On Limited-Delay Lossy Coding and Filtering of Individual Sequences*

Tsachy Weissman Neri Merhav

December 11, 2000

Abstract

We continue the study of adaptive schemes for the sequential lossy coding of individual sequences which was recently initiated by Linder and Lugosi. Specifically, we consider fixed-rate lossy coding systems of fixed (or zero) delay where the encoder (which is allowed to use randomization) and the decoder are connected via a noiseless channel of a given capacity. It is shown that for any finite set of such coding schemes of a given rate, there exists a source code (adhering to the same structural and delay limitations) with the same rate whose distortion is with high probability almost as small as that of the best scheme in that set, uniformly for all individual sequences. Applications of this result to reference classes of special interest are outlined. These include the class of scalar quantizers, trellis encoders with sliding block decoders, and DPCM-based source codes. In particular, for the class of all scalar quantizers, a source code is obtained with redundancy of order $n^{-1/3}\log n$. This improves the $n^{-1/5}\log n$ rate achieved by Linder and Lugosi. More importantly, the decoder here is deterministic and, in particular, does not assume a common randomization sequence available at both encoder and decoder. Finally, we consider the case where the individual sequence is corrupted by noise prior to reaching the coding system, whose goal now is to reconstruct a sequence with small distortion relative to the clean individual sequence. It is shown that for the case of a finite alphabet and an invertible channel transition probability matrix, for any finite set of sliding window schemes of a given rate, there exists a source code (allowed to use randomization yet adhering to the same delay constraints) whose performance is, with high probability, essentially as good as the best scheme in the class, for all individual sequences.

1 Introduction

The requirement for a limited decoding delay arises naturally in an increasing variety of coding applications. A standard model for a fixed-rate (R) source code with limited delay is the following. The source sequence $x_1, x_2, \ldots$ is transformed into channel symbols $y_1, y_2, \ldots$ taking values in $\{1, 2, \ldots, M\}$, $M = 2^R$, which are, in turn, transformed into a reconstruction sequence $\hat{x}_1, \hat{x}_2, \ldots$. The encoder-decoder is said to have overall delay of no more than

*Both authors are with the Department of Electrical Engineering, Technion- Israel Institute of Technology, Haifa 32000, Israel. tsachy@tx.technion.ac.il, merhav@ee.technion.ac.il.

d if each channel symbol y_n depends only on $x_1, \ldots, x_{n+d_1}$ and each reconstruction symbol $\hat{x}_n$ depends only on $y_1, \ldots, y_{n+d_2}$ such that $d_1 + d_2 \leq d$.

In the first part of this work, we consider such lossy source codes of limited delay in an individual sequence setting. Namely, the source sequence $x_1, x_2, \ldots$ is considered deterministic and statements are made which hold uniformly for all possible sequences. This setting naturally models situations where delay is crucial (e.g., cellular telephony, teleconferencing), yet very little is known about the source. Given a family of delay-constrained source codes (a reference class), our goal is to construct a code, adhering to the same constraint, which does essentially as well (i.e., has the same rate and incurs essentially the same average distortion) as the best code in the family, uniformly over all individual sequences. It is shown that this goal is achievable for an arbitrary finite reference class of delay-constrained codes having a finite-memory decoder. It is also illustrated how this result can be specialized to deal with infinite, parametrizable reference classes of codes, by obtaining coding schemes which compete with reference classes of particular practical interest.

The subsequent part of this work is dedicated to the case where the individual sequence is corrupted by a noisy memoryless channel (i.i.d. noise) prior to reaching the coder. I.e., the source sequence $x_1, x_2, \ldots$ is fed into a memoryless channel whose output is $Z_1, Z_2, \ldots$, and only the latter sequence is available to the encoder. This is a setting of practical interest which naturally models audio and imaging applications in which the underlying signal (which has no natural probabilistic model and, hence, is considered an individual sequence) is corrupted by noise, and delay limitations are inherent in the application. Typical schemes involving predictive coding of noisy images, for example, can be formulated to fall within this setting for zero delay.

It may be intuitively expected that the two goals associated with such a noisy scenario, namely, filtering and compression, will be in concord. This fact is well-known and has been made precise and exploited in the probabilistic context (without delay limitations) cf. [4, 10, 13]. This intuition is consolidated for the individual sequence setting as well by (constructively) establishing the existence of a code for this noisy scenario, which is guaranteed to do essentially as well as the best in a reference class. That is, the combined filtering-compression goal can be efficiently achieved in the individual sequence setting as well. A reference class which will be explicitly treated in this context is one that consists of time-invariant sliding-window encoder-decoder pairs, with the implication that other classes, possibly of codes having somewhat different structural properties, can be handled using similar ideas and tools.

The idea behind the construction of such codes is, in essence, quite similar to that underlying efficient prediction schemes for individual sequences: "Tracking" or trying to "imitate" those schemes in the reference class which have been proven efficient on the past sequence by exponentially weighting the extent to which each scheme is followed, according to its past performance. The implementation of such an approach in the lossy coding situation of the present setting, however, is not as straightforward as in the prediction setting. The main reason for that, in the noise-free setting, is the fact that the decoder, which accesses the reconstructed sequence only (and does not know the source sequence) does not have a precise picture of the past performance of each of the schemes in the class. In the noisy case, an additional level of difficulty is due to the fact that even the encoder, which accesses the noisy sequence, does not know the losses associated with the schemes in the reference class, as these depend on the unseen underlying clean individual sequence. As will be shown, however, this difficulty can be alleviated by employing suitable estimators for the unobserved cumulative distortion associated with the schemes in the reference class.

The idea of harnessing the exponential weighting approach to the present setting of delay-limited coding was instigated by Linder and Lugosi [11], where the study of zero-delay lossy source coding in the individual sequence setting was initiated. The main result of that work was a construction of a delayless sequential adaptive coding scheme, for bounded, real-valued individual sequences, which asymptotically achieves the average distortion of the best scalar quantizer matched to the sequence. The basic idea underlying the coding schemes that we construct here is similar to that presented in [11] in that both are based on the exponential weighting principle. There are, however, some essential differences which, among other things, eliminate the need for the availability of the common randomization sequence assumed in [11]. This point is further elaborated on in Section 4. The setting of the present work can be considered a generalization of that of [11] in several directions: Any finite delay (not necessarily zero) is allowed, richer and more general reference classes (than that of scalar quantizers), and the case where the original data sequence is corrupted by noise.

The remainder of the paper is organized as follows. In Section 2, a formal description of the problem for the noise-free setting is given. Section 3 is dedicated to a generic result which constructively establishes the existence of a delay-d code having essentially the same distortion as the best in a given finite family of delay-d codes. In Section 4, the result of Section 3 is applied to specific reference classes of practical interest. In particular, for the class of scalar quantizers (subsection 4.A), a delayless source code is obtained which, as will be discussed, is superior to that of [11]. Finally, Section 5 is devoted to the setting where the source sequence is corrupted by noise: Subsection 5.A formalizes the problem and subsection 5.B is dedicated to the construction and performance analysis of a delay-limited source code (or filtering scheme) which does essentially as well as the best sliding window encoder-decoder pair in a given set.

2 Problem Formulation For the Noise-Free Setting

Throughout the paper, for any integers $m \leq n$, we let a_m^n denote the vector $(a_m, \ldots, a_n)$ and $a^n = a_1^n$. Equalities and inequalities between random variables, when not explicitly specified, should be understood in the almost sure sense.

A delay-d (d a non-negative integer) sequential source code of fixed rate $R = \log M$ with a randomized encoder is given by a pair (E, D). The randomized encoder E is given by a sequence $\{E_i\}_{i=1}^{\infty}$, where $E_i : \mathcal{X}^{i+d} \times [0,1]^i \to \{1, 2, \ldots, M\}$, $\mathcal{X}$ being the source alphabet. The decoder D is given by a sequence $\{D_i\}_{i=1}^{\infty}$, where $D_i : \{1, 2, \ldots, M\}^i \to \hat{\mathcal{X}}$, $\hat{\mathcal{X}}$ being the reproduction alphabet. The source code operates as follows. The encoder produces the i-th channel symbol $y_i \in \{1, 2, \ldots, M\}$ based on x^{i+d} and on the random sequence U^i according to $y_i = E_i(x^{i+d}, U^i)$, where $\{U_i\}_{i=1}^{\infty}$ is a randomization sequence of i.i.d. random variables, uniformly distributed on $[0,1]$. The decoder emits the reconstructed sequence $\hat{x}_1, \hat{x}_2, \ldots$ according to $\hat{x}_i = D_i(y^i)$. We let $\mathcal{F}^d(R)$ denote the class of all such source codes[1]. We shall, in the sequel, assume a given fixed rate R and simply write $\mathcal{F}^d$.

The cumulative distortion of a source code $(E, D) \in \mathcal{F}^d$ is denoted by

$$d_{(E,D)}^n(\mathbf{x}) = \sum_{i=1}^{n} d(x_i, \hat{x}_i), \tag{1}$$

where $d : \mathcal{X} \times \hat{\mathcal{X}} \to [0, B]$ is a bounded distortion measure ($B < \infty$) and the $\hat{x}_i$'s on the right hand side are generated by feeding the sequence $\mathbf{x} = x_1, x_2, \ldots$ into (E, D) as described above. Note that, though the dependence is suppressed in the notation, $d_{(E,D)}^n(\mathbf{x})$ is a

[1]Note that there is no loss in restricting the decoder to "causality" since any finite delay source code with a non-causal decoder (of finite and bounded horizon) can be represented by an equivalent source code in $\mathcal{F}^d(R)$ for some d, by a time-shift.

random variable which depends on the realization of the randomization sequence U^n. We similarly denote, for $n_1 \le n_2$, $d_{(E,D)}^{n_1,n_2}(\mathbf{x}) = \sum_{i=n_1}^{n_2} d(x_i, \hat{x}_i)$.

A decoder D is said to be of finite-memory $s \ge 0$ if for all $i \ge s$ and all $y^i, z^i \in \hat{\mathcal{X}}^i$ such that $y_{i-s}^i = z_{i-s}^i$, $D_i(y^i) = D_i(z^i)$. We let $\mathcal{F}_s^d(R)$ denote the class of all source codes in $\mathcal{F}^d(R)$ with a decoder of finite-memory s. Note that, similarly to a finite-memory decoder, one can define a finite-*state* decoder (cf. [17]). While, admittedly, not every finite-state encoder is a finite-memory encoder, one can show, using the techniques of [5, 12], that finite-memory machines perform asymptotically as well as finite-state machines. Thus, the classes $\mathcal{F}_s^d(R)$ are relevant for the modeling of a variety of coding schemes in applications which require a finite (or zero) delay. In many such situations, practically any randomized encoder (situated, e.g., in some base station where algorithmic resources are abundant) can be implemented, yet the decoder (situated, e.g., in some small and low cost hand set) has no access to a randomization sequence and is limited in memory or in algorithmic resources.

3 Generic Result

We dedicate this section to the construction of a finite-delay coding scheme which competes with an arbitrary finite set of limited-delay schemes in the sense of operating at the same rate, and suffering a cumulative distortion which is, at most, negligibly higher than that of the best in the set, for all individual sequences. More precisely, we have the following:

Theorem 1 *Let $\mathcal{A}$ be a finite subset of $\mathcal{F}_s^d(R)$ for some $s \ge 0$, $d \ge 0$ and $R = \log M$. Then there exists a source code $(E, D) \in \mathcal{F}^d(R)$ such that for all $\mathbf{x} \in \mathcal{X}^\infty$, any $\varepsilon \ge 0$ and all N sufficiently large such that $N > C_1[(\log|\mathcal{A}|)(\log(|\mathcal{A}|e^{sR}))]/\varepsilon^3$*

$$\mathbb{E}\left\{\frac{1}{N}\left[d_{(E,D)}^N(\mathbf{x}) - \min_{(E',D')\in\mathcal{A}} d_{(E',D')}^N(\mathbf{x})\right]\right\} \le C_2[(\log|\mathcal{A}|)(\log(|\mathcal{A}|e^{sR}))]^{1/3} \cdot N^{-1/3}, \quad (2)$$

and

$$\Pr\left\{\frac{1}{N}\left[d_{(E,D)}^N(\mathbf{x}) - \min_{(E',D')\in\mathcal{A}} d_{(E',D')}^N(\mathbf{x})\right] \ge \varepsilon\right\} \le \exp\left\{-C_3\left(\frac{\log|\mathcal{A}|}{\log(|\mathcal{A}|e^{sR})}\right)^{1/3} \varepsilon^2 N^{2/3}\right\}, \quad (3)$$

where C_1, C_2, C_3 are positive constants which depend only on B and R.

Discussion: The explicit values of the constants C_1, C_2, C_3 appearing in the above theorem will be apparent in the proof. We note that the coding scheme suggested in the above theorem is a member of $\mathcal{F}^d(R)$. In particular, its delay is no more than those of the schemes in the reference class and it has *exactly* the same rate R (and not even a bit more as is sometimes the case in related universal coding scenarios).

We also remark that, while in most probabilistic contexts it is usually satisfactory and informative enough to make statements regarding the *expected* performance (distortion) of a source code, in the individual sequence setting considered here this is not the case. The whole point of the individual sequence setting is to have a complete picture of what is really happening (actual rather than expected distortion) for *every* possible sequence. This is why a point has been made to obtain the concentration inequality (3), which guarantees the actual performance of the source code, in addition to (2).

The source code (E, D) in the above theorem depends on the length N of the individual sequence to be encoded (this is referred to as "horizon-dependence" in the prediction literature). While in some applications (e.g., image coding) the length of the sequence is indeed known in advance, in others it may be desirable to have a source code guaranteed to be doing well at *all* points along the sequence (this is referred to as the "strong sequentiality" property). We merely remark here that it is a straightforward exercise to obtain a strongly sequential source code, for which (2) and (3) hold for *all* sufficiently large N. The Borel-Cantelli lemma can then be applied to obtain an almost sure performance guarantee for such a source code. This is done by employing the horizon-dependent source code of Theorem 1 on blocks of exponentially increasing lengths (cf., e.g., [1, 15] for typical examples of such constructions).

Proof: Fix an $l \ll N$ and divide the time-axis, $i = 1, 2, \ldots, N$, into $n = N/l$ consecutive non-overlapping blocks (assume l divides N). We construct the source code $(E, D) \in \mathcal{F}^d(R)$ as follows. At the beginning of the k-th block $1 \leq k \leq n$, i.e. at the $i = (k-1)l + 1$-th channel use, when x^{i+d} and U^i are available, the encoder uses U_i to generate $(E^{(k)}, D^{(k)})$, a $\mathcal{A}$-valued random variable with distribution satisfying almost surely

$$\Pr\{(E^{(k)}, D^{(k)}) = (E', D') | U_1^{(k-1)l}\} = \frac{\exp\left\{-\eta d_{(E',D')}^{(k-1)l}(\mathbf{x})\right\}}{\sum_{(\tilde{E},\tilde{D})\in\mathcal{A}} \exp\left\{-\eta d_{(\tilde{E},\tilde{D})}^{(k-1)l}(\mathbf{x})\right\}} \quad \forall (E', D') \in \mathcal{A}, \quad (4)$$

where $\eta > 0$ is a parameter to be chosen later. The conditioning on $U_1^{(k-1)l}$ on the left side of (33) is necessary due to the fact that the cumulative distortion of each of the schemes in the reference class which, by the definition of $\mathcal{F}_s^d(R)$, are allowed to use the randomization sequence, will, in general, depend on the realization of the randomization sequence. The encoder, operating at the same rate R, now dedicates the first $\lceil \frac{1}{R}\log|\mathcal{A}|\rceil$ channel symbols at the beginning of the k-th block, i.e., y_i for $i = (k-1)l+1,\ldots,(k-1)l+\lceil\frac{1}{R}\log|\mathcal{A}|\rceil$, to convey to the decoder the identity of $D^{(k)}$. At the remainder of the block, i.e., at times $i = (k-1)l+\lceil\frac{1}{R}\log|\mathcal{A}|\rceil+1,\ldots,kl$, the encoder produces the channel symbols $y_i = E_i^{(k)}(x^{i+d},U^i)$. At the same time, on the decoder's side, at the beginning of the block at times $i = (k-1)l+1,\ldots,(k-1)l+\lceil\frac{1}{R}\log|\mathcal{A}|\rceil+s-1$, the decoder outputs an arbitrary reproduction sequence of $\hat{x}_i$'s. From time $i = (k-1)l+\lceil\frac{1}{R}\log|\mathcal{A}|\rceil+s$ up to the end of the block $i = kl$, the decoder, knowing the identity of $D^{(k)}$ and the output of $E^{(k)}$ at least $s-1$ channel uses back, outputs the reproduction sequence according to $\hat{x}_i = D_i^{(k)}(y^i) = D_i^{(k)}(y_{i-s}^i)$ (the second equality is due to the fact that $(E^{(k)},D^{(k)})\in\mathcal{F}_s^d$). To establish equations (2) and (3), we use some standard ingredients from the theory of prediction of individual sequences (cf., e.g., [2, Theorem 1]). Define for each $k\geq 1$ the random variable

$$W_k = \sum_{(\tilde{E},\tilde{D})\in\mathcal{A}} \exp\left\{-\eta d_{(\tilde{E},\tilde{D})}^{(k-1)l}(\mathbf{x})\right\}. \tag{5}$$

We then have for $n = N/l$

$$\begin{aligned}
\log\frac{W_{n+1}}{W_1} &= \log\sum_{(\tilde{E},\tilde{D})\in\mathcal{A}}\exp\left\{-\eta d_{(\tilde{E},\tilde{D})}^{nl}(\mathbf{x})\right\} - \log|\mathcal{A}| \\
&\geq \log\left(\max_{(\tilde{E},\tilde{D})\in\mathcal{A}}\exp\left\{-\eta d_{(\tilde{E},\tilde{D})}^{nl}(\mathbf{x})\right\}\right) - \log|\mathcal{A}| \\
&= -\eta\min_{(\tilde{E},\tilde{D})\in\mathcal{A}} d_{(\tilde{E},\tilde{D})}^{N}(\mathbf{x}) - \log|\mathcal{A}|.
\end{aligned}\tag{6}$$

On the other hand, for each $1\leq k\leq n$

$$\begin{aligned}
\log\frac{W_{k+1}}{W_k} &= \log\frac{\sum_{(\tilde{E},\tilde{D})\in\mathcal{A}}\exp\left\{-\eta d_{(\tilde{E},\tilde{D})}^{(k-1)l+1,kl}(\mathbf{x})\right\}\cdot\exp\left\{-\eta d_{(\tilde{E},\tilde{D})}^{(k-1)l}(\mathbf{x})\right\}}{\sum_{(\tilde{E},\tilde{D})\in\mathcal{A}}\exp\left\{-\eta d_{(\tilde{E},\tilde{D})}^{(k-1)l}(\mathbf{x})\right\}} \\
&= \log\mathbb{E}_{Q_k}\exp\left\{-\eta d_{(\tilde{E},\tilde{D})}^{(k-1)l+1,kl}(\mathbf{x})\right\} \\
&\leq -\eta\mathbb{E}_{Q_k} d_{(\tilde{E},\tilde{D})}^{(k-1)l+1,kl}(\mathbf{x}) + \frac{\eta^2 l^2 B^2}{8} \\
&\leq -\eta\left\{\mathbb{E}\left[d_{(E,D)}^{(k-1)l+1,kl}(\mathbf{x})|U^{(k-1)l}\right] - B\left(\frac{1}{R}\log|\mathcal{A}|+s\right)\right\} + \frac{\eta^2 l^2 B^2}{8},
\end{aligned}\tag{7}$$

where $\mathbb{E}_{Q_k}$ denotes expectation with respect to the (random) distribution Q_k on $\mathcal{A}$ which assigns a probability proportional to $\exp\left\{-\eta d_{(\tilde{E},\tilde{D})}^{(k-1)l}(\mathbf{x})\right\}$ to each $(\tilde{E},\tilde{D})\in\mathcal{A}$. The first inequality follows from an application of Hoeffding's bound (cf. [3, Lemma 8.1]). The second inequality follows from the construction of the source code (E, D) described above (note that the cumulative distortion of (E, D) from the beginning of the k-th block up to time $i = (k-1)l + \lceil \frac{1}{R} \log |\mathcal{A}| \rceil + s - 1$ can be no more than $B\left(\frac{1}{R}\log|\mathcal{A}| + s\right)$ and from that time up to the end of the block it is exactly the loss of the pair $(E^{(k)}, D^{(k)})$ generated at the beginning of the block). Summing up over k we obtain almost surely

$$\log \frac{W_{n+1}}{W_1} \leq -\eta \sum_{k=1}^{n} \mathbb{E}\left[d_{(E,D)}^{(k-1)l+1,kl}(\mathbf{x})|U^{(k-1)l}\right] + \eta B n \left(\frac{1}{R}\log|\mathcal{A}| + s\right) + \frac{\eta^2 l^2 B^2 n}{8}. \quad (8)$$

Combining equations (8) and (6) gives almost surely

$$\begin{aligned} & \sum_{k=1}^{n} \mathbb{E}\left[d_{(E,D)}^{(k-1)l+1,kl}(\mathbf{x})|U^{(k-1)l}\right] - \min_{(\tilde{E},\tilde{D})\in\mathcal{A}} d_{(\tilde{E},\tilde{D})}^{N}(\mathbf{x}) \\ \leq\ & \frac{\log|\mathcal{A}|}{\eta} + \frac{\eta l^2 B^2 n}{8} + Bn\left(\frac{1}{R}\log|\mathcal{A}| + s\right) \\ =\ & B\sqrt{\log|\mathcal{A}|/2}\, l n^{1/2} + Bn\left(\frac{1}{R}\log|\mathcal{A}| + s\right), \end{aligned} \quad (9)$$

where the equality follows upon taking the minimizing value $\eta = \sqrt{8\log|\mathcal{A}|/(l^2B^2n)}$. For notational convenience, we now denote $\alpha = B\sqrt{\log|\mathcal{A}|/2}$ and $\beta = B\left(\frac{1}{R}\log|\mathcal{A}| + s\right)$ so that the right-hand side of (9) becomes $\alpha l n^{1/2} + \beta n = \alpha(N/n)n^{1/2} + \beta n = \alpha N n^{-1/2} + \beta n$. Minimizing with respect to n, we take $n = (\alpha/(2\beta))^{2/3}N^{2/3}$ and obtain an expression upper bounded by $3\alpha^{2/3}\beta^{1/3}N^{2/3}$. Plugging in the values of α, β, we finally obtain

$$\sum_{k=1}^{n} \mathbb{E}\left[d_{(E,D)}^{(k-1)l+1,kl}(\mathbf{x})|U^{(k-1)l}\right] - \min_{(\tilde{E},\tilde{D})\in\mathcal{A}} d_{(\tilde{E},\tilde{D})}^{N}(\mathbf{x}) \leq \frac{3B}{(2R)^{1/3}}[(\log|\mathcal{A}|)(\log(|\mathcal{A}|e^{sR}))]^{1/3}N^{2/3}. \quad (10)$$

Note that our choice of the number of blocks $n = (\alpha/(2\beta))^{2/3}N^{2/3}$ implies that the length of each block is $l = N/n = (2\beta/\alpha)^{2/3}N^{1/3}$. Since the above derivation assumes that the block length is greater than the overhead of $\left(\frac{1}{R}\log|\mathcal{A}| + s\right) = \beta/B$ channel uses at its beginning, we verify that this is indeed the case. We require $l = (2\beta/\alpha)^{2/3}N^{1/3} \geq \beta/B$ which is equivalent to $N \geq (\alpha^2\beta)/(4B^3) = \frac{1}{8R}(\log|\mathcal{A}|)(\log(|\mathcal{A}|e^{sR}))$, an inequalilty which holds by hypothesis (for a suitable constant C_1). Hence, taking expectation in (10) establishes (2).

Turning to establish (3), we denote

$$V_k = d_{(E,D)}^{(k-1)l+1,kl}(\mathbf{x}) - \mathbb{E}\left[d_{(E,D)}^{(k-1)l+1,kl}(\mathbf{x})|U^{(k-1)l}\right]$$

and observe that $\{V_k, \sigma(U^{kl})\}_{k\geq 1}$ is a martingale difference sequence with the V_k's almost surely bounded in magnitude by Bl. Applying Hoeffding's bound for martingale difference sequences (cf., e.g., [3, Theorem 9.1]) gives for any n and $\varepsilon > 0$

$$\begin{aligned}
\Pr\left\{\frac{1}{N}\left[d^N_{(E,D)}(\mathbf{x}) - \sum_{k=1}^{n}\mathbb{E}\left[d^{(k-1)l+1,kl}_{(E,D)}(\mathbf{x})|U^{(k-1)l}\right]\right] \geq \varepsilon\right\} &= \Pr\left\{\frac{1}{nl}\sum_{k=1}^{n} V_k \geq \varepsilon\right\} \\
&\leq \exp\left\{-\frac{n\varepsilon^2}{2B^2}\right\}. \qquad (11)
\end{aligned}$$

Finally, combining (10) and (11), we have for any N large enough such that the right-hand side of (10) is less than $N\varepsilon/2$, namely for $N \geq \frac{216B^3}{2R}[(\log|\mathcal{A}|)(\log(|\mathcal{A}|e^{sR}))]/(\varepsilon^3)$:

$$\begin{aligned}
& \Pr\left\{\frac{1}{N}\left[d^N_{(E,D)}(\mathbf{x}) - \min_{(E',D')\in\mathcal{A}} d^N_{(E',D')}(\mathbf{x})\right] \geq \varepsilon\right\} \\
\leq\ & \Pr\left\{\frac{1}{N}\left[d^N_{(E,D)}(\mathbf{x}) - \sum_{k=1}^{n}\mathbb{E}\left[d^{(k-1)l+1,kl}_{(E,D)}(\mathbf{x})|U^{(k-1)l}\right]\right] \geq \varepsilon/2\right\} \\
\leq\ & \exp\left\{-\frac{n(\varepsilon/2)^2}{2B^2}\right\} \\
=\ & \exp\left\{-\frac{(\alpha/(2\beta))^{2/3}N^{2/3}\varepsilon^2}{8B^2}\right\}, \qquad (12)
\end{aligned}$$

which, upon plugging in the values of α and β, establishes (3) and concludes the proof for the case where the above values of $n = (\alpha/(2\beta))^{2/3}N^{2/3}$ and $l = (2\beta/\alpha)^{2/3}N^{1/3}$ are integers. Otherwise, take $l = \lfloor(2\beta/\alpha)^{2/3}N^{1/3}\rfloor$ and let the coding scheme (E, D) behave arbitrarily on the last (incomplete) block. It is straightforward to verify that the above derivation carries over for this case as well with, possibly, slightly modified constants C_1, C_2, C_3 to accommodate the "edge effects". □

It might be tempting to simplify the source code constructed in Theorem 1 in the following way. Rather than generate the member of the reference class to be followed on the k-th block according to (33), simply decide, deterministically, to follow that member which has been proven most competent on the past sequence. This would eliminate the need for a randomization sequence altogether. Unfortunately, as is known from the theory of prediction of individual sequences, such a scheme does not have the necessary adaptivity properties. In fact, one can construct simple examples of reference classes containing as little as two source codes (analogously to simple examples from the prediction framework) for which the redundancy of such a deterministic scheme relative to the reference class is lower bounded by a non vanishing term.

4 Applications

We dedicate this section to an application of the generic result of Section 3 to a few representative cases.

A Scalar Quantizers

Though the reference class in the above theorem is finite, it can be straightforwardly applied to obtain delay-d source codes which compete with any reference class having a well-behaved effective covering number. One example for such an application is for the reference class of all scalar quantizers considered in [11]. Specifically, consider the case $\mathcal{X} = \hat{\mathcal{X}} = [0, 1]$ with the squared-error distortion $d(x_i, \hat{x}_i) = (x_i - \hat{x}_i)^2$. An M-level scalar quantizer Q is a measurable mapping $[0, 1] \to \mathcal{C}$ where $\mathcal{C}$ is any finite subset of $[0, 1]$ with cardinality $|\mathcal{C}| = M$. Following [11], we let $\mathcal{Q}$ denote the class of all M-level ($M = 2^R$) scalar quantizers and let $d_Q^N(\mathbf{x}) = \sum_{i=1}^N (x_i - Q(x_i))^2$. We then have

Corollary 2 *Let $\mathcal{X} = \hat{\mathcal{X}} = [0, 1]$ and $d(\cdot, \cdot)$ be the squared-error distortion measure. For any $\varepsilon > 0$ and N such that $N/(\log N)^2 > K_1/\varepsilon^3$, there exists a zero-delay source code $(E, D) \in \mathcal{F}^0$ such that for all $\mathbf{x} \in \mathcal{X}^\infty$*

$$\mathbb{E}\frac{1}{N} d_{(E,D)}^N(\mathbf{x}) - \inf_{Q \in \mathcal{Q}} \frac{1}{N} d_Q^N(\mathbf{x}) \le K_2 (\log N) N^{-1/3}, \tag{13}$$

and

$$\Pr\left\{ \frac{1}{N} \left[d_{(E,D)}^N(\mathbf{x}) - \inf_{Q \in \mathcal{Q}} d_Q^N(\mathbf{x}) \right] \ge \varepsilon \right\} \le \exp\left\{ -K_3 \varepsilon^2 N^{2/3} \right\}, \tag{14}$$

where K_1, K_2, K_3 are positive constants which depend only on R.

Note that Corollary 2 (inequality (13) in particular) improves the main result of [11] in two directions. The first is in establishing the existence of a source code with a redundancy term upper bounded by $C(\log N)N^{-1/3}$, where that associated with the source code of [11] was shown to be upper bounded by $C(\log N)N^{-1/5}$. The second, perhaps more essential improvement, is in the fact that the source code employed here belongs to $\mathcal{F}^{d=0}$. In particular, it does not use randomization at the decoder and, a fortiori, does not need to know the realization of the randomization sequence used by the encoder. This is in contrast to the source code constructed in [11] which, not only required the availability of a randomization sequence at the decoder's side as well, but required that both the encoder and the decoder dispose of the *same* randomizing sequence (as in models where a "subtractive dither" is assumed, cf. [18, 16]).

Proof: It is straightforward to show (cf. [11, Lemma 2]) that for any $A \geq 2$ there exists a set $\mathcal{Q}_A = \{Q_1, \ldots, Q_A\} \subset \mathcal{Q}$ such that for all N and $\mathbf{x} \in \mathcal{X}^\infty$

$$\min_{Q \in \mathcal{Q}_A} d_Q^N(\mathbf{x}) \leq \inf_{Q \in \mathcal{Q}} d_Q^N(\mathbf{x}) + N \frac{1}{A^{1/M} - 1}. \tag{15}$$

Therefore, letting $(E, D) \in \mathcal{F}^0$ be the source code of Theorem 1, tailored for the set $\mathcal{Q}_A \subset \mathcal{F}_{s=0}^{d=0}$, we obtain

$$\begin{aligned} \mathbb{E}\frac{1}{N} d_{(E,D)}^N(\mathbf{x}) - \inf_{Q \in \mathcal{Q}} \frac{1}{N} d_Q^N(\mathbf{x}) &\leq \mathbb{E}\frac{1}{N} d_{(E,D)}^N(\mathbf{x}) - \min_{Q \in \mathcal{Q}_A} \frac{1}{N} d_Q^N(\mathbf{x}) + \frac{1}{A^{1/M} - 1} \\ &\leq C_2 (\log A)^{2/3} N^{-1/3} + \frac{1}{A^{1/M} - 1}, \end{aligned} \tag{16}$$

where the second inequality follows from Theorem 1 (assuming N is sufficiently large, a fact we shortly verify to follow from the hypothesis of the corollary). Taking $A = \lceil \left(MC_2 N^{-1/3} \right)^{-M} \rceil$, we obtain:

$$\begin{aligned} \mathbb{E}\frac{1}{N} d_{(E,D)}^N(\mathbf{x}) - \inf_{Q \in \mathcal{Q}} \frac{1}{N} d_Q^N(\mathbf{x}) &\leq C_2[-M \log(MC_2 N^{-1/3})] N^{-1/3} + \frac{1}{(MC_2 N^{-1/3})^{-1} - 1} \\ &\leq K_2 (\log N) N^{-1/3}. \end{aligned} \tag{17}$$

We note that for (16), and hence for (17), to hold we need (by Theorem 1) to require $N > (C_1 (\log A)^2)/(\varepsilon^3)$. But, by our choice of A, $\log A$ is, up to a multiplicative constant (depending only on M), equivalent to $\log N$ and therefore it is enough to require $N > (K_1 (\log N)^2)/(\varepsilon^3)$ for some K_1 which establishes (13). To arrive at (14), it is easy to verify that the hypothesis $N > (K_1 (\log N)^2)/(\varepsilon^3)$ implies also $1/(A^{1/M} - 1) \leq \varepsilon/2$ (for a suitably chosen K_1). Hence

$$\begin{aligned} \Pr\left\{ \frac{1}{N} d_{(E,D)}^N(\mathbf{x}) - \inf_{Q \in \mathcal{Q}} \frac{1}{N} d_Q^N(\mathbf{x}) \geq \varepsilon \right\} &\leq \Pr\left\{ \frac{1}{N} d_{(E,D)}^N(\mathbf{x}) - \min_{Q \in \mathcal{Q}_A} \frac{1}{N} d_Q^N(\mathbf{x}) \geq \varepsilon/2 \right\} \\ &\leq \exp\left\{ -K_3 \varepsilon^2 N^{2/3} \right\}, \end{aligned} \tag{18}$$

where the first inequality follows from (15) and the second one follows from Theorem 1 (inequality (3)). $\square$

B Sliding Block and Trellis Source Codes

A class to which Theorem 1 can be directly applied is that of sliding-block codes (cf., e.g., [7]). Assume here an arbitrary distortion measure. A finite-constraint length, time-invariant encoder with constraint-length l, memory l_M, and delay $l_d = l - l_M - 1$ is a mapping $f : \mathcal{X}^l \to \{1, 2, \ldots, M\}$ yielding the channel symbols $\{y_i\}$ defined by $y_i = f(x_{i-l_M}^{i+l_d})$. Similarly, a sliding-block decoder with constraint length k is a mapping $g : \{1, 2, \ldots, M\}^k \to \hat{\mathcal{X}}$ yielding a reproduction process $\hat{x}_i = g(y_{i-k+1}^i)$. When the source and reproduction alphabets are finite, it is easy to see that the cardinality of the class of all such sliding-block source codes with respective constraint lengths of l and k is $M^{|\mathcal{X}|^l} \cdot |\hat{\mathcal{X}}|^{M^k}$. Therefore, Theorem 1 guarantees the existence of a source code with delay l_d which achieves the lowest distortion attainable by any sliding-block source code with constraint lengths l, k and delay l_d, uniformly for all individual sequences in the sense of equations (2) and (3) (with $|\mathcal{A}| = M^{|\mathcal{X}|^l} \cdot |\hat{\mathcal{X}}|^{M^k}$). For the case where the alphabets are not of finite cardinality, competing with *all* such sliding-block source codes of a given order is clearly an overambitious task. However, any sufficiently smoothly parametrizable subset of this class can be dealt with using appropriate grids, similarly as was done in the proof of Corollary 2.

Another important related family of source codes that are covered by Theorem 1 consists of the block trellis codes (cf., e.g., [7, 8]). In particular, note that a block trellis coding scheme of constraint length K and search depth L, which consists of a constraint-length K sliding-block decoder g and a depth L trellis search encoder matched to g, is a member of $\mathcal{F}_K^{L-1}$.

C Adaptive Differential Pulse Code Modulation

We dedicate this subsection to specializing the approach demonstrated in Theorem 1 for obtaining an adaptive scheme which "tracks" the Differential Pulse Code Modulator (D-PCM), out of a given family, which does best on the data. DPCM-based schemes are widely used in speech and image applications and are therefore of considerable practical interest. In particular, the study of such coding schemes within the individual-sequence framework which we consider here is especially relevant for many image compression situations in which there is no natural statistical model for the data.

The reader is referred, e.g., to [7, 9] and the references therein for a comprehensive account of the theory and practice of DPCM-based coding schemes. Following is a laconic description intended primarily to introduce notation for later use. A DPCM delayless source code of order s is an element of $\mathcal{F}^0$, which is fully characterized by a predictor $P : \hat{\mathcal{X}}^s \to \mathbb{R}$ and an M-level quantizer $Q \in \mathcal{Q}$. The encoder produces the i-th channel symbol representing the value of $Q(e_i)$, where $e_i = x_i - P(\hat{x}_{i-s}^{i-1})$. The decoder gives the

i-th reconstruction according to $\hat{x}_i = Q(e_i) + P(\hat{x}_{i-s}^{i-1})$. For concreteness in initializing the predictor take, say, $\hat{x}_{-s+1}^0 = (a, \ldots, a)$ for some $a \in \hat{\mathcal{X}}$. An important subset of this family is that where the source and reconstruction alphabets are (possibly subsets of) the real line and the predictor P is a linear time-invariant predictor[2]. We let $\mathcal{F}_{DPCM}(s, \delta, R)$ denote the class of all DPCM delayless source codes of order up to s having a linear time-invariant predictor with impulse response h satisfying $\|h\|_1 \leq 1 - \delta$ and operating at rate R. The reason we are interested in such impulse responses is the strong stability of the decoder (linear feedback system) that they induce. Specifically, this property ensures that the respective outputs of a DPCM (having such an h) fed with similar inputs will also be similar. Assume that the source and reconstruction alphabets in what follows consist of a bounded subset of the real line and that d is the squared-error distortion. In particular, we can assume, as before, that the distortion measure is bounded by $B > 0$.

Theorem 3 *Let $\mathcal{A}$ be a finite subset of $\mathcal{F}_{DPCM}(s, \delta, R)$ for some $s \geq 0$, R and $0 < \delta \leq 1$. Then for any $0 \leq \varepsilon \leq B$ and $N > C_1[(\log|\mathcal{A}|)(\log|\mathcal{A}| + C_4(s/\delta)^3 R)]/(\varepsilon^3)$ there exists a delayless source code $(E, D) \in \mathcal{F}^0(R)$ such that for all $\mathbf{x} \in \mathcal{X}^\infty$*

$$\mathbb{E}\left\{\frac{1}{N}\left[d_{(E,D)}^N(\mathbf{x}) - \min_{\mathcal{A}} d_{(E',D')}^N(\mathbf{x})\right]\right\} \leq C_2[(\log|\mathcal{A}|)(\log|\mathcal{A}| + C_4(s/\delta)^3 R)]^{1/3} \cdot N^{-1/3}, \quad (19)$$

and

$$\Pr\left\{\frac{1}{N}\left[d_{(E,D)}^N(\mathbf{x}) - \min_{\mathcal{A}} d_{(E',D')}^N(\mathbf{x})\right] \geq \varepsilon\right\} \leq \exp\left\{-C_3\left(\frac{\log|\mathcal{A}|}{\log|\mathcal{A}| + C_4(s/\delta)^3 R}\right)^{1/3} \varepsilon^2 N^{2/3}\right\}, \quad (20)$$

where C_1, C_2, C_3, C_4 are positive constants which depend only on B and R.

Note that Theorem 3 does not follow directly from Theorem 1 since a source code belonging to $\mathcal{F}_{DPCM}(s, \delta)$ will, in general, (for $s \geq 1$) not have a finite-memory decoder (note that the decoder's output depends on the present channel symbol and on past *reconstruction* symbols). The proof of Theorem 3, however, which we outline below, is very similar to that of Theorem 1. In particular, the construction of the source code (E, D) satisfying (19) and (20) is essentially the same as that in Theorem 1.

The reference class in the above theorem is an arbitrary finite subset of $\mathcal{F}_{DPCM}(s, \delta, R)$. Theorem 3 can be applied to obtain a scheme which competes with the whole set $\mathcal{F}_{DPCM}(s, \delta, R)$. This can be done by applying the theorem on a sufficiently dense finite subset of $\mathcal{F}_{DPCM}(s, \delta, R)$,

[2]Note that in this case the decoder is a simple linear time-invariant feedback system.

similarly as was done in subsection 4.A for the infinite family of scalar quantizers. More specifically, for $\varepsilon > 0$ we can let $\mathcal{F}^{\varepsilon}_{DPCM}(s,\delta,R)$ denote the finite set containing those elements of $\mathcal{F}_{DPCM}(s,\delta,R)$ with an associated impulse response h having coefficients that are integer multiples of ε. One can then show, using the simple structure of the encoder-decoder pair comprising a DPCM scheme, that

$$\min_{\mathcal{F}^{\varepsilon}_{DPCM}(s,\delta,R)} d^N_{(E',D')}(\mathbf{x}) - \inf_{\mathcal{F}_{DPCM}(s,\delta,R)} d^N_{(E',D')}(\mathbf{x}) \leq C\varepsilon N, \tag{21}$$

where C is a constant independent of ε, N and the bounded sequence $\mathbf{x}$. Moreover, the cardinality of $\mathcal{F}^{\varepsilon}_{DPCM}(s,\delta,R)$ is clearly upper bounded by $(2/\varepsilon)^s$. Hence, by applying the scheme of Theorem 3, tailored to the class $\mathcal{F}^{\varepsilon}_{DPCM}(s,\delta,R)$, where $\varepsilon \propto N^{-1/3}$, and using (21) one can obtain a scheme (E,D) satisfying for all $\mathbf{x} \in \mathcal{X}^{\infty}$

$$\mathbb{E}\left\{\frac{1}{N}\left[d^N_{(E,D)}(\mathbf{x}) - \inf_{\mathcal{F}_{DPCM}(s,\delta,R)} d^N_{(E',D')}(\mathbf{x})\right]\right\} \leq \tilde{C}(\log N)^2 \cdot N^{-1/3}, \tag{22}$$

where $\tilde{C}$ is a constant (possibly dependent on s, δ yet) independent on N. A similar analogue to inequality (20) can also be obtained in this way.

Proof sketch: The construction of (E,D) follows that in the proof of Theorem 1 essentially verbatim. The only difference is that now, at the decoder's side, from time $i = (k-1)l + \lceil \frac{1}{R}\log|\mathcal{A}|\rceil + 1$ up to the end of the block $i = kl$, the decoder, knowing the identity of $D^{(k)}$ (equivalently, of the linear predictor $h^{(k)}$ and of the quantizer $Q^{(k)}$), outputs a reproduction sequence by feeding the incoming ${Q^{(k)}}^{-1}(y_i)$, where ${Q^{(k)}}^{-1}(\cdot) : \{1,\ldots,M\} \to [0,1]$ denotes the inverse transformation of the quantizer $Q^{(k)}$, into the linear feedback system characterizing the decoder $D^{(k)}$, where the decoder D assumes that up to this point in time the input to the system was constant at say zero. The only thing left to verify is that the performance of this scheme on the k-th block is not too heavily deteriorated, relative to the performance of $(E^{(k)},D^{(k)})$, by the fact that the input to the decoder $D^{(k)}$ up to time i was in fact $\{{Q^{(k)}}^{-1}(y_j)\}_{j<i}$ and not constant at zero as the decoder D is assuming. The fact that $\|h\|_1 \leq 1-\delta$, however, can easily be shown to imply that the differences between the (unnormalized) cumulative distortions of (E,D) and $(E^{(k)},D^{(k)})$ on all blocks (of all lengths) are uniformly upper bounded by a quantity proportional to $(s/\delta)^3$. Therefore, an inequality is obtained which is analogous to (7), with s replaced by a constant times $(s/\delta)^3$. The remainder of the proof carries through verbatim. □

5 The Noisy Setting

In this section we consider the case where the individual sequence is corrupted by noise. Limited-delay coding schemes are sought, for this case, which operate on the noisy sequence and produce a reconstruction sequence which is judged with respect to the clean individual sequence. For concreteness and simplicity of the exposition, we shall focus on the case where the reference class consists of sliding window schemes. The approach we will present, however, can be similarly applied to handle reference classes of other types.

A Problem Formulation

We now formalize the notion of a limited-delay sequential source code for the case where the individual sequence to encode is corrupted by noise. Specifically, we assume now, as in the noise-free case, that there is an individual sequence $x_1, x_2, \ldots,\ x_i \in \mathcal{X}$ to encode. The encoder-decoder pair, however, accesses the sequence $Z_1, Z_2, \ldots,\ Z_i \in \mathcal{Z}$, which is the output of the fixed memoryless channel whose input is the individual sequence of interest $x_1, x_2, \ldots$. For simplicity of the exposition we assume that $\mathcal{X} = \mathcal{Z}$ is finite and we let M denote the $\mathcal{X} \times \mathcal{X}$ channel transition probability matrix, which we assume invertible. The approach we will present can be applied to more general cases.

For this setting we define a delay $d = d_e + d_d$ ($d_e, d_d \geq 0$) sequential scheme for combined filtering and compression of fixed rate $R = \log M$ with a randomized encoder by a pair (E, D). The randomized encoder E is given by a sequence $\{E_i\}_{i=1}^{\infty}$, where $E_i : \mathcal{X}^{i+d_e} \times [0,1]^i \to \{1, 2, \ldots, M\}$. The decoder D is given, as in the noise-free setting, by a sequence $\{D_i\}_{i=1}^{\infty}$, where $D_i : \{1, 2, \ldots, M\}^{i+d_d} \to \hat{\mathcal{X}}$. This scheme operates as follows. The encoder produces the i-th channel symbol $Y_i \in \{1, 2, \ldots, M\}$ based on Z^{i+d_e} and on the random sequence U^i according to $Y_i = E_i(Z^{i+d_e}, U^i)$, where $\{U_i\}_{i=1}^{\infty}$ is a randomization sequence of i.i.d. random variables, uniformly distributed on $[0, 1]$ and independent of $\{Z_i\}$. The decoder emits the reconstructed sequence $\hat{x}_1, \hat{x}_2, \ldots$ according to $\hat{x}_i = D_i(Y^{i+d_d})$. We continue to let $\mathcal{F}^d(R)$ ($\mathcal{F}^d$), where $R = \log M$, denote the class of all such combined filtering and compression schemes operating at the fixed rate R since they have exactly the same structure as in the noise-free case.

The definition of the cumulative distortion of a scheme $(E, D) \in \mathcal{F}^d$ remains as in the noise-free setting:

$$d^n_{(E,D)}(\mathbf{x}) = \sum_{i=1}^{n} d(x_i, \hat{x}_i), \tag{23}$$

where $d : \mathcal{X} \times \hat{\mathcal{X}} \to [0, B]$ is a bounded distortion measure ($B < \infty$) and here the $\hat{x}_i$'s on the right hand side are generated by feeding the noisy sequence $\mathbf{Z} = Z_1, Z_2, \ldots$ into (E, D) as described above. Note that $d^n_{(E,D)}(\mathbf{x})$ is a random variable which depends on the realization of the channel noise and of the randomization sequence U^n. As before, we similarly denote, for $n_1 \leq n_2$, $d^{n_1,n_2}_{(E,D)}(\mathbf{x}) = \sum_{i=n_1}^{n_2} d(x_i, \hat{x}_i)$.

B Time-Invariant Sliding-Window Schemes

A member (E, D) of $\mathcal{F}^d$ is a time-invariant sliding-window scheme if there exist $d_1, d_2, d_3, d_4 \geq 0$, $f : \mathcal{Z}^{d_1+d_2+1} \to \{1, 2, \ldots, M\}$ and $g : \{1, 2, \ldots, M\}^{d_3+d_4+1} \to \hat{\mathcal{X}}$ such that for all i, the i-th channel symbol generated by the encoder is given by $Y_i = f(Z_{i-d_1}^{i+d_2})$ and the i-th reconstruction symbol by $\hat{x}_i = g(Y_{i-d_3}^{i+d_4})$. We refer to d_1 (d_3) as the encoder- (decoder-) memory and to d_2 (d_4) as the encoder- (decoder-) horizon. We let $\mathcal{F}^{d_1,d_2,d_3,d_4} \subset \mathcal{F}^{d_2+d_4}$ denote the family of all such time-invariant sliding-window schemes.

B.1 Estimating the Distortion of $(E, D) \in \mathcal{F}^{d_1,d_2,d_3,d_4}$

Unfortunately, the cumulative distortion of a scheme $(E, D) \in \mathcal{F}^d$, which was defined in (23), depends on the individual sequence $\mathbf{x}$ and, therefore, is not available when only its noisy version $\mathbf{Z}$ is accessible. Motivated by the approach which guided the construction of the coding schemes in the noise-free setting, our first goal in the noisy setting is to obtain an efficient estimator for the distortion suffered by a scheme in $\mathcal{F}^{d_1,d_2,d_3,d_4}$, which is only based on the observed noisy sequence $\mathbf{Z}$. To this end, let $(E, D) \in \mathcal{F}^{d_1,d_2,d_3,d_4}$ be given and denote $\hat{x}_i = g(Y_{i-d_3}^{i+d_4}) = g(\{f(Z_{i'-d_1}^{i'+d_2})\}_{i'=i-d_3}^{i+d_4}) \equiv h(Z_{i-(d_1+d_3)}^{i+(d_2+d_4)})$. We can now write

$$\begin{aligned} d^n_{(E,D)}(\mathbf{x}) &= \sum_{i=1}^{n} d(x_i, h(Z_{i-(d_1+d_3)}^{i+(d_2+d_4)})) && (24)\\ &= \sum_{i=1}^{n} d(x_i, h(Z_{i-j}^{i+k})) && (25)\\ &= \sum_{(\mathbf{a},\mathbf{b}) \in \mathcal{X}^m \times \mathcal{Z}^m} d(a_{j+1}, h(\mathbf{b})) \cdot N^n(x_{i-j}^{i+k} = \mathbf{a}, Z_{i-j}^{i+k} = \mathbf{b}), && (26) \end{aligned}$$

where we denote $d_1 + d_3 = j$, $d_2 + d_4 = k$, $m = j + k + 1$, the $j + 1$-th component of $\mathbf{a}$ by a_{j+1}, and $N^n(x_{i-j}^{i+k} = \mathbf{a}, Z_{i-j}^{i+k} = \mathbf{b}) \stackrel{\triangle}{=} |\{1 \leq i \leq n : x_{i-j}^{i+k} = \mathbf{a}, Z_{i-j}^{i+k} = \mathbf{b}\}|$. Evidently, efficient estimators for the unobserved $\{N^n(x_{i-j}^{i+k} = \mathbf{a}, Z_{i-j}^{i+k} = \mathbf{b})\}_{(\mathbf{a},\mathbf{b})\in\mathcal{X}^m\times\mathcal{Z}^m}$ would lead to an efficient estimator for $d^n_{(E,D)}(\mathbf{x})$ by plugging these into (26).

Let $M(m)$ denote the transition matrix characterizing the memoryless noisy channel for vector inputs of length m, induced by the channel matrix M. Specifically, $M(m)$ is a $|\mathcal{X}|^m \times |\mathcal{X}|^m$ matrix whose entry at $(\mathbf{a},\mathbf{b}) \in \mathcal{X}^m \times \mathcal{X}^m$ (where the rows and columns of $M(m)$ are arranged according to some lexicographic ordering of $|\mathcal{X}|^m$) is given by the probability for noisy $Z_{i-j}^{i+k} = \mathbf{b}$ when $x_{i-j}^{i+k} = \mathbf{a}$. Since the channel is fixed and memoryless, this means $M(m)_{\mathbf{a},\mathbf{b}} = \prod_{k=1}^m M_{a_k,b_k}$. We will further let $M(m)^{-1}_{\mathbf{a},\mathbf{b}}$ denote the entry corresponding to $(\mathbf{a},\mathbf{b})$ in the inverse matrix of $M(m)$, which exists by the existence of M^{-1} which is assumed throughout (recall subsection A). The first observation we make towards constructing an estimator for $N^n(x_{i-j}^{i+k} = \mathbf{a}, Z_{i-j}^{i+k} = \mathbf{b})$ is the following.

Lemma 4 *For all $i \geq 1$ and $x_{i-j}^{i+k} \in \mathcal{X}^m$, the observable $M(m)_{\mathbf{a},\mathbf{b}} \cdot M(m)^{-1}_{Z_{i-j}^{i+k},\mathbf{a}}$ is an unbiased estimator for (the unobserved) $1_{\{x_{i-j}^{i+k}=\mathbf{a}, Z_{i-j}^{i+k}=\mathbf{b}\}}$. I.e.,*

$$E\left\{M(m)_{\mathbf{a},\mathbf{b}} \cdot M(m)^{-1}_{Z_{i-j}^{i+k},\mathbf{a}}\right\} = E1_{\{x_{i-j}^{i+k}=\mathbf{a}, Z_{i-j}^{i+k}=\mathbf{b}\}}. \tag{27}$$

Proof:

$$\begin{aligned}
E\left\{M(m)_{\mathbf{a},\mathbf{b}} \cdot M(m)^{-1}_{Z_{i-j}^{i+k},\mathbf{a}}\right\} &= E\left\{M(m)_{\mathbf{a},\mathbf{b}} \cdot \sum_{\mathbf{b}'\in|\mathcal{X}|^m} 1_{\{Z_{i-j}^{i+k}=\mathbf{b}'\}} M(m)^{-1}_{\mathbf{b}',\mathbf{a}}\right\} \\
&= E\left\{M(m)_{\mathbf{a},\mathbf{b}} \cdot \sum_{\mathbf{b}'\in|\mathcal{X}|^m} \left(\sum_{\mathbf{a}'\in|\mathcal{X}|^m} 1_{\{x_{i-j}^{i+k}=\mathbf{a}', Z_{i-j}^{i+k}=\mathbf{b}'\}}\right) M(m)^{-1}_{\mathbf{b}',\mathbf{a}}\right\} \\
&= M(m)_{\mathbf{a},\mathbf{b}} \cdot \sum_{\mathbf{b}'\in|\mathcal{X}|^m} \left(\sum_{\mathbf{a}'\in|\mathcal{X}|^m} 1_{\{x_{i-j}^{i+k}=\mathbf{a}'\}} \cdot M(m)_{\mathbf{a}',\mathbf{b}'}\right) M(m)^{-1}_{\mathbf{b}',\mathbf{a}} \\
&= M(m)_{\mathbf{a},\mathbf{b}} \cdot \sum_{\mathbf{b}'\in|\mathcal{X}|^m} M(m)_{x_{i-j}^{i+k},\mathbf{b}'} M(m)^{-1}_{\mathbf{b}',\mathbf{a}} \\
&= M(m)_{\mathbf{a},\mathbf{b}} \cdot I_{x_{i-j}^{i+k},\mathbf{a}} \\
&= M(m)_{\mathbf{a},\mathbf{b}} \cdot 1_{\{x_{i-j}^{i+k}=\mathbf{a}\}} \\
&= E1_{\{x_{i-j}^{i+k}=\mathbf{a}, Z_{i-j}^{i+k}=\mathbf{b}\}},
\end{aligned} \tag{28}$$

where the third equality follows from the fact that $1_{\{x_{i-j}^{i+k}=\mathbf{a}', Z_{i-j}^{i+k}=\mathbf{b}'\}} = 1_{\{x_{i-j}^{i+k}=\mathbf{a}'\}} \cdot M(m)_{\mathbf{a}',\mathbf{b}'}$ and where I denotes the identity matrix. □

We now let

$$\hat{N}^n_{\mathbf{a},\mathbf{b}}(Z^{n+k}) \stackrel{\triangle}{=} \sum_{i=1}^n M(m)_{\mathbf{a},\mathbf{b}} \cdot M(m)^{-1}_{Z_{i-j}^{i+k},\mathbf{a}} \tag{29}$$

be our estimator for the unobserved $N^n(x_{i-j}^{i+k} = \mathbf{a}, Z_{i-j}^{i+k} = \mathbf{b})$. Note that the estimation error,

$$\begin{aligned} N^n(x_{i-j}^{i+k} = \mathbf{a}, Z_{i-j}^{i+k} = \mathbf{b}) - \hat{N}_{\mathbf{a},\mathbf{b}}^n(Z^{n+k}) &= \sum_{i=1}^n \left(1_{\{x_{i-j}^{i+k}=\mathbf{a}, Z_{i-j}^{i+k}=\mathbf{b}\}} - M(m)_{\mathbf{a},\mathbf{b}} \cdot M(m)^{-1}_{Z_{i-j}^{i+k},\mathbf{a}} \right) \\ &\triangleq \sum_{i=1}^n \delta_i, \end{aligned} \tag{30}$$

is, by Lemma 4, a sum of zero-mean, bounded random variables. Furthermore, these random variables have a well-behaved dependence structure. In particular, for all $i \geq 1$, δ_i is independent of $\{\delta_j\}_{j>i+m}$. Hence, by standard techniques (cf., e.g., [14]), it can be shown that the estimation error will typically be $O(\sqrt{n})$, that the probability of its magnitude exceeding $n\varepsilon$, for any $\varepsilon > 0$, decays exponentially rapidly, pointwise asymptotic bounds on its magnitude in the order of $\sqrt{n \log \log n}$, etc. Motivated by this, we finally let our estimator for the unavailable cumulative distortion $d_{(E,D)}^n(\mathbf{x})$ be

$$\hat{d}_{(E,D)}^n(\mathbf{Z}) = \sum_{(\mathbf{a},\mathbf{b}) \in \mathcal{X}^m \times \mathcal{X}^m} d(a_{j+1}, h(\mathbf{b})) \cdot \hat{N}_{\mathbf{a},\mathbf{b}}^n(Z^{n+k}). \tag{31}$$

B.2 The Coding-Filtering Scheme for the Noisy Case

Equipped with an estimator for the unobserved cumulative distortion of a sliding window scheme in the noisy setting, we now construct an adaptive scheme which is similar, in principle, to that of Section 3. The essential difference is in the fact that the randomization is performed with respect to a distribution assigning weights which are exponentially proportional to the *estimated* past distortion. This approach has recently lead to efficient schemes in the context of prediction of individual sequences corrupted by noise [15].

Theorem 5 *Let $\mathcal{A}$ be a finite subset of $\mathcal{F}^{d_1,d_2,d_3,d_4}(R)$ for some $d_1, d_2, d_3, d_4 \geq 0$ and a given $R = \log M$. Then there exists a source code $(E, D) \in \mathcal{F}^d(R)$ $(d = d_2 + d_4)$ such that for any $\varepsilon > 0$, N sufficiently large so that $B\sqrt{\log|\mathcal{A}|/2} \cdot N^{-1/4} + BN^{-1/2}\left(\frac{1}{R}\log|\mathcal{A}| + d\right) < \varepsilon/3$ and all $\mathbf{x} \in \mathcal{X}^\infty$*

$$\Pr\left\{ \frac{1}{N}\left[d_{(E,D)}^N(\mathbf{x}) - \min_{(E',D') \in \mathcal{A}} d_{(E',D')}^N(\mathbf{x}) \right] \geq \varepsilon \right\} \leq 5\sqrt{N}|\mathcal{A}||\mathcal{X}|^{2m} \exp\left\{ -\sqrt{N}\varepsilon^2 C \right\}, \tag{32}$$

where C is a positive constant (which depends only on B, R, $\mathcal{A}$, m and $\max_{(\mathbf{a},\mathbf{b}) \in \mathcal{X}^m \times \mathcal{X}^m} M(m)^{-1}_{\mathbf{a},\mathbf{b}}$, the explicit value of which will be apparent in the proof).

Note that here, as in the noise-free setting, the scheme which we construct, in addition to complying with the delay limitations of the sliding-window schemes with which it competes, operates at the exact same rate.

We further remark, as we did in the noise-free case, that while the source code (E, D) depends on the length N of the sequence, it is straightforward to use it to obtain a strongly sequential (horizon-independent) scheme for which a concentration inequality in the form of (32) holds for all sufficiently large N. One can then apply the Borel-Cantelli lemma to obtain almost sure asymptotic performance guarantees. Finally, we remark that in the above theorem, $\mathcal{A}$ can be taken as the set of all sliding window schemes of up to a certain order (memory and horizon), which is a finite set in the finite alphabet case. By gradually increasing the order and employing the scheme of Theorem 5 taylored for this order on consecutive blocks of lengths which increase at the proper rate, a scheme can be obtained which asymptotically almost surely achieves the performance of any sliding-block scheme, no matter what the underlying individual sequence may be.

Proof of Theorem 5: The construction of $(E, D) \in \mathcal{F}^{d_2+d_4}$ which follows is similar to that in the proof of Theorem 1, with the essential difference that the past cumulative distortions of the schemes in $\mathcal{A}$, which are now unavailable to the encoder, are replaced by their respective estimates for the purpose of randomization.

Fix an $l \ll N$ and divide the time-axis $i = 1, 2, \ldots, N$ into $n = N/l$ consecutive non-overlapping blocks (assume l divides N). At the beginning of the k-th block $1 \leq k \leq n$, i.e. at the $i = (k-1)l + 1$-th channel use, after seeing Z^{i+d} and when U^i are available, the encoder uses U_i to generate $(E^{(k)}, D^{(k)})$, a $\mathcal{A}$-valued random variable with distribution satisfying almost surely

$$\Pr\{(E^{(k)}, D^{(k)}) = (E', D')|\mathcal{G}_Z\} = \frac{\exp\left\{-\eta \hat{d}^{(k-1)l}_{(E',D')}(\mathbf{Z})\right\}}{\sum_{(\tilde{E},\tilde{D})\in\mathcal{A}} \exp\left\{-\eta \hat{d}^{(k-1)l}_{(\tilde{E},\tilde{D})}(\mathbf{Z})\right\}} \quad \forall (E', D') \in \mathcal{A}, \tag{33}$$

where $\mathcal{G}_Z$ is the smallest sigma-algebra w.r.t. which all Z_i-s are measurable and $\eta > 0$ is taken as in Theorem 1. For convenience, we let $f^{(k)}$ and $g^{(k)}$ denote the sliding window encoder and decoder, respectively, characterizing the source code $(E^{(k)}, D^{(k)})$. The encoder now dedicates the first $\lceil \frac{1}{R} \log |\mathcal{A}| \rceil$ channel symbols at the beginning of the k-th block, i.e. Y_i for $i = (k-1)l + 1, \ldots (k-1)l + \lceil \frac{1}{R} \log |\mathcal{A}| \rceil$, to convey to the decoder the identity of $g^{(k)}$. At the remainder of the block, i.e. at channel uses $i = (k-1)l + \lceil \frac{1}{R} \log |\mathcal{A}| \rceil + 1, \ldots, kl$, the encoder produces the channel symbols $Y_i = f^{(k)}(Z_{i-d_1}^{i+d_2})$. On the decoder's side, at the beginning of the block, at channel uses $i = (k-1)l+1, \ldots, (k-1)l + \lceil \frac{1}{R}\log|\mathcal{A}|\rceil + d_3$, the decoder outputs an arbitrary reproduction sequence of $\hat{x}_i$'s. From the $i = (k-1)l + \lceil \frac{1}{R}\log|\mathcal{A}|\rceil + d_3 + 1$-th channel use up to the end of the block $i = kl$, the decoder, knowing the identity of $g^{(k)}$ and the output of $f^{(k)}$ at least d_3 channel uses back, outputs the reproduction sequence according to $\hat{x}_i = g^{(k)}(Y_{i-d_3}^{i+d_4})$.

To analyze the performance of the above scheme and to establish equation (32), we will consider a genie-aided scheme. Specifically, let (E_G, D_G) be a source code which is allowed to access the clean sequence $\mathbf{x}$ as well as its noisy version $\mathbf{Z}$. This scheme operates exactly as (E, D) described above, with the only difference that the pair $(E^{(k)}, D^{(k)})$ is generated (using the randomization sequence) according almost surely to

$$\Pr\{(E^{(k)}, D^{(k)}) = (E', D')|\mathcal{G}_Z\} = \frac{\exp\left\{-\eta d_{(E',D')}^{(k-1)l}(\mathbf{x})\right\}}{\sum_{(\tilde{E},\tilde{D})\in\mathcal{A}} \exp\left\{-\eta d_{(\tilde{E},\tilde{D})}^{(k-1)l}(\mathbf{x})\right\}} \quad \forall (E', D') \in \mathcal{A}, \tag{34}$$

rather than according to (33) (this is where the "genie" comes in since the generation of $(E^{(k)}, D^{(k)})$ from such a distribution requires knowledge of $\{d_{(\tilde{E},\tilde{D})}^{(k-1)l}(\mathbf{x})\}_{(\tilde{E},\tilde{D})\in\mathcal{A}}$ which depends on the clean sequence $\mathbf{x}$). All the stages of the proof of Theorem 1 carry over essentially verbatim for (E_G, D_G) (with s replaced by d). In particular, we have, as in (9), a.s.

$$\sum_{k=1}^{n} \mathbb{E}\left[d_{(E_G,D_G)}^{(k-1)l+1,kl}(\mathbf{x})|\mathcal{G}_Z\right] - \min_{(\tilde{E},\tilde{D})\in\mathcal{A}} d_{(\tilde{E},\tilde{D})}^{N}(\mathbf{x}) \leq B\sqrt{\log|\mathcal{A}|/2}\cdot ln^{1/2} + Bn\left(\frac{1}{R}\log|\mathcal{A}| + d\right). \tag{35}$$

In the Appendix we show that for arbitrary $(a_1, \ldots, a_N), (b_1, \ldots, b_N) \in \mathbb{R}^N$ and $C > 0$

$$\sup_{(\alpha_1,\ldots,\alpha_N)\in[0,C]^N} \left| \frac{\sum_{j=1}^{N} e^{-a_j}\alpha_j}{\sum_{j=1}^{N} e^{-a_j}} - \frac{\sum_{j=1}^{N} e^{-b_j}\alpha_j}{\sum_{j=1}^{N} e^{-b_j}} \right| \leq 2C \max_{1\leq j\leq N} |a_j - b_j|. \tag{36}$$

From the construction of the genie aided (E_G, D_G) and our legitimate (E, D), it follows by an application of (36) that for all $1 \leq k \leq n$

$$\left|\mathbb{E}\left[d_{(E,D)}^{(k-1)l+1,kl}(\mathbf{x})|\mathcal{G}_Z\right] - \mathbb{E}\left[d_{(E_G,D_G)}^{(k-1)l+1,kl}(\mathbf{x})|\mathcal{G}_Z\right]\right| \leq 2Bl\eta \max_{(E',D')\in\mathcal{A}} \left|\hat{d}_{(E',D')}^{(k-1)l}(\mathbf{Z}) - d_{(E',D')}^{(k-1)l}(\mathbf{x})\right| \quad \text{a.s.} \tag{37}$$

This leads a.s. to

$$\begin{aligned} &\sum_{k=1}^{n} \mathbb{E}\left[d_{(E,D)}^{(k-1)l+1,kl}(\mathbf{x})|\mathcal{G}_Z\right] - \min_{(\tilde{E},\tilde{D})\in\mathcal{A}} d_{(\tilde{E},\tilde{D})}^{N}(\mathbf{x}) \\ \leq\ & B\sqrt{\log|\mathcal{A}|/2}\cdot ln^{1/2} + Bn\left(\frac{1}{R}\log|\mathcal{A}| + d\right) \end{aligned}$$

$$+2Bl\eta\sum_{k=1}^{n}\max_{(E',D')\in\mathcal{A}}\left|\hat{d}^{(k-1)l}_{(E',D')}(\mathbf{Z})-d^{(k-1)l}_{(E',D')}(\mathbf{x})\right|, \tag{38}$$

where we recall that η is taken to be $\sqrt{8\log|\mathcal{A}|/(l^2B^2n)}$. To conclude, we show that the first term on the left side of (38) is, with high probability, close to $d^N_{(E,D)}(\mathbf{x})$ and that the last term on the right side of (38) is small with high probability. Specifically, it is shown in the Appendix that for all $\varepsilon>0$

$$\Pr\left\{\frac{1}{N}\left[d^N_{(E,D)}(\mathbf{x})-\sum_{k=1}^{n}\mathbb{E}\left[d^{(k-1)l+1,kl}_{(E,D)}(\mathbf{x})|\mathcal{G}_Z\right]\right]>\varepsilon\right\}\le\exp\left\{-\frac{2N^2\varepsilon^2}{n(2lB)^2}\right\}, \tag{39}$$

and

$$\Pr\left\{\frac{1}{N}2Bl\eta\sum_{k=1}^{n}\max_{(E',D')\in\mathcal{A}}\left|\hat{d}^{(k-1)l}_{(E',D')}(\mathbf{Z})-d^{(k-1)l}_{(E',D')}(\mathbf{x})\right|>\varepsilon\right\}\le 4n|\mathcal{A}||\mathcal{X}|^{2m}\exp\left\{-\frac{N}{n}C_1\varepsilon^2\right\}, \tag{40}$$

where $C_1=(32\log|\mathcal{A}|mB^2\tilde{B}^2|\mathcal{X}|^{4m})^{-1}$ and $\tilde{B}=\max_{(\mathbf{a},\mathbf{b})\in\mathcal{X}^m\times\mathcal{X}^m}M(m)^{-1}_{\mathbf{a},\mathbf{b}}$. Choosing this time $n=l=\sqrt{N}$ (as opposed to the $n\propto N^{2/3}$, $l\propto N^{1/3}$ regime of the noise-free case), the above two upper bounds give the same exponential speed of decay ($\sqrt{N}$). Combining inequalities (38), (39) and (40), taking $n=l=\sqrt{N}$, applying the union bound and noting that $C_1\le 1/(2B^2)$ gives

$$\begin{aligned}\Pr\left\{\frac{1}{N}\left[d^N_{(E,D)}(\mathbf{x})-\min_{(\tilde{E},\tilde{D})\in\mathcal{A}}d^N_{(\tilde{E},\tilde{D})}(\mathbf{x})\right]\ge\varepsilon\right\} &\le \exp\left\{-\sqrt{N}(\varepsilon/3)^2\frac{1}{2B^2}\right\} \\ &\quad +4\sqrt{N}|\mathcal{A}||\mathcal{X}|^{2m}\exp\left\{-\sqrt{N}(\varepsilon/3)^2C_1\right\} \\ &\le 5\sqrt{N}|\mathcal{A}||\mathcal{X}|^{2m}\exp\left\{-\sqrt{N}(\varepsilon/3)^2C_1\right\},\end{aligned} \tag{41}$$

for all N sufficiently large so that $B\sqrt{\log|\mathcal{A}|/2}\cdot N^{-1/4}+BN^{-1/2}\left(\frac{1}{R}\log|\mathcal{A}|+d\right)<\varepsilon/3$. □

Appendix: Technical stages in the proof of Theorem 5

Proof of (36): By rescaling, it clearly suffices to establish (36) for $C=1$. To this end, fix $(\alpha_1,\ldots,\alpha_N)\in[0,1]^N$ and let $\delta=\max_{1\le j\le N}|a_j-b_j|$. Then

$$\begin{aligned}\frac{\sum_{j=1}^N e^{-a_j}\alpha_j}{\sum_{j=1}^N e^{-a_j}}-\frac{\sum_{j=1}^N e^{-b_j}\alpha_j}{\sum_{j=1}^N e^{-b_j}} &= \frac{\sum_{j=1}^N e^{-a_j}\alpha_j}{\sum_{j=1}^N e^{-a_j}}-\frac{\sum_{j=1}^N e^{-a_j}e^{a_j-b_j}\alpha_j}{\sum_{j=1}^N e^{-a_j}e^{a_j-b_j}} \\ &\le \frac{\sum_{j=1}^N e^{-a_j}\alpha_j}{\sum_{j=1}^N e^{-a_j}}-\frac{\sum_{j=1}^N e^{-a_j}e^{-|a_j-b_j|}\alpha_j}{\sum_{j=1}^N e^{-a_j}e^{|a_j-b_j|}} \\ &\le \frac{\sum_{j=1}^N e^{-a_j}\alpha_j}{\sum_{j=1}^N e^{-a_j}}-\frac{\sum_{j=1}^N e^{-a_j}e^{-\delta}\alpha_j}{\sum_{j=1}^N e^{-a_j}e^{\delta}}\end{aligned}$$

$$
\begin{aligned}
&= \frac{\sum_{j=1}^N e^{-a_j}\alpha_j(1-e^{-2\delta})}{\sum_{j=1}^N e^{-a_j}} \\
&\le (1-e^{-2\delta}) \le 2\delta. \qquad \text{(A 1)}
\end{aligned}
$$

Symmetrically, we also have

$$
\frac{\sum_{j=1}^N e^{-b_j}\alpha_j}{\sum_{j=1}^N e^{-b_j}} - \frac{\sum_{j=1}^N e^{-a_j}\alpha_j}{\sum_{j=1}^N e^{-a_j}} \le 2\delta. \qquad \text{(A 2)}
$$

Therefore

$$
\left| \frac{\sum_{j=1}^N e^{-a_j}\alpha_j}{\sum_{j=1}^N e^{-a_j}} - \frac{\sum_{j=1}^N e^{-b_j}\alpha_j}{\sum_{j=1}^N e^{-b_j}} \right| \le 2\delta. \quad \square \qquad \text{(A 3)}
$$

Proof of (39): Since the randomization sequence $\mathbf{U}$ is independent of the channel's noisy output, for almost every realization of the noisy sequence $\mathbf{Z}$, the difference

$$
d^N_{(E,D)}(\mathbf{x}) - \sum_{k=1}^n \mathbb{E}\left[d^{(k-1)l+1,kl}_{(E,D)}(\mathbf{x})|\mathcal{G}_Z\right] = \sum_{k=1}^n d^{(k-1)l+1,kl}_{(E,D)}(\mathbf{x}) - \sum_{k=1}^n \mathbb{E}\left[d^{(k-1)l+1,kl}_{(E,D)}(\mathbf{x})|\mathcal{G}_Z\right],
$$

under $\Pr\{\cdot|\mathcal{G}_Z\}$ is distributed as a sum of n zero-mean independent random variables bounded in magnitude by $l \cdot B$. Applying Hoeffding's inequality [3, Th. 8.1] gives a.s.

$$
\Pr\left\{ \frac{1}{N}\left(d^N_{(E,D)}(\mathbf{x}) - \sum_{k=1}^n \mathbb{E}\left[d^{(k-1)l+1,kl}_{(E,D)}(\mathbf{x})|\mathcal{G}_Z\right]\right) > \varepsilon \middle| \mathcal{G}_Z \right\} \le \exp\left\{-\frac{2N^2\varepsilon^2}{n(2lB)^2}\right\}. \qquad \text{(A 4)}
$$

Taking expectation over the above inequality completes the proof. $\square$

Proof of (40): Using the union bound generously, we have for all $\varepsilon > 0$

$$
\begin{aligned}
&\Pr\left\{ \sum_{k=1}^n \max_{(E',D')\in\mathcal{A}} \left| \hat{d}^{(k-1)l}_{(E',D')}(\mathbf{Z}) - d^{(k-1)l}_{(E',D')}(\mathbf{x}) \right| > \varepsilon \right\} \\
\le\ & \sum_{k=1}^n \sum_{(E',D')\in\mathcal{A}} \Pr\left\{ \left| \hat{d}^{(k-1)l}_{(E',D')}(\mathbf{Z}) - d^{(k-1)l}_{(E',D')}(\mathbf{x}) \right| > \varepsilon/n \right\} \\
=\ & \sum_{k=1}^n \sum_{(E',D')\in\mathcal{A}} \\
& \Pr\left\{ \left| \sum_{(\mathbf{a},\mathbf{b})\in\mathcal{X}^m\times\mathcal{X}^m} d(a_{j+1},h(\mathbf{b}))\cdot\left(\hat{N}^{(k-1)l}_{\mathbf{a},\mathbf{b}}(Z^{n+k}) - N^{(k-1)l}(x^{i+k}_{i-j}=\mathbf{a}, Z^{i+k}_{i-j}=\mathbf{b})\right)\right| > \varepsilon/n \right\} \\
\le\ & \sum_{k=1}^n \sum_{(E',D')\in\mathcal{A}} \sum_{(\mathbf{a},\mathbf{b})\in\mathcal{X}^m\times\mathcal{X}^m} \\
& \Pr\left\{ \left| \hat{N}^{(k-1)l}_{\mathbf{a},\mathbf{b}}(Z^{(k-1)l+d}) - N^{(k-1)l}(x^{i+k}_{i-j}=\mathbf{a}, Z^{i+k}_{i-j}=\mathbf{b})\right| > \frac{\varepsilon}{Bn|\mathcal{X}|^{2m}} \right\}. \qquad \text{(A 5)}
\end{aligned}
$$

To bound each of the summands in (A 5), recall that it was established in section B.1 that for each $(\mathbf{a},\mathbf{b}) \in \mathcal{X}^m\times\mathcal{X}^m$ and all t, $\hat{N}^t_{\mathbf{a},\mathbf{b}}(Z^{t+d}) - N^t(x^{i+k}_{i-j}=\mathbf{a}, Z^{i+k}_{i-j}=\mathbf{b})$ is a sum of t random variables $\sum_{i=1}^{t}\delta_i$, where the δ_i-s are a.s. bounded by $\tilde{B} = \max_{(\mathbf{a},\mathbf{b})\in\mathcal{X}^m\times\mathcal{X}^m} M(m)^{-1}_{\mathbf{a},\mathbf{b}}$ and where $\{\delta_j\}_{j\le i}$ is independent of $\{\delta_j\}_{j>i+m}$ for all i. For a given t, we can divide the set of indices $\{1,\ldots,t\}$ into $\lceil t/(2m)\rceil$ "intervals" of length $2m$ each (except, possibly, for the last one which may be shorter). The sum from 1 to t can be decomposed into a sum over those indices that belong to the first m indices of some interval, and a sum over the complementary set of indices. By the aforementioned independence property of $\{\delta_i\}$, each of these two sums is a sum of (at the most) $\lceil t/(2m)\rceil$ independent, zero-mean random variables bounded in magnitude by $m\cdot\tilde{B}$. Therefore, applying Hoeffding's inequality to each of these two sums and employing the union bound gives for all $\varepsilon>0$, $(\mathbf{a},\mathbf{b})\in\mathcal{X}^m\times\mathcal{X}^m$, and all t

$$\begin{aligned}\Pr\left\{\left|\hat{N}^t_{\mathbf{a},\mathbf{b}}(Z^{(k-1)l+d}) - N^t(x_{i-j}^{i+k}=\mathbf{a}, Z_{i-j}^{i+k}=\mathbf{b})\right| > \varepsilon\right\} &\le 4\exp\left\{-\frac{2(\varepsilon/2)^2}{\lceil t/(2m)\rceil (m\tilde{B})^2}\right\} \\ &\le 4\exp\left\{-\frac{\varepsilon^2}{2tm\tilde{B}^2}\right\}. \qquad \text{(A 6)}\end{aligned}$$

Consequently, each of the summands in the right side of (A 5) is upper bounded by

$$4\exp\left\{-\frac{\left(\frac{\varepsilon}{Bn|\mathcal{X}|^{2m}}\right)^2}{2(k-1)lm\tilde{B}^2}\right\} \le 4\exp\left\{-\frac{\left(\frac{\varepsilon}{Bn|\mathcal{X}|^{2m}}\right)^2}{2Nm\tilde{B}^2}\right\}, \qquad \text{(A 7)}$$

which finally gives

$$\Pr\left\{\sum_{k=1}^{n}\max_{(E',D')\in\mathcal{A}}\left|\hat{d}^{(k-1)l}_{(E',D')}(\mathbf{Z}) - d^{(k-1)l}_{(E',D')}(\mathbf{x})\right| > \varepsilon\right\} \le 4n|\mathcal{A}||\mathcal{X}|^{2m}\exp\left\{-\frac{\varepsilon^2}{Nn^2mB^2\tilde{B}^2|\mathcal{X}|^{4m}}\right\}. \qquad \text{(A 8)}$$

Inequality (40) is exactly (A 8) with the assignment $\varepsilon\to\varepsilon N/(2Bl\eta)$. $\square$

References

[1] N. Cesa-Bianchi, Y. Freund, D.P. Helmbold, D. Haussler, R. Schapire, and M.K. Warmuth, "How To Use Expert Advice," *Journal of the ACM*, 44(3):427-485, 1997.

[2] N. Cesa-Bianchi and G. Lugosi, "On Sequential Prediction of Individual Sequences Relative to a Set of Experts," in *Ann. Statist.* vol. 27(6), pp. 1865-1895, 1999.

[3] L. Devroye, L. Györfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition* Springer-Verlag, New York, 1996.

[4] Y. Ephraim and R. M. Gray, "A Unified Approach for Encoding Clean and Noisy Sources by means of Waveform and Autoregressive Model Vector Quantization", *IEEE Trans. Inform. Theory*, vol. 34, no. 4, pp. 826-834, July 1988.

[5] M. Feder, N. Merhav, and M. Gutman, "Universal Prediction of Individual Sequences," *IEEE Trans. Inform. Theory*, vol. 38, pp. 1258-1270, July 1992.

[6] R. M. Gray, "Time-Invariant Trellis Encoding of Ergodic Discrete-Time Sources with a Fidelity Criterion," *IEEE Trans. Inform. Theory*, vol. IT-23, no. 1, pp. 71-83, Jan. 1977.

[7] R. M. Gray and D. L. Neuhoff, "Quantization," *IEEE Trans. Inform. Theory*, vol. 44, no. 6, pp. 2325-2383, Oct. 1998.

[8] M. E. Hellman, "Convolutional Source Encoding," *IEEE Trans. Inform. Theory*, vol. IT-21, no. 6, pp. 651-656, Nov. 1975.

[9] N. S. Jayant and P. Noll, *Digital Coding of Waveforms: Principles and Applications to Speech and Video.* Englewod Cliffs, NJ: Prentice-Hall, 1984.

[10] I. Lev, "Universal Signal Enhancement by Coding," Master's Thesis, Technion, April 1996.

[11] T. Linder and G. Lugosi, "A Zero-Delay Sequential Scheme for Lossy Coding of Individual Sequences," to appear in *IEEE Trans. Inform. Theory.* (See also *Proc. 2000 IEEE ISIT, Sorrento, Italy, June 25-30* pp. 125.)

[12] N. Merhav and M. Feder, "Universal Schemes for Sequential Decision from Individual Data Sequences," *IEEE Trans. Inform. Theory*, vol. 39, no. 4, pp. 1280-1292, July 1993.

[13] B. Natarajan, "Filtering Random Noise via Data Compression," Data Compression Conference, DCC '93., pp. 60-69, 1993.

[14] E. Eberlein and M. S. Taqqu (eds.), *Dependence in Probability and Statistics*, 1986.

[15] T. Weissman and N. Merhav, "Universal Prediction of Individual Binary Sequences in the Presence of Noise," accepted to *IEEE Trans. Inform. Theory.*

[16] R. Zamir and M. Feder, "On Universal Quantization by Randomized Uniform/Lattice Quantizers," *IEEE Trans. Inform. Theory*, vol. IT-38, pp. 428-436, March 1992.

[17] J. Ziv, "Distortion-Rate Theory for Individual Sequences," *IEEE Trans. Inform. Theory*, vol. IT-26, no. 2, pp. 137-143, March 1980.

[18] J. Ziv, "On Universal Quantization," *IEEE Trans. Inform. Theory*, vol. IT-31, pp. 344-347, May 1985.