ (here the expectation is taken with respect to the randomizing sequence $\mathbf{U}^T$).

While the EWA algorithm may choose a different action in each time step, in certain cases (e.g., in the coding scenario described in this paper) switching from one action to another has some extra cost, and so preference should be given to action sequences with fewer switches. The SD algorithm [11] addresses this problem and provides the same performance guarantee as EWA by controlling the number of switches between different actions, that is, the number of time instants when $\mathbf{i}_t \neq \mathbf{i}_{t-1}$. A modified version of this prediction method, called the modified SD (mSD) algorithm, is shown in Algorithm 1. The difference between the SD and the mSD algorithms is that mSD is horizon independent, which is achieved by introducing the constant c_t in the algorithm (setting $\eta_t \equiv \eta$ the mSD algorithm reduces to SD).

To see that the mSD algorithm is well-defined we have to show that $c_t \frac{w_{t,i}}{w_{t-1,i}} \le 1$ for all t and i. For $t = 1$, the statement follows from the definitions, since

Algorithm 1 The modified Shrinking Dartboard algorithm

1) Set $\eta_t > 0$ with $\eta_{t+1} \le \eta_t$ for all $t = 1, 2, \ldots$, $\eta_0 = \eta_1$, and $D_{0,i} = 0$ and $w_{0,i} = 1/N$ for all actions $i \in \mathcal{F}$.
2) **for** $t = 1, \ldots, T$ **do**
   a) Set $w_{t,i} = \frac{1}{N} e^{-\eta_t D_{t-1,i}}$ for all $i \in \mathcal{F}$.
   b) Set $p_{t,i} = \frac{w_{t,i}}{\sum_{j=1}^{N} w_{t,j}}$ for all $i \in \mathcal{F}$.
   c) Set $c_t = e^{(\eta_t - \eta_{t-1})(t-2)}$.
   d) With probability $c_t \frac{w_{t,\mathbf{i}_{t-1}}}{w_{t-1,\mathbf{i}_{t-1}}}$, set $\mathbf{i}_t = \mathbf{i}_{t-1}$ if $t \ge 2$, that is, do not change expert; otherwise choose $\mathbf{i}_t$ randomly according to the distribution $\{p_{t,1}, \ldots, p_{t,N}\}$.
   e) Observe the losses $d_{t,i}$ and set $D_{t,i} = D_{t-1,i} + d_{t,i}$ for all $i \in \mathcal{F}$.

   end for

$c_1 = 1$. For $t \ge 2$ it follows since

$$\begin{aligned}\frac{w_{t,i}}{w_{t-1,i}} &= \exp\left(\eta_{t-1} D_{t-2,i} - \eta_t D_{t-1,i}\right) \\ &\le \exp\left(\left(\eta_{t-1} - \eta_t\right) D_{t-2,i} - \eta_t d_{t-1,i}\right) \\ &\le \exp\left(\left(\eta_{t-1} - \eta_t\right)(t-2)\right) = 1/c_t.\end{aligned}$$

Note that the only difference between the mSD and the EWA prediction algorithms is the presence of the first random choice in step 2d of mSD: while the EWA algorithm chooses a new action in each time step t according to the distribution $\{p_{t,1}, \ldots, p_{t,N}\}$, the mSD algorithm sticks with the previously chosen action with some probability.

In the following, we state two results crucial for the analysis of the coding scheme that we will propose in the next section. The next lemma shows that the marginal distributions generated by the mSD and the EWA algorithms are the same. The lemma is obtained by a slight modification of the proof Lemma 1 in [11].

Lemma 1: Assume the mSD algorithm is run with $\eta_{t+1} \le \eta_t$ for all $t = 1, 2, \ldots, T$. Then the probability of selecting action i at time t satisfies $\mathbb{P}[\mathbf{i}_t = i] = p_{t,i}$ for all $t = 1, 2, \ldots$ and $i \in \mathcal{F}$.

As a consequence of this result, the expected regret of mSD matches that of EWA, so the performance bound of EWA, mentioned in the previous section, holds for the mSD algorithm as well [11, Lemma 2]).

Lemma 2: Assume $\eta_{t+1} \le \eta_t$ for all $t = 1, 2, \ldots, T$. Then the expected average regret of the mSD algorithm can be bounded as

$$\mathbb{E}[\mathbf{R}_T] \le \sum_{t=1}^{T} \frac{\eta_t}{8T} + \frac{\ln N}{T \eta_T}.$$

Setting $\eta_t = \sqrt{\ln N / T}$ optimally (as a function of the time horizon T), the bound becomes $\sqrt{\frac{\ln N}{2T}}$, while setting $\eta_t = 2\sqrt{\ln N / t}$ independent of T, we have $\mathbb{E}[\mathbf{R}_T] \le \sqrt{\ln N / T}$.

Let $\mathbf{S}_T = \{t : \mathbf{i}_t \ne \mathbf{i}_{t-1}, 2 \le t \le T\}$ denote the number of times the mSD algorithm switches between different actions. The next lemma, which is a slightly improved and generalized version of Lemma 2 from [11] gives an upper bound on $|\mathbf{S}_T|$.

Lemma 3: Let $\mathbf{S}_T$ denote the number of times the mSD algorithm switches between different actions. Then

$$\mathbb{E}[|\mathbf{S}_T|] \le \eta_T D^*_{T-1} + \ln N + \sum_{t=1}^{T-1} (\eta_t - \eta_T)$$

where $D^*_{T-1} = \min_{i \in \mathcal{F}} D_{T-1,i}$.

In particular, for $\eta_t = \sqrt{\ln N / T}$, we have $\mathbb{E}[|\mathbf{S}_T|] \le \sqrt{T \ln N} + \ln N$, while setting $\eta_t = 2\sqrt{\ln N / t}$, we obtain $\mathbb{E}[|\mathbf{S}_T|] \le 4\sqrt{T \ln N} + \ln N$.

III. Limited-delay limited-memory sequential source codes

A fixed-rate delay-δ (randomized) sequential source code of rate $\ln M$ is defined by an encoder-decoder pair connected via a discrete noiseless channel of capacity $\ln M$. Here δ is a nonnegative integer and $M \ge 2$ is a positive integer. The input to the encoder is a sequence $x_1, x_2, \ldots$ taking values in some source alphabet $\mathcal{X}$. At each time instant $t = 1, 2, \ldots$, the encoder observes x_t and a random number $\mathbf{U}_t$, where the randomizing sequence $\mathbf{U}_1, \mathbf{U}_2, \ldots$ is assumed to be independent with its elements uniformly distributed over the interval $[0, 1]$. At each time instant $t + \delta$, $t = 1, 2, \ldots$, based on the source sequence $x^{t+\delta} = (x_1, \ldots, x_{t+\delta})$ and the randomizing sequence $\mathbf{U}^t = (\mathbf{U}_1, \ldots, \mathbf{U}_t)$ received so far, the encoder produces a channel symbol $\mathbf{y}_t \in \{1, 2, \ldots, M\}$ which is then transmitted to the decoder. After receiving $\mathbf{y}_t$, the decoder outputs the reconstruction value $\hat{\mathbf{x}}_t \in \hat{\mathcal{X}}$ based on the channel symbols $\mathbf{y}^t = (\mathbf{y}_1, \ldots, \mathbf{y}_t)$ received so far, where $\hat{\mathcal{X}}$ is the reconstruction alphabet.

Formally, a code is given by a sequence of encoder-decoder functions $(f, g) = \{f_t, g_t\}_{t=1}^{\infty}$, where

$$f_t : \mathcal{X}^{t+\delta} \times [0, 1]^t \to \{1, 2, \ldots, M\}$$

and

$$g_t : \{1, 2, \ldots, M\}^t \to \hat{\mathcal{X}}$$

so that $\mathbf{y}_t = f_t(x^{t+\delta}, \mathbf{U}^t)$ and $\hat{\mathbf{x}}_t = g_t(\mathbf{y}^t)$, $t = 1, 2, \ldots$. Note that the total delay of the encoding and decoding process is δ.[1]

[1] Although we require the decoder to operate with zero delay, this requirement introduces no loss in generality, as any finite-delay coding system with δ_1 encoding and δ_2 decoding delay (described in Section I) can be represented equivalently in this way with $\delta_1 + \delta_2$ encoding and zero decoding delay [5].

Now let $\mathcal{F}$ be a finite set of reference codes with $|\mathcal{F}| = N$. The *cumulative distortion* of the sequential scheme after reproducing the first T symbols is given by

$$\hat{\mathbf{D}}_T(x^{T+\delta}) = \sum_{t=1}^{T} d(x_t, \hat{\mathbf{x}}_t),$$

where $d : \mathcal{X} \times \hat{\mathcal{X}} \to [0,1]$ is some distortion measure,[2] while the minimal cumulative distortion achievable by codes from $\mathcal{F}$ is

$$D^*_{\mathcal{F}}(x^{T+\delta}) = \min_{(f,g)\in\mathcal{F}} \sum_{t=1}^{T} d\left(x_t,\, g_t(y^t)\right)$$

where the sequence y^T is generated sequentially by (f,g), that is, $y_t = f_t\left(x^{t+\delta}, \mathbf{U}^t\right)$. Of course, in general it is impossible to come up with a coding scheme that attains this distortion without knowing the whole input sequence beforehand. Thus, our goal is to construct a coding scheme that asymptotically achieves the performance of the above encoder-decoder pair. Formally this means that we want to obtain a randomized coding scheme that minimizes the worst-case *expected normalized distortion redundancy*

$$\hat{R}_T = \max_{x^T\in\mathcal{X}^T} \frac{1}{T}\left\{\mathbb{E}\left[\hat{\mathbf{D}}_T\left(x^{T+\delta}\right)\right] - D^*_{\mathcal{F}}\left(x^{T+\delta}\right)\right\},$$

where the expectation is taken with respect to the randomizing sequence $\mathbf{U}^T$ of our coding scheme.

The decoder $\{g_t\}$ is said to be of memory $s \geq 0$ if $g_t(\hat{y}^t) = g_t(\tilde{y}^t)$ for all t and $\hat{y}^t, \tilde{y}^t \in \{0,\ldots,M\}^t$ such that $\hat{y}^t_{t-s} = \tilde{y}^t_{t-s}$, where $\hat{y}^t_{t-s} = (\hat{y}_{t-s}, \hat{y}_{t-s+1}, \ldots, \hat{y}_t)$ and $\tilde{y}^t_{t-s} = (\tilde{y}_{t-s}, \tilde{y}_{t-s+1}, \ldots, \tilde{y}_t)$. Let $\mathcal{F}^\delta$ denote the collection of all (randomized) delay-δ sequential source codes of rate $\ln M$, and let $\mathcal{F}^\delta_s$ denote the class of codes in $\mathcal{F}^\delta$ with memory s.

IV. The Algorithm

Next we describe a coding scheme, based on the mSD prediction algorithm, that adaptively creates blocks of variable length such that on the average $O(\sqrt{T})$ blocks are created, and so the overhead used to transmit code descriptions scales with $\sqrt{T}$ instead of $T^{2/3}$ in [5]. Assuming a finite reference class $\mathcal{F} \in \mathcal{F}^\delta_s$, our coding scheme works as follows.

At each time instant t the mSD algorithm selects one code $(\mathbf{f}^{(t)}, \mathbf{g}^{(t)})$ from the finite reference class $\mathcal{F}$; the loss in the mSD algorithm associated with $(f,g) \in \mathcal{F}$ is defined by

$$d_{t,(f,g)}(x^{t+\delta}) = d\left(x_t, g_t\left(y'^t\right)\right) \tag{1}$$

where $y'_1, y'_2, \ldots, y'_t$ is the sequence obtained by using the coding scheme (f,g) to encode x^t, that is, $y'_t = f_t(x^{t+\delta}, \mathbf{U}^t)$ (note that $d_{t,(f,g)}$ can be computed at the encoder at time $t+\delta$). The mSD algorithm splits the time into blocks $[1,t_1], [t_1+1,t_2], [t_2+1,t_3], \ldots$ in a natural way such that the decoder of the reference code chosen by the algorithm is constant over each block, that is, $\mathbf{g}^{(t_i+1)} = \mathbf{g}^{(t_i+2)} = \cdots = \mathbf{g}^{(t_{i+1})}$ and $\mathbf{g}^{(t_i)} \neq \mathbf{g}^{(t_i+1)}$ for all i (here we used the convention $t_0 = 0$). Since the beginning of a new block can only be noticed at the encoder, this event has to be communicated to the decoder. In order to do so, we select randomly a *new-block* signal $\mathbf{v}$ of length A (that is, $\mathbf{v} \in \{1,\ldots,M\}^A$), and $\mathbf{v}$ is transmitted over the channel in the first A time steps of each block. In the next B time steps of the block the identity of the decoder chosen by the mSD algorithm is communicated, where $B = \left\lceil \frac{\ln|\{g:(f,g)\in\mathcal{F}\}|}{\ln M} \right\rceil$ is the number of channel symbols required to describe uniquely all possible decoder functions. In the remainder of the block the selected encoder (or, possibly, more encoders) is used to encode the source symbols.

On the other hand, whenever the decoder observes $\mathbf{v}$ in the received channel symbol sequence $\mathbf{y}^t$, it starts a new block. In this block the decoder first receives the index of the reference decoder to be used in the block, and the received reference decoder is used in the remainder of the block to generate the reproduction symbols. One slight problem here is that the new-block signal may be obtained by encoding the input sequence; in this case, to synchronize with the decoder, a new block is started at the encoder. We can keep the loss introduced by these unnecessary new blocks low by a careful choice of the new-block signal. Clearly, if $\mathbf{v}$ is selected uniformly at random from $\{1,2,\ldots,M\}^A$ then for any fixed string $u \in \{1,2,\ldots,M\}^A$, $\mathbb{P}[\mathbf{v}=u] = 1/M^A$. Thus, setting $A = O(\ln T)$ makes $\mathbb{P}[\mathbf{v}=u] = O(1/T)$, and so the expected number of unnecessary new blocks is at most a constant in T time steps.

The next result shows that the normalized distortion redundancy of the proposed scheme is $O(\sqrt{\ln(T)/T})$.

Theorem 1: For any finite reference class $\mathcal{F} \subset \mathcal{F}^\delta_s$ and time horizon $T > 0$ there exists an adaptive sequential source coding scheme with expected normalized distortion redundancy bounded as

$$\hat{R}_T \leq 2\sqrt{\frac{\ln|\mathcal{F}|}{T}\left(\frac{17}{8} + \frac{\ln(T|\mathcal{F}|)}{\ln M} + s\right)} + O\left(\frac{\ln T}{T}\right).$$

In the above, the parameters $A = \left\lceil \frac{\ln T}{\ln M} \right\rceil$ and $\eta_t = \eta = O(1/\sqrt{T\ln T})$ are set as a function of the time horizon T. The proposed algorithm can be modified to be strongly sequential in the sense that it becomes horizon independent. The main difference is that the new-block signal will be time-variant: at time instants

[2] All results may be extended trivially for arbitrary bounded distortion measures

M^{k-1} the kth symbol $\mathbf{v}_k$ of $\mathbf{v}$ is transmitted, and at each time instant t the so far received new-block signal $\mathbf{v}^{A_t}$ of length $A_t = \left\lceil \frac{\ln t}{\ln M} \right\rceil$ is used. Setting $\eta_t = O(1/\sqrt{t \ln t})$, it can be shown that the modified algorithm has only a constant time larger regret than the original, horizon-dependent one.

V. Sequential zero-delay lossy source coding

An important and widely studied special case of the source coding problem considered is the case of on-line scalar quantization, that is, the problem of zero-delay lossy source coding with memoryless encoders and decoders [1], [5], [9], [10]. Here we assume for simplicity $\mathcal{X} = [0,1]$ and $d(x,\hat{x}) = (x-\hat{x})^2$. An M-level scalar quantizer Q (defined on $[0,1]$) is a measurable mapping $[0,1] \to C$, where the *codebook* C is a finite subset of $[0,1]$ with cardinality $|C| = M$. The elements of C are called the *code points*. The instantaneous squared distortion of Q for input x is $(x - Q(x))^2$. Without loss of generality we will only consider nearest neighbor quantizers Q satisfying $(x-Q(x))^2 = \min_{\hat{x}\in C}(x-\hat{x})^2$.

Let $\mathcal{Q}$ denote the collection of all M-level nearest neighbor quantizers. In this section our goal is to design a sequential coding scheme that asymptotically achieves the performance of the best scalar quantizer (from $\mathcal{Q}$) for all source sequences x^T. Note that the expected normalized distortion redundancy in this special case is defined as

$$\max_{x^T\in[0,1]^T} \frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T}(x_t-\hat{\mathbf{x}}_t)^2\right] - \min_{Q\in\mathcal{Q}} \frac{1}{T}\sum_{t=1}^{T}(x_t-Q(x_t))^2.$$

To be able to apply the results of the previous section, we approximate the infinite class $\mathcal{Q}$ with $\mathcal{Q}_K \subset \mathcal{Q}$, the set of M-level nearest neighbor scalar quantizers whose code points all belong to the set $\left\{\frac{1}{2K}, \frac{3}{2K}, \ldots, \frac{2K-1}{2K}\right\}$. It is shown in [9] that the distortion redundancy of any sequential coding scheme relative to $\mathcal{Q}$ is at least on the order of $T^{-1/2}$. The next theorem shows that the slightly larger $O(T^{-1/2}\ln T)$ normalized distortion redundancy is achievable.

Theorem 2: There exists an adaptive sequential coding scheme whose normalized expected distortion redundancy relative to the reference class $\mathcal{Q}$ satisfies, for any $T \geq 2$,

$$\hat{R}_T \leq \sqrt{\frac{2M\ln T}{T}\left(\frac{25}{8} + \frac{(M+2)\ln T}{2\ln M} + s\right)} + O\left(\frac{\ln^2 T}{T}\right)$$

and the coding scheme can be implemented with $O(T^2)$ time and $O(T)$ space complexity.

VI. Conclusion

We provided a sequential lossy source coding scheme that achieves an $O(\sqrt{\ln(T)/T})$ normalized distortion redundancy relative to any finite reference class of limited-delay limited-memory codes, improving the earlier results of $O(T^{-1/3})$. Applied to the case when the reference class is the (infinite) set of scalar quantizers, we showed that the algorithm achieves $O(\ln(T)/\sqrt{T})$ normalized distortion redundancy, which is almost optimal in view that the normalized distortion redundancy is known to be at least of order $1/\sqrt{T}$. The results can also be extended to the noisy-channel and the Wyner-Ziv settings [13].

Acknowledgment

This research was supported in part by the Hungarian Scientific Research Fund and the Hungarian National Office for Research and Technology (OTKA-NKTH CNK 77782), and the PASCAL2 Network of Excellence under EC grant no. 216886.

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*, (Rochester, NY), pp. 372–383, Morgan Kaufmann, Aug. 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] S. Matloub and T. Weissman, "Universal zero delay joint source-channel coding," *IEEE Transactions on Information Theory*, vol. 52, pp. 5240–5250, 2006.

[7] A. György, T. Linder, and G. Lugosi, "Tracking the best quantizer," *IEEE Transactions on Information Theory*, vol. 54, pp. 1604–1625, Apr. 2008.

[8] A. Reani and N. Merhav, "Efficient on-line schemes for encoding individual sequences with side information at the decoder," in *Proc. IEEE International Symposium on Information Theory (ISIT 2009)*, pp. 1025–1029, June 28–July 3 2009.

[9] A. György, T. Linder, and G. Lugosi, "Efficient algorithms and minimax bounds for zero-delay lossy source coding," *IEEE Transactions on Signal Processing*, vol. 52, pp. 2337–2347, Aug. 2004.

[10] A. György, T. Linder, and G. Lugosi, "A "follow the perturbed leader"-type algorithm for zero-delay quantization of individual sequences," in *Proc. Data Compression Conference*, (Snowbird, UT, USA), pp. 342–351, Mar. 2004.

[11] S. Geulen, B. Voecking, and M. Winkler, "Regret minimization for online buffering problems using the weighted majority algorithm," in *Proceedings of the 23rd Annual Conference on Learning Theory (COLT 2010)*, pp. 132–143, 2010.

[12] L. Györfi and G. Ottucsák, "Sequential prediction of unbounded stationary time series," *IEEE Transactions on Information Theory*, vol. 53, pp. 866–1872, May 2007.

[13] A. György and G. Neu, "Near-optimal rates for limited-delay universal lossy source coding." in preparation.