Universal Prediction of Individual Binary Sequences in the Presence of Noise *

Tsachy Weissman[†] Neri Merhav[‡]

November 4, 1999

Abstract

The problem of predicting the next outcome of an individual binary sequence, based on noisy observations of the past, is considered. The goal of the predictor is to perform, for each individual sequence, "almost" as well as the best in a set of experts, where performance is evaluated using a general loss function. A comprehensive approach to prediction in this noisy setting is presented and proven generally efficient under appropriate conditions. As an illustration of the applicability of the approach suggested for concrete situations, two important special cases are explicitly treated. The first is the case where the data-corrupting noise process is binary-valued (where the observed bit is the bitwise XOR of the clean bit and the noise bit). The second case is that of real-valued additive noise. It is shown that even in this more challenging situation, where the information available to the predictor regarding the past sequence is incomplete, a predictor can be guaranteed to successfully compete with a whole set of experts in considerably strong senses.

1 Introduction

Consider the problem of sequentially predicting an arbitrary binary sequence. At each time unit $t \geq 1$, after making a prediction P_t based on the previously-observed $x_1, x_2, \ldots, x_{t-1}$, one observes the t-th bit of the sequence, x_t. The loss at time t is given by a fixed loss function $L(x_t, P_t)$. The goal is to predict *any* sequence essentially as well as the best "expert" in a given set of experts in the sense of minimizing the cumulative loss $\sum_{t=1}^{n} L(x_t, P_t)$.

In this on-line prediction setting, the sequence of outcomes is assumed to be an individual, deterministic sequence and no assumptions are made on its structure or its statistical properties. The performance of the prediction scheme is judged in the worst case over all possible binary sequences and the performance of the predictor is compared to that of the

*This research, which is supported by the Israeli Science Foundation, is part of the D.Sc. dissertation of the first author. Both authors are with the Department of Electrical Engineering, Technion- Israel Institute of Technology, Haifa 32000, Israel.

[†]tsachy@tx.technion.ac.il.

[‡]merhav@ee.technion.ac.il.

best expert in the given set of experts $\mathcal{F}$, referred to as the *comparison class.* This setting is well connected to practical situations in which the sequence-generating mechanism is unknown, yet one is presented with a whole set of prediction models (experts), which may represent a set of different approaches or "beliefs" regarding the "right" prediction scheme for the particular problem. In other cases of practical interest, the experts in the comparison class may represent a set of prediction schemes which require a limited amount of memory, or algorithmic and computational resources, such as the class of finite-state machines of a certain order (cf. e.g. [39]). The problems that can be naturally formulated within this setting are innumerable, ranging from universal coding, gambling and portfolio selection, through on-line learning and artificial intelligence, to everyday earthly matters such as weather prediction. Accordingly, this problem has been studied by many researchers from various disciplines, such as information theory, learning theory, game theory, statistics, economics, control theory, and operations research.

To be more explicit, for every positive integer n and every possible binary sequence $x_1, \cdots, x_n$, we look at the excess total loss of the predictor P beyond that of the best expert. This excess total loss is referred to as *redundancy* (in the context of source coding, where the loss function is the logarithmic loss), *relative loss* or *regret.* Clearly, as the predictor must make its predictions on-line, it cannot know in advance which one of the experts in $\mathcal{F}$ will do best on the particular sequence. It turns out, however, that in many cases, the predictor can achieve remarkably small regret, no matter what the underlying individual sequence may be. In other words, the predictor can perform almost as well as if it had known ahead of time which expert will perform best. The degree to which such performance can be guaranteed and the meaning of "almost as well" have been, and are still being, studied by many researchers from different disciplines. The literature pertaining to this setting, under a variety of loss functions and expert classes, is far too plentiful to be specified briefly. Representative examples can be found in [2], [3], [10], [39], [14], [38], [4], 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.

One of the directions suggested for future research in [17], was an investigation of the case where only a noisy version of the underlying individual sequence is accessible. In accordance, the purpose of this paper is to revisit the above described setting of prediction of individual sequences relative to a set of experts, for the case where the predictor and experts access only a noisy version of the underlying individual sequence, yet the loss function remains in terms of the clean sequence. This setting is suitable to model situations where the underlying binary sequence is individual, with no probabilistic assumptions, yet it reaches the predictor after having passed through a noisy channel, whose characteristics may or may not be known to the predictor.

The setting considered here can also be regarded as a first step towards an alternative approach to the traditional setting of the filtering problem. In the traditional and well-established theory of linear and non-linear filtering, the clean discrete- or continuous-time signal is assumed to be governed by a given probabilistic law (e.g.: wide sense stationary signals with a given power spectral density, discrete-time random signals evolving from a given difference equation, continuous-time diffusion processes governed by a given stochastic differential equation, etc.) and one seeks an estimate, or a prediction, for the clean signal value, based on its noisy observation, which is assumed to be some given stochastic functional of the clean signal. In contrast, here we regard the clean signal as unknown but deterministic (individual). The set of experts may represent a natural class of prediction schemes, that are related to the particular problem at hand due to some assumptions on the clean sequence. For example, one may want to compete with the class of linear predictors. In other cases of interest, the class of experts may manifest practical, computational or algorithmic limitations imposed on a prediction scheme (e.g., the set of finite state machines up to a certain order). This revision is analogous to that made for the problem of prediction of noise-free time series. Historically, the original model considered by researchers was that where the sequence is assumed to be generated by some stochastic source belonging to a given family, and, in later research, this setting was (and is still being) revisited for the case where the sequence is regarded as deterministic. We have reason to believe that the new approach to universal filtering, induced by the setting considered in this work, will become a powerful and appealing alternative to the original filtering model. In particular, this setting can be used in many applications (e.g., learning in a noisy environment, portfolio selection based on inaccurate data, coding based on noisy observations, etc.) where the available information is incomplete.

To the best of our knowledge, the problem of prediction of individual sequences based on noisy observations has thus far been considered only in [1], [40] and [41], for the case where the individual sequence is corrupted by an i.i.d. binary noise process (and the observed bits are the modulo 2 sum of the clean bit and the noise bit), and in [42], [43] for the case of additive, real-valued noise.

The basic problem is the following: at each time instant $t \geq 1$, after making a prediction P_t based on $y_1, \ldots y_{t-1}$, one observes y_t, a *noisy version* of x_t, which is a random variable stochastically related to x_t. As before, the loss at time t is then given by $L(x_t, P_t)$. The goal of a predictor is, again, to predict *any* sequence $x_1, x_2, \ldots$ essentially as well as the best expert in a given set of experts, under criteria that will be made precise in the sequel. Does a predictor, fed by a noisy version of an individual sequence, have any hope of efficiently competing with a whole set of experts? The answer to this question turns out to be affirmative in some surprisingly strong meanings.

The philosophy underlying most of the existing prediction schemes for the noise-free deterministic setting, is that of predicting according to some combination of the experts' advice, by assigning weights according to past performance of each one of the experts. The problem with implementing this philosophy as is, in our case, is the lack of complete information on past performance of each expert (due to the noise). This difficulty can, however, be alleviated by a certain modification of the original approach.

We focus mostly on the case of finite comparison classes, though we give a few examples of using results of the finite expert class case to deal with classes of infinitely many experts. It should be emphasized that our results apply to *arbitrary* finite sets of experts without any assumptions on their structure or geometry. The class of loss functions considered is quite general and includes all standard loss functions arising in various applications with the exception of the logarithmic loss function. Future work is needed to determine whether the gap for the logarithmic loss can be bridged over.

The three main sections of this paper are as follows. Section 2 is dedicated to a presentation of our general approach to the prediction problem in this noisy setting and to a derivation of results assessing the merit of this approach. The succeeding two sections are dedicated to an illustration of how the general approach and results of Section 2 can be applied in concrete situations. The model of Section 3 addresses the important special case where the noise components are binary-valued, i.i.d., and the observations are the result of the bitwise XOR between the clean bit and the noise bit. In the second noise model, considered in Section 4, the noise components are zero-mean, real-valued, independent random variables with arbitrarily-varying distributions, and each observation is given by the sum of the clean bit (0 or 1) and the corresponding noise component. The approach to the exposition taken in the last two sections places clarity and simplicity before generality and strength. Limiting assumptions and hypotheses were imposed where these could help to avoid much work over non-essential technicalities and little attention was given to optimizing the constants in the bounds.

2 Main Results for the General Case

In this section, we present our general approach and results on the problem of prediction based on noisy observations. The general setting is the following. Let $L : \{0,1\} \times [0,1] \to [0,\infty]$ be a fixed *loss function*. In the noise-free setting, the cumulative loss of a predictor $F = \{F_t\}_{t\geq 1}$ (up to time n) on an individual binary sequence $x \in \{0,1\}^\infty$ is defined by

$$\begin{aligned} L_F(x^n) &\stackrel{\text{def}}{=} \sum_{t=1}^{n} L(x_t, F_t(x^{t-1})) && (1)\\ &= \sum_{t=1}^{n} (1-x_t)L_0(F_t(x^{t-1})) + x_t L_1(F_t(x^{t-1})), && (2) \end{aligned}$$

where $L_0(\cdot) \stackrel{\text{def}}{=} L(0,\cdot)$ and $L_1(\cdot) \stackrel{\text{def}}{=} L(1,\cdot)$, and $x^t = (x_1,\ldots,x_t)$ (first t bits of the individual sequence $x \in \{0,1\}^\infty$). Analogously, for the noisy setting, we define the cumulative loss of the predictor $F = \{F_t\}_{t\geq 1}$ (up to time n), fed by $y^n = (y_1,\ldots,y_n)$ and judged with respect to $x^n = (x_1,\ldots,x_n)$ by

$$\begin{aligned} L_F(y^n, x^n) &\stackrel{\text{def}}{=} \sum_{t=1}^{n} L(x_t, F_t(y^{t-1})) && (3)\\ &= \sum_{t=1}^{n} (1-x_t)L_0(F_t(y^{t-1})) + x_t L_1(F_t(y^{t-1})). && (4) \end{aligned}$$

We let $\{r_t\}_{t\geq 1}$ be the noise process, and $\mathcal{F}_n \stackrel{\text{def}}{=} \sigma(r^n)$ (the smallest σ-field with respect to which r^n is measurable) denote the history up to time n. The noisy sequence y takes its values in A^∞, where A can be any alphabet equipped with a σ-algebra, say $\mathcal{A}$. We further assume that the components of the noise process $\{r_t\}_{t\geq 1}$ are independent and that $y_t \in \sigma(r_t)$ for all $t \geq 1$ (that is, the t-th noise component affects the t-th observation only). A predictor (or an expert) F in this setting is a series of (measurable) functions $\{F_t : A^{t-1} \to [0,1]\}_{t\geq 1}$.

Suppose that there exists an unbiased estimator for x_t based on y_t, that is, a measurable function $h : A \to \mathbb{R}$ such that $Eh(y_t) = x_t$ for all x_t and all $t \geq 1$. We can then define the following estimator for the cumulative loss of any predictor F up to time n:

$$\hat{L}_F(y^n) \stackrel{\text{def}}{=} \sum_{t=1}^{n}(1-h(y_t))L_0(F_t(y^{t-1})) + h(y_t)L_1(F_t(y^{t-1})). \tag{5}$$

Suppose further that, for a given set of experts $\mathcal{F}$, one can find a predictor P for which a performance bound of the following type holds for all n and $y^n \in A^n$:

$$\hat{L}_P(y^n) - \inf_{F\in\mathcal{F}} \hat{L}_F(y^n) \le g_n(y^n) \tag{6}$$

for some sequence $\{g_n : A^n \to \mathbb{R}\}$. As will be seen in the examples of the next sections, $\hat{L}_F(y^n)$ can be identified as the noise-free loss of a predictor in the original cumulative setting, fed by y^n and judged w.r.t. y^n, under the loss function $L(y_t, F_t(y^{t-1})) = (1 - h(y_t))L_0(F_t(y^{t-1})) + h(y_t)L_1(F_t(y^{t-1}))$. In many cases, meaningful bounds in the form of equation (6) can be obtained by using performance bounds from the cumulative setting.

As previously mentioned, the basic idea is the following. We first present some results on $\hat{L}_F(y^n)$ defined in (5), and then use them to show that for any predictor that satisfies an inequality like (6), for some reasonably small $g_n(y^n)$, the real performance measure of interest, namely, the relative loss in terms of the real losses, has a meaningful upper bound as well.

For a given predictor F and $x^n \in \{0,1\}^n$, we let $\Delta_F(y^n, x^n) \stackrel{\text{def}}{=} L_F(y^n, x^n) - \hat{L}_F(y^n)$. As will be established in the Appendix, for any individual $x \in \{0,1\}^\infty$, $\{\Delta_F(y^n, x^n), \mathcal{F}_n\}_{n\ge1}$ is a martingale. Using results for martingales, the following three lemmas are proven in the Appendix. The first will be instrumental in assessing the expected loss as a performance criterion.

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

$$E\hat{L}_F(y^n) = EL_F(y^n, x^n). \tag{7}$$

In words, $\hat{L}_F(y^n)$ is an unbiased estimator for $EL_F(y^n, x^n)$.

The following lemma will be useful in establishing the almost-sure asymptotic performance of our predictors.

Lemma 2 *Suppose that L_0 and L_1 are upper bounded by $B > 0$ and that $|h(y_t)| \le C < \infty$ almost surely. Then for any predictor F*

$$\limsup_{n\to\infty} \frac{|\Delta_F(y^n, x^n)|}{\sqrt{n \log\log n}} \le 2\sqrt{2}C \cdot B \quad a.s. \tag{8}$$

The proof of Lemma 2 follows from an application of the upper bound in the martingale analogue of Kolmogorov's law of the iterated logarithm (cf. Theorem 1 of [32]).

The last lemma will be beneficial for characterizing large deviations performance.

Lemma 3 *Suppose that L_0 and L_1 are bounded by some $B > 0$. Suppose further that for a given constant β, $E(h(y_t) - x_t)^2 \leq \beta^2$ and that $E|h(y_t) - x_t|^k \leq \frac{k!}{2} \cdot E(h(y_t) - x_t)^2 \cdot c^{k-2}$ for all $k > 2$, $t \geq 1$ and some $0 < c < \infty$. Then for any predictor F, all $\varepsilon > 0$ and all $x^n \in \{0,1\}^n$*

$$\Pr\{\frac{1}{n}\Delta_F(y^n, x^n) \geq \varepsilon\} \leq 2\exp\{-n\frac{\varepsilon^2}{8B^2(\beta^2 + c\varepsilon/B)}\} \tag{9}$$

and similarly

$$\Pr\{\frac{1}{n}\Delta_F(y^n, x^n) \leq -\varepsilon\} \leq 2\exp\{-n\frac{\varepsilon^2}{8B^2(\beta^2 + c\varepsilon/B)}\}. \tag{10}$$

The main result that supports Lemma 3 is a new exponential inequality for martingales due to Peña [22, Theorem 1.2A], which is stronger than similar results of [43], that are based on a concentration inequality for martingales of bounded differences [6]. Lemma 3 is stronger than its counterparts in [43] in that it replaces the almost-sure boundedness assumption by the significantly milder moment condition, which means that the tail of the distribution of the estimation error would not be too heavy. The exponential decay rate on the right-hand sides of the inequalities is established in the bounds of [43] (and given in terms of the Kullback-Leibler divergence between two binary distributions) and can be shown to be faster in certain situations. This fact is not exploited at this point as we do not attempt here to characterize the best achievable exponential rate.

Using Lemmas 1-3, we can now prove the following performance bounds. Theorem 4 addresses the expected loss performance criterion.

Theorem 4 *Suppose that (6) holds for all n and $y^n \in A^n$ for a certain predictor P. Then for all n*

$$EL_P(y^n, x^n) - \inf_{F\in\mathcal{F}} EL_F(y^n, x^n) \leq Eg_n(y^n). \tag{11}$$

Proof:

$$\begin{aligned} EL_P(y^n, x^n) - \inf_{F\in\mathcal{F}} EL_F(y^n, x^n) &= E\hat{L}_P(y^n) - \inf_{F\in\mathcal{F}} E\hat{L}_F(y^n) && (12)\\ &\leq E\hat{L}_P(y^n) - E\inf_{F\in\mathcal{F}} \hat{L}_F(y^n) && (13)\\ &= E\{\hat{L}_P(y^n) - \inf_{F\in\mathcal{F}} \hat{L}_F(y^n)\} && (14)\\ &\leq Eg_n(y^n), && (15) \end{aligned}$$

where (12) follows from Lemma 1 and (15) follows from (6). □

For the asymptotic performance of a predictor satisfying (6), we have the following.

Theorem 5 *For a finite expert set $\mathcal{F}$, let P be a predictor that satisfies (6). Suppose that L_0 and L_1 are upper bounded by $B > 0$ and that $|h(y_t)| \leq C < \infty$ a.s. We then have a.s.*

$$\limsup_{n\to\infty} \frac{[L_P(y^n,x^n) - \inf_{F\in\mathcal{F}} L_F(y^n,x^n)]}{\sqrt{n\log\log n}} \leq 4\sqrt{2}C\cdot B + \limsup_{n\to\infty}\frac{g_n(y^n)}{\sqrt{n\log\log n}}. \qquad (16)$$

Note that this theorem is informative whenever $g_n(y^n) = O(\sqrt{n\log\log n})$. As will be seen in the sequel, we even have $g_n(y^n) = O(\sqrt{n})$ in many cases of interest.

Proof:

$$\begin{aligned}
L_P(y^n,x^n) - \min_{\mathcal{F}} L_F(y^n,x^n) &= L_P(y^n,x^n) - \hat{L_P}(y^n) + \\
&\quad \min_{\mathcal{F}} \hat{L_F}(y^n) - \min_{\mathcal{F}} L_F(y^n,x^n) + \\
&\quad \hat{L_P}(y^n) - \min_{\mathcal{F}} \hat{L_F}(y^n) && (17)\\
&\leq |L_P(y^n,x^n) - \hat{L_P}(y^n)| + && (18)\\
&\quad |\min_{\mathcal{F}} \hat{L_F}(y^n) - \min_{\mathcal{F}} L_F(y^n,x^n)| + \\
&\quad \hat{L_P}(y^n) - \min_{\mathcal{F}} \hat{L_F}(y^n) && (19)\\
&\leq |L_P(y^n,x^n) - \hat{L_P}(y^n)| + && (20)\\
&\quad \max_{\mathcal{F}} |\hat{L_F}(y^n) - L_F(y^n,x^n)| + \\
&\quad \hat{L_P}(y^n) - \min_{\mathcal{F}} \hat{L_F}(y^n) && (21)\\
&= |\Delta_P(y^n,x^n)| + \\
&\quad \max_{\mathcal{F}} |\Delta_F(y^n,x^n)| + \hat{L_P}(y^n) - \min_{\mathcal{F}} \hat{L_F}(y^n). && (22)
\end{aligned}$$

Dividing by $\sqrt{n\log\log n}$ and taking the lim sup over the two ends of the above inequality gives us almost surely

$$\begin{aligned}
\limsup_{n\to\infty} \frac{[L_P(y^n,x^n) - \min_{\mathcal{F}} L_F(y^n,x^n)]}{\sqrt{n\log\log n}} &\leq \limsup_{n\to\infty} \frac{|\Delta_P(y^n,x^n)|}{\sqrt{n\log\log n}} && (23)\\
&\quad + \limsup_{n\to\infty} \frac{\max_{\mathcal{F}} |\Delta_F(y^n,x^n)|}{\sqrt{n\log\log n}} && (24)\\
&\quad + \limsup_{n\to\infty} \frac{[\hat{L_P}(y^n) - \min_{\mathcal{F}} \hat{L_F}(y^n)]}{\sqrt{n\log\log n}} && (25)\\
&\leq \limsup_{n\to\infty} \frac{|\Delta_P(y^n,x^n)|}{\sqrt{n\log\log n}} && (26)\\
&\quad + \max_{\mathcal{F}} \limsup_{n\to\infty} \frac{|\Delta_F(y^n,x^n)|}{\sqrt{n\log\log n}} && (27)
\end{aligned}$$

$$
\begin{aligned}
&+\limsup_{n\to\infty}\frac{g_n(y^n)}{\sqrt{n\log\log n}} && (28)\\
\le\ & 2\sqrt{2}C\cdot B+2\sqrt{2}C\cdot B+ && (29)\\
&\limsup_{n\to\infty}\frac{g_n(y^n)}{\sqrt{n\log\log n}}, && (30)
\end{aligned}
$$

where the last inequality follows by Lemma 2 and the one before it follows from (6) and by the fact that $\mathcal{F}$ is finite.□

The following result establishes a large deviations performance bound.

Theorem 6 *Let equation (6) hold for a certain predictor P. Suppose that L_0 and L_1 are upper bounded by $B>0$, that $E(h(y_t)-x_t)^2\le\beta^2$, and assume that $E|h(y_t)-x_t|^k\le\frac{k!}{2}\cdot E(h(y_t)-x_t)^2\cdot c^{k-2}$ for all $k>2$, $t\ge 1$ and some $0<c<\infty$. Then for all $\varepsilon>0$ and $0\le\alpha<1/2$*

$$
\begin{aligned}
\Pr\left\{\frac{1}{n}[L_P(y^n,x^n)-\inf_{F\in\mathcal{F}}L_F(y^n,x^n)]\ge\varepsilon\right\} &\le 2(|\mathcal{F}|+1)\exp\left\{-n\frac{\alpha^2\varepsilon^2}{8B^2(\beta^2+c\alpha\varepsilon/B)}\right\}\\
&+\Pr\left\{\frac{g_n(y^n)}{n}\ge(1-2\alpha)\varepsilon\right\}. \qquad (31)
\end{aligned}
$$

Note that, in particular, whenever $g_n(y^n)=O(\sqrt{n})$ deterministically, as will be shown to be the case in many situations of interest, the second term on the right-hand side of (31) is annihilated for sufficiently large n.

Proof:

$$
\begin{aligned}
&\Pr\left\{\frac{1}{n}[L_P(y^n,x^n)-\min_{\mathcal{F}}L_F(y^n,x^n)]\ge\epsilon\right\} && (32)\\
=\ &\Pr\{\frac{1}{n}[L_P(y^n,x^n)-\hat{L_P}(y^n)]+ && (33)\\
&\frac{1}{n}[\min_{\mathcal{F}}\hat{L_F}(y^n)-\min_{\mathcal{F}}L_F(y^n,x^n)]+ && (34)\\
&\frac{1}{n}[\hat{L_P}(y^n)-\min_{\mathcal{F}}\hat{L_F}(y^n)]\ge\epsilon\} && (35)\\
\le\ &\Pr\{\frac{1}{n}[L_P(y^n,x^n)-\hat{L_P}(y^n)]\ge\alpha\epsilon\}+ && (36)\\
&\Pr\{\frac{1}{n}[\min_{\mathcal{F}}\hat{L_F}(y^n)-\min_{\mathcal{F}}L_F(y^n,x^n)]\ge\alpha\epsilon\}+ && (37)\\
&\Pr\{\frac{1}{n}[\hat{L_P}(y^n)-\min_{\mathcal{F}}\hat{L_F}(y^n)]\ge(1-2\alpha)\epsilon\} && (38)\\
\le\ &\Pr\{\frac{1}{n}[L_P(y^n,x^n)-\hat{L_P}(y^n)]\ge\alpha\epsilon\}+ && (39)\\
&\Pr\{\frac{1}{n}[\max_{\mathcal{F}}\{\hat{L_F}(y^n)-L_F(y^n,x^n)\}]\ge\alpha\epsilon\}+ && (40)\\
&\Pr\{\frac{1}{n}g_n(y^n)\ge(1-2\alpha)\epsilon\} && (41)
\end{aligned}
$$

$$
\begin{aligned}
&\leq \Pr\{\frac{1}{n}[L_P(y^n,x^n) - \hat{L_P}(y^n)] \geq \alpha\epsilon\} + && (42)\\
&\sum_{\mathcal{F}} \Pr\{\frac{1}{n}[\hat{L_F}(y^n) - L_F(y^n,x^n)] \geq \alpha\epsilon\} + && (43)\\
&\Pr\{\frac{1}{n} g_n(y^n) \geq (1-2\alpha)\epsilon\} && (44)\\
&= \Pr\{\frac{1}{n}\Delta_P(y^n,x^n) \geq \alpha\epsilon\} + && (45)\\
&\sum_{\mathcal{F}} \Pr\{-\frac{1}{n}\Delta_F(y^n,x^n) \geq \alpha\epsilon\} + && (46)\\
&\Pr\{\frac{1}{n} g_n(y^n) \geq (1-2\alpha)\epsilon\} && (47)\\
&\leq 2(|\mathcal{F}|+1)\exp\{-n\frac{\alpha^2\varepsilon^2}{8B^2(\beta^2+c\alpha\varepsilon/B)}\} + && (48)\\
&\Pr\{\frac{1}{n} g_n(y^n) \geq (1-2\alpha)\epsilon\}, && (49)
\end{aligned}
$$

where the last inequality follows from Lemma 3 and inequality (41) follows from (6).□

A. A Sufficient Condition for Twofold Universality

The above results disclose clear and simple guidelines for prediction in the noisy setting: construct a cumulative loss estimator using an unbiased estimator $h(\cdot)$ (equation (5)), and then use a predictor that meets equation (6). Suppose now that we have a family of noise processes satisfying the above assumptions and indexed by some set $\mathcal{I}$. Suppose further that there exists some h such that $E^\alpha h(y_t) = x_t$ for all t and all $\alpha \in \mathcal{I}$, where we let E^α denote expectation w.r.t. $\Pr^\alpha$, which is the probabilistic law governing the noise process indexed by α. In this situation, employing a prediction scheme that satisfies (6), will guarantee performance in the sense of Theorems 4-6, whatever the active noise source from the family $\mathcal{I}$ may be. A prediction scheme which efficiently competes with a set of experts, regardless of the active noise process within a family of processes, will be said to be *twofold universal* (once w.r.t. the experts and once w.r.t. the law which is governing the noise). Evidently, a sufficient condition for twofold universality w.r.t. a family of possible noise processes, is the existence of an estimator $h(\cdot)$ which is unbiased for every noise process in the class. This fact lies at the heart of the surprisingly strong sense of twofold universality which will be established for the predictors employed in Section 4. The next sections are dedicated to two important special cases which illustrate the strength and applicability of the above results.

3 Binary Valued Noise

In this section, we focus on the special case where $\{y_t\}$ is related to $\{x_t\}$ via a binary symmetric channel (BSC). Specifically, the t-th noisy observation is $y_t = x_t \oplus r_t$, where $r = \{r_t, t \geq 1\}$ 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 probability and expectation under this measure, respectively. The superscript p will be suppressed whenever there is no ambiguity.

Consider the unbiased estimator

$$h(y_t) = \frac{y_t - p}{1 - 2p}, \tag{50}$$

which leads to:

$$\begin{aligned} \hat{L}_F(y^n) &= \sum_{t=1}^{n} (1 - \frac{y_t - p}{1 - 2p}) L_0(F_t(y^{t-1})) + \frac{y_t - p}{1 - 2p} L_1(F_t(y^{t-1})) && (51) \\ &= \sum_{t=1}^{n} \frac{(1 - y_t) - p}{1 - 2p} L_0(F_t(y^{t-1})) + \frac{y_t - p}{1 - 2p} L_1(F_t(y^{t-1})). && (52) \end{aligned}$$

We now seek a predictor that satisfies (6). As insinuated in the previous section, we identify $\hat{L}_F(y^n)$ as the cumulative loss of F under the loss function

$$L(y_t, F_t(y^{t-1})) = \frac{(1 - y_t) - p}{1 - 2p} L_0(F_t(y^{t-1})) + \frac{y_t - p}{1 - 2p} L_1(F_t(y^{t-1})). \tag{53}$$

We now use results from the noise-free setting for this loss function. To this end, let us first recall a well-known result in universal prediction based on clean observations, [38], [37], [14].

A. Harnessing Vovk's Algorithm for the Noisy Setting

Given a loss function L such that L_0 and L_1 are twice differentiable, we recall from [14] the functions

$$S(z) = L_0'(z) L_1''(z) - L_1'(z) L_0''(z) \tag{54}$$

and

$$R(z) = \frac{L_0'(z) L_1'(z)^2 - L_1'(z) L_0'(z)^2}{S(z)}, \tag{55}$$

where L_i' and L_i'' are the first and second derivatives of L_i, respectively, $i = 0, 1$. Now, define

$$c_L = \sup_{0<z<1} R(z). \tag{56}$$

If $S(z) = 0$ for some $0 < z < 1$, we write $c_L = \infty$. Let $\mathcal{F}$ be a given finite set of experts and consider the following prediction algorithm:

Algorithm 3.1

Initialization: *Enumerate the experts arbitrarily as* $\mathcal{F} = \{F^{(i)}\}_{i=1}^N$ *and set the weights* $w_{1,i} = 1$, $1 \le i \le N$.

Prediction: *Let* $v_{t,i} = w_{t,i}/W_t$, *where* $W_t = \sum_{i=1}^N w_{t,i}$. *At the beginning of trial* t, *compute for* $x = 0$ *and* $x = 1$, *the value*

$$\Delta(x) = -c_L \ln \sum_{i=1}^N v_{t,i} e^{-L(x, F_t^{(i)}(x^{t-1}))/c_L}. \tag{57}$$

and predict with any value P_t *that satisfies, for* $x = 0$ *and* $x = 1$, *the condition*

$$L(x, P_t) \le \Delta(x). \tag{58}$$

If no value of P_t *meets (58) then the algorithm is said to fail.*

Update: *After observing* x_t, *let*

$$w_{t+1,i} = w_{t,i} e^{-L(x, F_t^{(i)}(x^{t-1}))/c_L}. \tag{59}$$

The relevant results of [14] are the following.

Theorem 7 ([14]) *Let* L *be a loss function such that* $L_0(0) = L_1(1) = 0$, L_0 *and* L_1 *are three times differentiable in* $(0,1)$, *and* $L_0'(z) > 0$ *and* $L_1'(z) < 0$ *for* $0 < z < 1$. *Assume that* c_L *is finite and* $S(z)$ *is positive for* $0 < z < 1$. *Then, for any finite set of experts* $\mathcal{F}$, *the strongly sequential predictor* P, *given in Algorithm 3.1, is guaranteed to never fail, and for all* n

$$\max_{x^n \in \{0,1\}^n} \{L_P(x^n) - \min_{F \in \mathcal{F}} L_F(x^n)\} \le c_L \ln |\mathcal{F}|. \tag{60}$$

Returning now to the noisy setting, for a given $p \in [0, 1/2)$, we define the following auxiliary loss function, $\tilde{L}$, in terms of the original one:

$$\tilde{L}_0(z) \stackrel{\text{def}}{=} \frac{1-p}{1-2p} L_0(z) - \frac{p}{1-2p} L_1(z) + \frac{p}{1-2p} L_1(0) \tag{61}$$

and

$$\tilde{L}_1(z) \stackrel{\text{def}}{=} \frac{-p}{1-2p} L_0(z) + \frac{1-p}{1-2p} L_1(z) + \frac{p}{1-2p} L_0(1). \tag{62}$$

so that

$$\tilde{L}(y,z) = (1-y)\tilde{L}_0(z) + y\tilde{L}_1(z). \tag{63}$$

Note that, up to the third terms on the right-hand sides of equations (61) and (62) (which do not depend on z), $\tilde{L}_0$ and $\tilde{L}_1$ play the roles of L_0 and L_1 corresponding to the loss function of (53). The additional constants were added so as to have $\tilde{L}_0(0) = \tilde{L}_1(1) = 0$. Let now $S_p(z)$, $R_p(z)$ and $c_L(p)$ be as in equations (54)-(56) for the loss function $\tilde{L}$, where if $S_p(z) = 0$ for some $0 < z < 1$, we take $c_L(p) = \infty$. We can now state the following.

Theorem 8 *Let L be a loss function such that: (i) $L_0(0) = L_1(1) = 0$, (ii) L_0 and L_1 are three times differentiable in $(0,1)$, and (iii) $L_0'(z) > 0$ and $L_1'(z) < 0$ for $0 < z < 1$. Assume further that $c_L(p) < \infty$ and that $S(z)$ is positive for $0 < z < 1$. For any expert class $\{F^{(i)} : 1 \le i \le N\}$, let P be the strongly sequential predictor given in Algorithm 3.1, with $\tilde{L}$ replacing L, and $F_t^{(i)}(y^{t-1})$ replacing $F_t^{(i)}(x^{t-1})$. Then P is guaranteed to never fail and we have for all n*

$$\max_{y^n \in \{0,1\}^n} \{\hat{L}_P(y^n) - \min_{1 \le i \le N} \hat{L}_{F^{(i)}}(y^n)\} \le c_L(p) \ln N. \tag{64}$$

Proof sketch (a full proof can be found in [43]): This result follows by applying Theorem 7 to the auxiliary loss function given in equations (61) and (62). The proof consists of a verification that the hypotheses of this theorem lead to the hypotheses of Theorem 7 for $\tilde{L}$. This assures us that, using the predictor suggested, we have for all $y^n \in \{0,1\}^n$

$$\sum_{t=1}^{n} \tilde{L}(y_t, P_t(y^{t-1})) - \min_{1 \le i \le N} \sum_{t=1}^{n} \tilde{L}(y_t, F_t^{(i)}(y^{t-1})) \le c_L(p) \ln N. \tag{65}$$

The proof is then completed by showing that the above upper bound for the difference in terms of the cumulative $\tilde{L}$-losses is also an upper bound for the difference in terms of the original loss estimators of equation (52).□

We next examine more closely a few examples.

Example 1 *α-powered loss function*

Consider the *α-powered* loss function defined by $L_\alpha(x,z) = |x-z|^\alpha$. Straightforward computations for $1 < \alpha \le 2$ give:

$$c_{L_\alpha}(p) = \frac{\alpha(1 - 6p^2 + 18p^3 - 20p^4 + 8p^5)}{2^\alpha(\alpha-1)(1-2p)^3}. \tag{66}$$

Example 2 *Hellinger loss function*

Consider the generalized Hellinger loss function given by $L_0(z) = 1 - (1-z)^\alpha$ and $L_1(z) = 1 - z^\alpha$. Computations similar to those for the previous example give for $1/2 \le \alpha < 1$:

$$c_{L_{H,\alpha}}(p) = \frac{\alpha(1 - 6p^2 + 18p^3 - 20p^4 + 8p^5)}{2^\alpha(1-\alpha)(1-2p)^3}. \tag{67}$$

As expected, the expressions in equations (66) and (67) coincide with those of $p = 0$ (cf. Example 3.2 of [14] for the square loss, which is obtained from the first example with $\alpha = 2$, and the Hellinger loss, which is obtained from the second example with $\alpha = 1/2$) and are easily seen to increase with p and become arbitrarily large for p sufficiently close to $1/2$.

Example 3 *The logarithmic loss function*

The logarithmic loss, also known as the *relative entropy loss*, is defined by $L(x,z) = x\ln(x/z) + (1-x)\ln((1-x)/(1-z))$, where $0 \ln 0 \stackrel{\text{def}}{=} 0$. This gives $L_0(z) = -\ln(1-z)$ and $L_1(z) = -\ln z$. A straightforward calculation for this case yields

$$R_p(z) = \frac{1-p+p^2}{(1-2p)^3} + \frac{(1-p)p(1-3z+3z^2)}{z-z^2}, \tag{68}$$

which is not bounded on $(0,1)$ whenever $p > 0$ and therefore $c_L(p) = \infty$ for all $p > 0$, and $c_L(0) = 1$ (as is known from the cumulative case and from universal coding), so that Theorem 8 is not useful in this case.

Combining Theorem 8 with Theorems 4-6, we can now harvest Corollaries 9-11, respectively.

Corollary 9 *Let the hypotheses of Theorem 8 hold. Then for any finite expert class $\mathcal{F}$ there exists a strongly sequential predictor P (that of Theorem 8) such that for all n and $x^n \in \{0,1\}^n$*

$$EL_P(y^n, x^n) - \inf_{F\in\mathcal{F}} EL_F(y^n, x^n) \le c_L(p) \ln |\mathcal{F}|.$$

Corollary 10 *Let the hypotheses of Theorem 8 hold and suppose further that L_0 and L_1 are upper bounded by $B > 0$. Then for any finite expert class $\mathcal{F}$ there exists a strongly sequential predictor P (that of Theorem 8) such that almost surely*

$$\limsup_{n\to\infty} \frac{[L_P(y^n, x^n) - \min_{F\in\mathcal{F}} L_F(y^n, x^n)]}{\sqrt{n \log\log n}} \le \frac{4\sqrt{2}B}{1-2p}. \tag{69}$$

Proof: Immediate from Theorem 5 by taking $C = 1/(1-2p) \geq (1-p)/(1-2p) \geq |(y_t-p)/(1-2p)| = |h(y_t)|$ and by the fact that for our case $g_n(y^n) = c_L(p)\ln|\mathcal{F}| = o(\sqrt{n\log\log n})$.□
Finally, combining Theorems 8 and 6 gives us

Corollary 11 *Under the conditions of Corollary 10, for any finite expert class $\mathcal{F}$ there exists a strongly sequential predictor P (that of Theorem 8) such that for all $\varepsilon > 0$, $0 \leq \alpha < 1/2$, and $n > c_L(p)\ln|\mathcal{F}|/((1-2\alpha)\varepsilon)$ we have*

$$\Pr\left\{\frac{1}{n}[L_P(y^n,x^n) - \inf_{F\in\mathcal{F}} L_F(y^n,x^n)] \geq \varepsilon\right\} \leq 2(|\mathcal{F}|+1)\exp\left\{-n\frac{\alpha^2\varepsilon^2 p(1-2p)^2}{4B^2(1-p)(4p^2-3p+1)}\right\}. \tag{70}$$

Proof: Since the assertion of the theorem follows trivially for $\varepsilon > B$, assume $\varepsilon \leq B$. We note first that our assumption on n, combined with Theorem 8, annihilates the second term on the right-hand side of (31). Next, we take $\beta^2 = E(h(y_t)-x_t)^2 = (1-p)\frac{p^2}{(1-2p)^2} + p\frac{(1-p)^2}{(1-2p)^2} = \frac{p(1-p)}{(1-2p)^2}$ and note that, since $|h(y_t)-x_t| \leq |\frac{1-p}{1-2p}|$ almost surely for all t, we have for all $k > 2$

$$\begin{aligned}
E|h(y_t)-x_t|^k &\leq \left(\frac{1-p}{1-2p}\right)^k && (71)\\
&= \frac{p(1-p)}{(1-2p)^2}\cdot\frac{(1-p)^{k-1}}{p(1-2p)^{k-2}} && (72)\\
&= E(h(y_t)-x_t)^2\cdot\frac{(1-p)^{k-1}}{p(1-2p)^{k-2}} && (73)\\
&\leq E(h(y_t)-x_t)^2\cdot\left[\frac{(1-p)^2}{p(1-2p)}\right]^{k-2} \\
&\leq \frac{k!}{2}E(h(y_t)-x_t)^2\cdot\left[\frac{(1-p)^2}{p(1-2p)}\right]^{k-2}. && (74)
\end{aligned}$$

Evidently, the hypotheses of Theorem 6 are satisfied with the above β^2 and $c = \frac{(1-p)^2}{p(1-2p)}$. Plugging these values of β^2 and c into the right-hand side of (31), we obtain

$$\begin{aligned}
&\Pr\left\{\frac{1}{n}[L_P(y^n,x^n) - \inf_{F\in\mathcal{F}} L_F(y^n,x^n)] \geq \varepsilon\right\} && (75)\\
\leq\;& 2(|\mathcal{F}|+1)\exp\{-n\frac{\alpha^2\varepsilon^2}{8B^2(\frac{p(1-p)}{(1-2p)^2} + \frac{(1-p)^2}{p(1-2p)}\alpha\varepsilon/B)}\} && (76)\\
\leq\;& 2(|\mathcal{F}|+1)\exp\{-n\frac{\alpha^2\varepsilon^2}{8B^2(\frac{p(1-p)}{(1-2p)^2} + \frac{(1-p)^2}{2p(1-2p)})}\}, && (77)
\end{aligned}$$

where, for the second inequality, we have used the fact that $\alpha\varepsilon/B \leq 1/2$. Simplifying the expression in the exponent of (77) gives the right-hand side of (70).□

B. Twofold Universality Under the Absolute Loss Function

The important special case of the absolute loss function, defined by $L(x_t, P_t(x^{t-1})) \stackrel{\text{def}}{=} |x_t - P_t(x^{t-1})|$, is not covered by the results of the previous subsection. This is because $c_L(p)$ is infinite in the case of the absolute loss (even in the noise-free case where $p = 0$, cf. [14], [2]). The case of the absolute loss function must therefore be treated separately. As it turns out, for this case, not only can a predictor efficiently compete with a set of experts in the sense of the performance criteria considered in the previous subsection, but also be independent of p. In other words, there exist predictors, for this case, which are twofold universal. This is in contrast to the predictors of the previous subsection, which assume knowledge of p, since the loss estimators that they employ are constructed with $h(\cdot)$, which depends on p (equation (50)).

Taking a closer look at the absolute loss function and setting $L_0(z) = z$ and $L_1(z) = 1-z$ into equations (4), (5), and (2), we obtain

$$L_F(y^n, x^n) = \sum_{t=1}^{n}(1 - x_t)F_t(y^{t-1}) + x_t(1 - F_t(y^{t-1})), \tag{78}$$

$$\hat{L}_F(y^n) = \sum_{t=1}^{n} \frac{1 - y_t - p}{1 - 2p} F_t(y^{t-1}) + \frac{y_t - p}{1 - 2p}(1 - F_t(y^{t-1})), \tag{79}$$

and

$$L_F(y^n) = \sum_{t=1}^{n}(1 - y_t)F_t(y^{t-1}) + y_t(1 - F_t(y^{t-1})), \tag{80}$$

respectively. The simple, yet essential, observation to make towards establishing universality of our predictor with respect to the noise distribution, which follows easily from the last two expressions, is that for any expert F, all n and $y^n \in \{0,1\}^n$, we have the following linear relation between $\hat{L}_F(y^n)$ and $L_F(y^n)$:

$$\hat{L}_F(y^n) = \frac{1}{1 - 2p}(L_F(y^n) - p). \tag{81}$$

Therefore, for *any* predictor P and expert class $\mathcal{F}$, we have

$$\begin{aligned} \hat{L}_P(y^n) - \inf_{F \in \mathcal{F}} \hat{L}_F(y^n) &= \frac{1}{1 - 2p}(L_P(y^n) - p) - \inf_{F \in \mathcal{F}} \frac{1}{1 - 2p}(L_F(y^n) - p) \\ &= \frac{1}{1 - 2p}[L_P(y^n) - \inf_{F \in \mathcal{F}} L_F(y^n)]. \end{aligned} \tag{82}$$

Consequently, any bound on the maximum regret of a predictor in the noise-free case, is also a bound for the noisy case after dividing by $1 - 2p$. This gives rise to all performance bounds of Theorems 4, 5, and 6. Therefore, for the case of absolute loss, one may use any good predictor for the noise-free setting and be guaranteed to perform well for every $p \in [0, 1/2)$. For example, for a finite set of experts $\mathcal{F}$, we can let

$$P_t(y^{t-1}) = \frac{\sum_{F \in \mathcal{F}} \exp\{-\eta_{k(t)} L_F(y^{t-1})_{2^{k(t)+1}+1}^{t-1}\} F_t(y^{t-1})}{\sum_{F \in \mathcal{F}} \exp\{-\eta_{k(t)} L_F(y^{t-1})_{2^{k(t)+1}+1}^{t-1}\}}, \tag{83}$$

where $k(t) = \lfloor \log t - 1 \rfloor$ and $\eta_k = \sqrt{8 \ln |\mathcal{F}| / 2^{k+1}}$. Note that

For this strongly sequential predictor, we have the following result from the noise-free setting:

Theorem 12 *For any finite expert set $\mathcal{F}$, using P of equation (83), we have for all n*

$$\max_{y^n \in \{0,1\}^n} \{L_P(y^n) - \min_{F \in \mathcal{F}} L_F(y^n)\} \leq \frac{\sqrt{n \ln |\mathcal{F}|}}{\sqrt{2} - 1}. \tag{84}$$

The strongly sequential predictor of equation (83) is constructed by using the (horizon-dependent) exponential weighting predictor in consecutive blocks of exponentially increasing length. The proof of Theorem 12 appears in [43, Theorem 3.6]. It is established by applying an upper bound ([2, Theorem 1]) on the regret for each block. Theorem 12, combined with equation (82), assure us that, using the predictor of equation (83), which is independent of p, we have, for all n and $y^n \in \{0,1\}^n$

$$\hat{L}_P(y^n) - \min_{F \in \mathcal{F}} \hat{L}_F(y^n) \leq \frac{\sqrt{n \ln |\mathcal{F}|}}{(\sqrt{2} - 1)(1 - 2p)}. \tag{85}$$

Consequently, letting $g_n(y^n) = \frac{\sqrt{n \ln |\mathcal{F}|}}{(\sqrt{2}-1)(1-2p)}$, Theorems 4, 5, and 6 give us the following respective three corollaries for the case of the absolute loss function, similarly as in the previous subsection. It should be emphasized that, while the predictor employed in the previous subsection depends on p, the predictor (83) does not. Specifically, we have:

Corollary 13 *For any expert class $\mathcal{F}$, let P be the strongly sequential predictor of equation (83). Then for all n, $x^n \in \{0,1\}^n$, and $p \in [0, 1/2)$*

$$E^p L_P(y^n, x^n) - \inf_{F \in \mathcal{F}} E^p L_F(y^n, x^n) \leq \frac{\sqrt{n \ln |\mathcal{F}|}}{(\sqrt{2} - 1)(1 - 2p)}. \tag{86}$$

Corollary 14 *For any expert class $\mathcal{F}$, let P be the strongly sequential predictor of equation (83). Then for all $p \in [0, 1/2)$,* $\Pr^p$*-almost surely*

$$\limsup_{n \to \infty} \frac{[L_P(y^n, x^n) - \min_{F \in \mathcal{F}} L_F(y^n, x^n)]}{\sqrt{n \log \log n}} \leq \frac{4\sqrt{2}}{1 - 2p}. \tag{87}$$

Corollary 15 *For any expert class $\mathcal{F}$, let P be the strongly sequential predictor of equation (83). Then for all $\varepsilon > 0$, $0 \le \alpha < 1/2$, $n > \ln|\mathcal{F}|/[\varepsilon(1-2\alpha)(\sqrt{2}-1)(1-2p)]^2$, and $p \in [0, 1/2)$ we have*

$$\Pr^p\left\{\frac{1}{n}[L_P(y^n, x^n) - \inf_{F\in\mathcal{F}} L_F(y^n, x^n)] \ge \varepsilon\right\} \le 2(|\mathcal{F}|+1)\exp\left\{-n\frac{\alpha^2\varepsilon^2 p(1-2p)^2}{4(1-p)(4p^2-3p+1)}\right\}. \tag{88}$$

C. Worst-Case Relative Expected Loss Asymptotics

Let us define now the *worst-case relative expected loss* of a predictor P by

$$R_n^p(P, \mathcal{F}) \stackrel{\text{def}}{=} \max_{x^n\in\{0,1\}^n}(E^p L_P(y^n, x^n) - \inf_{F\in\mathcal{F}} E^p L_F(y^n, x^n)). \tag{89}$$

The *minimax relative expected loss* achievable by any predictor with respect to a particular expert class $\mathcal{F}$ is defined by

$$V_n^p(\mathcal{F}) \stackrel{\text{def}}{=} \inf_P R_n^p(P, \mathcal{F}), \tag{90}$$

where the infimum is actually a minimum as the set of all distributions P over $\{0,1\}^n$ is compact (identify, e.g., each such P with a point in $\{g \in [0,1]^{2^n} : g_i \ge 0, \sum_{i=1}^{2^n} g_i = 1\}$), and $R_n^p(P, \mathcal{F})$ is a continuous functional of P. Finally, let

$$V^p(N, n) \stackrel{\text{def}}{=} \sup_{\mathcal{F}:|\mathcal{F}|=N} V_n^p(\mathcal{F}). \tag{91}$$

To denote the quantities analogous to the last three for the cumulative case, we simply omit the superscript p. In order to state our asymptotic order results for $V^p(N, n)$, we first recall the relevant asymptotic order results for the cumulative case, which can be found in [14, Theorems 3.1 and 3.3] (see Appendix A for precise definitions of the asymptotic notation):

Theorem 16 ([14]) *1. For a loss function satisfying the hypotheses of Theorem 8, we have*

$$V(N, n) = \Theta(\ln N). \tag{92}$$

2. For the absolute loss function, we have

$$V(N, n) = \Theta(\sqrt{n \log N}). \tag{93}$$

We make an important note, that serves as a basis for the proof of the lower bounds in the following theorem: The (randomly constructed) expert sets, used in [14] to prove the lower bounds in the above theorem, were actually sets of memoryless experts (i.e., prediction schemes which do not depend on the data). More precisely, wherever in the above theorem we have $V(N,n) = \Theta(g(n,N))$, it was actually established in [14] that there exists a constant $C > 0$ such that for sufficiently large n and N, there always exists a class $\mathcal{F}_s$ of memoryless experts of cardinality N for which $V_n(\mathcal{F}_s) \geq Cg(n,N)$. We now provide an analogue to Theorem 16 for the noisy case.

Theorem 17 *1. Let the loss function be such that the hypotheses of Theorem 8 hold for a given $B > 0$. Assume further that $L_0''(z) \geq 0$ and $L_1''(z) \geq 0$ for $0 < z < 1$. We then have*

$$V^p(N,n) = \Theta(\ln N). \tag{94}$$

2. For the absolute loss function, we have

$$V^p(N,n) = \Theta(\sqrt{n \log N}). \tag{95}$$

Item 2 of Theorem 17 can actually be shown to hold for more than just the case of the absolute loss function (cf. Theorem 3.11 in [43]).

Proof: The upper bound for the first item follows from Corollary 9 and that for the second follows from Corollary 13. To prove the lower bounds, one must simply establish the validity of the lower bounds of Theorem 16 for our noisy setting. To this end, note that for any predictor P and $x^n \in \{0,1\}^n$ one has

$$\begin{aligned}
E^p L_P(y^n, x^n) &= \sum_{t=1}^{n} (1-x_t) E^p L_0(P_t(y^{t-1})) + x_t E^p L_1(P_t(y^{t-1})) &(96)\\
&\geq \sum_{t=1}^{n} (1-x_t) L_0(E^p P_t(y^{t-1})) + x_t L_1(E^p P_t(y^{t-1})) &(97)\\
&= \sum_{t=1}^{n} (1-x_t) L_0(P_t^p(x^{t-1})) + x_t L_1(P_t^p(x^{t-1})) &(98)\\
&= L_{P^p}(x^n, x^n), &(99)
\end{aligned}$$

where for a given predictor P, we define the predictor P^p by

$$P_t^p(x^{t-1}) \stackrel{\text{def}}{=} E^p P_t(x^{t-1} \oplus r^{t-1}) \tag{100}$$

and the inequality follows from that of Jensen by the convexity of L_0 and L_1. In words, the expected cumulative loss of any predictor P is never less than the cumulative loss of the predictor P^p in the cumulative setting whenever the loss function is convex. On the other hand, for any memoryless (static) expert G, one clearly has

$$E^p L_G(y^n, x^n) = L_{G^p}(x^n, x^n) = L_G(x^n, x^n) \tag{101}$$

as the expectations in (99) and in (100) are degenerate for this case. From (99) and (101) it follows that when the loss function is convex, for any predictor P and any set of memoryless (static) experts $\mathcal{F}_s$ one has for all $x^n \in \{0,1\}^n$

$$E^p L_P(y^n, x^n) - \min_{F \in \mathcal{F}_s} E^p L_F(y^n, x^n) \geq L_{P^p}(x^n, x^n) - \min_{F \in \mathcal{F}_s} L_F(x^n, x^n). \tag{102}$$

Finally, since the right hand side of (102) can be identified with the case of cumulative prediction with respect to a class of static experts, it follows that under the hypotheses of this theorem

$$V_n^p(\mathcal{F}_s) \geq V_n(\mathcal{F}_s) \tag{103}$$

for any class of static experts $\mathcal{F}_s$. The asymptotic lower bounds in this theorem now follow from those of Theorem 16, the note following it, and inequality (103).□

We merely remark here that while the results of Theorem 17 are asymptotic, non-asymptotic bounds can be obtained for our noisy setup as well. For the upper-bounds this is clear, as our upper bounds (Corollaries 9 and 13) are non-asymptotic. As for non-asymptotic lower-bounds, the techniques of Cesa-Bianchi and Lugosi (cf. e.g. [2, Theorem 9]) can be extended to our setting of noisy observations and general convex loss functions, to obtain meaningful bounds.

D. Infinite Expert Classes and the Absolute Loss Function

We now demonstrate how the previous results for finite sets of experts extend to more general sets. Rather than giving a comprehensive treatment of the case of infinite expert classes, we illustrate, through one powerful example, for the case of the absolute loss function, how results from the cumulative setting, regarding infinite expert classes, can be combined with results for finite sets of experts in the noisy setting to obtain meaningful performance bounds. Specifically, we construct a predictor which competes with an arbitrary given expert class $\mathcal{F}$, not necessarily finite, which satisfies sufficient "integrability" and boundedness conditions. In light of the preceding subsection, which treated the special case of absolute loss and established the feasibility of twofold universality under this loss function, we introduce a predictor, which was originally constructed by Cesa-Bianchi and Lugosi in [2] for the noise-free case (up to some minor modifications which were made here for the sake of strong sequentiality). This predictor will be shown to be doing well in our noisy setup for all possible values of $p \in [0, 1/2)$.

We now describe the predictor of [2]. The argument builds on the assumption that the expert class $\mathcal{F}$ is δ-bounded, i.e., there exists $0 < \delta < 1/2$ such that for every $F \in \mathcal{F}$, t, and $y^{t-1} \in \{0,1\}^{t-1}$, $F_t(y^{t-1}) \in [\delta, 1-\delta]$. For any set $\mathcal{F}$ of experts, define the metric ρ by

$$\rho(F, G) \stackrel{\text{def}}{=} \sup_{t \geq 1} \sup_{y^{t-1} \in \{0,1\}^{t-1}} |F_t(y^{t-1}) - G_t(y^{t-1})|. \tag{104}$$

For any $\varepsilon > 0$, an *ε-cover* of $\mathcal{F}$ is a set $\mathcal{F}_\varepsilon$ of experts such that for all $F \in \mathcal{F}$, there exists $G \in \mathcal{F}_\varepsilon$ such that $\rho(F, G) < \varepsilon$. The *ε-covering number* $N_\infty(\mathcal{F}, \varepsilon)$ is the cardinality of the smallest ε-cover of $\mathcal{F}$. For all $k \geq 1$, define

$$\varepsilon_k = \frac{1 - 2\delta}{2(1 + 2\delta)} (2\delta)^k, \tag{105}$$

and

$$a_k = (1 - 2\delta)(2\delta)^{k-1}. \tag{106}$$

For any fixed class $\mathcal{F}$ of experts and for all $k \geq 1$, let $\mathcal{H}_k$ be an ε_k-cover of $\mathcal{F}$ of cardinality $N_k = N_\infty(\mathcal{F}, \varepsilon_k)$. Define $\mathcal{H}_0 = \{F^{(0)}\}$ as the singleton class containing the static expert $F^{(0)}$ such that $F_t^{(0)}(y^{t-1}) = 1/2$ for all $t \geq 1$ and all $y^{t-1} \in \{0,1\}^{t-1}$ (in particular, note that $\mathcal{H}_0$ is a 1/2-cover of $\mathcal{F}$).

For each $k \geq 1$ define $\mathcal{Q}_k$ as the class of all experts of the form

$$Q_t \stackrel{\text{def}}{=} \frac{1}{2} + \frac{1}{a_k} (F_t^{(k)} - F_t^{(k-1)}) \tag{107}$$

for all the pairs $(F^{(k-1)}, F^{(k)}) \in \mathcal{H}_{k-1} \times \mathcal{H}_k$ satisfying $\rho(F^{(k)}, F^{(k-1)}) \leq a_k/2$. Now, for each $k \geq 1$ we let $P^{(k)}$ be the predictor of equation (83), with the finite expert class $\mathcal{Q}_k$ playing the role of $\mathcal{F}$. We define P by

$$P_t \stackrel{\text{def}}{=} \sum_{k=1}^{\infty} a_k P_t^{(k)}. \tag{108}$$

The following Theorem (up to slight modifications for strong sequentiality) is due to Cesa-Bianchi and Lugosi [2].

Theorem 18 ([2]) *Let $\mathcal{F}$ be an arbitrary δ-bounded expert class. Then, using the predictor P of equation (108), we have for all $y^n \in \{0,1\}^n$*

$$L_P(y^n) - \min_{F \in \mathcal{F}} L_F(y^n) \leq \frac{\sqrt{2}(1+2\delta)\sqrt{n}\int_0^1 \sqrt{\ln N_\infty(\mathcal{F},\varepsilon)}d\varepsilon}{\delta(1-2\delta)(\sqrt{2}-1)}. \tag{109}$$

From this we easily obtain the following:

Theorem 19 *For any δ-bounded expert class $\mathcal{F}$, the predictor of equation (108) satisfies for all $n \geq 1$ and all $0 \leq p < 1/2$*

$$\max_{x^n \in \{0,1\}^n} \{E^p L_P(y^n, x^n) - \inf_{F \in \mathcal{F}} E^p L_F(y^n, x^n)\} \leq \frac{\sqrt{2}(1+2\delta)\sqrt{n}\int_0^1 \sqrt{\ln N_\infty(\mathcal{F},\varepsilon)}d\varepsilon}{\delta(1-2\delta)(1-2p)(\sqrt{2}-1)}. \tag{110}$$

Proof: For every $x^n \in \{0,1\}^n$ we have

$$\begin{aligned}
E^p L_P(y^n, x^n) - \inf_{F \in \mathcal{F}} E^p L_F(y^n, x^n) &= E^p \hat{L}_P(y^n) - \inf_{F \in \mathcal{F}} E^p \hat{L}_F(y^n) && (111)\\
&\leq E^p[\hat{L}_P(y^n) - \inf_{F \in \mathcal{F}} \hat{L}_F(y^n)] && (112)\\
&= \frac{1}{1-2p} E^p[L_P(y^n) - \inf_{F \in \mathcal{F}} L_F(y^n)] && (113)\\
&\leq \frac{\sqrt{2}(1+2\delta)}{\delta(1-2\delta)(1-2p)(\sqrt{2}-1)}\sqrt{n}\int_0^1 \sqrt{\ln N_\infty(\mathcal{F},\varepsilon)}d\varepsilon,
\end{aligned}$$

where the first equality follows from Lemma 1, the second equality follows from (82), and the last inequality follows from taking expectations over both sides of (109).□

The following, considerably stronger result assures us that, typically, the *actual* relative loss will not exceed the $\sqrt{n}$ order.

Theorem 20 *For a given δ-bounded expert class $\mathcal{F}$ with $\int_0^1 \sqrt{\ln N_\infty(\mathcal{F},\delta)}d\delta < \infty$, let P be the predictor of equation (108). Then for each $\varepsilon > 0$ and $0 \leq p < 1/2$ there exists $K = K(\int_0^1 \sqrt{\ln N_\infty(\mathcal{F},\delta)}d\delta, \varepsilon, p, \delta) < \infty$ such that for* all $n \geq 1$ *and $x^n \in \{0,1\}^n$*

$$\Pr{}^p\{\frac{1}{\sqrt{n}}[L_P(y^n, x^n) - \inf_{F \in \mathcal{F}} L_F(y^n, x^n)] \geq K\} \leq \varepsilon. \tag{114}$$

The following is the key result on which the proof of Theorem 20 is based.

Theorem 21 *For all classes $\mathcal{F}$ of experts, all $0 \leq p < 1/2$, all n and all $x^n \in \{0,1\}^n$, we have*

$$E^p \sup_{F \in \mathcal{F}} |\frac{1}{\sqrt{n}}(L_F(y^n, x^n) - \hat{L}_F(y^n))| \leq \frac{2}{1-2p}[1 + 16\int_0^1 \sqrt{\ln N_\infty(\mathcal{F},\delta)}d\delta]. \tag{115}$$

Proof of Theorem 21: Throughout the proof, we fix n and $x^n \in \{0,1\}^n$ and denote for convenience, for any expert F,

$$\Lambda_F = \Lambda_F(y^n, x^n) \stackrel{\text{def}}{=} \frac{1}{\sqrt{n}}(L_F(y^n, x^n) - \hat{L}_F(y^n)). \tag{116}$$

Our mission is to establish:

$$E^p \sup_{F\in\mathcal{F}} |\Lambda_F| \leq \frac{2}{1-2p}[1 + 16\int_0^1 \sqrt{\ln N_\infty(\mathcal{F},\delta)}d\delta]. \tag{117}$$

Note that $\sup_{F\in\mathcal{F}}|\Lambda_F|$ is a legitimate random variable no matter what the expert class $\mathcal{F}$ may be, since for a fixed individual sequence, x^n, the cardinality of the set in which it takes its values can be at most 2^n (corresponding to the possible values of y^n)[1]. The following Lemma is proven in the appendix:

Lemma 22 *For any $\varepsilon > 0$, $N \geq 2$ and set of $2N$ experts*

$$\{F^{(1)}, F^{(2)}, \ldots, F^{(N)}, G^{(1)}, G^{(2)}, \ldots, G^{(N)}\} \tag{118}$$

that satisfy

$$\rho(F^{(i)}, G^{(i)}) \leq \varepsilon \qquad \forall 1 \leq i \leq N, \tag{119}$$

we have:

$$E^p \max_{1\leq i\leq N} |\Lambda_{F^{(i)}} - \Lambda_{G^{(i)}}| \leq \varepsilon\frac{8}{1-2p}\sqrt{\ln N}. \tag{120}$$

Equipped with Lemma 22, we proceed to complete the proof of Theorem 21, using the so called *chaining* technique of empirical processes theory (cf. e.g., [23]). For all $k \geq 1$, let $\mathcal{G}_k$ be a 2^{-k}-cover of $\mathcal{F}$ of cardinality $N_k = N_\infty(\mathcal{F}, 2^{-k})$. Define $\mathcal{G}_0 = \{F^{(0)}\}$ as the singleton class containing the static expert $F^{(0)}$ such that $F_t^{(0)}(y^{t-1}) = 1/2$ for all $t \geq 1$ and all $y^{t-1} \in \{0,1\}^{t-1}$ (in particular, note that $\mathcal{G}_0$ is a $1/2$-cover of $\mathcal{F}$). For each $k \geq 0$ define the mapping

$$h_k : \mathcal{G}_{k+1} \to \mathcal{G}_k \tag{121}$$

such that for each $F \in \mathcal{G}_{k+1}$, $h_k(F)$ is an expert satisfying

$$h_k(F) = \arg\min_{G\in\mathcal{G}_k} \rho(F, G). \tag{122}$$

We now note that, by the triangle inequality, $\forall F \in \mathcal{G}_{k+1}$ and $k \geq 0$,

$$|\Lambda_F| \leq |\Lambda_{h_k(F)}| + |\Lambda_F - \Lambda_{h_k(F)}|. \tag{123}$$

[1]Generally, if $\Lambda = \{\Lambda_F\}_{F\in\mathcal{F}}$ is a random process indexed by the elements of $\mathcal{F}$, when $\mathcal{F}$ is not countable, the pointwise supremum, $\sup_{F\in\mathcal{F}}|\Lambda_F|$, may not be well defined. Cf., e.g., discussion in [16, Subsection 2.2].

The right hand side is, in turn, clearly upper bounded by

$$\max_{F\in\mathcal{G}_k}|\Lambda_F| + \max_{F\in\mathcal{G}_{k+1}}|\Lambda_F - \Lambda_{h_k(F)}| \tag{124}$$

so that we have $\forall F \in \mathcal{G}_{k+1}$ and $k \geq 0$,

$$|\Lambda_F| \leq \max_{F\in\mathcal{G}_k}|\Lambda_F| + \max_{F\in\mathcal{G}_{k+1}}|\Lambda_F - \Lambda_{h_k(F)}| \tag{125}$$

but since the right hand side of the above inequality is independent of the particular $F \in \mathcal{G}_{k+1}$, we obtain

$$\max_{F\in\mathcal{G}_{k+1}}|\Lambda_F| \leq \max_{F\in\mathcal{G}_k}|\Lambda_F| + \max_{F\in\mathcal{G}_{k+1}}|\Lambda_F - \Lambda_{h_k(F)}|. \tag{126}$$

Taking expectations gives

$$E^p\max_{F\in\mathcal{G}_{k+1}}|\Lambda_F| \leq E^p\max_{F\in\mathcal{G}_k}|\Lambda_F| + E^p\max_{F\in\mathcal{G}_{k+1}}|\Lambda_F - \Lambda_{h_k(F)}|. \tag{127}$$

Now, by the definition of the mapping h_k, for every F we have $\rho(F, h_k(F)) \leq 2^{-k}$ and therefore, applying Lemma 22 to the second term on the right hand side of equation (127) gives

$$\begin{aligned}
E^p\max_{F\in\mathcal{G}_{k+1}}|\Lambda_F| &\leq E^p\max_{F\in\mathcal{G}_k}|\Lambda_F| + 2^{-k}\frac{8}{1-2p}\sqrt{\ln|\mathcal{G}_{k+1}|} & (128)\\
&= E^p\max_{F\in\mathcal{G}_k}|\Lambda_F| + \frac{8}{1-2p}2^{-k}\sqrt{\ln N_\infty(\mathcal{F}, 2^{-(k+1)})}. & (129)
\end{aligned}$$

Exploiting the recursive form of inequality (129), repeated substitution for the first term on the right hand side eventually replaces it by a maximum over the singleton set $\mathcal{G}_0$, with the addition of one more error term for each level moved up:

$$\begin{aligned}
E^p\max_{F\in\mathcal{G}_{k+1}}|\Lambda_F| &\leq E^p|\Lambda_{F(0)}| + \frac{8}{1-2p}\sum_{j=0}^{k}2^{-j}\sqrt{\ln N_\infty(\mathcal{F}, 2^{-(j+1)})} & (130)\\
&\leq E^p|\Lambda_{F(0)}| + \frac{8}{1-2p}\sum_{j=0}^{\infty}2^{-j}\sqrt{\ln N_\infty(\mathcal{F}, 2^{-(j+1)})} & (131)\\
&= E^p|\Lambda_{F(0)}| + \frac{32}{1-2p}\sum_{j=0}^{\infty}2^{-(j+1)}\sqrt{\ln N_\infty(\mathcal{F}, 2^{-(j+1)})} & (132)\\
&\leq E^p|\Lambda_{F(0)}| + \frac{32}{1-2p}\int_0^1\sqrt{\ln N_\infty(\mathcal{F}, \delta)}d\delta, & (133)
\end{aligned}$$

where the integral bound for the infinite sum is possible since clearly $N_\infty(\mathcal{F}, \delta)$ is a non-increasing function of δ. Since the upper end of the above chain of inequalities is independent of k we have:

$$
\begin{aligned}
E^p \sup_{F\in\mathcal{F}} |\Lambda_F| &= \lim_{k\to\infty} E^p \max_{F\in\mathcal{G}_k} |\Lambda_F| &(134)\\
&\leq E^p|\Lambda_{F(0)}| + \frac{32}{1-2p}\int_0^1 \sqrt{\ln N_\infty(\mathcal{F},\delta)}d\delta, &(135)
\end{aligned}
$$

where the equality follows from the bounded convergence theorem (as argued previously, we always have $\sup_{F\in\mathcal{F}}|\Lambda_F| = \lim_{k\to\infty}\max_{F\in\mathcal{G}_k}|\Lambda_F|$ and it clearly follows by the definition of Λ_F that $\sup_{F\in\mathcal{F}}|\Lambda_F| \leq n^{1/2}2\frac{1-p}{1-2p}$ a.s., as the L_F and $\hat{L}_F$ are clearly bounded by $n\frac{1-p}{1-2p}$ each). Consequently, the proof is completed by establishing

Lemma 23

$$
E^p|\Lambda_{F(0)}| \leq \frac{2}{1-2p}, \tag{136}
$$

the proof of which is given in the appendix. □

Proof of Theorem 20: Note first that for all n and $y^n\in\{0,1\}^n$ we have, using the predictor of equation (108)

$$
\begin{aligned}
\hat{L}_P(y^n) - \inf_{F\in\mathcal{F}}\hat{L}_F(y^n) &= \frac{1}{1-2p}\{L_P(y^n) - \inf_{F\in\mathcal{F}} L_F(y^n)\} &(137)\\
&\leq \frac{\sqrt{2}(1+2\delta)}{\delta(1-2\delta)(1-2p)(\sqrt{2}-1)}\sqrt{n}\int_0^1 \sqrt{\ln N_\infty(\mathcal{F},\varepsilon)}d\varepsilon, &(138)
\end{aligned}
$$

where the equality follows from (82) and the inequality follows from Theorem 18. Therefore,

$$
\begin{aligned}
L_P(y^n,x^n) - \inf_{F\in\mathcal{F}} L_F(y^n,x^n) &= (L_P(y^n,x^n) - \hat{L}_P(y^n)) &(139)\\
&\quad +(\inf_{F\in\mathcal{F}}\hat{L}_F(y^n) - \inf_{F\in\mathcal{F}} L_F(y^n,x^n)) &(140)\\
&\quad +(\hat{L}_P(y^n) - \inf_{F\in\mathcal{F}}\hat{L}_F(y^n)) &(141)\\
&\leq (L_P(y^n,x^n) - \hat{L}_P(y^n)) &(142)\\
&\quad + \sup_{F\in\mathcal{F}}|L_F(y^n,x^n) - \hat{L}_F(y^n)| &(143)\\
&\quad +(\hat{L}_P(y^n) - \inf_{F\in\mathcal{F}}\hat{L}_F(y^n)) &(144)\\
&\leq (L_P(y^n,x^n) - \hat{L}_P(y^n)) &(145)\\
&\quad + \sup_{F\in\mathcal{F}}|L_F(y^n,x^n) - \hat{L}_F(y^n)| &(146)\\
&\quad + \frac{\sqrt{2}(1+2\delta)\int_0^1\sqrt{\ln N_\infty(\mathcal{F},\varepsilon)}d\varepsilon}{\delta(1-2\delta)(1-2p)(\sqrt{2}-1)}\sqrt{n}. &(147)
\end{aligned}
$$

Now, for any $\varepsilon>0$, if we take

$$
K \stackrel{\text{def}}{=} \max\left\{-8(1+(\frac{1-p}{1-2p})^2)^2\ln(\varepsilon/4), \frac{16}{\varepsilon(1-2p)}[1+16\int_0^1\sqrt{\ln N_\infty(\mathcal{F},\delta)}d\delta], \right. \tag{148}
$$

$$\frac{2\sqrt{2}(1+2\delta)}{\delta(1-2\delta)(1-2p)(\sqrt{2}-1)} \int_0^1 \sqrt{\ln N_\infty(\mathcal{F},\varepsilon)}d\varepsilon\Bigg\}, \tag{149}$$

then we have

$$\begin{aligned}
& \Pr^p(L_P(y^n,x^n) - \inf_{F\in\mathcal{F}} L_F(y^n,x^n) \geq K\sqrt{n}) & (150)\\
\leq\; & \Pr^p(\{(L_P(y^n,x^n) - \hat{L}_P(y^n)) + & (151)\\
& \sup_{F\in\mathcal{F}} |L_F(y^n,x^n) - \hat{L}_F(y^n)| + & (152)\\
& \frac{\sqrt{2}(1+2\delta)}{\delta(1-2\delta)(1-2p)(\sqrt{2}-1)}\sqrt{n}\int_0^1 \sqrt{\ln N_\infty(\mathcal{F},\varepsilon)}d\varepsilon \geq K\sqrt{n}\}) & (153)\\
=\; & \Pr^p(\{(L_P(y^n,x^n) - \hat{L}_P(y^n)) + & (154)\\
& \sup_{F\in\mathcal{F}} |L_F(y^n,x^n) - \hat{L}_F(y^n)| \geq & (155)\\
& [K - \frac{\sqrt{2}(1+2\delta)}{\delta(1-2\delta)(1-2p)(\sqrt{2}-1)}\int_0^1 \sqrt{\ln N_\infty(\mathcal{F},\varepsilon)}d\varepsilon]\sqrt{n}\}) & (156)\\
\leq\; & \Pr^p((L_P(y^n,x^n) - \hat{L}_P(y^n)) + & (157)\\
& \sup_{F\in\mathcal{F}} |L_F(y^n,x^n) - \hat{L}_F(y^n)| \geq \frac{K}{2}\sqrt{n}) & (158)\\
\leq\; & \Pr^p(\{L_P(y^n,x^n) - \hat{L}_P(y^n) \geq \frac{K}{4}\sqrt{n}\} \cup & (159)\\
& \{\sup_{F\in\mathcal{F}} |L_F(y^n,x^n) - \hat{L}_F(y^n)| \geq \frac{K}{4}\sqrt{n}\}) & (160)\\
\leq\; & \Pr^p(L_P(y^n,x^n) - \hat{L}_P(y^n) \geq \frac{K}{4}\sqrt{n}) + & (161)\\
& \Pr^p(\sup_{F\in\mathcal{F}} |L_F(y^n,x^n) - \hat{L}_F(y^n)| \geq \frac{K}{4}\sqrt{n}) & (162)\\
\leq\; & \Pr^p(n^{-1/2}(L_P(y^n,x^n) - \hat{L}_P(y^n)) \geq \frac{K}{4}) + & (163)\\
& \frac{4E^p \sup_{F\in\mathcal{F}} |n^{-1/2}(L_F(y^n,x^n) - \hat{L}_F(y^n))|}{K} & (164)\\
\leq\; & 2\exp(-\frac{2(K/4)^2}{(1+(\frac{1-p}{1-2p})^2)^2}) + & (165)\\
& \frac{1}{K}\frac{8}{1-2p}[1+16\int_0^1 \sqrt{\ln N_\infty(\mathcal{F},\delta)}d\delta] & (166)\\
\leq\; & \frac{\varepsilon}{2} + \frac{\varepsilon}{2}. & (167)
\end{aligned}$$

Inequality (153) follows from (147). Inequality (157) follows since, by the definition of K, we have

$$K \geq \frac{2\sqrt{2}(1+2\delta)}{\delta(1-2\delta)(1-2p)(\sqrt{2}-1)}\sqrt{n}\int_0^1 \sqrt{\ln N_\infty(\mathcal{F},\varepsilon)}d\varepsilon \tag{168}$$

which is equivalent to

$$K - \frac{\sqrt{2}(1+2\delta)}{\delta(1-2\delta)(1-2p)(\sqrt{2}-1)}\sqrt{n}\int_0^1 \sqrt{\ln N_\infty(\mathcal{F},\varepsilon)}d\varepsilon \geq K/2. \tag{169}$$

Inequality (163) follows from Chebychev's inequality. Inequality (165) follows by an upper bound on the first term of (163) which follows from the fact that, for any $\alpha > 0$ we have

$$\Pr^p\left\{n^{-1/2}(L_P(y^n,x^n) - \hat{L}_P(y^n)) \geq \alpha\right\} \leq 2\exp\left(-\frac{2\alpha^2}{(1+(\frac{1-p}{1-2p})^2)^2}\right), \tag{170}$$

which is obtained similarly to inequality (C.4) (see proof of Lemma 22 in the appendix). The upper bound on the second term of (163) follows from Theorem 21. Finally, inequality (167) follows from our definition of K.□

E. A Large Deviations Asymptotic Performance Bound for General Expert Classes

For a given loss function L, expert class $\mathcal{F}$ and $\varepsilon > 0$ we define

$$I_L^p(\mathcal{F},\epsilon) \stackrel{\text{def}}{=} -\inf_P \limsup_{n\to\infty}\frac{1}{n}\log \max_{x^n\in\{0,1\}^n} \Pr\{\frac{1}{n}[L_P(y^n,x^n) - \inf_{\mathcal{F}} L_F(y^n,x^n)] > \epsilon\}, \tag{171}$$

where the infimum is taken over all possible predictors. In words, $I_L^p(\mathcal{F},\epsilon)$ denotes the best exponential decay rate achievable by any predictor, uniformly for all individual sequences, for a particular expert class $\mathcal{F}$ under a particular loss function L. Equipped with the notation related to covering numbers introduced in subsection D., we can now prove the following:

Theorem 24 *Let the loss function L be such that the hypotheses of Theorem 8 hold, or be the absolute loss function. Suppose further that L_0 and L_1 are bounded by some $B > 0$ (if L is the absolute loss we can take $B = 1$). Then, for any expert class $\mathcal{F}$ satisfying*

$$N_\infty(\mathcal{F},\delta) < \infty \quad \forall\delta > 0, \tag{172}$$

we have

$$I_L^p(\mathcal{F},\epsilon) \geq \frac{\varepsilon^2 p(1-2p)^2}{16B^2(1-p)(4p^2-3p+1)} \quad \forall\varepsilon > 0. \tag{173}$$

Proof: Suppose first that $\mathcal{F}$ is finite. It then follows from Corollary 11 (respectively, Corollary 15 for the absolute loss function) by taking the logarithm, dividing by n, and taking the limit supremum, that

$$I_L^p(\mathcal{F},\varepsilon) \geq \frac{\alpha^2\varepsilon^2 p(1-2p)^2}{4B^2(1-p)(4p^2-3p+1)} \tag{174}$$

for all $0 \leq \alpha < 1/2$. Letting $\alpha \nearrow 1/2$ on the right-hand side of (174) establishes the assertion of the theorem for finite classes of experts.

Turning to infinite classes, we note first that, by the hypothesis of the Theorem, L_0 and L_1 are differentiable and bounded (by B) on $[0,1]$. This is more than enough to guarantee that the Lipschitz property is satisfied by L, namely, that for some $c > 0$ we have $|L(x_t, F_t) - L(x_t, G_t)| \leq c|F_t - G_t|$ for all $x_t \in \{0,1\}$ and $F_t, G_t \in [0,1]$. For any $\delta > 0$, let $\mathcal{F}_\delta$ be some δ-cover of $\mathcal{F}$ of finite cardinality (clearly one exists by (172)). Let P^δ be the predictor of Corollary 11 (respectively 15) tailored for the class $\mathcal{F}_\delta$. Clearly, the Lipschitz condition implies $\inf_{\mathcal{F}_\delta} L_F(y^n, x^n) \leq \inf_{\mathcal{F}} L_F(y^n, x^n) + n \cdot c \cdot \delta$ and therefore we have for all n and $x^n \in \{0,1\}^n$

$$\Pr\left\{\frac{1}{n}[L_{P^\delta}(y^n, x^n) - \inf_{\mathcal{F}} L_F(y^n, x^n)] > \epsilon\right\} \tag{175}$$

$$\leq \Pr\left\{\frac{1}{n}[L_{P^\delta}(y^n, x^n) - \inf_{\mathcal{F}_\delta} L_F(y^n, x^n)] > \epsilon - c\delta\right\}. \tag{176}$$

Consequently,

$$-I_L^p(\mathcal{F}, \epsilon) \tag{177}$$

$$\leq \limsup_{n\to\infty} \frac{1}{n} \log \max_{x^n \in \{0,1\}^n} \Pr\left\{\frac{1}{n}[L_{P^\delta}(y^n, x^n) - \inf_{\mathcal{F}} L_F(y^n, x^n)] > \epsilon\right\} \tag{178}$$

$$\leq \limsup_{n\to\infty} \frac{1}{n} \log \max_{x^n \in \{0,1\}^n} \Pr\left\{\frac{1}{n}[L_{P^\delta}(y^n, x^n) - \inf_{\mathcal{F}_\delta} L_F(y^n, x^n)] > \epsilon - c\delta\right\} \tag{179}$$

$$\leq -\frac{(\varepsilon - c\delta)^2 p(1-2p)^2}{16B^2(1-p)(4p^2 - 3p + 1)}, \tag{180}$$

where the first inequality follows from the definition of $I_L^p(\mathcal{F}, \epsilon)$, the second follows from (176), and the last follows since the assertion of the Theorem was already established above for finite sets. Finally, since $\delta > 0$ is arbitrary and $I_L^p(\mathcal{F}, \epsilon)$ independent of δ, then taking the limit in the right-hand side of (180) as $\delta \searrow 0$ gives

$$I_L^p(\mathcal{F}, \epsilon) \geq \frac{\varepsilon^2 p(1-2p)^2}{16B^2(1-p)(4p^2 - 3p + 1)}. \square \tag{181}$$

It should be noted that the above result is a worst-case result in a very strong sense. The right-hand side of (173) does not depend $\mathcal{F}$. In other words, $I_L^p(\mathcal{F}, \varepsilon)$ is uniformly lower bounded, for all $\mathcal{F}$ satisfying (172). To see that this exponential scale is meaningful (that is, $I_L^p(\mathcal{F}, \varepsilon) < \infty$), it is straightforward to show that for any non-degenerate expert class and loss function, and sufficiently small $\epsilon > 0$:

$$I_L^p(\mathcal{F}, \epsilon) \leq \log(1/p) \tag{182}$$

(roughly speaking, because a "bad" event contains at least one noise sequence of probability at least p^n).

4 Real Valued Noise: The Arbitrarily Varying, Memoryless, Additive Noise Channel

As was established for the general case in Section 2, and as was demonstrated for the case of binary-valued noise in the previous section, the prediction problem for the noisy setting essentially reduces to finding prediction schemes which efficiently compete with the given expert class in terms of the loss estimators. We dedicate this section to another demonstration of this idea for the case where the individual sequence $x = (x_1, x_2, \ldots) \in \{0,1\}^\infty$ is passed through a channel with zero-mean, additive, memoryless, arbitrarily varying noise. Specifically, the noisy observation accessible to the predictor, $y \in \{0,1\}^\infty$ is given by $y_t = x_t + r_t$, where here $r = \{r_t, t \geq 1\}$ is a zero-mean, memoryless, arbitrarily varying process. More formally, we assume that for every n, the p.d.f. governing r^n is of the form:

$$f(r^n|s^n) = \prod_{i=1}^{n} f(r_i|s_i), \tag{183}$$

where $s^n \in \mathcal{S}^n$ is some unknown *arbitrary* sequence of states, and $\mathcal{S}$, is any state-space such that for all $\sigma \in \mathcal{S}$ we have

$$\int_{-\infty}^{\infty} r \cdot f(r|\sigma) dr = 0. \tag{184}$$

We will let $\Pr^s$ and E^s denote the probability measure under which the noise is governed by the above described law for the state-sequence $s \in \mathcal{S}^\infty$ and the expectation w.r.t. this measure, respectively. Note that in typical real-life situations, where there is a binary clean sequence corrupted by additive noise (e.g. communication systems), the alphabet is more commonly given by $\{-A, A\}$ for some $A > 0$ (e.g. dictated by a transmission power constraint). It should be quite clear, however, that any result for the case where the alphabet is $\{0,1\}$ can easily be transformed into a similar result for the alphabet $\{-A, A\}$ by a straightforward shift and rescaling. We work in this section with the $\{0,1\}$-alphabet so as to manifest the fact that this setting is no more than a special case of the general one considered in Section 2, and to elucidate the direct applicability of its results for this case.

A predictor or expert F, in this setting, is a set of functions $F = \{F_t : \mathbb{R}^{t-1} \to [0,1]\}_{t \geq 1}$. As in the previous noise model, upon observing the $t-1$ noisy bits y^{t-1}, expert F predicts the value $F_t(y^{t-1})$, and the loss incurred by this prediction is $L(x_t, F_t(y^{t-1}))$. Note that this noise model, as that of the previous section, is a special case of that covered in Section 2, for the case where the observation alphabet $A = \mathbb{R}$, as the noise components are independent, and we clearly have $y_t \in \sigma(r_t)$ for all $t \geq 1$. Consequently, the general results of Section 2 can be applied, provided that we obtain upper bounds of the form of equation (6). The following subsection is dedicated to achieving this goal.

A. Competition in Terms of Loss Estimators

As the noise components are additive, with zero-mean, a natural choice for the function $h(\cdot)$ of Section 2, for this case, is the identity function, namely $h(y_t) = y_t$, which is clearly an unbiased estimator for x_t as $Eh(y_t) = Ey_t = E(x_t + r_t) = x_t + Er_t = x_t$. Using this $h(y_t)$ as our estimate for x_t, our cumulative loss estimator (equation (5)) for this case becomes:

$$\hat{L}_F(y^n) = \sum_{t=1}^{n}(1 - y_t)L_0(F_t(y^{t-1})) + y_t L_1(F_t(y^{t-1})). \tag{185}$$

The Absolute Loss Function

For the case of the absolute loss function, the estimator in (185) assumes the form

$$\hat{L}_F(y^n) = \sum_{t=1}^{n}(1 - y_t)F_t(y^{t-1}) + y_t(1 - F_t(y^{t-1})). \tag{186}$$

It turns out that the simple exponential weighting algorithm, traditionally used for the noise-free setting and the binary alphabet, can be effectively employed for the current setting, where the observations are real-valued. The proof of the following result is a straightforward adaptation of the proof of Theorem 1 of [2] and is given in [43, Theorem 4.1].

Theorem 25 *For any finite expert class $\mathcal{F}$, define the predictor P by*

$$P_t(y^{t-1}) \stackrel{\text{def}}{=} \frac{\sum_{F\in\mathcal{F}} e^{-\eta \hat{L_F}(y^{t-1})} F_t(y^{t-1})}{\sum_{F\in\mathcal{F}} e^{-\eta \hat{L_F}(y^{t-1})}}, \tag{187}$$

for $\eta = \sqrt{\frac{\ln|\mathcal{F}|}{n}} \cdot 8$. Then, for any $y^n \in \mathbb{R}^n$,

$$\hat{L}_P(y^n) - \min_{\mathcal{F}} \hat{L}_F(y^n) \leq \frac{\sqrt{n \ln|\mathcal{F}|}}{8} + \sqrt{\frac{\ln|\mathcal{F}|}{n}} \sum_{t=1}^{n}(1 - 2y_t)^2. \tag{188}$$

A strongly sequential predictor can be obtained from this one, similarly to the way one was obtained in Theorem 12, by employing that of Theorem 25 on blocks of exponentially increasing lengths (see Theorem 4.2 of [43] for a full proof):

Theorem 26 *Let*

$$P_t(y^{t-1}) = \frac{\sum_{F\in\mathcal{F}} \exp\{-\eta_{k(t)}\hat{L_F}(y^{t-1})_{2^{k(t)+1}+1}^{t-1}\}F_t(y^{t-1})}{\sum_{F\in\mathcal{F}} \exp\{-\eta_{k(t)}\hat{L_F}(y^{t-1})_{2^{k(t)+1}+1}^{t-1}\}}, \tag{189}$$

where $k(t) = \lfloor \log t - 1 \rfloor$ *and* $\eta_k = 8\sqrt{\frac{\ln|\mathcal{F}|}{2^{k+1}}}$. *Then, for any* $t \geq 1$ *and* $y^t \in \mathbb{R}^t$,

$$\hat{L_P}(y^t) - \min_{\mathcal{F}} \hat{L_F}(y^t) \leq \frac{\sqrt{t\ln|\mathcal{F}|}}{8} \cdot \frac{\sqrt{2}}{\sqrt{2}-1} + \sqrt{\ln|\mathcal{F}|}\sum_{j=0}^{\lfloor \log t\rfloor} \frac{1}{2^j} \sum_{\tau=2^j}^{2^{j+1}-1} (1-2y_\tau)^2. \tag{190}$$

Note that inequality (190) is exactly in the form of (6). The performance of the predictor of Theorem 26 will be assessed in the next subsection by combining (190) with Theorems 4-6.

General Loss Functions

We start with a summary of the results of [14] which are central for our purposes. In particular, the results for the case of an alphabet which is the whole set $[0,1]$ will be of interest to us here. For a loss function $L : [0,1]\times[0,1] \to [0,\infty]$ (note that here we allow the first argument to belong to the whole interval $[0,1]$), the bottom line of sections III-B and IV-A of [14], which will be of interest to us, can be encapsulated in the following theorem.

Theorem 27 ([14]) *Let* L *be a loss function such that* $L_0(0) = L_1(1) = 0$, L_0 *and* L_1 *are three times differentiable in* $(0,1)$, *and* $L_0'(z) > 0$, *and* $L_1'(z) < 0$ *for* $0 < z < 1$. *Assume that the constant* c_L *defined in equation (56) is finite, and* $S(z)$ *defined in (54) is positive for* $0 < z < 1$. *Letting* $g(y,a,b) = (L(y,a) - L(y,b))/c_L$, *assume further that for all* $y, a, b \in [0,1]$ *we have*

$$\frac{\partial^2 g(y,a,b)}{\partial y^2} + \left(\frac{\partial g(y,a,b)}{\partial y}\right)^2 \geq 0. \tag{191}$$

Let P *be the strongly sequential predictor given in Algorithm 3.1. Then* P *is guaranteed to never fail and we have*

$$\forall n \geq 1: \qquad \max_{x^n\in[0,1]^n} \{L_P(x^n) - \min_{F\in\mathcal{F}} L_F(x^n)\} \leq c_L \ln|\mathcal{F}|. \tag{192}$$

For a given $B > 0$, define now the following auxiliary "loss function" $L_B : [0,1]\times[0,1] \to [0,\infty]$:

$$L_B(u,z) = \{1 - [(2B+1)u - B]\} \cdot L_0(z) + [(2B+1)u - B] \cdot L_1(z) + BL_1(0). \tag{193}$$

Further let $S_B(z)$, $R_B(z)$ and $c_L(B)$ be defined as in equations (54), (55) and (56) respectively, with L_{B0} and L_{B1} replacing L_0 and L_1 where $L_{B0}(\cdot) = L_B(0,\cdot)$ and $L_{B1}(\cdot) = L_B(1,\cdot)$.

Before we proceed, let us motivate the introduction of the auxiliary loss function L_B. We recall that our cumulative loss estimator is

$$\hat{L}_F(y^n) = \sum_{t=1}^{n}(1-y_t)L_0(F_t) + y_t L_1(F_t). \tag{194}$$

Note also that the cumulative loss of a predictor under L_B, w.r.t. any sequence $u^n \in [0,1]^n$ is

$$L_{BF}(u^n) = \sum_{t=1}^{n}\{1-[(2B+1)u_t - B]\}\cdot L_0(F_t) + [(2B+1)u_t - B]\cdot L_1(F_t) + BL_1(0). \tag{195}$$

In particular, for any $y^n \in [-B, B+1]^n$ we can let $u_t = \frac{y_t+B}{2B+1}$, for $1 \leq t \leq n$, so that $u^n \in [0,1]^n$ and we have, plugging into (195),

$$L_{BF}(u^n) = (\sum_{t=1}^{n}(1-y_t)L_0(F_t) + y_t L_1(F_t)) + nBL_1(0) = \hat{L}_F(y^n) - nBL_1(0). \tag{196}$$

The following theorem is little more than a specialization of Theorem 27 for the auxiliary loss function L_B. The observations, which are allowed to belong to $[-B, B+1]$, are rescaled and shifted to belong to $[0,1]$ and thereby, to comply with the conditions of Theorem 27. The relation in (196) is then used to obtain an upper bound for the relative loss in terms of the cumulative loss estimators. A full proof can be found in [43] (Theorem 4.5).

Theorem 28 *Let $L : \{0,1\} \times [0,1] \to [0,\infty]$ be a loss function such that $L_1(0) = L_0(1) < \infty$, L_0 and L_1 are three times differentiable in $(0,1)$, and $L_0'(z) > 0$, $L_1'(z) < 0$ for $0 < z < 1$. Assume that $c_L(B)$ is finite and $S(z)$ is positive for $0 < z < 1$. Assume further that for all $u, a, b \in [0,1]$, the function g_B, defined by $g_B(u,a,b) = (L_B(u,a) - L_B(u,b))/c_L(B)$, satisfies*

$$\frac{\partial^2 g_B(u,a,b)}{\partial u^2} + \left(\frac{\partial g_B(u,a,b)}{\partial u}\right)^2 \geq 0. \tag{197}$$

Then there exists a strongly sequential predictor P for which

$$\forall n \geq 1: \quad \max_{y^n \in [-B,B+1]^n}\{\hat{L}_P(y^n) - \min_{F\in\mathcal{F}} \hat{L}_F(y^n)\} \leq c_L(B)\ln|\mathcal{F}|. \tag{198}$$

This predictor is given in the following algorithm (which is guaranteed to never fail so long as $y_t \in [-B, B+1]$ for all $1 \leq t \leq n$):

Initialization: *Set weights $w_{1,i} = 1$, $1 \leq i \leq N$.*

Prediction: *Let $v_{t,i} = w_{t,i}/W_t$, where $W_t = \sum_{i=1}^N w_{t,i}$. At the beginning of trial t, compute for $u = 0$ and $u = 1$ the value*

$$\Delta_B(u) = -c_L(B) \ln \sum_{i=1}^{N} v_{t,i} e^{-L_B(u, F_t^{(i)}(y^{t-1}))/c_L(B)} \tag{199}$$

and predict with any value P_t that satisfies for $u = 0$ and $u = 1$ the condition

$$L_B(u, P_t) \leq \Delta_B(u). \tag{200}$$

If no such P_t exists, the algorithm is said to fail.

Update: *After observing the t-th outcome y_t, let*

$$w_{t+1,i} = w_{t,i} e^{-L_B(\frac{y_t+B}{2B+1}, F_t^{(i)}(y^{t-1}))/c_L(B)}. \tag{201}$$

Note that the predictor of Theorem 28 and its performance bound in terms of the estimated losses are fit for the case $y^n \in [-B, B+1]^n$, i.e., when the noise is uniformly bounded. We would like, however, to relax the boundedness conditions, and seek a predictor whose performance w.r.t. the estimated loss is bounded similarly as in Theorem 26. Namely, we would like that bound to hold for all $y^n \in \mathbb{R}^n$ and to depend only on the sequence itself. As will be evident in subsection B., where the general results of Section 2 will be applied, such predictors can be shown to do well in quite general situations. As it turns out, the predictor of Theorem 28 can be used to obtain such a predictor, essentially by employing the former on consecutive blocks, and letting B grow from block to block at an appropriate rate. If $y_t \notin [-B, B+1]$, it is truncated to the nearest end-point. The rate at which B is allowed to grow depends on the behavior of $c_L(B)$. The following subsection is dedicated to an explicit construction of a predictor of this type and to assessing its performance in terms of the estimated loss. For concreteness, and in order to avoid cumbersome notation, we somewhat limit the generality and consider a specific class of loss functions, with the implication that any other loss function for which $c_L(B) < \infty$ for all $B > 0$ can be handled similarly.

A Concrete Case Study: α-Powered Loss for $1 < \alpha \leq 2$

Consider the loss function given by $L(x_t, F_t) = |x_t - F_t|^\alpha$, so that we have $L_0(z) = z^\alpha$ and $L_1(z) = (1-z)^\alpha$. Straightforward algebra gives

$$c_L(B) \stackrel{\text{def}}{=} \sup_{0<z<1} R_B(z) = R_B(1/2) = \frac{(2B+1)^2\alpha}{2^\alpha(\alpha-1)} \tag{202}$$

and

$$\frac{\partial^2 g_B(u,a,b)}{\partial u^2}+(\frac{\partial g_B(u,a,b)}{\partial u})^2=\left[\frac{2^\alpha(\alpha-1)[(1-a)^\alpha-a^\alpha-(1-b)^\alpha+b^\alpha]}{\alpha(2B+1)}\right]^2\geq 0. \quad (203)$$

Therefore, the conditions of Theorem 28 are satisfied for this case and we have:

Corollary 29 *Let the loss function be the α-powered loss for some $1<\alpha\leq 2$. Then for any given expert set $\mathcal{F}$, there exists a predictor P, given in Theorem 28, for which*

$$\forall n\geq 1: \quad \max_{y^n\in[-B,B+1]^n}\{\hat{L}_P(y^n)-\min_{F\in\mathcal{F}}\hat{L}_F(y^n)\}\leq\frac{(2B+1)^2\alpha}{2^\alpha(\alpha-1)}\ln|\mathcal{F}|. \quad (204)$$

For a given $B\geq 0$, let now $\phi_B:\mathbb{R}\to[-B,B+1]$ be the limiter function defined by

$$\phi_B(x)=\begin{cases} B+1 & \text{if } x>B+1\\ x & \text{if } x\in[-B,B+1]\\ -B & \text{otherwise}\end{cases}\quad . \quad (205)$$

For convenience of the formalism, for any $B<0$, we let ϕ_B be $\equiv 0$. We now consider the following prediction algorithm:

Algorithm 4.1

Initialization: *Set weights $w_{1,i}=1$, $1\leq i\leq N$.*

Prediction: *Let $v_{t,i}=w_{t,i}/W_t$, where $W_t=\sum_{i=1}^N w_{t,i}$. At the beginning of trial t, compute for $u=0$ and $u=1$ the value*

$$\Delta_{B_t}(u)=-c_L(B_t)\ln\sum_{i=1}^N v_{t,i}e^{-L_{B_t}(u,F_t^{(i)}(y^{t-1}))/c_L(B_t)} \quad (206)$$

where $B_t=\frac{\sqrt{\lfloor\log t+1\rfloor}-1}{2}$, $c_L(B)$ is given in (202) and $L_B(\cdot,\cdot)$ is as defined in (193). Predict with any value P_t that satisfies for $u=0$ and $u=1$ the condition

$$L_{B_t}(u,P_t)\leq\Delta_{B_t}(u). \quad (207)$$

If no such P_t exists, the algorithm is said to fail.

Update: *If $\log(t+1)$ is integer, let $w_{t+1,i}=1$ for all $1\leq i\leq N$. Otherwise, after observing the t-th outcome y_t, take $z_t=\phi_{B_t}(y_t)$ and let for each $1\leq i\leq N$*

$$w_{t+1,i}=w_{t,i}e^{-L_{B_t}(\frac{z_t+B_t}{2B_t+1},F_t^{(i)}(y^{t-1}))/c_L(B_t)}. \quad (208)$$

In words, this prediction scheme employs a predictor as that of Corollary 29 on each block, where the block lengths increase by a factor of 2. On the j-th block, (of length 2^j), the predictor of Corollary 29 is used with $B_j = \frac{\sqrt{j+1}-1}{2}$. Since such a predictor can accept only observations in $[-B_j, B_j+1]$, the observed values of y_t in the j-th block are truncated to that interval and only then fed into the predictor. We can now state the following:

Theorem 30 *Let L be the α-powered loss for some $1 < \alpha \leq 2$. Then for any given expert class $\mathcal{F}$ of finite cardinality, the strongly sequential predictor P, given in Algorithm 4.1, is guaranteed to never fail and $\forall y^n \in \mathbb{R}^n$ we have*

$$\hat{L}_P(y^n) - \min_{F\in\mathcal{F}} \hat{L}_F(y^n) \leq \ln|\mathcal{F}|\frac{\alpha}{2^{\alpha+1}(\alpha-1)}(\log n+2)^2 + \sum_{t=1}^{n}|y_t - \phi_{(\frac{\sqrt{\log t}-1}{2})}(y_t)|. \quad (209)$$

Note that the right hand side of (209) is given solely in terms of the observed sequence y^n, and is completely free of any assumptions or restrictions regarding its domain. The proof is a straightforward consequence of the following two lemmas, the proofs of which are deferred to the appendix.

Lemma 31 *Let P be the predictor of Theorem 30. Then $\forall y^n \in \mathbb{R}^n$*

$$\hat{L}_{P(y)}(z^n) - \min_{F\in\mathcal{F}} \hat{L}_F(y^n) \leq \ln|\mathcal{F}|\frac{\alpha}{2^{\alpha+1}(\alpha-1)}(\log n+2)^2 \quad (210)$$

where the t-th component of z^n is given by $z_t = \phi_{B_t}(y_t)$, $B_t = \frac{\sqrt{\lfloor \log t+1\rfloor}-1}{2}$ (as we recall from the algorithm of Theorem 30) and

$$\hat{L}_{P(y)}(z^n) \stackrel{\text{def}}{=} \sum_{t=1}^{n}(1-z_t)L_0(P_t(y^{t-1})) + z_t L_1(P_t(y^{t-1})). \quad (211)$$

Lemma 32 *Under the α-powered loss function, f or any P and all $y^n, z^n \in \mathbb{R}^n$ we have*

$$|\hat{L}_P(y^n) - \hat{L}_{P(y)}(z^n)| \leq \sum_{t=1}^{n}|y_t - z_t|.$$

Proof of Theorem 30: Using the predictor P of Theorem 30, and letting z^n be as in Lemma 31, namely $z_t = \phi_{(\frac{\sqrt{\lfloor \log t+1\rfloor}-1}{2})}(y_t)$, we have $\forall y^n \in \mathbb{R}^n$

$$\begin{aligned}
\hat{L}_P(y^n) - \min_{F\in\mathcal{F}} \hat{L}_F(y^n) &= (\hat{L}_P(y^n) - \hat{L}_{P(y)}(z^n)) + (\hat{L}_{P(y)}(z^n) - \min_{\mathcal{F}} \hat{L}_F(y^n)) \quad (212)\\
&\leq \sum_{t=1}^{n}|y_t - z_t| + \ln|\mathcal{F}|\frac{\alpha}{2^{\alpha+1}(\alpha-1)}(\log n+2)^2\\
&= \sum_{t=1}^{n}|y_t - \phi_{\frac{\sqrt{\lfloor \log t+1\rfloor}-1}{2}}(y_t)| + \ln|\mathcal{F}|\frac{\alpha}{2^{\alpha+1}(\alpha-1)}(\log n+2)^2\\
&\leq \sum_{t=1}^{n}|y_t - \phi_{\frac{\sqrt{\log t}-1}{2}}(y_t)| + \ln|\mathcal{F}|\frac{\alpha}{2^{\alpha+1}(\alpha-1)}(\log n+2)^2,
\end{aligned}$$

where the first inequality is obtained by upper bounding the first term of (212) with the bound of Lemma 32 and the second term with that of Lemma 31. The last inequality follows by replacing $\lfloor \log t + 1 \rfloor$ with $\log t$ and noting that, by the definition of $\phi_B(\cdot)$, $|y_t - \phi_{b_1}(y_t)| \leq |y_t - \phi_{b_2}(y_t)|$ whenever $b_1 \geq b_2$.□

B. Applying Theorems 4-6

Having dedicated the previous subsection to obtaining bounds in the form of equation (6), we can now harvest the following corollaries, which are obtained, similarly as in the previously treated binary-valued noise case, by combining the upper bounds in the form of (6) with Theorems 4-6.

For concreteness, we derive next corollaries for the case of the absolute loss function which was treated in subsection A.. In particular, the role of inequality (6) will be played by inequality (190) for applying Theorems 4-6. The case of, say, the α-powered loss function, is handled similarly by letting inequality (209) play the role of inequality (6) and applying the same theorems (the reader is referred to subsection 4.3.2 of [43] for results which are specialized to the case of the α-powered loss function).

For any i, we let $\mathcal{M}_i(B) \stackrel{\text{def}}{=} \{\sigma \in \mathcal{S} : \int_{-\infty}^{\infty} r^i f(r|\sigma) dr \leq B\}$. In words, for any $\sigma \in \mathcal{M}_i(B)$, a random variable governed by the p.d.f. $f(\cdot|\sigma)$ is one with zero mean and an i-th moment bounded by B. We can now state the following.

Corollary 33 *Let L be the absolute loss function. Let P be the strongly sequential predictor of equation (189). Then, for any $B > 0$, all $s \in \mathcal{M}_2(B)^\infty$, and all $t \geq 1$ we have*

$$\max_{x_t \in \{0,1\}^t} \{E^s L_P(y^t, x^t) - \min_{F \in \mathcal{F}} E^s L_F(y^t, x^t)\} \leq \frac{\sqrt{t \ln |\mathcal{F}|}}{8} \cdot \frac{\sqrt{2}}{\sqrt{2}-1} + \sqrt{\ln |\mathcal{F}|}(4B+1)\lfloor \log t + 1 \rfloor. \tag{213}$$

Proof: Combining inequality (190) with Theorem 4, we need merely establish

$$E^s\{\frac{\sqrt{t \ln |\mathcal{F}|}}{8} \cdot \frac{\sqrt{2}}{\sqrt{2}-1} + \sqrt{\ln |\mathcal{F}|} \sum_{j=0}^{\lfloor \log t \rfloor} \frac{1}{2^j} \sum_{\tau=2^j}^{2^{j+1}-1} (1 - 2y_\tau)^2\} \tag{214}$$

$$\leq \frac{\sqrt{t \ln |\mathcal{F}|}}{8} \cdot \frac{\sqrt{2}}{\sqrt{2}-1} + \sqrt{\ln |\mathcal{F}|}(4B+1)\lfloor \log t + 1 \rfloor, \tag{215}$$

or

$$E^s\{\sum_{j=0}^{\lfloor \log t \rfloor} \frac{1}{2^j} \sum_{\tau=2^j}^{2^{j+1}-1} (1 - 2y_\tau)^2\} \leq (4B+1)\lfloor \log t + 1 \rfloor. \tag{216}$$

But, for any $s \in \mathcal{M}_2(B)^\infty$ and $x^t \in \{0,1\}^t$ we have

$$\begin{aligned}
E^s\{\sum_{j=0}^{\lfloor \log t \rfloor} \frac{1}{2^j} \sum_{\tau=2^j}^{2^{j+1}-1} (1-2y_\tau)^2\} &= \sum_{j=0}^{\lfloor \log t \rfloor} \frac{1}{2^j} \sum_{\tau=2^j}^{2^{j+1}-1} E^s(1-2y_\tau)^2 && (217)\\
&\le \sum_{j=0}^{\lfloor \log t \rfloor} \frac{1}{2^j} \sum_{\tau=2^j}^{2^{j+1}-1} (4B+1) && (218)\\
&= \sum_{j=0}^{\lfloor \log t \rfloor} \frac{1}{2^j} \cdot (4B+1) \cdot 2^j && (219)\\
&= (4B+1)\lfloor \log t + 1 \rfloor, && (220)
\end{aligned}$$

where inequality (218) follows since $E^s(1-2y_\tau)^2 \le (4B+1)$. To see this, note that y_t is either $1+r_t$ or r_t according to as $x_t = 1$ or $x_t = 0$. Accordingly, $(2y_t - 1)^2$ is either $4r_t^2 + 4r_t + 1$ or $4r_t^2 - 4r_t + 1$. But clearly, for any $s \in \mathcal{M}_2(B)^\infty$, $E^s r_t^2 \le B$ and $E^s r_t = 0$. The rest is self-evident.□

As will be established in subsection C., this upper bound is of tight asymptotic order in t, namely, one also has a lower bound of order $\sqrt{t}$.

To present the next result, we now let $\mathcal{S}_B \stackrel{\text{def}}{=} \{\sigma \in \mathcal{S} : \int_{-B}^{B} f(r|\sigma)dr = 1\}$. Note that for any $B > 0$ and $s \in \mathcal{S}_B^\infty$, $\{r_t, t \ge 1\}$ is a process of independent components having zero mean, all restricted to $[-B, B]$, $\Pr^s$- almost surely. The following corollary assesses the asymptotic almost sure performance of the predictor of Corollary 33.

Corollary 34 *Let L be the absolute loss function. For a given expert class $\mathcal{F}$ of finite cardinality, let the predictor P be that of equation (189). Then, for any $B > 0$ and all $s \in \mathcal{S}_B^\infty$ we have*

$$\limsup_{n\to\infty} \frac{[L_P(y^n, x^n) - \inf_{F\in\mathcal{F}} L_F(y^n, x^n)]}{\sqrt{n \log\log n}} \le 4\sqrt{2}(B+1) \qquad \Pr^s - a.s. \tag{221}$$

Proof: Noting that $\Pr^s$-a.s. $|h(y_t)| = |y_t| = |x_t + r_t| \le x_t + |r_t| \le 1 + B$ for all $t \ge 1$ and $s \in \mathcal{S}_B^\infty$, and combining inequality (190) with Theorem 4, we need merely establish that

$$\frac{\sqrt{n \ln|\mathcal{F}|}}{8} \cdot \frac{\sqrt{2}}{\sqrt{2}-1} + \sqrt{\ln|\mathcal{F}|} \sum_{j=0}^{\lfloor \log n \rfloor} \frac{1}{2^j} \sum_{\tau=2^j}^{2^{j+1}-1} (1-2y_\tau)^2 = o(\sqrt{n \log\log n}) \quad \Pr^s - a.s., \tag{222}$$

or simply

$$\sum_{j=0}^{\lfloor \log n \rfloor} \frac{1}{2^j} \sum_{\tau=2^j}^{2^{j+1}-1} (1-2y_\tau)^2 = o(\sqrt{n \log\log n}) \quad \Pr^s - a.s. \tag{223}$$

But, for any $s \in \mathcal{S}_B^\infty$, $x \in \{0,1\}^\infty$ and $n \geq 1$, it follows from the definition of $\mathcal{S}_B$ that

$$\sum_{j=0}^{\lfloor \log n \rfloor} \frac{1}{2^j} \sum_{\tau=2^j}^{2^{j+1}-1} (1-2y_\tau)^2 \leq \sum_{j=0}^{\lfloor \log n \rfloor} \frac{1}{2^j} \sum_{\tau=2^j}^{2^{j+1}-1} 4B^2 \tag{224}$$

$$= 4B^2 \lfloor \log n + 1 \rfloor \quad \Pr^s - a.s., \tag{225}$$

which is more than enough to imply (223).□

Finally, by combining Theorem 6 with inequality (190), we next derive the following large deviations asymptotic upper bound. For this let $\mathcal{D}(\beta, c) \stackrel{\text{def}}{=} \{\sigma \in \mathcal{S} : \int_{-\infty}^{\infty} f(r|\sigma) r^2 dr \leq \beta^2, \int_{-\infty}^{\infty} f(r|\sigma)|r|^k dr \leq \frac{k!}{2} \int_{-\infty}^{\infty} f(r|\sigma) r^2 dr \cdot c^{k-2}, k > 2\}$.

Corollary 35 *Let L be the absolute loss function. For a given finite expert class $\mathcal{F}$, let P be the predictor of Theorem 26. Then for any $\beta > 0$, $c > 0$, $s \in \mathcal{D}(\beta, c)^\infty$, and $\varepsilon > 0$ we have:*

$$\limsup_{n\to\infty} \frac{1}{n} \log \max_{x^n \in \{0,1\}^n} \Pr^s \left\{ \frac{1}{n}[L_P(y^n, x^n) - \min_{\mathcal{F}} L_F(y^n, x^n)] > \epsilon \right\} \leq -\frac{\varepsilon^2}{32(\beta^2 + c\varepsilon/2)}. \tag{226}$$

Note that the moment condition imposed in this corollary is relatively mild. In particular, it allows for the important special case where the noise components are Gaussian (say, of arbitrarily varying but bounded variances). This is an improvement over previous results, (cf. [43, Subsection 4.3]) which required boundedness of the noise components or rather restrictive assumptions on the tails of their distributions.

Proof: By the definition of $\mathcal{D}(\beta, c)$, a direct application of Theorem 6, with $g_n(y^n)$ denoting the right-hand side of inequality (190), gives us for all $0 \leq \alpha < 1/2$

$$\max_{x^n \in \{0,1\}^n} \Pr^s \left\{ \frac{1}{n}[L_P(y^n, x^n) - \inf_{F \in \mathcal{F}} L_F(y^n, x^n)] \geq \varepsilon \right\} \tag{227}$$

$$\leq 2(|\mathcal{F}|+1) \exp\{-n \frac{\alpha^2 \varepsilon^2}{8B^2(\beta^2 + c\alpha\varepsilon/B)}\} \tag{228}$$

$$+ \max_{x^n \in \{0,1\}^n} \Pr^s \{ \frac{g_n(y^n)}{n} \geq (1-2\alpha)\varepsilon \}. \tag{229}$$

Taking logarithms, dividing by n, taking the limit supremum, and letting $\alpha \nearrow 1/2$ gives

$$\limsup_{n\to\infty} \frac{1}{n} \log \max_{x^n \in \{0,1\}^n} \Pr^s \{ \frac{1}{n}[L_P(y^n, x^n) - \min_{\mathcal{F}} L_F(y^n, x^n)] > \epsilon \} \tag{230}$$

$$\leq \max \left\{ -\frac{\varepsilon^2}{32(\beta^2 + c\varepsilon/2)}, \lim_{\alpha \to \frac{1}{2}^-} \limsup_{n\to\infty} \frac{1}{n} \log \max_{x^n \in \{0,1\}^n} \Pr^s \{ \frac{g_n(y^n)}{n} \geq (1-2\alpha)\varepsilon \} \right\}. \tag{231}$$

Consequently, it will be more than enough to establish for all $0 \le \alpha < 1/2$

$$\limsup_{n\to\infty} \frac{1}{n} \log \max_{x^n \in \{0,1\}^n} \Pr^s\{\frac{g_n(y^n)}{n} \ge (1-2\alpha)\varepsilon\} = -\infty, \tag{232}$$

which is deferred to the appendix.□

Remark: It is important to note that while the right-hand sides of (213), (221) and (226) in Corollaries 33, 34 and 35 may depend on parameters which limit the class of possible noise sources in some ways (B, β, c, etc.), the predictor itself does not.

C. Worst-Case Relative Expected Loss Asymptotics

Similarly to the case of binary noise, we define the *worst-case relative expected loss* of a predictor P by

$$R_n^s(P, \mathcal{F}) \stackrel{\text{def}}{=} \max_{x^n \in \{0,1\}^n} (E^s L_P(y^n, x^n) - \inf_{F \in \mathcal{F}} E^s L_F(y^n, x^n)). \tag{233}$$

The *minimax relative expected loss* achievable by any predictor with respect to a particular expert class $\mathcal{F}$ is defined by

$$V_n^s(\mathcal{F}) \stackrel{\text{def}}{=} \inf_P R_n^s(P, \mathcal{F}). \tag{234}$$

Finally, we let

$$V^s(N, n) \stackrel{\text{def}}{=} \sup_{\mathcal{F}: |\mathcal{F}| = N} V_n^s(\mathcal{F}). \tag{235}$$

The following gives the exact asymptotic order of $V^s(N, n)$ for a large class of loss functions and noise processes.

Theorem 36 *1. Let the loss function be such that the assumptions of Theorem 28 hold for a given $B > 0$. Assume further that $L_0''(z) \ge 0$ and $L_1''(z) \ge 0$ for $0 < z < 1$. We then have*

$$V^s(N, n) = \theta(\log N) \quad \forall s \in \mathcal{S}_B^\infty. \tag{236}$$

2. Under the absolute loss, for any $B > 0$ and all $s \in \mathcal{M}_2(B)^\infty$ we have

$$V^s(N, n) = \theta(\sqrt{n \log N}) \quad \forall s \in \mathcal{M}_2(B)^\infty. \tag{237}$$

Proof: The proof is a rerun of the argumentation used in the proof of Theorem 17. The upper bound for the first item follows by combining Theorem 28 (taking the role of equation (6)) with Theorem 4. Repeating now the argument given in the proof of Theorem 17 for the asymptotic lower-bound, one concludes that for any $s \in \mathcal{S}^\infty$, n and N we have

$V^s(N,n) \geq V(N,n)$. Consequently, the fact that $V(N,n) = \Omega(\log N)$ for this case (item 1 of Theorem 16), implies $V^s(N,n) = \Omega(\log N)$.

The upper bound for the second item follows from Theorem 33 and the lower bound follows, as the lower bound for the first item, by the facts that for any $s \in \mathcal{S}^\infty$, n and N we have $V^s(N,n) \geq V(N,n)$ and that, for the absolute loss, we have $V(N,n) = \Omega(\sqrt{n \log N})$ (item 2 of Theorem 16).□

5 Conclusion

In this paper, the traditional problem of prediction of individual binary sequences relative to a set of experts was revisited for the case where the predictors access a noisy observation of the past individual sequence. A general approach for obtaining efficient prediction schemes in this noisy setting was presented. The merit of this approach was illustrated by using the general guidelines that it suggests for the treatment of two important representative special cases, corresponding to two different types of noise processes corrupting the data.

Appendix A

Notation for Asymptotics: In this paper, we adopt the notation and definitions used by computer scientists. Namely, for non-negative functions f and g, writing $f(n,N) = O(g(n,N))$ is synonymous to: $\exists C > 0$ such that for all sufficiently large n, N, we have $f(n,N) \leq C \cdot g(n,N)$. Similarly, $f(n,N) = \Omega(g(n,N))$ is synonymous to: $\exists c > 0$ such that for all sufficiently large n, N, we have $f(n,N) \geq c \cdot g(n,N)$. Finally, $f(n,N) = \theta(g(n,N))$ if and only if we have both $f(n,N) = \Omega(g(n,N))$ *and* $f(n,N) = O(g(n,N))$.

Appendix B

Proofs of Lemmas 1 - 3: Throughout this section, we fix a predictor F and $x \in \{0,1\}^\infty$. By the definitions of L_F (equation (3)) and $\hat{L}_F$ (equation (5)) we have

$$\Delta_F(y^n, x^n) \stackrel{\text{def}}{=} L_F(y^n, x^n) - \hat{L}_F(y^n) \tag{B.1}$$

$$= \sum_{t=1}^{n} L(x_t, F_t(y^{t-1})) \tag{B.2}$$

$$- \sum_{t=1}^{n} (1 - h(y_t)) L_0(F_t(y^{t-1})) + h(y_t) L_1(F_t(y^{t-1})) \tag{B.3}$$

$$
\begin{aligned}
&= \sum_{t=1}^{n}(1-x_t)L_0(F_t(y^{t-1}))+x_tL_1(F_t(y^{t-1})) && \text{(B.4)}\\
&\quad -\sum_{t=1}^{n}(1-h(y_t))L_0(F_t(y^{t-1}))+h(y_t)L_1(F_t(y^{t-1})) && \text{(B.5)}\\
&= \sum_{t=1}^{n}(h(y_t)-x_t)L_0(F_t(y^{t-1}))+ && \text{(B.6)}\\
&\quad \sum_{t=1}^{n}(x_t-h(y_t))L_1(F_t(y^{t-1})). && \text{(B.7)}
\end{aligned}
$$

Letting $m_n \stackrel{\text{def}}{=} \sum_{t=1}^{n}(h(y_t)-x_t)L_0(F_t(y^{t-1}))$, we have for all $n \geq 1$

$$
\begin{aligned}
E(m_{n+1}|\mathcal{F}_n) &= E(\sum_{t=1}^{n+1}(h(y_t)-x_t)L_0(F_t(y^{t-1}))|\mathcal{F}_n) && \text{(B.8)}\\
&= E((h(y_{n+1})-x_{n+1})L_0(F_{n+1}(y^n))|\mathcal{F}_n)+ && \text{(B.9)}\\
&\quad E(\sum_{t=1}^{n}(h(y_t)-x_t)L_0(F_t(y^{t-1}))|\mathcal{F}_n) && \text{(B.10)}\\
&= L_0(F_{n+1}(y^n))E((h(y_{n+1})-x_{n+1})|\mathcal{F}_n)+ && \text{(B.11)}\\
&\quad \sum_{t=1}^{n}(h(y_t)-x_t)L_0(F_t(y^{t-1})) && \text{(B.12)}\\
&= L_0(F_{n+1}(y^n))E((h(y_{n+1})-x_{n+1}))+ && \text{(B.13)}\\
&\quad \sum_{t=1}^{n}(h(y_t)-x_t)L_0(F_t(y^{t-1})) && \text{(B.14)}\\
&= 0+m_n, && \text{(B.15)}
\end{aligned}
$$

where the last equation follows from the fact that $h(y_{n+1})$ is an unbiased estimator for x_{n+1}, 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 $h(y_{n+1})-x_{n+1}$ is $\sigma(r_{n+1})$-measurable). The equality before that follows from the measurability of $L_0(F_{n+1}(y^n))$ and of $\sum_{t=1}^{n}(h(y_t)-x_t)L_0(F_t(y^{t-1}))$ 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 (B.7), which, in turn, establishes the fact that $\{\Delta_F(y^n,x^n),\mathcal{F}_n\}_{n\geq 0}$ is a zero-mean martingale. In particular, this implies $E\Delta_F(y^n,x^n)=0$, which is precisely Lemma 1.

Proof of Lemma 2: By Equation (B.7) and the triangle inequality we have almost surely (actually, for all sample paths):

$$
\limsup_{n\to\infty}\frac{|\Delta_F(y^n,x^n)|}{\sqrt{n\log\log n}} \leq \limsup_{n\to\infty}\frac{|\sum_{t=1}^{n}(h(y_t)-x_t)L_0(F_t(y^{t-1}))|}{\sqrt{n\log\log n}} \tag{B.16}
$$

$$+\limsup_{n\to\infty}\frac{|\sum_{t=1}^{n}(x_t-h(y_t))L_1(F_t(y^{t-1}))|}{\sqrt{n\log\log n}}. \tag{B.17}$$

Consider the first term on the right hand side of (B.17) and let

$$s_n^2 \stackrel{\text{def}}{=} \sum_{t=1}^{n} E[(h(y_t)-x_t)^2 L_0(F_t(y^{t-1}))^2|\mathcal{F}_{t-1}] \tag{B.18}$$

$$= \sum_{t=1}^{n} L_0(F_t(y^{t-1}))^2 E[(h(y_t)-x_t)^2|\mathcal{F}_{t-1}] \tag{B.19}$$

$$\leq C^2\sum_{t=1}^{n} L_0(F_t(y^{t-1}))^2 \tag{B.20}$$

$$\leq C^2B^2n \quad \text{a.s.}, \tag{B.21}$$

where the last inequalities follow from the hypotheses concerning the almost sure boundedness of $|h(y_t)|$ and L_0. Now, since the magnitude of each of the summands in the first sum on the right-hand side of (B.17) is bounded by BC, the hypotheses of the upper-bound in the martingale analogue of Kolmogorov's Law of the Iterated Logarithm (cf., e.g., Theorem 1 of [32]) are satisfied and we have for all $\delta > 0$

$$\limsup_{n\to\infty}\frac{|\sum_{t=1}^{n}(h(y_t)-x_t)L_0(F_t(y_{t-1}))|}{\sqrt{2C^2B^2(1+\delta)n\log\log n}} \tag{B.22}$$

$$\leq \limsup_{n\to\infty}\frac{|\sum_{t=1}^{n}(h(y_t)-x_t)L_0(F_t(y_{t-1}))|}{\sqrt{2C^2B^2n\log\log C^2B^2n}} \tag{B.23}$$

$$\leq \limsup_{n\to\infty}\frac{|\sum_{t=1}^{n}(h(y_t)-x_t)L_0(F_t(y_{t-1}))|}{\sqrt{2s_n^2\log\log s_n^2}} \tag{B.24}$$

$$\leq 1 \quad \text{a.s.}, \tag{B.25}$$

where the second inequality follows by (B.21) and the last by the aforementioned law of the iterated logarithm. We therefore have for all $\delta > 0$

$$\limsup_{n\to\infty}\frac{|\sum_{t=1}^{n}(h(y_t)-x_t)L_0(F_t(y_{t-1}))|}{\sqrt{n\log\log n}} \leq \sqrt{2(1+\delta)}CB \quad \text{a.s.}, \tag{B.26}$$

which, by the arbitrariness of δ gives

$$\limsup_{n\to\infty}\frac{|\sum_{t=1}^{n}(h(y_t)-x_t)L_0(F_t(y_{t-1}))|}{\sqrt{n\log\log n}} \leq \sqrt{2}CB \quad \text{a.s.} \tag{B.27}$$

A similar bound is obtained for the second term on the right-hand side of (B.17), which, combined with (B.17), gives the desired result.□

The proof of Lemma 3 is a straightforward application of the following result:

Theorem 37 (Peña, [22]) *Let $\{d_i, \mathcal{F}_i\}$ be a martingale difference sequence with $E(d_j|\mathcal{F}_{j-1}) = 0$, $E(d_j^2|\mathcal{F}_{j-1}) = \sigma_j^2$, $V_n^2 = \sum_{j=1}^n \sigma_j^2$. Furthermore, assume that $E(|d_j|^k|\mathcal{F}_{j-1}) \leq (k!/2)\sigma_j^2 c^{k-2}$ a.e. for $k > 2$, $0 < c < \infty$. Then, for all $x, y > 0$,*

$$\Pr\left\{\sum_{i=1}^n d_i \geq x, V_n^2 \leq y \text{ for some } n\right\} \leq \exp\{-\frac{x^2}{2(y+cx)}\}. \tag{B.28}$$

Proof of Lemma 3: Setting $d_t = (h(y_t) - x_t)L_0(F_t(y^{t-1}))/B$ we have almost surely

$$E(d_j|\mathcal{F}_{j-1}) = \frac{L_0(F_j(y^{j-1}))}{B} E(h(y_j) - x_j|\mathcal{F}_{j-1}) = 0. \tag{B.29}$$

and

$$\begin{aligned}
\sigma_j^2 &= E(d_j^2|\mathcal{F}_{j-1}) && \text{(B.30)}\\
&= (\frac{L_0(F_j(y^{j-1}))}{B})^2 E((h(y_j) - x_j)^2|\mathcal{F}_{j-1}) && \text{(B.31)}\\
&= (\frac{L_0(F_j(y^{j-1}))}{B})^2 E(h(y_j) - x_j)^2 && \text{(B.32)}\\
&\leq \beta^2, && \text{(B.33)}
\end{aligned}$$

where the last inequality follows as, by the hypotheses of Lemma 3, we have $|L_0/B| \leq 1$ a.s. and $E(h(y_j) - x_j)^2 \leq \beta^2$. Also, for all $k > 2$, we have a.s.

$$\begin{aligned}
E(|d_j|^k|\mathcal{F}_{j-1}) &= |\frac{L_0(F_j(y^{j-1}))}{B}|^k E(|h(y_j) - x_j|^k|\mathcal{F}_{j-1}) && \text{(B.34)}\\
&= |\frac{L_0(F_j(y^{j-1}))}{B}|^k E|h(y_j) - x_j|^k && \text{(B.35)}\\
&\leq (\frac{L_0(F_j(y^{j-1}))}{B})^2 E|h(y_j) - x_j|^k && \text{(B.36)}\\
&\leq \frac{k!}{2}(\frac{L_0(F_j(y^{j-1}))}{B})^2 E(h(y_j) - x_j)^2 c^{k-2} && \text{(B.37)}\\
&= \frac{k!}{2}\sigma_j^2 c^{k-2}, && \text{(B.38)}
\end{aligned}$$

where the last inequality follows by the hypothesis and the inequality preceding it follows from the fact that $|\frac{L_0(F_j(y^{j-1}))}{B}| \leq 1$ a.s. Evidently, the conditions of Theorem 37 are satisfied by our $\{d_i, \mathcal{F}_i\}$ and we have

$$\begin{aligned}
& \Pr\{\frac{1}{n}\sum_{t=1}^n (h(y_t) - x_t)L_0(F_t(y_{t-1})) \geq \varepsilon/2\} && \text{(B.39)}\\
= \ & \Pr\{\sum_{t=1}^n d_t \geq n\varepsilon/(2B)\} && \text{(B.40)}
\end{aligned}$$

$$= \Pr\{\sum_{t=1}^{n} d_t \geq n\varepsilon/(2B), V_n^2 \leq n\beta^2\} \tag{B.41}$$

$$\leq \exp\{-n\frac{\varepsilon^2}{8B^2(\beta^2 + c\varepsilon/B)}\}, \tag{B.42}$$

where the last inequality follows by applying Theorem 37 with $x = n\varepsilon/(2B)$ and $y = n\beta^2$. The equality preceding it follows from the fact that, by (B.33), $V_n^2 \leq n\beta^2$ a.s., and the first inequality follows by the definition of d_t. Similarly, the right-hand side of (B.42) can be shown to upper bound $\Pr\{1/n\sum_{t=1}^{n}(x_t - h(y_t))L_1(F_t(y_{t-1})) \geq \varepsilon/2\}$. Combining the two upper-bounds with the union bound gives the first inequality claimed in Lemma 3. The second inequality is obtained similarly.□

Appendix C

Proof of Lemma 22: Define, for notational convenience, for each $1 \leq i \leq N$, $a_i \stackrel{\text{def}}{=} \Lambda_{F^{(i)}} - \Lambda_{G^{(i)}}$. So, by the definitions of Λ_F and of the absolute loss function we can write

$$a_i = n^{-1/2}[\sum_{t=1}^{n} k(x_t)(F_t^{(i)}(y^{t-1}) - G_t^{(i)}(y^{t-1})) + \sum_{t=1}^{n} l(x_t)(G_t^{(i)}(y^{t-1}) - F_t^{(i)}(y^{t-1}))], \tag{C.1}$$

where $\{k(x_1), k(x_2), \ldots\}$ is an independent set (as $k(x_t) \in \sigma(r_t)$) and likewise, $\{l(x_1), l(x_2), \ldots\}$ is an independent set. Furthermore, one can easily confirm that

$$\begin{aligned} k(0) &= \begin{cases} -\frac{p}{1-2p} & \text{w.p. } 1-p \\ \frac{1-p}{1-2p} & \text{w.p. } p \end{cases} \\ k(1) &= \begin{cases} \frac{p}{1-2p} & \text{w.p. } 1-p \\ \frac{p-1}{1-2p} & \text{w.p. } p \end{cases} \end{aligned} \tag{C.2}$$

and that

$$\begin{aligned} l(0) &= \begin{cases} \frac{p}{1-2p} & \text{w.p. } 1-p \\ \frac{p-1}{1-2p} & \text{w.p. } p \end{cases} \\ l(1) &= \begin{cases} \frac{-p}{1-2p} & \text{w.p. } 1-p \\ \frac{1-p}{1-2p} & \text{w.p. } p. \end{cases} \end{aligned} \tag{C.3}$$

Combining the facts that these random variables have zero expectation (and variance $\frac{p-p^2}{(1-2p)^2}$), that $k(x_t)$ and $l(x_t)$ are $\sigma(r_t)$-measurable and are therefore independent of $\sigma(r^{t-1})$ and in particular of $F_t^{(i)}(y^{t-1}) - G_t^{(i)}(y^{t-1})$ which is $\sigma(r^{t-1})$-measurable, establishes each of the sums on the right hand side of equation (C.1) as a martingale difference. So that it follows by the definitions of $k(x_t)$, $l(x_t)$, and a_i, and the fact that $\rho(F^{(i)}, G^{(i)}) \leq \varepsilon$ for $1 \leq i \leq N$, that $\sqrt{n}a_i\frac{1-2p}{(1-p)\varepsilon}$ is a sum of two martingales with differences bounded by one.

It therefore follows from inequality (2.4.10) of Corollary 2.4.7 in [6] (and the union bound) that for all $t \geq 0$,

$$\Pr^p(a_i \frac{1-2p}{(1-p)\varepsilon} \geq t) \leq 2\exp(-2\frac{(t/2)^2}{(1+1)^2}) = 2e^{-t^2/8}. \tag{C.4}$$

In other words, we have

$$\Pr^p(a_i \geq t) \leq 2\exp(-\frac{t^2}{8}(\frac{1-2p}{(1-p)\varepsilon})^2). \tag{C.5}$$

A similar bound is obtained for $\Pr^p(a_i \leq -t)$ so that we have by the union bound for all $t \geq 0$:

$$\overset{p}{\Pr}(|a_i| \geq t) \leq 4\exp(-\frac{t^2}{8}(\frac{1-2p}{(1-p)\varepsilon})^2). \tag{C.6}$$

Now, for any non-negative $H(\cdot)$ we have

$$\begin{aligned} E^p H(|a_i|) &= \int_0^\infty \Pr^p(H(|a_i|) \geq t)dt && \text{(C.7)}\\ &= \int_0^1 \Pr^p(H(|a_i|) \geq t)dt + \int_1^\infty \Pr^p(H(|a_i|) \geq t)dt && \text{(C.8)}\\ &\leq 1 + \int_1^\infty \Pr^p(H(|a_i|) \geq t)dt. && \text{(C.9)} \end{aligned}$$

The idea is to make H increase about as fast as our bound on the tail of $|a_i|$ can bear. Taking $H(x) = e^{\gamma x^2}$ for some $\gamma > 0$ (to be chosen shortly), it follows from the above inequality and inequality (C.6) that

$$\begin{aligned} E^p H(|a_i|) &\leq 1 + \int_1^\infty \overset{p}{\Pr}(H(|a_i|) \geq t)dt && \text{(C.10)}\\ &= 1 + \int_1^\infty \overset{p}{\Pr}(|a_i| \geq \sqrt{\frac{\ln t}{\gamma}})dt && \text{(C.11)}\\ &\leq 1 + 4\int_1^\infty \exp(-\frac{\ln t}{8\gamma}(\frac{1-2p}{(1-p)\varepsilon})^2)dt && \text{(C.12)}\\ &= 1 + 4\int_1^\infty \frac{1}{t^{\frac{1}{8\gamma}(\frac{1-2p}{(1-p)\varepsilon})^2}}dt. && \text{(C.13)} \end{aligned}$$

Taking $\gamma = (\frac{1-2p}{4(1-p)\varepsilon})^2$ we obtain

$$E^p H(|a_i|) \leq 1 + 4\int_1^\infty \frac{1}{t^2}dt = 5. \tag{C.14}$$

Now, for the particular H chosen we have

$$H(E^p \max_{1\leq i\leq N} |a_i|) \quad \leq \quad E^p H(\max_{1\leq i\leq N} |a_i|) \tag{C.15}$$

$$
\begin{aligned}
&= E^p \max_{1\le i\le N} H(|a_i|) && \text{(C.16)}\\
&\le E^p \sum_{i=1}^{N} H(|a_i|) && \text{(C.17)}\\
&= \sum_{i=1}^{N} E^p H(|a_i|) && \text{(C.18)}\\
&\le 5N && \text{(C.19)}\\
&\le N^4, && \text{(C.20)}
\end{aligned}
$$

where the first inequality follows that of Jensen, the equality following it holds by the monotonicity of H (in the positive side of the real line), the inequality following it is guaranteed by the positivity of H, the inequality before the last one follows from inequality (C.14) and the last inequality follows trivially for all $N \ge 2$. Consequently,

$$
E^p \max_{1\le i\le N} |a_i| \le H^{-1}(N^4) = \sqrt{\frac{\ln N^4}{\gamma}} = \frac{4(1-p)\varepsilon}{1-2p} 2\sqrt{\ln N} \le \varepsilon \frac{8}{1-2p}\sqrt{\ln N}.\square \qquad \text{(C.21)}
$$

Appendix D

Proof of Lemma 23: Using the notation of the previous proof and plugging in the prediction values of $F^{(0)}$ we have:

$$
n^{1/2}\Lambda_{F^{(0)}} = 1/2(\sum_{t=1}^{n} k(x_t) + \sum_{t=1}^{n} l(x_t)) \qquad \text{(D.1)}
$$

so that

$$
n(\Lambda_{F^{(0)}})^2 = 1/4[(\sum_{t=1}^{n} k(x_t))^2 + 2(\sum_{t=1}^{n} k(x_t))(\sum_{t=1}^{n} l(x_t)) + (\sum_{t=1}^{n} l(x_t))^2]. \qquad \text{(D.2)}
$$

Since $k(x_t)$ and $l(x_t)$ are memoryless processes of zero mean and since also $k(x_t)$ and $l(x_s)$ are uncorrelated for $t \neq s$, taking expectation over the right hand side of the above equality annihilates all the cross-products, as $E^p k(x_t)k(x_s) = E^p k(x_t)l(x_s) = E^p l(x_t)l(x_s) = 0$ whenever $s \neq t$, therefore:

$$
E^p n(\Lambda_{F^{(0)}})^2 = 1/4[\sum_{t=1}^{n} E^p k(x_t)^2 + 2\sum_{t=1}^{n} E^p k(x_t)l(x_t) + \sum_{t=1}^{n} E^p l(x_t)^2]. \qquad \text{(D.3)}
$$

Since the absolute value of $k(\cdot)$ and $l(\cdot)$ is upper bounded by $\frac{1-p}{1-2p}$, we obtain

$$
E^p n(\Lambda_{F^{(0)}})^2 \le n(\frac{1-p}{1-2p})^2 \le \frac{n}{(1-2p)^2}, \qquad \text{(D.4)}
$$

or simply

$$E^p(\Lambda_{F(0)})^2 \leq \frac{1}{(1-2p)^2}. \tag{D.5}$$

From the Cauchy-Schwartz inequality, we obtain

$$E^p|\Lambda_{F(0)}| \leq \sqrt{E^p(\Lambda_{F(0)})^2} \leq \sqrt{\frac{1}{(1-2p)^2}} = \frac{1}{1-2p}.\Box \tag{D.6}$$

Appendix E

Proof of Lemma 31: The proof is reminiscent of that of Theorem 26 in that here, as was done there, we obtain our bound by decomposing the quantity of interest, namely, the relative estimated losses, into the sum of the corresponding differences for each of the blocks on which a predictor, for which we already have a performance bound from a previous result, was employed. Specifically, summing over bounds for each of the segments on which B_t is constant (and the predictor of Corollary 29 is used), we have:

$$\begin{aligned}
L_{\hat{P(y)}}(z^n) - \min_{\mathcal{F}} \hat{L_F}(z^n) &= \sum_{j=0}^{\lfloor \log n - 1 \rfloor} L_{\hat{P(y)}}(z)_{2^j}^{2^{j+1}-1} + L_{\hat{P(y)}}(z)_{2^{\lfloor \log n \rfloor}}^{n} - && \text{(E.1)}\\
&\quad \min_{F \in \mathcal{F}} \{ \sum_{j=0}^{\lfloor \log n - 1 \rfloor} \hat{L_F}(z)_{2^j}^{2^{j+1}-1} + \hat{L_F}(z)_{2^{\lfloor \log n \rfloor}}^{n} \} && \text{(E.2)}\\
&\leq \sum_{j=0}^{\lfloor \log n - 1 \rfloor} [L_{\hat{P(y)}}(z)_{2^j}^{2^{j+1}-1} - \min_{\mathcal{F}} \hat{L_F}(z)_{2^j}^{2^{j+1}-1}] && \text{(E.3)}\\
&\quad + L_{\hat{P(y)}}(z)_{2^{\lfloor \log n \rfloor}}^{n} - \min_{\mathcal{F}} \hat{L_F}(z)_{2^{\lfloor \log n \rfloor}}^{n} && \text{(E.4)}\\
&\leq \sum_{j=0}^{\lfloor \log n \rfloor} (2B_{2^j}+1)^2 \frac{\alpha}{2^\alpha(\alpha-1)} \ln|\mathcal{F}| && \text{(E.5)}\\
&= \frac{\alpha}{2^\alpha(\alpha-1)} \ln|\mathcal{F}| \sum_{j=0}^{\lfloor \log n \rfloor} (j+1) && \text{(E.6)}\\
&= \frac{\alpha}{2^\alpha(\alpha-1)} \ln|\mathcal{F}| \frac{\lfloor \log n + 1 \rfloor \cdot \lfloor \log n + 2 \rfloor}{2} && \text{(E.7)}\\
&\leq \frac{\alpha}{2^{\alpha+1}(\alpha-1)} \ln|\mathcal{F}| (\log t + 2)^2, && \text{(E.8)}
\end{aligned}$$

Where for a quantity, say Q, defined as a sum running from 1 to n, Q_a^b denotes the analogous quantity with the sum running from a to b. Equation (E.2) follows by breaking the overall loss estimators into estimators for consecutive blocks on which B_t is constant. Inequality (E.4) follows by the fact that the sum over the minima cannot be larger than the minimum over the sums. Inequality (E.5) follows from applying inequality (204) with $B \to B_{2^j}$ to each of the summands in (E.4). This is justified by the facts that, by construction, on the j-th block, our predictor coincides with that of Corollary (29) applied to this block (as if time started at the beginning of the block) and that, since $z_t = \phi_{B_t}(y_t)$, for each t in the j-th interval, z_t is guaranteed to belong to $[-B_{2^j}, B_{2^j}+1]$. $\square$

Appendix F

Proof of Lemma 32:

$$\begin{aligned}
|\hat{L}_P(y^n) - \hat{L}_{P(y)}(z^n)| &= |\sum_{t=1}^{n}[(1-y_t)L_0(F_t(y^{t-1})) + y_t L_1(F_t(y^{t-1}))] & \text{(F.1)}\\
&\quad -\sum_{t=1}^{n}[(1-z_t)L_0(F_t(y^{t-1})) + z_t L_1(F_t(y^{t-1}))] & \text{(F.2)}\\
&= |\sum_{t=1}^{n}(z_t - y_t)L_0(F_t(y^{t-1})) + (y_t - z_t)L_1(F_t(y^{t-1}))| & \text{(F.3)}\\
&\leq \sum_{t=1}^{n}|y_t - z_t||L_0(F_t(y^{t-1})) - L_1(F_t(y^{t-1}))| & \text{(F.4)}\\
&\leq \sum_{t=1}^{n}|y_t - z_t|, & \text{(F.5)}
\end{aligned}$$

where the last inequality follows from the fact that $|L_0(\cdot) - L_1(\cdot)|$ is clearly upper bounded by 1 for the α-powered loss function.$\square$

Appendix G

Proof of Equation (232): Throughout the proof we assume a fixed $s \in \mathcal{D}(\beta, c)^\infty$ and suppress the superscript s in the notation for the probability measure. We recall that the role of $g_n(y^n)$ is played here by

$$\frac{\sqrt{n \ln|\mathcal{F}|}}{8} \cdot \frac{\sqrt{2}}{\sqrt{2}-1} + \sqrt{\ln|\mathcal{F}|}\sum_{j=0}^{\lfloor \log n \rfloor}\frac{1}{2^j}\sum_{\tau=2^j}^{2^{j+1}-1}(1-2y_\tau)^2. \tag{G.1}$$

As the first term is deterministically $o(n)$, we will be done if we establish the fact that for all $\varepsilon > 0$,

$$\Pr\{\frac{1}{n}\sum_{j=0}^{\lfloor \log n \rfloor}\frac{1}{2^j}\sum_{\tau=2^j}^{2^{j+1}-1}(1-2y_\tau)^2 > \varepsilon\} \tag{G.2}$$

decays at a superexponential rate. Easy arithmetic gives us $E(1-2y_t)^2 = 5-4x_t+4Er_t^2 \stackrel{\text{def}}{=} e_t$ for all t. Denoting $z_t \stackrel{\text{def}}{=} (1-2y_t)^2 - e_t$, we have for sufficiently large n

$$\begin{aligned}
& \Pr\{\frac{1}{n}\sum_{j=0}^{\lfloor \log n \rfloor} \frac{1}{2^j} \sum_{\tau=2^j}^{2^{j+1}-1} (1-2y_\tau)^2 > \varepsilon\} && \text{(G.3)}\\
= \; & \Pr\{\frac{1}{n}\sum_{j=0}^{\lfloor \log n \rfloor} \frac{1}{2^j} \sum_{\tau=2^j}^{2^{j+1}-1} z_t + \frac{1}{n}\sum_{j=0}^{\lfloor \log n \rfloor} \frac{1}{2^j} \sum_{\tau=2^j}^{2^{j+1}-1} e_t > \varepsilon\} && \text{(G.4)}\\
\leq \; & \Pr\{\frac{1}{n}\sum_{j=0}^{\lfloor \log n \rfloor} \frac{1}{2^j} \sum_{\tau=2^j}^{2^{j+1}-1} z_t > \varepsilon/2\} + && \text{(G.5)}\\
& \Pr\{\frac{1}{n}\sum_{j=0}^{\lfloor \log n \rfloor} \frac{1}{2^j} \sum_{\tau=2^j}^{2^{j+1}-1} (5+4\beta^2) > \varepsilon/2\} && \text{(G.6)}\\
= \; & \Pr\{\frac{1}{n}\sum_{j=0}^{\lfloor \log n \rfloor} \frac{1}{2^j} \sum_{\tau=2^j}^{2^{j+1}-1} z_t > \varepsilon/2\} + && \text{(G.7)}\\
& \Pr\{\frac{\lfloor \log n + 1 \rfloor (5+4\beta^2)}{n} > \varepsilon/2\} && \text{(G.8)}\\
\leq \; & \sum_{j=0}^{\lfloor \log n \rfloor} \Pr\{\frac{1}{2^j} \sum_{\tau=2^j}^{2^{j+1}-1} z_t > \frac{\varepsilon n}{2\lfloor \log n + 1 \rfloor}\}, && \text{(G.9)}
\end{aligned}$$

where the first inequality follows since, by the definitions of e_t, $\mathcal{D}(\beta, c)$ (and the fact that $s \in \mathcal{D}(\beta, c)^\infty$), we are assured that $e_t \leq 5 + 4\beta^2$. Inequality (G.9) follows since the second term of (G.8) is clearly annihilated for sufficiently large n. We note now that, by their definitions, the z_t-s have zero-mean and are independent. Furthermore, it easily follows from the definition of $\mathcal{D}(\beta, c)$ that $\{z_t, \mathcal{F}_t\}$ satisfies the conditions of Theorem 37. Applying Theorem 37 to $\{z_t\}$, it is easy to show (as was done in the proof of Lemma 3) that there exists a constant K such that for all m, n, and $\varepsilon > 0$ we have

$$\Pr\{\frac{1}{n}\sum_{t=m+1}^{m+n} z_t > \varepsilon\} \leq \exp\{-nK\varepsilon^2\}. \tag{G.10}$$

Using this to upper-bound (G.9), we have

$$\begin{aligned}
\Pr\{\frac{1}{n}\sum_{j=0}^{\lfloor \log n \rfloor} \frac{1}{2^j} \sum_{\tau=2^j}^{2^{j+1}-1} (1-2y_\tau)^2 > \varepsilon\} \;&\leq\; \sum_{j=0}^{\lfloor \log n \rfloor} \exp\{-2^j K \frac{\varepsilon^2 n^2}{4\lfloor \log n + 1 \rfloor^2}\} && \text{(G.11)}\\
&\leq\; \sum_{j=1}^{\infty} \exp\{-jK \frac{\varepsilon^2 n^2}{4\lfloor \log n + 1 \rfloor^2}\} && \text{(G.12)}\\
&\leq\; \tilde{K} \exp\{-K \frac{\varepsilon^2 n^2}{4\lfloor \log n + 1 \rfloor^2}\}, && \text{(G.13)}
\end{aligned}$$

which establishes the superexponential decay rate desired.□

References

[1] A. Baruch, "Universal Algorithms for Sequential Decision in the Presence of Noisy Observations," submitted to the senate of the technion (Master's thesis), February 1999. (See also Proc. ISIT 98', p. 331, Cambridge, MA, August 1998.)

[2] 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).

[3] 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.

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

[5] L. D. Davisson, "Universal Noiseless Coding," *IEEE Trans. Inform. Theory*, vol. IT-19, no. 6, pp. 783-795, November 1973.

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

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

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

[9] M. Feder and N. Merhav, "Hierarchical Universal Coding," *IEEE Trans. Inform. Theory,* vol. 42, no. 5, pp. 1354-1364, September 1996.

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

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

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

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

[14] 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.

[15] W. Hoeffding, "Probability Inequalities for Sums of Bounded Random Variables," *Amer. Statist. Assoc.*, vol. 58, pp. 13-30, 1963.

[16] M. Ledoux and M. Talagrand, *Probability in Banach Spaces* Springer-Verlag, New York, 1991.

[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, "A Strong Version of the Redundancy-Capacity Theorem of Universal Coding," *IEEE Trans. Inform. Theory,* vol. 41, no. 3, pp. 714-722, May 1995.

[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] V. H. D. L. Peña, "A General Class of Exponential Inequalities for Martingales and Ratios," *Ann. of Prob.*, vol. 27, no. 1, pp. 537-564, 1999.

[23] D. Pollard, "Asymptotics via Empirical Processes," *Statistical Science*, vol. 4, pp. 341-366, 1989.

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

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

[26] J. Rissanen, "Fisher Information and Stochastic Complexity," *IEEE Trans. Inform. Theory*, vol. 42, no. 1, pp. 40-47, January 1996.

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

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

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

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

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

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

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

[34] M. Talagrand, "Sharper Bounds for Gaussian and Empirical Processes," *Ann. of Prob.*, vol. 22, no. 1, pp. 28-76, 1994.

[35] V. N. Vapnik and A. Y. Chervonenkis, "On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities," *Theory of Prob. and its Applications*, 16 (1971), no. 2, 264-280.

[36] 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.

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

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

[39] 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.

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

[41] 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.

[42] T. Weissman and N. Merhav, "Universal Prediction of Individual Binary Sequences in the Presence of Arbitrarily Varying, Memoryless Additive Noise," submitted to ISIT 2000.

[43] T. Weissman and N. Merhav, "Universal Prediction of Individual Binary Sequences in the Presence of Noise," Internal report, Department of Electrical Engineering, Technion, EE PUB #1224, CC PUB #293, October, 1999.

[44] L. Wu, "Moderate Deviations of Dependent Random Variables Related to CLT," *Ann. of Prob.*, vol. 23, no. 1, pp. 420-445, 1995.

[45] L. Wu, "Large Deviations, Moderate Deviations and LIL for Empirical Processes," *Ann. of Prob.*, vol. 22, no. 1, pp. 17-27, 1994.