Source Coding Exponents for Zero-Delay Coding with Finite Memory

Neri Merhav* Ioannis Kontoyiannis†

October 7, 2001

Abstract

Fundamental limits on the source coding exponents (or large deviations) performance of zero-delay finite-memory (ZDFM) lossy source codes are studied. Our main results are the following: For any memoryless source, a suitably designed encoder that time-shares (at most two) memoryless scalar quantizers is as good as any time-varying fixed-rate ZDFM code, in that it can achieve the fastest exponential rate of decay for the probability of excess distortion. A dual result is shown to apply to the probability of excess code-length, among all fixed-distortion ZDFM codes with variable rate. Finally, it is shown that if the scope is broadened to ZDFM codes with variable rate *and* variable distortion, then a *time-invariant* entropy-coded memoryless quantizer (without time-sharing) is asymptotically optimal under a "fixed-slope" large deviations criterion (introduced and motivated here in detail) corresponding to a linear combination of the code-length and the distortion. These results also lead to single-letter characterizations for the source coding error-exponents of ZDFM codes.

Index Terms: causal source codes, zero-delay source codes, sliding-block codes, finite-memory codes, source coding exponents, large deviations, time-sharing.

*N. Merhav is with the Department of Electrical Engineering, Technion – Israel Institute of Technology, Technion City, Haifa 32000, Israel. E-mail: `merhav@ee.technion.ac.il`. Web: `http://tiger.technion.ac.il/users/merhav/`. N. Merhav was supported by the Israeli Science Foundation administered by the Israeli Academy of Sciences and Humanities.

†I. Kontoyiannis is with the Division of Applied Mathematics and the Department of Computer Science, Brown University, Box F, 182 George St., Providence, RI 02912, USA. Email: `yiannis@dam.brown.edu` Web: `www.dam.brown.edu/people/yiannis/` I. Kontoyiannis was supported by NSF grants #0073378-CCR and DMS-9615444.

1 Introduction

Zero-delay (or delayless) codes for lossy data compression form a subclass of the class of causal codes, namely, codes for which the reproduction data stream depends on the source data stream in a causal manner. Zero-delay codes are causal codes with the additional property that each reproduction symbol is entropy-coded separately (rather than in long blocks), and thus encoding and decoding can be carried out in an instantaneous manner. Causal codes that are not zero-delay are also of practical importance because they can still be implemented with low delay by harnessing an arithmetic code for the entropy-coding part. The "price of causality" (and hence also of zero delay) is well-known to be the inherent inability to approach the rate-distortion function at strictly positive distortion levels, although there are examples where causal codes may come fairly close, especially in the high resolution regime.

Lloyd [16], who was the first to study causal source codes, derived a lower bound on the best achievable compression rate for causal codes applied to the binary symmetric memoryless source with respect to (w.r.t.) the Hamming distortion measure. Subsequently, Piret [19] proved that this bound could be achieved by causal sliding-block codes with feedback, and Neuhoff and Gilbert [18] later showed that for memoryless sources, optimum rate-distortion performance among all causal source codes can be attained by time-sharing no more than two memoryless codes; see also [7] for an extension to Markov sources.

More recently, Linder and Zamir [15] have shown that, in the high resolution limit, Neuhoff and Gilbert's result continues to hold for all stationary sources with finite differential entropy. Therefore, in the case of high resolution the price of causality is the same as the "space filling loss" of the uniform scalar quantizer, i.e., $\frac{1}{2}\log_2(2\pi e/12) \approx 0.255$ bits. For the subclass of zero-delay codes, Ericson [4] and Gaarder and Slepian [5],[6] have shown that optimal performance is achieved by optimal (Lloyd-Max) scalar quantization for the given memoryless source. Recently, zero-delay [14] and limited-delay [21] codes have also been investigated in the individual-sequence setting.

While causal and zero-delay source codes have evidently been studied quite extensively under the average rate-distortion performance criterion (of expected code-length vs. expected distortion), we are not aware of any existing results on the "large deviations" or "error exponents" performance of these codes. In particular, it is natural to ask what are the best error-exponents achievable by causal or by zero-delay codes, in analogy to Marton's well-known error-exponents results for general block codes; see [17], and also [10] for a generalization to Gaussian sources. It is also worth noting that, so far, almost all of the error-exponent characterizations that have been derived (even for non-causal block codes), are asymmetric: Either the maximum distortion is kept fixed and the optimal exponent of decay for the probability of the code-length is examined, or, vice versa, the maximum code-length is fixed and the tail behavior of the distortion is examined. An exception to this is the recent work [22], where various trade-offs between the large deviations exponents of the code-length and the distortion are studied, with the rate and distortion being treated in a more symmetric manner.

1.1 Motivation and Discussion of Main Results

In this paper we derive results in the spirit of Neuhoff and Gilbert, but from a large deviations perspective. We consider the class of zero-delay finite-memory (ZDFM) codes without feedback, namely, those codes for which each reproduction symbol depends (possibly, in a time-varying manner) on the current input source symbol and on an arbitrarily large but fixed number of past source symbols. The two main problems we treat are descried is some more detail in the following two paragraphs.

In the first problem (Section 2), we begin by fixing an arbitrary point (R, D) in the rate-distortion plane, lying above the rate-distortion curve of the class of zero-delay codes. The objective is to characterize the best achievable rate of decay of the *excess distortion probability*, $\Pr\{\text{distortion} \geq nD\}$, n being the data length, subject to the constraint of fixed-rate coding, namely, among all codes whose *excess code-length probability*, $\Pr\{\text{code-length} \geq nR\}$, is zero. Our main result here is that for any memoryless source, a suitably-designed encoder that time-shares (at most two) memoryless scalar quantizers, is as good as any time-varying ZDFM code. A dual result is obtained in Subsection 3.1, where the roles of rate and distortion are interchanged: The objective now is to maximize the exponential decay rate of the excess code-length probability, subject to the constraint of fixed-distortion coding, i.e., the excess distortion probability being zero.

In the second problem (Section 3, Subsection 3.2), we consider a broader class of zero-delay codes, codes that have variable-rate *and* variable-distortion. Here we seek the fastest exponential rate of decay of the probability that a given linear combination of the code-length and the distortion exceeds a threshold nR_0, $R_0 > 0$ being a given constant, i.e.,

$$\Pr\Big\{(\text{code-length}) + \lambda \cdot (\text{distortion}) \geq nR_0\Big\}, \tag{1}$$

where $\lambda > 0$ is a constant which plays the role of a Lagrange multiplier. We henceforth refer to this probability as the *excess Lagrangian probability*. Our main finding here is that, again, it is enough to seek the optimum solution among entropy-coded memoryless scalar quantizers, but this time without even time-sharing, i.e., one time-invariant memoryless code maximizes the exponential decay rate of (1).

At this point a few remarks are in order with regard to the motivation behind the criterion of the excess Lagrangian probability, and on a meaningful choice of the parameters R_0 and λ. In geometric terms, the probability in (1) can be thought of as the probability that the rate-distortion performance of a given code lies in the half-plane above the line $R = R_0 - \lambda D$ in the rate-distortion plane (after normalizing by n). Note that the rate-distortion function $R(D)$ (in the ordinary, expected-rate, expected-distortion sense) of a discrete memoryless source (DMS) w.r.t. a *finite* collection of codes, such as the family of ZDFM codes considered here, is nonincreasing, convex, and *piecewise linear* in D.[1] Therefore, if we choose the parameters λ and R_0 so that the line $R = R_0 - \lambda D$ is parallel to and slightly above one of the linear segments of $R(D)$ (see Fig. 1), then the probability in (1) is the probability that the performance of a given code will exceed $R(D)$ by a certain amount at this linear segment. Now, due to the convexity of $R(D)$, the line $R = R_0 - \lambda D$ lies below $R(D)$ at essentially all its other linear segments, hence, any code that operates in the region of one of the other segments must have a very high probability of exceeding the line $R = R_0 - \lambda D$. Therefore, the excess Lagrangian criterion in (1) identifies codes that operate very close to the aimed linear segment with slope $-\lambda$.

In summary, while the first problem we described was posed in the fixed-rate regime, and its dual counterpart was in the fixed-distortion regime, the second problem corresponds to a *fixed slope* of the rate-distortion curve.

[1]To see why this true, observe that every time-varying code given by a certain sequence of members of the given finite family of codes, operates at a point in the rate-distortion plane given by a certain convex combination of the rate-distortion points of time-invariant codes in the family, where the weights correspond to the relative frequencies of usage of each member in the family (time-sharing). The rate-distortion function is then the lower boundary of the convex polygon spanned by these points.

1.2 Technical Aspects

We now give a brief description of the main ideas behind the proof of one of our main results (Theorem 1); the arguments in the proofs of our other results follow roughly the same outline. We also discuss a few other techncial issues.

Let us consider the case of ZDFM codes with a fixed-rate constraint. The cumulative distortion of such codes satisfies a large deviations principle (LDP), with a rate function that can be expressed in terms of the moment-generating function of the cumulative distortion.[2] First we show that the moment-generating function of the cumulative distortion is always minimized by a sequence of memoryless codes (due to the zero-delay assumption), and then we use the fact that the rate-function of the above LDP is determined by this moment-generating function to show that our objective function is optimized by memoryless encoders.

The main technical tools in the above argument are the non-asymptotic large deviations bounds we develop for arbitrary varying sources (Lemmas 1 and 2 in Appendix B). In a more general setting, asymptotic versions of these results have also been investigated in [11]. In the context of lossy data compression, similar LDPs for independent but not identically distributed random variables have been used in [2] and [12].

At this point it should be pointed out that in the asymptotic arguments of our main results it is always assumed that the memory length (henceforth denoted by $(k-1)$) of the competing ZDFM codes remains fixed, while the data length n grows without bound.[3] In other words, we assert that time-sharing of properly chosen memoryless quantizers is as good as any ZDFM code, as long as k is fixed and finite. But we do *not* claim that this continues to hold for arbitrary zero-delay codes with infinite memory, or even for finite-memory codes with $k = n$. It remains an open problem to assess the optimum attainable large deviations performance if all zero-delay source codes are allowed to compete. In fact, we conjecture that our main results may not hold in this case.

On the other hand, we expect that our results should continue to hold for a subclass of infinite-memory zero-delay codes, whose memory fades away sufficiently rapidly – so that

[2]Of course, the cumulative code-length, and any linear combination between the distortion and the code-length, also satisfy corresponding LDPs.

[3]It is these two assumptions – instantaneous entropy-coding and fixed memory length k – which give rise to the aforementioned LDPs. Similar results can also be obtained if the memory length k is allowed to grow sufficiently slowly with n.

these codes can be well-approximated by ZDFM codes with long enough, finite memory. For example, ZDFM codes with feedback, where the output reproduction symbol depends on finitely many past inputs and outputs (like in predictive encoders), may fall in this category. We will elaborate on this more in the sequel. The important point to keep in mind here is that the ZDFM codes with finite memory w.r.t. the input, but with no feedback, can be thought of as approximations of more general zero-delay codes with fading memory.

2 Fixed-Rate ZDFM Codes

We begin with some notation and definitions. Throughout the paper, random variables (RV's) are denoted by capital letters, specific realizations of them are denoted by the corresponding lower-case letters, and their alphabets are written as the respective calligraphic letters. For example, a RV X may take on any value $x \in \mathcal{X}$. For a sequence of letters $\{x_t\}$, the substring $(x_t, x_{t+1}, \ldots, x_\tau)$, where $t \leq \tau$, will be denoted by x_t^τ. A similar convention applies to RV's with capital letters replacing lower-case letters.

Consider a discrete memoryless source (DMS) $\ldots, X_{-1}, X_0, X_1, X_2, \ldots$ with distribution P over a finite alphabet $\mathcal{X}$ of size $A \geq 2$. Without loss of generality, we assume throughout that $P(x) > 0$ for all $x \in \mathcal{X}$. A *fixed-rate, zero-delay, finite-memory (ZDFM) encoder with memory of size* $(k-1)$, is a sequence of *reproduction functions* $\{f_t\}_{t\geq 1}$, where, at each time $t \geq 1$, f_t maps the source string x_{t-k+1}^t into a reproduction letter $\hat{x}_t = f_t(x_{t-k+1}^t) \in \hat{\mathcal{X}}$, where $\hat{\mathcal{X}}$ is the reproduction alphabet of size B. We assume throughout that $2 \leq B \leq A$. Note that, although the entire source output $X_{-\infty}^\infty$ is available to the encoder, only X_1^∞ is to be coded.

For each $t \geq 1$ and any past string $x_{t-k+1}^{t-1} \in A^{k-1}$, let $\hat{\mathcal{X}}_t(x_{t-k+1}^{t-1}) \subseteq \hat{\mathcal{X}}$ denote the range of f_t, so that $\hat{\mathcal{X}}_t(x_{t-k+1}^{t-1}) = \{f_t(x_{t-k+1}^t) : \ x_t \in A\}$. To ensure that the sequence $\{f_t\}$ yields a decodable code, we assume for now that the ranges $\hat{\mathcal{X}}_t = \hat{\mathcal{X}}_t(x_{t-k+1}^{t-1})$ are independent of the past x_{t-k+1}^{t-1}, yet they still may depend on t (this assumption will be partially relaxed shortly). Writing $\|f_t\| = |\hat{\mathcal{X}}_t|$ for the size of the range of f_t, the *instantaneous rate* of f_t is $\log \|f_t\|$ bits per symbol, i.e., no entropy-coding is performed on the reproduction symbols $\hat{x}_t$ (here and throughout the paper, log denotes the base 2 logarithm and ln denotes the natural logarithm).

A word of clarification is in order at this point: The term *fixed*-rate code here means that the instantaneous rate is independent of the data (as opposed to the variable-rate codes of Section 3), yet it still may vary with t. Our restriction will be over the total code-length over n points in time, $\sum_{t=1}^{n} \log \|f_t\|$, which we will assume to be always bounded by a fixed nR bits, (hence "fixed-rate").

The sequence $\{\hat{x}_t\}$, where $\hat{x}_t = f_t(x_{t-k+1}^t)$, $t = 1, 2, \ldots$, is referred to as the *reproduction* of the source sequence $\{x_t\}$. Note that there is only a finite number, namely $r(k) \stackrel{\triangle}{=} B^{A^k}$, of distinct reproduction functions with memory size $(k-1)$. The distortion between x_1^n and its reproduction $\hat{x}_1^n$ is defined as $\sum_{t=1}^{n} \rho(x_t, \hat{x}_t)$, where $\rho : \mathcal{X} \times \hat{\mathcal{X}} \to \mathbb{R}_+$ is an arbitrary single-letter distortion measure. Note that, since $\mathcal{X}$ and $\hat{\mathcal{X}}$ are finite sets, we have implicitly assumed that $D_{\max} \stackrel{\triangle}{=} \max_{x,\hat{x}} \rho(x, \hat{x})$ is finite.

The following elementary observation shows that, for *any* data sequence $\{x_t\}$ (and hence also any source, not necessarily memoryless), the performance of an arbitrary sequence of (possibly infinite-memory) reproduction functions, can also be achieved by a sequence of *memoryless* reproduction functions (corresponding to $k = 1$):

Remark 1.

For any reproduction function $f : \mathcal{X}^\infty \to \hat{\mathcal{X}}$ (with possibly infinite memory), there is a memoryless reproduction function $g : \mathcal{X} \to \hat{\mathcal{X}}$ with $\|g\| \leq \|f\|$, such that

$$\rho(x_t, g(x_t)) \leq \rho(x_t, f(x_{t-k+1}^t)), \tag{2}$$

for all source sequences x_{-t-k+1}^t.

To see why this is true, note that for any such f we can define a memoryless reproduction function g by the nearest-neighbor rule: $g(x) = \operatorname{argmin}_{\hat{x} \in \hat{\mathcal{X}}(f)} \rho(x, \hat{x})$. Then (2) trivially holds, and the range of g is clearly no larger than that of f.

2.1 Optimal Error Exponents for Fixed-Rate ZDFM Codes

We now turn to the more general case where $\hat{\mathcal{X}}_t(x_{t-k+1}^{t-1})$, the range of f_t given x_{t-k+1}^{t-1}, is allowed to depend on x_{t-k+1}^{t-1}, but $\|f_t\| = \|\hat{\mathcal{X}}_t(x_{t-k+1}^{t-1})\|$, the size of this range, is still independent of x_{t-k+1}^{t-1}.[4] Given a positive integer n and an average rate $R \in [0, \log B]$, we let $\mathcal{E}_n^k(R)$ denote the set all fixed-rate encoders $f_1^n = (f_1, f_2, \ldots, f_n)$ with memory size

[4]The reader may wonder how decoding can be possible when the reproduction alphabet depends on past source symbols, since these are not available to the decoder. This can be the case, for example, if the dependence is only via past reconstruction symbols, which are available to the decoder (e.g., in certain types of predictive encoders). Moreover, for the purpose of converse theorems, it is legitimate to allow any competing class of "genie-aided" decoders. And in any case, as Theorem 1 will show, the performance of even such unrealistic codes can be matched by memoryless ones.

$(k-1)$, achieving an average rate R:

$$\mathcal{E}_n^k(R) = \left\{ f_1^n : \sum_{t=1}^{n} \log \|f_t\| \leq nR \right\}.$$

For a given distortion level D, the *source coding exponent function for fixed-rate, time-varying ZDFM codes with memory size* $(k-1)$ is defined by

$$\mathcal{F}^k(D,R) \stackrel{\triangle}{=} \limsup_{n\to\infty} \left[-\frac{1}{n} \ln \min_{f_1^n \in \mathcal{E}_n^k(R)} \Pr\left\{ \sum_{t=1}^{n} \rho(X_t, f_t(X_{t-k+1}^t)) \geq nD \right\} \right].$$

As we show next, here it also turns out that memoryless codes are at least as good as codes with finite memory, although the reasons for this are no longer as obvious as in Remark 1. In Theorem 1 we show that $\mathcal{F}^k(D,R) = \mathcal{F}^1(D,R)$ for all $k < \infty$, we provide a single-letter characterization for $\mathcal{F}^k(D,R)$, and we point out the structure of optimal encoders.

To be precise, let $\mathcal{G}_1 = \{g_1, \ldots, g_{r_1}\}$, $g_i : A \to B$, $1 \leq i \leq r_1$, $r_1 \stackrel{\triangle}{=} r(1) = B^A$, be the set of all memoryless reproduction functions (corresponding to $k=1$). Let $\theta = (\theta_1, \ldots, \theta_{r_1})$ lie in the set $\Theta(R, r_1)$, where

$$\Theta(R,r) \stackrel{\triangle}{=} \left\{ \theta = (\theta_1, \ldots, \theta_r) : \sum_{s=1}^{r} \theta_s = 1, \sum_{s=1}^{r} \theta_s \log \|g_s\| \leq R, \ \theta_s \geq 0, \ s = 1, \ldots, r \right\},$$

and for $\theta \in \Theta(R, r_1)$ define

$$F(D,\theta) \stackrel{\triangle}{=} \sup_{\xi \geq 0} \left[\xi D - \sum_{s=1}^{r_1} \theta_s \ln E \exp\{\xi \rho(X, g_s(X))\} \right]. \tag{3}$$

Theorem 1, our main result in this section, states that $\mathcal{F}^k(D,R)$ is equal to the following expression:

$$F^*(D,R) = \sup_{\theta \in \Theta(R,r_1)} F(D,\theta). \tag{4}$$

Remarks:

2. Since $F(D,\theta)$ is a convex function of D for every θ, it is continuous in D in the interior of the domain where it is finite. Similarly, $F(D,\theta)$ is convex in θ for each D, and hence continuous in θ in the (relative) interior of the set of θ's for which it is finite. As we discuss in Appendix A, $F(D,\theta)$ is finite iff $D \leq \sum_s \theta_s D_s$, where

$$D_s = \max_x \rho(x, g_s(x)). \tag{5}$$

Moreover, $F(D,\theta)$ is left-continuous at the boundary point $D = \sum_s \theta_s D_s$.

3. Let $R \in [0, \log B]$. As explained in Appendix A, $F^*(D,R)$ is infinite when D is greater than $D^*(R)$, where

$$D^*(R) \stackrel{\triangle}{=} \inf_{\theta \in \Theta(R,r_1)} \sum_s \theta_s D_s, \tag{6}$$

with D_s as in (5). But in that range we also have $\mathcal{F}^k(D,R) = \infty$. To see this, note that for n large enough we can approximate the achieving θ^* in (6) by an n-type $\theta^{(n)} \in \Theta(R, r_1)$, such that $D > \sum_s \theta_s^{(n)} D_s$; the existence of an achieving θ^* is of course obvious from (6). Therefore, concatenating $(n\theta_s^{(n)})$ copies of each of the memoryless reproduction functions g_s, $s = 1, 2, \ldots, r_1$, we obtain a memoryless ZDFM encoder $f_1^n \in \mathcal{E}_n^1(R)$, whose distortion on *any* data sequence x_1^n is bounded above by $\sum_s \theta_s^{(n)} D_s < D$. Thus, f_1^n achieves $\Pr\{\text{distortion} \geq nD\} = 0$, which implies that $\mathcal{F}^1(D,R) = \infty$, and, therefore, $\mathcal{F}^k(D,R) = F^*(D,R) = \infty$ whenever D exceeds $D^*(R)$.

In view of Remark 3 above, in Theorem 1 we restrict attention to the interesting range of distortion values D below $D^*(R)$.

Theorem 1 *For all $R \in (0, \log B)$, all $D \in (0, D^*(R))$, and any positive integer $k < \infty$,*

$$\mathcal{F}^k(D,R) = F^*(D,R).$$

Remarks:

4. Since $F(D,\theta)$ is convex in D it follows that $F^*(D,R)$ is also convex in D, and also it is easily seen to be concave in R. Therefore, $F^*(D,R)$ is continuous in both D and R, as long as they lie in the interior of the corresponding intervals on which $F^*(D,R)$ is finite.

5. An alternative expression for $F(D,\theta)$, of a more information-theoretic flavor, is the following: For every $s = 1, \ldots, r_1$, let P_s denote the probability mass function (PMF) of $Y_s \stackrel{\triangle}{=} \rho(X, g_s(X))$, and let Q_s be an arbitrary PMF with the same support. Then, alternatively we can define

$$F(D,\theta) = \inf \sum_{s=1}^{r_1} \theta_s D(Q_s \| P_s), \tag{7}$$

where the infimum is over all sets of PMF's $\{Q_s\}$ such that $\sum_{s=1}^{r_1} \theta_s E_{Q_s} Y_s \geq D$, with $E_{Q_s}(\cdot)$ denoting the expectation under Q_s. In Appendix A we outline a proof of the fact that the two expressions in (4) and (7) are indeed equal. Also note that the expression in (7) can be arrived at by using the method of types to prove Lemmas 1 and 2 in Appendix B.

6. The theorem indicates that time-sharing among memoryless reproduction functions (in proportions corresponding to the optimal θ^* for which $F^*(D,R) = F(D,\theta^*)$) achieves the best attainable distortion error-exponent within the class of *all* fixed-rate, time-varying ZDFM codes with finite memory. In fact, after the proof we show that time-sharing among no more than two memoryless reproduction functions is always sufficient.

7. In view of Remark 1, if we only consider ZDFM codes f_t whose ranges $\hat{\mathcal{X}}_t(x_{t-k+1}^{t-1})$ do not depend on the past x_{t-k+1}^{t-1}, then the result of the theorem remains valid in the case of infinite-memory codes, corresponding to $k = \infty$.

Proof. Choose and fix a rate $R \in (0, \log B)$, a distortion level $D \in (0, D^*(R))$, and an integer k. Observe that $D^*(R)$ is always finite, and that, as a function of R, it is nonincreasing and convex, which implies that it is also continuous for $R > 0$. Let an arbitrarily small $\epsilon > 0$ be given, and assume it is small enough so that $D < D^*(R+\epsilon)$ and $D + \epsilon < D^*(R)$. All the above quantities will remain fixed throughout the proof.

Direct part. We first prove that $\mathcal{F}^k(D,R) \geq F^*(D,R)$. For any sequence of memoryless reproduction functions $(f_1, \ldots, f_n)$, the process $\{Z_t\}$, where $Z_t = \rho(X_t, f_t(X_t))$, $f_t \in \mathcal{G}_1$, is clearly a memoryless arbitrarily varying source (AVS), with r_1 states corresponding to the different choices of f_t. For each $s = 1, 2, \ldots, r_1$, let θ_s denote the relative frequency of the reproduction function g_s among $(f_1, \ldots, f_n)$.

Pick a $\theta^* \in \Theta(R, r_1)$ such that $F(D-\epsilon, \theta^*) \geq F^*(D-\epsilon, R) - \epsilon$. For each n we can approximate θ^* by an n-type $\theta^{(n)} \in \Theta(R, r_1)$, where the sequence $\{\theta^{(n)}\}$ is chosen so that $\theta^{(n)} \to \theta^*$ and $F(D-\epsilon, \theta^{(n)}) \to F(D-\epsilon, \theta^*)$, as $n \to \infty$. [To see that this can be done, note that $F(D-\epsilon, \cdot)$ is continuous in the (relative) interior of $\Theta(R+\epsilon, r_1)$. Also, its restriction to the interior of any of the lower-dimensional boundaries defined by combinations of the constraints $\{\theta_s = 0\}$ is continuous as well. Finally, observe that if some of the components of θ^* are actually equal to zero, then we can take the corresponding components of $\theta^{(n)}$ to be zero too, for all n.]

For every n we consider a code consisting of a sequence $(f_1^*, \ldots, f_n^*)$ of reproduction functions, where $n\theta_s^{(n)}$ of them are equal to g_s, for each $s = 1, 2, \ldots, r_1$. Obviously, $(f_1^*, \ldots, f_n^*) \in \mathcal{E}_n^1(R)$. We are interested in assessing the probability of the event $\{\sum_{t=1}^n Z_t \geq nD\}$, where $\{Z_t\}$ is the AVS defined by $Z_t = \rho(X_t, f_t^*(X_t))$, $t \geq 1$. This is a problem that is generically handled in Appendix B. In fact, applying Lemma 2 of Appendix B with $\Delta_{\max} = D_{\max}$, $r = r_1$, and $\phi = F$, we have for every positive integer ℓ:

$$\Pr\left\{\sum_{t=1}^n \rho(X_t, f_t^*(X_t)) \geq nD\right\} \leq \exp\left\{-nF\left(D - D_{\max}\left[\frac{\ell}{n} + \frac{r_1}{\ell}\right], \theta^{(n)}\right) + \ell F(D, \theta^{(n)})\right\}.$$

[Note that the assumption $D + \epsilon < D^*(R)$ guarantees that we can apply Lemma 2, at least for large enough n.] Since $F(D, \theta)$ is nondecreasing in D (for each fixed θ), we can pick ℓ and then M large enough, such that, for all $n \geq M$, the above right-hand-side (r.h.s.) is bounded above by

$$\exp\left\{-nF\left(D - \epsilon, \theta^{(n)}\right) + \ell F^*(D, R)\right\}.$$

Letting n to infinity (for a fixed ℓ) and recalling the choice of the sequence $\{\theta^{(n)}\}$, yields that $\mathcal{F}^1(D, R) \geq F(D - \epsilon, \theta^*) \geq F^*(D - \epsilon, R) - \epsilon$. And since $\epsilon > 0$ was arbitrary, by the continuity of $F^*(\cdot, R)$ we get $\mathcal{F}^1(D, R) \geq F^*(D, R)$. Finally, since memoryless reproduction functions form a subset of the set of finite-memory reproduction functions, then by definition, $\mathcal{F}^k(D, R) \geq F^*(D, R)$ for all k,

Converse part. We now prove that $\mathcal{F}^k(D, R) \leq F^*(D, R)$. Fix an integer $p \gg k$ large enough so that

$$D < \left(\frac{p}{p+k}\right) D^*(R + \epsilon), \tag{8}$$

and let $N \gg q \stackrel{\triangle}{=} p + k - 1$ be so large that

$$\min_{f_1^N \in \mathcal{E}_N^k(R)} \Pr\left\{\sum_{t=1}^N \rho(X_t, f_t(X_{t-k+1}^t)) \geq ND\right\} \leq \exp\{-N[\mathcal{F}^k(D, R) - \epsilon]\}.$$

Denote by $(f_1^*, f_2^*, \ldots, f_N^*)$ a sequence of reproduction functions attaining the minimum in the left-hand-side (l.h.s.) of the last inequality. Let $n = \lfloor N/q \rfloor$, and define

$$Z_t = \sum_{\tau=(t-1)q+k}^{tq} \rho(X_\tau, f_\tau^*(X_{\tau-k+1}^\tau)), \quad t = 1, 2, \ldots, n. \tag{9}$$

Note that $\{Z_t\}$ are independent RV's since they are functions of non-overlapping q-blocks of the independent X_τ's. Since there are $r(k) = B^{A^k}$ distinct reproduction functions with memory of size $(k-1)$, it is clear that $\{Z_t\}$ is a memoryless AVS with $r(k,p) \stackrel{\triangle}{=} [r(k)]^p = B^{A^k p}$ states, corresponding to the $[r(k)]^p$ different possible combinations of p reproduction functions $\{f_\tau^*\}_{\tau=(t-1)q+k}^{tq}$.

Now, for each $\boldsymbol{s} \in \{1,\ldots,r(k,p)\}$ let $\boldsymbol{\theta_s}$ denote the proportion of times $t \in \{1,2,\ldots,n\}$ during which $\{f_\tau^*\}_{\tau=(t-1)q+k}^{tq}$ coincides with the particular set of p reproduction functions indexed by $\boldsymbol{s}$, and let $\boldsymbol{\theta}$ denote the vector $\{\boldsymbol{\theta_s}\}$. For convenience, we visualize the state $\boldsymbol{s}$ as a p-vector $(\sigma_1,\ldots,\sigma_p)$, where each $\sigma_\tau \in \{1,\ldots,r(k)\}$ designates the index of the τ-th reproduction function within the class $\mathcal{G}_k$ of reproduction functions with memory of size $(k-1)$. We will further assume that $n \gg r(k,p)$, which means that $N \gg qB^{A^k p}$.

Moreover, we write $R_\sigma \stackrel{\triangle}{=} \log\|g_\sigma\|$ for the (instantaneous) rate of the reproduction function g_σ, and we take q (and hence n) large enough so that the average rate achieved by $\boldsymbol{\theta}$, call it $\tilde{R}$, satisfies

$$\tilde{R} \stackrel{\triangle}{=} \sum_{\text{all } \boldsymbol{s}=(\sigma_1,\ldots,\sigma_p)} \boldsymbol{\theta_s} \sum_{\tau=1}^{p} R_{\sigma_\tau} \stackrel{(a)}{\leq} \frac{NR}{n(q-k+1)} < R+\epsilon\,, \tag{10}$$

where (a) follows from the observation that, having omitted $N-n(q-k+1)$ of the original N reproduction functions, the average rate cannot increase by more than a factor of $N/[n(q-k+1)]$.

Let us define

$$F_p(D,\boldsymbol{\theta}) \stackrel{\triangle}{=} \sup_{\xi\geq 0}\left[\xi D - \sum_{\boldsymbol{s}=1}^{r(k,p)} \boldsymbol{\theta_s} \ln E\exp\left\{\xi\sum_{\tau=1}^{p}\rho(X_\tau, g_{\sigma_\tau}(X_{\tau-k+1}^{\tau}))\right\}\right]$$

where $\boldsymbol{\theta} \in \Theta(\tilde{R}, r(k,p))$. Our first step will be to show that

$$F_p(pD,\boldsymbol{\theta}) \leq p\cdot F^*(D,\tilde{R})\,, \quad \text{for all } \boldsymbol{\theta} \in \Theta(\tilde{R}, r(k,p)).$$

Let us denote the set of all p-tuples of functions from $\mathcal{G}_k$, by $\mathcal{G}_k^p = [\mathcal{G}_k]^p$. First of all, we show that for every $\xi \geq 0$ and every $(g_1,\ldots,g_p) \in \mathcal{G}_k^p$, there exists $(\tilde{g}_1,\ldots,\tilde{g}_p) \in \mathcal{G}_1^p$ such that

$$E\exp\left\{\xi\sum_{j=1}^{p}\rho(X_j,\tilde{g}_j(X_j))\right\} \leq E\exp\left\{\xi\sum_{j=1}^{p}\rho(X_j, g_j(X_{j-k+1}^{j}))\right\}, \tag{11}$$

and at the same time $\|\tilde{g}_j\| \leq \|g_j\|$ for all $j=1,\ldots,p$. To see why this is true, we use a simple idea similar to the one used by Stiglitz [20], where he proved that memoryless channels are least favorable in terms of the jamming game of the error-exponent: Let us rewrite the r.h.s. of eq. (11), call it W, as follows:

$$\begin{aligned}W &= \sum_{x^0_{-k+2}} P(x^0_{-k+2})\sum_{x_1} P(x_1)\exp\{\xi\rho(x_1,g_1(x^0_{-k+2},x_1))\}\times\\ &\quad \sum_{x_2} P(x_2)\exp\{\xi\rho(x_2,g_2(x^1_{-k+1},x_2))\}\times\cdots\times\\ &\quad \sum_{x_p} P(x_p)\exp\{\xi\rho(x_p,g_p(x^{p-1}_{p-k},x_p))\}.\end{aligned} \tag{12}$$

Consider first the part of the expression that depends on g_p, namely, only the last summation over x_p. Note that in this part of the expression, x^{p-1}_{p-k} can simply be thought of as an index of a function of x_p from $\mathcal{X}\to\hat{\mathcal{X}}$ (and the dependence on the past is only via this index). Therefore, for any x^{p-1}_{p-k}, this summation over $\{x_p\}$ cannot be smaller than the minimum of

$$\sum_x P(x)\exp\{\xi\rho(x,\tilde{g}(x))\}$$

over all $\tilde{g}\in\mathcal{G}_1$ such that $\|\tilde{g}\|\le\|g_p\|$, or equivalently, $\log\|\tilde{g}\|\le\log\|g_p\|=R_p$. Note that the minimizer $\tilde{g}_p$ depends on p only via R_p. Denoting the value of the minimum by $m(R_p)$, we have bounded the r.h.s. of eq. (11) below by

$$\begin{aligned}&m(R_p)\cdot\sum_{x^0_{-k+2}} P(x^0_{-k+2})\sum_{x_1} P(x_1)\exp\{\xi\rho(x_1,g_1(x^0_{-k+2},x_1))\}\times\\ &\sum_{x_2} P(x_2)\exp\{\xi\rho(x_2,g_2(x^1_{-k+1},x_2))\}\times\cdots\times\\ &\sum_{x_{p-1}} P(x_{p-1})\exp\{\xi\rho(x_{p-1},g_{p-1}(x^{p-2}_{p-k-1},x_{p-1}))\},\end{aligned}$$

for which the summation over x_{p-1} is similarly bounded below by $m(R_{p-1})$, and so on. Continuing this way until the summation over x_1, our conclusion is that we found a p-tuple of memoryless reproduction functions $(\tilde{g}_1,\ldots,\tilde{g}_p)\in\mathcal{G}_1^p$ for which the moment-generating function of the associated distortion, at a given value of ξ, does not exceed that of a given $(g_1,\ldots,g_p)\in\mathcal{G}_k^p$, while maintaining the instantaneous rates $R_1,\ldots,R_p$, and hence also the total rate $R_1+\cdots+R_p$. It follows immediately that

$$\begin{aligned}F_p(pD,\boldsymbol{\theta}) &\le \sup_{\boldsymbol{\theta}\in\Theta(\tilde{R},r(k,p))} F_p(pD,\boldsymbol{\theta})\\ &\le \sup_{\boldsymbol{\theta}\in\Theta(\tilde{R},r(1,p))}\sup_{\xi\ge 0}\left[\xi pD-\sum_{\boldsymbol{s}=1}^{r(1,p)}\boldsymbol{\theta}_{\boldsymbol{s}}\ln E\exp\{\xi\sum_{j=1}^{p}\rho(X_j,g_{\sigma_j}(X_j))\}\right]\\ &= \sup_{\boldsymbol{\theta}\in\Theta(\tilde{R},r(1,p))}\sup_{\xi\ge 0}\left[\xi pD-\sum_{\boldsymbol{s}=1}^{r(1,p)}\boldsymbol{\theta}_{\boldsymbol{s}}\ln\prod_{j=1}^{p}E\exp\{\xi\rho(X_j,g_{\sigma_j}(X_j))\}\right]\end{aligned}$$

$$
\begin{aligned}
&= \sup_{\boldsymbol{\theta}\in\Theta(\tilde{R},r(1,p))} \sup_{\xi\geq 0}\left[\xi p D - \sum_{\boldsymbol{s}=1}^{r(1,p)} \boldsymbol{\theta}_{\boldsymbol{s}} \sum_{j=1}^{p} \ln E\exp\{\xi\rho(X_j, g_{\sigma_j}(X_j))\}\right] \\
&= \sup_{\boldsymbol{\theta}\in\Theta(\tilde{R},r(1,p))} \sup_{\xi\geq 0}\left[\xi p D - p\sum_{\sigma=1}^{r_1}\left(\frac{1}{p}\sum_{j=1}^{p}\sum_{\boldsymbol{s}:\sigma_j=\sigma}\boldsymbol{\theta}_{\boldsymbol{s}}\right)\ln E\exp\{\xi\rho(X, g_\sigma(X))\}\right] \\
&= p\cdot \sup_{\theta\in\Theta(\tilde{R},r_1)} \sup_{\xi\geq 0}[\xi D - \sum_{\sigma=1}^{r_1}\theta_\sigma \ln E\exp\{\xi\rho(X,g_\sigma(X))\}] \\
&= p\cdot F^*(D,\tilde{R}). \qquad (13)
\end{aligned}
$$

¿From the choice of N and by an application of Lemma 1 (Appendix B) with $\{Z_t\}$ as defined above, $\Delta_{\max} = pD_{\max}$, $r = r(k,p)$, and $\phi = F_p$, we have the following: Let $\delta > 0$ be sufficiently small, $\ell \geq \lceil 2pr(k,p)D_{\max}/\delta\rceil$,

$$
n \geq \left\lceil \frac{\ell^3 p^2 D_{\max}^2}{\delta^2 \ln 2}\right\rceil,
$$

and assume for the moment that D lies in the range within which Lemma 1 can be applied (we come back to justify this assumption at the end of the proof). Then,

$$
\begin{aligned}
&\exp\{-N[\mathcal{F}^k(D,R)-\epsilon]\} \qquad (14) \\
&\geq \Pr\left\{\sum_{t=1}^{N}\rho(X_t, f_t^*(X_{t-k+1}^t)) \geq ND\right\} \\
&\geq \Pr\left\{\sum_{t=1}^{n} Z_t \geq n(q+1)D\right\} \\
&\stackrel{(a)}{\geq} \exp\Big\{-n\Big[F_p\Big((q+1)D + \zeta_\ell(\delta p, pD_{\max}),\boldsymbol{\theta}\Big) + \nu_n(\ell,\delta p, pD_{\max}, r(k,p))\Big]\Big\} \\
&\geq \exp\Big\{-\frac{N}{q}\Big[F_p\Big((p+k)D + \zeta_\ell(\delta p, pD_{\max}),\boldsymbol{\theta}\Big) + \nu_n(\ell,\delta p, pD_{\max}, r(k,p))\Big]\Big\}
\end{aligned}
$$

and using (13),

$$
\begin{aligned}
&\mathcal{F}^k(D,R) \\
&\leq \frac{1}{q}\Big[F_p((p+k)D + \zeta_\ell(\delta p, pD_{\max}),\boldsymbol{\theta}) + \nu_n(\ell,\delta p, pD_{\max}, r(k,p))\Big] + \epsilon \\
&\leq \frac{p}{p+k-1}\left[F^*\left((1+k/p)D + \frac{1}{p}\zeta_\ell(\delta p, pD_{\max}),\tilde{R}\right) + \nu_n(\ell,\delta p, pD_{\max}, r(k,p))\right] + \epsilon \\
&\leq \frac{p}{p+k-1}\left[F^*\left((1+k/p)D + \frac{1}{p}\zeta_\ell(\delta p, pD_{\max}),R+\epsilon\right) + \nu_n(\ell,\delta p, pD_{\max}, r(k,p))\right] + \epsilon,
\end{aligned}
$$

where the last step follows from recalling (10) and that $F^*(D,R)$ is an increasing function of R. Taking the limit $N\to\infty$ (and hence $n\to\infty$), then $\ell\to\infty$, and finally $p\to\infty$, we get, by the continuity of $F^*(D,R)$ in D, that $\mathcal{F}^k(D,R)\leq F^*(D,R+\epsilon)+\epsilon$. Since $\epsilon>0$ is arbitrarily small and $F^*(D,R)$ is also continuous in R, we have $\mathcal{F}^k(D,R) \leq F^*(D,R)$. This completes the proof, subject to justifying the application of Lemma 1 in step (a) above.

For that, it suffices to show that $D < D^*(R)$ lies in the range allowed by Lemma 1, namely that $(q+1)D < \overline{\Delta}_p(\theta)$, where

$$\overline{\Delta}_p(\boldsymbol{\theta}) \stackrel{\triangle}{=} \sum_{\boldsymbol{s}=1}^{r(k,p)} \boldsymbol{\theta}_{\boldsymbol{s}} \max_{x_1^q} \sum_{j=k}^{q} \rho(x_j, g_{\sigma_j}(x_{j-k+1}^j)).$$

Moving the sum over j in front of the maximization, fixing an arbitrary x_1^{k-1}, and sequentially maximizing over x_j, for $j = k, \ldots, q$, with a slight abuse of notation we get that

$$\overline{\Delta}_p(\boldsymbol{\theta}) \geq \sum_{\boldsymbol{s}=1}^{r(k,p)} \boldsymbol{\theta}_{\boldsymbol{s}} \sum_{j=k}^{q} \max_{x_j} \rho(x_j, g_{\sigma_j}(x_j * x_{j-k+1}^{j-1})), \tag{15}$$

where $*$ denotes concatenation of strings. For each j, after x_{j-k+1}^{j-1} has been fixed, we can think of $g_{\sigma_j}(x_j * x_{j-k+1}^{j-1}))$ as a memoryless reproduction function applied to x_j, call it $g'_{\sigma_j}(x_j)$, so that

$$\max_{x_j} \rho(x_j, g_{\sigma_j}(x_j * x_{j-k+1}^{j-1})) = \Delta_{\sigma_j} \stackrel{\triangle}{=} \max_{x} \rho(x, g'_{\sigma_j}(x)).$$

Note also that each such memoryless g'_{σ_j} has rate no greater that of the corresponding g_{σ_j}. Rearranging the terms in (15) and rewriting the memoryless reproduction functions g' in terms of the earlier enumeration g_s, $s = 1, 2, \ldots, r_1$, we can rewrite the lower bound in (15) as

$$\overline{\Delta}_p(\boldsymbol{\theta}) \geq \sum_{s=1}^{r_1} \theta'_s D_s,$$

with D_s as in (5), and the vector θ' sums to $(q-k+1) = p$. Or, alternatively,

$$\overline{\Delta}_p(\boldsymbol{\theta}) \geq p \sum_{s=1}^{r_1} \theta''_s D_s$$

where $\sum_\sigma \theta''_\sigma = 1$. Moreover, tracing these steps backwards and recalling (10), it is easy to see that we actually have $\theta'' \in \Theta(R+\epsilon, r_1)$. Therefore, by the definition of $D^*(\cdot)$,

$$\overline{\Delta}_p(\boldsymbol{\theta}) \geq pD^*(R+\epsilon) > (q+1)D,$$

where the last inequality follows from (8). This shows that $(q+1)D < \overline{\Delta}_p(\boldsymbol{\theta})$, as required, thereby completing the proof. $\square$

Remarks:

8. Observe that in the proof of the fact that the exponential moment of the cumulative distortion is minimized by a memoryless code, we have not used the fact that k is finite. This means that, if the objective function was the moment-generating function of the distortion (instead of the probability that the distortion exceeds nD), then the finite memory limitation could be relaxed.

9. Note that in the definition of $\{Z_t\}$ in the converse part we have created "guard spaces" of k time units between successive segments in order to avoid dependence. If, more generally, the code has infinite memory that fades away fast enough to make the process $\{\rho(X_t, \hat{X}_t)\}$ sufficiently rapidly mixing, then for sufficiently large "guard spaces" (depending only on the code and not on N), the distribution of $\{Z_t\}$ can be well-approximated by a product distribution on an exponential scale, and our converse result will continue to hold.

2.2 Time-Sharing Between Two Codes is Enough

Here we show that $F^*(D,R)$ can be attained be a vector θ with no more than two nonzero components, and therefore, at most two memoryless codes need to be time-shared.

Let us rewrite $F^*(D,R)$ as

$$F^*(D,R) = \max_{\xi \geq 0} \left[\xi D - \min_{\theta \in \Theta(R, r_1)} \sum_{s=1}^{r_1} \theta_s \ln E \exp\{\xi \rho(X, g_s(X))\} \right].$$

For a given value of ξ consider the inner minimization of $\sum_{s=1}^{r_1} \theta_s \ln E \exp\{\xi \rho(X, g_s(X))\}$ over θ. With no loss of optimality, instead of considering all $r_1 = B^A$ members of $\mathcal{G}_1$, one may confine attention only to the subset $\mathcal{G}^*$ of $r^* \leq B$ *representative* reproduction functions that have the smallest value of $\ln E \exp\{\xi \rho(X, g(X))\}$ among all g of the same rate or less, including all possible time-sharings among such g with the same average rate or less. In other words, a reproduction function g is representative iff $\ln E \exp\{\xi \rho(X, g(X))\} = \mu(\log \|g\|)$, where

$$\mu(R) \stackrel{\triangle}{=} \min \left\{ \sum_{s=1}^{r_1} \theta_s \ln E \exp\{\xi \rho(X, g_s(X))\} : \sum_{s=1}^{r_1} \theta_s \log \|g_s\| \leq R \right\}.$$

Let us re-index the representative reproduction functions as $g_1, \ldots, g_{r^*}$, according to increasing rates. Writing $\theta = (\theta_1, \ldots, \theta_{r^*})$, $R_s \stackrel{\triangle}{=} \log \|g_s\|$, and $\mu_s \stackrel{\triangle}{=} \mu(R_s)$, we have the following linear programming problem:

$$\min_{\theta} \sum_{s=1}^{r^*} \theta_s \mu_s$$

subject to:

$$\begin{aligned} \theta_s &\geq 0, \quad s = 1, \ldots, r^*, \\ \sum_{s=1}^{r^*} \theta_s &= 1, \\ \sum_{s=1}^{r^*} \theta_s R_s &\leq R. \end{aligned}$$

The necessary and sufficient Kuhn-Tucker conditions for the optimality of $\theta^* = (\theta_1^*, \ldots, \theta_{r^*}^*)$ are that there exists a constant $\alpha \geq 0$ equal to zero if $\sum_{s=1}^{r^*} \theta_s^* R_s < R$, and a constant β such that, for all $s = 1, \ldots, r^*$,

$$\mu_s + \alpha R_s \geq \beta,$$

with equality for all s for which $\theta_s^* > 0$. Note, however, that the minimum of $\sum_{s=1}^{r^*} \theta_s \mu_s$ subject to the aforementioned constraints is a convex function of R, and as such it is strictly decreasing in the range $R \in (0, \log B)$, which means that the inequality constraint $\sum_{s=1}^{r^*} \theta_s^* R_s \leq R$ is always attained with equality in this range. By the definition of the subset of representative reproduction functions, it is clear that the piecewise-linear function formed by the straight lines that connect (R_1, μ_1) to (R_2, μ_2), then (R_2, μ_2) to (R_3, μ_3), etc., and finally, (R_{r^*-1}, μ_{r^*-1}) to (R_{r^*}, μ_{r^*}), is a convex function of R, and the slopes of these straight lines are $\gamma_s = -(\mu_s - \mu_{s+1})/(R_{s+1} - R_s)$, $s = 1, \ldots, r^* - 1$. Convexity then implies that $\gamma_1 \leq \gamma_2 \leq \ldots \leq \gamma_{s-1}$. Now, unless two or more elements γ_r are equal, eq. (2.2) cannot be attained with equality simultaneously by more than two values of s, the reason being that for any pair (s, s') for which there is equality, α must be the negative slope of straight line that connects (R_s, μ_s) to $(R_{s'}, \mu_{s'})$, which is, in general, different for every such pair. This means that at most two coordinates of θ^* need to be strictly positive. In particular, any given rate $R \in [0, \log B]$ can be represented as $\eta R_s + (1 - \eta) R_{s+1}$, for some $\eta \in [0, 1]$ and some $s = 1, \ldots, r^* - 1$, and then $\theta_s^* = \eta$, $\theta_{s+1}^* = 1 - \eta$, and all the remaining coordinates of θ^* are zero.

Finally, recall that the above optimization over θ is defined for a given value of ξ, and so, the indices of the two memoryless reproduction functions that take part in the time-sharing may depend on ξ. Specifically, we can write $F^*(D,R)$ as follows:

$$F^*(D,R) = \max_{\xi \geq 0}[\xi D - \eta \ln E \exp\{\xi \rho(X, g_{s_1(\xi)}(X)\} - (1-\eta)\ln E \exp\{\xi \rho(X, g_{s_2(\xi)}(X)\}].$$

Thus, after carrying out the optimization over ξ, the optimal reproduction functions that are time-shared are $g_{s_1(\xi^*)}$ and $g_{s_2(\xi^*)}$, where ξ^* achieves the maximum in the last expression.

3 Variable-Rate ZDFM Codes

In this section, we consider the problem of determining the best achievable error-exponents for zero-delay *variable-rate* codes with finite memory. The model we adopt is the same as before, with the difference that we now allow for variable-rate lossless compression (or "entropy-coding") of the reproduction symbols $\{\hat{x}_t\}$. In analogy to the finite-memory assumption that we made regarding the reproduction functions $\{f_t\}$, we will also assume that the associated entropy-coders have finite memory: Every reproduction symbol $\hat{x}_t$ will be described using $L_t(\hat{x}_t|\hat{x}_{t-k+1}^{t-1})$ bits, depending on the $(k-1)$ previously decoded reproduction outcomes,[5] where $L_t(\cdot|\hat{x}_{t-k+1}^{t-1}) : \hat{\mathcal{X}}_t \to \boldsymbol{Z}_+$, is the length-function of a uniquely decipherable, fixed-to-variable length code, satisfying Kraft's inequality:

$$\sum_{\hat{x}_t \in \hat{\mathcal{X}}_t} 2^{-L_t(\hat{x}_t|\hat{x}_{t-k+1}^{t-1})} \leq 1, \qquad \text{for all } \hat{x}_{t-k+1}^{t-1}.$$

Note that, since each $\hat{x}_t = f_t(x_{t-k+1}^t)$, the overall memory of the entropy-coder cascaded with the reproduction function, is $(k'-1) = 2(k-1)$.

Without loss of generality, from now on we restrict attention to *admissible* entropy-coders, that is, those whose performance cannot be strictly dominated by another entropy-coder. Formally, an entropy-coder with length-function $L_t(\cdot|\hat{x}_{t-k+1}^{t-1})$ is *inadmissible* if there exists another entropy-coder with length-function $L'_t(\cdot|\hat{x}_{t-k+1}^{t-1})$ such that $L'_t(\hat{x}_t|\hat{x}_{t-k+1}^{t-1}) \leq L_t(\hat{x}_t|\hat{x}_{t-k+1}^{t-1})$ for all $\hat{x}_t$, and $L'_t(\hat{x}_t|\hat{x}_{t-k+1}^{t-1}) < L_t(\hat{x}_t|\hat{x}_{t-k+1}^{t-1})$ for at least one $\hat{x}_t$. Note that all admissible coders have $\max_{\hat{x}_t} L'_t(\hat{x}_t|\hat{x}_{t-k+1}^{t-1}) \leq (B-1)$, and therefore there are only finitely many, say r_2, of them.

[5]Although there is no reason to assume *a priori* that the memory length of the entropy-coder is the same as the memory of f_t, this assumption is made here for the sake of simplicity. It is only a straightforward exercise to extend all our subsequent results to the case of codes with entropy-coders and reproduction functions of different memory lengths.

3.1 Variable-Rate/Fixed-Distortion ZDFM Codes

Following the exact same steps as in the derivation of the best exponent for fixed-rate codes, it is easy to derive the fastest exponential rate of decay of

$$\Pr\left\{\sum_{t=1}^{n} L_t(\hat{X}_t|\hat{X}_{t-k+1}^{t-1}) \geq nR\right\} \tag{16}$$

subject to the constraint that the overall distortion achieved is no greater than nD, i.e.,

$$\sum_{t=1}^{n} \max_{x_{t-k+1}^{t}} \rho(x_t, f_t(x_{t-k+1}^{t})) \leq nD.$$

The best exponent in this case can be characterized as follows. Consider again the set $\mathcal{G}_1$ of all memoryless reproduction functions $\{g_1, g_2, \ldots, g_{r_1}\}$. For a given reproduction function $g_s \in \mathcal{G}_1$ and any $\xi \geq 0$, let $L_s^{\xi} : \hat{\mathcal{X}} \to \boldsymbol{Z}_+$ minimize $E\exp\{\xi L(g_s(X))\}$ over all (admissible) memoryless length-functions $L(\cdot)$ that satisfy Kraft's inequality $\sum_{\hat{x}} 2^{-L(\hat{x})} \leq 1$, where the summation is over the range of g_s. For simplicity, we will sometimes use the shorthand notation $L_s^{\xi}(X)$ for $L_s^{\xi}(g_s(X))$. For any vector θ in the set

$$\Theta'(D, r_1) \stackrel{\triangle}{=} \left\{\theta \in [0,1]^{r_1} : \sum_{s=1}^{r_1} \theta_s = 1,\ \sum_{s=1}^{r_1} \theta_s D_s \leq D,\ \theta_s \geq 0,\quad s = 1, \ldots, r_1\right\},$$

where $D_s = \max_x \rho(x, g_s(x))$ as before, we define:

$$G(R, \theta) = \sup_{\xi \geq 0}\left[\xi R - \sum_{s=1}^{r_1} \theta_s \ln E\exp\{\xi L_s^{\xi}(X)\}\right].$$

Then, the best achievable code-length exponent is given by $G^*(D, R)$, the supremum of $G(R, \theta)$ over all $\theta \in \Theta'(D, r_1)$. [See Remark 11 below for the range of validity of this result.]

Alternatively, $G^*(D, R)$ can be defined in a manner that more closely parallels the definition of the optimal exponent $F^*(D, R)$ in the fixed-rate case: Let

$$\Theta''(D, r_1, r_2) \stackrel{\triangle}{=} \left\{\theta \in [0,1]^{r_1 \cdot r_2} : \sum_{s=(1,1)}^{(r_1,r_2)} \theta_s = 1,\ \sum_{s=(1,1)}^{(r_1,r_2)} \theta_s D_s \leq D\right\},$$

where the vectors $\theta = (\theta_s) \in \Theta''(D, r_1, r_2)$ are indexed by pairs of indices $s = (s_1, s_2) \in \{1, \ldots, r_1\} \times \{1, \ldots, r_2\}$, which run over all $r_1 r_2$ possible combinations (g_{s_1}, L_{s_2}) of reproduction functions in $\mathcal{G}_1$ and admissible entropy-coders. To avoid cumbersome notation, we will denote g_{s_1} and L_{s_2} simply by g_s and L_s, respectively, with the understanding that the former depends only on the first component, s_1, of s and the latter depends only on the second component s_2. For $\theta \in \Theta''(D, r_1, r_2)$ we let

$$\tilde{G}(R, \theta) = \sup_{\xi \geq 0}\left[\xi R - \sum_{s=(1,1)}^{(r_1,r_2)} \theta_s \ln E\exp\{\xi L_s(g_s(X))\}\right],$$

so that we can define

$$G^*(D,R) \stackrel{\triangle}{=} \sup_{\theta\in\Theta''(D,r_1,r_2)} \tilde{G}(R,\theta).$$

It is straightforward to see that the two definitions of $G^*(D,R)$ are equivalent.

Remarks:

10. Arguing precisely as in the case of $F(D,\theta)$ (see Remark 2), it is seen that $\tilde{G}(R,\theta)$ is finite iff $R \leq \sum_s \theta_s \rho_s$, where

$$\rho_s \stackrel{\triangle}{=} \max_x L_s(g_s(x)),$$

Moreover, for each given θ, it is continuous in R for $0 < R < \sum_s \theta_s\rho_s$, and it is left-continuous at $R = \sum_s \theta_s \rho_s$.

11. As discussed above, $G^*(D,R)$ is the best achievable rate at which the probability in (16) decays to zero, over all fixed-distortion ZDFM codes achieving maximum distortion no greater than nD. As before, this result holds for all interesting values of D and R, namely, for all $D \in (0, D_{\max})$ and all $R \in (0, R^*(D))$, where

$$R^*(D) \stackrel{\triangle}{=} \inf_{\theta\in\Theta''(D,r_1,r_2)} \sum_s \theta_s\rho_s.$$

3.2 Variable-Rate/Variable-Distortion ZDFM Codes

We now turn to the more general case of variable-rate zero-delay codes with variable distortion. Our goal here, as explained in detail in the Introduction, is to determine the fastest asymptotic rate of decay of the "error probability,"

$$\Pr\left\{\sum_{t=1}^n L_t(\hat{X}_t|\hat{X}_{t-k}^{t-1}) + \lambda\sum_{t=1}^n \rho(X_t,\hat{X}_t) \geq nR_0\right\}, \tag{17}$$

for given constants λ and R_0.

More specifically, for any $\lambda > 0$ and $R_0 > 0$, the *source coding exponent function for variable-rate/variable-distortion, time-varying ZDFM codes with memory size* $(k-1)$ is defined by

$$\mathcal{H}^k(\lambda,R_0) = \limsup_{n\to\infty} -\frac{1}{n}\ln \min_{\{(f_t,L_t)\}} \Pr\left\{\sum_{t=1}^n [L_t(\hat{X}_t|\hat{X}_{t-k}^{t-1}) + \lambda\rho(X_t,\hat{X}_t)] \geq nR_0\right\} \tag{18}$$

where the minimum is over *all* encoders $\{(f_t,L_t)\}_{t=-k+1}^n$ with memory parameter k (without any further restrictions on their instantaneous rate, distortion, or alphabets).

Next we define

$$H(\lambda, R_0) = \sup_\theta H(\lambda, R_0, \theta),$$

where the supremum w.r.t. θ is over all $(r_1 r_2)$-dimensional vectors whose components are nonnegative and sum to unity (without any additional constraints), and

$$H(\lambda, R_0, \theta) = \sup_{\xi \geq 0} \left[\xi R_0 - \sum_s \theta_s \ln E \exp\{\xi[L_s(g_s(X)) + \lambda\rho(X, g_s(X))]\} \right].$$

Interchanging the two suprema in the definition of $H(\lambda, R_0)$, and observing that the supremum over θ is attained by the vector θ that puts all its mass on the pair $s^* = (s_1^*, s_2^*)$ that minimizes $E \exp\{\xi[L_s(g_s(X)) + \lambda\rho(X, g_s(X))]\}$, we also have

$$\begin{aligned} H(\lambda, R_0) &= \sup_{\xi \geq 0} \left[\xi R_0 - \min_s \ln E \exp\{\xi[L_s(g_s(X)) + \lambda\rho(X, g_s(X))]\} \right] \\ &= \sup_{\xi \geq 0} \left[\xi R_0 - \ln E \exp\{\xi[L_{s^*}(g_{s^*}(X)) + \lambda\rho(X, g_{s^*}(X))]\} \right]. \end{aligned}$$

Remarks:

12. Arguing as in the case of $F(D, \theta)$ (see Remark 2), we see that for any given $\lambda > 0$, $H(\lambda, R_0, \theta)$ is finite iff $R_0 \leq \sum_s \theta_s W_s$, where

$$W_s \stackrel{\triangle}{=} \max_x \left[L_s(g_s(x)) + \lambda\rho(x, g_s(x)) \right].$$

This implies that $H(\lambda, R_0)$ is infinite for R_0 greater than

$$\min_\theta \sum_s \theta_s W_s = \min_s W_s. \tag{19}$$

But for R_0 in that range it is easy to see that the memoryless encoder (g_s, L_s) with s achieving the minimum in (19) achieves zero probability of "error" (as in (17)), implying that for R_0 greater than $\min_s W_s$ we also have $\mathcal{H}^k(\lambda, R_0) = \infty$.

13. We note that $H(\lambda, R_0, \theta)$ is a convex function of R_0, and hence so is $H(\lambda, R_0)$. Therefore, the function $H(\lambda, R_0)$ is continuous in R_0 for all $R_0 \in (0, \min_s W_s)$.

In view of the above discussion, in our main result of this section, Theorem 2, we restrict attention to the interesting range of values of $R \in (0, \min_s W_s)$.

Theorem 2 *For all $\lambda > 0$, all $R_0 \in (0, \min_s W_s)$, and every positive integer $k < \infty$,*

$$\mathcal{H}^k(\lambda, R_0) = H(\lambda, R_0).$$

Remark 14.

Note that here, unlike the fixed-rate (or fixed-distortion) case, there is no need for time-sharing: Optimal performance can be achieved by using a *single* memoryless encoder (g, L). This is because, as noted above, the supremum over θ in the definition of $H(\lambda, R_0)$ is always achieved by a vector θ with only one nonzero component.

Proof. The direct part asserting that $\mathcal{H}^k(\lambda, R_0) \geq H(\lambda, R_0)$, is easily established by analyzing the performance of the memoryless encoder (g_{s^*}, L_{s^*}) that achieves $H(\lambda, R_0)$, using Lemma 2 (or simply applying the Chernoff bound), as in the proof of the direct part of Theorem 1.

For the converse part we also apply a method similar to the one in Theorem 1. As noted earlier, the concatenation of reproduction functions and entropy-coders all having memory k yields encoders with overall memory $k' - 1 = 2(k-1)$. In view of this, we repeat the same construction as in the proof of Theorem 1, but with $k' = (2k-1)$ replacing k.

Let an arbitrarily small $\epsilon > 0$ be given, and choose and fix an integer $q \gg k'$ such that

$$R_0 < \left(\frac{q - k' + 1}{q+1}\right) \min_s W_s. \tag{20}$$

Write $p = (q - k' + 1)$, and let $N \gg q$ be sufficiently large so that for the optimum encoder $\{(f_t^*, L_t^*)\}_{t=-k+1}^N$ achieving the minimum in (18) with $n = N$, has:

$$\Pr\left\{\sum_{t=1}^N [L_t^*(\hat{X}_t|\hat{X}_{t-k+1}^{t-1}) + \lambda\rho(X_t, \hat{X}_t)] \geq NR_0\right\} \leq \exp\{-N[\mathcal{H}^k(\lambda, R_0) - \epsilon]\}. \tag{21}$$

Let $n = \lfloor N/q \rfloor$, recall the definition of the AVS $\{Z_t\}$ from (9), and for $t = 1, 2, \ldots, n$ similarly define Z_t' as the sum

$$\sum_{\tau=(t-1)q+k'}^{tq} \left[L_\tau^*(f_\tau^*(X_{\tau-k+1}^\tau)|f_{\tau-1}^*(X_{\tau-k}^{\tau-1}), \ldots, f_{\tau-k+1}^*(X_{\tau-2k+2}^{\tau-k+1})) + \lambda\rho(X_\tau, f_\tau^*(X_{\tau-k+1}^\tau)\right].$$

Clearly $\{Z_t'\}$ is also an AVS, where the number of different states $r(k,p)$ is now upper bounded by $B^{A^k q} \cdot [B^{k-1}(B-1)!]^p$, corresponding to all possible combinations of $B^{A^k q}$ q-vectors of reproduction functions $(f_{(t-1)q+1}, \ldots, f_{tq})$ together with all $(B-1)!$ possible binary trees of prefix codes with at most B leaves for every context $\hat{x}_{\tau-k+1}^{t-1} \in \hat{\mathcal{X}}^{k-1}$ and every $\tau = (q-1)t + 2k - 1, \ldots, tq$.

Continuing as in the proof of Theorem 1, we let $\boldsymbol{\theta_s}$ denote the relative frequency of $\boldsymbol{s} \in \{1, \ldots, r(k,p)\}$, namely, the proportion of times t during which the vector of reproduction functions $\{f_\tau^*\}_{\tau=(t-1)q+k'}^{tq}$ and the vector of entropy-coders $\{L_\tau^*\}_{\tau=(t-1)q+k'}^{tq}$ all coincide with the particular set of p pairs of reproduction functions and entropy-coders indexed by $\boldsymbol{s}$. For convenience, we visualize the state $\boldsymbol{s}$ as a p-vector $(\sigma_1, \ldots, \sigma_p)$, where each $\sigma_\tau \in \{1, \ldots, r(k,1)\}$ designates the index of the τ-th pair of a reproduction function and an entropy-coder.

In analogy to the definition of $F_p(D, \boldsymbol{\theta})$ in the proof of Theorem 1, we now define $H_p(\lambda, R_0, \boldsymbol{\theta})$ as

$$\sup_{\xi \geq 0} \left[\xi R_0 - \sum_{\boldsymbol{s}=1}^{r(k,p)} \boldsymbol{\theta}_s \ln E \exp \left\{ \xi \sum_{\tau=1}^{p} [L_{\sigma_\tau}(g_{\sigma_\tau}(X_{\tau-k+1}^{\tau}) | X_{\tau-2k+2}^{\tau-1}) + \lambda \rho(X_\tau, g_{\sigma_\tau}(X_{\tau-k+1}^{\tau}))] \right\} \right]$$

where we use the shorthand notation

$$L_{\sigma_\tau}(g_{\sigma_\tau}(X_{\tau-k+1}^{\tau}) | X_{\tau-2k+2}^{\tau-1}) \stackrel{\triangle}{=} L_{\sigma_\tau}(g_{\sigma_\tau}(X_{\tau-k+1}^{\tau}) | g_{\sigma_{\tau-1}}(X_{\tau-k+1}^{\tau-1}), \ldots, g_{\sigma_{\tau-k+1}}(X_{\tau-2k+2}^{\tau-k+1})).$$

In order to further lower bound the l.h.s. of eq. (21), our first step will be to show that

$$H_p(\lambda, pR_0, \boldsymbol{\theta}) \leq p \cdot H(\lambda, R_0). \tag{22}$$

To this end, consider the expression

$$C \stackrel{\triangle}{=} E \exp \left\{ \xi \sum_{\tau=1}^{p} [L_{\sigma_\tau}(g_{\sigma_\tau}(X_{\tau-k+1}^{\tau}) | X_{\tau-2k+2}^{\tau-1}) + \lambda \rho(X_\tau, g_{\sigma_\tau}(X_{\tau-k+1}^{\tau}))] \right\},$$

which is obviously part of the l.h.s. of eq. (22). We next show that C is minimized by a pair consisting of a memoryless reproduction function and a memoryless entropy-coder. Expanding C as

$$\begin{aligned}
& E \exp \left\{ \xi \sum_{\tau=1}^{p} \left[L_{\sigma_\tau}(g_{\sigma_\tau}(X_{\tau-k+1}^{\tau}) | X_{\tau-2k+2}^{\tau-1}) + \lambda \rho(X_\tau, g_{\sigma_\tau}(X_{\tau-k+1}^{\tau})) \right] \right\} \\
&= \sum_{x_{-2k+3}^{0}} P(x_{-2k+3}^{0}) \times \\
&\quad \sum_{x_1} P(x_1) \exp \left\{ \xi [L_{\sigma_1}(g_{\sigma_\tau}(x_{-k+2}^{1}) | x_{-2k+3}^{0})) + \lambda \rho(x_1, g_{\sigma_1}(x_{-k+2}^{1}))] \right\} \times \\
&\quad \sum_{x_2} P(x_2) \exp \left\{ \xi [L_{\sigma_2}(g_{\sigma_2}(x_{-k+3}^{2}) | x_{-2k+4}^{1})) + \lambda \rho(x_2, g_{\sigma_2}(x_{-k+3}^{2}))] \right\} \times \cdots \times \\
&\quad \sum_{x_p} P(x_p) \exp \left\{ \xi [L_{\sigma_p}(g_{\sigma_p}(x_{p-k+1}^{p}) | x_{p-2k+2}^{p-1})) + \lambda \rho(x_p, g_{\sigma_p}(x_{p-k+1}^{p}))] \right\},
\end{aligned} \tag{23}$$

and arguing as in the proof of Theorem 1, the last summation cannot be smaller than the minimum of

$$\sum_x P(x) \exp \{ \xi [L(g(x)) + \lambda \rho(x, g(x))] \} = E \exp\{ \xi [L(g(X) + \lambda \rho(X, g(X))] \} \tag{24}$$

over all pairs (g,L) corresponding to memoryless encoders. Repeating this argument for the summation over $\{x_{p-1}\}$, and continuing inductively, it follows that C is bounded below by the expression in (24) raised to the power of p. Thus,

$$\begin{aligned}
&H_p(\lambda, pR_0, \boldsymbol{\theta}) \\
&\leq \sup_{\xi \geq 0} \Big(\xi p R_0 - p \min_s \ln E \exp\{\xi[L_s(g_s(X)) + \lambda\rho(X, g_s(X))\}\Big) \\
&= p \cdot H(\lambda, R_0). \qquad (25)
\end{aligned}$$

Next we will apply Lemma 1 (Appendix B) to the AVS $\{Z'_t\}$, with $\Delta_{\max} = p[(B-1) + \lambda D_{\max}]$, $r = r(k,p)$, and $\phi = H_p$ as follows: Take $\delta > 0$ sufficiently small, and let $\ell \geq \lceil 2pr(k,p)[(B-1) + \lambda D_{\max}]/\delta \rceil$, and $n \geq \lceil (\ell^3 \Delta_{\max}^2)/(\delta^2 \ln 2) \rceil$. Assuming for now that R_0 lies in the range within which Lemma 1 applies (we come back to justify this shortly) and proceeding exactly as in the proof of Theorem 1, we obtain

$$\begin{aligned}
&\exp\{-N[\mathcal{H}^k(\lambda, R_0) - \epsilon]\} \\
&\geq \exp\left\{-\frac{N}{q}[H_p\left(\lambda, (p+k')R_0 + \zeta_\ell(\delta p, \Delta_{\max}), \boldsymbol{\theta}\right) + \nu_n(\ell, \delta p, \Delta_{\max}, r(k,p))]\right\}.
\end{aligned}$$

And using (25),

$$\begin{aligned}
&\mathcal{H}^k(\lambda, R_0) \\
&\leq \frac{p}{p+k'-1}\left[H\left(\lambda, (1+k'/p)R_0 + \frac{1}{p}\zeta_\ell(\delta p, \Delta_{\max})\right) + \nu_n(\ell, \delta p, \Delta_{\max}, r(k,p))\right] + \epsilon.
\end{aligned}$$

Taking the limit $N \to \infty$ (and hence $n \to \infty$), then $\ell \to \infty$, and finally $q \to \infty$ (hence also $p \to \infty$), by the continuity of $H(\lambda, R_0)$ in R_0 (see Remark 13) we get that $\mathcal{H}^k(\lambda, R_0) \leq H(\lambda, R_0) + \epsilon$. Since $\epsilon > 0$ was arbitrary, this implies that $\mathcal{H}^k(\lambda, R_0) \leq H(\lambda, R_0)$ and completes the proof, subject to justifying the application of Lemma 1 in step (a) above.

To do this, it suffices to show that R_0 lies in the range allowed by Lemma 1, namely that

$$(q+1)R_0 < \sum_{\boldsymbol{s}=1}^{r(k,p)} \boldsymbol{\theta}_{\boldsymbol{s}} \max_{x_1^q} \sum_{j=k'}^{q} \left[L_{\sigma_j}(g_{\sigma_j}(x_{j-k+1}^j) | x_{j-2k+2}^{j-1}) + \lambda\rho(x_j, g_{\sigma_j}(x_{j-k+1}^j))\right]. \qquad (26)$$

But, arguing as in the proof of Theorem 1, we see the r.h.s. above is no smaller than $\min_s[(q-k'+1)W_s]$, and by the choice of q in (20) we see that (26) is trivially satisfied, thereby completing the proof. □

Finally we note that the above characterization of the best achievable exponent in the fixed-slope case can easily be extended to characterize the fastest possible exponential decay rate of the probability of the "error event":

$$\left\{\sum_{t=1}^{n}[L_t(\hat{X}_t|\hat{X}_{t-k}^{t-1})+\lambda\rho(X_t,\hat{X}_t)]\geq nR_0,\ \sum_{t=1}^{n}[L_t(\hat{X}_t|\hat{X}_{t-k}^{t-1})+\lambda'\rho(X_t,\hat{X}_t)]\geq nR_0'\right\}. \quad (27)$$

As before, this can be treated by considering the moment-generating function of linear combinations of distortion and code-length as in (27). Recalling the discussion in the Introduction where the formulation of Theorem 2 was motivated, we note that, here, a reasonable choice of the parameters $(\lambda, R_0, \lambda', R_0')$ would correspond to two adjacent linear segments of the rate-distortion function $R(D)$.

4 Conclusions and Future Research

In this paper we have analyzed the best achievable exponents of ZDFM source codes for lossy compression under three different regimes: fixed-rate, fixed-distortion, and fixed-slope. Our main finding, in all three of them, was that the best large deviations performance is achieved by memoryless codes (in the case of fixed-slope) or by time-sharing between at most two such codes (in the cases of fixed-rate and fixed-distortion). At the heart of the analysis lies a simple "onion-peeling" argument (cf. eqs. (12), (23)), which tells us that the moment-generating function of the code-length (or of the distortion, or of any linear combination between the two), is always minimized by memoryless codes. Since the code-length and the distortion of ZDFM codes satisfy an LDP, the optimal exponents (corresponding to the large deviations rate-functions of the error probabilities), are similarly maximized by the same memoryless codes.

It is natural to expect that this "onion-peeling" technique may be useful in other problem areas in communications and information theory, particularly in the context of zero-delay systems. T. Weissman has suggested to us that this might be the case in joint source-channel coding of memoryless sources through memoryless channels, where both the encoder and the decoder are (possibly, stochastic and time-varying) ZDFM systems. Indeed, Let $\{U_t\}$ be a DMS, encoded by a stochastic ZDFM code characterized by the product distribution $\prod_t P_t^e(x_t|u_{t-k+1}^t)$, and let the encoder output $\{X_t\}$ be transmitted via a discrete memoryless channel (DMC) with distribution $\prod_t P(y_t|x_t)$, whose output $\{Y_t\}$ is decoded by a stochastic ZDFM decoder $\prod_t P_t^d(v_t|y_{t-k+1}^t)$. The probability of excess distortion, $\Pr\{\sum_{t=1}^n \rho(U_t,V_t)\geq nD\}$, may be estimated using its moment-generating function, whose t-th 'layer' in our onion-peeling argument is given by

$$\sum_{u_t,x_t,y_t,v_t} P(u_t)P_t^e(x_t|u_{t-k+1}^t)P(y_t|x_t)P_t^d(v_t|y_{t-k+1}^t)\exp\{\xi\rho(u_t,v_t)\}.$$

As in eqs. (12) and (23), u_{t-k+1}^{t-1} and y_{t-k+1}^{t-1} can be thought of as 'indices', and so the above expression cannot be smaller than the minimum of

$$\sum_{u,x,y,v} P(u)P^e(x|u)P(y|x)P^d(v|y)\exp\{\xi\rho(u,v)\},$$

over all memoryless systems $\{P^e(x|u)\}$ and $\{P^d(v|y)\}$. Moreover, since this is a linear functional of P^e and P^d, this minimum is achieved by some deterministic encoder $P^e(x|u) = \delta(x - f(u))$ and deterministic decoder $P^d(v|y) = \delta(v - g(y))$. Thus, for DMSs and DMCs, time-invariant deterministic memoryless encoders and decoders are as good as any time-varying stochastic ZDFM encoders and decoders in the sense of the excess distortion exponent.

Finally we list a number of open questions and possible directions for future research which might be interesting to consider:

- *Zero-delay codes with infinite memory.* This the open problem presented in the Introduction: Is it possible to relax the finite-memory assumption and extend Theorems 1 and 2 to infinite-memory zero-delay codes?

- *Excess-distortion vs. excess-rate exponents.* When considering variable-rate/variable-distortion ZDFM codes, we may alternatively choose to examine (as in [22]) the best achievable trade-off between the asymptotic exponents of the probabilities of the events

$$\left\{\sum_{t=1}^{n}\rho(X_t,\hat{X}_t)\geq nD\right\} \quad \text{and} \quad \left\{\sum_{t=1}^{n}L_t(\hat{X}_t|\hat{X}_{t-k}^{t-1})\geq nR\right\}.$$

  For any $\alpha > 0$, the corresponding source-coding exponent function is defined by

$$\mathcal{U}^k(D,R,\alpha) \stackrel{\triangle}{=} \limsup_{n\to\infty}\left[-\frac{1}{n}\min\ln\Pr\left\{\sum_{t=1}^{n}\rho(X_t,f_t(X_{t-k+1}^t))\geq nD\right\}\right],$$

  where the minimum is over all variable-rate/variable-distortion ZDFM codes with memory parameter k, such that

$$\Pr\left\{\sum_{t=1}^{n}L_t(f_t(X_{t-k+1}^t)|f_{t-1}(X_{t-k}^{t-1}),\ldots,f_{t-k+1}(X_{t-2k+2}^{t-k+1}))\geq nR\right\}\leq e^{-n\alpha}.$$

In view of our results in Sections 2 and 3 above, the natural guess for a single-letter expression characterizing $\mathcal{U}^k(D,R,\alpha)$ is the function

$$U(D,R,\alpha) \stackrel{\triangle}{=} \sup_{\{\theta: G(R,\theta)\geq\alpha\}} F(D,\theta).$$

Unfortunately we have not been able to show that indeed $\mathcal{U}^k(D,R,\alpha) = U(D,R,\alpha)$. This is due to the following two new subtleties arising here. First, it appears that $U(D,R,\alpha)$ may not necessarily be jointly continuous in all three of its arguments. But it can be shown to be continuous at *almost* all such triplets (D,R,α), and following the same argument as in the proofs of the direct parts of Theorems 1 and 2 it can be shown that $\mathcal{U}^k(D,R,\alpha) \leq U(D,R,\alpha)$ at all continuity points of U.

The second and more serious subtlety is that the main argument in the proofs of the converses in Theorems 1 and 2 (the "onion-peeling" argument of eqs. (12), (23)) does not generalize in a straightforward manner to this case. Nevertheless, it does generalize (exactly as in the proof of Theorem 1) to the case of ZDFM codes with memoryless entropy-coding, showing that the best achievable exponent achieved by such codes is indeed $U(D,R,\alpha)$ (at all continuity points of U).

It would be interesting to settle the conjecture above, stating that $\mathcal{U}^k(D,R,\alpha) = U(D,R,\alpha)$ for ZDFM codes. This would provide a symmetric characterization of the best distortion error exponent versus the best rate error exponent.

- *Codes with finite anticipation.* How can our results be extended to the richer class of finite-memory codes with finite anticipation (delay)?

- *More general sources.* It would be interesting to extend our results to more general sources. It is natural to expect that in the case of a Markov source, for example, the "onion-peeling" argument regarding the moment generating function would identify an optimal encoder as one whose memory length is equal to the order of the Markov source. An LDP for Markov sources will then have to be invoked in order to make a statement regarding the large deviations performance. However, if time–varying ZDFM codes are allowed in the competing class of codes, this is not straightforward.

- *Universal zero-delay coding.* Perhaps the most intriguing direction for future research is to investigate the existence of *universal* zero-delay schemes for memoryless sources.

Is there a zero-delay code that achieves the optimal source coding exponent for any memoryless source? While in the non-causal case [17] such codes exist, in the zero-delay case the answer is not obvious. If it turned out that there is some unavoidable price for universality, then the question would be how to minimize it in some uniform sense across the class of all memoryless sources.

Acknowledgments

Interesting discussions with Jacob Ziv and with Tsachy Weissman are gratefully acknowledged.

Appendix A

An Information-Theoretic Expression for $F(D, \theta)$

We would like to show that

$$F(D, \theta) = \sup_{\xi \geq 0} [\xi D - \Lambda(\xi)], \tag{A.1}$$

where $\Lambda(\xi) = \sum_s \theta_s \ln E_{P_s} e^{\xi Y_s}$, is equal to

$$\tilde{F}(D, \theta) \stackrel{\triangle}{=} \inf_{\{Q_s\} : \sum_s \theta_s E_{Q_s} Y_s \geq D} \sum_s \theta_s D(Q_s \| P_s). \tag{A.2}$$

First define

$$\underline{\Delta} = \underline{\Delta}(\theta) = \sum_s \theta_s E_{P_s} Y_s$$

and

$$\overline{\Delta} = \overline{\Delta}(\theta) = \sum_s \theta_s \Delta_s,$$

where $\Delta_s = \max\{z : \; P_s(z) > 0\}$.

The proof of the equality between $F(D, \theta)$ and $\tilde{F}(D, \theta)$ follows very closely the corresponding proof of Proposition 1 (ii) in [13, Appendix II]. Here we outline the necessary modifications to that proof. First, it easy to check that $F(D, \theta) = \tilde{F}(D, \theta) = 0$ for $D \leq \underline{\Delta}$. Also it is straightforward to show that $F(D, \theta) = \tilde{F}(D, \theta) = \infty$ for $D > \overline{\Delta}$, which implies that, as claimed in Section 2, indeed $F^*(D, R) = \infty$ for $D > D^*(R)$. Next we observe that $\Lambda(\xi)$ is differentiable in ξ, with

$$\Lambda'(\xi) = \sum_s \theta_s E_{P_s} \left[\frac{Y_s e^{\xi Y_s}}{E_{P_s} e^{\xi Y_s}} \right],$$

and $\Lambda''(\xi) \geq 0$, for all $\xi > 0$.

In the range $\underline{\Delta} < D < \overline{\Delta}$ it is easy to show that the supremum in (A.1) is achieved by the unique $\xi^* \geq 0$ satisfying $\Lambda'(\xi^*) = D$. Fix θ, $D \in (\underline{\Delta}, \overline{\Delta})$, and a corresponding ξ^*, and define a new family of distributions

$$\mu_s(y) \stackrel{\triangle}{=} \frac{e^{\xi^* y}}{E_{P_s}[e^{\xi^* Y_s}]} P_s(y).$$

Then,

$$\sum_s \theta_s E_{\mu_s} Y_s = \Lambda'(\xi^*) = D$$

and therefore

$$\begin{aligned}
\tilde{F}(D,\theta) &\leq \sum_s \theta_s D(\mu_s \| P_s) \\
&= \sum_s \theta_s \sum_y \mu_s(y) \ln \frac{e^{\xi^* y}}{E_{P_s}[e^{\xi^* Y_s}]} \\
&= \xi^* \Lambda'(\xi^*) - \Lambda(\xi^*) \\
&= F(D,\theta). \qquad\qquad (A.3)
\end{aligned}$$

Conversely, take any candidate $\{Q_s\}$ as in (A.2). Then for each s (by Lemma 6.2.13 in [3], after taking $\phi(y) = \xi^* y$ in the definition of Λ^*), we have that

$$D(Q_s \| P_s) \geq E_{Q_s}[\xi^* Y_s] - \ln E_{P_s}[e^{\xi^* Y_s}],$$

so multiplying both sides by θ_s and summing over s, we get

$$\sum_s \theta_s D(Q_s \| P_s) \geq \xi^* \sum_s \theta_s E_{Q_s} Y_s - \Lambda(\xi^*) \geq \xi^* D - \Lambda(\xi^*) = F(D,\theta).$$

Taking the infimum over all $\{Q_s\}$ as in (A.2), we get $\tilde{F}(D,\theta) \geq F(D,\theta)$. This together with the upper bound in (A.3) shows that $\tilde{F}(D,\theta) = F(D,\theta)$ for $\underline{\Delta} < D < \overline{\Delta}$.

Finally, if $D = \overline{\Delta} > \underline{\Delta}$, then both $F(D,\theta)$ and $\tilde{F}(D,\theta)$ can be evaluated explicitly, and they are both equal to

$$\lim_{\xi \to \infty} [\xi \overline{\Delta} - \Lambda(\xi)] = \sum_s \theta_s \ln[1/P_s(\Delta_s)]. \qquad (A.4)$$

And, moreover, $F(D,\theta)$ is left-continuous at the point $D = \overline{\Delta} > \underline{\Delta}$. To see this recall that $F(D,\theta)$ is nondecreasing in D, and also

$$\liminf_{D \uparrow \overline{\Delta}} F(D,\theta) = \liminf_{\epsilon \downarrow 0} \sup_\xi [\xi(\overline{\Delta} - \epsilon) - \Lambda(\xi)]$$

$$\begin{aligned}&\overset{(a)}{\geq} \quad \liminf_{\epsilon \downarrow 0}[\epsilon^{-1/2}(\overline{\Delta}-\epsilon)-\Lambda(\epsilon^{-1/2})]\\ &= \quad \lim_{\xi\to\infty}[\xi\overline{\Delta}-\Lambda(\xi)]\\ &\overset{(b)}{=} \quad F(\overline{\Delta},\theta),\end{aligned}$$

where (a) follows by taking $\xi = \epsilon^{-1/2}$ in the supremum, and (b) follows from (A.4). □

Appendix B

Large Deviations Analysis for Arbitrarily Varying Sources

In this appendix, we present and prove two auxiliary lemmas that give upper and lower bounds on the probability of a certain large deviations event associated with an arbitrarily-varying source (AVS). These lemmas, which are used in the proofs of the main results, are quite standard except for the fact they hold for every sample size and not merely asymptotically. The importance of this feature lies in the fact that the main term in the exponent of the large deviations probability under consideration depends on the relative frequencies of the various states of the AVS, which may not stabilize, in general, as the sample size grows.

Consider an AVS with r states, emitting symbols $Z_1, Z_2, \ldots, Z_n$ from a finite subset $\mathcal{Z}$ of $\mathbb{R}_+$, according to the probability law

$$P(z_1,\ldots,z_n) = \prod_{t=1}^{n} P(z_t|s_t),$$

where $s_1,\ldots,s_n$, $s_t \in \mathcal{S} = \{1,\ldots,r\}$ (r a positive integer) is an arbitrary (deterministic) sequence of states. Let $\theta_s \in [0,1]$ denote the relative frequency of $s_t = s$ along $s_1,\ldots,s_n$, i.e., $\theta_s = |\{t:\ s_t = s\}|/n$, $s \in \mathcal{S}$. For a given $\xi \geq 0$, define

$$M_s(\xi) = \sum_{z\in\mathcal{Z}} P(z|s)e^{\xi z}, \quad s \in \mathcal{S}.$$

Let $\Delta_s = \max\{z:\ P(z|s) > 0\}$, $\overline{\Delta} = \sum_{s\in\mathcal{S}} \theta_s\Delta_s$, and $\Delta_{\max} = \max_s \Delta_s$. For a given $D \in [0, \Delta_s)$, let

$$\phi_s(D) = \sup_{\xi\geq 0}[\xi D - \ln M_s(\xi)] = \max_{\xi\geq 0}[\xi D - \ln M_s(\xi)], \quad s \in \mathcal{S}.$$

For a given $\delta > 0$, let $\xi_s(\delta) < \infty$ be the (usually unique) value of ξ that achieves $\phi_s(\Delta_s - \delta)$, and let $\xi(\delta) = \max_{s\in\mathcal{S}} \xi_s(\delta)$. Finally, for a given $\delta > 0$, and $D \in [0, \overline{\Delta} - 2\delta]$, define

$$\phi(D,\theta) = \sup_{\xi\geq 0}[\xi D - \sum_{s\in\mathcal{S}} \theta_s \ln M_s(\xi)],$$

where $\theta \stackrel{\triangle}{=} (\theta_1, \ldots, \theta_r)$.

The following two lemmas provide lower and upper bounds on $\Pr\{\sum_{t=1}^n Z_t \geq nD\}$.

Lemma 1 *Let $\delta \in (0, \Delta_{\max}/(2\ln 2))$, $D \in [0, \overline{\Delta} - 2\delta]$, and let ℓ be a positive integer at least as large as $\lceil 2r\Delta_{\max}/\delta \rceil$. Then, for all*

$$n \geq \left\lceil \frac{\ell^3 \Delta_{\max}^2}{\delta^2 \ln 2} \right\rceil + 2\ell,$$

we have:

$$Pr\left\{\sum_{t=1}^{n} Z_t \geq nD\right\} \geq \exp\left\{-n\left[\phi\left(D + \zeta_\ell(\delta, \Delta_{\max}), \theta\right) + \nu_n(\ell, \delta, \Delta_{\max}, r)\right]\right\}$$

where

$$\zeta_\ell(\delta, \Delta_{\max}) = \frac{2\delta^2 \ln 2}{\ell^2 \Delta_{\max}}$$

and

$$\nu_n(\ell, \delta, \Delta_{\max}, r) = \xi(\delta)\Delta_{\max}\left(\frac{r}{\ell} + 2\sqrt{\frac{\ell \ln 2}{n}}\right) + \frac{\ln 2}{n}.$$

Lemma 2 *Let $D \in (0, \overline{\Delta})$. Then, for every positive integer ℓ:*

$$Pr\left\{\sum_{t=1}^{n} Z_t \geq nD\right\} \leq \exp\left\{-n\phi\left(D - \Delta_{\max}\left[\frac{\ell}{n} + \frac{r}{\ell}\right], \theta\right) + \ell\phi(D, \theta)\right\}.$$

Proof of Lemma 1: We begin with a lower bound for the case $r = 1$, where $\{Z_t\}$ are independent and all drawn from the same PMF P, whose moment-generating function is $M(\xi) = \sum_z P(z)e^{\xi z}$, and where the maximum value of z, with positive probability is Δ. Let $\phi_Z(D) = \sup_{\xi \geq 0}[\xi D - \ln M(\xi)]$, and for a given $\delta > 0$, $D \in [0, \Delta - 2\delta)$, $\epsilon \in (0, \delta]$, let $\xi^* \in [0, \infty)$ achieve $\phi_Z(D + \epsilon)$, namely, be the unique maximizer of $[\xi(D + \epsilon) - \ln M(\xi)]$. Note that since $D + \epsilon < \Delta - \delta$, ξ^* is upper-bounded by $\xi(\delta) < \infty$ (the value of ξ that achieves $\phi_Z(\Delta - \delta)$). Let us now denote the event of interest by

$$E = \left\{z^n : \sum_{t=1}^{n} z_t \geq nD\right\}$$

and let

$$F = \left\{z^n : \sum_{t=1}^{n} z_t \leq n(D + 2\epsilon)\right\}.$$

Now, defining the PMF $P_\xi(z) = P(z)e^{\xi z}/M(\xi)$, we have:

$$\begin{aligned}
\Pr\{E\} &\geq \Pr\{E \cap F\} \\
&= \sum_{z^n \in E\cap F} \prod_{t=1}^n P(z_t) \\
&= [M(\xi^*)]^n \cdot \sum_{z^n \in E\cap F} \prod_{t=1}^n e^{-\xi^* z_t} P_{\xi^*}(z_t) \\
&\geq [M(\xi^*) \cdot \exp\{-\xi^*(D+2\epsilon)\}]^n \cdot \sum_{z^n \in E\cap F} \prod_{t=1}^n P_{\xi^*}(z_t).
\end{aligned}$$

As for the first factor, we have

$$\begin{aligned}
[M(\xi^*) \cdot \exp\{-\xi^*(D+2\epsilon)\}]^n &\geq \exp\{-\xi^* \epsilon n\} \cdot \exp\{-n\phi_Z(D+\epsilon)\} \\
&\geq \exp\{-2\xi(\delta)\epsilon n\} \cdot \exp\{-n\phi_Z(D)\}.
\end{aligned}$$

For the second factor of the r.h.s. of eq. (B.1), we first apply the union bound:

$$\sum_{z^n \in E\cap F} \prod_{t=1}^n P_{\xi^*}(z_t) \geq 1 - \sum_{z^n \in E^c} \prod_{t=1}^n P_{\xi^*}(z_t) - \sum_{z^n \in F^c} \prod_{t=1}^n P_{\xi^*}(z_t) \stackrel{\triangle}{=} 1 - \alpha - \beta.$$

To further lower-bound the last expression, we upper-bound both α and β using Hoeffding's inequality [9], which asserts that given $\epsilon > 0$ and n independent RV's $Y_1, \ldots, Y_n$, ranging over an interval of size Δ,

$$\Pr\left\{\sum_{t=1}^n (Y_t - EY_t) \geq n\epsilon\right\} \leq \exp\left\{-\frac{2n\epsilon^2}{\Delta^2}\right\}.$$

Applying this inequality for α (with $Y_t = -Z_t$, $EY_t = -(D+\epsilon)$, $t = 1, \ldots, n$) and for β (with $Y_t = Z_t$, and $EY_t = D+\epsilon$, $t = 1, \ldots, n$), we see that both α and β are upper-bounded by $\exp\{-2n\epsilon^2/\Delta^2\}$. Therefore,

$$\begin{aligned}
\Pr\{E\} &\geq \sup_{\epsilon \in (0,\delta]} \left[1 - 2\exp\left\{-\frac{2n\epsilon^2}{\Delta^2}\right\}\right] \cdot \exp\{-n\phi_Z(D) - 2\xi(\delta)\epsilon\} \\
&\geq \exp\left\{-n\phi_Z(D) - 2\xi(\delta)\Delta\sqrt{n \ln 2} - \ln 2\right\},
\end{aligned}$$

where the last inequality follows from setting $\epsilon = \Delta\sqrt{\ln 2/n}$, which is in the allowed range $(0, \delta]$ for all

$$n \geq \left\lceil \frac{\Delta^2}{\delta^2 \ln 2} \right\rceil. \tag{B.1}$$

We now return to the case of an r-state AVS with relative frequencies of states $\theta_1, \ldots, \theta_r$. Let us re-index the RV's $\{Z_t\}$ as $\{W_\tau^s,\ \tau = 1, \ldots, \theta_i n,\ s \in \mathcal{S}\}$, where W_τ^s is Z_t, with t being the τ-th occurrence of $s_t = s$. Fix a (large) positive integer ℓ, and let $\ell_s \stackrel{\triangle}{=} \lfloor \theta_s \ell \rfloor$. Now generate $m = \lfloor n/\ell \rfloor$ i.i.d. RV's, $Y_1, \ldots, Y_m$, according to

$$Y_t = \sum_{s \in \mathcal{S}} \sum_{\tau=(t-1)\ell_s+1}^{t\ell_s} W_\tau^s, \quad t = 1, \ldots, m$$

where if $\ell_s = 0$, the inner summation is defined as zero. Now, obviously, since $\{Z_t\}$ are nonnegative random variables:

$$\begin{aligned} \Pr\left\{\sum_{t=1}^n Z_t \geq nD\right\} &\geq \Pr\left\{\sum_{t=1}^m Y_t \geq nD\right\} \\ &\geq \Pr\left\{\sum_{t=1}^m Y_t \geq m\frac{\ell D}{1-\ell/n}\right\}. \end{aligned}$$

Thus, it is enough to lower-bound the r.h.s., which corresponds to the i.i.d. RV's $\{Y_t\}$, all having a PMF whose moment-generating function is

$$M_Y(\xi) = \prod_{s \in \mathcal{S}} [M_s(\xi)]^{\ell_s}.$$

Let us define

$$\phi_Y(D) = \sup_{\xi \geq 0}[\xi D - \ln M_Y(\xi)] = \sup_{\xi \geq 0}\left[\xi D - \sum_{s \in \mathcal{S}} \ell_s \ln M_s(\xi)\right].$$

It is easy to see that $\xi(\delta)$ is an upper bound to the value of ξ that achieves $\phi_Y(\sum_{s \in \mathcal{S}} \ell_s(\Delta_s - \delta))$. This means that if a certain value of D satisfies

$$\ell D \leq \sum_{s \in \mathcal{S}} \ell_s(\Delta_s - \delta),$$

then

$$\begin{aligned} \phi_Y(\ell D) &= \sup_{\xi \geq 0}\left[\xi \ell D - \sum_{s \in \mathcal{S}} \ell_s \ln M_s(\xi)\right] \\ &= \sup_{0 \leq \xi \leq \xi(\delta)}\left[\xi \ell D - \sum_{s \in \mathcal{S}} \ell_s \ln M_s(\xi)\right] \\ &\leq \sup_{0 \leq \xi \leq \xi(\delta)}\left[\xi \ell D - \sum_{s \in \mathcal{S}} (\ell\theta_s - 1) \ln M_s(\xi)\right] \\ &\leq \ell\phi(D, \theta) + \sum_{s \in \mathcal{S}} \ln M_s(\xi(\delta)) \\ &\leq \ell\phi(D, \theta) + r\xi(\delta)\Delta_{\max}. \end{aligned}$$

We now apply the lower bound (B.1) to $\{Y_t\}$, with n, $\phi_Z(D)$, and $\Delta_{\max}$ replaced by m, $\phi_Y(\ell D/(1 - \ell/n))$, and $\ell\Delta_{\max}$, respectively, under the condition

$$\frac{n}{\ell} \geq \frac{\ell^2 \Delta_{\max}^2}{\delta^2 \ln 2} + 2, \tag{B.2}$$

which parallels the earlier condition (B.1) we had for $r = 1$. First observe that the assumptions $\ell \geq \lceil 2r\Delta_{\max}/\delta \rceil$ and (B.2) guarantee that eq. (B.2) is applicable for $D = D/(1-\ell/n)$, whenever $D \leq \overline{\Delta} - 2\delta$, provided that $\delta < \Delta_{\max}/(2\ln 2)$. We therefore have that the probability

$$\Pr\left\{\sum_{t=1}^{n} Z_t \geq nD\right\} \tag{B.3}$$

is bounded below by

$$\begin{aligned}
& \exp\left\{-m\phi_Y\left(\ell\frac{D}{1-\ell/n}\right) - 2\xi(\delta)\ell\Delta_{\max}\sqrt{m\ln 2} - \ln 2\right\} \\
\geq\ & \exp\left\{-\frac{n}{\ell}\cdot\left[\ell\phi\left(\frac{D}{1-\ell/n},\theta\right) + r\xi(\delta)\Delta_{\max}\right] - 2\xi(\delta)\ell\Delta_{\max}\sqrt{\frac{n\ln 2}{\ell}} - \ln 2\right\} \\
\geq\ & \exp\left\{-n\left[\phi\left(\frac{D}{1-\ell/n},\theta\right) + \xi(\delta)\Delta_{\max}\frac{r}{\ell}\left(+2\sqrt{\frac{\ell\ln 2}{n}}\right) + \frac{\ln 2}{n}\right]\right\}.
\end{aligned}$$

Since we assume that n is so large that

$$\frac{\ell}{n} \leq \frac{\delta^2\ln 2}{\ell^2\Delta_{\max}^2 + 2\delta^2\ln 2},$$

we have

$$\frac{D}{1-\ell/n} \leq D + \frac{2\delta^2\ln 2}{\ell^2\Delta_{\max}},$$

which in turn implies that the probability in (B.3) is bounded below by

$$\exp\left\{-n\left[\phi\left(D+\zeta_\ell(\delta,\Delta_{\max}),\theta\right) + \xi(\delta)\Delta_{\max}\left[\frac{r}{\ell} + 2\sqrt{\frac{\ell\ln 2}{n}}\right] + \frac{\ln 2}{n}\right]\right\}.$$

Thus, Lemma 1 is proved. □

Proof of Lemma 2. Let us define $\{Y_t\}_{t=1}^{m}$ similarly as in the proof of Lemma 1. First, observe that the total number of $\{Z_t\}$ which are accounted for within $Y_1, \ldots, Y_m$ is

$$\left\lfloor\frac{n}{\ell}\right\rfloor\sum_{s\in\mathcal{S}}\lfloor\theta_s\ell\rfloor \geq \left(\frac{n}{\ell}-1\right)\sum_{s\in\mathcal{S}}(\theta_s\ell-1) = \left(\frac{n}{\ell}-1\right)(\ell-r) > n-\ell-\frac{rn}{\ell},$$

namely, no more than $\ell + rn/\ell$ terms $\{Z_t\}$ are omitted from the summation. Therefore,

$$\begin{aligned}
\Pr\left\{\sum_{t=1}^{n} Z_t \geq nD\right\} &\leq \Pr\left\{\sum_{t=1}^{m} Y_t \geq n\left[D - \Delta_{\max}\left(\frac{\ell}{n}+\frac{r}{\ell}\right)\right]\right\} \\
&\leq \Pr\left\{\sum_{t=1}^{m} Y_t \geq m\ell\left[D - \Delta_{\max}\left(\frac{\ell}{n}+\frac{r}{\ell}\right)\right]\right\} \\
&\leq \exp\left\{-m\phi_Y\left(\ell\left[D - \Delta_{\max}\left(\frac{\ell}{n}+\frac{r}{\ell}\right)\right]\right)\right\}
\end{aligned}$$

where the last step follows from the Chernoff bound. Next observe that for every D:

$$\begin{aligned}
\phi_Y(\ell D) &= \sup_{\xi \geq 0} \left[s\ell D - \sum_{s \in \mathcal{S}} \ell_s \ln M_s(\xi) \right] \\
&\geq \sup_{\xi \geq 0} \left[s\ell D - \sum_{s \in \mathcal{S}} \theta_s \ell \ln M_s(\xi) \right] \\
&= \ell \phi(D, \theta).
\end{aligned}$$

Thus, the probability of the event of interest is further upper-bounded by

$$\begin{aligned}
\Pr\left\{ \sum_{t=1}^{n} Z_t \geq nD \right\} &\leq \exp\left\{ -m\ell\phi\left(D - \Delta_{\max}\left[\frac{\ell}{n} + \frac{r}{\ell} \right], \theta \right) \right\} \\
&\leq \exp\left\{ -(n-\ell)\phi\left(D - \Delta_{\max}\left[\frac{\ell}{n} + \frac{r}{\ell} \right], \theta \right) \right\} \\
&\leq \exp\left\{ -n\phi\left(D - \Delta_{\max}\left[\frac{\ell}{n} + \frac{r}{\ell} \right], \theta \right) + \ell\phi(D, \theta) \right\}
\end{aligned}$$

which proves the lemma. □

References

[1] T. Berger and J. K.-Y. Lau, "On binary sliding block codes," *IEEE Trans. Inform. Theory*, vol. IT–23, no. 3, pp. 343–353, May 1977.

[2] A. Dembo and I. Kontoyiannis. The asymptotics of waiting times between stationary processes, allowing distortion. *Ann. Appl. Probab.*, 9:413–429, 1999.

[3] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*, Jones & Bartlett Publishers, 1993.

[4] T. Ericson, "A result on delay–less information transmission," *Int. Symposium on Information Theory*, Grignano, Italy, June 1979.

[5] N. T. Gaarder and D. Slepian, "On optimal finite-state digital transmission systems," *Int. Symposium on Information Theory*, Grignano, Italy, June 1979.

[6] N. T. Gaarder and D. Slepian, "On optimal finite-state digital transmission systems," *IEEE Trans. Inform. Theory*, vol. IT–28, no. 2, pp. 167–186, March 1982.

[7] R. K. Gilbert and D. L. Neuhoff, "Bounds to the performance of causal codes for Markov sources," *Proc. Allerton Conf. Comm. Contr. and Comput.*, pp. 284–292, 1979.

[8] R. M. Gray, "Sliding block source coding," *IEEE Trans. Inform. Theory*, vol. IT–21, no. 4, pp. 357–368, July 1975.

[9] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *J. Amer. Statist. Assoc.,* vol. 58, pp. 13–30, 1963.

[10] Sh. Ihara and M. Kubo, "Error exponent for coding of memoryless Gaussian sources with a fidelity criterion," *IEICE Trans. Fundamentals,* vol. E83-A, no. 10, pp. 1891–1897, 2000.

[11] R. Kiesel and U. Stadtmüller, "A large deviation principle for weighted sums of independent identically distributed random variables," *J. Math. Anal. Appl.,* vol. 251, pp. 929–939, 2000.

[12] I. Kontoyiannis. An implementable lossy version of the Lempel-Ziv algorithm – Part I: Optimality for memoryless sources. *IEEE Trans. Inform. Theory*, 45(7):2293–2305, November 1999.

[13] I. Kontoyiannis, "Pointwise redundancy in lossy data compression and universal lossy data compression," *IEEE Trans. Inform. Theory* vol. IT–46, no. 1, pp. 136–152, January 2000.

[14] T. Linder and G. Lugosi, "A zero–delay sequential scheme for lossy coding of individual sequences," *IEEE Trans. Inform. Theory* vol. IT–47, no. 6, pp. 2533–2538, September 2001.

[15] T. Linder and R. Zamir, "Causal source coding of stationary sources with high resolution," *2001 International Symposium on Information Theory.*

[16] S. P. Lloyd, "Rate vs. fidelity for the binary source," *Bell Systems Technical Journal*, vol. 56, no. 3, pp. 427–437, March 1977.

[17] K. Marton, "Error exponent for source coding with a fidelity criterion," *IEEE Trans. Inform. Theory*, vol. IT–20, no. 2, pp. 197–199, March 1974.

[18] D. L. Neuhoff and R. K. Gilbert, "Causal source codes," *IEEE Trans. Inform. Theory*, vol. IT–28, no. 5, pp. 701–713, September 1982.

[19] P. Piret, "Causal sliding block encoders with feedback," *IEEE Trans. Inform. Theory*, vol. IT–25, no. 2, pp. 237–240, March 1979.

[20] I. G. Stiglitz, "A coding theorem for a class of unknown channels," *IEEE Trans. Inform. Theory*, vol. IT–13, no. 2, pp. 217–220, April 1967.

[21] T. Weissman and N. Merhav, "On limited–delay lossy coding and filtering of individual sequences," to appear in *IEEE Trans. Inform. Theory*. (Available at [`http://tiger.technion.ac.il/users/merhav`].)

[22] T. Weissman and N. Merhav, "Tradeoffs between the excess code-length exponent and the excess distortion exponent in lossy source coding," to appear in *IEEE Trans. Inform. Theory.* (Available at [`http://tiger.technion.ac.il/users/merhav`].)

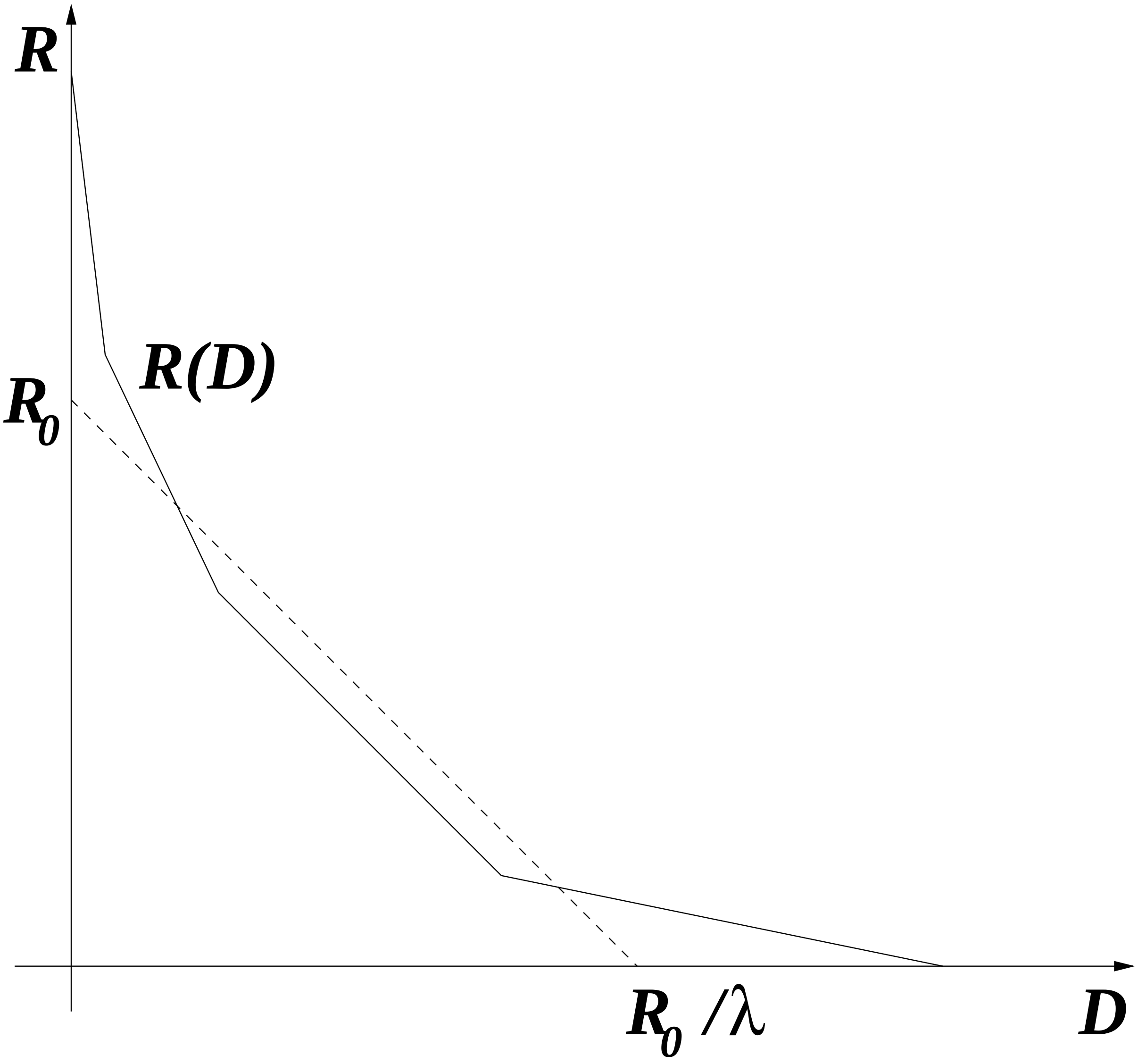

Figure 1: The rate distortion function (solid line) and the line $R = R_0 - \lambda D$ (dashed line).