Proceedings of the
44th IEEE Conference on Decision and Control, and
the European Control Conference 2005
Seville, Spain, December 12-15, 2005

Limited-Delay Coding of Individual Sequences with Piecewise Different Behavior

(Invited Paper)

András György
Informatics Laboratory
Computer and Automation Research Institute
Hungarian Academy of Sciences
Lágymányosi u. 11, 1111 Budapest, Hungary
E-mail: gya@szit.bme.hu

Tamás Linder
Department of Mathematics and Statistics
Queen's University
Kingston, Ontario
Canada, K7L 3N6
E-mail: linder@mast.queensu.ca

Gábor Lugosi
Department of Economics
Pompeu Fabra University
Ramon Trias Fargas 25-27
08005 Barcelona, Spain
E-mail: lugosi@upf.es

***Abstract*— Limited delay lossy coding schemes are considered for individual sequences. We address the problem of tracking the best code (from a given reference class) which is adaptively matched to the source sequence with piecewise different behavior. A general randomized algorithm is presented which can perform, on any source sequence, asymptotically as well as the best combined coding scheme matched to the sequence that is allowed to change the employed code (from a finite reference class of limited delay codes) several times during the coding procedure. In particular, a low complexity algorithm is presented for the special case where the reference class is the set of scalar quantizers.**

I. Introduction

In this paper we consider limited-delay lossy coding schemes for individual sequences. Such schemes are of obvious interest in many applications; in particular, limited-delay lossy compression schemes are of importance in real-time feedback control problems, where the feedback information has to be transmitted under strict delay requirements over a low-rate channel. More generally, the analysis and construction of limited (or zero) delay source codes is of interest in distributed control problems under communication constraint, a recent challenge in control theory that has been receiving increasing attention; see, e.g., [1] and the references therein.

Our goal is to provide a universal coding method which can dynamically adapt to the changes of the feedback data, providing the controller with as much information as possible. We 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 when very little is known about the source to be encoded.

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 $b_1, b_2, \ldots$ 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 delay δ if the reproduction symbol $\hat{x}_n$ can be decoded at most δ time instants after x_n was available at the encoder. A general model for this situation is that each channel symbol b_n depends only on the source symbols $x_1, \ldots, x_{n+\delta}$, and the reproduction $\hat{x}_n$ for the source symbol x_n depends only on the channel symbols $b_1, \ldots, b_n$. Thus, the encoder produces b_n as soon as $x_{n+\delta}$ is available, and the decoder can produce $\hat{x}_n$ when b_n is received. (Note that this setup can also be used to model the situation where there is limited delay both at the encoder and at the decoder.)

The performance of a scheme is measured with respect to a reference class of coding schemes, and the goal is to perform, on any source sequence, asymptotically as well as the best scheme in the reference class. Thus, the performance is measured by the distortion redundancy defined as the maximum of the difference of the normalized cumulative distortion of the applied scheme and the normalized cumulative distortion of the best scheme in the reference class over all source sequences of length n.

Limited delay lossy sequential coding of individual sequences was studied first in [2] for the special case of zero-delay coding with the reference class of scalar quantizers. The coding scheme of [2] is based on a generalization of exponentially weighted average prediction of individual sequences (see, e.g., Littlestone and Warmuth [3]), and achieves a distortion redundancy of order $n^{-1/5} \log n$, with common randomization at the encoder and the decoder. This result was improved and generalized by Weissman and Merhav [4]. 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. The resulting scheme has distortion redundancy $O(n^{-1/3} \log^{2/3} N)$, where N is the size of the reference class. To our knowledge, this is the best known redundancy bound for this problem. In the special case where the reference class is the (infinite) set of scalar quantizers, an $O(n^{-1/3} \log n)$ distortion redundancy

This research was supported in part by the Natural Sciences and Engineering Research Council (NSERC) of Canada, the János Bolyai Research Scholarship of the Hungarian Academy of Sciences, Spanish Ministry of Science and Technology and FEDER, grant BMF2003-03324, and by the PASCAL Network of Excellence under EC grant no. 506778.

0-7803-9568-9/05/$20.00 ©2005 IEEE

can be achieved by approximating the reference class by an appropriately chosen finite set of quantizers.

Although both schemes have the attractive property of performing uniformly well on individual sequences, they are computationally inefficient: in their straightforward implementation, they require a computational time of order $n^{c\,2^R}$, where R is the rate of the coding scheme, and $c = 1/5$ for the scheme in [2] and $c = 1/3$ for the scheme in [4]. This prohibitive complexity comes from the fact that, in order to approximate the performance of the best scalar quantizer, 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.

For more general finite reference classes, the method of [4] has to maintain a weight for each of the N reference codes. This results in a computational complexity of order nN, which allows the use of only small reference classes. When the reference class is an infinite set of codes, the method is applied to a finite approximation of the reference class, which, as we have seen above, results in a prohibitively large N even when we want to compete with scalar quantizers.

Recently, using the special structure of scalar quantizers, we provided an efficient implementation of the algorithm of [4] (for the reference class of scalar quantizers) with encoding complexity $O(n^{4/3})$ and distortion redundancy $O(n^{-1/3}\log n)$ [5]. The complexity can be made linear in the sequence length at the price of increasing the distortion redundancy to $O(n^{-1/4}\sqrt{\log n})$. A different method was introduced in [6], where, based on the "follow-the-perturbed-leader" prediction method of Hannan [7], a conceptually simpler algorithm was provided with linear encoding complexity and slightly increased distortion redundancy $O(n^{-1/4}\log n)$.

It was identified as an interesting open problem in [4] to find an algorithm of low complexity that is able to approximate the performance of the best scheme from a larger reference class. In this paper we consider a more general reference class in which each scheme partitions the input sequence into contiguous segments and may employ a different delay-δ code in each segment from a finite base reference class $\mathcal{F}$. (In the probabilistic setting, Shamir and Merhav [8] and Shamir and Costello [9] considered the similar problem of low-complexity sequential lossless coding for piecewise-stationary memoryless sources.) If a combined scheme can change the applied code m times, then the number of such schemes is $\sum_{j=0}^{m}\binom{n}{j}|\mathcal{F}|(|\mathcal{F}|-1)^j$. As in the algorithm of Weissman and Merhav [4] one has to maintain a weight for each reference code, the implementation of that method is infeasible even for a very small $\mathcal{F}$. In this paper we overcome this problem by utilizing the special structure of the reference class via the combination of the "tracking-the-best-expert" prediction method of Herbster and Warmuth [10] with the algorithm of [4]. As it will be shown later, the resulting algorithm requires only the maintenance of $|\mathcal{F}|$ weights. However, for rich base reference classes, this method may still prove to be too complex. In the special case when $\mathcal{F}$ is the set of scalar quantizers, we can combine our recent efficient implementation of the "tracking-the-best-expert" prediction method [11] with the efficient implementation in [5] to obtain low-complexity coding schemes (these methods were presented in a less general context in [15]).

The rest of the paper is organized as follows. First, in Section II, the "tracking-the-best-expert" prediction method is revisited and slightly modified to suit our intended application. Section III formalizes the problem of sequential lossy source coding with delay constraints and introduces a class of combined reference coding schemes where each scheme is allowed to change the employed code (from a given base reference class) a given number of times during the coding procedure. In Section IV we use the prediction framework of Section II to construct a limited-delay online coding algorithm. We analyze the complexity and performance of the algorithm, and show that it has relatively modest complexity and performs nearly as well as the best of the reference schemes matched to the entire input sequence. Finally, in Section V, for the special case when the reference coding schemes are combinations of scalar quantizers, a low-complexity zero-delay scheme is provided which performs essentially as well as the best scalar quantization scheme which can change the employed quantizer from time to time (the complexity of this scheme can be made linear in the sequence length).

II. A variation of the tracking-the-best-expert prediction method

In this section we consider the following sequential decision problem. Suppose we want to perform a sequence of decisions from a set $\mathcal{D}$ without the knowledge of the future. The state of the system is described by a sequence $y_1, y_2, \ldots$ taking values in some set $\mathcal{Y}$. We assume that the predictor has access to a sequence $U_1, U_2, \ldots$ of independent random variables distributed uniformly over the interval $[0,1]$. At each time instant $t = 1, 2, \ldots$, the predictor observes U_t, and based on U_t and the past states $y^{t-1} = (y_1, \ldots, y_{t-1})$ makes a decision $\hat{y}_t \in \mathcal{D}$. Then the predictor can observe the next state y_t, and suffers a loss $\ell(y_t, \hat{y}_t)$ for some bounded loss function $\ell : \mathcal{Y} \times \mathcal{D} \to [0, B]$ ($B > 0$).

The predictor is supported by N experts: At each time instant t expert i forms its decision $\hat{y}_{i,t} \in \mathcal{D}$, and the predictor can observe the decisions of all experts before producing its own decision.

Formally, at each time instant $t = 1, 2, \ldots$, first the decision $\hat{y}_{i,t} \in \mathcal{D}, i = 1, \ldots, N$, of each expert is revealed, then the predictor observes the random variable U_t and makes a decision $\hat{y}_t \in \mathcal{D}$, and finally the state of the system is revealed and the predictor suffers loss $\ell(y_t, \hat{y}_t)$.

The *expected cumulative loss* of the sequential scheme at time T is given by

$$\mathbb{E}L_T(f) = \mathbb{E}\left[\sum_{t=1}^{T}\ell(y_t, \hat{y}_t)\right]$$

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

The goal of the predictor is to achieve a cumulative loss (almost) as small as the best "tracking" of the N (base) experts. More precisely, to describe the loss the predictor is compared to, consider the following "m-partition" decision making scheme: the sequence of examples is partitioned into $m+1$ contiguous segments, and on each segment the scheme assigns exactly one of the N experts. Formally, an m-partition $\mathcal{P}(T,m,\mathbf{t},\mathbf{e})$ of the first T samples is given by an m-tuple $\mathbf{t}=(t_1,\ldots,t_m)$ such that $t_0=1<t_1<\cdots<t_m<T+1=t_{m+1}$, and an $(m+1)$-vector $\mathbf{e}=(e_0,\ldots,e_m)$ where $e_i\in\{1,\ldots,N\}$. At each time instant t, $t_i\le t<t_{i+1}$ expert e_i is used to perform the decision $\hat{y}_t$. The cumulative loss of a partition $\mathcal{P}_{T,m,\mathbf{t},\mathbf{e}}$ is

$$L(\mathcal{P}(T,m,\mathbf{t},\mathbf{e}))=\sum_{i=0}^{m}\sum_{t=t_i}^{t_{i+1}-1}\ell(y_t,\hat{y}_{i,t})=\sum_{i=0}^{m}L([t_i,t_{i+1}),e_i)$$

where, for any time interval I, $L(I,e_i)=\sum_{t\in I}\ell(y_t,\hat{y}_{i,t})$ denotes the cumulative loss of expert i in I.

The goal of the predictor is to perform as well as the best partition, that is, to keep the normalized expected redundancy

$$\frac{1}{T}\left(\mathbb{E}[L_T(f)]-\min_{\mathbf{t},\mathbf{e}}L(\mathcal{P}(T,m,\mathbf{t},\mathbf{e}))\right)$$

as small as possible for all possible outcome sequences.

As the number of "m-partition" schemes is $\sum_{k=0}^{m}\binom{T}{k}N(N-1)^k$, it is computationally infeasible to apply the exponentially weighted average prediction method to the above problem, as there it is required to store and update a weight for each of the "m-partition" schemes. However, exploiting certain structural properties of the problem, Herbster and Warmuth [10] provided an efficient solution which only requires to store the weights of the (base) experts.

Here we present a slightly modified version of the "fixed-share" share update algorithm of [10]. While this modification also appeared in [12], the performance bounds provided there are insufficient for our purposes.

Algorithm 1: Fix the positive numbers η and $0<\alpha<1$, and initialize weights $w^s_{1,i}=1/N$ for $i=1,\ldots,N$. At time instants $t=1,2,\ldots,T$ let $v_t^{(i)}=w^s_{t,i}/W_t$ where $W_t=\sum_{i=1}^{N}w^s_{t,i}$, and decide $\hat{y}_t$ randomly according to the distribution

$$\mathbb{P}\{\hat{y}_t=\hat{y}_{i,t}\}=v_t^{(i)}. \tag{1}$$

After receiving y_t, for all $i=1,\ldots,N$ let

$$w^m_{t,i}=w^s_{t,i}e^{-\eta\ell(y_t,\hat{y}_{i,t})}$$

and

$$w^s_{t+1,i}=\frac{\alpha W_{t+1}}{N}+(1-\alpha)w^m_{t,i}$$

where $W_{t+1}=\sum_{i=1}^{N}w^m_{t,i}$.

Note that $\sum_{i=1}^{N}w^s_{t+1,i}=\sum_{i=1}^{N}w^m_{t,i}=W_{t+1}$, thus W_{t+1} is uniquely defined.

Next we present a bound on the loss of the algorithm. The proof is almost identical to that in [10] with some necessary modifications introduced by the random choice (1), which can be treated using standard methods (see, e.g., [13]).

Theorem 1: For all positive integers $m<T$, real numbers $0<\alpha<1$ and $\eta>0$, and for any sequence $y_1,\ldots,y_T$ taking values in $[0,B]$ with some $B>0$, the expected redundancy $\mathbb{E}L_T(f)$ of Algorithm 1 can be bounded as

$$\begin{aligned}&\mathbb{E}[L_T(f)]-\min_{\mathbf{t},\mathbf{e}}L(\mathcal{P}(T,m,\mathbf{t},\mathbf{e}))\\ &\quad\le\ \frac{1}{\eta}\ln\left(\frac{N^{m+1}}{\alpha^m(1-\alpha)^{T-m-1}}\right)+\frac{T\eta B^2}{8}\,.\end{aligned}$$

In particular, if $\alpha=\frac{m}{T-1}$ and η is chosen to minimize the above bound, we have

$$\begin{aligned}&\mathbb{E}[L_T(f)]-\min_{\mathbf{t},\mathbf{e}}L(\mathcal{P}(T,m,\mathbf{t},\mathbf{e}))\\ &\quad\le\ T^{1/2}\frac{B}{\sqrt{2}}\sqrt{(m+1)\ln N+m\ln\frac{T-1}{m}+m}.\end{aligned} \tag{2}$$

III. Finite-delay finite-memory sequential source codes

A fixed-rate delay-δ sequential source code of rate $R=\log M$ is defined by an encoder-decoder pair connected via a discrete noiseless channel of capacity R. (Here δ is a nonnegative integer, M is a positive integer and $\log$ denotes base-2 logarithm.) 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 some source alphabet $\mathcal{X}$. At each time instant $i=1,2,\ldots$, the encoder observes x_i. At each time instant $i+\delta$, $i=1,2,\ldots$, the encoder also observes the random number U_i, and based on the source sequence $x^{i+\delta}=(x_1,\ldots,x_{i+\delta})$ and the randomizing sequence $U^i=(U_1,\ldots,U_i)$ received so far, the encoder produces a channel symbol $b_i\in\{1,2,\ldots,M\}$ which is then transmitted to the decoder. After receiving b_i, the decoder outputs the reconstruction value $\hat{x}_i\in\hat{\mathcal{X}}$ based on the channel symbols $b^i=(b_1,\ldots,b_i)$ received so far, where $\hat{\mathcal{X}}$ is the reconstruction alphabet.

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

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

and

$$g_i:\{1,2,\ldots,M\}^i\to\hat{\mathcal{X}}$$

so that $b_i=f_i(x^{i+\delta},U^i)$ and $\hat{x}_i=g_i(b^i)$, $i=1,2,\ldots$. Note that the total delay of the encoding and decoding process is δ. 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 can be equivalently represented in this way with $\delta_1+\delta_2$ encoding and zero decoding delay.

The *expected normalized cumulative distortion* of the sequential scheme after reproducing the first n symbols is

given by

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

where $d : \mathcal{X}\times\hat{\mathcal{X}} \to [0,1]$ is some distortion measure, and the expectation is taken with respect to the randomizing sequence $U^n = (U_1,\ldots,U_n)$. (All results may be extended trivially for arbitrary bounded distortion measures.)

The decoder $\{g_i\}$ is said to be of finite memory $s \geq 0$ if $g_i(b^i) = g_i(\hat{b}^i)$ for all i and $b^i, \hat{b}^i \in \{0,\ldots,M\}^i$ such that $b_{i-s}^i = \hat{b}_{i-s}^i$, where $b_{i-s}^i = (b_{i-s}, b_{i-s+1},\ldots,b_i)$ and $\hat{b}_{i-s}^i = (\hat{b}_{i-s},\hat{b}_{i-s+1},\ldots,\hat{b}_i)$. To simplify the notation sometimes we will write $g_i(b_{i-s}^i)$ instead of $g_i(b^i)$ for such decoders. Let $\mathcal{F}^\delta$ denote the collection of all (randomized) delay-δ sequential source codes of rate R, and let $\mathcal{F}_s^\delta$ denote the class of codes in $\mathcal{F}^\delta$ with memory s.

Let $\mathcal{F} \subset \mathcal{F}_s^\delta$ be a finite class of base reference codes. Our goal is to construct a coding scheme that performs, for every sequence x^n, asymptotically as well as the best coding scheme which employs codes from $\mathcal{F}$ and is allowed to change the code m times. Formally, a code in this class $\mathcal{F}_{m,n}$ is given by integers $1 \leq i_1 < i_2 < \ldots < i_m < n$ and codes $\{(f_i^{(j)}, g_i^{(j)}\}_{i=1}^\infty$, $j = 0,\ldots,m$ such that x_i is encoded to $b_i = f_i^{(j)}(x^{i+\delta})$ if $i_j < i \leq i_{j+1}$, where $i_0 = 0$ and $i_{m+1} = n$. The "idealized" minimum normalized cumulative distortion achievable by such schemes for n reproduction values is

$$D_{\mathcal{F},m,n}^*(\mathbf{x}) = \frac{1}{n}\min_{1\leq i_1<i_2<\ldots<i_m<n}\sum_{j=0}^{m}\min_{(f,g)\in\mathcal{F}}\left\{\sum_{i=i_j+1}^{i_{j+1}} d\left(x_i, g_i\left(f_{i-s}(x^{i+\delta-s}),\ldots,f_i(x^{i+\delta})\right)\right)\right\} \quad (3)$$

where $\mathbf{x} = (x_1, x_2, \ldots)$ denotes the entire sequence. Note that to find the best scheme achieving this minimum one has to know the sequence $x^{n+\delta}$ in advance. The minimum above is idealized (and so optimistic). This is because when the code is changed, any real coding system has to wait s symbols to be able to fully utilize the decoder's memory, but in the formula above we assume that the decoder can operate correctly immediately after the change.

IV. Tracking the best finite-delay finite-memory source code

In this section a low-complexity coding scheme is provided to track the best code from $\mathcal{F}_{m,n}$. The scheme is a combination of the coding scheme of [4] and the decision scheme of [10] discussed in Section II.

The scheme works as follows. Divide the source sequence x^n into non-overlapping blocks of length l (for simplicity assume that l divides n). At the beginning of the kth block, that is, at time instants $t = kl+1$, $k = 0,\ldots,n/l-1$, a coding scheme $(f^{(k)}, g^{(k)}) = \{f_i^{(k)}, g_i^{(k)}\}_{i=1}^\infty$ is chosen randomly from the reference class $\mathcal{F}$. The exact distribution of $(f^{(k)}, g^{(k)})$ will be specified later based on the results in Section II. Then the encoder uses the first $\lceil \frac{1}{R}\log|\mathcal{F}|\rceil$ time instants of the block to describe the selected coding scheme $(f^{(k)}, g^{(k)})$ to the receiver ($\lceil x\rceil$ denotes the smallest integer not less than x), that is, for time instants

$$i = (k-1)l+1,\ldots,(k-1)l+\left\lceil\frac{1}{R}\log|\mathcal{F}|\right\rceil$$

an index identifying $(f^{(k)}, g^{(k)})$ is transmitted. In the rest of the block, that is, for time instants

$$i = (k-1)l+\left\lceil\frac{1}{R}\log|\mathcal{F}|\right\rceil+1,\ldots,kl$$

the encoder uses $f_i^{(k)}$ to produce and transmit $b_i = f_i^{(k)}(x^{i+\delta}, U^i)$ to the receiver. In the first

$$h = \left\lceil\frac{1}{R}\log|\mathcal{F}|\right\rceil + s$$

time instants of the kth block, that is, while the index of the coding scheme $(f^{(k)}, g^{(k)})$ is communicated and the first s correct channel symbols are received, the decoder emits an arbitrary reproduction symbol $\hat{x}_i = \hat{x}$ with distortion upper bounded by

$$\hat{d} = \sup_{x\in\mathcal{X}} d(x,\hat{x}) \leq 1.$$

In the remainder of the block, the decoder uses $g_i^{(k)}$ to decode the transmitted channel symbols as

$$\hat{x}_i = g_i^{(k)}(b^i) = g_i^{(k)}(b_{i-s}^i)$$

where $b_{i-s}^i = (b_{i-s}, b_{i-s+1},\ldots,b_i)$ (recall that the decoder $g^{(k)}$ has finite memory s).

Now except for the distortion induced by communicating the quantizer index and the first s correct code symbols at the beginning of each block, the above scheme can easily be fitted in the sequential decision framework. We want to make a sequence of decisions concerning the sequence $\{y_k\}$ with $y_k = (x_{(k-1)l+h+1},\ldots,x_{kl})$ for $k = 1,\ldots,n/l$. We consider any $(f,g) \in \mathcal{F}$ an expert whose prediction is $\hat{y}_k^{(f,g)} = (\hat{x}_{(k-1)l+h+1}^{(f,g)},\ldots,\hat{x}_{kl}^{(f,g)})$, where $\hat{x}_i^{(f,g)} = g_i(\bar{b}^i)$ and $\bar{b}_i = f_i(x^{i+\delta})$, incurring loss

$$\ell(y,\hat{y}) = \sum_{j=1}^{l-h} d(x(j),\hat{x}(j))$$

where $y = (x(1),\ldots,x(l-h))$ and $\hat{y} = (\hat{x}(1),\ldots,\hat{x}(l-h))$. Then

$$\sum_{i=1}^{n} d(x_i,\hat{x}_i) \leq \sum_{i=1}^{n/l}\ell(y_i,\hat{y}_i) + \frac{nh\hat{d}}{l}$$

where the second term comes from the fact that in each block the distortion of each of the first h symbols is at most $\hat{d}$.

Choosing $\{f^{(k)}, g^{(k)}\}$ according to Algorithm 1 for the above model, the following performance bound can be proved based on Theorem 1.

Theorem 2: Let $\mathcal{F} \subset \mathcal{F}_s^\delta$ be a subclass of codes with delay δ and memory s. Assume that m, n, l, M and s are positive integers such that $m < n/l$, $h = \lceil \log|\mathcal{F}|/\log M\rceil + s \leq l$,

and l divides n, and let $\eta > 0$ and $0 < \alpha < 1$. Then the normalized cumulative distortion of the above coding scheme can be bounded for any sequence $\mathbf{x}$ and $n \geq 0$ as

$$\begin{aligned}\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} d(x_i,\hat{x}_i)\right] - D^*_{\mathcal{F},m,n}(\mathbf{x}) \\ \leq \quad \frac{h\hat{d}}{l} + \frac{1}{\eta n}\ln\left(\frac{|\mathcal{F}|^{m+1}}{\alpha^m(1-\alpha)^{n/l-m-1}}\right) \\ + \frac{\eta(l-h)^2}{8l} + \frac{m(l-1)}{n}.\end{aligned}$$

Letting $\alpha = m/(n/l-1)$, choosing η to minimize the above bound, and setting $l = c_1(n/m)^{1/3}\log^{2/3}|\mathcal{F}|/\log(n|\mathcal{F}|/m)$, similarly to (2), we obtain the following corollary.

Corollary 1: Assume that $n/m > (\lceil \log|\mathcal{F}|/R\rceil + s)^3$. Then there is a zero-delay sequential coding scheme with normalized distortion redundancy

$$\begin{aligned}\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} d(x_i,\hat{x}_i)\right] - D^*_{\mathcal{F},m,n}(\mathbf{x}) \\ \leq \quad O\left(\frac{m}{n}\log|\mathcal{F}|\ \log\frac{n|\mathcal{F}|}{m}\right)^{1/3}.\end{aligned}$$

As mentioned in the introduction, the straightforward implementation of this algorithm requires the maintenance of only $|\mathcal{F}|$ weights. In contrast, the straightforward implementation of the algorithm of [4] for the combined reference class $\mathcal{F}_{m,n}$ would require $\sum_{j=0}^{m}\binom{n}{j}|\mathcal{F}|(|\mathcal{F}|-1)^j$ weights. Thus, the proposed algorithm results in a drastic reduction of both the space and time complexity, while keeping the performance bound essentially the same. On the other hand, if the base class $\mathcal{F}$ is too complex, the algorithm may still be computationally too expensive to implement. However, if $\mathcal{F}$ has some special structure, then it may be utilized to obtain more efficient implementations. In the next section we provide an example when $\mathcal{F}$ is the set of scalar quantizers.

V. Tracking the best scalar quantizer

In this section we consider the special situation where the base reference class $\mathcal{F}$ is the class of scalar quantizers. Note that scalar quantization is a zero delay coding scheme. To be able to utilize the special structure of scalar quantizers, we introduce an alternative implementation of the on-line decision algorithm of Section II. In a recent work [11], we have shown that the random choice of the decision according to (1) can be performed in two steps. First we choose a random time τ_t, which specifies how many most recent samples we are going to use for the prediction, then we choose the decision according to the exponentially weighted average prediction for these samples.

Algorithm 2: For $t = 1$, choose $\hat{y}_1$ uniformly from the set $\{\hat{y}_{1,1},\dots,\hat{y}_{N,1}\}$. For $t \geq 2$, choose τ_t randomly according to the distribution

$$\mathbb{P}\{\tau_t = t'\} = \begin{cases}\frac{(1-\alpha)^{t-1}Z_{1,t-1}}{NW_t} & \text{for } t'=1\\ \frac{\alpha(1-\alpha)^{t-t'}W_{t'}Z_{t',t-1}}{NW_t} & \text{for } t'=2,\dots,t\end{cases}$$

where $Z_{t',t-1} = \sum_{i=1}^{N} e^{-\eta L([t',t-1],i)}$ for $1 \leq t' \leq t-1$, and $Z_{t,t-1} = N$; $W_1 = 1$ and

$$W_t = \frac{\alpha}{N}\sum_{t'=2}^{t-1}(1-\alpha)^{t-1-t'}W_{t'}Z_{t',t-1} + \frac{(1-\alpha)^{t-2}}{N}Z_{1,t-1}.$$

Given $\tau_t = t'$, choose $\hat{y}_t$ randomly according to the probabilities

$$\mathbb{P}\{\hat{y}_t = \hat{y}_{i,t}|\tau_t = t'\} = \begin{cases}\frac{e^{-\eta L([t',t-1],i)}}{Z_{t',t-1}} & \text{for } t'=1,\dots,t-1\\ \frac{1}{N} & \text{for } t'=t.\end{cases}$$

It can be shown that Algorithm 2 provides an alternative implementation of Algorithm 1.

Theorem 3 ([11]): Algorithm 1 and Algorithm 2 are equivalent in the sense that the generated decision sequences have the same distribution. In particular, the sequence $(\hat{y}_1,\dots,\hat{y}_T)$ generated by Algorithm 2 satisfies

$$\mathbb{P}\{\hat{y}_t = \hat{y}_{i,t}\} = v_t^{(i)}$$

for all t and i, where $v_t^{(i)}$ are the normalized weights generated by Algorithm 1.

The implementation of Algorithm 2 is useful if efficient algorithms are available to compute the constants $Z_{k',k}$. Note that $Z_{k',k}$ is the total weight assigned to the problem by the exponentially weighted average prediction method in the interval $[k',k]$. We will use the fact that these constants can be efficiently computed for the related scalar quantization problem [5].

An M-level scalar quantizer Q is a measurable mapping $\mathbb{R} \to C$, where the *codebook* C is a finite subset of $\mathbb{R}$ 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 $(Q(x)-x)^2 = \min_{\hat{x}\in C}(x-\hat{x})^2$. 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. Our goal is to design a coding scheme that performs, for any sequence x^n, asymptotically as well as the best coding scheme which employs M-level scalar quantizers and is allowed to change quantizer m times. Formally, a code in this class $\mathcal{Q}_{m,n}$ is given by integers $1 \leq i_1 < i_2 < \dots < i_m < n$ and M-level scalar quantizers $q_0,\dots,q_m \in \mathcal{Q}$ such that x_i is encoded to $q_j(x_i)$ if $i_j < i \leq i_{j+1}$, where $i_0 = 0$ and $i_{m+1} = n$. The minimum normalized cumulative distortion achievable by such schemes is

$$D^*_{m,n}(x^n) = \frac{1}{n}\min_{1\leq i_1<\dots<i_m<n}\sum_{j=0}^{m}\min_{q\in\mathcal{Q}}\sum_{i=i_j+1}^{i_{j+1}}(x_i-q(x_i))^2.$$

Note that to find the best scheme achieving this minimum, one has to know the entire sequence x^n in advance.

The expected *normalized distortion redundancy* of a scheme (with respect to the class $\mathcal{Q}_{m,n}$) is the quantity

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

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 order to be able to apply the coding procedure of Section III, we need to find a finite class of codes that is sufficiently close to $\mathcal{Q}$. Let $\mathcal{Q}_K$ denote the class of nearest-neighbor scalar quantizers whose code points all belong to the finite grid $\{1/(2K), 3/(2K), \ldots, (2K-1)/(2K)\}$. It is easy to see that for any quantizer $Q \in \mathcal{Q}$ there is a quantizer $Q^* \in \mathcal{Q}_K$ such that $\sup_{x\in[0,1]}\left|(x-Q(x))^2-(x-Q^*(x))^2\right| \le 1/K$, thus the best quantizers in $\mathcal{Q}_k$ have essentially the same performance as the best quantizers in $\mathcal{Q}$. Thus, when applying the coding scheme of Section III, at time instances $t = kl+1, k = 0.\ldots, n/l-1$, we will choose a code (quantizer) $Q_k = (f^{(k)}, g^{(k)})$ from $\mathcal{Q}_K$.

Moreover, to reduce complexity, Algorithm 2 is used to compute Q_k instead of Algorithm 1. The efficient computation of the constants

$$Z_{k',k} = \sum_{Q\in\mathcal{Q}_K} e^{-\eta\sum_{i=(k'-1)l+1}^{kl}(x_i-Q(x_i))^2}$$

for all $k' \le k$ is possible by a recent efficient implementation of the scheme of [4] given in [5]. The method is very similar to the weight pushing algorithm (see, e.g., [14]), and relies on a correspondence between optimal quantizer design and shortest path search in weighted digraphs. Instead of randomly choosing the quantizer Q_k at once, its code points are chosen in an increasing order, one at a time; this approach provides a substantial reduction in the complexity. The complexity of the algorithm can be further reduced if one uses the finely quantized version $\bar{x}_i = q_K(x_i)$ of the input, where q_K is a K level uniform quantizer on $[0,1]$, instead of the original values. While this modification results in a slightly deteriorated performance, it makes possible to implement the algorithm in linear time.

Using different choices of K and l, the following bounds can be given on the performance and computational complexity of the algorithm.

Corollary 2: Assume that $n > m$ and $n/m > \left\lceil M\log(n/m)/(3R)\right\rceil^3$. Then there is a zero-delay sequential coding scheme with normalized distortion redundancy

$$\mathbb{E}\left\{\frac{1}{n}\sum_{i=1}^{n}(x_i-\hat{x}_i)^2\right\}-D^*_{m,n}(x^n) \le O\left(\left(\frac{m}{n}\right)^{\frac{1}{3}}\log\frac{n}{m}\right)$$

and computational complexity $O((M+m)n^2)$. On the other hand, if $m = o(n^{1/3}/\log n)$, then there is a zero-delay sequential scheme with computational complexity $O(Mn)$ (i.e., linear in time), and normalized distortion redundancy

$$\mathbb{E}\left\{\frac{1}{n}\sum_{i=1}^{n}(x_i-\hat{x}_i)^2\right\}-D^*_{m,n}(x^n) \le O\left(\frac{m^{1/2}\log^{1/2} n}{n^{1/6}}\right).$$

VI. Conclusion

We provided low-complexity randomized limited-delay lossy source coding schemes which can perform, on any source sequence, asymptotically as well as the best combined coding scheme which is allowed to change the employed code from a finite class of limited-delay finite memory coding schemes from time to time. When the reference codes are combinations of scalar quantizers, even further complexity reduction was achieved, and a coding scheme with linear complexity was presented. Extensions to other special reference classes, as well as the proofs of the results of this paper can be found in [16]. These include the case of multiple description scalar quantization that can be used in situations where the feedback information has to be transmitted over a non-reliable lossy packet network.

References

[1] S. Tatikonda and S. Mitter, "Control over noisy channels,"*IEEE Trans. Automatic Control*, vol. 49, no. 7, pp. 1196–1201, Jul. 2004.

[2] 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.

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

[4] 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.

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

[6] 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.

[7] 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.

[8] G. I. Shamir and N. Merhav, "Low-complexity sequential lossless coding for piecewise-stationary memoryless sources," *IEEE Trans. Inform. Theory*, vol. IT-45, pp. 1498–1519, July 1999.

[9] G. I. Shamir and D. J. Costello, Jr., "Asymptotically optimal low-complexity sequential lossless coding for piecewise-stationary memoryless sources – part i: The regular case," *IEEE Trans. Inform. Theory*, vol. IT-46, pp. 2444–2467, Nov. 2000.

[10] M. Herbster and M. K. Warmuth, "Tracking the best expert," *Machine Learning*, pp. 151–178, 1998.

[11] A. György, T. Linder, and G. Lugosi, "Tracking the best of many experts," in *COLT 2005* (P. Auer and R. Meir, eds.), LNAI 3559, (Berlin–Heidelberg), pp. 204-216, Springer-Verlag, 2005.

[12] O. Bousquet and M. K. Warmuth, "Tracking a small set of experts by mixing past posteriors," *Journal of Machine Learning Research*, vol. 3, pp. 363–396, Nov. 2002.

[13] N. Cesa-Bianchi and G. Lugosi, "On prediction of individual sequences," *Annals of Statistics*, vol. 27, pp. 1865–1895, 1999.

[14] E. Takimoto and M. K. Warmuth, "Path kernels and multiplicative updates," in *COLT 2002* (J. Kivinen and R. H. Sloan, eds.), LNAI 2375, (Berlin–Heidelberg), pp. 74–89, Springer-Verlag, 2002.

[15] A. György, T. Linder, and G. Lugosi, "Tracking the best quantizer," *2005 IEEE Int. Symp. Inform. Theory*, Adelaide, Australia, Sep. 2005.

[16] A. György, T. Linder, and G. Lugosi, "Tracking the best quantizer." in preparation.