Large Deviations Performance of Predictors for Markov Sources

Erez Sabbag* Neri Merhav*

November 30, 2003

Abstract

Performance and design of finite-memory predictors for Markov sources are investigated under the large deviations regime of the probability of excess cumulative loss beyond a certain threshold. It is shown that time-varying predictors with memory size corresponding to the Markov order of the source are as good as any finite memory predictor. In addition, we present a procedure for designing a sequence of predictors with error exponent arbitrarily close to the optimal error exponent. The computational complexity of this procedure is linear in the sequence length. Upper and lower bounds on performance are given.

1 Introduction

Consider the problem of sequentially predicting symbols emitted from a Markov source, based on k previously-observed symbols, where, at each time instant, we can choose a different prediction function. This problem has been extensively studied under the expected-loss criterion where it is clear that the optimal performance can be obtained using a time-invariant predictor whose memory size is identical to the Markov order of source.

By contrast, large deviations (LD) performance analysis of predictors received relatively little attention (e.g., [1, Sec.III] where predictors for binary memoryless sources were considered), and it is not even a priori obvious, that optimum LD performance can be achieved by a predictor whose memory length is as the Markov order of the source. In this work, we investigate the LD performance of finite-memory time-varying predictors under the large-deviations criterion, namely, the exponential decay rate of the probability of excessive loss. It should be noted that the above problem for infinite memory predictors is still open as it is not clear if our result continues to hold. The techniques used are in the spirit of those of [2] and [3] where LD performance of zero-delay finite memory lossy source codes were studied.

*The authors are with the Faculty of the Electrical Engineering, Technion - I.I.T., Haifa 32000, Israel. E-mail addresses: erezs@tx.technion.ac.il, merhav@ee.technion.ac.il.

2 Notations and Conventions

We begin with some notations and definitions. Throughout this work, capital letters represent random variables (RVs), specific realizations of them are denoted by the corresponding lowercase letters, and their alphabets are written as the respective calligraphic letters. For a sequence of symbols $\{x_t\}$, 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 RVs with capital letters replacing lowercase letters. For the sake of simplicity we confine ourselves to the case of first order Markov sources. However, the extension to higher orders of the following results is straightforward.

Consider a first-order, homogeneous, irreducible, aperiodic Markov source, $\{X_t\}_{t\in\mathbb{Z}}$, over a finite alphabet $\mathcal{X}$ of size $A \geq 2$ with a stationary distribution $\{\pi(a), \forall a \in \mathcal{X}\}$ and a transition matrix Π whose elements are $\{\Pi(i|j) = \Pr\{X_t = i|X_{t-1} = j\}, \forall i, j \in \mathcal{X}\}$.

A *time-varying, finite-memory predictor* with memory size k is a sequence of *prediction functions* $\{F_t\}_{t\geq 1}$, where $F_t : \mathcal{X}^k \to \hat{\mathcal{X}}$ predicts the next source symbol, X_t, based on the past k source symbols, X_{t-k}^{t-1}, and $\hat{\mathcal{X}}$ denoted the prediction alphabet of size B. Let $\mathcal{F}_k$ denote the set of all prediction functions with memory size k, where $|\mathcal{F}_k| = r(k) \stackrel{\triangle}{=} B^{A^k}$.

The sequence $\{\hat{x}_t\}_{t\geq 1}$, where $\hat{x}_t = F_t(x_{t-k}^{t-1})$, is referred to as the *prediction* of the source sequence $\{x_t\}_{t\geq 1}$. The loss between x_1^n and its prediction $\hat{x}_1^n$ 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 loss function. We assume that $d_{\max} \stackrel{\triangle}{=} \max_{x,\hat{x}} \rho(x, \hat{x})$ is finite.

3 Prediction Exponent of Finite-Memory Predictors

For a given loss level d, the *prediction error exponent function for time-varying predictors with memory size k* is defined by

$$\mathcal{J}^k(d) \stackrel{\triangle}{=} \limsup_{n\to\infty} \left[-\frac{1}{n} \ln \min_{\substack{\{F_t\}_{t=1}^n \\ F_t \in \mathcal{F}_k}} \Pr\left\{ \sum_{t=1}^n \rho\big(X_t, F_t(X_{t-k}^{t-1})\big) \geq nd \right\} \right].$$

As in the case of finite memory predictors under expected loss criterion, it turns out that predictors with memory size equal to the Markov order of the source are at least as good as any other predictors with finite memory, namely:

Theorem 1. *For every $d \in (0, d_{\max})$, and any positive integer k,*

$$\mathcal{J}^k(d) = \mathcal{J}^1(d).$$

First, we prove that $\mathcal{J}^k(d) = \mathcal{J}^1(d)$. Next, based on this proof, we propose a procedure for designing a sequence of predictors, each of memory size one, with error exponent arbitrarily close to $\mathcal{J}^1(d)$.

3.1 Proof Outline

The direct part is trivial since $\mathcal{F}_1 \subset \mathcal{F}_k$ which implies that $\mathcal{J}^k(d) \geq \mathcal{J}^1(d)$. Therefore, we focus on the converse part, namely, proving that $\mathcal{J}^1(d) \geq \mathcal{J}^k(d)$. Given an integer N, fix integers p and c such that $p \gg c \geq k$ and let $n = \lfloor N/q \rfloor$ where $q = p + c - 1$, and define

$$Z_t = \sum_{\tau=(t-1)q+c}^{tq} \rho\big(X_\tau, F_\tau(X_{\tau-k}^{\tau-1})\big), \quad 1 \leq t \leq n\,, \tag{3.1}$$

where $\{F_\tau\}_{\tau=(t-1)q+c}^{tq}$ is a sequence of prediction functions with memory k. Note that Z_t is a RV representing the total loss incurred by applying $\{F_\tau\}_{\tau=(t-1)q+c}^{tq}$ on the last p symbols of the t'th sub-block, $X_{(t-1)q+c}^{tq}$. The basic uniformly-strongly-mixing property of finite-homogeneous-irreducible-aperiodic Markov process [4, p.175],[5] imply that $\{Z_t\}_1^n$ obeys the following condition:

$$\frac{1}{\kappa^n}\prod_{t=1}^{n} P(z_t|s_t) \leq \Pr(Z_1 = z_1, \ldots, Z_n = z_n) \leq \kappa^n \prod_{t=1}^{n} P(z_t|s_t), \tag{3.2}$$

where s_t represents a sequence of p prediction functions with memory size k (one out of $r(k,p) \stackrel{\triangle}{=} [r(k)]^p = B^{A^k p}$ different sequences) associated with Z_t, namely $\{F_\tau\}_{\tau=(t-1)q+c}^{qt}$, and κ, a function of c (the gap between each two sub-blocks of size p), converges to 1 exponentially fast in c [5, Th.5] (i.e., $\kappa = 1+\exp\{-\omega c\}$, $\omega > 0$). This condition enables us to analyze the cumulative loss (represented by Z_t) of each sub-block's as "almost" independent of the other sub-blocks. Since each sub-block contributes no more than κ to the product term we obtain (3.2). Using the above condition, we can derive [3, Lem.1] a lower bound on the probability of excess loss in terms of the moment-generating function (MGF) of Z_t given the sequence of prediction functions.

In the second part of the proof, we show that for any sequence of prediction functions with memory size k, the moment-generating function (MGF) of the cumulative loss can be minimized by a sequence of prediction functions with memory size 1, i.e.,

$$E\exp\left\{\xi\sum_{j=1}^{p}\rho(X_j, \tilde{F}_j(X_{j-1}))\right\} \leq E\exp\left\{\xi\sum_{j=1}^{p}\rho(X_j, F_j(X_{j-k}^{j-1}))\right\}, \tag{3.3}$$

where $\{\tilde{F}_i\}_{i=1}^p \in \mathcal{F}_1^p$ and $\xi \geq 0$ is an arbitrary constant. In order to show that, we use a "onion-peeling" argument, originally used by Stiglitz [6], but in a rather different context. Let us rewrite the r.h.s. of (3.3), call it M, as follows:

$$\begin{aligned} M \;=\; & \sum_{x_{-k+1}^0 \in \mathcal{X}^k} P(x_{-k+1}^0) \sum_{x_1} \Pi(x_1|x_0) \exp\Big\{\xi\rho\big(x_1, F_1(x_{-k+1}^{-1}, x_0)\big)\Big\} \times \\ & \sum_{x_2} \Pi(x_2|x_1) \exp\Big\{\xi\rho\big(x_2, F_2(x_{-k+2}^{0}, x_1)\big)\Big\} \times \cdots \times \\ & \sum_{x_p} \Pi(x_p|x_{p-1}) \exp\Big\{\xi\rho\big(x_p, F_p(x_{p-k}^{p-2}, x_{p-1})\big)\Big\}. \end{aligned} \tag{3.4}$$

Consider first, the part of the expression that depends on F_p, namely, only the last summation over x_p. Note that in this part of the expression, x_{p-k}^{p-2} can simply be thought of as

an index of a function of x_{p-1} from $\mathcal{X} \to \hat{\mathcal{X}}$ (prediction function with memory size 1 which depends on the past only via this index). Therefore, for any x_{p-k}^{p-2} and any given x_{p-1}, this summation over $\{x_p\}$ cannot be smaller than

$$m(x_{p-1}) = \min_{\tilde{F}_p \in \mathcal{F}_1} \sum_{x \in \mathcal{X}} \Pi(x|x_{p-1}) \exp\Big\{\xi\rho\big(x, \tilde{F}_p(x_{p-1})\big)\Big\}. \tag{3.5}$$

Repeating this argument over the rest of the "peels" proves eq.(3.3). So far we showed that a p-tuple of prediction functions $(\tilde{F}_1, \ldots, \tilde{F}_p) \in \mathcal{F}_1^p$ for which the moment-generating function of the associated loss, at a given value of ξ, does not exceed that of a given $(F_1, \ldots, F_p) \in \mathcal{F}_k^p$.

Combining eq.(3.3) with condition (3.2) enables us to follow the proof in [3]. Note that unlike the source-coding exponent for memoryless sources, presented in [2], single-letter expression can not be given to the prediction error exponent since the prediction functions are time-variant and the source is not memoryless. In Remark 1 below, we present the error exponent of time-invariant predictor.

3.2 Predictor Design

In quest for an optimal predictor achieving the optimal prediction exponent for Markov sources, Theorem 1 confines our search to sequence of prediction functions with memory size corresponding to the Markov order of the source. Nevertheless, the way of finding that sequence of functions still needs to be specified. In the following theorem, we propose a procedure for designing a predictor with memory size 1, having an exponent arbitrarily close to the optimal exponent. This procedure is derived from the above proof of the converse part, where the MGF of the prediction loss is recursively minimized.

Theorem 2. *Given an integer $p > 2$ there exist a procedure for designing a sequence of prediction functions $\{F_i\}_{i=1}^p$, where $F_i \in \mathcal{F}_1$, satisfying the following properties:*

(i) The computational complexity is linear in p.

(ii) The designed code achieves an exponent $\mathcal{J}^1(d) - \epsilon(p)$, where $\epsilon(p) \xrightarrow[p\to\infty]{} 0$.

Procedure description:
Given an integer p calculate:

$$J_{p,1}(d) = \max_{\xi} \Big[\xi p d - M(\xi)\Big], \tag{3.6}$$

where $M(\xi)$ is evaluated in the following way:

1. For $i = 0, \ldots, p-1$ compute:

   $$m(x_{p-i-1}) = \min_{F_{p-i} \in \mathcal{F}_1} \sum_{x_{p-i}} m(x_{p-i}) \Pi(x_{p-i}|x_{p-i-1}) \cdot \exp\Big\{\xi\rho\big(x_{p-i}, F_{p-i}(x_{p-i-1})\big)\Big\},$$

   where $m(x_p) \equiv 1$.

2. Compute:

$$M(\xi) = \ln\Big(\sum_{x_0} \pi(x_0)m(x_0)\Big).$$

The procedure returns the sequence of prediction functions, $\{F_i\}_{i=1}^{p}$, which achieves $J_{p,1}(d)$.

The resulting sequence of prediction functions will be employed periodically on sub-blocks of size p, with gaps of size c^* between each two sub-blocks, on an input sequence. During the gaps, arbitrarily prediction function is employed, and the value of c^* is given by

$$c^* = \text{Round}\bigg(\arg\max_{c>2} \frac{1}{p+c}\Big[J_{p,1}(pd - c\cdot d_{\max}) - \ln\big(1+\exp(-\omega c)\big)\Big]\bigg),$$

It is easy to see that the computational complexity of the evaluation of $M(\xi)$ is linear in p, where in each step, a minimizing function is sought out of a finite set of functions, where the set size is not greater than B^A. The fact that $M(\xi)$ is piecewise convex, continuous, and $M(0) = 0$ implies that $[\xi pd - M(\xi)]$ is continuous and piecewise concave therefore $J_{p,1}(d)$ can easily evaluated using simple optimization methods as the quadratic interpolation algorithm.

To evaluate the performance of the procedure when $n \to \infty$ (i.e., the number of sub-blocks tends to infinity while the size of the sub-blocks, p, is fixed) We use [3, Lem.1,2] which gives upper and lower bounds on the prediction exponent in terms of the MGF.

Denote by ϵ the difference between the optimal performance, $\mathcal{J}^1(d)$, and the actual predictor performance, we can conclude that:

$$\epsilon < \frac{1}{q}\big[J_{p,1}(pd + cd) - J_{p,1}(pd - cd_{\max}) + 2\ln\kappa\big].$$

Note that ϵ tends to zero as $p \to \infty$.

It remains an open question whether the sequence of predictors gives rise, in general, to repetitive usage of a single predictor as $p \to \infty$.

Remark 1. When restricting attention to a fixed prediction function taken out of $\mathcal{F}_1$, it is easier to calculate its prediction exponent. For a given function $F \in \mathcal{F}_1$, define the matrix $\Pi_\xi(F)$ whose elements are

$$\Pi_\xi(F) = \{\Pi(a|b)\exp\{\xi\rho(a, F(b))\}\ ,\ \forall b \in \mathcal{X}\}.$$

Let $\lambda\big(\Pi_\xi(F)\big)$ denote the Perron-Frobenius eigenvalue of the matrix $\Pi_\xi(F)$ (Perron-Forbenius eigenvalue exist for irreducible matrices [7, Th.3.1.1]), Then by applying Theorem 3.1.2 in [7]:

$$J_F(d) = \sup_{\xi\geq 0}\big[\xi d - \ln\lambda\big(\Pi_\xi(F)\big)\big],$$

where $J_F(d)$ is the prediction exponent of F, for distortion level d. Therefore, the optimal error exponent of time-invariant predictor with memory size 1 is

$$J(d) = \max_{F\in\mathcal{F}_1} J_F(d).$$

Remark 2. The above proof technique may be useful to other problems in information theory. In [3, Ch.4], it is shown that zero-delay finite-memory encoders/decoders with memory size equals to the Markov order of the source/channel are as good as any finite memory encoders/decoders. The procedure proposed earlier can be modified and used to design a sequence of encoders and decoders with performance close to the optimal error exponent.

References

[1] N. Merhav, M. Feder, and M. Gutman, "Some properties of universal sequential predictors for binary Markov sources," *IEEE Trans. Information Theory*, vol. 39, no. 3, pp. 887–892, May 1993.

[2] N. Merhav and I. Kontoyiannis, "Source coding exponents for zero-delay coding with finite memory," *IEEE Trans. Information Theory*, vol. 49, no. 3, pp. 609–625, March 2003, available at [`http://www.ee.technion.ac.il/~merhav`].

[3] E. Sabbag, "Large deviations performance of zero-delay finite-memory lossy source codes and source–channel codes," M.Sc. thesis, Technion, I.I.T., November 2003.

[4] R. C. Bradley, "Mixing condition," in *Dependence in Probability and Statistics: A Survey of Recent Results*, ser. Progress in Probability and Statistics Series, E. Eberlein and M. S. Taqqu, Eds. Boston: Birkhäuser, 1986, vol. 11, pp. 163–192.

[5] J. R. Blum, D. L. Hanson, and L. H. Koopmans, "On the strong law of large numbers for a class of stochastic processes," *Z. Wahrsch. Verw. Gebiete*, vol. 2, pp. 1–11, 1963.

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

[7] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*, 2nd ed., I. Karatzas and M. Yor, Eds. Springer-Verlag, 1998.