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

András György, *Member, IEEE*, and Gergely Neu

***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, and the same redundancy is achievable, up to logarithmic factors, when the reference class is the set of scalar quantizers. It has also been shown that the distortion redundancy is at least of order $1/\sqrt{T}$ in the latter case, and the lower bound can easily be extended to sufficiently powerful (possibly finite) reference coding schemes. In this paper, we narrow the gap between the upper and lower bounds, and give a compression scheme whose normalized distortion redundancy is $O(\sqrt{\ln(T)/T})$ relative to any finite class of reference schemes, only a logarithmic factor larger than the lower bound. The method is based on the recently introduced shrinking dartboard prediction algorithm, a variant of exponentially weighted average prediction. The algorithm is also extended to the problem of joint source-channel coding over a (known) stochastic noisy channel and to the case when side information is also available to the decoder (the Wyner–Ziv setting). The same improvements are obtained for these settings as in the case of a noiseless channel. Our method is also applied to the problem of zero-delay scalar quantization, where $O(\ln(T)/\sqrt{T})$ normalized distortion redundancy is achieved relative to the (infinite) class of scalar quantizers of a given rate, almost achieving the known lower bound of order $1/\sqrt{T}$. The computationally efficient algorithms known for scalar quantization and the Wyner–Ziv setting carry over to our (improved) coding schemes presented in this paper.**

***Index Terms*— Distortion redundancy, individual sequences, joint source-channel coding, lossy source coding, sequential coding, switching cost, universal coding.**

Manuscript received January 21, 2013; revised October 17, 2013; accepted December 12, 2013. Date of publication February 20, 2014; date of current version April 17, 2014. This work was supported in part by the Hungarian Scientific Research, in part by the Hungarian National Office for Research and Technology under Grant OTKA-NKTH CNK 77782, in part by the PASCAL2 Network of Excellence through EC under Grant 216886, in part by the Alberta Innovates Technology Futures, and in part by the Natural Sciences and Engineering Research Council of Canada. This work was presented in part at the 2011 IEEE International Symposium on Information Theory.

A. György was with the Machine Learning Research Group, Computer and Automation Research Institute, Hungarian Academy of Sciences, Budapest 1111, Hungary. He is now with the Department of Computing Science, University of Alberta, Edmonton, AB T6G 2E8, Canada (e-mail: gyorgy@ualberta.ca).

G. Neu was with the Department of Computer Science and Information Theory, Budapest University of Technology and Economics, Budapest 1117, Hungary, and also with the Computer and Automation Research Institute, Hungarian Academy of Sciences, Budapest 1111, Hungary. He is now with INRIA Lille - Nord Europe, Villeneuve d'Ascq 59650, France (e-mail: neu.gergely@gmail.com).

Communicated by T. Weissman, Associate Editor for Shannon Theory.

Digital Object Identifier 10.1109/TIT.2014.2307062

## 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 $\delta_1$ encoding and $\delta_2$ 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 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 (in particular, 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, they show 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]–[4]: at each time instant a coding scheme (a scalar quantizer) is selected based on its "estimated" performance. A major problem in this approach is that the prediction, and hence the choice of the quantizer at each time instant, is performed based on the source sequence which is not known exactly at the decoder. Therefore, in [1] information about the source sequence that is used in the random choice of the quantizers is also transmitted over the channel, reducing the available capacity for actually encoding the source symbols.

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, selected at the source, whose identity is conveyed to the receiver 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 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 disadvantage of this method is that the EWA prediction algorithm keeps one weight for each code in the reference class, and so the computational complexity of the method becomes prohibitive even for relatively simple and small reference classes. Computationally efficient solutions to the case of zero-delay scalar quantization were given by György, Linder and Lugosi using dynamic programming [6] and EWA prediction in [7] and based on the "follow-the-perturbed-leader" prediction method (see [8], [9]) in [10]. Over a channel with alphabet size $M$, the first method achieves the $O(T^{-1/3} \ln T)$ redundancy of Weissman and Merhav with $O(MT^{4/3})$ computational and $O(T^{2/3})$ space complexity and a somewhat worse $O(T^{-1/4}\sqrt{\ln T})$ distortion redundancy with linear $O(MT)$ time and $O(T^{1/2})$ space complexity, while the second method achieves $O(T^{-1/4} \ln T)$ distortion redundancy with the same $O(MT)$ linear time complexity and $O(MT^{1/4})$ space complexity.

Matloub and Weissman [11] extended the problem to allow a discrete stochastic channel between the encoder and the decoder, while Reani and Merhav [12] extended the model to the Wyner-Ziv case (i.e., when side information is also available at the decoder). The performance bound in both cases are based on [5] while low-complexity solutions for the zero-delay scalar quantization case are provided based on [10] and [7], respectively. Finally, the case when the reference class is a set of time-varying limited-delay limited-memory coding scheme was analyzed in [13], and efficient solutions were given for the zero-delay case for both traditional and network (multiple-description and multi-resolution) scalar quantization.

Since most of 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 [7], leaving a gap between the best known upper and lower bounds. Furthermore, if the identity of the used coding scheme were communicated as side information (before the encoded symbol is revealed), that is, no channel bandwidth were needed to be devoted to communicate the identity of the decoder, the employed EWA prediction method would guarantee an $O(\sqrt{\ln |\mathcal{F}|/T})$ distortion redundancy for any finite reference coder class $\mathcal{F}$ (of limited delay and limited memory), in agreement with the lower bound.[1]

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 than blockwise coding. This goal can be achieved by the recent Shrinking Dartboard (SD) algorithm of Geulen, Voecking, and Winkler [14], a modified version of the EWA prediction method that is designed to control the number of expert switches while keeping the same marginal distributions for the predictions as EWA, and so provides similar performance guarantees.

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. These bounds are only logarithmic factors larger than the corresponding lower bound. Note that Devroye, Lugosi, and Neu [15] has recently introduced a "follow the perturbed leader"-type prediction method that also keeps the number of expert switches low. Applying this algorithm in place of the mSD algorithm in our coding schemes would yield similar results.

In Section II we revisit the SD algorithm of [14] 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) scalar quantization in Section V. Extensions to the noisy channel and the Wyner-Ziv settings are given 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 [14]. 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 $d_{t,\mathbf{i}_t}$. After 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 $T$.

[1] This follows from the fact that if the chosen decoding function is known at the receiver, the problem becomes an instance of the prediction with expert advice problem. Similarly, the $O(\sqrt{\ln |\mathcal{F}|/T})$ bound can be obtained by analyzing the bounds of the block-coding methods devised based on [5] and setting the term resulting from the communication overhead to zero.

**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 $D_{0,i} = 0$ 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
```

It is assumed that $\{d_{t,i}\}$, the sequence of losses, is fixed in advance for all $i \in \mathcal{F}$ and $t = 1, 2, \ldots$, but it is unknown to the decision maker a priori, who only learns the values $d_{t,i}, i \in \mathcal{F}$ after $\mathbf{i}_t$ has been selected. It is also 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 $\mathbf{i}_t$ at time step $t$ depends only on $\mathbf{U}^t = (\mathbf{U}_1, \ldots, \mathbf{U}_t)$ and $d_{\tau,i}, \tau = 1, \ldots, t-1, i \in \mathcal{F}$.

A well-known solution to this problem (which is optimal under various conditions) is the EWA prediction method that, at time step $t$, chooses action $i$ with probability proportional to $e^{-\eta_t D_{t-1,i}}$ for some sequence of positive step size parameters $\{\eta_t\}_{t=1}^{T}$ [2]–[4].[2] It can be shown (using techniques developed in [17] and [18]) 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 [14] addresses this problem and provides the same performance guarantee as EWA while controlling the number of switches between different actions, that is, the number of time instants when $\mathbf{i}_t \ne \mathbf{i}_{t-1}$. 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 $c_1 = 1$. For $t \ge 2$ it follows since

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

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. By precise tuning of this probability, the method guarantees that actions are changed over time only at most $O(\sqrt{T})$ times in $T$ time steps, while maintaining the same marginal distributions over the actions as the EWA algorithm. The latter fact guarantees that the expected regret of the two algorithms are the same; in particular, the same parameter setting gives the optimal $O(\sqrt{\ln N/T})$ expected regret.

In the following we formalize the above statements concerning the mSD algorithm. We state two results crucial for the analysis of the coding scheme that we will propose in the next section. Since the proofs are obtained by minor modifications of existing results, they are deferred to the appendix.

The first 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 [14, Lemma 1].

*Lemma 1:* For all $t = 1, 2, \ldots$ and $i \in \mathcal{F}$, the mSD algorithm selects action $i$ at time $t$ with probability $p_{t,i}$, that is, $\mathbb{P}[\mathbf{i}_t = i] = p_{t,i}$.

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 [14, Lemma 2]. That is, the following result can be obtained by a slight modification of the proof of [17, Lemma 1] for EWA (the same bound for the specific time-dependent choice of $\eta_t$ discussed after the lemma follows directly as a special case of [18, Theorem 2]).

*Lemma 2:* For any $T \ge 1$, 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 $\eta_t = \sqrt{8 \ln N/T}$ optimally (as a function of the time horizon $T$), the bound becomes $\sqrt{\ln N/(2T)}$, while setting $\eta_t = 2\sqrt{\ln N/t}$ independent of $T$, we have $\mathbb{E}[\mathbf{R}_T] \le \sqrt{\ln N/T}$ (here we used $\sum_{t=1}^{T} 1/\sqrt{t} \le 2\sqrt{T}$ and optimized the constant in setting $\eta_t = const\sqrt{\ln N/t}$).

Let $\mathbf{S}_T = |\{t : \mathbf{i}_t \ne \mathbf{i}_{t-1}, 1 < t \le T\}|$ 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 [14] gives an upper bound on $\mathbf{S}_T$.

*Lemma 3:* The expected number of times the mSD algorithm switches between different actions in $T$ time steps

[2] EWA is probably the best-known algorithm for the sequential prediction problem considered here, also known as the problem of prediction with expert advice. It is a special case of both generally used approaches to solve such problems, the follow the regularized leader and the mirror descent algorithms. In the lossless data compression scenario, when the predictions and experts define probability distributions for a source sequence, and the loss is measured as the negative logarithm of the probability of the observed symbol, EWA is just the Bayesian mixture predictor for the models defined by the experts. For more details, see, for example, [16].

can be bounded as

$$\mathbb{E}\left[\mathbf{S}_T\right] \le \min\left\{ \eta_T D^*_{T-1} + \ln N + \sum_{t=2}^{T-1}(\eta_t - \eta_T), \sum_{t=2}^{T}(2\eta_t - \eta_T)\right\}, \quad (1)$$

where $D^*_{T-1} = \min_{i\in\mathcal{F}} D_{T-1,i}$.
The second expression in the above minimum is better by a $\ln N$ term when all the $\eta_t$ are the same and $D^*_{T-1}$ is bounded by $T-1$, but the first expression is preferable for the typical time-varying $\eta_t$. In particular, for $\eta_t = \sqrt{\ln N/T}$, we have $\mathbb{E}\left[\mathbf{S}_T\right] \le \sqrt{T \ln N}$, while setting $\eta_t = 2\sqrt{\ln N/t}$, we obtain $\mathbb{E}\left[\mathbf{S}_T\right] \le 4\sqrt{T\ln N} + \ln N$ (using again $\sum_{t=1}^{T} 1/\sqrt{t} \le 2\sqrt{T}$).

## III. Limited-Delay Limited-Memory Sequential Source Codes

A fixed-rate delay-$\delta$ (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 $\delta$ 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 $\hat{\mathbf{x}}_t \in \widehat{\mathcal{X}}$ based on the channel symbols $\mathbf{y}^t = (\mathbf{y}_1, \ldots, \mathbf{y}_t)$ received so far, where $\widehat{\mathcal{X}}$ is the reconstruction alphabet.

Formally, a delay-$\delta$ (randomized) sequential source code of rate $\ln M$ 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

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

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 $\delta$.[3] To simplify the notation we will omit the randomizing sequence from $f_t(\cdot, \mathbf{U}^t)$ and write $f_t(\cdot)$ instead.

We will denote by $\mathcal{F}^\delta$ the collection of all deterministic delay-$\delta$ sequential source codes of rate $\ln M$. Similarly to [5], we will consider decoders of limited memory. A decoder $\{g_t\}$ is said to be of memory $s \ge 0$ 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 $\tilde{y}^t_{t-s} = (\tilde{y}_{t-s}, \tilde{y}_{t-s+1}, \ldots, \tilde{y}_t)$.

In what follows, $\mathcal{F}^\delta_s$ will denote the class of codes in $\mathcal{F}^\delta$ whose decoders are of memory $s$.

Now let $\mathcal{F} \subset \mathcal{F}^\delta$ be a finite set of reference codes with $|\mathcal{F}| = N$. Note that here we implicitly made the simplifying assumption that $\mathcal{F}$ contains only deterministic coding schemes. This assumption is only used for notational convenience: all of our results can easily be extended to randomized reference coding schemes (which use independent randomization) by conditioning on the randomization used by the reference codes and applying our results to the resulting deterministic schemes.

The *cumulative distortion* of a sequential scheme after reproducing the first $T$ symbols is given by

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

where $d : \mathcal{X} \times \widehat{\mathcal{X}} \to [0, 1]$ is some distortion measure,[4] 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, $y_t = f_t\left(x^{t+\delta}\right)$. 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*

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

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

Weissman and Merhav [5] proved that there exists a randomized coding scheme such that, for any $\delta \ge 0$ and $s \ge 0$ and for any finite class $\mathcal{F} \subset \mathcal{F}^\delta_s$ of reference codes, the normalized distortion redundancy with respect to $\mathcal{F}$ is of order $T^{-1/3} \ln^{2/3} |\mathcal{F}|$. This coding scheme splits the source sequence into blocks of length $O(T^{1/3})$. At the beginning of each block a code is selected from $\mathcal{F}$ using EWA prediction and the identity of the selected reference decoder function is communicated to the decoder. During these first steps, the decoder emits arbitrary reproduction symbols, while the chosen code is used in the rest of the block. The formation of the blocks ensures that only a limited fraction of the available channel capacity is used for describing codes, while the limited memory property ensures that not transmitting real data at the beginning of each block has only a limited effect on decoding the rest of the block.

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

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

## 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, non-empty reference class $\mathcal{F} \subseteq \mathcal{F}_s^\delta$, our coding scheme, given in Algorithm 2, 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 $\mathcal{F}$, and the loss associated with a code $(f, g) \in \mathcal{F}$ at this time instant is defined by

$$d_{t,(f,g)}(x^{t+\delta}) = d\left(x_t, g_t\left(y^t\right)\right), \tag{2}$$

where $y^t$ is the sequence obtained by using the coding scheme $(f, g)$ to encode $x^t$, that is, $y_t = f_t(x^{t+\delta})$ for all $t$ (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 function 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 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 function chosen by the mSD algorithm is communicated, where

$$B = \left\lceil \frac{\ln |\{g : (f, g) \in \mathcal{F}\}|}{\ln M} \right\rceil \tag{3}$$

is the number of channel symbols required to describe uniquely all possible decoder functions. In the remainder of the block the selected encoder is used to compress 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. However, this does not hold if $\mathbf{v}$ is not selected independently of $u$, for example, if the beginning of $u$ is a postfix of $\mathbf{v}$. As an illustration, consider the case when $\mathbf{v}$ is the all-one vector: then, after its transmission, the probability that the last $A$ symbols equal $\mathbf{v}$ is increased for $A - 1$ steps. To avoid these situations, we ensure that no new-block signal is sent too soon after another one has been transmitted; specifically, we wait $B + A - 1$ steps, and so the receiver does not have to check for the new-block signal for $B + A - 1$ steps after one is received. We use $B$ steps to transmit the decoder function index and $A - 1$ steps to ensure that when the receiver first checks for a new-block signal, the last $A$ symbols are completely independent of the new-block signal $\mathbf{v}$ (note that since the starting positions of the blocks may depend on $\mathbf{v}$, so do the symbols transmitted in the decoder function index).

In summary, the algorithm works in blocks of variable length as follows: At the beginning of the block an algorithm is selected using the mSD prediction algorithm and a new-block signal and the identity of the chosen decoder function is communicated to the receiver. In the next time steps, as long as the mSD algorithm selects the same decoder function, the chosen code is used to encode the source symbols at the sender and used for decoding at the receiver. When the mSD method selects a different decoder function, or a new-block signal is transmitted by chance, a new block is started both at the encoder and the decoder. Note that the encoder and the decoder use a slightly different blocking: the blocks of the encoder start with a new-block signal, while the blocks on the decoder side end with the new-block signal. The method is shown in Algorithm 2.

The next theorem gives a bound on the performance of our proposed coding scheme.

*Theorem 1:* Let $T \geq 1$ and $\eta_t = \eta > 0$ for all $1 \leq t \leq T$. Then the expected normalized distortion redundancy of Algorithm 2 for any finite, non-empty reference class $\mathcal{F} \subset \mathcal{F}_s^\delta$ can be bounded as

$$\widehat{R}_T \leq \frac{\ln |\mathcal{F}|}{T\eta} + \eta\left(\frac{1}{8} + A + B + s\right) + \frac{(A + B + s)\left(1 + \frac{T-A}{M^A}\right)}{T},$$

where $B$ is defined in (3).

Setting the parameters of the algorithm appropriately, we immediately see that the normalized distortion redundancy of the proposed scheme becomes $O(\sqrt{\ln(T)/T})$:

*Corollary 1:* Let $\mathcal{F} \subset \mathcal{F}_s^\delta$ be a finite, non-empty reference class of delay-$\delta$ memory-$s$ codes, and, for time horizon $T \geq 1$, set $A = \lceil \ln T / \ln M \rceil$ and

$$\eta_t = \eta = \sqrt{\frac{\ln |\mathcal{F}|}{T\left(\frac{17}{8} + \frac{\ln(T|\mathcal{F}|)}{\ln M} + s\right)}} \tag{4}$$

for all $1 \leq t \leq T$. Then the expected normalized distortion redundancy of Algorithm 2 can be bounded as

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

*Remark (Unknown time horizon):* In the above, the parameters $A = \left\lceil \frac{\ln T}{\ln M} \right\rceil$and $\eta = O(1/\sqrt{T \ln T})$ have been 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, that is, its parameters do not depend on $T$. The simplest way to achieve this is to use the

**Algorithm 2** A Near-Optimal Algorithm for Adaptive Sequential Lossy Source Coding

```
Encoder:
  1) Input: A finite, non-empty reference class $\mathcal{F} \subset \mathcal{F}_s^{\delta}$, positive integer $A$, and time horizon $T$.
  2) Initialization
     a) Draw a new-block signal $\mathbf{v}$ uniformly at random from $\{1, \ldots, M\}^A$, the set of channel symbol sequences of length $A$.
     b) Initialize the mSD algorithm for $\mathcal{F}$ and set $B$ according to (3).
  3) For each block do
     a) Observe $x_{t+\delta}$.
     b) For all time instants $(t+\delta)$ run the mSD algorithm:
        i) Feed the mSD algorithm with losses $d_{t,(f,g)}(x^{t+\delta})$ for each code $(f, g) \in \mathcal{F}$.
        ii) Let $(\mathbf{f}^{(t)}, \mathbf{g}^{(t)})$ denote the choice of the mSD algorithm.
     c) In the first $A$ time steps of the block transmit $\mathbf{v}$.
     d) After the first $A$ time steps set $(\mathbf{f}, \mathbf{g}) = (\mathbf{f}^{(t)}, \mathbf{g}^{(t)})$, the output of the mSD algorithm in this time step.
     e) In time steps $A+1, \ldots, A+B$ of the block send the index describing $\mathbf{g}$.
     f) If $(t+\delta)$ belongs to steps $A+B+1, A+B+2, \ldots$ of the block then
        i) if $\mathbf{g}^{(t)} = \mathbf{g}$ then transmit $\mathbf{y}_t = \mathbf{f}_t(x^{t+\delta})$;
        ii) else start a new block with the same time index.
     g) If $(t+\delta)$ belongs to steps $2A+B, 2A+B+1, \ldots$ of the block and $(\mathbf{y}_{t-A+1}, \ldots, \mathbf{y}_t) = \mathbf{v}$ then start a new block and declare the current time instant as the $A$th step of the new block.
Decoder:
  1) Input: A finite, non-empty reference class $\mathcal{F} \subset \mathcal{F}_s^{\delta}$, positive integers $A$, $B$, time horizon $T$.
  2) For $t = 1, \ldots, A$
     a) Observe $\mathbf{y}_t$ and output an arbitrary symbol $\hat{\mathbf{x}}_t \in \widehat{\mathcal{X}}$.
     b) At time $t = A$ set $\mathbf{v} = \mathbf{y}^A$ and declare a new block.
  3) For each block do
     a) Observe $\mathbf{y}_t$.
     b) In the first $B$ time steps of the block receive the index of the decoder to be used and output an arbitrary symbol $\hat{\mathbf{x}}_t \in \widehat{\mathcal{X}}$. At time step $B$ of the block set the decoder $\mathbf{g}$ according to the symbols received so far.
     c) In time steps $B+1, B+2, \ldots$ of the block output $\hat{\mathbf{x}}_t = \mathbf{g}(\mathbf{y}^t) = \mathbf{g}(\mathbf{y}_{t-s+1}^t)$.
     d) In time steps $A+B, A+B+1, \ldots$ of the block declare a new block if $(\mathbf{y}_{t-A+1}, \ldots, \mathbf{y}_t) = \mathbf{v}$.
```

so-called doubling trick [16], by running the algorithm from scratch over time intervals of known, exponentially increasing (doubling) lengths. A more preferable way to achieve strong sequentiality is to smoothly modify the algorithm over time while avoiding resets. This can be done by setting $\eta_t$ to depend on $t$ instead of $T$, and by introducing a new-block signal whose length increases over time (independently of the unknown time horizon $T$). At time instant $t+\delta$, the length of the new-block signal is set to $A_t = \left\lceil \frac{\ln t}{\ln M} \right\rceil$, and its symbols are transmitted at fixed time instants $t + \delta = M^{k-1}, k = 1, 2, \ldots$. That is, at time instant $M^{k-1}$, the $k$th symbol $\mathbf{v}_k$ of the new-block signal is selected uniformly at random (independently of any other randomization used beforehand in the coding process) and is transmitted to the decoder as $\mathbf{y}_{M^{k-1}} = \mathbf{v}_k$. The other parts of the coding process skip these time instants, that is, they are not concerned with encoding and decoding source symbols, nor with the transmission or reception of new-block signals. When encoding $x_t$ time instants $M^{k-1} < t < M^k$, the coding scheme uses the length-$A_t$ new-block signal $\mathbf{v}^{A_t} = (\mathbf{v}_1, \ldots, \mathbf{v}_{A_t})$ (note that $A_t = k$ for the selected values of $t$). 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.

*Proof of Theorem 1:* Let $\hat{x}_{(f,g),1}, \ldots, \hat{x}_{(f,g),T}$ denote the reproduction sequence generated by the reference code $(f, g) \in \mathcal{F}$ when applied to the source sequence $x^T$, and let $\tilde{\mathbf{x}}_t = \hat{x}_{(\mathbf{f}^{(t)},\mathbf{g}^{(t)}),t}$. That is, $\tilde{\mathbf{x}}^T$ is the reproduction sequence our coding scheme would generate if it did not have to transmit the identity of the chosen reference decoder, and the correct past $s$ symbols were also available at the decoder (in the current setting when the reference decoder changes we have to wait $s$ channel symbols to have the decoder operating correctly, as it may require $s$ past symbols due to its memory).

Decomposing the cumulative distortion we get

$$\begin{aligned}
&\sum_{t=1}^{T} d_t(x_t, \hat{\mathbf{x}}_t) \\
&= \sum_{t:1 \le t \le T, \hat{\mathbf{x}}_t = \tilde{\mathbf{x}}_t} d_t(x_t, \tilde{\mathbf{x}}_t) + \sum_{t:1 \le t \le T, \hat{\mathbf{x}}_t \ne \tilde{\mathbf{x}}_t} d_t(x_t, \hat{\mathbf{x}}_t) \\
&\le \sum_{t=1}^{T} d_{t,(\mathbf{f}^{(t)},\mathbf{g}^{(t)})}(x^{t+\delta}) + \left|\left\{t : \hat{\mathbf{x}}_t \ne \tilde{\mathbf{x}}_t, 1 \le t \le T\right\}\right|. \quad (5)
\end{aligned}$$

The expectation of the first term can be bounded using Lemma 2 as

$$\begin{aligned}\mathbb{E}&\left[\sum_{t=1}^{T} d_{t,(\mathbf{f}^{(t)},\mathbf{g}^{(t)})}(x^{t+\delta})\right]\\ &\le D^*_{\mathcal{F}}(x^{T+\delta}) + \frac{\ln|\mathcal{F}|}{\eta_T} + \sum_{t=1}^{T}\frac{\eta_t}{8}\\ &= D^*_{\mathcal{F}}(x^{T+\delta}) + \frac{\ln|\mathcal{F}|}{\eta} + \frac{\eta T}{8}. \qquad (6)\end{aligned}$$

It is easy to see that $\hat{\mathbf{x}}_t \neq \tilde{\mathbf{x}}_t$ may happen only at the first $A+B+s$ steps of each block. Indeed, if the mSD algorithm does not change the code to be used in the first $A+B+s$ steps of the block, the receiver becomes completely synchronized and so it decodes $\hat{\mathbf{x}}_t \neq \tilde{\mathbf{x}}_t$ from step $A+B+s+1$. If mSD decides to change the code before step $A+B+s+1$, the length of the block is at most $A+B+s$. New blocks are started at the beginning of the communication, and when either the mSD algorithm decides to start one, or when a new-block signal is transmitted by chance. It may also happen that when the encoder starts to transmit a new-block signal, the receiver encounters an unintentional new-block signal whose last symbols are the first symbols of the just transmitted new-block signal; in this case the new block is started as planned, only this happens with less overhead communication. We consider this case as an "intentionally" started block. Letting $\mathbf{S}_T$ and $\mathbf{N}_T$ denote the number of new blocks, up to time $T$, started "intentionally" by the mSD algorithm (except for the first block) and, respectively, "unintentionally" by chance (starting a new block in step 3g of the encoding algorithm), we have

$$\left|\{t : 1 \le t \le T, \hat{\mathbf{x}}_t \neq \tilde{\mathbf{x}}_t\}\right| \le (\mathbf{S}_T + 1 + \mathbf{N}_T)(A+B+s). \quad (7)$$

$\mathbf{S}_T$ can be bounded by $\eta(T-1)$ using Lemma 3. To bound $\mathbf{N}_T$, the number of blocks started unintentionally, consider the sequence $\tilde{\mathbf{y}}_t = \mathbf{f}_t(x^{t+\delta})$, that is, the sequence of channel symbols generated by the "idealized" coding scheme that does not need to transmit the new-block signals and the identity of the decoder function, nor needs to worry about synchronizing the decoder. Let $\mathbf{n}_t = \mathbb{I}_{\{\mathbf{v}=(\tilde{\mathbf{y}}_{t-A+1},\ldots,\tilde{\mathbf{y}}_t)\}}$ denote the indicator function of the event $\mathbf{v} = (\tilde{\mathbf{y}}_{t-A+1},\ldots,\tilde{\mathbf{y}}_t)$. Then clearly $\mathbf{N}_T \le \sum_{t=A}^{T}\mathbf{n}_t$, since unintentional new blocks may only be started based on $\tilde{\mathbf{y}}^T$. Since $\mathbf{v}$ is independent of $\tilde{\mathbf{y}}^{t-1}$,

$$\mathbb{P}\left[\mathbf{v} = (\tilde{\mathbf{y}}_{t-A+1},\ldots,\tilde{\mathbf{y}}_t)\right] = 1/M^A$$

for any $A \le t \le T$, and so

$$\mathbb{E}\left[\mathbf{N}_T\right] \le (T-A)/M^A. \quad (8)$$

Now taking expectations in (7), the second expression in Lemma 3 and (8) yield

$$\begin{aligned}\mathbb{E}&\left[\left|\{t : 1 \le t \le T,\ \hat{\mathbf{x}}_t \neq \tilde{\mathbf{x}}_t\}\right|\right]\\ &= (A+B+s)\left(\eta(T-1) + 1 + \frac{T-A}{M^A}\right).\end{aligned}$$

Combining the above with (5) and (6) proves the statement of the theorem. ■

## V. Sequential Zero-Delay Scalar Quantization

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], [7], [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 performance of $Q$ is measured by the squared distortion,[5] and the instantaneous distortion of $Q$ for input $x$ is defined as $(x-Q(x))^2$. 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 $\left\{\frac{1}{2K}, \frac{3}{2K}, \ldots, \frac{2K-1}{2K}\right\}$ for some positive integer $K$. Note that the number of quantizers in $\mathcal{Q}_K$ is $|\mathcal{Q}_K| = \binom{K}{M}$. It is shown in [7] 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:* Relative to the reference class $\mathcal{Q}$, the expected normalized distortion redundancy of Algorithm 2 applied to $\mathcal{Q}_{\lfloor\sqrt{T}\rfloor}$ with appropriate parameters satisfies, for any $T \ge 2$,

$$\begin{aligned}\widehat{R}_T \le &\sqrt{\frac{2M\ln T}{T}\left(\frac{17}{8}+\frac{(M+2)\ln T}{2\ln M}\right)}\\ &+\frac{1}{\sqrt{T}} + O\left(\frac{M\ln T}{T}\right)\end{aligned}$$

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

The theorem is obtained as a combination of the EWA-based efficient quantization scheme of [7] with the mSD-based coding scheme of the previous section. Similar results could be obtained by combining the "follow the perturbed leader"-based low-complexity quantization scheme of [10] with a seldom changing version of the "follow the perturbed leader" prediction method recently introduced in [15].

*Proof:* The proof is based on results developed in [7]. It is easy to see that for any quantizer $Q \in \mathcal{Q}$ there exists

[5] More general distortion measures could be considered in the same way as in [13, Section 5].

a quantizer $Q_K \in \mathcal{Q}_K$ such that

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

Thus, in this sense, the class $\mathcal{Q}$ is well approximated by $\mathcal{Q}_K$. Therefore, for any sequence $x^T \in [0,1]^T$,

$$\begin{aligned} &\min_{Q \in \mathcal{Q}_K} \frac{1}{T} \sum_{t=1}^{T} (x_t - Q(x_t))^2 \\ &\leq \min_{Q \in \mathcal{Q}} \frac{1}{T} \sum_{t=1}^{T} (x_t - Q(x_t))^2 + \frac{1}{K}. \end{aligned}$$

Applying Algorithm 2 to the reference class $\mathcal{F} = \mathcal{Q}_K$ we obtain by Corollary 1 that the normalized distortion redundancy relative to the class $\mathcal{Q}_K$ can be bounded as

$$\begin{aligned} &\max_{x^T \in [0,1]^T} \frac{1}{T} \mathbb{E}\left[ \sum_{t=1}^{T} (x_t - \hat{x}_t)^2 - \min_{Q \in \mathcal{Q}_K} \sum_{t=1}^{T} (x_t - Q(x_t))^2 \right] \\ &\leq 2\sqrt{\frac{\ln \binom{K}{M}}{T} \left( \frac{17}{8} + \frac{\ln \left( T \binom{K}{M} \right)}{\ln M} \right)} + O\left( \frac{\ln \left( T \binom{K}{M} \right)}{T} \right), \end{aligned}$$

where we used that the size of the reference class $\mathcal{Q}_K$ is $|\mathcal{Q}_K| = \binom{K}{M}$ and that scalar quantization is memoryless, that is, $s = 0$. Combining the above results and substituting $K = \lfloor \sqrt{T} \rfloor$ gives the performance bound of the theorem, taking into account that $\binom{\lfloor \sqrt{T} \rfloor}{M} \leq T^{M/2}$ and, for all $T > 1$, $1/K = 1/\lfloor \sqrt{T} \rfloor < 1/(\lfloor \sqrt{T} \rfloor - 1) = 1/\sqrt{T} + O(1/T)$.

It is shown in [7] that the random choice of a quantizer according to the EWA prediction algorithm in one time step can be performed with $O(MK^2)$ time and $O(K^2)$ space complexity (in essence, the quantization problem is reduced to the online shortest path problem, as explained in [13]). Applying the same method in our algorithm we obtain the desired complexity results (also note that one quantization step requires $O(\ln K)$ operations). ■

*Remark (Complexity of the algorithm):* One may think that most operations in the implementation of the encoder of Algorithm 2 for the online quantization problem are spent on choosing the quantizer, and since here we only need to choose $O(\sqrt{T})$ quantizers on expectation, the required time complexity may be reduced. However, this is not exactly the case: in each time step of the algorithm, $O(K^2)$ operations are needed to update some weights corresponding to the cumulative distortion of possible cells of the quantizers belonging to $\mathcal{Q}_K$ ($K = \lfloor \sqrt{T} \rfloor$), and $O(MK^2)$ operations are used to randomly choose a quantizer according to EWA. The random choice in mSD whether the previous quantizer $Q_{t-1}$ should be kept at time $t$ only requires to determine the distortion of $Q_{t-1}$ on the last source symbol, since $c_t w_{t,Q_{t-1}} / w_{t-1,Q_{t-1}} = e^{-\eta d(x_{t-1}, Q_{t-1}(x_{t-1}))}$, which can be computed in constant time since $Q_{t-1}(x_{t-1})$ is already known at the beginning of time step $t$. Thus, since the expected number of blocks is bounded by $\eta(T-1) + 2$, and using that $\eta = O(\sqrt{\ln M / T})$ by (4), the overall expected computational complexity of the scheme is $O(TK^2 + MK^2\sqrt{T \ln M})$, which is still $O(T^2)$. However, we can use another trick from [7] to reduce complexity on the price of slightly increasing the distortion redundancy. The idea is that the source sequence $x^T$ can be pre-quantized using a uniform $K$ level quantizer, that is $\overline{x}_t = \widehat{Q}_K(x_t)$, where $\widehat{Q}_K$ is a uniform $K$-level quantizer on $[0,1]$, and the distortions in encoding $\overline{x}^T$ are used in step 3(b)i of the encoder in Algorithm 2. This introduces only a $2/K \leq 2/(\sqrt{T} - 1)$ term in the normalized distortion redundancy, since for any quantizer $Q \in \mathcal{Q}$,

$$\max_{x \in [0,1]} \left| (x - Q(x))^2 - (\overline{x} - Q(\overline{x}))^2 \right| \leq \frac{1}{K}.$$

The advantage of working with $\overline{x}^T$ instead of $x^T$ is that in this case the histogram of $\overline{x}^t$ can be updated in constant time in every time step $t$, and the cell weights can be computed from the histogram in $O(MK^2)$ time whenever a new block starts and a new quantizer has to be chosen by EWA. In this way, since we still need to encode each $x_t$ and $\overline{x}_t$ in $O(\ln K)$ time, the expected total computational complexity of the algorithm becomes $O(T \ln K + MK^2 \sqrt{T \ln M}) = O(T^{3/2} M \sqrt{\ln M})$.

## VI. Extensions

In the previous sections we assumed that the encoder and the decoder communicate over a noiseless channel. Following Matloub and Weissman [11], we can extend the results to the case of stochastic channels with positive error exponents. We assume that the communication channel has finite memory $r$ for some integer $r \geq 0$, and its output also depends on some stationary noise process $\ldots, Z_{-1}, Z_0, Z_1, \ldots$ with known distribution such that if the channel input up to time $t$ is $y^t$ for some $t \geq r$, then the output of the channel is a function of $y_{t-r+1}^t$ and $Z^t$. Moreover, it is assumed that for some rate $R > 0$ there exists a constant $\sigma > 0$ such that for any block length $b$ there exists a channel code $\mathcal{C}_b$ that can discriminate $\lfloor e^{bR} \rfloor$ messages with maximum error probability $e^{-\sigma b}$ in $b$ channel uses. These assumptions are not restrictive and hold for all channels with positive capacity and error exponent.

Formally, denoting the channel input and output alphabet by $\mathcal{M} = \{1, \ldots, M\}$ and $\widehat{\mathcal{M}}$, a delay-$\delta$ sequential joint source-channel code is given by a sequence of encoder-decoder functions $(f, g) = \{f_t, g_t\}_{t=1}^{\infty}$ with $f_t : \mathcal{X}^{t+\delta} \times [0,1]^t \to \mathcal{M}$ and $g_t : \mathcal{M}^t \to \hat{\mathcal{X}}$. Matloub and Weissman [11] used a channel code $\mathcal{C}_b$ (minimizing the maximum error probability) to communicate the decoder function at the beginning of each block, as well as replaced the distortion $d_{t,(f,g)}(x^{t+\delta})$ with its expectation $\bar{d}_{t,(f,g)}(x^{t+\delta}) = \mathbb{E}\left[ d_{t,(f,g)}(x^{t+\delta}) \right]$. Note that the randomness in $d_{t,(f,g)}(x^{t+\delta})$ is only due to its dependence on $Z^t$; in particular, $d_{t,(f,g)}(x^{t+\delta})$ and $\mathbf{U}^t$ are independent. Also note that $\bar{d}_{t,(f,g)}(x^{t+\delta})$ can be computed at the encoder at time step $t + \delta$ since the distribution of $Z^t$ is known. In our case a further modification is needed, as the new-block signal also has to be communicated using channel coding.

Thus, we need to do the following modifications in Algorithm 2 to make it suitable for the joint source-channel coding scenario: First, in step 3(b)i of the encoder in Algorithm 2, $d_{t,(f,g)}(x^{t+\delta})$ has to be replaced with $\bar{d}_{t,(f,g)}(x^{t+\delta})$. Furthermore, during the whole communication process, the new-block signal $\mathbf{v}$ and the indices of the decoder functions $\mathbf{g}$ are transmitted using channel coding, with codes $\mathcal{C}_A$ and $\mathcal{C}_B$, respectively. These codes are used at the decoder to identify the beginning of a new block and determining the decoder function. Accordingly, the new-block signal $\mathbf{v}$ is selected uniformly at random from the set $\{1, 2, \ldots, \lfloor e^{AR} \rfloor\}$, and

$$\lfloor e^{BR} \rfloor \geq |\{g : (f, g) \in \mathcal{F}\}|. \tag{9}$$

Note that before each use of the channel code $C_A$, the encoder uses $r$ symbols to reset the memory of the channel, that is, transmitting the new-block signal actually takes $A + r$ time steps. For simplicity, assuming $r \leq A$, after the receipt of the new block signal, the last $r$ symbols are both known to the encoder and the decoder, so the transmission of the index of the decoder function can be started immediately, using the channel code $C_B$. Furthermore, unlike to the noiseless channel case, the encoder is not able to determine if the decoder would receive a new-block signal by chance, since it depends on the channel noise; therefore, we omit step 3g of the encoding algorithm. While this modification makes the algorithm simpler, it can also ruin its performance if such an accident occurs since the scheme has no built-in method to recover from such an error. However, by a careful selection of the new-block signal, we can guarantee that this disaster happens only with very small, specifically $O(1/T^2)$ probability. Similarly, we set $A$ and $B$ large enough so that the probability of incorrectly decoding a single new-block signal or code index also becomes $O(1/T^2)$. It is straightforward to modify the algorithm so that it can avoid such complete failures by communicating the identity of the decoding schemes in $O(\sqrt{T})$ additional *deterministically chosen* time windows. The analysis of this modified algorithm is straightforward, and is omitted to preserve clarity.

By analyzing the performance of the above coding scheme with appropriately set parameters, the next theorem shows that $O(\sqrt{\ln(T)/T})$ normalized distortion redundancy is achievable in the joint source-channel coding problem:

*Theorem 3:* Let $\mathcal{F}$ be a finite, non-empty class of delay-$\delta$ memory-$s$ sequential joint source-channel codes. Then, for any time horizon $T \geq 1$, under our assumptions on the communication channel with $r \leq \lceil 2 \ln T / \min\{\sigma, R\} \rceil$, the expected normalized distortion redundancy of the above coding scheme relative to $\mathcal{F}$, with appropriate parameter settings, satisfies

$$\widehat{R}_T \leq \sqrt{\frac{\ln |\mathcal{F}|}{T}\left(\left(\frac{4}{\sigma}+\frac{2}{R}\right)\ln T+\frac{\ln(|\mathcal{F}|+1)}{R}+r+s+\frac{17}{8}\right)} + \frac{3}{2T}.$$

*Remark (Wyner-Ziv setting):* Before giving the proof of the theorem, let us discuss the implication of the above result to the Wyner-Ziv setting considered by Reani and Merhav [12]. In this problem there is a noiseless communication channel between the encoder and the decoder, and the decoder also has access to a side information signal that is a noisy observation of the current source symbol $x_t$ through a memoryless channel. This setup can be treated as a special case of the above joint source channel coding problem with a restricted set of encoders and a special channel: the channel is composed of a noiseless part and a noisy side information channel, and each encoder has to transmit the actual source symbol uncoded over the side information channel. In fact, this setup is simpler, as there is no need to use error protection for communicating the indices of the decoders and the new-block signals; however, replacing $d_t$ by $\bar{d}_t$ is still necessary. Thus, the above $O(\sqrt{\ln(T)/T})$ normalized distortion redundancy is also achievable in this case. Moreover, Reani and Merhav also gave an efficient implementation for the zero-delay scalar quantization case based on an efficient implementation of the EWA algorithm. This efficient algorithm can easily be incorporated in our method in the same way as the efficient algorithms for scalar quantization (provided by [7], [13]) were used in Section V.

*Proof of Theorem 3:* The proof follows very closely the proof of Theorem 1, so we will emphasize the differences and skip some details. As in the proof of Theorem 1, defining $\{\hat{\mathbf{x}}^T\}$ to be the real reproduction sequence, and $\tilde{\mathbf{x}}_t = \hat{x}_{(\mathbf{f}^{(t)}, \mathbf{g}^{(t)}), t}$ to be the idealized reproduction sequence, the decomposition (5) of the regret obviously holds in the joint source-channel coding scenario considered. That is,

$$\sum_{t=1}^{T} d_t(x_t, \hat{\mathbf{x}}_t) \leq \sum_{t=1}^{T} d_{t,(\mathbf{f}^{(t)},\mathbf{g}^{(t)})}(x^{t+\delta}) + \left|\left\{t : \hat{\mathbf{x}}_t \neq \tilde{\mathbf{x}}_t, 1 \leq t \leq T\right\}\right|. \tag{10}$$

Since $\left(\mathbf{f}^{(t)}, \mathbf{g}^{(t)}\right)$ are obtained using mSD with the losses $\bar{d}_{t,(f,g)}(x^{t+\delta})$, Lemma 2 implies

$$\mathbb{E}\left[\sum_{t=1}^{T} \bar{d}_{t,(\mathbf{f}^{(t)},\mathbf{g}^{(t)})}(x^{t+\delta})\right] \leq \overline{D}^{*}_{\mathcal{F}}(x^{T+\delta}) + \frac{\ln |\mathcal{F}|}{\eta} + \frac{\eta T}{8} \tag{11}$$

with

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

where $y_t = f_t(x^{t+\delta})$ for all $t \geq 1$. Combining (10) and (11), we see that to bound the distortion redundancy, we need to analyze the expectation of the last term in (10). This term is influenced by the communication overhead for conveying the identity of the decoder function, as well as by errors in the communication, that is, incorrectly determining the blocks and making an error in decoding the identity of the decoder function. Let $\mathcal{B}_t$ denote the event that a new block is started at the encoder at time $t$. Let $\mathcal{E}_{t,nb}$ denote the event that the corresponding new-block signal is decoded incorrectly; and let $\mathcal{E}_{t,i}$ denote the event that the decoder function to be used in the block is determined incorrectly. Note that $\mathcal{E}_{1,nb}$ means that the new-block signal is incorrectly decoded at the beginning of the whole communication process, while, for $t \geq 2$, $\mathcal{E}_{t,nb}$ means that a new-block is not noticed at the decoder given that $\mathcal{E}_{1,nb}$ does not hold (i.e., the new-block signal is correctly known at the decoder). By our assumptions on the channel code, $\mathbb{P}\left[\mathcal{E}_{t,nb}\middle|\mathcal{B}_t\right] \le e^{-\sigma A}$ and $\mathbb{P}\left[\mathcal{E}_{t,i}\middle|\mathcal{B}_t\right] \le e^{-\sigma B}$. The other source of error in the decoding process is the event that the decoder mistakenly declares a new block by decoding the last $A$ symbols seen on the channel by $C_A$ to $\mathbf{v}$. If the decoder correctly identifies all blocks before, the last $A$ symbols are independent of $\mathbf{v}$, and so, as in the noiseless case, the probability of finding a new-block signal when the encoder has not sent one is bounded by $1/\lfloor e^{AR}\rfloor$. Using the pessimistic bound that $\hat{\mathbf{x}}_t \neq \tilde{\mathbf{x}}_t$ after any of the above errors occur, and taking into account that if a block is transmitted correctly, $\hat{\mathbf{x}}_t \neq \tilde{\mathbf{x}}_t$ happens at most in the first $r + A + B + s$ steps of the block, we obtain the following bound:

$$
\begin{aligned}
&\mathbb{E}\left[\left|\left\{t : \hat{\mathbf{x}}_t \neq \tilde{\mathbf{x}}_t, 1 \le t \le T\right\}\right|\right] \\
&\quad\le \sum_{t=1}^{T}(T-t+1)\mathbb{P}\left[\mathcal{B}_t\right]\mathbb{P}\left[\mathcal{E}_{t,nb}\middle|\mathcal{B}_t\right] \\
&\qquad+\sum_{t=1}^{T}(T-t+1)\left(\mathbb{P}\left[\mathcal{B}_t\right]\mathbb{P}\left[\mathcal{E}_{t,i}\middle|\mathcal{B}_t\right]+\frac{1}{\lfloor e^{AR}\rfloor}\right) \\
&\qquad+\sum_{t=1}^{T}\mathbb{P}\left[\mathcal{B}_t\right](r+A+B+s) \\
&\quad\le \sum_{t=1}^{T}(T-t+1)\left(e^{-\sigma A}+e^{-\sigma B}+1/(e^{AR}-1)\right) \\
&\qquad+\mathbb{E}\left[S_T\right](r+A+B+s) \\
&\quad\le \left(e^{-\sigma A}+e^{-\sigma B}+\frac{1}{e^{AR}-1}\right)\frac{T(T-1)}{2} \\
&\qquad+\eta(T-1)(r+A+B+s). \qquad (12)
\end{aligned}
$$

Now selecting

$$
A = \left\lceil \frac{2\ln T}{\min\{\sigma, R\}} \right\rceil
$$

and

$$
B = \left\lceil \max\left\{\frac{2\ln T}{\sigma}, \frac{\ln(|\mathcal{F}|+1)}{R}\right\} \right\rceil
$$

(recall that $B$ has to satisfy condition (9)) ensures that the first term in (12) is bounded by $3/2$, and using $\max\{a, b\} \le a + b$ for $a, b \ge 0$, we get

$$
\begin{aligned}
&\mathbb{E}\left[\left|\left\{t : \hat{\mathbf{x}}_t \neq \tilde{\mathbf{x}}_t, 1 \le t \le T\right\}\right|\right] \\
&\quad\le 3/2+\eta T\left(\left(\frac{4}{\sigma}+\frac{2}{R}\right)\ln T+\frac{\ln(|\mathcal{F}|+1)}{R}+r+s+2\right).
\end{aligned}
$$

Combining this inequality with (10) and (11) we obtain that the expected normalized distortion redundancy can be bounded as

$$
\begin{aligned}
\widehat{R}_T \le\ & \frac{\ln|\mathcal{F}|}{T\eta}+\eta\left(\left(\frac{4}{\sigma}+\frac{2}{R}\right)\ln T+\frac{\ln(|\mathcal{F}|+1)}{R}+r+s+\frac{17}{8}\right) \\
&+\frac{3}{2T}.
\end{aligned}
$$

Optimizing over $\eta$ proves the statement of the theorem. ■

## VII. 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 $O(T^{-1/3})$ results. 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 were also extended to joint source-channel coding and coding with side information at the decoder (the Wyner-Ziv setting).

## Appendix

### A. Proof of Lemma 1

We will use the notation $W_t = \sum_{i\in\mathcal{F}} w_{t,i}$ (note that $W_t \le 1$ for all $t \ge 1$ since $w_{t,i} \le 1/N$). We prove the lemma by induction on $1 \le t \le T$. For $t = 1$, the statement follows from the definition of the algorithm. Now assume that $t \ge 2$ and the hypothesis holds for $t - 1$. We have

$$
\begin{aligned}
\mathbb{P}\left[\mathbf{i}_t = i\right] &= \mathbb{P}\left[\mathbf{i}_{t-1} = i\right] c_t\frac{w_{t,i}}{w_{t-1,i}} \\
&\qquad+p_{t,i}\sum_{j\in\mathcal{F}}\mathbb{P}\left[\mathbf{i}_{t-1} = j\right]\left(1-c_t\frac{w_{t,j}}{w_{t-1,j}}\right) \\
&= p_{t-1,i}c_t\frac{w_{t,i}}{w_{t-1,i}} \\
&\qquad+p_{t,i}\sum_{j\in\mathcal{F}}p_{t-1,j}\left(1-c_t\frac{w_{t,j}}{w_{t-1,j}}\right) \\
&= c_t\frac{w_{t-1,i}}{W_{t-1}}\frac{w_{t,i}}{w_{t-1,i}} \\
&\qquad+\frac{w_{t,i}}{W_t}\sum_{j\in\mathcal{F}}\frac{w_{t-1,j}}{W_{t-1}}\left(1-c_t\frac{w_{t,j}}{w_{t-1,j}}\right) \\
&= c_t\frac{w_{t,i}}{W_{t-1}}+\frac{w_{t,i}}{W_t}-c_t\frac{w_{t,i}}{W_t}\frac{W_t}{W_{t-1}} \\
&= \frac{w_{t,i}}{W_t} = p_{t,i}.
\end{aligned}
$$

■

### B. Proof of Lemma 2

Introduce the following notation:

$$
w'_{t,i} = \frac{1}{N}e^{-\eta_{t-1}D_{t-1,i}},
$$

where $D_{t-1,i} = \sum_{s=1}^{t-1} d_{s,i}$. Note that the difference between $w_{t,i}$ and $w'_{t,i}$ is that $\eta_t$ is replaced by $\eta_{t-1}$ in the latter. We will also use $W'_t = \sum_{i\in\mathcal{F}} w'_{t,i}$. First, we have

$$
\begin{aligned}
\frac{1}{\eta_t}\ln\frac{W'_{t+1}}{W_t} &= \frac{1}{\eta_t}\ln\frac{\sum_{i\in\mathcal{F}} w_{t,i}e^{-\eta_t d_{t,i}}}{W_t} \\
&= \frac{1}{\eta_t}\ln\sum_{i\in\mathcal{F}} p_{t,i}e^{-\eta_t d_{t,i}} \\
&\le -\sum_{i\in\mathcal{F}} p_{t,i}d_{t,i}+\frac{\eta_t}{8} \\
&= -\mathbb{E}\left[d_{t,\mathbf{i}_t}\right]+\frac{\eta_t}{8},
\end{aligned}
$$

where the next-to-last step follows from Hoeffding's inequality (see [16, Lemma A.1])[6] and the fact that $d_{t,i} \in [0,1]$, and the last equality is a consequence of Lemma 1. After rearranging, we get

$$\mathbb{E}\left[d_{t,\mathbf{i}_t}\right] \le -\frac{1}{\eta_t} \ln \frac{W'_{t+1}}{W_t} + \frac{\eta_t}{8}.$$

Rewriting the first term on the right hand side, we obtain

$$\mathbb{E}\left[d_{t,\mathbf{i}_t}\right] \le \left(\frac{\ln W_t}{\eta_t} - \frac{\ln W_{t+1}}{\eta_{t+1}}\right) + \left(\frac{\ln W_{t+1}}{\eta_{t+1}} - \frac{\ln W'_{t+1}}{\eta_t}\right) + \frac{\eta_t}{8}. \qquad (13)$$

The first term can be telescoped as

$$\begin{aligned}\sum_{t=1}^{T} &\left(\frac{\ln W_t}{\eta_t} - \frac{\ln W_{t+1}}{\eta_{t+1}}\right) \\ &= \frac{\ln W_1}{\eta_1} - \frac{\ln W_{T+1}}{\eta_{T+1}} \le -\frac{\ln w_{T+1,i}}{\eta_{T+1}} \\ &= -\frac{1}{\eta_{T+1}} \ln \frac{1}{N} e^{-\eta_{T+1} D_{T,i}} = D_{T,i} + \frac{\ln N}{\eta_{T+1}}, \qquad (14)\end{aligned}$$

for any $i \in \mathcal{F}$, where we used that $W_{T+1} \ge w_{t+1,i}$ and $W_1 = 1$ since $w_{1,j} = 1/N$ by definition for all $j \in \mathcal{F}$. To deal with the second term, observe that

$$\begin{aligned}W_{t+1} &= \sum_{i\in\mathcal{F}} \frac{1}{N} e^{-\eta_{t+1} D_{t,i}} = \sum_{i\in\mathcal{F}} \frac{1}{N} \left(e^{-\eta_t D_{t,i}}\right)^{\frac{\eta_{t+1}}{\eta_t}} \\ &\le \left(\sum_{i\in\mathcal{F}} \frac{1}{N} e^{-\eta_t D_{t,i}}\right)^{\frac{\eta_{t+1}}{\eta_t}} = \left(W'_{t+1}\right)^{\frac{\eta_{t+1}}{\eta_t}},\end{aligned}$$

where we applied Jensen's inequality to the concave function $x^{\frac{\eta_{t+1}}{\eta_t}}$, $x \in \mathbb{R}$ (the latter function is concave since $\eta_{t+1} \le \eta_t$ by our assumptions). Taking logarithms in the above inequality, we obtain

$$\frac{\ln W_{t+1}}{\eta_{t+1}} - \frac{\ln W'_{t+1}}{\eta_t} \le 0.$$

This shows that the second term on the right hand side of (13) is non-positive. Thus, summing up (13) for all $t = 1, 2, \ldots, T$ and using (14) we obtain

$$\sum_{t=1}^{T} \mathbb{E}\left[d_{t,\mathbf{i}_t}\right] \le D_{T,i} + \sum_{t=1}^{T} \frac{\eta_t}{8} + \frac{\ln N}{\eta_{T+1}}.$$

Finally, since the losses $d_{t,i}, i \in \mathcal{F}$ and $d_{t,\mathbf{i}_t}$ do not depend on $\eta_{T+1}$ for $t \le T$, we can choose, without loss of generality $\eta_{T+1} = \eta_T$, and the statement of the lemma follows. ■

### C. Proof of Lemma 3

The probability of switching experts in step $t \ge 2$ is

$$\begin{aligned}\alpha_t &\overset{\text{def}}{=} \mathbb{P}\left[\mathbf{i}_{t-1} \ne \mathbf{i}_t\right] \\ &= \sum_{i\in\mathcal{F}} \mathbb{P}\left[\mathbf{i}_{t-1} = i\right] \left(1 - c_t \frac{w_{t,i}}{w_{t-1,i}}\right) \left(1 - p_{t,i}\right) \\ &\le \sum_{i\in\mathcal{F}} \mathbb{P}\left[\mathbf{i}_{t-1} = i\right] \left(1 - c_t \frac{w_{t,i}}{w_{t-1,i}}\right) \\ &= 1 - \sum_{i\in\mathcal{F}} \frac{w_{t-1,i}}{W_{t-1}} c_t \frac{w_{t,i}}{w_{t-1,i}} \\ &= 1 - c_t \frac{W_t}{W_{t-1}},\end{aligned}$$

where the next-to-last equality is due to Lemma 1. Reordering gives $W_t \le \frac{1-\alpha_t}{c_t} W_{t-1}$ and thus

$$W_T \le W_1 \prod_{t=2}^{T} \frac{1-\alpha_t}{c_t} = \prod_{t=2}^{T} \frac{1-\alpha_t}{c_t}.$$

On the other hand,

$$W_T \ge \max_{j\in\mathcal{F}} w_{T,j} = \max_{j\in\mathcal{F}} \frac{1}{N} e^{-\eta_T D_{T-1,j}} = \frac{1}{N} e^{-\eta_T D^*_{T-1}},$$

where $D^*_{T-1} = \min_{j\in\mathcal{F}} D_{T-1,j}$. Taking logarithms of both inequalities and putting them together, we get

$$-\ln N - \eta_T D^*_{T-1} \le \sum_{t=2}^{T} \ln(1-\alpha_t) - \sum_{t=2}^{T} \ln c_t.$$

Now using $\ln(1-x) \le -x$ for all $x \in [0,1)$ and $0 \le \alpha_t < 1$, we obtain

$$\mathbb{E}\left[\mathbf{S}_T\right] = \sum_{t=2}^{T} \alpha_t \le \eta_T D^*_{T-1} + \ln N - \sum_{t=2}^{T} \ln c_t.$$

Now the statement of the lemma for the first expression in the minimum in (1) follows since

$$-\sum_{t=2}^{T} \ln c_t = \sum_{t=2}^{T} (\eta_{t-1} - \eta_t)(t-2) = \sum_{t=2}^{T-1} (\eta_t - \eta_T).$$

To prove that the second expression in (1) is also an upper bound on the expected number of switches, we start with the following bound:

$$\begin{aligned}\mathbb{P}\left[\mathbf{i}_t = \mathbf{i}_{t-1}\right] &= \sum_{i\in\mathcal{F}} \mathbb{P}\left[\mathbf{i}_t = i \,|\, \mathbf{i}_{t-1} = i\right] \mathbb{P}\left[\mathbf{i}_{t-1} = i\right] \\ &\ge \sum_{i\in\mathcal{F}} c_t \frac{w_{t,i}}{w_{t-1,i}} p_{t-1,i}.\end{aligned}$$

The elements in the sum can be bounded as

$$\begin{aligned}c_t \frac{w_{t,i}}{w_{t-1,i}} &= e^{\eta_{t-1} D_{t-2,i} - \eta_t D_{t-1,i} + (\eta_t - \eta_{t-1})(t-2)} \\ &= e^{(\eta_{t-1} - \eta_t) D_{t-2,i} - \eta_t d_{t-1,i} + (\eta_t - \eta_{t-1})(t-2)} \\ &\ge e^{(\eta_t - \eta_{t-1})(t-2) - \eta_t} \\ &\ge 1 - \eta_t + \left(\eta_t - \eta_{t-1}\right)(t-2),\end{aligned}$$

[6] Hoeffding's inequality states that if $X$ is a random variable with $a \le X \le b$ then the inequality $\ln\left[e^{sX}\right] \le s\mathbb{E}[X] + s^2(b-a)^2/8$ holds for any real number $s$ (see [16, Lemma A.1]). The inequality is applied for a random variable $X$ with distribution $\mathbb{P}\left[X = -d_{t,i}\right] = p_{t,i}$.

for all $i \in \mathcal{F}$, where we used $1 + x \leq e^x$. Thus, using that $\sum_{i \in \mathcal{F}} p_{t-1,i} = 1$, we obtain

$$\mathbb{P}\left[\mathbf{i}_t \neq \mathbf{i}_{t-1}\right] \leq \eta_t + (\eta_{t-1} - \eta_t)(t-2).$$

Summing up for $t = 2, \ldots, T$ gives

$$\begin{aligned}\mathbb{E}\left[\mathbf{S}_T\right] &= \sum_{t=2}^{T} \mathbb{P}\left[\mathbf{i}_t \neq \mathbf{i}_{t-1}\right] \\ &\leq \sum_{t=2}^{T} \eta_t + \sum_{t=2}^{T} (\eta_{t-1} - \eta_t)(t-2) \\ &= \sum_{t=2}^{T} \eta_t + \sum_{t=2}^{T-1} (\eta_t - \eta_T) = \sum_{t=2}^{T} (2\eta_t - \eta_T),\end{aligned}$$

completing the proof. ■

## Acknowledgment

The authors would like to thank the anonymous reviewers for their comments that helped improve the presentation of the material.

## References

[1] T. Linder and G. Lugosi, “A zero-delay sequential scheme for lossy coding of individual sequences,” *IEEE Trans. Inf. Theory*, vol. 47, no. 6, pp. 2533–2538, Sep. 2001.

[2] V. Vovk, “Aggregating strategies,” in *Proc. 3rd Annu. Workshop Comput. Learn. Theory*, Rochester, NY, USA, Aug. 1990, pp. 372–383.

[3] V. Vovk, “A game of prediction with expert advice,” *J. Comput. Syst. Sci.*, vol. 56, no. 2, pp. 153–173, 1998.

[4] N. Littlestone and M. K. Warmuth, “The weighted majority algorithm,” *Inf. Comput.*, vol. 108, no. 2, pp. 212–261, 1994.

[5] T. Weissman and N. Merhav, “On limited-delay lossy coding and filtering of individual sequences,” *IEEE Trans. Inf. Theory*, vol. 48, no. 3, pp. 721–733, Mar. 2002.

[6] E. Takimoto and M. K. Warmuth, “Path kernels and multiplicative updates,” in *Proc. 15th Annu. Conf. Comput. Learn. Theory (COLT)*, Jul. 2002, pp. 74–89.

[7] A. György, T. Linder, and G. Lugosi, “Efficient adaptive algorithms and minimax bounds for zero-delay lossy source coding,” *IEEE Trans. Signal Process.*, vol. 52, no. 8, pp. 2337–2347, Aug. 2004.

[8] J. Hannan, “Approximation to bayes risk in repeated plays,” in *Contributions to the Theory of Games*, vol. 3, M. Dresher, A. Tucker, and P. Wolfe, Eds. Princeton, NJ, USA: Princeton Univ. Press, 1957, pp. 97–139.

[9] A. Kalai and S. Vempala, “Efficient algorithms for the online decision problem,” in *Proc. 16th Annu. Conf. Learn. Theory and the 7th Kernel Workshop, COLT-Kernel 2003*, Aug. 2003, pp. 26–40.

[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 Compress. Conf.*, Washington, DC, USA, Mar. 2004, pp. 342–351.

[11] S. Matloub and T. Weissman, “Universal zero delay joint source-channel coding,” *IEEE Trans. Inf. Theory*, vol. 52, no. 12, pp. 5240–5250, Dec. 2006.

[12] A. Reani and N. Merhav, “Efficient on-line schemes for encoding individual sequences with side information at the decoder,” *IEEE Trans. Inf. Theory*, vol. 57, no. 10, pp. 6860–6876, Oct. 2011.

[13] A. György, T. Linder, and G. Lugosi, “Tracking the best Quantizer,” *IEEE Trans. Inf. Theory*, vol. 54, no. 4, pp. 1604–1625, Apr. 2008.

[14] S. Geulen, B. Voecking, and M. Winkler, “Regret minimization for online buffering problems using the weighted majority algorithm,” in *Proc. 23rd Annu. Conf. Learn. Theory (COLT)*, 2010, pp. 132–143.

[15] L. Devroye, G. Lugosi, and G. Neu, “Prediction by random-walk perturbation,” in *Proc. 26th Annu. Conf. Learn. Theory (COLT)*, vol. 30. 2013, pp. 460–473.

[16] N. Cesa-Bianchi and G. Lugosi, *Prediction, Learning, and Games*. Cambridge, U.K.: Cambridge Univ. Press, 2006.

[17] L. Györfi and G. Ottucsák, “Sequential prediction of unbounded stationary time series,” *IEEE Trans. Inf. Theory*, vol. 53, no. 5, pp. 866–1872, May 2007.

[18] A. V. Chernov and F. Zhdanov, “Prediction with expert advice under discounted loss,” in *Proc. 21st Int. Conf. Alg. Learn. Theory (ALT)*, 2010, pp. 255–269.

**András György** (S’01–A’03–M’04) received the M.Sc. (Eng.) degree (with distinction) in technical informatics from the Technical University of Budapest, in 1999, the M.Sc. (Eng.) degree in mathematics and engineering from Queen’s University, Kingston, ON, Canada, in 2001, and the Ph.D. degree in technical informatics from the Budapest University of Technology and Economics in 2003.

He was a Visiting Research Scholar in the Department of Electrical and Computer Engineering, University of California, San Diego, USA, in the spring of 1998. In 2002–2011 he was with the Computer and Automation Research Institute of the Hungarian Academy of Sciences, where, from 2006, he was a Senior Researcher and Head of the Machine Learning Research Group. In 2003–2004 he was also a NATO Science Fellow in the Department of Mathematics and Statistics, Queen’s University. He also held a part-time research position at GusGus Capital Llc., Budapest, Hungary, in 2006–2011. Since 2012 he has been with the Department of Computing Science, University of Alberta, Edmonton, AB, Canada. His research interests include machine learning and information theory.

Dr. György received the Gyula Farkas prize of the János Bolyai Mathematical Society in 2001 and the Academic Golden Ring of the President of the Hungarian Republic in 2003.

**Gergely Neu** received the M.Sc. degree in Electrical Engineering and the Ph.D. degree in Technical Informatics from the Budapest University of Technology and Ecomomics (Hungary) in 2008 and 2013, respectively.

In 2006–2013, he was a Junior Researcher at the Computer and Automation Research Institute of the Hungarian Academy of Sciences. Between 2008 and 2013, he has spent multiple time periods spanning 10 months as a Visiting Researcher at the Department of Computing Science, University of Alberta, Edmonton, AB, Canada. Since 2013, he is a Marie Curie postdoctoral fellow at the SequeL team of INRIA Lille-Nord Europe. His research interests include reinforcement learning, online learning, and bandit problems.

In 2009, Dr. Neu received the Pro Scientia Gold Medal for his excellent research activity from the president of the Hungarian Academy of Sciences.