Worst-case Bounds for the Logarithmic Loss of Predictors *

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

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

Abstract. We investigate on-line prediction of individual sequences. Given a class of predictors, the goal is to predict as well as the best predictor in the class, where the loss is measured by the self information (logarithmic) loss function. The excess loss (regret) is closely related to the redundancy of the associated lossless universal code. Using Shtarkov's theorem and tools from empirical process theory, we prove a general upper bound on the best possible (minimax) regret. The bound depends on certain metric properties of the class of predictors. We apply the bound to both parametric and nonparametric classes of predictors. Finally, we point out a suboptimal behavior of the popular Bayesian weighted average algorithm.

Keywords: Universal prediction, universal coding, empirical processes, on-line learning, metric entropy.

1. Introduction

Assume that elements of an arbitrary sequence $y_1, \ldots, y_n$ are revealed one by one, where the elements y_t belong to some set $\mathcal{Y}$, which, in the simplest case, is assumed to be finite. At each time $t = 1, \ldots, n$, before revealing an element y_t, we are asked to assign a probability mass function p_t on $\mathcal{Y}$ and then observe y_t incurring the logarithmic loss $-\ln p_t(y_t)$. Our total loss at the end is the sum of the losses suffered at each round. As we know the prefix $y_1, \ldots, y_{t-1}$ before choosing each probability assignment p_t, we may view each p_t as the conditional $p(\cdot \mid y_1, \ldots, y_{t-1})$ of some joint distribution p that we choose before the game begins. We call p a *prediction strategy*. Any strategy for playing this game is equivalent to a probability distribution on $\mathcal{Y}^n$.

Our goal is to predict (almost) as well as the best strategy in a given "reference" set of strategies. We will call "experts" the strategies in the

* Both authors acknowledge partial support of ESPRIT Working Group EP 27150, Neural and Computational Learning II (NeuroCOLT II).

reference set. In other words, we intend to accumulate a loss not much larger than that of the best expert, regardless of what the sequence $y_1, \ldots, y_n$ might be.

In this paper we investigate the minimum excess loss, with respect to the total loss of the best expert, achievable on any sequence. This quantity, known as minimax regret (under logarithmic loss), will turn out to depend on certain metric properties of the class $\mathcal{F}$ of experts.

It is well-known, via arithmetic coding (Rissanen, [12]), that every sequential prediction strategy may be converted into a sequential lossless source code. Conversely, every uniquely decodable code over $\mathcal{Y}^n$ defines a probability distribution. Thus, the prediction problem under logarithmic loss is formally equivalent to the problem of sequential universal coding in data compression. In this context, the subject of our study is the smallest achievable worst-case redundancy of a sequential lossless code, with respect to a general class of reference codes. The study of the worst-case regret was pioneered by Shtarkov [15], and later studied from various points of view by De Santis *et al.* [14], Vovk [17, 18], Haussler and Barron [8], Weinberger, Merhav and Feder [19], Yamanishi [20], Rissanen [13], Haussler, Kivinen, and Warmuth [9], and others. Merhav and Feder summarize the relevant history in their recent survey [10].

The notion of minimax regret has natural applications in gambling and portfolio selection. This connection was explored by Cover [3], Feder [6], Cover and Ordentlich [4], Barron and Xie [2], and others.

Definitions. Let $\mathcal{Y}$ be a measurable set equipped with a σ-algebra $\mathcal{A}$ and σ-finite measure μ. Let n be any fixed positive integer denoting the length of the sequence or, equivalently, the number of game rounds. Let $\langle \mathcal{Y}^n, \boldsymbol{\mathcal{A}}, \nu \rangle$ denote the probability space obtained as the n-fold product of $\langle \mathcal{Y}, \mathcal{A}, \mu \rangle$. Throughout the paper, all densities on $\mathcal{Y}$ and $\mathcal{Y}^n$ are understood with respect to the measures μ and ν, respectively. Moreover, all integrals are computed over the set $\mathcal{Y}^n$ unless explicitly specified. (If $\mathcal{Y}$ is a countable set, then μ is usually the counting measure, and all densities are understood as probabilities.)

For any integer $t \geq 0$, we use y^t to denote a sequence of t elements from $\mathcal{Y}$ (where y^0 is the empty sequence). In this context, a *prediction strategy* is a density p on $\mathcal{Y}^n$. Upon observing the prefix y^{t-1}, the strategy p uses the conditional density $p(\cdot \mid y^{t-1})$ as a probability assignment for the next element y_t of the sequence.

Fix a class $\mathcal{F}$ of "reference" strategies, called here *experts.* The *worst-case regret* of a strategy p (with respect to the class $\mathcal{F}$) is defined by

$$R_n(p,\mathcal{F}) = \sup_{y^n \in \mathcal{Y}^n} \left(\sum_{t=1}^{n} \ln \frac{1}{p(y_t|y^{t-1})} - \inf_{f\in\mathcal{F}} \sum_{t=1}^{n} \ln \frac{1}{f(y_t|y^{t-1})} \right)$$

or, equivalently, in terms of the joint densities

$$R_n(p,\mathcal{F}) = \sup_{y^n} \ln \frac{\sup_{\mathcal{F}} f(y^n)}{p(y^n)} \ .$$

In other words, $R_n(p,\mathcal{F})$ is the worst-case difference between the log-likelihood of y^n under the density p and the log-likelihood of y^n under its maximum likelihood estimator (MLE) in the class $\mathcal{F}$. The smallest worst-case regret achievable by any predictor is the *minimax regret*

$$R_n(\mathcal{F}) = \inf_{p} \sup_{y^n} \ln \frac{\sup_{\mathcal{F}} f(y^n)}{p(y^n)}$$

where the infimum is taken over all densities p on $\mathcal{Y}^n$.

The main contribution of this paper is a general upper bound on the minimax regret $R_n(\mathcal{F})$ in terms of some metric structure of the expert class $\mathcal{F}$. In previous works, Rissanen [13] obtained asymptotical upper bounds for general parametric classes, which was generalized considerably by Yamanishi [21]. Opper and Haussler [11] were the first to prove upper bounds for nonparametric classes. However, their bounds are restricted to classes of *static* experts, that is, experts which correspond to product distributions. Our main result, Theorem 3 below, extends both results in different ways: (1) in the parametric case we are able to obtain nonasymptotical bounds under much simpler conditions than those needed by Rissanen; (2) in the nonparametric case our results extend those of Opper and Haussler to classes of arbitrary, not just static, experts.

The rest of the paper is organized as follows: In Section 2 we review Shtarkov's optimal prediction strategy p^*, whose regret $R_n(p^*,\mathcal{F})$ is always equal to the minimax regret $R_n(\mathcal{F})$. In Section 3 we establish our main result: a general upper bound on the minimax regret for any class of experts. In Section 4 we apply our upper bound in concrete situations, which could not be handled by any of the previous methods. Finally, in Section 5 we point out that for certain classes of experts, prediction strategies based on mixture of experts may have a regret which is significantly larger than that of Shtarkov's optimal predictor.

2. Shtarkov's theorem, mixture strategies

Shtarkov proved the remarkable fact that the density corresponding to the normalized MLE achieves the minimax regret for any class of experts.

PROPOSITION 1. *(Shtarkov, [15]) For any class $\mathcal{F}$ of experts, the density (normalized MLE)*

$$p^*(y^n) = \frac{\sup_{\mathcal{F}} f(y^n)}{\int \sup_{\mathcal{F}} f(x^n)\, d\nu(x^n)}$$

is a minimax strategy, that is,

$$R_n(p^*, \mathcal{F}) = R_n(\mathcal{F}) \ .$$

Moreover, p^* is an equalizer. *That is, for all $y^n \in \mathcal{Y}^n$*

$$\ln \frac{\sup_{\mathcal{F}} f(y^n)}{p^*(y^n)} = \int \sup_{\mathcal{F}} \ln f(x^n)\, d\nu(x^n) = R_n(\mathcal{F}) \ . \qquad (1)$$

Note that the equalizer property (1) implies that the minimax regret may be expressed as

$$R_n(\mathcal{F}) = \int \left(\sup_{\mathcal{F}} \ln \frac{f(y^n)}{p^*(y^n)} \right) p^*(y^n)\, d\nu(y^n) \ . \qquad (2)$$

The above expression is at the basis of the proof of the main result of this paper, see Theorem 3 below.

Even though by Shtarkov's theorem we may explicitly compute the minimax optimal predictor, its practical use is limited by the hardness of computing each conditional $p^*(y|y^t)$. The most common way to define more easily computable prediction strategies is to consider mixture strategies of the form

$$p(y^n) = \int_{\Theta} f_\theta(y^n) dw(\theta) \ ,$$

where Θ is a set of parameters by which the experts are parametrized: $\mathcal{F} = \{f_\theta : \theta \in \Theta\}$, and w is a probability measure over Θ. For an exhaustive survey of related results, we again refer to [10].

A simple example of a mixture strategy is when $\mathcal{F}$ is a finite class and w is the uniform distribution over $\mathcal{F}$. In this case, the conditionals of the mixture strategy take the simple form

$$p(y|y^{t-1}) = \frac{\sum_{f \in \mathcal{F}} f(y|y^{t-1}) f(y^{t-1})}{\sum_{g \in \mathcal{F}} g(y^{t-1})} \ . \qquad (3)$$

This is just the weighted average (WA) algorithm of De Santis *et al.* [14], see also [8, 9, 18, 20].

Besides being computationally easier to handle than p^*, mixture strategies are (in general) *universal*, that is, their conditionals can be computed without knowing the sequence length n in advance. On the other hand, there are simple finite classes $\mathcal{F}$ on which mixture strategies perform very poorly compared to the optimal predictor. We will discuss this further in Section 5.

We close this section by recalling the simple and elegant analysis of the regret of the WA strategy. This result will be used in Section 3.

PROPOSITION 2. *(De Santis et al., [14]) For the WA strategy p and for any finite class $\mathcal{F}$ of experts,*

$$R_n(p, \mathcal{F}) \leq \ln |\mathcal{F}| \ .$$

Proof. Let $W_1 = |\mathcal{F}|$, and $W_t = \sum_{f \in \mathcal{F}} f(y^{t-1})$, $t \geq 2$. Then, on the one hand,

$$\ln \frac{W_{n+1}}{W_1} = \ln \left(\sum_{f \in \mathcal{F}} f(y^n) \right) - \ln |\mathcal{F}| \geq \ln \max_{f \in \mathcal{F}} f(y^n) - \ln |\mathcal{F}| \ ,$$

and the other hand,

$$\begin{aligned} \ln \frac{W_{n+1}}{W_1} &= \sum_{t=1}^{n} \ln \frac{W_{t+1}}{W_t} = \sum_{t=1}^{n} \ln \frac{\sum_{f \in \mathcal{F}} f(y^t)}{\sum_{f \in \mathcal{F}} f(y^{t-1})} \\ &= \sum_{t=1}^{n} \ln p(y_t | y^{t-1}) = \ln p(y^n) \ . \end{aligned}$$

Thus, we obtain

$$R_n(p, \mathcal{F}) \leq \sup_{y^n} \ln \frac{\max_{f \in \mathcal{F}} f(y^n)}{p(y^n)} \leq \ln |\mathcal{F}| \ .$$

□

3. Main result

We start with some definitions. The diameter of a totally bounded metric space (S, ρ) is

$$\sup_{x,y \in S} \rho(x, y) \ .$$

Let $T \subseteq S$. Then for any $\varepsilon > 0$, the ε-covering number $N_\rho(T, \varepsilon)$ of T is the cardinality of the smallest subset $T' \subseteq S$ such that

$$(\forall x \in T)(\exists x' \in T') \quad \rho(x, x') \leq \varepsilon \ .$$

To any class $\mathcal{F}$ of experts, we associate the metric d defined by

$$d(f,g) = \sqrt{\sum_{t=1}^{n} \sup_{y^t} \left(\ln f(y_t|y^{t-1}) - \ln g(y_t|y^{t-1})\right)^2} \ . \tag{4}$$

We use $N(\mathcal{F}, \varepsilon)$ to denote the ε-covering number of $\mathcal{F}$ under the metric d.

THEOREM 3. *For any class $\mathcal{F}$ of experts,*

$$R_n(\mathcal{F}) \leq \inf_{\varepsilon > 0} \left(\ln N(\mathcal{F}, \varepsilon) + 24 \int_0^\varepsilon \sqrt{\ln N(\mathcal{F}, \delta)}\, d\delta \right) \ .$$

Remark. The main Theorem in Opper and Haussler [11] has a similar form. In particular, they showed that if every expert f in the class $\mathcal{F}$ has the special form $f(y_t|y^{t-1}) = f'(y_t)$ (i.e., every expert f corresponds to the product of n identical distributions f' on $\mathcal{Y}$ — we call such experts *static*), then for some constant K,

$$R_n(\mathcal{F}) \leq \inf_{\varepsilon > 0} \left(\ln N_\rho(\mathcal{F}, \varepsilon) + K \int_0^\varepsilon \sqrt{\ln N_\rho(\mathcal{F}, \delta)}\, d\delta + n\varepsilon^2 \right) \tag{5}$$

where ρ is a metric on $\mathcal{F}$. With Theorem 3 above, we show that in the upper bound (5) the term $n\varepsilon^2$ is unnecessary. We also show that the metric ρ can be replaced by another metric d, defined in (4), which satisfies $N_d(\mathcal{F}, \delta) \leq N_\rho(\mathcal{F}, \delta)$ for any $\delta > 0$ and any class $\mathcal{F}$ of static experts. (Note that the relative weakness of the bound (5) did not prevent Opper and Haussler from obtaining upper bounds of the right order in their applications.) Most importantly, however, we extend the result of [11] to classes of arbitrary, not just product experts. Our proof of Theorem 3 shows some similarities with that of Opper and Haussler, in that we also use techniques from empirical process theory. Nevertheless, their proof does not seem to be extendable to handle the general case treated here.

Remark. Theorem 3 requires that the expert class be finitely coverable in the metric d. This in turn requires that all conditional densities be bounded away from zero. It is unclear whether such a condition is necessary in general. In fact, for certain special parametric cases like the family of all i.i.d. measures over a finite alphabet, such boundedness conditions are not needed (see the remark after Corollary 6 for further discussion).

In order to prove the theorem, first we recall some well-known notions from empirical process theory. A family

$$\{T_f \,:\, f \in \mathcal{F}\}$$

of zero mean random variables (indexed by a metric space $(\mathcal{F}, \rho)$) is called *subgaussian* in the metric ρ whenever

$$\mathbf{E}\left[e^{\lambda(T_f - T_g)}\right] \leq e^{\lambda^2 \rho(f,g)^2/2}$$

holds for any $f, g \in \mathcal{F}$ and $\lambda > 0$. We also assume that the family is *sample continuous*, that is, for any sequence $f_1, f_2, \ldots \in \mathcal{F}$ converging to some $f \in \mathcal{F}$, we have $T_{f_n} - T_f \to 0$ almost surely.

The main tool used in our proofs is the following result of empirical process theory, stating that the expected supremum over a subgaussian family is governed by geometrical properties of the family in an appropriate metric. The result is a simple version of Dudley's classical metric entropy bound (see, e.g., [16]), whose proof is given in the Appendix for completeness. Note that we ignore measurability issues here, by implicitly assuming the measurability for all suprema.

PROPOSITION 4. *If $\{T_f \,:\, f \in \mathcal{F}\}$ is subgaussian and sample continuous in the metric ρ, then*

$$\mathbf{E}\left[\sup_{\mathcal{F}} T_f\right] \leq 12 \int_0^{D/2} \sqrt{\ln N_\rho(\mathcal{F}, \varepsilon)}\, d\varepsilon$$

where D is the diameter of $\mathcal{F}$.

We use Proposition 4 to obtain a first (weak) bound on $R_n(\mathcal{F})$ based on a direct analysis of Sharkov's strategy p^*. This will be later used as a tool to prove the stronger bound of Theorem 3, which is based on the analysis of a variant of p^*.

LEMMA 5. *For any class $\mathcal{F}$ of experts,*

$$R_n(\mathcal{F}) \leq 24 \int_0^{D/2} \sqrt{\ln N(\mathcal{F}, \varepsilon)}\, d\varepsilon$$

where D is the diameter of $\mathcal{F}$.

Proof. Using (2), we write

$$\begin{aligned} R_n(\mathcal{F}) &= \int \left(\sup_{\mathcal{F}} \ln \frac{f(y^n)}{p^*(y^n)} \right) p^*(y^n)\, d\nu(y^n) \\ &= \mathbf{E}\left[\sup_{\mathcal{F}} \ln \frac{f(Y^n)}{p^*(Y^n)} \right] \\ &\qquad \text{(where } Y^n = (Y_1, \ldots, Y_n) \text{ is a vector of random} \\ &\qquad\ \text{variables distributed according to } p^*) \\ &= \mathbf{E}\left[\sup_{\mathcal{F}} \sum_{t=1}^{n} \ln \frac{f(Y_t|Y^{t-1})}{p^*(Y_t|Y^{t-1})} \right] \\ &\le \mathbf{E}\left[\sup_{\mathcal{F}} \sum_{t=1}^{n} \left(\ln \frac{f(Y_t|Y^{t-1})}{p^*(Y_t|Y^{t-1})} - \mathbf{E}\left[\ln \frac{f(Y_t|Y^{t-1})}{p^*(Y_t|Y^{t-1})} \Big| Y^{t-1} \right] \right) \right] \end{aligned}$$

where the last step follows from the nonnegativity of the Kullback-Leibler divergence of the conditional densities (see, e.g., [5]):

$$\mathbf{E}\left[\ln \frac{p^*(Y_t|Y^{t-1} = y^{t-1})}{f(Y_t|Y^{t-1} = y^{t-1})} \right] \ge 0 \ . \tag{6}$$

Now, for each $f \in \mathcal{F}$ let

$$T_f(y^n) = \frac{1}{2} \sum_{t=1}^{n} \left(\ln \frac{f(y_t|y^{t-1})}{p^*(y_t|y^{t-1})} - \mathbf{E}\left[\ln \frac{f(Y_t|Y^{t-1})}{p^*(Y_t|Y^{t-1})} \Big| Y^{t-1} \right] \right)$$

so that we have $R_n(\mathcal{F}) \le 2\mathbf{E}\left[\sup_{\mathcal{F}} T_f\right]$.

To apply Proposition 4, we need to show that $\{T_f \,:\, f \in \mathcal{F}\}$ is indeed a subgaussian family under the metric d. (Sample continuity of the process is obvious.) To this end, note that for any $f, g \in \mathcal{F}$,

$$T_f(y^n) - T_g(y^n) = \sum_{t=1}^{n} Z_t(y^t) \ ,$$

where

$$Z_t(y^t) = \frac{1}{2} \left(\ln \frac{f(y_t \mid y^{t-1})}{g(y_t \mid y^{t-1})} - \mathbf{E}\left[\ln \frac{f(Y_t \mid Y^{t-1} = y^{t-1})}{g(Y_t \mid Y^{t-1} = y^{t-1})} \right] \right) \ .$$

Now it is easy to see that $T_f - T_g = T_f(y^n) - T_g(y^n)$ is a sum of bounded martingale differences, that is, each term Z_t has zero conditional mean and range bounded by $2d_t(f,g)$. Then the Hoeffding-Azuma inequality [1] implies that, for all $\lambda > 0$,

$$\mathbf{E}\left[e^{\lambda(T_f - T_g)} \right] \le \exp\left(\frac{\lambda^2}{2} d(f,g)^2 \right) \ .$$

Thus, the family $\{T_f : f \in \mathcal{F}\}$ is indeed subgaussian. Hence, recalling that $R_n(\mathcal{F}) \le 2\mathbf{E}\left[\sup_{\mathcal{F}} T_f\right]$ and applying Proposition 4 we obtain the statement of the lemma.

□

Lemma 5 provides a sharp bound on the regret of p^* if the diameter D of $\mathcal{F}$ is very small. However, inequality (6) becomes very loose for experts f far away from p^*. To avoid such situations, we prove our general bound by analyzing the following prediction strategy (different from p^*): $\mathcal{F}$ is partitioned into small subclasses and the minimax predictor is calculated for each subclass (in which Lemma 5 may be applied). Finally, these predictors are combined using the WA algorithm.

Proof. (Theorem 3) Fix an arbitrary $\varepsilon > 0$ and let $\mathcal{G}$ be an ε-covering of $\mathcal{F}$ of minimum size $N = N(\mathcal{F}, \varepsilon)$. Let $\mathcal{F}_1, \ldots, \mathcal{F}_N$ be the cells of the Voronoi tessellation of $\mathcal{F}$, under the metric d, having the elements of $\mathcal{G}$ as cell centers (remember that $\mathcal{F}$ and $\mathcal{G}$ live in the same metric space, but $\mathcal{G}$ does not have to be a subset of $\mathcal{F}$). Then $\mathcal{F}_1, \ldots, \mathcal{F}_N$ is a partition of $\mathcal{F}$. For each $i = 1, \ldots, N$, let $g^{(i)}$ be Shtarkov's optimal predictor for $\mathcal{F}_i$,

$$g^{(i)}(y^n) = \frac{\sup_{\mathcal{F}_i} f(y^n)}{\int \sup_{\mathcal{F}_i} f(x^n)\, d\nu(x^n)} \ .$$

Now let the predictor p_ε be the WA algorithm defined in (3) run over the set of "experts" $g^{(1)}, \ldots, g^{(N)}$. Clearly, $R_n(\mathcal{F}) \le \inf_{\varepsilon > 0} R_n(p_\varepsilon, \mathcal{F})$. So all we have to do is to bound the regret of p_ε.

To this end, fix any $y^n \in \mathcal{Y}^n$ and let $k = k(y^n)$ be such that

$$\ln \sup_{\mathcal{F}} f(y^n) = \ln \sup_{\mathcal{F}_k} f(y^n) \ .$$

Then,

$$\ln \frac{\sup_{\mathcal{F}} f(y^n)}{p_\varepsilon(y^n)} = \ln \frac{g^{(k)}(y^n)}{p_\varepsilon(y^n)} + \ln \frac{\sup_{\mathcal{F}_k} f(y^n)}{g^{(k)}(y^n)} \ . \tag{7}$$

As $k = k(y^n)$ ranges in $\{1, \ldots, N\}$, by Proposition 2 we get

$$\sup_{y^n} \ln \frac{g^{(k)}(y^n)}{p_\varepsilon(y^n)} \le \ln N \ . \tag{8}$$

Furthermore

$$\sup_{y^n} \ln \frac{\sup_{\mathcal{F}_k} f(y^n)}{g^{(k)}(y^n)} \le \max_{1 \le i \le N} \sup_{y^n} \ln \frac{\sup_{\mathcal{F}_i} f(y^n)}{g^{(i)}(y^n)} = \max_{1 \le i \le N} R_n(\mathcal{F}_i) \ . \tag{9}$$

Hence, combining (7), (8), and (9) we get

$$R_n(p_\varepsilon, \mathcal{F}) \le \ln N + \max_{1 \le i \le N} R_n(\mathcal{F}_i) \ . \tag{10}$$

Now note that the diameter of each element of the partition $\mathcal{F}_1, \ldots, \mathcal{F}_N$ is at most 2ε. Hence, applying Lemma 5 to each $\mathcal{F}_i$ in (10) we find that

$$\begin{aligned} R_n(p_\varepsilon, \mathcal{F}) &\leq \ln N + \max_{1\leq i\leq N} 24 \int_0^\varepsilon \sqrt{\ln N(\mathcal{F}_i, \delta)}\, d\delta \\ &\leq \ln N(\mathcal{F}, \varepsilon) + 24 \int_0^\varepsilon \sqrt{\ln N(\mathcal{F}, \delta)}\, d\delta \end{aligned}$$

concluding the proof.

□

Remark. Similarly to an analogous derivation in [11], Theorem 3 could be also proven by direct manipulation of the minimax regret in the form

$$\ln \int \sup_{\mathcal{F}} f(y^n)\, d\nu(y^n) \ .$$

This is done by partitioning $\mathcal{F}$ as in the proof of Theorem 3 and then replacing the derivation of the bound (10) with the following:

$$\begin{aligned} R_n(\mathcal{F}) &= \ln \int \sup_{\mathcal{F}} f(y^n)\, d\nu(y^n) \\ &\leq \ln \int \left(\sum_{i=1}^N \sup_{\mathcal{F}_i} f(y^n) \right) d\nu(y^n) \\ &\leq \ln N + \max_{1\leq i\leq N} \ln \int \sup_{\mathcal{F}_i} f(y^n)\, d\nu(y^n) \\ &= \ln N + \max_{1\leq i\leq N} \ln R_n(\mathcal{F}_i) \ . \end{aligned}$$

Though a bit more concise, this proof ignores the algorithmical meaning of the right-hand side of (10).

Remark. It is interesting to note that, while strategies like p_ε can have a regret close to the optimal value $R_n(\mathcal{F})$, p^* is the *unique* strategy with regret equal to $R_n(\mathcal{F})$, and this is precisely due to the fact that p^* is an equalizer. To show this, pick any $\mathcal{F}$ and assume there exists p' such that $p' \neq p^*$ and yet $R_n(p', \mathcal{F}) = R_n(p^*, \mathcal{F}) = R_n(\mathcal{F})$. As p is normalized, $p' \neq p^*$ implies that $p(y^n) < p^*(y^n)$ for some y^n. Hence, $\sup_{\mathcal{F}} f(y^n)/p'(y^n) > \sup_{\mathcal{F}} f(y^n)/p^*(y^n)$ for this y^n. But (2) implies that $\sup_{\mathcal{F}} f(y^n)/p^*(y^n) = R_n(\mathcal{F})$ for any y^n. Hence $\sup_{\mathcal{F}} f(y^n)/p'(y^n) > R_n(\mathcal{F})$ contradicting the assumption $R_n(p', \mathcal{F}) = R_n(\mathcal{F})$.

4. Applications

In this Section we illustrate some natural applications of our upper bounds that, to the best of our knowledge, could not be obtained with previous techniques.

4.1. Parametric classes

As a first example, consider classes $\mathcal{F}$ such that there exist positive constants k and c such that for all $\varepsilon > 0$,

$$\ln N(\mathcal{F}, \varepsilon) \leq k \ln \frac{c\sqrt{n}}{\varepsilon} . \tag{11}$$

This is the case for most "parametric" classes, that is, classes which can be parametrized by a bounded subset of $\mathcal{R}^k$ in some "smooth" way. Asymptotic expressions for $R_n(\mathcal{F})$ were established by Rissanen [13] for such classes under certain general conditions. In particular, Rissanen showed under his conditions that $R_n(\mathcal{F}) \approx (k/2)\ln n$. However, these conditions are difficult to check in some situations, and they are asymptotic in nature. Theorem 3 allows us to derive a simple nonasymptotic bound under the sole metric condition (11).

COROLLARY 6. *Assume that the covering numbers of the class $\mathcal{F}$ satisfy (11). Then for each n so large that*

$$c\sqrt{n} \geq 48\sqrt{2}\sqrt{\ln(c\sqrt{n})/k} ,$$

we have

$$R_n(\mathcal{F}) \leq \frac{k}{2}\ln n + \frac{k}{2}\ln\frac{\ln(c\sqrt{n})}{k} + k\ln c + 6k .$$

Proof. Substituting (11) in the upper bound of Theorem 3, the first term of the expression is bounded by $\frac{k}{2}\ln n + k\ln c - k\ln\varepsilon$. Then the second term may be bounded as follows:

$$\begin{aligned}
24\int_0^\varepsilon \sqrt{\ln N(\mathcal{F},\delta)}\,d\delta \;&\leq\; 48c\sqrt{kn}\int_{a_n}^\infty x^2 e^{-x^2}dx \\
&\qquad \text{(by substituting } x = \sqrt{\ln(c\sqrt{n}/\delta)} \\
&\qquad\quad \text{and writing } a_n = \sqrt{\ln(c\sqrt{n}/\varepsilon)}) \\
&=\; 48c\sqrt{kn}\left[\frac{a_n}{2c\sqrt{n}/\varepsilon} + \frac{1}{2}\int_{a_n}^\infty e^{-x^2}dx\right] \\
&\qquad \text{(by integrating by parts)}
\end{aligned}$$

$$
\begin{aligned}
&\leq\ 48c\sqrt{kn}\left[\frac{a_n}{2c\sqrt{n}/\varepsilon}+\frac{1}{2a_nc\sqrt{n}/\varepsilon}\right]\\
&\qquad\text{(by using the gaussian tail estimate}\\
&\qquad\ \textstyle\int_t^\infty e^{-x^2}dx \leq e^{-t^2}/(2t))\\
&\leq\ 48\sqrt{k}a_n\varepsilon \quad \text{(whenever } e\varepsilon \leq c\sqrt{n}\text{)}\\
&\leq\ 48\sqrt{2}\varepsilon\sqrt{k\ln(c\sqrt{n})} \quad \text{(whenever } \varepsilon \geq 1/(c\sqrt{n})\text{)}
\end{aligned}
$$

The obtained upper bound is minimized for

$$
\varepsilon = \frac{1}{48\sqrt{2}}\sqrt{\frac{k}{\ln(c\sqrt{n})}}\ ,
$$

which yields the desired result.

□

Remark. The main term $(k/2)\ln n$ is known to be the best possible for most k-dimensional parametric families such as the family of all i.i.d. measures over a finite alphabet $\mathcal{Y}$ of $k+1$ elements [15], or, if $k = 2^m$, for the family of all m-th order stationary Markov measures over a binary alphabet [19]. The lower-order term in the Corollary above is however not the best possible in some cases, when much sharper estimates are available (see, e.g., Barron and Xie [2], Freund [7]). In fact, typical specific upper bounds, have the form $(k/2)\ln n + O(1)$ (see [2, 7, 13]). We do not know if, in the generality treated here, the second $O(k\ln\ln n)$ term is necessary. Also, Corollary 6 may only be used if all conditional densities are bounded away from zero. For example, in the case of the examples mentioned above, we need to restrict the range of the parameters so that all probabilities become bounded away from zero. Such condition is not necessary in some of the parametric examples mentioned above. For example, Barron and Xie [2] and Freund [7] do not need any restriction when $\mathcal{F}$ is the class of all i.i.d. measures over a finite alphabet $\mathcal{Y}$ of $k+1$ elements. Note, however, that to obtain his general parametric bound, Rissanen assumes that the Fisher information is uniformly bounded away from zero, which, in specific cases, leads to a similar restriction of the class as the one we implicitly need by assuming (11). On the other hand, the general condition under which the Corollary holds makes it useful in situations where all previously known techniques fail. This is illustrated in the next simple example.

Example: Fading-memory predictors. Let $\mathcal{Y} = \{0, 1\}$, and consider the one-parameter class $\mathcal{F}$ of distributions on $\{0, 1\}^n$ containing all experts

$f^{(a)}$ with $a \in [0,1]$, where each $f^{(a)}$ is defined by its conditionals as: $f_1^{(a)}(1) = 1/2$, $f_2^{(a)}(1|y_1) = y_1$, and

$$f_t^{(a)}(1|y^{t-1}) = \frac{1}{t-1} \sum_{i=1}^{t-1} y_i \left(1 + \frac{a(2i-t)}{t-2}\right),$$

for all $y^{t-1} \in \{0,1\}^{t-1}$ and for all $t > 2$. In other words, each expert predicts according to a weighted average of the past outcomes with linearly decaying weights as we go back in the past. The parameter a determines the slope of the decay. Unfortunately, Theorem 3 cannot handle this class because the values of $f_t^{(a)}(1|y^{t-1})$ and $f_t^{(a)}(0|y^{t-1})$ may be arbitrarily close to zero, and therefore the covering numbers of this class with respect to the metric d are infinite. To avoid this difficulty, we slightly modify the experts by considering the class $\mathcal{G}$ of all experts g of the form

$$g_t^{(a)}(1|y^{t-1}) = \tau(f_t^{(a)}(1|y^{t-1})) ,$$

where

$$\tau(x) = \begin{cases} x & \text{if } x \in [\delta, 1-\delta] \\ \delta & \text{if } x < \delta \\ 1-\delta & \text{if } x > 1-\delta \end{cases}$$

for some fixed $0 < \delta < 1/2$. Now clearly, for all $t \geq 1$, and $a, b \in [0,1]$,

$$\begin{aligned}
d_t(g^{(a)}, g^{(b)}) &= \max_{y^{t-1}\in\{0,1\}^{t-1}} \left| \ln g_t^{(a)}(1|y^{t-1}) - \ln g_t^{(b)}(1|y^{t-1}) \right| \\
&\leq \frac{1}{\delta} \max_{y^{t-1}\in\{0,1\}^{t-1}} \left| g_t^{(a)}(1|y^{t-1}) - g_t^{(b)}(1|y^{t-1}) \right| \\
&\leq \frac{1}{\delta} \max_{y^{t-1}\in\{0,1\}^{t-1}} \left| f_t^{(a)}(1|y^{t-1}) - f_t^{(b)}(1|y^{t-1}) \right| \\
&= \frac{1}{\delta} \max_{y^{t-1}\in\{0,1\}^{t-1}} \left| \frac{1}{t-1} \sum_{i=1}^{t-1} y_i \frac{(a-b)(2i-t)}{t-2} \right| \\
&= \frac{1}{\delta} \max_{y^{t-1}\in\{0,1\}^{t-1}} \left| (a-b) \frac{1}{t-1} \sum_{i=1}^{t-1} y_i \frac{(2i-t)}{t-2} \right| \\
&\leq \frac{1}{\delta} \max_{y^{t-1}\in\{0,1\}^{t-1}} |a-b| \left| \frac{1}{t-1} \sum_{i=1}^{t-1} y_i \right| \\
&\leq \frac{|a-b|}{\delta} .
\end{aligned}$$

Therefore, we immediately see that for all $\varepsilon > 0$,

$$\ln N(\mathcal{G}, \varepsilon) \leq \ln \frac{\sqrt{n}}{\varepsilon\delta} ,$$

so Corollary 6 yields

$$R_n(\mathcal{G}) \leq \frac{1}{2}\ln n + \frac{1}{2}\ln\ln\frac{\sqrt{n}}{\delta} + \ln\frac{1}{\delta} + 6 \ .$$

Note that this class cannot be handled by Rissanen's asymptotic expansion, which requires that the MLE in the class satisfy a uniform central limit theorem condition. In fact, the experts in $\mathcal{G}$ are nonstationary, and reach far back in the past, so proving a central limit theorem for the MLE of a would be extremely difficult (let alone a uniform one!), even if we had known what the MLE was.

4.2. Nonparametric classes

Next, we illustrate on two examples how Theorem 3 can be applied for very large, nonparametric classes. The first example shows that nontrivial bounds may be obtained even for utterly huge classes of predictors.

Example: Lipschitz-Markov predictors. Assume, for simplicity, that the alphabet is $\mathcal{Y} = [0,1]$. Let $\mathcal{C}$ be a class of densities (with respect to the Lebesgue measure) on $[0,1]$ such that its covering number $N_\rho(\mathcal{C},\varepsilon)$ with respect to the metric

$$\rho(p,p') = \sup_{x\in[0,1]} |\ln p(x) - \ln p'(x)|$$

satisfies $\ln N_\rho(\mathcal{C},\varepsilon) \leq c\varepsilon^{-a}$ for some $a, c > 0$. (An example of a nonparametric class of densities satisfying this condition is the class of all Lipschitz densities which are uniformly bounded away from zero, a class also considered in [11].)

Now consider the class $\mathcal{F}$ of *all* k-th order Markov measures on $[0,1]^n$ such that for all $t \leq n$ and

$$y_{t-k}^{t-1} = (y_{t-k},\ldots,y_{t-1}) \in [0,1]^k \ ,$$

the conditional densities satisfy $f_t(\cdot|y_{t-k}^{t-1}) \in \mathcal{C}$, and moreover, for all $t \leq n$ and $y_{t-k}^{t-1}, z_{t-k}^{t-1} \in [0,1]^k$,

$$\sup_{x\in[0,1]} \left|\ln f_t(x|y_{t-k}^{t-1}) - \ln f_t(x|z_{t-k}^{t-1})\right| \leq \max_{t-k\leq s\leq t-1} |y_s - z_s| \ .$$

The last condition requires that a small change in the past does not cause a drastic change in the log of the conditional density. Notice that all these are quite natural smoothness assumptions, and the resulting class of experts is very large.

To use Theorem 3 it suffices to observe that $N_{\mathcal{F}}(\varepsilon)$ may easily be bounded by

$$N_{\mathcal{F}}(\varepsilon) \leq [N_{\mathcal{C}}(\varepsilon/2)]^{(c_1\sqrt{n}/\varepsilon)^k} ,$$

where c_1 is a positive constant. Now it is a matter of routine calculation to obtain the bound

$$R_n(\mathcal{F}) = O\left(n^{\frac{a+k}{2+a+k}}\right) .$$

Example: Monotone predictors. Let $\mathcal{Y} = \{0,1\}$ be a binary alphabet, and consider the class $\mathcal{F}$ of all experts $f = \prod_t f_t$ such that $f(1|y^{t-1}) = f_t(1) \in [\delta, 1-\delta]$, where $\delta \in (0,1/2)$ is some fixed constant, and for each $t = 2,3,\ldots,n$, $f_t(1) \geq f_{t-1}(1)$. In other words, $\mathcal{F}$ contains all *static* experts (i.e., experts which predict independently of the past data) which assign a probability to the outcome "1" in a monotonically increasing manner. This class is again "nonparametric", but here the richness of the class is not due to the richness of the conditional densities, but rather to the nonstationarity of the experts in $\mathcal{F}$. To estimate the covering number of $\mathcal{F}$, consider the finite subclass $\mathcal{G}$ of $\mathcal{F}$ containing only those monotone experts $g = \prod_t g_t$ which take values of the form $g_t(1) = \delta + (i/k)(1-2\delta)$, $i = 0,\ldots,k$, where k is a positive integer to be specified later. It is easy to see that $|\mathcal{G}| = \binom{n+k}{k} \leq (2n)^k$ if $k \leq n$, and $|\mathcal{G}| \leq 2^k$ otherwise. On the other hand, for any $f \in \mathcal{F}$, if g is the expert in $\mathcal{G}$ which is closest to f, then for each $t \leq n$,

$$\begin{aligned}
\max_{y\in\{0,1\}} |\ln f_t(y) - \ln g_t(y)| \;&\leq\; \frac{1}{\delta} \max_{y\in\{0,1\}} |f_t(y) - g_t(y)| \\
&=\; \frac{1}{\delta}|f_t(1) - g_t(1)| \\
&\leq\; \frac{1}{\delta k}.
\end{aligned}$$

Thus, $d(f,g) \leq \sqrt{n}/(\delta k)$, where the metric d is defined in (4). By taking $k = \sqrt{n}/(\delta\epsilon)$, it follows that the covering number of $\mathcal{F}$ is bounded as

$$N(\mathcal{F},\varepsilon) \leq \begin{cases} (2n)^{\sqrt{n}/(\delta\varepsilon)} & \text{if } \epsilon \geq \frac{1}{\delta\sqrt{n}} \\ 2^{\sqrt{n}/(\delta\varepsilon)} & \text{otherwise.} \end{cases}$$

Substituting this bound into Theorem 3, it is a matter of straightforward calculation to obtain

$$R_n(\mathcal{F}) = O\left(n^{1/3}\delta^{-2/3}\ln^{2/3} n\right) .$$

Note that the radius optimizing the bound of Theorem 3 is about $\epsilon \approx n^{1/6}\delta^{-1/3}\ln^{1/3} n$.

5. Suboptimality of mixture predictors

As we have pointed it out in the introduction, instead of the minimax predictor given by Proposition 1, often mixture predictors are used. In some cases, the worst-case regret of mixture predictors, in particular, the WA predictor (3), is very close to the optimal value $R_n(\mathcal{F})$, see [2, 7, 9]. The purpose of this section is to point out that this is not necessarily so. In fact, even for very simple classes of static experts, the ratio of the minimax regret of the WA algorithm and that of the optimal algorithm can be arbitrarily large. Note that this does not contradict Theorem 3, where the WA algorithm was run on a special set of predictors derived from $\mathcal{F}$, instead of being run directly on the expert class $\mathcal{F}$, as prescribed by (3).

THEOREM 7. *For every $n > 1$ there exists a class $\mathcal{F}_n$ of two static experts such that, if p denotes the predictor defined in (3), then*

$$\frac{R_n(p, \mathcal{F}_n)}{R_n(\mathcal{F}_n)} \geq c\sqrt{n} \ ,$$

where c is a universal constant.

Proof. Let $\mathcal{F}_n$ contain the two experts f, g defined over the binary alphabet $\mathcal{Y} = \{0,1\}$ by

$$f(1|y^{t-1}) = \frac{1}{2} \quad \text{and} \quad g(1|y^{t-1}) = \frac{1}{2} + \frac{1}{2n}$$

for all $t \leq n$ and $y^{t-1} \in \{0,1\}^{t-1}$. We may easily estimate the minimax regret $R_n(\mathcal{F}_n)$ using Lemma 5. The diameter of $\mathcal{F}_n$ is easily seen to be

$$D = d(f,g) = \sqrt{n} \ln\left(1 + \frac{1}{n}\right) \leq \frac{1}{\sqrt{n}} \ .$$

Also, since $N(\mathcal{F}_n, \epsilon) \leq 2$ for all $\epsilon > 0$, Lemma 5 provides the upper bound

$$R_n(\mathcal{F}_n) \leq \frac{12\sqrt{\ln 2}}{\sqrt{n}} \ . \tag{12}$$

On the other hand, the definition of the WA algorithm in (3) shows that

$$p(y^n) = \frac{f(y^n) + g(y^n)}{2} \ .$$

The relative loss of p is

$$\begin{aligned} R_n(p, \mathcal{F}_n) &= \ln \max_{y^n} \frac{\max(f(y^n), g(y^n))}{p(y^n)} \\ &= \ln \max_{y^n} \frac{2\max(f(y^n), g(y^n))}{f(y^n) + g(y^n)} \end{aligned}$$

$$
\begin{aligned}
&\geq \ln \max_{y^n} \frac{2f(y^n)}{f(y^n)+g(y^n)} \\
&= \ln \max_{0\leq k\leq n} 2\frac{2^{-n}}{2^{-n}+\left(\frac{1}{2}-\frac{1}{2n}\right)^k\left(\frac{1}{2}+\frac{1}{2n}\right)^{n-k}} \\
&= \ln \max_{0\leq k\leq n} \frac{2}{1+\left(1-\frac{1}{n}\right)^k\left(1+\frac{1}{n}\right)^{n-k}} \\
&= \ln \frac{2}{1+\left(1-\frac{1}{n}\right)^n} \\
&\geq \ln \frac{2}{1+\frac{1}{e}} \ .
\end{aligned}
$$

Comparing this lower bound with (12), we obtain the statement of the theorem with $c = \ln\left(\frac{2}{1+\frac{1}{e}}\right) \Big/ 12\sqrt{\ln 2} \approx 0.038$.

□

Appendix

To prove Proposition 4, we use the following simple lemma, whose elegant proof was shown to one of us by Pascal Massart.

LEMMA 8. *Let $\sigma > 0$, and let $X_1,\ldots,X_N$ be real-valued random variables such that for all $\lambda > 0$ and $1 \leq i \leq N$, $\mathbf{E}\left[e^{\lambda X_i}\right] \leq e^{\lambda^2\sigma^2/2}$. Then*

$$
\mathbf{E}\left[\max_{i\leq N} X_i\right] \leq \sigma\sqrt{2\ln N} \ .
$$

Proof. By Jensen's inequality, for all $\lambda > 0$,

$$
\begin{aligned}
e^{\lambda \mathbf{E}\left[\max_{i\leq N} X_i\right]} &= \mathbf{E}\left[e^{\lambda \max_{i\leq N} X_i}\right] = \mathbf{E}\left[\max_{i\leq N} e^{\lambda X_i}\right] \\
&\leq \sum_{i=1}^{N} \mathbf{E}\left[e^{\lambda X_i}\right] \leq N e^{\lambda^2\sigma^2/2} \ .
\end{aligned}
$$

Thus,

$$
\mathbf{E}\left[\max_{i\leq N} X_i\right] \leq \frac{\ln N}{\lambda} + \frac{\lambda\sigma^2}{2} \ ,
$$

and taking $\lambda = \sqrt{2\ln N/\sigma^2}$ yields the result.

□

Proof. (Proposition 4) For each $k = 0, 1, 2, \ldots$, let $\mathcal{F}^{(k)}$ be a minimal cover of $\mathcal{F}$ of radius $D2^{-k}$. Note that $|\mathcal{F}^{(k)}| = N(\mathcal{F}, D2^{-k})$. Denote the unique element of $\mathcal{F}^{(0)}$ by f_0.

Let Ω be the common domain where the r.v.'s T_f, $f \in \mathcal{F}$, are defined. Pick $\omega \in \Omega$ and let $f^* \in \mathcal{F}$ be such that $\sup_{f \in \mathcal{F}} T_f(\omega) = T_{f^*}(\omega)$. (Here we implicitly assume that such an element exists. The modification of the proof for the general case is straightforward.)

For each $k \geq 0$, let f_k^* denote an element of $\mathcal{F}^{(k)}$ whose distance to f^* is minimal. Clearly, $\rho(f^*, f_k^*) \leq D2^{-k}$, and therefore, by the triangle inequality, for each $k \geq 1$,

$$\rho(f_{k-1}^*, f_k^*) \leq \rho(f^*, f_k^*) + \rho(f^*, f_{k-1}^*) \leq 3D2^{-k} \ . \tag{13}$$

Clearly, $\lim_{k \to \infty} f_k^* = f^*$, and so by the sample continuity of the process,

$$\sup_f T_f(\omega) = T_{f^*}(\omega) = T_{f_0}(\omega) + \sum_{k=1}^{\infty} \left(T_{f_k^*}(\omega) - T_{f_{k-1}^*}(\omega) \right) \ ,$$

and therefore

$$\mathbf{E}\left[\sup_f T_f\right] \leq \sum_{k=1}^{\infty} \mathbf{E}\left[\max_{f,g} (T_f - T_g)\right]$$

where the max is taken over all pairs $(f, g) \in \mathcal{F}^{(k)} \times \mathcal{F}^{(k-1)}$ such that $\rho(f, g) \leq 3D2^{-k}$.

Since these pairs are at most $N(\mathcal{F}, D2^{-k})^2$, and recalling that $\{T_f : f \in \mathcal{F}\}$ is subgaussian in the metric ρ, we can apply Lemma 8 using (13). Thus, for each $k \geq 1$,

$$\mathbf{E}\left[\max_{f,g} (T_f - T_g)\right] \leq 3D2^{-k} \sqrt{2 \ln N(\mathcal{F}, D2^{-k})^2} \ .$$

Summing over k, we obtain

$$\begin{aligned}
\mathbf{E}\left[\sup_f T_f\right] \ &\leq \ \sum_{k=1}^{\infty} 3D2^{-k} \sqrt{2 \ln N(\mathcal{F}, D2^{-k})^2} \\
&= \ 12 \sum_{k=1}^{\infty} D2^{-(k+1)} \sqrt{\ln N(\mathcal{F}, D2^{-k})} \\
&\leq \ 12 \int_0^{D/2} \sqrt{\ln N_\rho(\mathcal{F}, \varepsilon)}\, d\varepsilon \ ,
\end{aligned}$$

as desired.

□

Acknowledgements

Thanks to Manfred Opper for fruitful comments.

References

1. Azuma, K.: 1967, 'Weighted Sums of Certain Dependent Random Variables'. *Tohoku Mathematical Journal* **68**, 357–367.
2. Barron, A. and Q. Xie: 1996, 'Asymptotic Minimax Regret for Data Compression, Gambling, and Prediction'. in: *Proceedings of the Ninth Annual Conference on Computational Learning Theory*
3. Cover, T.: 1991, 'Universal Portfolios'. *Mathematical Finance* **1**, 1–29.
4. Cover, T. and E. Ordentlich: 1996, 'Universal Portfolios with Side Information'. *IEEE Transactions on Information Theory* **42**(2), 348–363.
5. Cover, T. and J. Thomas: 1991, *Elements of Information Theory*. John Wiley and Sons.
6. Feder, M.: 1991, 'Gambling Using a Finite State Machine'. *IEEE Transactions on Information Theory* **37**, 1459–1465.
7. Freund, Y.: 1996, 'Predicting a Binary Sequence Almost as Well as the Optimal Biased Coin'. In: *Proceedings of the 9th Annual Conference on Computational Learning Theory*. pp. 89–98.
8. Haussler, D. and A. Barron: 1993, 'How Well Does the Bayes Method Work in On-line Predictions of $\{+1, -1\}$ Values?'. In: *Proceedings of 3rd NEC Symposium*. pp. 74–100.
9. Haussler, D., J. Kivinen, and M. Warmuth: 1998, 'Sequential Prediction of Individual Sequences Under General Loss Functions'. *IEEE Transactions on Information Theory* **44**, 1906–1925.
10. Merhav, N. and M. Feder: 1998, 'Universal Prediction'. *IEEE Transactions on Information Theory* **44**(6), 2124–2147.
11. Opper, M. and D. Haussler: 1997, 'Worst Case Prediction over Sequences under Log Loss'. In: *The Mathematics of Information Coding, Extraction, and Distribution*. Springer Verlag.
12. Rissanen, J.: 1976, 'Generalized Kraft's Inequality and Arithmetic Coding'. *IBM Journal of Research and Development* **20**, 198–203.
13. Rissanen, J.: 1996, 'Fischer Information and Stochastic Complexity'. *IEEE Transactions on Information Theory* **42**, 40–47.
14. Santis, A. D., G. Markowski, and M. Wegman: 1988, 'Learning Probabilistic Prediction Functions'. In: *Proceedings of the 1st Annual Workshop on Computational Learning Theory*. pp. 312–328.
15. Shtarkov, Y.: 1987, 'Universal Sequential Coding of Single Messages'. Translated from: *Problems in Information Transmission* **23**(3), 3–17.
16. Talagrand, M.: 1996, 'Majorizing Measures: the Generic Chaining'. *Annals of Probability* **24**, 1049–1103. (Special Invited Paper).
17. Vovk, V.: 1990, 'Aggregating Strategies'. In: *Proceedings of the 3rd Annual Workshop on Computational Learning Theory*. pp. 372–383.
18. Vovk, V.: 1998, 'A Game of Prediction with Expert Advice'. *Journal of Computer and System Sciences* **56**(2), 153–173.

19. Weinberger, M., N. Merhav, and M. Feder: 1994, 'Optimal Sequential Probability Assignment for Individual Sequences'. *IEEE Transactions on Information Theory* **40**, 384–396.
20. Yamanishi, K.: 1995, 'A Loss Bound Model for On-line Stochastic Algorithms'. *Information and Computation* **119**(1), 39–54.
21. Yamanishi, K.: 1998, 'A Decision-theoretic Extension of Stochastic Complexity and its Application to Learning'. *IEEE Transactions on Information Theory* **44**, 1424–1440.

Address for Offprints:
Nicolò Cesa-Bianchi
Department of Information Sciences,
University of Milan
Via Comelico 39,
20135 Milano, Italy