On Prediction of Individual Sequences*

Nicolò Cesa-Bianchi
Dept. of Information Sciences,
University of Milan,
Via Comelico 39,
20135 Milano, Italy
cesabian@dsi.unimi.it

Gábor Lugosi
Department of Economics,
Pompeu Fabra University
Ramon Trias Fargas 25-27,
08005 Barcelona, Spain
lugosi@upf.es

October 14, 1999

Abstract

Sequential randomized prediction of an arbitrary binary sequence is investigated. No assumption is made on the mechanism of generating the bit sequence. The goal of the predictor is to minimize its relative loss (or regret), i.e., to make almost as few mistakes as the best "expert" in a fixed, possibly infinite, set of experts. We point out a surprising connection between this prediction problem and empirical process theory. First, in the special case of static (memoryless) experts, we completely characterize the minimax regret in terms of the maximum of an associated Rademacher process. Then we show general upper and lower bounds on the minimax regret in terms of the geometry of the class of experts. As main examples, we determine the exact order of magnitude of the minimax regret for the class of autoregressive linear predictors and for the class of Markov experts.

***AMS 1991 subject classifications.** Primary-62C20 (Minimax procedures); secondary-60G25 (Prediction theory). Key words and phrases: Universal prediction, prediction with experts, absolute loss, empirical processes, covering numbers, finite-state machines. A preliminary version of this work appeared in the Proceedings of the 11th Annual ACM Conference on Computational Learning Theory. ACM Press, 1998. Both authors gratefully acknowledge support from ESPRIT Working Group EP 27150, Neural and Computational Learning II (NeuroCOLT II). The work of the second author was also supported by DGES grant PB96-0300.

1 Introduction

Consider the problem of predicting sequentially an arbitrary binary sequence of length n. At each time unit $t = 1, \ldots, n$, after making a guess, one observes the t-th bit of the sequence. Predictions are allowed to depend on the outcome of biased coin flips. The loss at time t is defined as the probability (with respect to the coin flip) of predicting incorrectly the t-th bit of the sequence. The goal is to predict any sequence almost as well as the best "expert" in a given set of experts. In this paper we investigate the minimum number of excess mistakes, with respect to the mistakes of the best expert, achievable in a worst-case sense; that is, when no assumptions are made on the mechanism generating the binary sequence.

To formally define the prediction problem, we introduce the notion of *expert*. An expert F is a sequence of functions $F_t : \{0,1\}^{t-1} \to [0,1]$, $t \geq 1$. Each expert defines a prediction strategy in the following way: upon observing the first $t-1$ bits $y^{t-1} = (y_1, \ldots, y_{t-1}) \in \{0,1\}^{t-1}$, expert F predicts that the next bit y_t is 1 with probability $F_t(y^{t-1})$.

We now describe the binary prediction problem as an *iterated game* between a *predictor* and the *environment* (see also [29]). This game is parametrized by a positive integer n (number of game rounds to play) and by a set $\mathcal{F}$ of experts (the expert class). On each round $t = 1, \ldots, n$:

1. The predictor picks a number $P_t \in [0,1]$.
2. The environment picks a bit $y_t \in \{0,1\}$.
3. Each $F \in \mathcal{F}$ incurs loss $|F_t(y^{t-1}) - y_t|$ and the predictor incurs loss $|P_t - y_t|$.

For each positive integer n, one may view an expert F as a probability distribution over the set $\{0,1\}^n$ of binary strings of length n such that, for each $y^{t-1} \in \{0,1\}^{t-1}$, $F_t(y^{t-1})$ stands for the conditional probability of $y_t = 1$ given the past y^{t-1}. (Here y^0 denotes the empty sequence.) In this respect, the loss $|F_t(y^{t-1}) - y_t|$ of expert F at time t may be intepreted as the probability of error $\mathbf{P}\{X_t \neq y_t\}$ if the expert's guess $X_t \in \{0,1\}$ were to be drawn randomly according to the probability $\mathbf{P}\{X_t = 1\} = F_t(y^{t-1})$.

As our goal is to compare the loss of the predictor with the loss of the best expert in $\mathcal{F}$, we find it convenient to define the strategy P of the predictor in the same way as we defined experts. That is, P is a sequence of functions $P_t : \{0,1\}^{t-1} \to [0,1]$, $t \geq 1$, and P predicts that y_t is 1 with probability $P_t(y^{t-1})$, where y^{t-1} is the sequence of previously observed bits. Note that the predictor's strategy P may (and in general will) be defined in terms of the given expert class $\mathcal{F}$. Finally, as we did with experts, for each $n \geq 1$ we may view the predictor's strategy as a distribution P over $\{0,1\}^n$ and interpret the loss $|P_t(y^{t-1}) - y_t|$ as the error probability $\mathbf{P}\{\widehat{Y}_t \neq y_t\}$, where the prediction $\widehat{Y}_t \in \{0,1\}$ is randomly drawn according to the probability $\mathbf{P}\{\widehat{Y}_t = 1\} = P_t(y^{t-1})$.

We now move on to define some quantities characterizing the performance of a strategy P in the prediction game. The *cumulative loss* of each expert $F \in \mathcal{F}$ is defined by

$$L_F(y^n) \stackrel{\text{def}}{=} \sum_{t=1}^{n} |F_t(y^{t-1}) - y_n| \ ,$$

and the cumulative loss of the predictor using strategy P is

$$L_P(y^n) \stackrel{\text{def}}{=} \sum_{t=1}^{n} |P_t(y^{t-1}) - y_n| \ .$$

The goal of the predictor P is to minimize its *worst-case regret*, defined by

$$R_n(P,\mathcal{F}) \stackrel{\text{def}}{=} \max_{y^n\in\{0,1\}^n} \left(L_P(y^n) - \inf_{F\in\mathcal{F}} L_F(y^n) \right) \ .$$

Finally, we define the *minimax regret* as the smallest worst-case regret achievable by any predictor,

$$V_n(\mathcal{F}) \stackrel{\text{def}}{=} \min_P R_n(P,\mathcal{F}),$$

where the minimum is taken over the compact set of all distributions P over $\{0,1\}^n$. In the rest of this work, we show that $V_n(\mathcal{F})$ is characterized by metric properties of the class $\mathcal{F}$, and we give examples of this characterization for specific choices of $\mathcal{F}$.

The first study of the quantity $V_n(\mathcal{F})$, though for a very specific choice of the expert class $\mathcal{F}$, goes back to a 1965 paper by Cover [8]. He proves that $V_n(\mathcal{F}) = \Theta(\sqrt{n})$ when $\mathcal{F}$ contains two experts: one always predicting 0 and the other always predicting 1. A remarkable extension was achieved by Feder, Merhav, and Gutman [11], who considered the class of all finite-state experts. In particular, they show that $V_n(\mathcal{F}) = O\left(\sqrt{2^k n}\right)$ when $\mathcal{F}$ contains all k-th order Markov experts (a subclass of all finite-state experts). Cesa-Bianchi et al. [3], building on results of Vovk [28] and Littlestone and Warmuth [19], consider arbitrary finite classes of experts and prove that the minimax regret is bounded from above as

$$V_n(\mathcal{F}) \le \sqrt{(n/2)\ln|\mathcal{F}|} \ . \tag{1}$$

(In fact, the bound shown in [3] has a slightly different form. The above inequality is proved in Theorem 1 below.) This surprising result shows that there exists a prediction algorithm such that the number of mistakes is only a constant times $\sqrt{n}$ larger than the number of errors committed by the best expert, regardless of the outcome of the bit sequence. (Typically, the number of mistakes made by the best expert grows linearly with n.) Moreover, the constant is proportional to the logarithm of the size of the expert class.

In [3] it is also shown that the upper bound (1) is asymptotically tight in a *worst-case* sense. That is, for each $N \ge 1$ there exists an expert class $\mathcal{F}_N$ of cardinality N such that

$$\liminf_{N\to\infty} \liminf_{n\to\infty} \frac{V_n(\mathcal{F}_N)}{\sqrt{(n/2)\ln N}} = 1 \ .$$

The approach taken in this paper is different. We treat $\mathcal{F}$ as a fixed class and try to estimate the size of the minimax regret $V_n(\mathcal{F})$ for this class. It turns out that for a fixed class, the order of magnitude of $V_n(\mathcal{F})$ may be significantly smaller than $\sqrt{(n/2)\ln|\mathcal{F}|}$. (Just consider a class $\mathcal{F}$ of two experts such that the predictions of the two experts are always the same except in the first time instant $t=1$. In this case it is easy to see that $V_n(\mathcal{F}) = 1/2$.) In general, the value of $V_n(\mathcal{F})$ depends on the geometry of the expert class. As the above-mentioned examples of Cover and Feder *et al.* show, even for infinite classes of experts one may be able to determine meaningful upper bounds.

The paper is organized as follows: In Section 2 we give a new simple proof of inequality (1) and extend it to expert classes defined as a convex hull of finitely many experts. In Section 3 a special type of experts is considered. These so-called static experts predict according to prespecified probabilities, independently of the past bits of the sequence. In this special case it is possible to characterize the minimax regret $V_n(\mathcal{F})$ by the maximum of a Rademacher process, which highlights an intriguing connection to empirical process theory. In Section 4 we use the insight provided by the example of static experts to define a general algorithm for prediction, and derive a general upper bound for $V_n(\mathcal{F})$. In Section 5 lower bounds for $V_n(\mathcal{F})$ are derived. To demonstrate the tightness of the upper and lower bounds, we consider two main examples: In Section 6 we derive matching upper and lower bounds for the class of k-th order autoregressive linear predictors. Finally, in Section 7 we take another look at the class of Markov experts of Feder, Merhav, and Gutman.

2 Finite classes and their convex hulls

We start with a new short proof of (1). The algorithm is a simple version of a "weighted majority" method proposed by Vovk [28] and Littlestone and Warmuth [19]. However, whereas previous analyses of this algorithm did not get the best leading costant of (1), we are able to do so using a proof technique similar to that used for proving [5, Theorem 5].

Theorem 1 *For any finite expert class $\mathcal{F}$, define the predictor strategy P by*

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

where $\eta > 0$ is a parameter. If $\eta = \sqrt{8\ln|\mathcal{F}|/n}$ then, for any $y^n \in \{0,1\}^n$,

$$L_P(y^n) - \min_{\mathcal{F}} L_F(y^n) \le \sqrt{\frac{n}{2}\ln|\mathcal{F}|} \ .$$

Proof. Fix an arbitrary sequence $y^n \in \{0,1\}^n$. Define $W_1 = \ln|\mathcal{F}|$, and for all $t > 1$ let

$$W_t = \sum_{F\in\mathcal{F}} e^{-\eta L_F(y^{t-1})} \ .$$

Then

$$\begin{aligned}
\ln\frac{W_{n+1}}{W_1} &= \ln\left(\sum_{F\in\mathcal{F}} e^{-\eta L_F(y^n)}\right) - \ln|\mathcal{F}| \\
&\ge \ln\left(\max_{F\in\mathcal{F}} e^{-\eta L_F(y^n)}\right) - \ln|\mathcal{F}| \\
&= -\eta\min_{F\in\mathcal{F}} L_F(y^n) - \ln|\mathcal{F}| \ .
\end{aligned} \tag{2}$$

On the other hand, for each $t = 1, \ldots, n$

$$
\begin{aligned}
\ln \frac{W_{t+1}}{W_t} &= \ln \frac{\sum_{\mathcal{F}} e^{-\eta|F_t(y^{t-1})-y_t|} e^{-\eta L_F(y^{t-1})}}{\sum_{\mathcal{F}} e^{-\eta L_F(y^{t-1})}} \\
&= \ln \mathbf{E}_{F\sim Q_t}\left[e^{-\eta|F_t(y^{t-1})-y_t|}\right],
\end{aligned}
$$

where $\mathbf{E}_{F\sim Q_t}$ denotes expectation taken with respect to a distribution Q_t on $\mathcal{F}$ which assigns a probability proportional to $e^{-\eta L_F(y^{t-1})}$ to each $F \in \mathcal{F}$. Recalling Hoeffding's bound [17] on the moment-generating function of bounded random variables (see also [9, Lemma 8.1]), we observe that

$$
\begin{aligned}
&\ln \mathbf{E}_{F\sim Q_t}\left[e^{-\eta|F_t(y^{t-1})-y_t|}\right] \\
&\le -\eta \mathbf{E}_{F\sim Q_t}\left|F_t(y^{t-1}) - y_t\right| + \frac{\eta^2}{8} \\
&= -\eta \left|\mathbf{E}_{F\sim Q_t}[F_t(y^{t-1})] - y_t\right| + \frac{\eta^2}{8} \qquad (3)\\
&= -\eta \left|P_t(y^{t-1}) - y_t\right| + \frac{\eta^2}{8}, \qquad (4)
\end{aligned}
$$

where equation (3) holds because $y_t \in \{0,1\}$, and (4) holds by definition of P_t.

Summing over $t = 1, \ldots, n$ we get

$$
\ln \frac{W_{n+1}}{W_1} \le -\eta L_P(y^n) + \frac{\eta^2}{8} n .
$$

Combining this with (2) and solving for $L_P(y^n)$ we find that

$$
L_P(y^n) \le \min_{\mathcal{F}} L_F(y^n) + \frac{\ln|\mathcal{F}|}{\eta} + \frac{\eta}{8} n .
$$

Finally, choosing $\eta = \sqrt{8 \ln|\mathcal{F}|/n}$ yields the desired bound. □

The upper bound of Theorem 1 may be easily extended to infinite classes of experts which may be written as a convex hull of finitely many experts:

Theorem 2 *Let $F^{(1)}, \ldots, F^{(N)}$ be arbitrary experts, and consider the class $\mathcal{F}$ of all convex combinations of $F^{(1)}, \ldots, F^{(N)}$, that is,*

$$
\mathcal{F} = \left\{ \sum_{j=1}^N q_j F^{(j)} : \; q_1, \ldots, q_N \ge 0, \; \sum_{j=1}^N q_j = 1 \right\}.
$$

Then

$$
V_n(\mathcal{F}) \le \sqrt{(n/2) \ln N}.
$$

Proof. The theorem immediately follows from Theorem 1 and the simple fact that for any bit sequence $y^n \in \{0,1\}^n$ and expert $G = \sum_{j=1}^N q_j F^{(j)} \in \mathcal{F}$ there exists an expert among

$F^{(1)}, \ldots, F^{(N)}$ whose loss on y^n is not larger than that of G. To see this note that

$$
\begin{aligned}
L_G(y^n) &= \sum_{t=1}^n |G_t(y^{t-1}) - y_t| \\
&= \sum_{t=1}^n \left| \sum_{j=1}^N q_j F_t^{(j)}(y^{t-1}) - y_t \right| \\
&= \sum_{t=1}^n \sum_{j=1}^N q_j \left| F_t^{(j)}(y^{t-1}) - y_t \right| \\
&= \sum_{j=1}^N q_j \sum_{t=1}^n \left| F_t^{(j)}(y^{t-1}) - y_t \right| \\
&= \sum_{j=1}^N q_j L_{F^{(j)}}(y^n) \\
&\ge \min_{j=1,\ldots,N} L_{F^{(j)}}(y^n) .
\end{aligned}
$$

□

3 Prediction with static experts

In this section we study the important special case when every $F \in \mathcal{F}$ is such that the prediction of F at time t depends only on t but not on the past y^{t-1}. We use $\bar{F}_t \in [0,1]$ to denote the prediction at time t of a static expert F. Thus, $F_t(y^{t-1}) = \bar{F}_t$ for all $t = 1, \ldots, n$ and all $y^{t-1} \in \{0,1\}^{t-1}$. Such experts are called *static* in [3]. When interpreting experts as probability distributions on $\{0,1\}^n$, this means that every expert corresponds to a product distribution. Note that every static expert is determined by a vector $(\bar{F}_1, \ldots, \bar{F}_n)$, and $\mathcal{F}$ may be thought of as a subset of $[0,1]^n$.

Next we derive a formula for the minimax regret of any (finite or infinite) *fixed* class $\mathcal{F}$ of static experts. This enables us to derive sharp upper bounds, as well as corresponding lower bounds, in terms of the geometry of $\mathcal{F}$. We start by observing that a previous characterization of the minimax regret, implicitely shown in [3], corresponds to the expected supremum of a class of random variables.

Theorem 3 *For any class $\mathcal{F}$ of static experts,*

$$
V_n(\mathcal{F}) = \mathbf{E}\left[\sup_{F\in\mathcal{F}} \sum_{t=1}^n \left(\frac{1}{2} - \bar{F}_t\right)(1 - 2Y_t) \right]
$$

where $Y_1, \ldots, Y_n$ are independent Bernoulli $(1/2)$ random variables.

Proof. From the statement of [3, Theorem 3.1.2, page 441], we have that if $Y_1, \ldots, Y_n$ are independent Bernoulli (1/2) random variables, then

$$\begin{aligned}
V_n(\mathcal{F}) &= \frac{n}{2} - \mathbf{E}\left[\inf_{F\in\mathcal{F}} \sum_{t=1}^n |\bar{F}_t - Y_t|\right] \\
&= \mathbf{E}\left[\sup_{F\in\mathcal{F}} \sum_{t=1}^n \left(\frac{1}{2} - |\bar{F}_t - Y_t|\right)\right] \\
&= \mathbf{E}\left[\sup_{F\in\mathcal{F}} \sum_{t=1}^n \left(\frac{1}{2} - \bar{F}_t\right)(1-2Y_t)\right] .
\end{aligned}$$

□

Remark. The fact that $V_n(\mathcal{F}) \ge (n/2) - \mathbf{E}\left[\inf_{F\in\mathcal{F}} \sum_{t=1}^n |\bar{F}_t - Y_t|\right]$ may be proven very easily, and, in fact, it can be extended to classes of arbitrary experts (see Theorem 10 below.) The more involved proof of the other direction, given in [3], is based on an inductive argument and exhibits a minimax optimal prediction algorithm achieving $V_n(\mathcal{F})$ for any class of static experts $\mathcal{F}$. The algorithm (described in this form in [7, Section 2.6.2]) is as follows: Suppose $y^n \in \{0,1\}^n$ is the sequence to predict. Then the prediction at each time $t = 1, \ldots, n$ is

$$P_t(y^{t-1}) = \frac{1}{2} + \mathbf{E}\left[\frac{\inf_{\mathcal{F}} L_F(y^{t-1}0Y^{n-t}) - \inf_{\mathcal{F}} L_F(y^{t-1}1Y^{n-t})}{2}\right] .$$

Theorems 4 and 5 below provide performance bounds for this algorithm.

With a more compact notation, Theorem 3 states

$$V_n(\mathcal{F}) = \frac{1}{2}\mathbf{E}\left[\sup_{F\in\mathcal{F}} \sum_{t=1}^n \tilde{F}_t Z_t\right] , \tag{5}$$

where $Z_1, \ldots, Z_n$ are independent Rademacher random variables (i.e., $\mathbf{P}\{Z_t = -1\} = \mathbf{P}\{Z_t = +1\} = 1/2$) and each $\tilde{F}_t$ is the constant $1 - 2\bar{F}_t$. Rademacher averages of this type appear in the study of uniform deviations of averages from their means, and they have been thoroughly studied in empirical process theory. (For excellent surveys on empirical process theory we refer to Pollard [23] and Giné [14].)

Based on the characterization (5) of $V_n(\mathcal{F})$ as a Rademacher average, we get the following two results, which give useful upper and lower bounds for the minimax loss in terms of certain covering numbers of the class of experts. These covering numbers are defined as follows. For any class $\mathcal{F}$ of static experts, let $N_2(\mathcal{F}, r)$ be the minimum cardinality of a set $\mathcal{F}_r$ of static experts (possibly not all belonging to $\mathcal{F}$) such that

$$(\forall F \in \mathcal{F})\ (\exists G \in \mathcal{F}_r) \qquad \sqrt{\sum_{t=1}^n \left(\bar{F}_t - \bar{G}_t\right)^2} \le r .$$

Theorem 4 *For any class $\mathcal{F}$ of static experts,*

$$V_n(\mathcal{F}) \le 12 \int_0^{\sqrt{n}/2} \sqrt{\ln(N_2(\mathcal{F}, r) + 1)}\, dr .$$

Proof. According to (5), it suffices to derive the upper bound for the Rademacher process characterizing $V_n(\mathcal{F})$. Such bounds are well-known, the bound given here is Dudley's metric entropy bound and may be proved by the technique of "chaining", as explained very nicely in Pollard [23]. □

Theorem 4 may be used to obtain bounds which are not achievable by earlier methods (see examples below).

Theorem 5 *Let $\mathcal{F}$ be a class of static experts containing F and G such that $\bar{F}_t = 0$ and $\bar{G}_t = 1$ for all $t = 1, \ldots, n$. Then, for some universal constant $K > 0$,*

$$V_n(\mathcal{F}) \ge K \sup_r r\sqrt{\ln N_2(\mathcal{F}, r)} .$$

Theorem 5 is a direct consequence of (5) and Corollary 4.14 in Ledoux and Talagrand [18]. In most cases of static experts, Theorem 5 gives a lower bound matching (up to a constant factor) the upper bound obtained by Theorem 4. (See Talagrand [26] for a detailed discussion about the tightness of such bounds and possible improvements.) Now we describe two natural examples which show how to use the above results in concrete situations. We start with the simplest case.

Let $\mathcal{F}$ be the class of all experts $F \in \mathcal{F}$ of the form $F_t(y^{t-1}) = p$ regardless of t and the past outcomes y^{t-1}. The class contains all such experts with $p \in [0,1]$. Since each expert in $\mathcal{F}$ is static, and $N_2(\mathcal{F}, r) \le \sqrt{n}/r$, we may use Theorem 4 to bound the minimax value from above. After simple calculations we obtain

$$V_n(\mathcal{F}) \le c\sqrt{n}$$

for some constant c. Note that in this case one may obtain a better constant by different methods. For example, it follows from Theorem 2 that

$$V_n(\mathcal{F}) \le \sqrt{(n/2)\ln 2} .$$

Thus the bound of Theorem 4 does not give optimal constants, but it almost always gives bounds which have the correct order of magnitude. In this special case the sharpest upper and lower bounds may be directly obtained from Theorem 3, since

$$\begin{aligned}
V_n(\mathcal{F}) &= \mathbf{E}\left[\sup_{p\in[0,1]} \sum_{t=1}^n \left(\frac{1}{2} - p\right)(1-2Y_t)\right] \\
&= \mathbf{E}\left[\max\left\{\sum_{t=1}^n \left(\frac{1}{2} - Y_t\right), \sum_{t=1}^n \left(Y_t - \frac{1}{2}\right)\right\}\right] \\
&= \mathbf{E}\left|\sum_{t=1}^n \left(\frac{1}{2} - Y_t\right)\right| .
\end{aligned}$$

Now on the one hand, by the Cauchy-Schwarz inequality,

$$\mathbf{E}\left|\sum_{t=1}^n \left(\frac{1}{2} - Y_t\right)\right| \le \sqrt{\mathbf{E}\left(\sum_{t=1}^n \left(\frac{1}{2} - Y_t\right)\right)^2} = \frac{\sqrt{n}}{2},$$

and on the other hand, Khinchine's inequality (see Szarek [25]) for the expected deviation of a binomial random variable from its mean implies

$$\mathbf{E}\left|\sum_{t=1}^{n}\left(\frac{1}{2}-Y_t\right)\right| \geq \sqrt{\frac{n}{8}}.$$

Summarizing the upper and lower bounds, for every n, we have

$$0.3535 \leq \frac{V_n(\mathcal{F})}{\sqrt{n}} \leq 0.5.$$

(For example, for $n = 100$, there exists a prediction strategy such that for any sequence $y_1, \ldots, y_{100}$ the number of mistakes is not more than that of the best expert plus 5, but for any prediction strategy there exists a sequence $y_1, \ldots, y_{100}$ such that the number of excess errors is at least 3.) Note that by the central limit theorem, $V_n(\mathcal{F})/\sqrt{n} \to 1/\sqrt{2\pi} \approx 0.3989$ (this exact asymptotical value was originally shown by Cover [8].) It is easy to see that Theorem 5 also gives a lower bound of the right order of magnitude, though with a suboptimal constant.

We now show a case where Theorem 4 yields an upper bound significantly better than those obtainable with any of the previous techniques. Let $\mathcal{F}$ be the class of all static experts which predict in a monotonic way, that is, for each $F \in \mathcal{F}$, either $\bar{F}_t \leq \bar{F}_{t+1}$ for all $t \geq 1$ or $\bar{F}_t \geq \bar{F}_{t+1}$ for all $t \geq 1$. In view of applying Theorem 4, we upper bound the log of $N_2(\mathcal{F}, r)$ for any $0 < r < \sqrt{n}$. Consider the class $\mathcal{F}_r$ of all monotone static experts taking values in

$$\left\{(2k+1)r/\sqrt{n} \,:\, k = 0, 1, \ldots, m\right\}$$

where m is the largest integer such that $(2m+1)r/\sqrt{n} \leq 1$. Then $m \leq \lfloor\sqrt{n}/(2r)\rfloor$. Let $d = m + 1$ be the cardinality of the range of the functions in $\mathcal{F}_r$. Clearly, $N_2(\mathcal{F}, r) \leq |\mathcal{F}_r| \leq 2\binom{n+d}{d}$. Using

$$\ln\binom{n+d}{d} \leq d(\ln(1 + n/d) + 1)$$

and $n/d \leq 2r\sqrt{n}$ we get $\ln N_2(\mathcal{F}, r) = O\left((\sqrt{n}/r)\ln(rn)\right)$. Hence, applying Theorem 4, we obtain

$$V_n(\mathcal{F}) = O\left(\sqrt{n\log n}\right) .$$

Note that this is a large "nonparametric" class of experts, yet, we have been able to derive a bound which is just slightly larger than those obtained for finite classes.

In the special case of static classes $\mathcal{F}$ such that $\bar{F}_t \in \{0, 1\}$ for each $F \in \mathcal{F}$, a quantity which may be used to obtain good bounds on the covering numbers is the Vapnik-Chervonenkis (VC) dimension [27]. If a class $\mathcal{F}$ has VC-dimension bounded by a positive constant d, then, using a result of Haussler [16, Theorem 1], one may show that

$$N_2(\mathcal{F}, r) \leq e(d+1)\left(\frac{2en}{r^2}\right)^d .$$

For such classes, Theorem 4 gives $V_n(\mathcal{F}) = O\left(\sqrt{dn}\right)$, which was not obtainable with previous techniques. Note that this bound can not be improved in general. In fact, Haussler [16,

Theorem 2] exhibits, for each positive integer $d \geq 1$ and for each n integer multiple of d, a class $\mathcal{F}_d$ with VC-dimension d and such that

$$N_2(\mathcal{F}_d, r) \geq \left(\frac{n}{2e(2r^2 + d)}\right)^d .$$

Hence, the above discussion and Theorem 5 together yield $V_n(\mathcal{F}_d) = \Theta\left(\sqrt{dn}\right)$.

4 Chaining: a general prediction algorithm

In this section we obtain general upper bounds for binary prediction without assuming that the experts in $\mathcal{F}$ are static. Our aim is to extend the upper bound (1) to arbitrary (i.e., not necessarily finite) expert classes. The simplest way to do this is by discretizing the expert class, that is, taking a finite set of experts that approximately represent the whole class, and use the algorithm described in Theorem 1 for the finite class. As we will shortly see, this leads to a simple but suboptimal upper bound. To state this bound, we need to introduce a notion of covering numbers for a nonstatic expert class: The r-covering number $N_1(\mathcal{F}, r)$ of a class $\mathcal{F}$ of experts is the minimum cardinality of a set $\mathcal{F}_r$ of experts such that

$$(\forall F \in \mathcal{F})\ (\forall y^n \in \{0,1\}^n)\ (\exists G \in \mathcal{F}_r) \qquad \sum_{t=1}^{n}\left|F_t(y^{t-1}) - G_t(y^{t-1})\right| \leq r .$$

Then we have the following easy consequence of (1).

Corollary 6 *For any expert class $\mathcal{F}$,*

$$V_n(\mathcal{F}) \leq \inf_{r>0}\left(r + \sqrt{\frac{n\ln N_1(\mathcal{F}, r)}{2}}\right) .$$

Proof. First note that for each $r > 0$, the r-covering $\mathcal{F}_r$ satisfies $V_n(\mathcal{F}) \leq r + V_n(\mathcal{F}_r)$. Applying (1) to bound $V_n(\mathcal{F}_r)$ yields the upper bound. □

Note that for finite classes this bound is at least as good as (1) as, trivially, $N_1(\mathcal{F}, r) \leq |\mathcal{F}|$. Corollary 6 provides a quite acceptable upper bound in many cases, though sometimes it is way off-mark. One such example is the class of "monotone" experts considered in Section 3. For this class Theorem 4 implies $V_n(\mathcal{F}) = O(\sqrt{n\log n})$, while the best bound one can get by Corollary 6 is about the order of $n^{2/3}$. Indeed, it is easy to see that $\ln N_1(\mathcal{F}, r) = O\left((n/r)\ln r\right)$. Taking $r = n^{2/3}$ shows that $V_n(\mathcal{F}_r) = O\left(n^{2/3}(\ln n)^{1/3}\right)$.

To achieve the near-optimal upper bound of Theorem 4 in the case of static experts, a technique called "chaining" is used, a standard method in empirical process theory. For nonstatic experts, however, unfortunately we do not have a characterization of the minimax regret by an empirical process. Still, the idea of chaining turns out to be useful even in this case. Next we use this idea to define a prediction algorithm, and derive a performance bound, which, in turn, leads to a general upper bound on $V_n(\mathcal{F})$ of the same form as the upper bound stated in Theorem 4 for static experts, albeit we use a somewhat stronger notion of covering numbers.

In the argument below we need the assumption that the expert class $\mathcal{F}$ is such that all predictors are bounded away from 0 and 1. That is, there exists $0 < \delta < 1/2$ such that for every $F \in \mathcal{F}$, $t \le n$, and $y^{t-1} \in \{0,1\}^{t-1}$, $F_t(y^{t-1}) \in [\delta, 1-\delta]$. We call such a class *$\delta$-bounded.* We do not know whether it is possible to drop this assumption, but the predictor presented here does not seem to work for classes of experts not satisfying such a property. In the remainder of this section we assume that $\mathcal{F}$ is δ-bounded.

We start by defining the covering number used in the bound. For any class $\mathcal{F}$ of experts, define the metric ρ by

$$\rho(F,G) \stackrel{\text{def}}{=} \max_{\substack{1\le t\le n \\ y^n \in \{0,1\}^n}} \left| F_t(y^{t-1}) - G_t(y^{t-1}) \right| .$$

For any $\varepsilon > 0$, an ε-cover of $\mathcal{F}$ (with respect to the metric ρ) is a set $\mathcal{F}_\varepsilon \subset \mathcal{F}$ of experts on $\{0,1\}^n$ such that for all $F \in \mathcal{F}$, there exists an expert $G \in \mathcal{F}_\varepsilon$ such that $\rho(F,G) \le \varepsilon$. The ε-covering number $N_\infty(\mathcal{F}, \varepsilon)$ is the cardinality of the smallest ε-cover of $\mathcal{F}$.

We now move on to the description of the predictor P for a δ-bounded class of experts $\mathcal{F}$. For all $k \ge 1$, define

$$\varepsilon_k = \frac{1-2\delta}{2(1+2\delta)}(2\delta)^k \quad \text{and} \quad a_k = (1-2\delta)(2\delta)^{k-1} .$$

Let $\mathcal{G}_k$ be an ε_k-cover of $\mathcal{F}$ of cardinality $N_k = N_\infty(\mathcal{F}, \epsilon_k)$. Define $\mathcal{G}_0 = \{F^{(0)}\}$ as the singleton class containing the static expert $F^{(0)}$ such that $F_t^{(0)} = 1/2$ for each $t = 1, \ldots, n$.

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

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

for each pair $(F^{(k-1)}, F^{(k)}) \in \mathcal{G}_{k-1} \times \mathcal{G}_k$ satisfying the condition $\rho(F^{(k)}, F^{(k-1)}) \le a_k/2$. By definition of the metric ρ and by the above condition,

$$\frac{1}{a_k}\left| F_t^{(k)}(y^{t-1}) - F_t^{(k-1)}(y^{t-1}) \right| \le \frac{1}{a_k}\rho(F^{(k)}, F^{(k-1)}) \le \frac{1}{2} .$$

This implies that each $Q \in \mathcal{Q}_k$ satifies $0 \le Q_t(y^{t-1}) \le 1$ and, therefore, is indeed a *bona fide* expert.

Now, for each $k \ge 1$ we choose a predictor $P^{(k)}$, for example, the one defined in Theorem 1, such that its minimax regret with respect to the finite class $\mathcal{Q}_k$ achieves the bound (1). Namely, for all $y^n \in \{0,1\}^n$

$$L_{P^{(k)}}(y^n) - \min_{Q\in\mathcal{Q}_k} L_{Q^{(k)}}(y^n) \le \sqrt{\frac{n}{2}\ln|\mathcal{Q}_k|} \le \sqrt{n \ln N_k} . \tag{6}$$

where the last inequality holds because $|\mathcal{Q}_k| \le N_k N_{k-1} \le N_k^2$. Define P by

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

Note that $\sum_{k=1}^\infty a_k = 1$, and therefore P is indeed a valid predictor, taking values in $[0,1]$.

Theorem 7 *For all classes $\mathcal{F}$ of δ-bounded experts and for the predictor P defined above*

$$L_P(y^n) - \inf_{F\in\mathcal{F}} L_F(y^n) \le \frac{1+2\delta}{\delta(1-2\delta)}\sqrt{n}\int_0^1 \sqrt{\ln N_\infty(\mathcal{F},\epsilon)}\, d\epsilon .$$

Proof. We fix a sequence y^n and, to simplify notation, we write F_t instead of $F_t(y^{t-1})$. Let $F^* \in \mathcal{F}$ be an expert for which $L_{F^*}(y^n) = \inf_{F\in\mathcal{F}} L_F(y^n)$. (If no such F^* exists, for any small $\epsilon > 0$ we may consider an F^* such that $L_{F^*}(y^n) < \inf_{F\in\mathcal{F}} L_F(y^n) + \epsilon$ and the same proof works.) Consider the "chain"

$$F^{(0)}, F^{(1)}, \ldots$$

formed by experts $F^{(k)} \in \mathcal{G}_k$ satisfying

$$\rho(F^{(k)}, F^*) \le \epsilon_k , \qquad k \ge 1.$$

Note that such a chain exists by the definition of the $\mathcal{G}_k$'s. Now, as

$$\lim_{k\to\infty} \rho\left(F^{(k)}, F^*\right) = 0 ,$$

we may write

$$\begin{aligned} F_t^* &= \lim_{k\to\infty} F_t^{(k)} \\ &= F_t^{(0)} + \sum_{k=1}^{\infty}\left(F_t^{(k)} - F_t^{(k-1)}\right) \\ &= \frac{1}{2} + \sum_{k=1}^{\infty}\left(F_t^{(k)} - F_t^{(k-1)}\right) \\ &= \sum_{k=1}^{\infty} a_k Q_t^{(k)} \end{aligned}$$

where

$$Q_t^{(k)} = \frac{1}{2} + \frac{1}{a_k}(F_t^{(k)} - F_t^{(k-1)}) . \tag{7}$$

Claim 8 *For each $k \ge 1$, the expert $Q^{(k)}$ defined in (7) belongs to $\mathcal{Q}_k$.*

Proof (of the claim). We must check that $\rho(F^{(k)}, F^{(k-1)}) \le a_k/2$ for each $k \ge 1$. For $k = 1$, $Q^{(1)}$ belongs to $\mathcal{Q}_1$ because

$$\rho(F, F^{(0)}) \le \frac{1}{2} - \delta = \frac{a_1}{2}$$

for each $F \in \mathcal{G}_1$, due to the δ-boundedness of $\mathcal{F}$ (recall that $F^{(0)}$ predicts $1/2$ all the time). Now take $k \ge 2$. By the triangle inequality we clearly have

$$\rho(F^{(k)}, F^{(k-1)}) \le \rho(F^{(k)}, F^*) + \rho(F^*, F^{(k-1)}) \le \epsilon_k + \epsilon_{k-1} = \frac{a_k}{2}$$

and this concludes the proof of the claim. $\square$

Thus, for all $t = 1, \ldots, n$,

$$\begin{aligned} |P_t - y_t| - |F_t^* - y_t| &= \left|\sum_{k=1}^{\infty} a_k P_t^{(k)} - y_t\right| - \left|\sum_{k=1}^{\infty} a_k Q_t^{(k)} - y_t\right| \\ &= \sum_{k=1}^{\infty} a_k \left[\left|P_t^{(k)} - y_t\right| - \left|Q_t^{(k)} - y_t\right| \right] . \qquad (8) \end{aligned}$$

Therefore, summing the above equality over $t = 1, \ldots, n$ and applying (6) to each predictor $P^{(k)}$, $k \geq 1$, we obtain

$$\begin{aligned} L_P(y^n) - \inf_{F \in \mathcal{F}} L_F(y^n) &= \sum_{k=1}^{\infty} a_k \left(L_{P^{(k)}}(y^n) - L_{Q^{(k)}}(y^n) \right) \\ &\leq \sum_{k=1}^{\infty} a_k \left(L_{P^{(k)}}(y^n) - \min_{Q \in \mathcal{Q}_k} L_Q(y^n) \right) \\ &\leq \sum_{k=1}^{\infty} a_k \sqrt{n \ln N_k} \\ &= \sum_{k=1}^{\infty} a_k \sqrt{n \ln N_\infty\left(\mathcal{F}, \epsilon_k\right)} \\ &= \sqrt{n} \sum_{k=1}^{\infty} \frac{1+2\delta}{\delta(1-2\delta)} (\epsilon_k - \epsilon_{k+1}) \sqrt{\ln N_\infty\left(\mathcal{F}, \epsilon_k\right)} \\ &\leq \sqrt{n} \frac{1+2\delta}{\delta(1-2\delta)} \int_0^1 \sqrt{\ln N_\infty\left(\mathcal{F}, \epsilon\right)}\, d\epsilon, \end{aligned}$$

as desired. $\square$

The predictor P can be easily modified so to avoid the infinite sum. Define the new predictor P' by

$$P_t' \stackrel{\text{def}}{=} \sum_{k=1}^{t-1} a_k P_t^{(k)} + \sum_{k=t}^{\infty} a_k . \qquad (9)$$

Here, only $P^{(1)}, \ldots, P^{(n-1)}$ are actually used to predict any sequence y^n. More precisely, each $P^{(k)}$ is only used to predict the $n-k$ bits $y_{k+1}, \ldots, y_n$. The performance of P' is derived from (8) as follows

$$|P_t' - y_t| - |F_t^* - y_t| \leq \sum_{k=1}^{t-1} a_k \left[\left|P_t^{(k)} - y_t\right| - \left|Q_t^{(k)} - y_t\right| \right] + \sum_{k=t}^{\infty} a_k . \qquad (10)$$

Note that

$$\sum_{k=t}^{\infty} a_k = (2\delta)^{t-1} .$$

Therefore, summing (10) over $t = 1, \ldots, n$, we get

$$\begin{aligned} L_{P'}(y^n) - \inf_{F \in \mathcal{F}} L_F(y^n) &\leq \sum_{t=2}^{n} \sum_{k=1}^{t-1} a_k \left[\left|P_t^{(k)} - y_t\right| - \left|Q_t^{(k)} - y_t\right| \right] + \sum_{t=1}^{n} (2\delta)^{t-1} \\ &\leq \sum_{k=1}^{n-1} \sum_{t=k+1}^{n} a_k \left[\left|P_t^{(k)} - y_t\right| - \left|Q_t^{(k)} - y_t\right| \right] + \frac{1}{1-2\delta} \\ &= \sum_{k=1}^{n-1} a_k \left(L_{P^{(k)}}(y_{k+1}^n) - L_{Q^{(k)}}(y_{k+1}^n) \right) + \frac{1}{1-2\delta} \\ &\leq \sum_{k=1}^{n-1} a_k \left(L_{P^{(k)}}(y_{k+1}^n) - \min_{Q \in \mathcal{Q}_k} L_Q(y_{k+1}^n) \right) + \frac{1}{1-2\delta} \\ &\leq \sum_{k=1}^{n-1} a_k \left(\sqrt{(n-k) \ln N_k} \right) + \frac{1}{1-2\delta} . \end{aligned}$$

(Here and in what follows, $y_p^q = (y_p, y_{p+1}, \ldots, y_q)$ for $p \leq q$.)

Hence, concluding the proof as we did in Theorem 7, we obtain the following.

Corollary 9 *For all classes $\mathcal{F}$ of δ-bounded experts and for the predictor P' defined above*

$$L_{P'}(y^n) - \inf_{F \in \mathcal{F}} L_F(y^n) \leq \sqrt{n} \frac{1+2\delta}{\delta(1-2\delta)} \int_0^1 \sqrt{\ln N_\infty(\mathcal{F}, \epsilon)}\, d\epsilon + \frac{1}{1-2\delta} .$$

5 Lower bounds

In this section we derive lower bounds for the minimax regret $V_n(\mathcal{F})$ of general classes of experts. Recall that in the case of static experts, by Theorem 3, we have the following characterization:

$$V_n(\mathcal{F}) = \mathbf{E}\left[\sup_{F \in \mathcal{F}} \sum_{t=1}^{n} \left(\frac{1}{2} - \bar{F}_t \right) (1 - 2Y_t) \right],$$

where the Y_i's are independent Bernoulli (1/2) random variables. Unfortunately, this equality is not true for general classes of experts. However, the right-hand side is always a lower bound for $V_n(\mathcal{F})$, and the following inequality is our starting point.

Theorem 10 *For any expert class $\mathcal{F}$,*

$$V_n(\mathcal{F}) \geq \mathbf{E}\left[\sup_{F \in \mathcal{F}} \sum_{t=1}^{n} \left(\frac{1}{2} - F_t(Y^{t-1}) \right) (1 - 2Y_t) \right],$$

where $Y_1, \ldots, Y_n$ are independent Bernoulli (1/2) random variables,

Proof. For any prediction strategy P, if Y_t is a Bernoulli (1/2) random variable then $\mathbf{E}|P_t(y^{t-1}) - Y_t| = 1/2$ for each y^{t-1}. Hence

$$\begin{aligned} V_n(\mathcal{F}) &\geq R_n(P, \mathcal{F}) \\ &= \max_{y^n \in \{0,1\}^n} \left(L_P(y^n) - \inf_{F \in \mathcal{F}} L_F(y^n) \right) \\ &\geq \mathbf{E}\left[L_P(Y^n) - \inf_{F \in \mathcal{F}} L_F(Y^n) \right] \\ &= \frac{n}{2} - \mathbf{E}\left[\inf_{F \in \mathcal{F}} L_F(Y^n) \right] \\ &= \mathbf{E}\left[\sup_{F \in \mathcal{F}} \sum_{t=1}^{n} \left(\frac{1}{2} - F_t(Y^{t-1}) \right) (1 - 2Y_t) \right] . \end{aligned}$$

□

Theorem 10 will be directly applied in Section 6 to the class of linear experts. Next, we prove a different lower bound on $V_n(\mathcal{F})$ in terms of the packing number of $\mathcal{F}$ with respect to a random metric. This result will be applied in Section 7. We make use of the following notations.

For any class $\mathcal{F}$ of experts and for all $y^n \in \{0,1\}^n$, $\mathcal{F}_{|y^n}$ is the class of static experts F' such that $\bar{F}'_t = F_t(y^{t-1})$.

For each expert F, for all $t = 1, \ldots, n$, and for all $z^{t-1} = (z_1, \ldots, z_{t-1}) \in \{-1,+1\}^{t-1}$, we define $\tilde{F}_t$ by $\tilde{F}_t(z^{t-1}) = 1 - 2F_t\left((1-z_1)/2, \ldots, (1-z_{t-1})/2\right)$.

As we did in the case of static experts, we write the inequality proven in Theorem 10 using the more compact notation

$$V_n(\mathcal{F}) \geq \frac{1}{2}\,\mathbf{E}\left[\sup_{F\in\mathcal{F}} \sum_{t=1}^{n} \tilde{F}_t(Z^{t-1}) Z_t\right] , \tag{11}$$

where $Z_1, \ldots, Z_n$ are now independent Rademacher random variables.

Theorem 11 *Let $\mathcal{F}$ be a class of experts containing two static experts F and G such that $\bar{F}_t = 0$ and $\bar{G}_t = 1$ for all t. If there exists a positive constant c such that*

$$\mathbf{E}\left[\sup_{F\in\mathcal{F}} \sum_{t=1}^{n} \tilde{F}_t(Z^{t-1}) U_t\right] \leq c\,\mathbf{E}\left[\sup_{F\in\mathcal{F}} \sum_{t=1}^{n} \tilde{F}_t(Z^{t-1}) Z_t\right] \tag{12}$$

where $Z_1, \ldots, Z_n, U_1, \ldots, U_n$ are independent Rademacher random variables, then

$$V_n(\mathcal{F}) \geq \frac{K}{2c}\mathbf{E}\left[\sup_r r\sqrt{\ln N_2(\mathcal{F}_{|Y^n}, r)}\right] ,$$

where $Y_1, \ldots, Y_n$ are independent Bernoulli $(1/2)$ random variables and K is the universal constant appearing in Theorem 5. In particular, if in addition to the above, for some $r > 0$, there exists a set $\mathcal{G}_r$ of experts such that with probability at least $1/2$ it is an r-packing with respect to the (random) metric

$$d_{Y^n}(F,G) = \sqrt{\sum_{t=1}^{n} \left(F_t(Y^{t-1}) - G_t(Y^{t-1})\right)^2} ,$$

then

$$V_n(\mathcal{F}) \geq \frac{Kr}{4c}\sqrt{\ln|\mathcal{G}_r|} .$$

Proof. For each fixed $y^n \in \{0,1\}^n$, we can apply Theorem 5 to the static class $\mathcal{F}_{|y^n}$ and prove

$$\mathbf{E}\left[\sup_{F\in\mathcal{F}} \sum_{t=1}^{n} \tilde{F}_t(z^{t-1}) U_t\right] \geq K \sup_r r\sqrt{\ln N_2(\mathcal{F}_{|y^n}, r)} ,$$

where $z_t = 1 - 2y_t$. Then, by averaging over y^n, we get

$$\begin{aligned} V_n(\mathcal{F}) &\geq \frac{1}{2}\mathbf{E}\left[\sup_{F\in\mathcal{F}} \sum_{t=1}^{n} \tilde{F}_t(Z^{t-1}) Z_t\right] && \text{by inequality (11)} \\ &\geq \frac{1}{2c}\mathbf{E}\left[\sup_{F\in\mathcal{F}} \sum_{t=1}^{n} \tilde{F}_t(Z^{t-1}) U_t\right] && \text{by hypothesis} \\ &\geq \frac{K}{2c}\mathbf{E}\left[\sup_r r\sqrt{\ln N_2(\mathcal{F}_{|Y^n}, r)}\right] \end{aligned}$$

concluding the proof of the first statement. The second statement is a trivial consequence. □

6 Linear predictors

In this section we study the class $\mathcal{L}_k$ of k-th order autoregressive linear predictors, where $k \geq 2$ is a fixed positive integer. As each prediction of a k-th order autoregressive linear predictor is determined by the last k bits observed, we add an arbitrary prefix $y_{-k+1}, \ldots, y_0$ to the sequence y^n to be predicted. We use y^n_{1-k} to denote the resulting sequence of $n+k$ bits. Accordingly, for such predictors the minimax regret is re-defined by

$$V_n(\mathcal{F}) = \min_P \max_{y^n_{1-k}\in\{0,1\}^{n+k}} \left(L_P(y^n_{1-k}) - \inf_{F\in\mathcal{F}} L_F(y^n_{1-k}) \right) ,$$

where

$$L_P(y^n_{1-k}) \sum_{t=1}^{n} \left| P_t(y^{t-1}_{1-k}) - y_t \right|$$

and $L_F(y^n_{1-k})$ is defined analogously. Similar remarks hold for the next section.

The class $\mathcal{L}_k$ contains all experts F such that

$$F_t(y^{t-1}_{1-k}) = \sum_{i=1}^{k} q_i y_{t-i}$$

for some $q_1, \ldots, q_k \geq 0$ with $\sum_{i=1}^{k} q_i = 1$. In other words, an expert F predicts according to a convex combination of the k most recent outcomes of the sequence. Convexity of the coefficients q_i assures that $F_t(y^{t-1}_{1-k}) \in [0,1]$. The same class of experts (without the convexity assumption) was considered also by Singer and Feder [24] who studied a rather different problem. The main result of this section determines the exact order of magnitude of the minimax regret for $\mathcal{L}_k$.

Theorem 12 *For any positive integers n and $k \geq 2$,*

$$V_n(\mathcal{L}_k) \leq \sqrt{\frac{n \ln k}{2}} \approx 0.707\sqrt{n \ln k}$$

and for all $k > 5$

$$\liminf_{n\to\infty} \frac{V_n(\mathcal{L}_k)}{\sqrt{n}} \geq \left(\frac{1}{4} - \frac{1}{4e}\right)\sqrt{\ln k} - \sqrt{\frac{1}{8\pi}} \approx 0.158\sqrt{\ln k} - 0.199 \ . \tag{13}$$

Moreover,

$$\liminf_{k\to\infty}\liminf_{n\to\infty} \frac{V_n(\mathcal{L}_k)}{\sqrt{n\ln k}} = \frac{1}{\sqrt{2}} \approx 0.707.$$

Remarks. For small k, the lower bound (13) holds vacuously. However, for all $k \geq 2$, one can prove that $\liminf_{n\to\infty} V_n(\mathcal{L}_k)/\sqrt{n} \geq 1/6$. With more work is also possible to obtain nonasymptotical lower bounds of the "right" order $\sqrt{n \ln k}$ by studying the rate of convergence in the martingale central theorem used in the proof below. Such nonasymptotical bounds will be derived for the class Markov experts in Section 7. Finally note that the last statement implies that the upper bound cannot be improved: the constant $1/\sqrt{2}$ is optimal.

Proof. The first statement is a straighforward consequence of Theorem 2 if we observe that $\mathcal{L}_k$ is the convex hull of the k experts $F^{(1)}, \ldots, F^{(k)}$ defined by

$$F_t^{(i)}(y^{t-1}) = y_{t-i}, \quad i = 1, \ldots, k.$$

We prove the second statement by applying directly Theorem 10 in the compact form (11),

$$V_n(\mathcal{L}_k) \geq \frac{1}{2}\mathbf{E}\left[\sup_{F\in\mathcal{L}_k}\sum_{t=1}^n \tilde{F}_t(Z^{t-1})Z_t\right] = \frac{1}{2}\mathbf{E}\left[\max_{1\leq i\leq k}\sum_{t=1}^n Z_t Z_{t-i}\right]$$

where $Z_{-k+1}, \ldots, Z_n$ are independent Rademacher variables and the last step holds simply because a linear function over a convex polygon takes its maximum in one of the vertices of the polygon.

Consider now the k-vector $X_n = (X_{n,1}, \ldots, X_{n,k})$ of components

$$X_{n,i} \stackrel{\text{def}}{=} \frac{1}{\sqrt{n}}\sum_{t=1}^n Z_t Z_{t-i}, \quad i = 1, \ldots, k.$$

By the "Cramér-Wold device" (see, e.g., Billingsley [2, p. 48]), the sequence of vectors $\{X_n\}$ converges in distribution to a vector random variable $N = (N_1, \ldots, N_k)$ if and only if $\sum_{i=1}^k a_i X_{n,i}$ converges in distribution to $\sum_{i=1}^k a_i N_i$ for all possible choices of the coefficients $a_1, \ldots, a_k$. Thus, consider

$$\sum_{i=1}^k a_i X_{n,i} = \frac{1}{\sqrt{n}}\sum_{t=1}^n Z_t \sum_{i=1}^k a_i Z_{t-i}.$$

It is easy to see that the sequence of random variables $\sqrt{n}X_{n,i}$, $n = 1, 2, \ldots$, forms a martingale with respect to the sequence of σ-algebras $\mathcal{G}_t$ generated by $Z_{-k+1}, \ldots, Z_t$. Furthermore, by the martingale central limit theorem (see, e.g., Hall and Heyde [15, Theorem 3.2]) $\sum_{i=1}^k a_i X_{n,i}$ converges in distribution, as $n \to \infty$, to a zero-mean normal random variable

with variance $\sum_{i=1}^k a_i^2$. Then, by the Cramér-Wold device, as $n \to \infty$ the vector X_n converges in distribution to $N = (N_1, \ldots, N_k)$, where $N_1, \ldots, N_k$ are independent standard normal random variables.

Convergence in distribution implies that for any bounded continuous function $\psi : \mathcal{R}^k \to \mathcal{R}$,

$$\lim_{n\to\infty} \mathbf{E}\left[\psi(X_{n,1}, \ldots, X_{n,k})\right] = \mathbf{E}\left[\psi(N_1, \ldots, N_k)\right]. \tag{14}$$

Consider, in particular, the function $\psi(x_1, \ldots, x_k) = \phi_L(\max_i x_i)$, where $L > 0$, and ϕ_L is the "thresholding" function

$$\phi_L(x) = \begin{cases} -L & \text{if } x < -L \\ x & \text{if } |x| \leq L \\ L & \text{if } x > L. \end{cases}$$

Clearly, ϕ_L is bounded and continuous. Hence, by (14), we conclude

$$\lim_{n\to\infty} \mathbf{E}\left[\phi_L\left(\max_{1\leq i\leq k} X_{n,i}\right)\right] = \mathbf{E}\left[\phi_L\left(\max_{1\leq i\leq k} N_i\right)\right] \ .$$

Now note that for any $L > 0$,

$$\mathbf{E}\left[\max_{1\leq i\leq k} X_{n,i}\right] \geq \mathbf{E}\left[\phi_L\left(\max_{1\leq i\leq k} X_{n,i}\right)\right] + \mathbf{E}\left[\left(L + \max_{1\leq i\leq k} X_{n,i}\right) I_{\{\max_{1\leq i\leq k} X_{n,i} < -L\}}\right],$$

where

$$\begin{aligned}
\left|\mathbf{E}\left[\left(L + \max_{1\leq i\leq k} X_{n,i}\right) I_{\{\max_{1\leq i\leq k} X_{n,i} < -L\}}\right]\right| &\leq \mathbf{E}\left[\left(\left|\max_{1\leq i\leq k} X_{n,i}\right| - L\right) I_{\{|\max_{1\leq i\leq k} X_{n,i}| - L > 0\}}\right] \\
&= \int_0^\infty \mathbf{P}\left\{\left|\max_{1\leq i\leq k} X_{n,i}\right| > L + u\right\} du \\
&= \int_L^\infty \mathbf{P}\left\{\left|\max_{1\leq i\leq k} X_{n,i}\right| > u\right\} du \\
&\leq \int_L^\infty k \max_{1\leq i\leq k} \mathbf{P}\left\{|X_{n,i}| > u\right\} du \\
&\leq 2k\int_L^\infty e^{-u^2/2} du \\
&\quad \text{(by the Hoeffding-Azuma inequality)} \\
&\leq 2k\int_L^\infty \left(1 + \frac{1}{u^2}\right) e^{-u^2/2} du \\
&= \frac{2k}{L} e^{-L^2/2}.
\end{aligned}$$

Therefore, we have that for any $L > 0$,

$$\liminf_{n\to\infty} \mathbf{E}\left[\max_{1\leq i\leq k} X_{n,i}\right] \geq \mathbf{E}\left[\phi_L\left(\max_{1\leq i\leq k} N_i\right)\right] - \frac{2k}{L} e^{-L^2/2} \ .$$

Letting $L \to \infty$ on the right-hand side, we see that

$$\liminf_{n\to\infty} \mathbf{E}\left[\max_{1\leq i\leq k} X_{n,i}\right] \geq \mathbf{E}\left[\max_{1\leq i\leq k} N_i\right] \ .$$

(Note that one can similarly show that, in fact, $\mathbf{E}\left[\max_{1\le i\le k} X_{n,i}\right] \to \mathbf{E}\left[\max_{1\le i\le k} N_i\right]$ as $n\to\infty$.) Using a standard estimate for the expected value of the maximum of k independent standard normal variables (see, e.g., [18, p. 80]), which holds for $k > 5$, we obtain

$$\liminf_{n\to\infty} \frac{V_n(\mathcal{L}_k)}{\sqrt{n}} \ge \frac{1}{2}\mathbf{E}\left[\max_{1\le i\le k} N_i\right] \ge \left(\frac{1}{4} - \frac{1}{4e}\right)\sqrt{\ln k} - \sqrt{\frac{1}{8\pi}} .$$

The last statement now follows from the fact that

$$\lim_{k\to\infty} \frac{\mathbf{E}\left[\max_{1\le i\le k} N_i\right]}{\sqrt{\ln k}} = \sqrt{2},$$

see, for example, Galambos [12]. □

7 Markov experts

This section is devoted to an important family of examples (i.e., k-th order Markov experts), another example of how the upper and lower bounds obtained in Sections 4 and 5 may be used in concrete situations. The same class of experts was also considered by Feder, Merhav, and Gutman [11], who derived an upper bound. The main novelty of this section is a matching nonasymptotical lower bound, obtained via Theorem 11, revealing the exact value (to within constant factors) of the minimax regret for Markov experts.

For an arbitrary $k \ge 1$, we consider the class $\mathcal{M}_k$ of experts that, when considered as probability measures on $\{0,1\}^n$, represent all stationary k-th order Markov measures. The rigorous definition is as follows (as each prediction of a k-th order Markov expert is determined by the last k bits observed, we add a prefix $y_{-k+1},\dots,y_0$ to the sequence to predict in the same way we did in the previous section for the autoregressive experts): The class $\mathcal{M}_k$ is indexed by the set $[0,1]^{2^k}$ so that the index of any $F \in \mathcal{M}_k$ is the vector $f = (f_0, f_1, \dots, f_{2^k-1})$ with $f_s \in [0,1]$ for $0 \le s < 2^k$. If F has index f then $F_t(y_{-k+1}^{t-1}) = f_s$ for all $1 \le t \le n$ and for all $y_{-k+1}^{t-1} \in \{0,1\}^{t+k-1}$, where s has binary expansion $y_{t-k},\dots,y_{t-1}$. (Note that, due to the need of adding a prefix $y_{-k+1},\dots,y_0$ to the sequence to predict, the function F_t is now defined over the set $\{0,1\}^{t+k-1}$.)

As mentioned in Section 1, Feder *et al.* [11] showed that

$$V_n(\mathcal{M}_k) \le C\sqrt{2^k n} ,$$

where C is a universal constant. Interestingly, both Theorem 2 and Theorem 7 imply the same upper bound. The best constant $C = \sqrt{(\ln 2)/2}$ is achieved by Theorem 2. (To see why Theorem 7 implies a bound of the same order of magnitude, just observe that $N_\infty(\mathcal{F},\delta) \le \delta^{-2^k}$ for all $\delta \in (0,1)$.) We now complement this result by showing a matching lower bound on $V_n(\mathcal{M}_k)$ that holds for all $k \ge 1$ and for all sufficiently large n.

Theorem 13 *There exist universal constants C_1 and C_2 such that for all $k \ge 1$ and for all $n \ge C_1 k^2 2^{2k}$*

$$V_n(\mathcal{M}_k) \ge C_2\sqrt{2^k n}$$

Remark. In the proof, we will show that the theorem holds for $C_1 = 78941$ and $C_2 = K/690$ (where K is the universal constant appearing in Theorem 5) provided that n is also a multiple of 2^k. It should be clear from the analysis that this condition on n could be indeed removed by a suitable increase of C_1 and C_2.

We prove this theorem by applying the lower bound of Theorem 11. However, instead of checking directly that Markov experts satisfy condition (12), we proceed as follows: First, we define two simple properties (symmetry and contraction) whose conjunction is shown to imply condition (12). Second, we prove that Markov experts have both of these properties.

Definition 14 (Symmetry) *An expert class $\mathcal{F}$ is* symmetric *if for each $F \in \mathcal{F}$ and for each $y^n \in \{0,1\}^n$ there exists $F' \in \mathcal{F}$ such that $F'_t(y^{t-1}) = 1 - F_t(y^{t-1})$ for each $t = 1,\dots,n$.*

This condition is quite mild, and even if it is not satisfied, one may easily "symmetrize" the expert class by adding to $\mathcal{F}$ a "symmetric" expert $F' = 1 - F$ for each $F \in \mathcal{F}$. This operation just slightly increases the size of $\mathcal{F}$.

Definition 15 (Contraction) *Let c be a positive constant. An expert class $\mathcal{F}$ is* c-contractive *if*

$$\mathbf{E}\left[\sup_{F\in\mathcal{F}} \sum_{t=1}^n \tilde{F}_t(Z^{t-1}) Z_t Y_t\right] \le c\,\mathbf{E}\left[\sup_{F\in\mathcal{F}} \sum_{t=1}^n \tilde{F}_t(Z^{t-1}) Z_t\right]$$

where $Z_1,\dots,Z_n,Y_1,\dots,Y_n$ are independent random variables such that each Z_t is Rademacher and each Y_t is Bernoulli $(1/2)$.

The next result shows that symmetry and c-contraction imply condition (12) with constant $2c+1$.

Lemma 16 *If an expert class $\mathcal{F}$ is symmetric, c-contractive, and contains some static expert F such that $\tilde{F}_t = 1$ for all t, then*

$$\mathbf{E}\left[\sup_{F\in\mathcal{F}} \sum_{t=1}^n \tilde{F}_t(Z^{t-1}) U_t\right] \le (2c+1)\,\mathbf{E}\left[\sup_{F\in\mathcal{F}} \sum_{t=1}^n \tilde{F}_t(Z^{t-1}) Z_t\right]$$

where $Z_1,\dots,Z_n,U_1,\dots,U_n$ are independent Rademacher random variables.

Proof. Consider the chain of inequalities

$$\begin{aligned}
&(2c+1)\mathbf{E}\left[\sup_{F\in\mathcal{F}} \sum_{t=1}^n \tilde{F}_t(Z^{t-1}) Z_t\right] - \mathbf{E}\left[\sup_{F\in\mathcal{F}} \sum_{t=1}^n \tilde{F}_t(Z^{t-1}) U_t\right] \\
&= 2c\mathbf{E}\left[\sup_{F\in\mathcal{F}} \sum_{t=1}^n \tilde{F}_t(Z^{t-1}) Z_t\right] - \left(\mathbf{E}\left[\sup_{F\in\mathcal{F}} \sum_{t=1}^n \tilde{F}_t(Z^{t-1}) U_t\right] - \mathbf{E}\left[\sup_{F\in\mathcal{F}} \sum_{t=1}^n \tilde{F}_t(Z^{t-1}) Z_t\right]\right) \\
&\ge 2c\mathbf{E}\left[\sup_{F\in\mathcal{F}} \sum_{t=1}^n \tilde{F}_t(Z^{t-1}) Z_t\right] - \mathbf{E}\left[\sup_{F\in\mathcal{F}} \sum_{t=1}^n \tilde{F}_t(Z^{t-1}) (U_t - Z_t)\right] .
\end{aligned}$$

Now pick independent Bernoulli (1/2) random variables $Y_1,\ldots,Y_n$. We further bound as follows.

$$
\begin{aligned}
&2c\mathbf{E}\left[\sup_{F\in\mathcal{F}}\sum_{t=1}^n \tilde{F}_t(Z^{t-1})Z_t\right] - \mathbf{E}\left[\sup_{F\in\mathcal{F}}\sum_{t=1}^n \tilde{F}_t(Z^{t-1})(U_t-Z_t)\right] \\
&= 2c\mathbf{E}\left[\sup_{F\in\mathcal{F}}\sum_{t=1}^n \tilde{F}_t(Z^{t-1})Z_t\right] - 2\mathbf{E}\left[\sup_{F\in\mathcal{F}}\sum_{t=1}^n \tilde{F}_t(Z^{t-1})(-Z_t)Y_t\right] \qquad (15)\\
&= 2c\mathbf{E}\left[\sup_{F\in\mathcal{F}}\sum_{t=1}^n \tilde{F}_t(Z^{t-1})Z_t\right] - 2\mathbf{E}\left[\sup_{F\in\mathcal{F}}\sum_{t=1}^n \tilde{F}_t(Z^{t-1})Z_tY_t\right] \qquad (16)\\
&\geq 0\ . \qquad (17)
\end{aligned}
$$

To show (15) fix any z^n and notice that, for each t, the distribution of $U_t - z_t$ is the same as the distribution of $-2Y_tz_t$. Finally, symmetry of $\mathcal{F}$ implies (16) and contractiveness implies (17). $\square$

As $\mathcal{M}_k$ is clearly symmetric and contains some static expert F such that $\bar{F}_t = 1$ for all t, all we have to show, by Lemma 16, is that $\mathcal{M}_k$ is contractive, and then the existence of a packing set $\mathcal{G}_r$ with the required property. These properties are stated by the next two lemmas.

Lemma 17 *For all $k \geq 1$ and all $n \geq 78941k^2 2^{2k}$ multiple of 2^k, $\mathcal{M}_k$ is $\left(10\sqrt{2}\right)$-contractive.*

To prove this result we need some preliminary definitions and a few technical sublemmas. Fix any $k \geq 1$. For each $s \in \{-1,1\}^k$, for each $z = z_{-k+1}^n \in \{-1,1\}^{n+k}$, and for each $t = 1,\ldots,n$ define

$$
a_t(s,z) \stackrel{\text{def}}{=} \begin{cases} 1 & \text{if } z_{t-k}^{t-1} = s \text{ and } z_t = 1, \\ -1 & \text{if } z_{t-k}^{t-1} = s \text{ and } z_t = -1, \\ 0 & \text{otherwise.} \end{cases}
$$

Recall that, by definition of k-th order Markov expert, for any $F \in \mathcal{M}_k$ the quantity $\tilde{F}_t(z_{-k+1}^{t-1})$ depends only on the subsequence z_{t-k}^{t-1}. Hence,

$$
\max_{F\in\mathcal{M}_k}\sum_{t=1}^n \tilde{F}_t(z_{-k+1}^{t-1})z_t = \sum_{s\in\{-1,1\}^k}\left|\sum_t a_t(s,z)\right| \ .
$$

So, showing that $\mathcal{M}_k$ is $\left(10\sqrt{2}\right)$-contractive amounts to showing that

$$
\sum_{s\in\{-1,1\}^k}\mathbf{E}\left|\sum_t a_t(s,Z)Y_t\right| \leq \left(10\sqrt{2}\right)\sum_{s\in\{-1,1\}^k}\mathbf{E}\left|\sum_t a_t(s,Z)\right| \qquad (18)
$$

where $Z = (Z_{-k+1},\ldots,Z_n)$ is a vector of $n+k$ independent Rademacher random variables. Define $m(s,z) = \sum_{t=1}^n |a_t(s,z)|$, so $m(s,z)$ is just the number of times s occurs in z. We now prove the following.

Lemma 18 *For all $s \in \{-1,1\}^k$,*

$$
\begin{aligned}
\mathbf{Var}\left|\sum_t a_t(s,Z)\right| &\leq \mathbf{E}\left[\left(\sum_t a_t(s,Z)\right)^2\right] = \mathbf{E}[m(s,Z)]\ , \qquad (19)\\
\mathbf{E}\left|\sum_t a_t(s,Z)Y_t\right| &\leq \sqrt{\frac{\mathbf{E}[m(s,Z)]}{2}}\ , \qquad (20)\\
\mathbf{E}[m(s,Z)] &= \frac{n}{2^k}\ . \qquad (21)
\end{aligned}
$$

Proof. We start with (19). Note that

$$
\begin{aligned}
\mathbf{E}\left[\left(\sum_t a_t(s,Z)\right)^2\right] &= \mathbf{E}\left[\sum_t a_t(s,Z)^2\right] + \mathbf{E}\left[\sum_{t\neq v} a_t(s,Z)a_v(s,Z)\right] \\
&= \mathbf{E}[m(s,Z)] + \sum_{t\neq v}\mathbf{E}\left[a_t(s,Z)a_v(s,Z)\right]\ .
\end{aligned}
$$

To investigate one term of the sum on the right-hand side, assume without loss of generality that $t > v$ and write

$$
\begin{aligned}
\mathbf{E}\left[a_t(s,Z)a_v(s,Z)\right] &= \mathbf{E}\left[\mathbf{E}\left[a_t(s,Z)a_v(s,Z) \mid Z_{-k+1}^{t-1}\right]\right] \\
&= \mathbf{E}\left[a_v(s,Z)\mathbf{E}\left[a_t(s,Z) \mid Z_{-k+1}^{t-1}\right]\right] \\
&\qquad \text{(since } a_v(s,Z) \text{ is determined by } Z_{-k+1}^{t-1}\text{)} \\
&= 0
\end{aligned}
$$

and this concludes the proof of (19). To prove (20) fix $z \in \{-1,1\}^{n+k}$ and consider the chain of inequalities

$$
\begin{aligned}
\mathbf{E}\left|\sum_t a_t(s,z)Y_t\right| &\leq \sqrt{\mathbf{E}\left[\left(\sum_t a_t(s,z)Y_t\right)^2\right]} \\
&= \sqrt{\mathbf{E}\left[\sum_t a_t(s,z)^2Y_t^2\right] + \mathbf{E}\left[\sum_{t\neq v} a_t(s,z)a_v(s,z)Y_tY_v\right]} \\
&= \sqrt{\frac{1}{2}\sum_t a_t(s,z)^2 + \frac{1}{4}\sum_{t\neq v} a_t(s,z)a_v(s,z)} \\
&= \frac{1}{2}\sqrt{\left(\sum_t a_t(s,z)\right)^2 + m(s,z)}\ .
\end{aligned}
$$

Averaging both sides with respect to $z \in \{-1,1\}^{n+k}$ yields

$$
\mathbf{E}\left|\sum_t a_t(s,Z)Y_t\right| \leq \frac{1}{2}\mathbf{E}\sqrt{\left(\sum_t a_t(s,Z)\right)^2 + m(s,Z)}
$$

$$\le \frac{1}{2}\sqrt{\mathbf{E}\left[\left(\sum_t a_t(s,Z)\right)^2\right] + \mathbf{E}[m(s,Z)]}$$
$$= \sqrt{\frac{\mathbf{E}[m(s,Z)]}{2}} .$$

Finally, to prove (21) just observe

$$\mathbf{E}[m(s,Z)] = \sum_{t=1}^{n} \mathbf{P}\left\{Z_{t-k}^{t-1} = s\right\} = n/2^k .$$

□

Lemma 19 *For all $n \ge 78941k^2 2^{2k}$ multiple of 2^k and all $s \in \{-1,1\}^k$,*

$$\mathbf{P}\left\{\frac{\left|\sum_{t=1}^n a_t(s,Z)\right|}{\sqrt{\mathbf{E}[m(s,Z)]}} \ge \frac{1}{4}\right\} \ge \frac{1}{5} .$$

Proof. Recall that the functions $a_1,\ldots,a_n$ are such that for all $s \in \{-1,1\}^k$ and $z \in \{-1,1\}^{n+k}$, each $a_t(s,z)$ is determined by s and $(z_{t-k},\ldots,z_t)$ only. Hence, we can extend each function's domain by adding infinitely many arbitrary components $z_{n+1}, z_{n+2},\ldots$ to the vector z without affecting the function's value on a pair (s,z). Call $a'_1,\ldots,a'_n$ these extended functions, which are now defined on pairs (s,z') where $z' = (z'_{-k+1},\ldots,z'_0,z'_1,\ldots)$. Thus, we have that for each such pair (s,z'),

$$a'_t(s,z') = a_t\left(s,(z')_{-k+1}^n\right)$$

where, as usual, $(z')_{-k+1}^n = (z'_{-k+1},\ldots,z'_n)$. As the function m is the sum of $a_1,\ldots,a_n$, we can extend m to m' so that $m'(s,z') = \sum_{t=1}^n |a'_t(s,z')|$. Now, if $Z = (Z_{-k+1},\ldots,Z_n)$ is a vector of $n+k$ independent Rademacher r.v.'s and $Z' = (Z'_{-k+1},\ldots,Z'_0,Z'_1,\ldots)$ is a vector of infinitely many independent Rademacher r.v.'s, then

$$\begin{aligned}
\mathbf{P}\left\{\frac{\left|\sum_{t=1}^n a_t(s,Z)\right|}{\sqrt{\mathbf{E}[m(s,Z)]}} \ge \frac{1}{4}\right\} &= \mathbf{P}\left\{\frac{\left|\sum_{t=1}^n a_t\left(s,(Z')_{-k+1}^n\right)\right|}{\sqrt{\mathbf{E}\left[m\left(s,(Z')_{-k+1}^n\right)\right]}} \ge \frac{1}{4}\right\} \\
&= \mathbf{P}\left\{\frac{\left|\sum_{t=1}^n a'_t(s,Z')\right|}{\sqrt{\mathbf{E}\left[m'(s,Z')\right]}} \ge \frac{1}{4}\right\} . \qquad (22)
\end{aligned}$$

We now lower bound (22). To keep the notation tidy, in the rest of the proof we will write a_t, z_t, and Z_t instead of a'_t, z'_t, and Z'_t.

Fix any $s \in \{-1,1\}^k$. Define the random variables $T_1,\ldots,T_n$ such that $T_i(z) = t$ iff z_{t-k+1}^t is the i-th occurrence of s in z. More formally,

$$T_1(z) = \inf\{t \ge 0 \,:\, z_p = s_p,\ t-k < p \le t\}$$

and, inductively for each $i > 1$,

$$T_i(z) = \inf\{t > T_{i-1}(z) \,:\, z_p = s_p,\ t-k < p \le t\} .$$

Now, for all $1 \le \ell \le n$, let

$$S_\ell(z) \stackrel{\text{def}}{=} \sum_{i=1}^{\ell} z_{T_i(z)+1} .$$

Then

$$\sum_{t=1}^n a_t(s,z) = \sum_{i=1}^{m(s,z)} z_{T_i(z)+1} = S_{m(s,z)}(z) .$$

To control $S_{m(s,z)}(z)$, we use a technique due to Doeblin and Ascombe (see, e.g., [6, Theorem 1, page 322]). The use of this technique relies on the following key observation.

Claim 20 *For any nonnegative integer m, the random variables $Z_{T_1+1},\ldots,Z_{T_m+1}$ are independent Rademacher.*

Proof (of the claim). We use the following idea due to Merhav and Weissman [21]. Consider the nested sequence $\{\Omega,\emptyset\} = \mathcal{A}_0 \subseteq \mathcal{A}_1 \subseteq \ldots$ of σ-algebras, where each $\mathcal{A}_t$ is the σ-algebra generated by the r.v.'s $Z_1,\ldots,Z_t$. To see that $Z_{T_1+1},\ldots,Z_{T_m+1}$ are indeed Rademacher note that

$$\begin{aligned}
\mathbf{P}\{Z_{T_i+1} = 1\} &= \sum_{t=0}^{\infty} \mathbf{P}\{Z_{T_i+1} = 1 \mid T_i = t\}\,\mathbf{P}\{T_i = t\} \\
&= \sum_{t=0}^{\infty} \mathbf{P}\{Z_{t+1} = 1 \mid T_i = t\}\,\mathbf{P}\{T_i = t\} \\
&= \frac{1}{2}\sum_{t=0}^{\infty} \mathbf{P}\{T_i = t\} \\
&\qquad \text{(as } \{T_i = t\} \in \mathcal{A}_t \text{ and } Z_{t+1} \text{ is independent of } \mathcal{A}_t\text{)} \\
&= \frac{1}{2} .
\end{aligned}$$

Independence is proven by induction as follows: The random variable Z_{T_1+1} is trivially independent. Now assume by induction that $Z_{T_1+1},\ldots,Z_{T_j+1}$ are independent. By construction, $Z_{T_1+1},\ldots,Z_{T_j+1}$ are all $\mathcal{A}_{T_j+1}$-measurable. Now the fact that $Z_{T_{j+1}+1}$ is independent of $\mathcal{A}_{T_j+1}$ (and therefore of $Z_{T_1+1},\ldots,Z_{T_j+1}$) follows from observing that $T_{j+1} > T_j$ and from applying [6, Lemma 2, page 138] to the finite stopping time T_j. This concludes the proof of the Claim. □

Let $k_n = \mathbf{E}[m(s,Z)]$. Note that (21) in Lemma 18 and the assumption that 2^k divides n imply that k_n is integer. We have

$$\begin{aligned}
\mathbf{P}\left\{\frac{\left|\sum_{t=1}^n a_t(s,Z)\right|}{\sqrt{\mathbf{E}[m(s,Z)]}} \ge \frac{1}{4}\right\} &= \mathbf{P}\left\{\frac{\left|S_{m(s,Z)}(Z)\right|}{\sqrt{k_n}} \ge \frac{1}{4}\right\} \\
&\ge \mathbf{P}\left\{|S_{k_n}(Z)| \ge \frac{\sqrt{k_n}}{2} \text{ and } |S_{m(s,Z)}(Z) - S_{k_n}(Z)| \le \frac{\sqrt{k_n}}{4}\right\} \qquad (23) \\
&\ge \mathbf{P}\left\{|S_{k_n}(Z)| \ge \frac{\sqrt{k_n}}{2}\right\} - \mathbf{P}\left\{|S_{m(s,Z)}(Z) - S_{k_n}(Z)| \ge \frac{\sqrt{k_n}}{4}\right\} \qquad (24)
\end{aligned}$$

where (23) holds since

$$|S_{m(s,Z)}(Z)| \geq |S_{k_n}(Z)| - |S_{m(s,Z)}(Z) - S_{k_n}(Z)|$$

and (24) holds since $\mathbf{P}\{A \cap B\} \geq \mathbf{P}\{A\} - \mathbf{P}\{B^c\}$. Note that $S_{k_n}(Z) = \sum_{i=1}^{k_n} Z_{T_i}(Z)$, where k_n is constant and, by Claim 20, $Z_{T_1}, \ldots, Z_{T_{k_n}}$ are independent Rademacher random variables. Hence, if I_A is the indicator function of the event A,

$$\begin{aligned}
\mathbf{P}\left\{|S_{k_n}(Z)| \geq \frac{\sqrt{k_n}}{2}\right\} &\geq 2\,\mathbf{P}\left\{\sum_{i=1}^{k_n} I_{\{Z_{T_i}=1\}} \geq \sum_{i=1}^{k_n} I_{\{Z_{T_i}=-1\}} + \frac{\sqrt{k_n}}{2}\right\} \\
&= 2\,\mathbf{P}\left\{\sum_{i=1}^{k_n} I_{\{Z_{T_i}=1\}} \geq \frac{k_n}{2} + \frac{\sqrt{k_n}}{4}\right\} \\
&\geq 2\left(1 - \Phi(1/2) - \frac{1}{\sqrt{k_n}}\right) \qquad (25)
\end{aligned}$$

where the last inequality follows from the Berry-Esséen theorem (see Chow and Teicher [6, Corollary 4, page 305] where we took $c_\delta = 1$) with Φ being the Normal distribution function. As $1-\Phi(1/2) > 3/10$, the quantity in (25) is at least $2/5$ for $k_n \geq 100$, that is for $n \geq (100)2^k$. Now, for any $\alpha > 0$,

$$\begin{aligned}
&\mathbf{P}\left\{|S_{m(s,Z)}(Z) - S_{k_n}(Z)| \geq \frac{\sqrt{k_n}}{4}\right\} \\
&\leq \mathbf{P}\left\{|S_{m(s,Z)}(Z) - S_{k_n}(Z)| \geq \frac{\sqrt{k_n}}{4} \text{ and } |m(s,Z) - k_n| \leq \alpha k_n\right\} \qquad (26) \\
&\quad + \mathbf{P}\left\{|m(s,Z) - k_n| > \alpha k_n\right\} \; . \qquad (27)
\end{aligned}$$

We start to bound (26) by establishing the following:

$$\left(|S_{m(s,z)}(z) - S_{k_n}(z)| \geq \frac{\sqrt{k_n}}{4}\right) \wedge \left(|m(s,z) - k_n| \leq \alpha k_n\right)$$

implies

$$\begin{aligned}
&\left(\max_{k_n \leq j \leq (1+\alpha)k_n} |S_j(z) - S_{k_n}(z)| \geq \frac{\sqrt{k_n}}{4}\right) \\
&\quad \vee \left(\max_{(1-\alpha)k_n \leq j \leq k_n} |S_j(z) - S_{k_n}(z)| \geq \frac{\sqrt{k_n}}{4}\right) \; .
\end{aligned}$$

Hence we have

$$\begin{aligned}
&\mathbf{P}\left\{|S_{m(s,Z)}(Z) - S_{k_n}(Z)| \geq \frac{\sqrt{k_n}}{4} \;\wedge\; |m(s,Z) - k_n| \leq \alpha k_n\right\} \\
&\leq \mathbf{P}\left\{\max_{k_n \leq j \leq (1+\alpha)k_n} |S_j(Z) - S_{k_n}(Z)| \geq \frac{\sqrt{k_n}}{4}\right\} \\
&\quad + \mathbf{P}\left\{\max_{(1-\alpha)k_n \leq j \leq k_n} |S_j(Z) - S_{k_n}(Z)| \geq \frac{\sqrt{k_n}}{4}\right\} \; .
\end{aligned}$$

Note that, again by Claim 20, $|S_j(Z) - S_{k_n}(Z)|$ is the absolute value of the sum of at most $\lfloor \alpha k_n \rfloor$ independent Rademacher random variables. Hence, by Kolmogorov's inequality,

$$\begin{aligned}
&\mathbf{P}\left\{\max_{k_n \leq j \leq (1+\alpha)k_n} |S_j(Z) - S_{k_n}(Z)| \geq \frac{\sqrt{k_n}}{4}\right\} \\
&\quad + \mathbf{P}\left\{\max_{(1-\alpha)k_n \leq j \leq k_n} |S_j(Z) - S_{k_n}(Z)| \geq \frac{\sqrt{k_n}}{4}\right\} \\
&\leq \frac{16}{k_n}\mathbf{E}\left[\left(S_{k_n+\lfloor \alpha k_n \rfloor}(Z) - S_{k_n}(Z)\right)^2\right] \\
&\quad + \frac{16}{k_n}\mathbf{E}\left[\left(S_{k_n}(Z) - S_{k_n-\lfloor \alpha k_n \rfloor}(Z)\right)^2\right] \\
&\leq 16\alpha + 16\alpha \; .
\end{aligned}$$

Now we bound (27). As $m(s,z)$ can change by at most k by changing the value of z_t for at most one $1 \leq t \leq n$, we can apply McDiarmid's inequality [20] (see also [9, p. 136]) and conclude, recalling that $\mathbf{E}[m(s,Z)] = 2^k/n$ by (21) in Lemma 18,

$$\mathbf{P}\left\{|m(s,Z) - \mathbf{E}[m(s,Z)]| > \alpha \mathbf{E}[m(s,Z)]\right\} \;\leq\; 2\exp\left(-\frac{2\alpha^2 n}{k^2 2^{2k}}\right) = \delta \; .$$

Hence,

$$\mathbf{P}\left\{\frac{|\sum_{t=1}^n a_t(s,Z)|}{\sqrt{\mathbf{E}[m(s,Z)]}} \geq \frac{1}{4}\right\} \geq \frac{2}{5} - 32\alpha - \delta.$$

By choosing $\alpha = 1/165$ and $n \geq 78941k^2 2^{2k}$, we get $\delta \leq 1/165$ implying $32\alpha + \delta \leq 1/5$. □

Proof of Lemma 17. To prove the lemma, by (18) it suffices to show that, for any $s \in \{-1,1\}^k$ and for all n multiple of 2^k and larger or equal than $78941k^2 2^{2k}$,

$$\mathbf{E}\left|\sum_{t=1}^n a_t(s,Z)Y_t\right| \leq \left(10\sqrt{2}\right)\mathbf{E}\left|\sum_{t=1}^n a_t(s,Z)\right| \; .$$

Lemma 19 and Markov's inequality imply

$$\frac{1}{5} \leq \mathbf{P}\left\{\left|\sum_{t=1}^n a_t(s,Z)\right| \geq \frac{1}{4}\sqrt{\mathbf{E}[m(s,Z)]}\right\} \leq \frac{4\mathbf{E}\left|\sum_{t=1}^n a_t(s,Z)\right|}{\sqrt{\mathbf{E}[m(s,Z)]}} \; .$$

Now, from (20) in Lemma 18,

$$\mathbf{E}\left|\sum_{t=1}^n a_t(s,Z)Y_t\right| \leq \frac{\sqrt{\mathbf{E}[m(s,Z)]}}{\sqrt{2}} \;\leq\; \left(10\sqrt{2}\right)\mathbf{E}\left|\sum_{t=1}^n a_t(s,Z)\right|$$

as desired. □

Lemma 21 *If $n \geq 2^{k+5}$, there exists a subset of Markov experts of cardinality $2^{2^k/3}$ which, with probability at least $1/2$, is an $r = \sqrt{n/8}$-packing of $\mathcal{F}$ with respect to the random metric*

$$d_{Y^n}(F,G) = \sqrt{\sum_{t=1}^n \left(F_t(Y^{t-1}) - G_t(Y^{t-1})\right)^2} \; .$$

Proof. The key tool is Gilbert's [13] packing bound which states that if $A(\ell, r)$ is the largest number of sequences of length ℓ in $\{0,1\}^\ell$ such that the Hamming distance (i.e., the number of disagreements) between any two of them is at least $2r+1$, then

$$A(\ell, r) \geq \frac{2^\ell}{\sum_{i=1}^{2r} \binom{\ell}{i}}.$$

In particular,

$$A(\ell, \ell/8) \geq \frac{2^\ell}{\sum_{i=1}^{\ell/4} \binom{\ell}{i}} \geq 2^{\ell - \ell h(1/4)} \geq 2^{\ell/3}, \tag{28}$$

where h is the binary entropy function.

We need to prove the existence of a set

$$\mathcal{G}_r = \{F^{(1)}, \ldots, F^{(M)}\} \subset \mathcal{F}$$

such that

$$\mathbf{P}\left\{ \min_{\substack{i,j \leq M \\ i \neq j}} d_{Y^n}(F^{(i)}, F^{(j)}) > r \right\} \geq \frac{1}{2}$$

where $M = 2^{2^k/3}$ and $r = \sqrt{n/8}$. We choose the packing set $\mathcal{G}_r$ as follows. Let $\mathcal{M}_k'$ contain all $F \in \mathcal{M}_k$ such that $F_t(y^{t-1}) \in \{0,1\}$ for all $1 \leq t \leq n$ and for all y^{t-1}. By the Gilbert lower bound (28), there exists a set $\mathcal{G}_r \subseteq \mathcal{M}_k'$ of cardinality $M = 2^{2^k/3}$ so that any two distinct $F^{(i)}, F^{(j)} \in \mathcal{G}_r$ are indexed by vectors $f^{(i)}, f^{(j)} \in \{0,1\}^{2^k}$ that disagree on at least $2^k/4$ components. Then,

$$\begin{aligned}
\mathbf{E}\left[\left(d_{Y^n}(F^{(i)}, F^{(j)})\right)^2\right] &= \frac{n}{2^k} \sum_{s=0}^{2^k-1} \left(f_s^{(i)} - f_s^{(j)}\right)^2 \\
&\geq \frac{n}{2^k} \frac{2^k}{4} = \frac{n}{4} ,
\end{aligned}$$

and therefore

$$\begin{aligned}
&\mathbf{P}\left\{ \min_{\substack{i,j \leq M \\ i \neq j}} d_{Y^n}(F^{(i)}, F^{(j)}) \leq r \right\} \\
&\leq \left(2^{2^k/3}\right)^2 \max_{i,j \leq M} \mathbf{P}\left\{ \sum_{t=1}^n \left(F_t^{(i)}(Y_{-k+1}^{t-1}) - F_t^{(j)}(Y_{-k+1}^{t-1})\right)^2 \leq r^2 \right\} \\
&\leq 2^{2^k} \max_{i,j \leq M} \mathbf{P}\left\{ \sum_{t=1}^n \left(F_t^{(i)}(Y_{-k+1}^{t-1}) - F_t^{(j)}(Y_{-k+1}^{t-1})\right)^2 \right. \\
&\qquad \left. - \mathbf{E}\left[\left(d_{Y^n}(F^{(i)}, F^{(j)})\right)^2\right] \leq r^2 - \frac{n}{4} \right\} \\
&\qquad \text{(by the above inequality for the expected value)} \\
&= 2^{2^k} \max_{i,j \leq M} \mathbf{P}\left\{ \sum_{t=1}^n \left(F_t^{(i)}(Y_{-k+1}^{t-1}) - F_t^{(j)}(Y_{-k+1}^{t-k})\right)^2 \right. \\
&\qquad \left. - \mathbf{E}\left[\left(d_{Y^n}(F^{(i)}, F^{(j)})\right)^2\right] \leq -\frac{n}{8} \right\} \\
&\qquad \text{(choosing } r^2 = n/8\text{)} \\
&\leq 2^{2^k} e^{-n/32}
\end{aligned}$$

where at the last step we used the Hoeffding-Azuma inequality for sums of bounded martingale differences [1, 17], (see also [9, Theorem 9.1]). This upper bound is less than $1/2$ whenever $n \geq 2^{k+5}$, which is guaranteed by assumption. The proof is now complete. $\square$

8 Conclusion and remarks

In this work we demonstrate that ideas and results from empirical process theory can be successfully applied to the problem of predicting arbitrary binary sequences given a fixed set of experts. For general expert classes, we prove upper and lower bounds on the minimax regret in terms of the metric entropy of the expert class. In the special case of static experts, the prediction problem turns out to be precisely equivalent to a Rademacher process; hence we can prove tighter upper and lower bounds on the corresponding minimax regret. Furthemore, applications of our results to the classes of autoregressive linear predictors, Markov experts, and (static) monotone experts yield bounds that were not apparently obtainable with any of the previous techniques.

As we noted before, the loss function considered here may be interpreted as the expected loss $|\widehat{Y}_t - y_t|$ of a randomized prediction strategy whose prediction at time t is the binary random variable $\widehat{Y}_t$, where $\widehat{Y}_1, \ldots, \widehat{Y}_n$ are independent, and $\mathbf{P}\{\widehat{Y}_t = 1\} = 1 - \mathbf{P}\{\widehat{Y}_t = 0\} = P_t$. Then an obvious question is how the actual (random) loss $\sum_{t=1}^n |\widehat{Y}_t - y_t|$ relates to its expected value $L_P(y^n)$. Luckily, this difference may be easily bounded by general concentration-of-measure inequalities. Since all prediction algorithms considered in this paper calculate P_t by looking at the *expected* losses of the experts up to time $t-1$, it is easy to see that changing the value of one $\widehat{Y}_t$ cannot change the cumulative loss by more than one. Therefore, for example, McDiarmid's inequality [20] (see also [9, p. 136]) implies that for any $u > 0$,

$$\mathbf{P}\left\{ \left| \sum_{t=1}^n |\widehat{Y}_t - y_t| - L_P(y^n) \right| > u \right\} \leq 2e^{-2u^2/n}.$$

In other words, the random loss $\sum_{t=1}^n |\widehat{Y}_t - y_t|$ with very large probability is at most $O(\sqrt{n})$-away from $L_P(y^n)$, regardless of the expert class.

The loss function considered here is by no means the only interesting one. The most popular loss function considered in the literature is the so-called "log loss"

$$-\log\left(P_t(y^{t-1}) I_{\{y_t=1\}} + (1 - P_t(y^{t-1})) I_{\{y_t=0\}}\right),$$

which has several interesting interpretations in coding theory, gambling, and stock-market prediction. Instead of surveying the literature, we refer to the excellent recent review paper of Merhav and Feder [10]. For the log loss with static experts, and under some additional conditions, Opper and Haussler [22] bounded the minimax regret with an expression whose form is similar to Theorem 7. This result was recently extended to more general (nonstatic) experts in [4].

Acknowledgement. We thank Tsachy Weissman for pointing out an error in a previous version of the manuscript. We also thank the reviewers for the insightful comments.

References

[1] K. Azuma. Weighted sums of certain dependent random variables. *Tohoku Mathematical Journal*, 68:357–367, 1967.

[2] P. Billingsley. *Convergence of Probability Measures.* John Wiley, New York, 1968.

[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] N. Cesa-Bianchi and G. Lugosi. Minimax regret under log loss for general classes of experts. In *Proceedings of the 12th Annual Conference on Computational Learning Theory.* ACM Press, 1999.

[5] N. Cesa-Bianchi. Analysis of two gradient-based algorithms for on-line regression. In *Proceedings of the 10th Annual Conference on Computational Learning Theory*, pages 163–170. ACM Press, 1997.

[6] Y.S. Chow and H. Teicher. *Probability Theory, Independence, Interchangeability, Martingales* (2nd edition). Springer-Verlag, New York, 1988.

[7] T.H. Chung. *Minimax Learning in Iterated Games via Distributional Majorization.* PhD thesis, Stanford University, 1994.

[8] T.M. Cover. Behavior of sequential predictors of binary sequences. In *Proceedings of the 4th Prague Conference on Information Theory, Statistical Decision Functions, Random Processes*, pages 263–272. Publishing House of the Czechoslovak Academy of Sciences, Prague, 1965.

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

[10] N. Merhav and M. Feder. Universal prediction. *IEEE Transactions on Information Theory*, 44(6):2124–2147, 1998.

[11] M. Feder, N. Merhav, and M. Gutman. Universal prediction of individual sequences. *IEEE Transactions on Information Theory*, 38:1258–1270, 1992.

[12] J. Galambos. *The Asymptotic Theory of Extreme Order Statistics.* R.E. Kreiger, 1987.

[13] E.N. Gilbert. A comparison of signalling alphabets. *Bell System Technical Journal*, 31:504–522, 1952.

[14] E. Giné, Empirical processes and applications: an overview. *Bernoulli*, 2:1–28, 1996.

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

[16] D. Haussler. Sphere Packing Numbers for Subsets of the Boolean n-cube with Bounded Vapnik-Chervonenkis Dimension. *Journal of Combinatorial Theory, Series A*, 69:217–232, 1995.

[17] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58:13–30, 1963.

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

[19] N. Littlestone and M.K. Warmuth. The weighted majority algorithm. *Information and Computation*, 108:212–261, 1994.

[20] C. McDiarmid. On the method of bounded differences. In *Surveys in Combinatorics 1989*, pages 148–188. Cambridge University Press, Cambridge, 1989.

[21] N. Merhav and T. Weissman. Some results on prediction in the presence of noise. Unpublished manuscript, 1998.

[22] M. Opper and D. Haussler. Worst case prediction over sequences under log loss. In *The Mathematics of Information Coding, Extraction and Distribution*, Springer-Verlag, New York, 1998.

[23] D. Pollard. Asymptotics via empirical processes. *Statistical Science*, 4:341–366, 1989.

[24] A.C. Singer and M. Feder. Universal linear prediction over parameters and model orders. Submitted for publication, 1997.

[25] S.J. Szarek. On the best constants in the Khintchine inequality. *Studia Mathematica*, 63:197–208, 1976.

[26] M. Talagrand. Majorizing measures: the generic chaining. *Annals of Probability*, 24:1049–1103, 1996. (Special Invited Paper).

[27] V.N. Vapnik and A.Y. Chervonenkis. On the Uniform Convergence of Relative Frequencies of Events to their Probabilities. *Theory of Probability and its Applications*, 16:264–280, 1971.

[28] V.G. Vovk. Aggregating strategies. In *Proceedings of the 3rd Annual Workshop on Computational Learning Theory*, pages 372–383, 1990.

[29] V.G. Vovk. A game of prediction with expert advice. *Journal of Computer and System Sciences*, 56:153–173, 1998.