Twofold Universal Prediction Schemes for Achieving the Finite-State Predictability of a Noisy Individual Binary Sequence*

Tsachy Weissman Neri Merhav Anelia Baruch

May 8, 2000

Abstract

The problem of predicting the next outcome of an individual binary sequence corrupted by noise using finite memory, is considered. The conditional finite-state (FS) predictability of an infinite individual sequence given its noisy version is defined as the minimum fraction of errors that can be made by any finite-state predictor fed by the noisy version. It is proved that the conditional FS predictability can be attained almost surely by universal sequential prediction schemes in the case where the noisy version is the output of a binary symmetric channel (BSC) whose input is the clean individual sequence. In particular, universal predictors of the original noise-free setting, which operate on the noisy sequence, have this property. Moreover, these universal predictors do not depend on the crossover probability characterizing the BSC. It is seen that the noisy setting gives rise to additional criteria by which the performance of prediction schemes can be assessed. Finally, a closer look is taken at the conditional FS predictability, and this quantity is proposed as an additional measure of the complexity of a sequence, perhaps finer and more informative than the predictive complexity of the noise-free setting.

1 Introduction

Research pertaining to the problem of sequential decision-making with an arbitrary deterministic binary sequence corrupted by noise, dates back to the fifties and sixties, where the *compound decision problem* was thoroughly investigated by statisticians, cf. [7], [14], [20], [24], [25], [27], [28]-[34], [38], and references therein. With the dawning of the information age, the last decade has witnessed a wave of intensive interest in the problem of sequential prediction of individual sequences by researchers of many disciplines, such as information theory, learning theory, game theory, statistics, economics, control theory and operations research. Accordingly, the literature pertaining to the work done in this setting over the

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

last decade is far too plentiful to be specified briefly. Representative examples can be found in [3], [4], [11], [41], [15], [39], [40], [6], and in the many references therein. The reader is referred to [17] for an overview of this setting and for a more comprehensive list of references. The important case where the individual sequence is corrupted by noise, however, has not yet received much attention in the context of individual sequences, with the exception of [1], [42]-[45].

In order to ask meaningful questions in this deterministic setting, one must limit the freedom allowed in the choice of the predictors (cf. discussion in Section 4 of [17]). The class of allowable predictors is referred to as the *comparison class.* Roughly speaking, the idea in universal prediction is to find a *single* universal predictor that competes with the best predictor in the comparison class, simultaneously for every individual sequence, in the sense that its normalized cumulative loss asymptotically never exceeds that of the best predictor in the comparison class.

One of the findings for this individual sequence prediction setting, was the existence of universal predictors for the case where the comparison class consists of all strategies that are implementable by finite-state machines. Analogously to the FS compressibility defined in [47], the FS predictability of an infinite individual sequence was defined in [11] as the minimum asymptotic fraction of errors that can be made by any FS predictor. This quantity takes on values between zero and a half, where zero corresponds to perfect predictability and a half corresponds to total unpredictability. While the definition of FS predictability enables a different optimal FS predictor for each sequence, the main result of [11] was to establish the existence of *universal* predictors, independent of the particular sequence, that always attain the FS predictability.

The purpose of this paper is to revisit the setting considered in [11] for the case where the predictors access only a noisy version of the underlying binary individual sequence. In particular, we consider the case where the individual sequence reaches the predictor after having passed through a binary symmetric channel (BSC). Analogously to the FS predictability defined in [11] for the original noise-free setting, we define the conditional FS predictability of an infinite individual sequence given its noisy version as the minimum asymptotic fraction of errors that can be made by any FS predictor allowed to base its prediction for time $t+1$ on the past t *noisy* bits. This quantity, as its analogue of the noise-free setting, takes on values between zero and a half.

The main contribution of this paper is in establishing the existence of universal predictors that always attain the conditional FS predictability in the noisy setting, in a very strong, almost-sure sense. Furthermore, it will be shown that the universal predictors of the noise-free setting continue to be universal in the noisy setting. In other words, a universal predictor tailored to the noise-free setting, continues to be asymptotically optimal even when the sequence is noisy. As these universal predictors do not depend on the crossover probability of the BSC that corrupts the clean sequence, they are rendered *twofold universal*: first, in the usual sense of universality, namely, with respect to (w.r.t.) all individual sequences (and FSM's), and secondly, w.r.t. all BSC's with crossover probabilities $p \in [0, 1/2)$. In particular, the universality of the predictors presented in [11], namely, the Markovian predictor and the incremental parsing predictor, will be assessed in much stronger and deeper senses than had previously been realized.

In a recent work [45], the problem of prediction of individual binary sequences relative to a set of experts in the presence of noise was considered. The main focus of that work was on the case of an arbitrary but finite comparison class (expert set). The essential difference between the setting considered in [45] and that of the present work is that in the former, the comparison classes were limited enough to allow the existence of predictors with uniformly vanishing relative losses (redundancy rates). The comparison class considered in this work, however, namely, that of all finite-state machines (FSM's), is clearly too rich to allow the existence of predictors with uniform redundancy rates. Hence, the target performance for a universal predictor must be modified appropriately. The results of this work join those of [45] in establishing the remarkably strong senses in which a predictor's performance can be guaranteed in the noisy setting.

The alphabet considered in this work is binary. It will be clear from the derivation of the main results, however, that predictors which are universal in the sense of the present setting can be similarly constructed for the case of any finite alphabet (cf. [45] for a discussion of general guidelines for the construction of universal predictors in the noisy setting). Yet, it is much less clear whether, for a general finite alphabet (and a reasonable generalization of a BSC for an alphabet of more than two letters), there will exist *twofold* universal predictors which will not depend on the channel parameters.

The outline of the paper is as follows. Section 3 is devoted to a brief presentation of the original noise-free setting of universal prediction for the case where the comparison class consists of the finite-state machines (FSM's), and to a summary of the main results of [11]. Section 4, is dedicated to the introduction of the noisy setting for universal prediction, of the notion of finite-state predictability corresponding to this setting, and to a presentation of the main results of this work. In Section 5, a closer look will be taken at the finite-state predictability defined for the noisy setting. It will be established that this quantity is a new and meaningful complexity measure for individual binary sequences. Finally, Section 6 summarizes the paper along with a few directions for future research.

2 Notation Conventions

Throughout the paper, for any positive integers $m < n$, we let x_m^n denote a deterministic binary vector $(x_m, \ldots, x_n) \in \{0,1\}^{n-m+1}$. One-sided binary sequences will be denoted by boldface letters, e.g., $\mathbf{x} = (x_1, x_2, \ldots) \in \{0,1\}^\infty$. Random elements will be denoted by capital letters, while their sample values will be denoted by the respective lower case letters. Thus, for example, if $Y_m^n = (Y_m, \ldots, Y_n)$ denotes a random vector, then $y_m^n = (y_m, \ldots, y_n)$ would designate a specific realization of Y_m^n. Similarly, $\mathbf{y}$ would designate a specific realization of $\mathbf{Y}$, which is a random element of $\{0,1\}^\infty$. For any pair of finite (of the same length) or half-infinite binary vectors a, b, we let $a \oplus b$ denote the vector obtained by component-wise addition modulo 2.

3 Finite-State Predictability in the Noise-free Setting

The most convenient way to formally define the prediction problem for the noise-free setting is to introduce the notion of a *predictor*. A predictor F is a sequence of functions $F_t : \{0,1\}^{t-1} \to [0,1]$, $t \geq 1$. For any predictor F and individual binary sequence $\mathbf{x} = (x_1, x_2, \ldots) \in \{0,1\}^\infty$, we define the *noiseless per-bit loss up to time* n by

$$L_F(x_1^n) \stackrel{\text{def}}{=} \frac{1}{n} \sum_{t=1}^{n} |F_t(x^{t-1}) - x_t|, \tag{1}$$

and the *asymptotic noiseless per-bit loss* by

$$L_F(\mathbf{x}) \stackrel{\text{def}}{=} \limsup_{n\to\infty} L_F(x_1^n) = \limsup_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} |F_t(x^{t-1}) - x_t|. \tag{2}$$

Note that we may view the predictor's strategy as a distribution F over $\{0,1\}^\infty$ and interpret the loss $|F_t(x^{t-1}) - x_t|$ as the error probability $\Pr\{\hat{X}_t \neq x_t\}$, where the prediction $\hat{X}_t \in$

$\{0,1\}$ is randomly drawn according to the probability $\Pr\{\hat{X}_t = 1\} = F_t(x^{t-1})$. Under this interpretation, $L_F(x_1^n)$ can be thought of as the fraction of prediction errors made by such a randomized prediction scheme. Turning to the predictors that are implementable by FSM's, for any individual sequence $\mathbf{x} \in \{0,1\}^\infty$, the *prediction rule* of an FS predictor is defined by

$$\hat{x}_{t+1} = f(s_t), \tag{3}$$

where $\hat{x}_{t+1} \in [0,1]$ is the predicted value for x_{t+1}, and s_t is the current state which takes on values in a finite set $\mathcal{S} = \{1, 2, \cdots, S\}$. The state sequence of the finite-state machine (FSM) is generated recursively according to

$$s_{t+1} = g(x_t, s_t), \tag{4}$$

where $g : \{0,1\} \times \mathcal{S} \to \mathcal{S}$ is called the *next-state* function of the FSM. Thus, a FS predictor is defined by a pair (f, g) and an initial state s_1. We will denote the noiseless per-bit loss up to time n of the FSM corresponding to (f, g) by $L_{(f,g)}(x_1^n)$.

Consider first a finite sequence $x_1^n = (x_1, \cdots, x_n)$, and suppose that the initial state s_1 and the next-state function g (and hence also the state sequence) are provided. In this case, it is easy to show (cf. [10]) that the best prediction rule for the sequence x_1^n, in the sense of minimizing (1), is given by

$$\hat{x}_{t+1} = f(s_t) = \begin{cases} \text{“0”}, & \text{if } N_n(s_t, 0) > N_n(s_t, 1), \\ \text{“1”}, & \text{otherwise,} \end{cases} \tag{5}$$

where $N_n(s, x)$, $s \in \mathcal{S}$, $x \in \{0,1\}$ is the joint count of $s_t = s$ and $x_{t+1} = x$ along the sequence x_1^n. Applying (5) to x_1^n, the minimum fraction of prediction errors achievable by a finite-state machine with next-state function g is

$$\begin{aligned} \pi(g; x_1^n) &\stackrel{\text{def}}{=} \inf_{f \in F_S} L_{(f,g)}(x_1^n) \\ &= \min_{f \in F_S^D} L_{(f,g)}(x_1^n) \qquad (6) \\ &= \frac{1}{n} \sum_{s=1}^{S} \min\{N_n(s,0), N_n(s,1)\}, \qquad (7) \end{aligned}$$

where F_S is the set of all prediction rules $f : \{1, 2, \cdots, S\} \to [0,1]$, F_S^D is the set of all 2^S prediction rules taking values in $\{0,1\}$. Equation (6) is justified by the fact that, as is evident from equation (5), there is no loss of optimality in confining our attention to the

FSM's with predictions taking values in $\{0,1\}$. The *S-state predictability* of x_1^n is defined as the minimum fraction of prediction errors with respect to all FSM's with S states:

$$\begin{aligned} \pi_S(x_1^n) &\stackrel{\text{def}}{=} \min_{g \in G_S} \pi(g; x_1^n) \\ &= \min_{(f,g) \in F_S^D \times G_S} L_{(f,g)}(x_1^n), \end{aligned} \tag{8}$$

where G_S is the set of all S^{2S} next-state functions corresponding to FSM's with no more than S states. Note that the initial state s_1 can be chosen arbitrarily since the search over G_S allows for state permutations. The *asymptotic S-state predictability* of the infinite sequence $\mathbf{x} = (x_1, x_2, \cdots)$ is defined as

$$\pi_S(\mathbf{x}) = \limsup_{n \to \infty} \pi_S(x_1^n). \tag{9}$$

Finally, the *FS predictability* is defined by

$$\pi(\mathbf{x}) = \lim_{S \to \infty} \pi_S(\mathbf{x}), \tag{10}$$

where the limit always exists as $\pi_S(\mathbf{x})$ is non-increasing with S. Observe that, by its definition, $\pi(\mathbf{x})$ is attained by a sequence of FSM's that depends on the particular sequence $\mathbf{x}$. The main result of [11] was to (constructively) establish the existence of two prediction schemes that are universal in the sense of being independent of $\mathbf{x}$ and yet asymptotically achieve $\pi(\mathbf{x})$. For completeness and concreteness in future reference, we now present one of these prediction schemes and briefly mention the other.

A The Increasing-Order Markov Predictor

An important subclass of FS predictors is the class of so-called *Markov predictors*. A Markov predictor of order k is an FS predictor with 2^k states where $s_t = (x_{t-1}, \cdots, x_{t-k})$. Similarly to (8), define the *kth-order Markov predictability* of the finite sequence x_1^n as

$$\mu_k(x_1^n) = \frac{1}{n} \sum_{x^k \in \{0,1\}^k} \min\{N_n(x^k, 0), N_n(x^k, 1)\}, \tag{11}$$

where $N_n(x^k, x) = N_n(x^{k+1})$, $x = 0, 1$, is the number of times the symbol x follows the binary string x^k in x_1^n, and where for the initial Markov state the cyclic convention $x_{-i} = x_{n-i}$, $i = 1, \cdots, k$, is used. The *asymptotic kth-order Markov predictability* of the infinite sequence $\mathbf{x}$ is defined as

$$\mu_k(\mathbf{x}) = \limsup_{n \to \infty} \mu_k(x_1^n), \tag{12}$$

and the *Markov predictability* of the sequence $\mathbf{x}$ is defined as

$$\mu(\mathbf{x}) = \lim_{k\to\infty} \mu_k(\mathbf{x}), \tag{13}$$

where the limit exists as clearly $\mu_k(\mathbf{x})$ is monotonically non-increasing with k.

One of the important findings of [11] is that the FS predictability, not only lower bounds the Markov predictability (which is immediate from the definitions and the fact that the Markov predictors are a subclass of the FS predictors), but is also an upper bound, hence the two quantities are, in fact, equal. Specifically, the following was established.

Theorem 1 ([11], Theorem 2) *For all integers $k \geq 0$, $S \geq 1$ and for any finite sequence $x_1^n \in \{0,1\}^n$,*

$$\mu_k(x_1^n) \leq \pi_S(x_1^n) + \sqrt{\frac{\ln S}{2(k+1)}}. \tag{14}$$

Note that the statement of the above theorem is non-asymptotic and holds for arbitrary integers n, k and S. In particular, taking the limit supremum as $n \to \infty$, then the limit $k \to \infty$, and finally, the limit $S \to \infty$, one obtains $\mu(\mathbf{x}) \leq \pi(\mathbf{x})$, which together with the obvious fact that $\mu(\mathbf{x}) \geq \pi(\mathbf{x})$, leads to:

$$\mu(\mathbf{x}) = \pi(\mathbf{x}). \tag{15}$$

Theorem 1 suggests that any universal prediction scheme attaining $\mu(\mathbf{x})$ is guaranteed to attain $\pi(\mathbf{x})$. It should be noted that, in the context of individual sequences, this finding is not at all trivial. This is because, in an individual sequence setting, there is no apparent analogue to mixing-type assumptions of the stochastic setting, where it is intuitively plausible that the remote past is essentially immaterial for prediction. This result is established by appealing to pure information-theoretic considerations.

We next present the sequential universal prediction scheme of [11] which exploits this fact and, by employing a Markov predictor of time-varying order, attains $\mu(\mathbf{x})$ and thus, $\pi(\mathbf{x})$. For a fixed k, consider the predictor

$$\hat{x}_{t+1} = 1 - \phi(\hat{p}_t(0|x_t, \cdots, x_{t-k+1})), \tag{16}$$

where

$$\hat{p}_t(0|x_t, \cdots, x_{t-k+1}) = \frac{N_t(x_{t-k+1} \cdots x_t 0) + 1}{N_t(x_{t-k+1} \cdots x_t) + 2}, \tag{17}$$

$\phi(\cdot)$ is given by

$$\phi(\alpha) = \begin{cases} 0, & 0 \le \alpha \le \frac{1}{2} - \varepsilon \\ \frac{1}{2\varepsilon}\left[\alpha - \frac{1}{2}\right] + \frac{1}{2}, & \frac{1}{2} - \varepsilon \le \alpha \le \frac{1}{2} + \varepsilon \\ 1, & \frac{1}{2} + \varepsilon < \alpha \le 1, \end{cases} \tag{18}$$

and $\phi(\cdot)$ is taken with $\varepsilon_{N_t(x_{t-k+1}\cdots x_t)}$, where $\varepsilon_t = 1/(2\sqrt{t+2})$ (see discussion in [11] for the motivation for defining ϕ in this way). The universal predictor is now obtained as follows: suppose that the observed data is divided into non-overlapping segments $\mathbf{x} = \mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \cdots$ and apply the kth order predictor (16) to the kth segment, $\mathbf{x}^{(k)}$. Choose a sequence α_k such that $\alpha_k \to \infty$ monotonically as $k \to \infty$, and let the length of the kth segment, denoted n_k, be at least $\alpha_k \cdot 2^k$. Letting P denote this universal predictor, we have

Theorem 2 ([11], Theorem 3) *For any finite S, and for any $x_1^n \in \{0,1\}^n$*

$$L_P(x_1^n) \le \pi_S(x_1^n) + \xi^*(n), \tag{19}$$

where $\xi^*(n) \to 0$ as $n \to \infty$.

Actually, the derivation in [11] leading to Theorem 2 shows that

$$\xi^*(n) = \frac{1}{k_n} \sum_{k=1}^{k_n} \frac{c}{\sqrt{\alpha_k}} + \sqrt{\frac{\ln S}{2(k_n+1)}}, \tag{20}$$

where c is some constant independent of x_1^n and k_n is the number of segments in x_1^n, so that if one takes, e.g., $\alpha_k = k$, then $\xi^*(n) = O(\sqrt{(\ln S)/(\ln n)})$ uniformly in $\mathbf{x} \in \{0,1\}^\infty$. Note, in particular, that Theorem 2 implies that for every individual sequence

$$L_P(\mathbf{x}) \le \pi(\mathbf{x}). \tag{21}$$

B The Incremental Parsing Predictor

The incremental parsing predictor presented in [11] is a prediction scheme based on the well-known incremental parsing algorithm introduced by Ziv and Lempel [47] in the context of data compression. The basic idea behind it is that the incremental parsing algorithm induces a technique for gradually changing the Markov order with time at an appropriate rate. The reader is referred to section V of [11] for a full description of this prediction scheme, a heuristic explanation of why it works, and for a proof of the following result:

Theorem 3 ([11], Theorem 4) *Let P be the incremental parsing predictor, then for every sequence x_1^n and any integer $k \ge 0$,*

$$L_P(x_1^n) \le \mu_k(x_1^n) + \eta(n,k), \tag{22}$$

where for a fixed k, $\eta(n,k) = O(1/\sqrt{\log n})$.

Clearly, Theorem 3, combined with Theorem 1, implies that with the incremental parsing predictor, as with the predictor of the previous subsection, for all $\mathbf{x} \in \{0,1\}^\infty$, inequality (21) holds. In other words, the incremental parsing predictor is universal as well.

4 Finite-State Predictability in the Noisy Setting

Consider now the case where the individual sequence is passed through a BSC prior to reaching the predictor. Specifically, we assume that the predictors must base their predictions for the $(t+1)$-th clean bit x_{t+1} on $Y_1^t = (Y_1, Y_2, \cdots, Y_t)$ where, for each t, $Y_t = x_t \oplus R_t$, where $\mathbf{R} = \{R_t, t \geq 1\}$ is an i.i.d. Bernoulli(p) process for some $p \in [0, 1/2)$ and $\oplus$ denotes binary (modulo 2) addition. Let $\Pr^p$ and E^p denote the probability measure under which $\mathbf{R} = \{R_t, t \geq 1\}$ is Bernoulli(p), and the expectation under this measure, respectively. The superscript p will be suppressed whenever there is no room for ambiguity. For any predictor F, we let

$$L_F(Y_1^n, x_1^n) \stackrel{\text{def}}{=} \frac{1}{n} \sum_{t=1}^{n} |F_t(Y_1^{t-1}) - x_t| \tag{23}$$

denote the *noisy per-bit loss up to time n.* Note that the noisy per-bit loss is a random variable which depends on the noise realization. Analogously to the noise-free setting, we now let

$$L_F(\mathbf{Y}, \mathbf{x}) \stackrel{\text{def}}{=} \limsup_{n\to\infty} L_F(Y_1^n, x_1^n) = \limsup_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} |F_t(Y_1^{t-1}) - x_t| \tag{24}$$

denote the *asymptotic noisy per-bit loss.* As in the noise-free case, a FS predictor is defined by

$$\hat{x}_{t+1} = f(s_t), \tag{25}$$

where $\hat{x}_{t+1} \in [0, 1]$ is the predicted value for x_{t+1}, and s_t is the current state. In contrast to the noise-free case, however, as the FS predictor accesses only the noisy sequence rather than the clean one, the state sequence of the FSM is now generated recursively according to

$$s_{t+1} = g(Y_t, s_t). \tag{26}$$

Analogously to the noise-free setting, we will denote the noisy per-bit loss up to time n of the FSM corresponding to (f, g) by $L_{(f,g)}(Y_1^n, x_1^n)$.

A Expected Loss Performance

The first, seemingly natural, performance criterion is the expected loss. To this end, we denote the expectation of the noisy per-bit loss of any predictor up to time n by

$$\begin{aligned} L_F^p(x_1^n) &\stackrel{\text{def}}{=} E^p L_F(Y_1^n, x_1^n) \\ &= E^p L_F(x_1^n \oplus R_1^n, x_1^n), \end{aligned} \tag{27}$$

which is a function of the clean individual sequence x_1^n and of the Bernoulli parameter p which is governing the noise. The quantities $\pi^{E^p}(g; x_1^n)$, $\pi_S^{E^p}(x_1^n)$, $\pi_S^{E^p}(\mathbf{x})$, and $\pi^{E^p}(\mathbf{x})$, can now quite naturally be defined analogously to the definitions given for the noise-free case, by replacing everywhere $L_{(f,g)}(x_1^n)$ with $L_{(f,g)}^p(x_1^n)$, where

$$L_{(f,g)}^p(x_1^n) \stackrel{\text{def}}{=} E^p L_{(f,g)}(Y_1^n, x_1^n). \tag{28}$$

We now proceed to show that the universal predictor of the noise-free setting of Theorem 2 is also universal for the noisy setting, under the above defined expected per-bit loss performance measure. To this end, let us first define the following. For any predictor F and $y_1^n \in \{0,1\}^n$, we define

$$\hat{L}_F(y_1^n) \stackrel{\text{def}}{=} \frac{L_F(y_1^n) - p}{1 - 2p}, \tag{29}$$

where $L_F(y_1^n)$ on the right-hand side of (29) is the per-bit loss of F, as defined in (1), evaluated on the *noisy* sequence. In other words, it is the per-bit loss of F when fed with the *noisy* sequence and judged w.r.t. the *noisy* sequence. Though this may not be apparent from the definition, $\hat{L}_F(Y_1^n)$ may be thought of as our estimator for the actual, unobserved, $L_F(Y_1^n, x_1^n)$ (cf. [45] for a broader discussion of this point). An initial justification for this observation is given in the following lemma, whose proof is deferred to the Appendix.

Lemma 4 *For any predictor F, n and $x_1^n \in \{0,1\}^n$ we have*

$$E^p \hat{L}_F(Y_1^n) = L_F^p(x_1^n). \tag{30}$$

In words, $\hat{L}_F(Y_1^n)$ is an unbiased estimator for $L_F^p(x_1^n)$.

We can now state the following.

Theorem 5 *Let P be the universal predictor of Theorem 2. Then, for any finite S, for any $x_1^n \in \{0,1\}^n$, and all $p \in [0, 1/2)$*

$$L_P^p(x_1^n) \leq \pi_S^{E^p}(x_1^n) + \frac{1}{1-2p}\xi^*(n), \tag{31}$$

where $\xi^*(n)$ is as in Theorem 2.

Proof:

$$\begin{aligned}
L_P^p(x_1^n) - \pi_S^{E^p}(x_1^n) &= E^p L_P(Y_1^n, x_1^n) - \pi_S^{E^p}(x_1^n) \\
&= E^p L_P(Y_1^n, x_1^n) - \min_{(f,g)\in F_S^D \times G_S} E^p L_{(f,g)}(Y_1^n, x_1^n) \\
&= E^p \hat{L}_P(Y_1^n) - \min_{(f,g)\in F_S^D \times G_S} E^p \hat{L}_{(f,g)}(Y_1^n) \qquad (32)\\
&\le E^p \left\{ \hat{L}_P(Y_1^n) - \min_{(f,g)\in F_S^D \times G_S} \hat{L}_{(f,g)}(Y_1^n) \right\} \\
&= E^p \left\{ \frac{L_P(Y_1^n) - p}{1-2p} - \min_{(f,g)\in F_S^D \times G_S} \frac{L_{(f,g)}(Y_1^n) - p}{1-2p} \right\} \\
&= \frac{1}{1-2p} E^p \left\{ L_P(Y_1^n) - \min_{(f,g)\in F_S^D \times G_S} L_{(f,g)}(Y_1^n) \right\} \\
&= \frac{1}{1-2p} E^p \left\{ L_P(Y_1^n) - \pi_S(Y_1^n) \right\} \\
&\le \frac{1}{1-2p} \xi^*(n), \qquad (33)
\end{aligned}$$

where (32) follows from Lemma 4 and (33) follows from Theorem 2.□

The following corollary is immediate.

Corollary 6 *Let P be the universal predictor of Theorem 2. Then for all $\mathbf{x} \in \{0,1\}^\infty$ and $p \in [0, 1/2)$*

$$L_P^p(\mathbf{x}) \le \pi^{E^p}(\mathbf{x}), \qquad (34)$$

where we let

$$L_P^p(\mathbf{x}) \stackrel{\text{def}}{=} \limsup_{n\to\infty} L_P^p(x_1^n). \qquad (35)$$

In words, P is also universal in the expectation sense for the noisy setting.

One can note that the proof of Theorem 5 did not involve the specific structure of the universal predictor. It is therefore easy to extend the proof of Theorem 5 and to show that any universal predictor of the noise-free case, namely, one for which

$$L_P(\mathbf{x}) \le \pi(\mathbf{x}) \qquad (36)$$

for all $\mathbf{x} \in \{0,1\}^\infty$, is also universal, in the expectation sense, for the noisy setting, namely, such a predictor satisfies

$$L_P^p(\mathbf{x}) \le \pi^{E^p}(\mathbf{x}) \qquad (37)$$

for all $\mathbf{x} \in \{0,1\}^\infty$ and $p \in [0, 1/2)$.

Theorem 5 and its corollary serve to give us initial indication for the fact that the universal predictors from the noise-free setting remain relevant for the noisy setting as well. The expected per-bit loss of a prediction scheme seems like a reasonable measure by which to evaluate the performance of a predictor. In particular, it seems like an acceptable yardstick for a comparison between different predictors, and especially, for evaluation of the performance of a predictor w.r.t. the class of FSM's. However, a second glance reveals that it is not sufficiently strong. To see this, let $\mathcal{F}$ be an arbitrary set of predictors, and suppose we are given a predictor P for which we know that for all x_1^n

$$L_P^p(x_1^n) - \inf_{F \in \mathcal{F}} L_F^p(x_1^n) \leq B, \tag{38}$$

for some constant $B > 0$, that is independent of x_1^n. Note now that

$$\begin{aligned} L_P^p(x_1^n) - \inf_{F \in \mathcal{F}} L_F^p(x_1^n) &= E^p L_P(Y_1^n, x_1^n) - \inf_{F \in \mathcal{F}} E^p L_F(Y_1^n, x_1^n) \\ &\leq E^p \left\{ L_P(Y_1^n, x_1^n) - \inf_{F \in \mathcal{F}} L_F(Y_1^n, x_1^n) \right\} \end{aligned} \tag{39}$$

and therefore, the upper-bound of (38) is even less informative than

$$E^p \left\{ L_P(Y_1^n, x_1^n) - \inf_{F \in \mathcal{F}} L_F(Y_1^n, x_1^n) \right\} \leq B. \tag{40}$$

But even had we known inequality (40) to hold, this would still give us relatively little information. Looking at inequality (40) alone, even for B small, it is conceivable that the expectation on the left-hand side of (40) is small, yet with considerable probability, the actual quantity of interest, namely $L_P(Y_1^n, x_1^n) - \inf_{F \in \mathcal{F}} L_F(Y_1^n, x_1^n)$, is very large. A bound as that given in inequality (40) would be valuable if we were guaranteed to be working with the same x_1^n on many repeated independent occasions. Then, combining inequality (40) with the law of large numbers would guarantee that our performance, averaged over all these occasions, is satisfactory (for B small). However, such a scenario of repeated prediction sessions with the same sequence x_1^n does not agree with the philosophy which is at the heart of the individual-sequence setting. The whole point in the individual-sequence setting is that, at each prediction session, we are working with a sequence on which we are willing to assume nothing, a fortiori, we cannot assume to be working with the same sequence on repeated occasions, as on each such occasion it could be an entirely different individual sequence.

Thus, statements that would really be appropriate and meaningful for the individual-sequence setting under consideration, are those that would guarantee the *actual* performance of a predictor on each individual sequence. As will be established in the following two subsections, such statements can indeed be made and the actual performance of the predictor on each individual sequence can indeed be guaranteed in considerably strong senses.

B Almost Sure Asymptotic Performance

In this section, the actual, rather than the expected loss of the predictors will serve as the basis for comparison and evaluation of the performance of universal predictors. We first give the appropriate analogues of the predictability quantities of the noise-free setting to our noisy setting.

Note first that, as in the noise-free setting, given a next-state function g, the best prediction rule for the sequence x_1^n is given by

$$\hat{x}_{t+1} = f(s_t) = \begin{cases} \text{“0”}, & \text{if } N_n(s_t, 0) > N_n(s_t, 1), \\ \text{“1”}, & \text{otherwise,} \end{cases} \tag{41}$$

where $N_n(s,x)$, $s \in \mathcal{S}$, $x \in \{0,1\}$ is the joint count of $s_t = s$ and $x_{t+1} = x$ along the sequence x_1^n (only now $N_n(s,x)$ is a random variable which depends on the evolution of the state sequence which, in turn, depends on the noisy sequence through (26)). For a given next-state function $g(\cdot,\cdot)$ and $y_1^n, x_1^n \in \{0,1\}^n$, the minimum fraction of prediction errors, similarly to the noise-free case, is

$$\begin{aligned} \pi_g(x_1^n|y_1^n) &\stackrel{\text{def}}{=} \inf_{f \in F_S} L_{(f,g)}(y_1^n, x_1^n) \\ &= \min_{f \in F_S^D} L_{(f,g)}(y_1^n, x_1^n) \\ &= \frac{1}{n} \sum_{s=1}^{S} \min\{N_n(s,0), N_n(s,1)\}, \end{aligned} \tag{42}$$

where, as in the noise-free case, equality (42) is justified by (41). Define the minimum loss incurred by the best FSM with up to S states as the *conditional S-state predictability* of x_1^n given y_1^n,

$$\begin{aligned} \pi_S(x_1^n|y_1^n) &\stackrel{\text{def}}{=} \min_{g \in G_S} \pi_g(x_1^n|y_1^n) \\ &= \min_{(f,g) \in F_S^D \times G_S} L_{(f,g)}(y_1^n, x_1^n). \end{aligned} \tag{43}$$

The *asymptotic conditional S-state predictability* of the infinite sequence $\mathbf{x}$ given $\mathbf{y}$ is defined as

$$\pi_S(\mathbf{x}|\mathbf{y}) = \limsup_{n\to\infty} \pi_S(x_1^n|y_1^n). \tag{44}$$

Finally, we define the *conditional FS predictability* of $\mathbf{x}$ given $\mathbf{y}$ by

$$\pi(\mathbf{x}|\mathbf{y}) = \lim_{S\to\infty} \pi_S(\mathbf{x}|\mathbf{y}), \tag{45}$$

where the limit always exists as $\pi_S(\mathbf{x}|\mathbf{y})$ is clearly non-increasing with S. Observe that, for each individual sequence $\mathbf{x}$, the quantities defined in equations (42)-(45) and evaluated on the (random) noisy sequence $\mathbf{Y}$ are random variables. It turns out, however, as will be shown in Section 5, that the conditional *FS predictability* defined in (45) is with probability one a deterministic constant for every individual sequence $\mathbf{x}$ and for every parameter $p \in [0,1]$ governing the noise process. The main result of this paper is the following:

Theorem 7 *Let P be any predictor satisfying*

$$L_P(\mathbf{x}) \leq \pi(\mathbf{x}), \ \forall \mathbf{x} \in \{0,1\}^\infty. \tag{46}$$

Then for every $p \in [0, 1/2)$

$$L_P(\mathbf{Y},\mathbf{x}) \leq \pi(\mathbf{x}|\mathbf{Y}) \ \mathrm{Pr}^p\text{-}a.s. \ \forall \mathbf{x} \in \{0,1\}^\infty. \tag{47}$$

In words, if P is any universal predictor of the noise-free setting, then P is also a universal predictor for the noisy setting in the almost-sure sense.

Discussion: The merit of this result lies first and foremost in the fact that it answers the previously open question of whether the finite-state predictability is at all achievable in the noisy setting by any sequential prediction scheme. Theorem 7 tells us that the answer to this question is affirmative, in a very strong, almost-sure sense. It should be emphasized that the answer to this question was a priori not at all trivial. This is because the efficient universal predictors from the noise-free setting (not only in the context of the FSM's but for arbitrary comparison classes, cf. e.g. [3], [15], [21]), including those of [11] presented in Section 3, can be viewed as prediction schemes which strive to imitate the predictions of those predictors in the comparison class which have been proven "reliable" in the past. In the noisy setting, however, implementation of this approach is problematic, due to the lack of complete information on the past performance of each of the predictors in the comparison class.

Yet Theorem 7 is even deeper and more surprising. It tells us that, not only is the FS predictability achievable in the noisy setting, but that we do not have to look for new prediction schemes which achieve it. Namely, we can use the universal predictors of the original noise-free setting and be guaranteed to almost surely achieve the FS predictability. In particular, this means that the universal predictors of Section 3, originally tailored for the noise-free setting, are universal in a much stronger sense than had previously been realized, namely, they are also universal w.r.t. the parameter of the noisy channel through which the clean individual sequence is passed, hence they are entitled *twofold universal.* The operative significance of this finding is that the universal predictors presented in Section 3 are completely robust in the sense that one can employ them, being completely ignorant of the possibility that the sequence may be corrupted by a noisy channel, and still be guaranteed to achieve the FS predictability.

Before proving Theorem 7, we state the following result (whose proof is deferred to the Appendix):

Lemma 8 *For any predictor F, $p \in [0, 1/2)$, and $\mathbf{x} \in \{0,1\}^\infty$, we have*

$$\limsup_{n\to\infty} \sqrt{\frac{n}{\log\log n}} |\hat{L}_F(Y_1^n) - L_F(Y_1^n, x_1^n)| \leq \frac{2\sqrt{2p}}{1-2p} \ \mathrm{Pr}^p\text{-}a.s., \tag{48}$$

where $\hat{L}_F(Y_1^n)$ was defined in (29).

The following is an immediate corollary to Lemma 8, further justifying our view of $\hat{L}_F(Y_1^n)$ as an estimator for $L_F(Y_1^n, x_1^n)$, which will be used in the proof of Theorem 7.

Corollary 9 *For any predictor F, $p \in [0, 1/2)$, and $\mathbf{x} \in \{0,1\}^\infty$, we have*

$$\lim_{n\to\infty} |\hat{L}_F(Y_1^n) - L_F(Y_1^n, x_1^n)| = 0 \ \mathrm{Pr}^p\text{-}a.s. \tag{49}$$

Proof of Theorem 7: Fix arbitrary $p \in [0, 1/2)$ and $\mathbf{x} \in \{0,1\}^\infty$. It will clearly suffice to establish the fact that

$$L_P(\mathbf{Y}, \mathbf{x}) \leq \pi_S(\mathbf{x}|\mathbf{Y}) \ \mathrm{Pr}^p\text{-a.s.} \tag{50}$$

for all S. To this end, fix S and note first that inequality (46) assures us that for all $\mathbf{y} \in \{0,1\}^\infty$

$$\begin{aligned} \limsup_{n\to\infty} L_P(y_1^n) &= L_P(\mathbf{y}) \\ &\leq \pi_S(\mathbf{y}) \end{aligned}$$

$$\begin{aligned} &= \limsup_{n\to\infty} \pi_S(y_1^n) \\ &= \limsup_{n\to\infty} \min_{(f,g)\in F_S^D\times G_S} L_{(f,g)}(y_1^n), \end{aligned} \tag{51}$$

which, by subtracting p from both ends of the above chain and then dividing by $1-2p$, gives for all $\mathbf{y}\in\{0,1\}^\infty$

$$\limsup_{n\to\infty} \hat{L}_P(y_1^n) \le \limsup_{n\to\infty} \min_{(f,g)\in F_S^D\times G_s} \hat{L}_{(f,g)}(y_1^n). \tag{52}$$

Note now that Corollary 9 implies that whatever the underlying clean individual sequence $\mathbf{x}$ may be, we have

$$\limsup_{n\to\infty} \hat{L}_P(Y_1^n) = \limsup_{n\to\infty} L_P(Y_1^n, x_1^n) = L_P(\mathbf{Y},\mathbf{x}) \ \mathrm{Pr}^p\text{-a.s.} \tag{53}$$

Furthermore, as $|F_S^D\times G_S| = (2S^2)^S < \infty$, Corollary 9 implies also that

$$\begin{aligned} \limsup_{n\to\infty} \min_{(f,g)\in F_S^D\times G_s} \hat{L}_{(f,g)}(Y_1^n) &= \limsup_{n\to\infty} \min_{(f,g)\in F_S^D\times G_s} L_{(f,g)}(Y_1^n, x_1^n) \\ &= \limsup_{n\to\infty} \pi_S(x_1^n|Y_1^n) \\ &= \pi_S(\mathbf{x}|\mathbf{Y}) \ \mathrm{Pr}^p\text{-a.s.} \end{aligned} \tag{54}$$

Combining (53) and (54) with (52) gives (50). □

Note that the proof of Theorem 7 made use of Lemma 8 only through Corollary 9. Using Lemma 8 directly, we can establish considerably finer almost sure asymptotic results of the following type:

Theorem 10 *1. Let P be a predictor satisfying for any finite S, and for any $x_1^n\in\{0,1\}^n$*

$$L_P(x_1^n) - \pi_S(x_1^n) \le a_n, \tag{55}$$

where $\{a_n\}$ is any sequence. Then for any $p\in[0,1/2)$, any finite S and any $\mathbf{x}\in\{0,1\}^\infty$

$$L_P(Y_1^n, x_1^n) - \pi_S(x_1^n|Y_1^n) \le O(\max\{a_n, \sqrt{\frac{\log\log n}{n}}\}) \ \mathrm{Pr}^p\text{-}a.s. \tag{56}$$

2. Let P be a predictor satisfying, for any integer $k\ge 0$, and for any $x_1^n\in\{0,1\}^n$

$$L_P(x_1^n) - \mu_k(x_1^n) \le b_n, \tag{57}$$

where $\{b_n\}$ is any sequence. Then for any $p\in[0,1/2)$, any finite k and any $\mathbf{x}\in\{0,1\}^\infty$

$$L_P(Y_1^n, x_1^n) - \mu_k(x_1^n|Y_1^n) \le O(\max\{b_n, \sqrt{\frac{\log\log n}{n}}\}) \ \mathrm{Pr}^p\text{-}a.s., \tag{58}$$

where

$$\mu_k(x_1^n|y_1^n) \stackrel{\text{def}}{=} \min_{F\in\mathcal{M}_k} L_F(y_1^n, x_1^n), \tag{59}$$

and $\mathcal{M}_k$ denotes the set of all Markov predictors of order up to k.

Proof:

$$\begin{aligned}
& L_P(Y_1^n, x_1^n) - \pi_S(x_1^n|Y_1^n) \\
= \;& L_P(Y_1^n, x_1^n) - \min_{(f,g)\in F_S^D\times G_S} L_{(f,g)}(Y_1^n, x_1^n) \\
= \;& \hat{L}_P(Y_1^n) - \hat{L}_P(Y_1^n) + L_P(Y_1^n, x_1^n) - \min_{(f,g)\in F_S^D\times G_S} L_{(f,g)}(Y_1^n, x_1^n) - \\
& \min_{(f,g)\in F_S^D\times G_S} \hat{L}_{(f,g)}(Y_1^n) + \min_{(f,g)\in F_S^D\times G_S} \hat{L}_{(f,g)}(Y_1^n) \\
\leq \;& \hat{L}_P(Y_1^n) - \min_{(f,g)\in F_S^D\times G_S} \hat{L}_{(f,g)}(Y_1^n) + |L_P(Y_1^n, x_1^n) - \hat{L}_P(Y_1^n)| + \\
& \max_{(f,g)\in F_S^D\times G_S} |L_{(f,g)}(Y_1^n, x_1^n) - \hat{L}_{(f,g)}(Y_1^n)| \\
= \;& \frac{L_P(Y_1^n) - p}{1-2p} - \min_{(f,g)\in F_S^D\times G_S} \frac{L_{(f,g)}(Y_1^n) - p}{1-2p} + |L_P(Y_1^n, x_1^n) - \hat{L}_P(Y_1^n)| + \\
& \max_{(f,g)\in F_S^D\times G_S} |L_{(f,g)}(Y_1^n, x_1^n) - \hat{L}_{(f,g)}(Y_1^n)| \\
= \;& \frac{1}{1-2p}\left[L_P(Y_1^n) - \min_{(f,g)\in F_S^D\times G_S} L_{(f,g)}(Y_1^n)\right] + |L_P(Y_1^n, x_1^n) - \hat{L}_P(Y_1^n)| \\
& + \max_{(f,g)\in F_S^D\times G_S} |L_{(f,g)}(Y_1^n, x_1^n) - \hat{L}_{(f,g)}(Y_1^n)| \\
= \;& \frac{1}{1-2p}\left[L_P(Y_1^n) - \pi_S(Y_1^n)\right] + |L_P(Y_1^n, x_1^n) - \hat{L}_P(Y_1^n)| + \\
& \max_{(f,g)\in F_S^D\times G_S} |L_{(f,g)}(Y_1^n, x_1^n) - \hat{L}_{(f,g)}(Y_1^n)| \\
\leq \;& a_n + |L_P(Y_1^n, x_1^n) - \hat{L}_P(Y_1^n)| + \max_{(f,g)\in F_S^D\times G_S} |L_{(f,g)}(Y_1^n, x_1^n) - \hat{L}_{(f,g)}(Y_1^n)|,
\end{aligned} \tag{60}$$

where the last inequality follows from (55). The last two terms in (60) are $O(\sqrt{(\log\log n)/n})$ $\Pr^p$-a.s. by Lemma 8 (for the last term this is guaranteed as the maximization is over $F_S^D \times G_S$ which is a finite set of cardinality $(2S^2)^S$). Consequently, we have $\Pr^p$-a.s.

$$L_P(Y_1^n, x_1^n) - \pi_S(x_1^n|Y_1^n) \leq a_n + O(\sqrt{\frac{\log\log n}{n}}) = O(\max\{a_n, \sqrt{\frac{\log\log n}{n}}\}), \tag{61}$$

which concludes the proof of the first item. The second item is proven similarly. □

In particular, for the universal predictors of [11], Theorem 10 gives:

Corollary 11 *1. Let P be the increasing order Markov predictor presented in subsection 3.A. Then for any $p \in [0, 1/2)$, any finite S and any $\mathbf{x} \in \{0,1\}^\infty$*

$$L_P(Y_1^n, x_1^n) - \pi_S(x_1^n|Y_1^n) \leq O(1/\sqrt{\log n}) \ \mathrm{Pr}^p\text{-}a.s. \tag{62}$$

2. Let P be the incremental parsing predictor of subsection 3.B. Then for any $p \in [0, 1/2)$, any finite k and any $\mathbf{x} \in \{0,1\}^\infty$

$$L_P(Y_1^n, x_1^n) - \pi_k(x_1^n|Y_1^n) \leq O(1/\sqrt{\log n}) \ \mathrm{Pr}^p\text{-}a.s. \tag{63}$$

Proof: The first item is immediate by combining the first item of Theorem 10 with Theorem 2. The second is immediate by combining the second item of Theorem 10 with Theorem 3. □

C Large Deviations Performance

The results of the previous subsection established the universality of the predictors presented in [11] (and in Section 3) for the noisy setting as well as the original noise-free setting, in an almost sure sense. We dedicate this section to a further substantiation of the universality of these predictors for the noisy setting. The introduction of noise into the model gives rise to another performance criterion - the large deviations criterion. With this new performance criterion, we establish an additional dimension of universality associated with our universal prediction schemes. The techniques used in the proof of the main result of this subsection are similar to those used for a result of a similar type in [45, subsection 3.E] in the context of prediction relative to an arbitrary expert set. The asymptotic regime considered here, however, differs from that considered in [45] so as to avoid trivialities associated with very large comparison classes of the type discussed in the introduction. Specifically, for a given predictor F, we look at the asymptotic exponential decay rate that it achieves relative to all FSM's of order S, and only then do we send S to infinity. This is made precise in what follows.

For any predictor F, define the *S-state relative loss exponent associated with F* by

$$I_S^F(p, \varepsilon) \stackrel{\mathrm{def}}{=} -\limsup_{n \to \infty} \frac{1}{n} \log \max_{x_1^n \in \{0,1\}^n} \mathrm{Pr}^p\{L_F(Y_1^n, x_1^n) - \pi_S(x_1^n|Y_1^n) > \varepsilon\}. \tag{64}$$

It should be noted that the above defined S-state relative loss exponent, though giving an asymptotic exponential decay rate, is non-asymptotic in the type of information that it conveys regarding the performance of the predictor F relative to the best FSM of order S. Note that, for each finite n, a maximum is taken over all individual binary sequences $x_1^n \in \{0,1\}^n$. This means that if we know, e.g., that $I_S^F(p,\varepsilon) > c$ this tells us that, for n large enough, the probability that the performance of the predictor F will be worse than that of the best FSM of up to S states is upper-bounded by $\exp(-nc)$ *uniformly* over all $x_1^n \in \{0,1\}^n$. Note that, as the maximization in the definition of $I_S^F(p,\varepsilon)$ is taken over all sequences for each n, even for a given S and $\varepsilon > 0$, the existence of a predictor F for which $I_S^F(p,\varepsilon)$ is positive is not a trivial fact. Nevertheless, we now proceed to define the *FS relative loss exponent associated with* F by

$$I^F(p,\varepsilon) \stackrel{\text{def}}{=} \lim_{S\to\infty} I_S^F(p,\varepsilon), \tag{65}$$

where the limit exists as $I_S^F(p,\varepsilon)$ is clearly non-increasing in S. Finally, let

$$I(p,\varepsilon) \stackrel{\text{def}}{=} \sup_F I^F(p,\varepsilon) \tag{66}$$

denote the *FS relative loss exponent*, where the supremum in the definition is over all possible predictors. We then have the following lower bound to $I(p,\varepsilon)$.

Theorem 12 *For every $p \in [0,1/2)$ and $\varepsilon > 0$, we have*

$$I(p,\varepsilon) \geq D\left(\frac{\frac{\varepsilon}{4}+\nu}{1+\nu} \| \frac{\nu}{1+\nu}\right), \tag{67}$$

where $\nu = \left[\frac{1-p}{1-2p}\right]^2$ and $D(\alpha\|\beta) = \alpha\log\frac{\alpha}{\beta} + (1-\alpha)\log\frac{1-\alpha}{1-\beta}$.

Theorem 12 guarantees that for any $\varepsilon > 0$ and $p \in [0,1/2)$, there exists a predictor F such that $I^F(p,\varepsilon) \geq D(\frac{\frac{\varepsilon}{4}+\nu}{1+\nu}\|\frac{\nu}{1+\nu})$. In other words, by the definition of $I^F(p,\varepsilon)$, this implies the existence of a predictor for which the S-state relative loss exponent $I_S^F(p,\varepsilon)$ is uniformly lower bounded for all S. This is a highly non-trivial fact as, with increasing S, the FSMs of order S is a family of predictors that becomes richer and richer. Theorem 12 follows directly from the following result, which establishes the universality of the original predictors presented in Section 3 for the noisy case in the exponential sense, in addition to the almost-sure asymptotic sense established in the previous subsection.

Theorem 13 *Let P be any of the two universal predictors presented in Section 3. Then For every $p \in [0,1/2)$ and $\varepsilon > 0$, we have*

$$I^P(p,\varepsilon) \geq D\left(\frac{\frac{\varepsilon}{4}+\nu}{1+\nu} \| \frac{\nu}{1+\nu}\right). \tag{68}$$

Remarks: Note first that Theorem 13 trivially implies Theorem 12 as

$$I(p,\varepsilon) \stackrel{\text{def}}{=} \sup_F I^F(p,\varepsilon) \geq I^P(p,\varepsilon). \tag{69}$$

Theorem 13, however, is a much stronger statement as it assesses the universality of our predictors in the exponential sense of this subsection as well. Observe that, whereas the definition of $I(p,\varepsilon)$ allows to have a different minimizing predictor F for each p and ε, Theorem 13 guarantees the performance of the universal predictors from the noise-free setting under the above defined exponential criterion, for all $p \in [0,1/2)$ and $\varepsilon > 0$, with an exponential rate being no worse than the right-hand side of (68). We merely remark here that, as may be expected, the right-hand side of (68) vanishes for $\varepsilon = 0$ and is increasing with ε. It is decreasing with p and goes to 0 for $p \nearrow 1/2$. Furthermore, it is a convex (upward) function of ε, though not of p. We forbear from a further exploration of this function as, at this point, we have no claims regarding the tightness of inequality (68). Future work will be dedicated to addressing the question of the tightness of the bounds in Theorems 12 and 13. One particularly interesting question to address is whether the inequality $I(p,\varepsilon) \geq I^P(p,\varepsilon)$, where P is one of the universal predictors of Theorem 13, is strict or tight. This question is analogous to that asked in [18] in the context of universal coding, where it was established that the LZ data compression scheme is optimal in the sense of achieving the best rate of exponential decay of the probability that the codeword length exceed a certain threshold. The analogue of this fact for our context would be the equality $I(p,\varepsilon) = I^P(p,\varepsilon)$.

The following lemma, whose proof is deferred to the Appendix, will be instrumental in the proof of Theorem 13:

Lemma 14 *For any predictor F, n, $x_1^n \in \{0,1\}^n$ and $\varepsilon > 0$ we have*

$$\Pr{}^p\{|\hat{L}_F(Y_1^n) - L_F(Y_1^n, x_1^n)| > \varepsilon\} \leq 4\exp\{-nD(\frac{\frac{\varepsilon}{2}+\nu}{1+\nu}\|\frac{\nu}{1+\nu})\}, \tag{70}$$

where $\hat{L}_F(\cdot)$ is as defined in (29).

Proof of Theorem 13: Let P be the increasing order Markov predictor presented in subsection 3.A. Fix n, $x_1^n \in \{0,1\}^n$, $p \in [0,1/2)$, $\varepsilon > 0$, $\alpha \in [0,1/2)$, and S. We have

$$\begin{aligned} & L_P(Y_1^n, x_1^n) - \pi_S(x_1^n|Y_1^n) \\ = \; & L_P(Y_1^n, x_1^n) - \min_{(f,g)\in F_S^D \times G_S} L_{(f,g)}(Y_1^n, x_1^n) \end{aligned}$$

$$
\begin{aligned}
&= \hat{L}_P(Y_1^n) - \min_{(f,g)\in F_S^D\times G_S} \hat{L}_{(f,g)}(Y_1^n) + L_P(Y_1^n, x_1^n) - \hat{L}_P(Y_1^n) - \\
&\quad \left(\min_{(f,g)\in F_S^D\times G_S} L_{(f,g)}(Y_1^n, x_1^n) - \min_{(f,g)\in F_S^D\times G_S} \hat{L}_{(f,g)}(Y_1^n) \right) \\
&\leq \hat{L}_P(Y_1^n) - \min_{(f,g)\in F_S^D\times G_S} \hat{L}_{(f,g)}(Y_1^n) + |L_P(Y_1^n, x_1^n) - \hat{L}_P(Y_1^n)| + \\
&\quad \max_{(f,g)\in F_S^D\times G_S} |L_{(f,g)}(Y_1^n, x_1^n) - \hat{L}_{(f,g)}(Y_1^n)|. \qquad (71)
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
& \Pr{}^p\{L_P(Y_1^n, x_1^n) - \pi_S(x_1^n|Y_1^n) > \varepsilon\} \\
\leq\ & \Pr{}^p\left\{ \hat{L}_P(Y_1^n) - \min_{(f,g)\in F_S^D\times G_S} \hat{L}_{(f,g)}(Y_1^n) + |L_P(Y_1^n, x_1^n) - \hat{L}_P(Y_1^n)| + \right. \\
& \left. \max_{(f,g)\in F_S^D\times G_S} |L_{(f,g)}(Y_1^n, x_1^n) - \hat{L}_{(f,g)}(Y_1^n)| > \varepsilon \right\} \\
\leq\ & \Pr{}^p\left\{ \hat{L}_P(Y_1^n) - \min_{(f,g)\in F_S^D\times G_S} \hat{L}_{(f,g)}(Y_1^n) > (1-2\alpha)\varepsilon \right\} + \\
& \Pr{}^p\{|L_P(Y_1^n, x_1^n) - \hat{L}_P(Y_1^n)| > \alpha\varepsilon\} + \\
& \Pr{}^p\left\{ \max_{(f,g)\in F_S^D\times G_S} |L_{(f,g)}(Y_1^n, x_1^n) - \hat{L}_{(f,g)}(Y_1^n)| > \alpha\varepsilon \right\} \\
\leq\ & \Pr{}^p\left\{ \frac{L_P(Y_1^n) - p}{1-2p} - \min_{(f,g)\in F_S^D\times G_S} \frac{L_{(f,g)}(Y_1^n) - p}{1-2p} > (1-2\alpha)\varepsilon \right\} + \\
& \Pr{}^p\{|L_P(Y_1^n, x_1^n) - \hat{L}_P(Y_1^n)| > \alpha\varepsilon\} + \\
& |F_S^D \times G_S| \max_{(f,g)\in F_S^D\times G_S} \Pr{}^p\left\{ |L_{(f,g)}(Y_1^n, x_1^n) - \hat{L}_{(f,g)}(y_1^n)| > \alpha\varepsilon \right\} \\
\leq\ & \Pr{}^p\left\{ L_P(Y_1^n) - \min_{(f,g)\in F_S^D\times G_S} L_{(f,g)}(Y_1^n) > (1-2p)(1-2\alpha)\varepsilon \right\} \qquad (72) \\
& +4[1+(2S^2)^S]\exp\{-nD(\frac{\frac{\alpha\varepsilon}{2}+\nu}{1+\nu}\|\frac{\nu}{1+\nu}) \\
=\ & \Pr{}^p\{L_P(Y_1^n) - \pi_S(Y_1^n) > (1-2p)(1-2\alpha)\varepsilon\} \qquad (73) \\
& +4[1+(2S^2)^S]\exp\{-nD(\frac{\frac{\alpha\varepsilon}{2}+\nu}{1+\nu}\|\frac{\nu}{1+\nu}), \qquad (74)
\end{aligned}
$$

where inequality (72) follows from Lemma 14. But, from Theorem 2 and the remark proceeding it, we know that $L_P(y_1^n) - \pi_S(y_1^n) \leq o(1)$ uniformly in all $\mathbf{y}$ and therefore, for sufficiently large n, the first term in (73) annihilates. Consequently, for n sufficiently large we have

$$
\Pr{}^p\{L_P(Y_1^n, x_1^n) - \pi_S(x_1^n|Y_1^n) > \varepsilon\} \leq 4[1+(2S^2)^S]\exp\{-nD(\frac{\frac{\alpha\varepsilon}{2}+\nu}{1+\nu}\|\frac{\nu}{1+\nu}). \qquad (75)
$$

Since the right-hand side of (75) is independent of x_1^n, this gives

$$\max_{x_1^n \in \{0,1\}^n} \Pr^p\{L_P(y_1^n, x_1^n) - \pi_S(x_1^n|Y_1^n) > \varepsilon\} \leq 4[1+(2S^2)^S] \exp\{-nD(\frac{\frac{\alpha\varepsilon}{2}+\nu}{1+\nu}\|\frac{\nu}{1+\nu}). \quad (76)$$

Taking the logarithm, dividing by n, and sending n to infinity gives

$$\begin{aligned} I_S^P(p,\varepsilon) &= -\limsup_{n\to\infty} \frac{1}{n} \log \max_{x_1^n \in \{0,1\}^n} \Pr^p\{L_P(Y_1^n, x_1^n) - \pi_S(x_1^n|Y_1^n) > \varepsilon\} \\ &\geq D(\frac{\frac{\alpha\varepsilon}{2}+\nu}{1+\nu}\|\frac{\nu}{1+\nu}). \end{aligned} \quad (77)$$

Now, since $\alpha \in [0, 1/2)$ was arbitrary, (77) gives us

$$I_S^P(p,\varepsilon) \geq \lim_{\alpha \nearrow 1/2} D(\frac{\frac{\alpha\varepsilon}{2}+\nu}{1+\nu}\|\frac{\nu}{1+\nu}) = D(\frac{\frac{\varepsilon}{4}+\nu}{1+\nu}\|\frac{\nu}{1+\nu}). \quad (78)$$

Finally, since the S on the left-hand side of (78) was arbitrary, we have

$$I^P(p,\varepsilon) = \lim_{S\to\infty} I_S^P(p,\varepsilon) \geq D(\frac{\frac{\varepsilon}{4}+\nu}{1+\nu}\|\frac{\nu}{1+\nu}). \quad (79)$$

The proof for the incremental parsing predictor of subsection 3.B is handled similarly. □

5 The Conditional Finite-State Predictability as a New Complexity Measure

The conditional FS predictability defined in subsection 4.B, $\pi(\mathbf{x}|\mathbf{y})$, is a function of both $\mathbf{x}$ and $\mathbf{y}$. Therefore, $\pi(\mathbf{x}|\mathbf{Y})$ is a random variable which depends on the noise realization. As it turns out, however, $\pi(\mathbf{x}|\mathbf{Y})$ is almost surely a deterministic constant for every individual sequence $\mathbf{x}$ and for every parameter $p \in [0,1]$ which may be governing the noise process. More precisely, we have the following:

Theorem 15 *For every $p \in [0,1]$ and $\mathbf{x} \in \{0,1\}^\infty$, there exists $c(p,\mathbf{x})$ such that*

$$\pi(\mathbf{x}|\mathbf{Y}) = c(p,\mathbf{x}) \ \Pr^p\text{-}a.s. \quad (80)$$

Theorem 15 will be proven by combining Kolmogorv's zero-one law with the following result (proof of which is deferred to the Appendix):

Lemma 16 *For any $\mathbf{x}, \tilde{\mathbf{y}}, \hat{\mathbf{y}} \in \{0,1\}^\infty$ such that*

$$\sum_{t=1}^{\infty} \tilde{y}_t \oplus \hat{y}_t < \infty, \quad (81)$$

we have

$$\pi(\mathbf{x}|\tilde{\mathbf{y}}) = \pi(\mathbf{x}|\hat{\mathbf{y}}). \quad (82)$$

In words, $\pi(\mathbf{x}|\mathbf{y})$ is invariant to finitely many substitutions of coordinates of $\mathbf{y}$.

Proof of Theorem 15: Let $\tau_n = \mathcal{F}(R_n, R_{n+1}, \ldots)$ and $\tau = \cap_{n \geq 1} \tau_n$ (the tail σ-field). Lemma 16 implies that for all n, $\pi(\mathbf{x}|\mathbf{Y}) \in \tau_n$ (as, clearly, the value of $\pi(\mathbf{x}|\mathbf{Y})$ is invariant to changes in the first $n-1$ bits of $\mathbf{Y}$). Therefore $\pi(\mathbf{x}|\mathbf{Y}) \in \tau$. Now, under $\Pr^p$, $\{R_t, t \geq 1\}$ is an i.i.d. Bernoulli(p) process, a-fortiori of independent components. It therefore follows from the Kolmogorov zero-one law (cf., e.g., [9, Theorem 7] or [2, Theorem 3.12]) that there exists some constant $c(p, \mathbf{x})$ such that $\pi(\mathbf{x}|\mathbf{Y}) = c(p, \mathbf{x})$ $\Pr^p$-a.s. $\square$

As is now justified by Theorem 15, we define the *p-conditional FS predictability* of $\mathbf{x}$ as $c(p, \mathbf{x})$. In [11], the (noise-free) FS-predictability was suggested as a complexity measure, analogously to the complexity definitions of Solomonoff [35], Kolmogorov [16], and Chaitin [5]. The p-conditional FS predictability may be seen as a new complexity measure, generalizing the predictive complexity suggested in [11], interpreted as the minimum fraction of errors made by a universal Turing machine in sequentially predicting the future of the sequence, based on its noisy past.

The relationship between the original noiseless predictability $\pi(\mathbf{x})$ and the p-conditional predictability $c(p, \mathbf{x})$ is an interesting and non-trivial question. Observe that, for a given $\mathbf{x} \in \{0,1\}^\infty$, $c(\cdot, \mathbf{x})$ is not even necessarily increasing with p. Let, for example, $\mathbf{x}$ be any periodic sequence of period S. Clearly, the S-state predictability of such a sequence in the noise-free as well as in the p-conditional noisy sense is zero. Therefore, for such a sequence, we have for all p

$$c(p, \mathbf{x}) = \pi(\mathbf{x}). \tag{83}$$

This observation excludes the possibility of a meaningful lower-bound on $c(p, \mathbf{x})$ in terms of $\pi(\mathbf{x})$ alone, in the spirit of "Mrs. Gerber's Lemma" [46] which upper bounds the entropy rate of the output of a BSC in terms of the entropy rate of its input.

One may hastily argue that, although not necessarily strictly, $c(\cdot, \mathbf{x})$ is increasing for $p \in [0, 1/2]$ for the following reason. Suppose the contrary is true. Namely, that there exist $\mathbf{x} \in \{0,1\}^\infty$ and $\tilde{p} < \hat{p} \leq 1/2$ such that $c(\tilde{p}, \mathbf{x}) > c(\hat{p}, \mathbf{x})$. This, one may argue, would lead to a contradiction as the noisy sequence emitted from the BSC with parameter $\tilde{p}$ can be passed through another BSC, prior to prediction, which will make it equivalent to a sequence that had been passed through a BSC with parameter $\hat{p}$. Note that this line of argumentation takes the possibility of generating an i.i.d. Bernoulli sequence (which is what passing the noisy sequence through another BSC really means) for granted. The validity of this assumption in the FSM-context, however, is questionable, as an FSM (of the deterministic type that we consider in this framework) cannot really generate anything random. Thus, as can be seen, the attempt to quantify and characterize the p-conditional FS predictability is not a trivial task, which gives rise to questions and considerations of notable depth.

The following result gives initial indication for the fact that $c(p, \mathbf{x})$ is indeed a meaningful and nondegenerate generalization of the original noiseless predictive complexity $\pi(\mathbf{x})$. In particular, it establishes the fact that $c(p, \mathbf{x})$ has a finer resolution and a higher ability to separate and order individual sequences than it noise-free origin $\pi(\mathbf{x})$.

Claim 17 *For any $p \in [0, 1/2]$ there exist sequences $\tilde{\mathbf{x}}, \hat{\mathbf{x}} \in \{0,1\}^\infty$ for which*

$$\pi(\tilde{\mathbf{x}}) = \pi(\hat{\mathbf{x}}) = \frac{1}{4}, \tag{84}$$

$$c(p, \tilde{\mathbf{x}}) = \frac{1}{4}, \tag{85}$$

and

$$c(p, \hat{\mathbf{x}}) = \frac{1}{4} + \frac{p}{2}. \tag{86}$$

In particular, $\tilde{\mathbf{x}}$ and $\hat{\mathbf{x}}$ have the same noiseless predictive complexity yet have different values for their p-conditional FS predictability.

Actually, it will be established that for any $p \in [0, 1/2]$ there exist "many" pairs of sequences which satisfy Theorem 17, where the meaning of "many" will be made precise in what follows. To set and formalize the scene for the proof, we let P^x denote the probability measure on $\Omega_x = \{(x_1, x_2, \ldots) : x_i \in \{0,1\}\}$ (equipped with the product sigma algebra $\mathcal{F} = \mathcal{S} \times \mathcal{S} \times \ldots$, where $\mathcal{S} = \{\phi, \{0\}, \{1\}, \{0,1\}\}$) under which $\mathbf{X}$ is an i.i.d. Bernoulli(1/2) sequence. Similarly, let P^n denote the probability measure on $(\Omega_n = \{(n_1, n_2, \ldots) : n_i \in \{0,1\}\}, \mathcal{F})$ under which $\mathbf{N}$ is an i.i.d Bernoulli(p) sequence[1] (where $p \in [0, 1/2]$ is fixed throughout). We will be working with the probability space $(\Omega_x \times \Omega_n, \mathcal{F} \times \mathcal{F}, P^x \times P^n)$, under which $\mathbf{X}$ and $\mathbf{N}$ are independent. Using $\mathbf{X} = (X_1, X_2, \ldots)$, we construct two sequences: $\tilde{\mathbf{X}} = (X_1, 0, X_2, 0, X_3, 0, \ldots)$ and $\hat{\mathbf{X}} = (X_1, X_1, X_2, X_2, X_3, X_3, \ldots)$. We further let $\tilde{\mathbf{Y}} = \tilde{\mathbf{X}} \oplus \mathbf{N}$ and $\hat{\mathbf{Y}} = \hat{\mathbf{X}} \oplus \mathbf{N}$. As will be clear from the proof of Theorem 17 below, not only does there exist an $\mathbf{x} \in \{0,1\}^\infty$ for which the pair $(\tilde{\mathbf{x}}, \hat{\mathbf{x}})$ satisfies equations (84)-(86), but, this holds

[1] The script n is used here in the notation for the probability measure and the sample space, P^n and Ω_n, of the process $\mathbf{N}$ and should not be confused with a sequence length (as we henceforth deal with infinite sequences only).

true for P^x- almost every $\mathbf{x}$. Appendix C is dedicated to establishing

$$\pi(\tilde{\mathbf{X}}|\tilde{\mathbf{Y}}) = \frac{1}{4} \quad P^x \times P^n \ \ -\text{a.s.} \tag{87}$$

and

$$\pi(\hat{\mathbf{X}}|\hat{\mathbf{Y}}) = \frac{1}{4} + \frac{p}{2} \quad P^x \times P^n \ \ -\text{a.s.} \tag{88}$$

Equipped with equations (87) and (88), we can now proceed to establish Claim 17.

Proof of Claim 17: Note first that, by taking $p = 0$, equation (87) also gives us

$$\pi(\tilde{\mathbf{X}}) = \frac{1}{4} \quad P^x \times P^n \ \ -\text{a.s.} \tag{89}$$

and, similarly, equation (88) also gives us

$$\pi(\hat{\mathbf{X}}) = \frac{1}{4} \quad P^x \times P^n \ \ -\text{a.s.} \tag{90}$$

Letting $A = \{(\mathbf{x}, \mathbf{n}) : \pi(\tilde{\mathbf{x}}) = 1/4\}$, $B = \{(\mathbf{x}, \mathbf{n}) : \pi(\hat{\mathbf{x}}) = 1/4\}$, $C = \{(\mathbf{x}, \mathbf{n}) : \pi(\tilde{\mathbf{x}}|\tilde{\mathbf{y}}) = 1/4\}$, $D = \{(\mathbf{x}, \mathbf{n}) : \pi(\hat{\mathbf{x}}|\hat{\mathbf{y}}) = 1/4 + p/2\}$ and $I = A \cap B \cap C \cap D$, it follows from (89), (90), (87), (88) that

$$P^x \times P^n(I) = 1. \tag{91}$$

This clearly implies also that for P^x-almost every $\mathbf{x} \in \{0,1\}^\infty$

$$P^n(\mathbf{n} : (\mathbf{x}, \mathbf{n}) \in I) = 1. \tag{92}$$

From the definition of the set I it now follows that we have established the fact that for P^x-almost every $\mathbf{x} \in \{0,1\}^\infty$, the sequences $\tilde{\mathbf{x}}, \hat{\mathbf{x}}$ satisfy: $\pi(\tilde{\mathbf{x}}) = \pi(\hat{\mathbf{x}}) = 1/4$, $\pi(\tilde{\mathbf{x}}|\tilde{\mathbf{Y}}) = 1/4$ P^n-almost surely, and $\pi(\hat{\mathbf{x}}|\hat{\mathbf{Y}}) = 1/4 + p/2$ P^n-almost surely. Finally, by Theorem 15, this implies that for P^x-almost every $\mathbf{x} \in \{0,1\}^\infty$ we have: $\pi(\tilde{\mathbf{x}}) = \pi(\hat{\mathbf{x}}) = c(p, \tilde{\mathbf{x}}) = 1/4$ and $c(p, \hat{\mathbf{x}}) = 1/4 + p/2$. □

6 Conclusion and Future Directions

The problem of prediction of noisy versions of individual binary sequences was studied for the comparison class of all finite-state predictors. It was shown that the universal predictors of the noise-free setting remain valid for the noisy setting in surprisingly strong meanings. It was also seen how the introduction of a stochastic noise process into a setting which had originally been deterministic gives rise to questions concerning performance criteria, as the large deviations performance criterion, which have no analogues from the noise-free setting.

The conditional FS predictability of a binary individual sequence was defined as the best asymptotic performance achievable by any finite-state machine in predicting the sequence when observing the output of a BSC whose input is the sequence. Though defined in a stochastic setting, this quantity was shown to essentially be a deterministic constant $c(p, \mathbf{x})$, depending on the noise parameter p and on the individual sequence. It was then argued that $c(p, \mathbf{x})$, termed "the noisy predictive complexity of $\mathbf{x}$" is a new meaningful complexity measure. In particular, it was shown that $c(p, \mathbf{x})$ has a finer resolution in the sense of having a better ability to separate and order individual sequences than the predictive complexity $\pi(\mathbf{x})$ of [11].

One of the directions for future work is an investigation of the tightness of the large deviations upper-bound established in subsection 4.C. Another interesting direction is to complete the picture associated with prediction of binary sequences in the noisy setting by addressing the case of probabilistic rather than individual sequences, as was done in [47, Theorem 4] in the context of data compression and in [19, Theorem 3] in the context of noise-free prediction. These issues, along with additional related ones, are currently under investigation.

Appendix A

Proofs of Lemmas 4, 8 and 14: Throughout this section, we fix a predictor F, $\mathbf{x} \in \{0,1\}^\infty$, and $p \in [0, 1/2)$. As there is no room for ambiguity, we omit the superscripts from $\Pr^p$ and E^p. Define first

$$\Delta_F(Y_1^n, x_1^n) \stackrel{\text{def}}{=} n\left(L_F(Y_1^n, x_1^n) - \hat{L}_F(Y_1^n)\right). \tag{A.1}$$

We then have

$$\begin{aligned}
\Delta_F(Y_1^n, x_1^n) &= n\left(L_F(Y_1^n, x_1^n) - \hat{L}_F(Y_1^n)\right) \\
&= \sum_{t=1}^n |F_t(Y_1^{t-1}) - x_t| - \frac{n[L_F(Y_1^n) - p]}{1-2p} \\
&= \sum_{t=1}^n |F_t(Y_1^{t-1}) - x_t| - \sum_{t=1}^n \frac{|F_t(Y_1^{t-1}) - Y_t| - p}{1-2p} \\
&= \sum_{t=1}^n \left\{|F_t(Y_1^{t-1}) - x_t| - \frac{|F_t(Y_1^{t-1}) - Y_t| - p}{1-2p}\right\} \\
&= \sum_{t=1}^n \{(1 - x_t)F_t(Y_1^{t-1}) + x_t(1 - F_t(Y_1^{t-1}))
\end{aligned}$$

$$
\begin{aligned}
&-\frac{(1-Y_t)F_t(Y_1^{t-1})+Y_t(1-F_t(Y_1^{t-1}))-p}{1-2p}\Bigg\} \\
&= \sum_{t=1}^{n}\frac{(1-x_t)(1-2p)+p+Y_t-1}{1-2p}F_t(Y_{t-1})+ \\
&\quad \sum_{t=1}^{n}\frac{x_t(1-2p)-Y_t+p}{1-2p}(1-F_t(Y_{t-1})). \qquad \text{(A.2)}
\end{aligned}
$$

Denoting $k(x_t, Y_t) = \frac{(1-x_t)(1-2p)+p+Y_t-1}{1-2p}$ and $l(x_t, Y_t) = \frac{x_t(1-2p)-Y_t+p}{1-2p}$, we have

$$
\Delta_F(Y_1^n, x_1^n) = \sum_{t=1}^{n} k(x_t, Y_t)F_t(Y_{t-1}) + \sum_{t=1}^{n} l(x_t, Y_t)(1-F_t(Y_{t-1})). \qquad \text{(A.3)}
$$

Clearly, $\{k(x_1, Y_1), k(x_2, Y_2), \ldots\}$ are independent (as $k(x_t, Y_t) \in \sigma(R_t)$) and likewise, $\{l(x_1, Y_1), l(x_2, Y_2), \ldots\}$ are independent. Furthermore, one can easily confirm that

$$
\begin{aligned}
k(x_t = 0, Y_t) &= \begin{cases} -\frac{p}{1-2p} & \text{w.p. } 1-p \\ \frac{1-p}{1-2p} & \text{w.p. } p \end{cases} \\
k(x_t = 1, Y_t) &= \begin{cases} \frac{p}{1-2p} & \text{w.p. } 1-p \\ \frac{p-1}{1-2p} & \text{w.p. } p \end{cases}
\end{aligned} \qquad \text{(A.4)}
$$

and that

$$
\begin{aligned}
l(x_t = 0, Y_t) &= \begin{cases} \frac{p}{1-2p} & \text{w.p. } 1-p \\ \frac{p-1}{1-2p} & \text{w.p. } p \end{cases} \\
l(x_t = 1, Y_t) &= \begin{cases} \frac{-p}{1-2p} & \text{w.p. } 1-p \\ \frac{1-p}{1-2p} & \text{w.p. } p. \end{cases}
\end{aligned} \qquad \text{(A.5)}
$$

Letting $m_n \stackrel{\text{def}}{=} \sum_{t=1}^{n} k(x_t, Y_t)F_t(Y_1^{t-1})$, we have for all $n \geq 1$

$$
\begin{aligned}
E(m_{n+1}|\mathcal{F}_n) &= E(\sum_{t=1}^{n+1} k(x_t, Y_t)F_t(Y_1^{t-1})|\mathcal{F}_n) \\
&= E(k(x_{n+1}, Y_{n+1})F_{n+1}(Y_1^n)|\mathcal{F}_n) + \\
&\quad E(\sum_{t=1}^{n} k(x_t, Y_t)F_t(Y_1^{t-1}|\mathcal{F}_n) \\
&= F_{n+1}(Y_1^n)E(k(x_{n+1}, Y_{n+1})|\mathcal{F}_n) + \\
&\quad \sum_{t=1}^{n} k(x_t, Y_t)F_t(Y_1^{t-1}) \\
&= F_{n+1}(Y_1^n)E(k(x_{n+1}, Y_{n+1})) + \\
&\quad \sum_{t=1}^{n} k(x_t, Y_t)F_t(Y_1^{t-1}) \\
&= 0 + m_n, \qquad \text{(A.6)}
\end{aligned}
$$

where the last equation follows from the fact that $Ek(x_{n+1}, Y_{n+1}) = 0$ (this is clear from (A.4)), and the definition of m_n. The one before follows from the independence of $\mathcal{F}_n$

and $\sigma(R_{n+1})$ (and the fact that $k(x_{n+1}, Y_{n+1})$ is $\sigma(R_{n+1})$-measurable). The equality before that follows from the measurability of $F_{n+1}(Y_1^n)$ and of $\sum_{t=1}^n k(x_t, Y_t) F_t(Y_1^{t-1}$ w.r.t. $\mathcal{F}_n$. Consequently, we have established the fact that $\{m_n, \mathcal{F}_n\}_{n\geq 0}$ (with $m_n \equiv 0$ and $\mathcal{F}_0$ being the trivial σ-algebra) is a zero-mean martingale. A similar argument establishes the same for the second term on the right-hand side of (A.3), which, in turn, establishes the fact that $\{\Delta_F(Y_1^n, x_1^n), \mathcal{F}_n\}_{n\geq 0}$ is a zero-mean martingale. In particular, this implies $E\Delta_F(Y_1^n, x_1^n) = 0$, which is precisely Lemma 4 as it implies

$$E^p \hat{L}_F(Y_1^n) = E^p L_F(Y_1^n, x_1^n) = L_F^p(x_1^n). \tag{A.7}$$

Proof of Lemma 8: By Equation (A.3), the triangle inequality and the fact that the lim sup of a sum is not greater than the sum of the lim sups we have for all sample paths:

$$\begin{aligned}\limsup_{n\to\infty} \frac{|\Delta_F(Y^n, x^n)|}{\sqrt{n \log\log n}} \quad \leq \quad & \limsup_{n\to\infty} \frac{|\sum_{t=1}^n k(x_t, Y_t) F_t(Y_{t-1})|}{\sqrt{n \log\log n}} + \\ & \limsup_{n\to\infty} \frac{|\sum_{t=1}^n l(x_t, Y_t)(1 - F_t(Y_{t-1}))|}{\sqrt{n \log\log n}}. \end{aligned} \tag{A.8}$$

Consider the first term on the right hand side of (A.8) and let

$$\begin{aligned} s_n^2 &= \sum_{t=1}^n E[k(x_t, Y_t)^2 (F_t(Y^{t-1}))^2 | \mathcal{F}_{t-1}] && \text{(A.9)} \\ &= \sum_{t=1}^n (F_t(Y^{t-1}))^2 E[k(x_t, Y_t)^2 | \mathcal{F}_{t-1}] \\ &= \frac{p - p^2}{(1-2p)^2} \sum_{t=1}^n (F_t(Y^{t-1}))^2 && \text{(A.10)} \\ &\leq \frac{(p - p^2)}{(1-2p)^2} n, && \text{(A.11)} \end{aligned}$$

where (A.10) follows by the fact that $k(x_t, Y_t)$ is independent of $\mathcal{F}_{t-1}$ and that, as is easily verifiable from (A.4), we have

$$Ek(x_t, Y_t)^2 = \frac{p - p^2}{(1-2p)^2}. \tag{A.12}$$

Now, since the magnitude of each of the summands in the first sum on the right hand side of (A.3) is bounded by $\frac{1-p}{1-2p}$, the hypotheses of the martingale analogue of Kolmogorov's law of the iterated logarithm (cf. [36, Theorem 1]) are satisfied and we have

$$\limsup_{n\to\infty} \frac{|\sum_{t=1}^n k(x_t, Y_t) F_t(Y_{t-1})|}{\sqrt{\frac{2p}{(1-2p)^2} n \log\log n}}$$

$$\leq \limsup_{n\to\infty} \frac{|\sum_{t=1}^n k(x_t,Y_t)F_t(Y_{t-1})|}{\sqrt{2\frac{(p-p^2)}{(1-2p)^2}n\log\log\frac{(p-p^2)}{(1-2p)^2}n}}$$

$$\leq \limsup_{n\to\infty} \frac{|\sum_{t=1}^n k(x_t,Y_t)F_t(Y_{t-1})|}{\sqrt{2s_n^2\log\log s_n^2}} \leq 1 \ \Pr^p\text{-a.s.}, \tag{A.13}$$

where the second inequality follows by (A.9) and we therefore have

$$\limsup_{n\to\infty} \frac{|\sum_{t=1}^n k(x_t,Y_t)F_t(Y_{t-1})|}{\sqrt{n\log\log n}} \leq \frac{\sqrt{2p}}{1-2p} \ \Pr^p\text{-a.s.} \tag{A.14}$$

A similar bound is obtained for the second term on the right hand side of (A.8), which, combined with (A.8), gives

$$\limsup_{n\to\infty} \frac{|\Delta_F(Y^n,x^n)|}{\sqrt{n\log\log n}} \leq \frac{2\sqrt{2p}}{1-2p} \ \Pr^p\text{-a.s.}, \tag{A.15}$$

which, by the definition of $\Delta_F(Y^n,x^n)$, is precisely Lemma 8. □

Prior to proving Lemma 14, we recall the following concentration inequality for bounded martingale differences:

Theorem 18 ([8], Corollary 2.4.7) *Suppose $\nu > 0$ and the real valued random variables $\{Z_n : n = 1,2,\ldots\}$ are such that both $Z_n \leq 1$ almost surely, and $E[Z_n|S_{n-1}] = 0$, $E[Z_n^2|S_{n-1}] \leq \nu$ for $S_n \stackrel{\text{def}}{=} \sum_{j=1}^n Z_j$, $S_0 = 0$. Then, for all $\alpha > 0$,*

$$\Pr(n^{-1}S_n \geq \alpha) \leq \exp\left(-nD(\frac{\alpha+\nu}{1+\nu}||\frac{\nu}{1+\nu})\right). \tag{A.16}$$

Proof of Lemma 14: By the union bound, it will clearly suffice to establish for all $\varepsilon > 0$

$$\Pr\left\{\frac{1}{n}\Delta_F(Y^n,x^n) \geq \epsilon\right\} \leq 2\exp\left(-nD(\frac{\frac{\epsilon}{2}+\nu}{1+\nu}||\frac{\nu}{1+\nu})\right) \tag{A.17}$$

and

$$\Pr\left\{\frac{1}{n}\Delta_F(Y^n,x^n) \leq -\epsilon\right\} \leq 2\exp\left(-nD(\frac{\frac{\epsilon}{2}+\nu}{1+\nu}||\frac{\nu}{1+\nu})\right). \tag{A.18}$$

Letting $Z_n = k(x_n,Y_n)F_n(Y_{n-1})$ and

$$S_n = \sum_{t=1}^n k(x_t,Y_t)F_t(Y_{t-1}), \tag{A.19}$$

it follows from Theorem 18, from equation (A.4) and from the fact that $F_t(\cdot) \leq 1$, that

$$\Pr\left\{\frac{1}{n}\sum_{t=1}^n k(x_t,Y_t)F_t(Y_{t-1}) \geq \epsilon/2\right\} \leq \exp\left(-nD(\frac{\frac{\epsilon}{2}+\nu}{1+\nu}||\frac{\nu}{1+\nu})\right). \tag{A.20}$$

Similarly, one obtains

$$\Pr\left\{\frac{1}{n}\sum_{t=1}^{n} l(x_t, Y_t)(1 - F_t(Y_{t-1})) \geq \epsilon/2\right\} \leq \exp\left(-nD(\frac{\frac{\epsilon}{2}+\nu}{1+\nu}||\frac{\nu}{1+\nu})\right). \tag{A.21}$$

Consequently,

$$\begin{aligned}
\Pr\left\{\frac{1}{n}\Delta_F(Y^n, x^n) \geq \epsilon\right\} &= \Pr\left\{\frac{1}{n}\left[\sum_{t=1}^{n} k(x_t, Y_t)F_t(Y_{t-1}) + \sum_{t=1}^{n} l(x_t, Y_t)(1 - F_t(Y_{t-1}))\right] \geq \epsilon\right\} \\
&\leq \Pr\left\{\frac{1}{n}\sum_{t=1}^{n} k(x_t, Y_t)F_t(Y_{t-1}) \geq \epsilon/2\right\} + \qquad (A.22) \\
&\quad \Pr\left\{\frac{1}{n}\sum_{t=1}^{n} l(x_t, Y_t)(1 - F_t(Y_{t-1})) \geq \epsilon/2\right\} \\
&\leq 2\exp\left(-nD(\frac{\frac{\epsilon}{2}+\nu}{1+\nu}||\frac{\nu}{1+\nu})\right).
\end{aligned}$$

Inequality (A.18) is obtained similarly. □

Appendix B

Proof of Lemma 16: Fix $\mathbf{x}, \tilde{\mathbf{y}}, \hat{\mathbf{y}} \in \{0,1\}^\infty$ satisfying the hypothesis. It will suffice to establish the fact that for any $k \in \mathbb{N}$, $a, b \in \{0,1\}^k$ and $\mathbf{y}, \mathbf{x} \in \{0,1\}^\infty$ we have

$$\pi(b\mathbf{x}|a\mathbf{y}) = \pi(\mathbf{x}|\mathbf{y}), \tag{B.1}$$

where, e.g., $a\mathbf{y}$ denotes the usual concatenation. To see this, note that the hypothesis of the lemma implies the existence of m, $A, B, C, D \in \{0,1\}^m$ and $\mathbf{y}_{\mathrm{tr}}, \mathbf{x}_{\mathrm{tr}} \in \{0,1\}^\infty$ such that

$$\pi(\mathbf{x}|\tilde{\mathbf{y}}) = \pi(B\mathbf{x}_{\mathrm{tr}}|A\mathbf{y}_{\mathrm{tr}}) \tag{B.2}$$

and

$$\pi(\mathbf{x}|\hat{\mathbf{y}}) = \pi(D\mathbf{x}_{\mathrm{tr}}|C\mathbf{y}_{\mathrm{tr}}). \tag{B.3}$$

Combining (B.2) and (B.3) with (B.1) will clearly lead to the desired conclusion. So to establish (B.1) we now proceed to show

(I)

$$\pi(b\mathbf{x}|a\mathbf{y}) \geq \pi(\mathbf{x}|\mathbf{y}) \tag{B.4}$$

and

(II)

$$\pi(b\mathbf{x}|a\mathbf{y}) \le \pi(\mathbf{x}|\mathbf{y}). \tag{B.5}$$

Proof of (I): For fixed n and S, let g be an achieving next-state function in the definition of $\pi_S(bx_1^n|ay_1^n)$, i.e.,

$$\pi_g(bx_1^n|ay_1^n) = \pi_S(bx_1^n|ay_1^n), \tag{B.6}$$

and assume, without loss of generality, that the initial state for g in (B.6) is $s_1 = 1$. Then

$$\begin{aligned}
n\pi_S(x_1^n|y_1^n) &\stackrel{\text{def}}{=} n \min_{g' \in G_S} \pi_{g'}(x_1^n|y_1^n) \\
&\le n\pi_g(g(1,a); x_1^n|y_1^n) \\
&\le (n+k)\pi_g(bx_1^n|ay_1^n) && \text{(B.7)} \\
&= (n+k)\pi_S(bx_1^n|ay_1^n), && \text{(B.8)}
\end{aligned}$$

where $\pi_g(s; x_1^n|y_1^n)$ denotes the minimum fraction of prediction errors made by an FSM with next-state function g and initial state s, and $g(1,a)$ denotes the state of the machine with initial state 1, at the end of the block a. Inequality B.7 holds since the loss of g on the first k bits of bx_1^n (namely, on b) is clearly non-negative. Since n was arbitrary, we obtain

$$\begin{aligned}
\pi_S(\mathbf{x}|\mathbf{y}) &= \limsup_{n\to\infty} \pi_S(x_1^n|y_1^n) \\
&\le \limsup_{n\to\infty} \frac{n+k}{n} \pi_S(bx_1^n|ay_1^n) \\
&= \limsup_{n\to\infty} \pi_S(bx_1^n|ay_1^n) \\
&= \pi_S(b\mathbf{x}|a\mathbf{y}). && \text{(B.9)}
\end{aligned}$$

Letting now $S \to \infty$ at both ends of the above chain finally gives

$$\begin{aligned}
\pi(\mathbf{x}|\mathbf{y}) &= \lim_{S\to\infty} \pi_S(\mathbf{x}|\mathbf{y}) \\
&\le \lim_{S\to\infty} \pi_S(b\mathbf{x}|a\mathbf{y}) \\
&= \pi(b\mathbf{x}|a\mathbf{y}). && \text{(B.10)}
\end{aligned}$$

Proof of (II): Fix n and S and let g be an achieving next-state function in the definition of $\pi_S(x_1^n|y_1^n)$, i.e.,

$$\pi_g(x_1^n|y_1^n) = \pi_S(x_1^n|y_1^n). \tag{B.11}$$

We then have

$$
\begin{aligned}
(n+k)\pi_{S+k}(bx_1^n|ay_1^n) &\stackrel{\text{def}}{=} (n+k)\min_{g'\in G_{S+k}}\pi_{g'}(bx_1^n|ay_1^n) \\
&\le (n+k)\pi_{\tilde{g}}(bx_1^n|ay_1^n) &\text{(B.12)}\\
&\le k+n\pi_g(x_1^n|y_1^n) &\text{(B.13)}\\
&= k+n\pi_S(x_1^n|y_1^n), &\text{(B.14)}
\end{aligned}
$$

where, for concreteness, let $\tilde{g}$ be the next-state function which gives $s_t \equiv 1$ for the first k bits, and then coincides with g of (B.11). Clearly, such a $\tilde{g}$ is implementable with no more than $S+k$ states and therefore $\tilde{g} \in G_{S+k}$. Inequality (B.13) holds as, clearly, the best FSM fit for $\tilde{g}$ in (B.12), will suffer (non-normalized) loss no greater than k on the first k bits. Consequently, we have

$$
\begin{aligned}
\pi_{S+k}(b\mathbf{x}|a\mathbf{y}) &= \limsup_{n\to\infty}\pi_{S+k}(bx_1^n|ay_1^n) &\text{(B.15)}\\
&\le \limsup_{n\to\infty}\left\{\frac{k}{n+k}+\frac{n}{n+k}\pi_S(x_1^n|y_1^n)\right\} &\text{(B.16)}\\
&\le 0+1\cdot\limsup_{n\to\infty}\pi_S(x_1^n|y_1^n)\\
&= \pi_S(\mathbf{x}|\mathbf{y}), &\text{(B.17)}
\end{aligned}
$$

where (B.16) follows from the preceding chain. Taking $S \to \infty$ in (B.15) and (B.17) completes the proof of (II). $\square$

Appendix C

For convenience, we start by quoting a (rather weak) version of a strong law of large numbers for bounded martingale differences which will be used in the proofs of equations (87) and (88).

Theorem 19 ([13], Theorem 2.19) *Let $\{A_t, t \ge 1\}$ be a sequence of random variables and $\{\mathcal{F}_t, t \ge 1\}$ an increasing sequence of σ-fields with A_t measurable with respect to $\mathcal{F}_t$ for each t. Assume further that for each t $|A_t| \le c$ almost surely for some constant c. Then almost surely*

$$\lim_{m\to\infty}\frac{1}{m}\sum_{t=1}^{m}[A_t - E(A_t|\mathcal{F}_{t-1})] = 0. \tag{C.1}$$

Actually, if to be more loyal to the origin, Theorem 2.19 in [13] is a rather stronger and more general statement than that given in Theorem 19 above. The statement of Theorem 19, however, will suffice for our needs in the proofs that follow, and is presented in this limited generality for simplicity. Note that, in particular, when $\{\delta_t, \mathcal{F}_t\}$ is any bounded martingale difference sequence (namely, δ_t measurable with respect to $\mathcal{F}_t$ and $E(\delta_{t+1}|\mathcal{F}_t) = 0$ for each t), Theorem 19 gives us almost surely

$$\lim_{m\to\infty} \frac{1}{m}\sum_{t=1}^{m} \delta_t = 0. \tag{C.2}$$

Proof of equation (87): Let F^0 be the predictor which constantly predicts 0 at all times, regardless of past input. Since such a predictor is trivially implementable by an FSM (with one state), we clearly have $\pi(\tilde{\mathbf{X}}|\tilde{\mathbf{Y}}) \leq L_{F^0}(\tilde{\mathbf{Y}}, \tilde{\mathbf{X}})$ $P^x \times P^n$-almost surely. But, letting $O_m \stackrel{\text{def}}{=} \{t : 1 \leq t \leq m,\ t \text{ odd}\}$ and $E_m \stackrel{\text{def}}{=} \{t : 2 \leq t \leq m,\ t \text{ even}\}$, we have $P^x \times P^n$-almost surely

$$\begin{aligned}
L_{F^0}(\tilde{\mathbf{Y}}, \tilde{\mathbf{X}}) &= \limsup_{m\to\infty} \frac{1}{m}\sum_{t=1}^{m} |F_t^0(\tilde{Y}_1^{t-1}) - \tilde{X}_t| \\
&= \limsup_{m\to\infty} \frac{1}{m}\sum_{t=1}^{m} |0 - \tilde{X}_t| \\
&= \limsup_{m\to\infty} \frac{1}{m}\left[\sum_{O_m} \tilde{X}_t + \sum_{E_m} \tilde{X}_t\right] \\
&= \limsup_{m\to\infty} \frac{1}{m}\left[\sum_{t=1}^{\lceil m/2\rceil} X_t + \sum_{E_m} 0\right] \\
&= \frac{1}{2}\cdot\frac{1}{2},
\end{aligned} \tag{C.3}$$

where the last equality follows from the strong law of large numbers and those preceding it by the definition of F^0 and the construction of $\tilde{\mathbf{X}}$. Consequently, we have

$$\pi(\tilde{\mathbf{X}}|\tilde{\mathbf{Y}}) \leq \frac{1}{4} \quad P^x \times P^n \ -\text{a.s.} \tag{C.4}$$

Turning to establish the reverse inequality, we note that Theorem 7 implies the existence of a predictor F (we shall use here F instead of P so as not to confuse with the notation for probability measures) for which $L_F(\tilde{\mathbf{Y}}, \tilde{\mathbf{X}}) \leq \pi(\tilde{\mathbf{X}}|\tilde{\mathbf{Y}})$ $P^x \times P^n$- almost surely. Consequently, it will suffice to show that, for *any* predictor F, we have

$$L_F(\tilde{\mathbf{Y}}, \tilde{\mathbf{X}}) \geq \frac{1}{4} \quad P^x \times P^n \ -\text{a.s.} \tag{C.5}$$

To this end, fix any predictor F and for $t \in \mathbb{N}$ let $\delta_t = |F_{2t-1}(\tilde{Y}_1^{2t-2}) - \tilde{X}_{2t-1}| - 1/2$ and $\tilde{\mathcal{F}}_t = \sigma(\tilde{X}_1^{2t-1}, \tilde{Y}_1^{2t-1})$. Clearly, for each t, δ_t is measurable w.r.t. $\tilde{\mathcal{F}}_t$. Furthermore, we have

$P^x \times P^n$- almost surely

$$
\begin{aligned}
E(\delta_{t+1}|\tilde{\mathcal{F}}_t) &= E\{|F_{2t+1}(\tilde{Y}_1^{2t}) - \tilde{X}_{2t+1}| - 1/2|\tilde{\mathcal{F}}_t\} \\
&= E\left\{E\{|F_{2t+1}(\tilde{Y}_1^{2t}) - \tilde{X}_{2t+1}| - 1/2|\sigma(\tilde{X}_1^{2t-1}, \tilde{Y}_1^{2t})\}|\tilde{\mathcal{F}}_t\right\} \qquad \text{(C.6)} \\
&= E\left\{0|\tilde{\mathcal{F}}_t\right\} \qquad \text{(C.7)} \\
&= 0, \qquad \text{(C.8)}
\end{aligned}
$$

where (C.6) follows since clearly $\tilde{\mathcal{F}}_t \subset \sigma(\tilde{X}_1^{2t-1}, \tilde{Y}_1^{2t})$. Equality (C.7) follows since, conditioned on $\sigma(\tilde{X}_1^{2t-1}, \tilde{Y}_1^{2t})$, $F_{2t+1}(\tilde{Y}_1^{2t})$ is a constant and $\tilde{X}_{2t+1}$ is Bernoulli(1/2) (as, clearly, $\tilde{X}_{2t+1}$ is independent of $\sigma(\tilde{X}_1^{2t-1}, \tilde{Y}_1^{2t})$). Therefore, the inner expectation in (C.6) gives $|F_{2t+1}(\tilde{Y}_1^{2t}) - 0|\frac{1}{2} + |F_{2t+1}(\tilde{Y}_1^{2t}) - 1|\frac{1}{2} - \frac{1}{2} = 0$. Thus, we have established $\{\delta_t, \tilde{\mathcal{F}}_t\}_{t\geq 1}$ as a (clearly bounded) martingale difference sequence. Equation C.2 now assures us that

$$
\lim_{m\to\infty} \frac{1}{m} \sum_{t=1}^{m} \delta_t = 0 \quad P^x \times P^n \ \text{– a.s.}, \qquad \text{(C.9)}
$$

which, plugging in the specific form of the δ_t's for our case, translates into

$$
\lim_{m\to\infty} \frac{1}{m} \sum_{t=1}^{m} |F_{2t-1}(\tilde{Y}_1^{2t-2}) - \tilde{X}_{2t-1}| = \frac{1}{2} \quad P^x \times P^n \ \text{– a.s.} \qquad \text{(C.10)}
$$

Finally, we have $P^x \times P^n$-almost surely

$$
\begin{aligned}
L_F(\tilde{\mathbf{Y}}, \tilde{\mathbf{X}}) &= \limsup_{m\to\infty} \frac{1}{m} \sum_{t=1}^{m} |F_t(\tilde{Y}_1^{t-1}) - \tilde{X}_t| \\
&= \limsup_{m\to\infty} \frac{1}{m} \left[\sum_{O_m} |F_t(\tilde{Y}_1^{t-1}) - \tilde{X}_t| \right. \\
&\quad \left. + \sum_{E_m} |F_t(\tilde{Y}_1^{t-1}) - \tilde{X}_t| \right] \\
&\geq \limsup_{m\to\infty} \frac{1}{m} \sum_{O_m} |F_t(\tilde{Y}_1^{t-1}) - \tilde{X}_t| \\
&= \limsup_{m\to\infty} \frac{1}{m} \sum_{t=1}^{\lceil m/2 \rceil} |F_{2t-1}(\tilde{Y}_1^{2t-2}) - \tilde{X}_{2t-1}| \\
&= \frac{1}{4}, \qquad \text{(C.11)}
\end{aligned}
$$

where the last equality follows from (C.10). This establishes (C.5), which, in turn, establishes the reverse inequality to (C.4) and completes the proof. □

Proof of equation (88): Let here F^0 denote the predictor which, at all odd times predicts 0, regardless of past input and at all even times predicts the same value as the previous noisy input. Since such a predictor is implementable by an FSM (with three states), we clearly have $\pi(\hat{\mathbf{X}}|\hat{\mathbf{Y}}) \leq L_{F^0}(\hat{\mathbf{Y}}, \hat{\mathbf{X}})$ $P^x \times P^n$-almost surely. But we have $P^x \times P^n$-almost surely

$$\begin{aligned}
L_{F^0}(\hat{\mathbf{Y}}, \hat{\mathbf{X}}) &= \limsup_{m\to\infty} \frac{1}{m} \sum_{t=1}^{m} |F_t^0(\hat{Y}_1^{t-1}) - \hat{X}_t| \\
&= \limsup_{m\to\infty} \frac{1}{m} \left[\sum_{O_m} \hat{X}_t + \sum_{E_m} |\hat{Y}_{t-1} - \hat{X}_t| \right] \\
&= \limsup_{m\to\infty} \frac{1}{m} \left[\sum_{O_m} \hat{X}_t + \sum_{E_m} |\hat{Y}_{t-1} - \hat{X}_{t-1}| \right] \\
&= \limsup_{m\to\infty} \frac{1}{m} \left[\sum_{O_m} \hat{X}_t + \sum_{E_m} N_{t-1} \right] \\
&= \limsup_{m\to\infty} \frac{1}{m} \left[\sum_{t=1}^{\lceil m/2 \rceil} X_t + \sum_{t=1}^{\lfloor m/2 \rfloor} N_{2t-1} \right] \\
&= \frac{1}{2} \cdot \frac{1}{2} + \frac{1}{2} \cdot p, \qquad (C.12)
\end{aligned}$$

where the last equality follows from the strong law of large numbers and those preceding it by the definition of F^0 and the construction of $\hat{\mathbf{X}}$. Consequently, we have

$$\pi(\hat{\mathbf{X}}|\hat{\mathbf{Y}}) \leq \frac{1}{4} + \frac{p}{2} \quad P^x \times P^n \ - \text{a.s.} \qquad (C.13)$$

Turning to establish the reverse inequality, we note that it follows from the same line of argumentation as in the proof of equation (87) above, that it will suffice to show that for *any* predictor F, we have

$$L_F(\hat{\mathbf{Y}}, \hat{\mathbf{X}}) \geq \frac{1}{4} + \frac{p}{2} \quad P^x \times P^n \ - \text{a.s.} \qquad (C.14)$$

To this end, we note that for any predictor F

$$\begin{aligned}
L_F(\hat{\mathbf{Y}}, \hat{\mathbf{X}}) &= \limsup_{m\to\infty} \frac{1}{m} \sum_{t=1}^{m} |F_t(\hat{Y}_1^{t-1}) - \hat{X}_t| \\
&= \limsup_{m\to\infty} \frac{1}{m} \left[\sum_{O_m} |F_t(\hat{Y}_1^{t-1}) - \hat{X}_t| \right. \\
&\quad \left. + \sum_{E_m} |F_t(\hat{Y}_1^{t-1}) - \hat{X}_t| \right] \\
&= \limsup_{m\to\infty} \frac{1}{m} \left[\sum_{t=1}^{\lceil m/2 \rceil} |F_{2t-1}(\hat{Y}_1^{2t-2}) - \hat{X}_{2t-1}| \right.
\end{aligned}$$

$$+\sum_{t=1}^{\lfloor m/2\rfloor}|F_{2t}(\hat{Y}_1^{2t})-\hat{X}_{2t}|\Bigg]$$

$$\geq \liminf_{m\to\infty}\frac{1}{m}\left[\sum_{t=1}^{\lceil m/2\rceil}|F_{2t-1}(\hat{Y}_1^{2t-2})-\hat{X}_{2t-1}|+\sum_{t=1}^{\lfloor m/2\rfloor}|F_{2t}(\hat{Y}_1^{2t})-\hat{X}_{2t}|\right]$$

$$\geq \liminf_{m\to\infty}\frac{1}{m}\sum_{t=1}^{\lceil m/2\rceil}|F_{2t-1}(\hat{Y}_1^{2t-2})-\hat{X}_{2t-1}|$$

$$+\liminf_{m\to\infty}\frac{1}{m}\sum_{t=1}^{\lfloor m/2\rfloor}|F_{2t}(\hat{Y}_1^{2t-1})-\hat{X}_{2t}|. \tag{C.15}$$

For the first term in (C.15), letting $\delta_t=|F_{2t-1}(\hat{Y}_1^{2t-2})-\hat{X}_{2t-1}|-1/2$ and $\hat{\mathcal{F}}_t=\sigma(\hat{X}_1^{2t-1},\hat{Y}_1^{2t-1})$, $\{\delta_t,\hat{\mathcal{F}}_t\}_{t\geq 1}$ can be shown to be a bounded martingale difference sequence very similarly to the way $\{\delta_t,\tilde{\mathcal{F}}_t\}_{t\geq 1}$ was shown to be one in the proof of equation (87) above. This, in turn, leads to

$$\lim_{m\to\infty}\frac{1}{m}\sum_{t=1}^{m}|F_{2t-1}(\hat{Y}_1^{2t-2})-\hat{X}_{2t-1}|=\frac{1}{2}\quad P^x\times P^n\ \ -\text{a.s.}, \tag{C.16}$$

similarly as (C.10) was obtained in the proof above. This finally establishes the fact that the first term in (C.15) equals 1/4 $P^x\times P^n$-almost surely. To handle the second term in (C.15) we let $A_t=|F_{2t}(\hat{Y}_1^{2t-1})-\hat{X}_{2t}|$ and $\mathcal{F}_t=\sigma(\hat{Y}^{2t+1},\hat{X}^{2t})$. Since A_t is clearly bounded and measurable with respect to $\mathcal{F}_t$, Theorem 19 gives us $P^x\times P^n$-almost surely

$$\begin{aligned}
&\lim_{m\to\infty}\frac{1}{m}\sum_{t=1}^{m}\Big[|F_{2t}(\hat{Y}_1^{2t-1})-\hat{X}_{2t}|-E[|F_{2t}(\hat{Y}_1^{2t-1})-\hat{X}_{2t}||\mathcal{F}_{t-1}]\Big]\\
=\;&\lim_{m\to\infty}\frac{1}{m}\sum_{t=1}^{m}[A_t-E(A_t|\mathcal{F}_{t-1})]\\
=\;&0.
\end{aligned} \tag{C.17}$$

We can now write $P^x\times P^n$-almost surely

$$\begin{aligned}
&\liminf_{m\to\infty}\frac{1}{m}\sum_{t=1}^{\lfloor m/2\rfloor}|F_{2t}(\hat{Y}_1^{2t-1})-\hat{X}_{2t}|\\
=\;&\liminf_{m\to\infty}\frac{1}{m}\sum_{t=1}^{\lfloor m/2\rfloor}E[|F_{2t}(\hat{Y}_1^{2t-1})-\hat{X}_{2t}||\mathcal{F}_{t-1}]\\
=\;&\liminf_{m\to\infty}\frac{1}{m}\sum_{t=1}^{\lfloor m/2\rfloor}\Big\{|F_{2t}(\hat{Y}_1^{2t-1})-0|\cdot\Pr\{\hat{X}_{2t}=0|\mathcal{F}_{t-1}\}\\
&+|F_{2t}(\hat{Y}_1^{2t-1})-1|\cdot\Pr\{\hat{X}_{2t}=1|\mathcal{F}_{t-1}\}\Big\}\\
=\;&\liminf_{m\to\infty}\frac{1}{m}\sum_{t=1}^{\lfloor m/2\rfloor}\Big\{F_{2t}(\hat{Y}_1^{2t-1})\cdot\Pr\{\hat{X}_{2t}=0|\mathcal{F}_{t-1}\}
\end{aligned}$$

$$
\begin{aligned}
& +(1 - F_{2t}(\hat{Y}_1^{2t-1})) \cdot (1 - \Pr\{\hat{X}_{2t} = 0 | \mathcal{F}_{t-1}\})\Big\} \\
\geq \; & \liminf_{m\to\infty} \frac{1}{m} \sum_{t=1}^{\lfloor m/2 \rfloor} \min\Big\{\Pr\{\hat{X}_{2t} = 0|\mathcal{F}_{t-1}\}, 1 - \Pr\{\hat{X}_{2t} = 0|\mathcal{F}_{t-1}\}\Big\} \\
= \; & \liminf_{m\to\infty} \frac{1}{m} \sum_{t=1}^{\lfloor m/2 \rfloor} p \qquad \text{(C.18)} \\
= \; & p/2. \qquad \text{(C.19)}
\end{aligned}
$$

The first equality follows from (C.17). The second equality follows since, conditioned on $\mathcal{F}_{t-1}$, $F_{2t}(\hat{Y}_1^{2t-1})$ is almost surely a constant. The inequality follows by the fact that a convex combination of any two reals is lower bounded by their minimum. To see why equality (C.18) holds, observe that

$$
\begin{aligned}
\Pr\{\hat{X}_{2t} = 0|\mathcal{F}_{t-1}\} &= \Pr\{\hat{X}_{2t} = 0|\sigma(\hat{Y}^{2t-1}, \hat{X}^{2t-2})\} \\
&= (1-p)1_{\{\hat{Y}_{2t-1}=0\}} + p1_{\{\hat{Y}_{2t-1}=1\}}, \qquad \text{(C.20)}
\end{aligned}
$$

where (C.20) holds since it clearly follows from the construction of $\hat{\mathbf{X}}$ that the distribution of $\hat{X}_{2t}$, conditioned on $\sigma(\hat{Y}^{2t-1}, \hat{X}^{2t-2})$, is Bernoulli$(1-p)$ when $\hat{Y}_{2t-1} = 0$ and Bernoulli(p) when $\hat{Y}_{2t-1} = 1$. Therefore, $P^x \times P^n$-almost surely

$$
\begin{aligned}
& \min\Big\{\Pr\{\hat{X}_{2t} = 0|\mathcal{F}_{t-1}\}, 1 - \Pr\{\hat{X}_{2t} = 0|\mathcal{F}_{t-1}\}\Big\} \\
= \; & 1_{\{\hat{Y}_{2t-1}=0\}} \min\{1-p, p\} + 1_{\{\hat{Y}_{2t-1}=1\}} \min\{p, 1-p\} \\
= \; & p, \qquad \text{(C.21)}
\end{aligned}
$$

justifying (C.18) and, consequently, establishing the fact that for any predictor

$$
\liminf_{m\to\infty} \frac{1}{m} \sum_{t=1}^{\lfloor m/2 \rfloor} |F_{2t}(\hat{Y}_1^{2t-1}) - \hat{X}_{2t}| \geq \frac{p}{2} \quad P^x \times P^n \; - \text{a.s.} \qquad \text{(C.22)}
$$

This joins equation (C.16) in establishing the fact that (C.15) is lower bounded $P^x \times P^n$-almost surely by $1/4 + p/2$ which, in turn, implies inequality (C.14) and completes the proof. □

References

[1] A. Baruch, "Universal Algorithms for Sequential Decision in the Presence of Noisy Observations," M.Sc. dissertation, Dept. of EE, Technion - I.I.T., Haifa, Israel, February 1999.

[2] L. Breiman, *Probability* SIAM, Philadelphia, 1992.

[3] N. Cesa-Bianchi and G. Lugosi, "On Sequential Prediction of Individual Sequences Relative to a Set of Experts," accepted for publication in *Ann. Statist.* (A preliminary version appeared in the Proceedings of the 11th Annual ACM Conference on Computational Learning Theory. ACM Press, 1998).

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

[5] G. J. Chaitin, "A Theory of Program Size Formally Identical to Information Theory," *J. ACM,* vol. 22, pp. 329-340, 1975.

[6] T. M. Cover, "Universal Portfolios," *Mathematical Finance,* vol. 1(1), pp. 1-29, January 1991.

[7] T. M. Cover and A. Shenhar, "Compound Bayes Predictors for Sequences with Apparent Markov Structure," *IEEE Trans. Syst. Man Cybern.*, vol. SMC-7, pp. 421-424, May/June 1977.

[8] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications* 2nd ed. Springer-Verlag, New York, 1998.

[9] R. Durret, *Probability: Theory and Examples.* Belmont, California: Duxbury Press, 1991.

[10] M. Feder, "Gambling Using a Finite-State Machine," *IEEE Trans. Inform. Theory*, vol. 37, pp. 1459-1465, Sept. 1991.

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

[12] R. G. Gallager, "Source Coding with Side Information and Universal Coding," unpublished manuscript, Sept. 1976.

[13] P. Hall and C.C. Heyde, *Martingale Limit Theory and its Application* Academic Press, New York, 1980.

[14] J. F. Hannan and H. Robbins, "Asymptotic Solutions of the Compound Decision Problem for two completely specified distributions," *Ann. Math. Statist.* 26 37-51.

[15] D. Haussler, J. Kivinen, and M.K. Warmuth, "Sequential Prediction of Individual Sequences Under General Loss Functions," *IEEE Trans. Inform. Theory*, vol. 44, pp. 1906-1925, September 1998.

[16] A. N. Kolmogorov, "Three Approaches to the Quantitative Definition of Information," *Probl. Inform. Transm.*, vol. 1, pp. 4-7, 1965.

[17] N. Merhav and M. Feder, "Universal Prediction," *IEEE Trans. Inform. Theory,* vol. IT-44, no. 6, pp. 2124-2147, October 1998.

[18] N. Merhav, "Universal Coding with Minimum Probability of Codeword Length Overflow," *IEEE Trans. Inform. Theory*, vol. 37, pp. 556-563, May 1991.

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

[20] Y. Nogami, "The k-extended Set-Compound Estimation Problem in Nonregular Family of Distributions," *Ann. Inst. Stat. Math.*, vol. 31A, pp. 169-176, 1979.

[21] M. Opper and D. Haussler, "Worst Case Prediction Over Sequences Under Log Loss," *The Mathematics of Information Coding, Extraction, and Distribution* George Cybenko, Dianne P. O'Leary, Jorma Rissanen Editors, pp. 81-90, Springer, 1996.

[22] E. Rio, "The Functional Law of the Iterated Logarithm for Stationary Strongly Mixing Sequences," *Ann. Math. Statist.* 33 pp. 659-680 (1995).

[23] J. Rissanen, "Universal Coding, Information, Prediction, and Estimation," *IEEE Trans. Inform. Theory*, vol. IT-30, no. 4, pp. 629-636, July 1984.

[24] H. Robbins, "Asymptotically Subminimax Solutions of Compound Statistical Decision Problems," *Proc. Second Berkeley Symp. Math. Statis. Prob.*, 131-148. Univ. of California Press.

[25] H. Robbins and E. Samuel, "Testing Statistical Hypotheses - The "Compound" Approach," *Recent Developments in Information and Decision Procedures,* ed. Machol and Gray, Macmillan NY 1962.

[26] B. Y. Ryabko, "Encoding a Source with Unknown but Ordered Probabilities," *Probl. Inform. Transmission,* pp. 134-138, Oct. 1979.

[27] J. van Ryzin, "The Sequential Compound Decision Problem with $m \times n$ Finite Loss Matrix," *Ann. Math. Statist.*, vol. 37, pp. 954-975, 1966.

[28] E. Samuel, "An Empirical Bayes Approach to the Testing of Certain Parametric Hypotheses," *Ann. Math. Statist.*, vol. 34, no. 4, pp. 1370-1385, 1963.

[29] E. Samuel, "Asymptotic Solutions of the Sequential Compound Decision Problem," *Ann. Math. Statist.*, pp. 1079-1095, 1963.

[30] E. Samuel, "Convergence of the Losses of Certain Decision Rules for the Sequential Compound Decision Problem," *Ann. Math. Statist.*, pp. 1606-1621, 1964.

[31] E. Samuel, "On Simple Rules for the Compound Decision Problem," *J. Royal Stat. Society*, series B-27, pp. 238-244, 1965.

[32] E. Samuel, "Note on a Sequential Classification Problem," *Ann. Math. Statist.*, vol. 34, no. 3, pp. 1095-1097, 1963.

[33] E. Samuel, "Sequential Compound Estimators," *Ann. Math. Statist.*, vol. 36, no. 3, pp. 879-889, 1965.

[34] E. Samuel, "The Compound Decision Problem in the Opponent Case," *Israel Journal of Mathematics*, vol. 3, no. 3, Sept. 1965.

[35] R. J. Solomonoff, "A Formal Theory of Inductive Inference, pts. 1 and 2," *Inform. Contr.*, vol. 7, pp. 1-22 and pp. 224-254, 1964.

[36] W.F. Stout, "A Martingale Analogue of Kolmogorov's Law of the Iterated Logarithm," *Z. Wahrsch. Verw. Gebiete* 15, pp. 279-290, 1970.

[37] W.F. Stout, "The Hartman-Wintner Law of the Iterated Logarithm for Martingales," *Ann. Math. Statist.* 41, pp. 2158-2160, 1970.

[38] S. B. Vardeman, "Admissible Solutions of k-Extended Finite State Set and the Sequence Compound Decision Problems," *J. Multiv. Anal.*, vol. 10, pp. 426-441, 1980.

[39] V. Vovk, "A Game of Prediction with Expert Advice," *Proc. 8th Annu. Workshop on Computational Learning Theory.* pp. 51-60, New York NY, 1995.

[40] V. Vovk, "Aggregating Strategies," in *Proc. 3rd Annu. Workshop on Computational Learning Theory,* pp. 371-383, San Mateo, CA: Kaufmann, 1990.

[41] M.J. Weinberger, N. Merhav, and M. Feder, "Optimal Sequential Probability Assignment for Individual Sequences," *IEEE Trans. Inform. Theory*, vol. 40, pp. 384-396, March 1994.

[42] T. Weissman and N. Merhav, "On Prediction in the Presence of Noise," unpublished manuscript, 1999.

[43] T. Weissman and N. Merhav, "On Prediction of Individual Sequences Relative to a Set of Experts in the Presence of Noise," in *Proc. 12th Annu. Workshop on Computational Learning Theory,* pp. 19-28, New York: ACM, 1999.

[44] T. Weissman and N. Merhav, "Universal Prediction of Individual Binary Sequences in the Presence of Arbitrarily Varying, Memoryless Additive Noise," accepted for presentation in ISIT 2000.

[45] T. Weissman and N. Merhav, "Universal Prediction of Individual Binary Sequences in the Presence of Noise," Submitted to *IEEE Trans. Inform. Theory*, Nov. 1999.

[46] A. D. Wyner and J. Ziv, "A Theorem on the Entropy of Certain Binary Sequences and Applications: Part I," *IEEE Trans. Inform. Theory*, vol. IT-19, no. 6, pp. 769-772, Nov. 1973.

[47] J. Ziv and A. Lempel, "Compression of Individual Sequences Via Variable Rate Coding," *IEEE Trans. Inform. Theory*, vol. IT-24, pp. 530-536, Sept. 1978.