# WORST CASE PREDICTION OVER SEQUENCES UNDER LOG LOSS

MANFRED OPPER* AND DAVID HAUSSLER†

**Abstract.** We consider the game of sequentially assigning probabilities to future data based on past observations under logarithmic loss. We are not making probabilistic assumptions about the generation of the data, but consider a situation where a player tries to minimize his loss relative to the loss of the (with hindsight) best distribution from a target class for the worst sequence of data. We give bounds on the minimax regret in terms of the metric entropies of the target class with respect to suitable distances between distributions.

## 1. Introduction.

The assignment of probabilities to the possible outcomes of future data which is based on past observations has important applications to prediction, data compression and gambling. In a scenario where the data are assumed to occur at random with an unknown probability distribution, this problem can be treated as a well known statistical estimation problem. Optimal strategies can be found within a game theoretic approach, where a statistician (in the following called 'learner') tries to minimize a certain average loss in a game played against 'nature', which chooses unfavourable probabilities for the data out of a given family.

In some cases, however, the assumption of randomness and of a *true* distribution for the data may not be fulfilled, or, there may not be enough prior information to specify a reasonable target class of distributions to which the true one belongs. Recently, new approaches to the prediction on sequences of data that avoid the assumption of randomness have found a great deal of interest in computational learning theory (see e.g. [2, 3, 4]) and information theory [5, 6, 7, 14]. For a collection of recent work, see the webpage `http:// www-stat.wharton.upenn.edu/Seq96/` of a workshop on prediction over sequences, held at UC Santa Cruz in 1996.

Instead of considering average losses, the goal is to find strategies which achieve a small accumulated loss for arbitrary sequences of data in a game of sequential prediction. Assuming a target family of predictors, often called *experts*, which is hopefully well suited to the data, the learner tries to find a strategy, which for any sequence guarantees a total loss that is not much larger than that of the expert which is best with hindsight.

For the important case of logarithmic loss, and families of experts which assign probabilities independently of past data, we will give general bounds for the minimal relative loss or regret which can be achieved with such a strategy. Our work differs from the other recent work in this area in

*Institut für Theoretische Physik III, Universität Würzburg, D-8700 Würzburg, Germany.

†Department of Computer and Information Sciences, University of California, Santa Cruz, CA 95064.

that it applies to both finite dimensional and infinite dimensional families of (simple) experts.

**2. Notation.** The following notation and assumptions will be used throughout the paper.

Let $Y$ be a complete separable metric space. All probability distributions on $Y$ discussed in this paper are assumed to be defined on the $\sigma$-algebra of Borel sets of $Y$. Let $\Theta$ be a set, and for each $\theta \in \Theta$, let $P_\theta$ be a probability distribution on $Y$. We assume that for any $\theta \neq \theta^* \in \Theta$, the distributions associated with $\theta$ and $\theta^*$ are distinct in the sense that there is a Borel set $S \subset Y$ such that $P_\theta(S) \neq P_{\theta^*}(S)$. In addition, we assume there is a fixed $\sigma$-finite measure $\nu$ on $Y$ that dominates $P_\theta$ for all $\theta \in \Theta$ (i.e. for any Borel set $S \subseteq Y$, $\nu(S) = 0$ implies $P_\theta(S) = 0$). We will also make (implicitly) the assumption that any other distribution $Q$ on $Y$ mentioned in the results below is also dominated by $\nu$. Radon-Nikodym derivatives (densities) with respect to $\nu$ will be denoted by lower case symbols like $q = \frac{dQ}{d\nu}$ and $p_\theta = \frac{dP_\theta}{d\nu}$.

**3. A sequential prediction game.** Suppose that $n$ symbols, $y^n = y_1, \ldots, y_n$ are observed sequentially, i.e. one after the other. After each observation $y_{t-1}$, where $t = 1, \ldots, n$, a learner is asked how likely each value $y \in Y$ is to be the *next* observation. I.e., the learner's goal is to assign a probability distribution $q(y|y^{t-1})$ over the possible outcomes $y$ of the next observation, based on the previous values. When at the next time step $t$, the actual new observation $y_t$ is received, the learner suffers a loss, which, throughout this article, will be the *logarithmic loss* $-\log q(y_t|y^{t-1})$.

Logarithmic loss has an important meaning in data compression, where any assignment of probabilities to data values can be considered as an assignment of possible codelengths to the data using a uniquely decodable code [8]. The total log loss is (ignoring the problems of truncating continuous data values and the rounding of integers) proportional to the length of the compressed sequence of data. For an interpretation of logarithmic loss in terms of the wealth achieved in gambling, where probabilities stand for the relative amount of money bet on future data values, see [8, 9]

At the end of the game, the learner has suffered a total loss

$$L(q, y^n) = -\sum_{t=1}^{n} \log q(y_t|y^{t-1}).$$

All $n$ predictions $q(y_t|y^{t-1})$, $t = 1, \ldots, n$ can be composed into a single joint distribution

$$q(y^n) = \prod_{t=1}^{n} q(y_t|y^{t-1}).$$

On the other hand, any joint distribution $q(y^n)$ defines a sequence of predictive distributions from its conditionals $q(y_t|y^{t-1})$. Hence, the learners goal can be understood as an assignment of a distribution $q$ to the set of all possible sequences $y^n \in Y^n$ and the loss can be written as

$$L(q, y^n) = -\log q(y^n).$$

If the sequences were known to be randomly drawn from a distribution $p(y^n)$, then it is easy to see that in order to achieve a minimal *expected loss*, the learner should predict with the conditional distributions $p(y|y^{t-1})$ for $t = 1, \dots, n-1$. If the true distribution is not known, only the fact that it belongs to a family of distributions $p_\theta$, $\theta \in \Theta$, then a possible strategy would be to minimize the expected extra loss above the minimum for the worst $\theta$. This means that the learner, trying to be prepared for the worst *distribution* of sequences, should minimize the *risk*[1]

$$\sup_\theta \int d\nu(y^n) p_\theta(y^n) \left\{-\log q(y^n) + \log p_\theta(y^n)\right\}.$$

See [12] and papers cited there for a discussion this average loss framework and the results that can be obtained there.

We will now go beyond the average loss framework and analyze a strategy which aims at performing well on *individual sequences.* In this approach, no probabilistic assumptions about the generation of sequences is made. The target family of distributions $p_\theta$ can now be seen as a family of experts which is hoped to be well suited to the sequences. The goal of the learner is now to find a distribution $q$, which makes the loss on the sequence $y^n$ not much bigger than that of the best expert in the target class $p_\theta$, $\theta \in \Theta$. This best expert, which achieves a loss $-\sup_{\theta\in\Theta} \log p_\theta(y^n)$, depends on the *entire* sequence, and will not be known to the learner (before the last symbol $y_n$ is observed). In a worst case scenario, the learner has to choose a distribution $q$ which minimizes the *regret* that is the difference between his loss and the loss of the best expert,

$$R(q, y^n) = -\log q(y^n) + \sup_{\theta\in\Theta} \log p_\theta(y^n).$$

The minimal regret achievable by such a strategy is

$$R_n = \inf_q \sup_{y^n} R(q, y^n) = \inf_q \sup_{y^n} \left\{\log \frac{\sup_{\theta\in\Theta} p_\theta(y^n)}{q(y^n)}\right\}. \tag{1}$$

Bounds and asymptotic expressions for this minimax regret have been obtained for finite dimensional parametric families of distributions, such as probability mass functions over a finite alphabet or distributions which are smooth functions of the parameters [4, 5, 6, 7].

[1] Here and in what follows, $\nu(y^n)$ and $p_\theta(y^n)$ are used to denote the $n$-fold products of the distributions $\nu$ and $p_\theta$ respectively, evaluated at the point $y^n$.

In this paper, we give a general upper bound on $R_n$ for target families of product distributions

$$p_\theta(y^n) = \prod_{t=1}^{n} p_\theta(y_t),$$

which are uniformly bounded away from zero and infinity. The bounds can be applied to nonparametric families of distributions, where, to our knowledge, minimax results for arbitrary sequences have not been obtained.

Our calculation is based on the explicit solution to the minimax problem (1), which was given by Shtarkov [5]. He found that the distribution

$$\hat{q}_n(y^n) = \frac{\sup_{\theta\in\Theta} p_\theta(y^n)}{\int d\nu(z^n) \sup_{\theta\in\Theta} p_\theta(z^n)}$$

minimizes the worst case regret $\sup_{y^n} R(q, y^n)$. It is easily seen that the regret for this distribution, $R(\hat{q}_n, y^n)$, does not depend on the sequence $y^n$ and for all $y^n$ it satisfies

$$R(\hat{q}_n, y^n) = R_n = \log \int d\nu(y^n) \sup_{\theta\in\Theta} p_\theta(y^n). \tag{2}$$

From this, the proof that $\hat{q}_n$ achieves optimality is simple. For any $q \neq \hat{q}_n$, there will be at least one sequence $z^n$, for which $q(z^n) < \hat{q}_n(z^n)$ (note that both distributions are normalized!). Hence, we have

$$\sup_{y^n} R(q, y^n) \geq R(q, z^n) > R(\hat{q}, z^n) = \sup_{y^n} R(\hat{q}_n, y^n).$$

**4. Upper Bound on the Minimax Regret.** For the following, the definition of metric entropy, also called Kolmogorov $\epsilon$-entropy, is needed [1].

DEFINITION 1. *Let $D$ be a metric and $(S, D)$ be a complete separable metric space. A partition $\Pi$ of $S$ is a collection $\{\pi_i\}$ of Borel subsets of $S$ that are pairwise disjoint and whose union is $S$. The diameter of a set $A \subseteq S$ is given by diam$(A) = \sup_{x,y\in A} D(x, y)$. The diameter of a partition is the supremum of the diameters of the sets in the partition. For $\epsilon > 0$, by $\mathcal{D}(\epsilon, S, D)$ we denote the cardinality of the smallest finite partition of $S$ of diameter at most $\epsilon$, or $\infty$ if no such finite partition exists. The metric entropy of $(S, D)$ is defined by*

$$\mathcal{K}(\epsilon, S, D) = \log \mathcal{D}(\epsilon, S, D).$$

*We say $S$ is* totally bounded *if $\mathcal{D}(\epsilon, S, D) < \infty$ for all $\epsilon > 0$.*

DEFINITION 2. *For $\epsilon > 0$, an $\epsilon$-*separated subset *of $S$ is a subset $A \subseteq S$ such that for all distinct $x, y \in A$, $D(x, y) > \epsilon$. By the packing number $\mathcal{M}(\epsilon, S, D)$ we denote the cardinality of the largest finite $\epsilon$-separated subset* *of $S$, or $\infty$ if arbitrarily large such sets exist.* The following lemma is easily verified [1].

LEMMA 1. *For any $\epsilon > 0$,*

$$\mathcal{M}(2\epsilon, S, D) \le \mathcal{D}(2\epsilon, S, D) \le \mathcal{M}(\epsilon, S, D).$$

It follows that the metric entropy $\mathcal{K}$ (and the condition defining total boundedness) can also be defined by packing numbers in place of $\mathcal{D}$, to within a constant factor in $\epsilon$.

THEOREM 1. *Let $0 < \underline{c} \le p_\theta(y) \le \overline{c} < \infty$ for all $y \in Y$ and all $\theta \in \Theta$. Set further*

$$D_\infty(\theta, \theta') \doteq \sup_y |\log p_\theta(y) - \log p_{\theta'}(y)|$$

*for all $\theta, \theta' \in \Theta$. Then there exists a positive universal constant $A$ such that for $n \ge 1$ and for all $\epsilon > 0$,*

$$R_n \le \mathcal{K}(\epsilon, \Theta, D_\infty) + A\sqrt{n} \int_0^\epsilon \sqrt{\mathcal{K}(\delta, \Theta, D_\infty)}d\delta + 8n\epsilon^2.$$

The proof of (1) is given is a series of lemmas. We begin with some elementary steps that cast the problem into a form where the tools of empirical process theory [11, 10] can be applied. We first construct a minimal partition of $\Theta$ with $D_\infty$ diameter at most $\epsilon$ consisting of the subsets $\Theta_k,\ k = 1, \ldots, \mathcal{D}(\epsilon, \Theta, D_\infty)$ and try to control the sup in equation (2) inside of each set.

Let us fix a probability distribution $p_{\theta_k}, \theta_k \in \Theta_k$ for each set in the partition of $\Theta$. For our first lemma, we define the expectation $\mathbb{E}_k$ for each $k$ by setting

$$\mathbb{E}_k(F) = \int d\nu(y^n) p_{\theta_k}(y^n) F(y^n)$$

for any function $F(y^n)$. For each $\theta$, let

$$Z_{n,\theta}^{(k)}(y^n) = \sum_{i=1}^n \left( \log \frac{p_\theta(y_i)}{p_{\theta_k}(y_i)} - \mathbb{E}_k \log \frac{p_\theta(y_i)}{p_{\theta_k}(y_i)} \right).$$

Let further

$$S_n^{(k)}(y^n) = \sup_{\theta \in \Theta_k} |Z_{n,\theta}^{(k)}|$$

LEMMA 2. *For all $\epsilon > 0$,*

$$R_n \le \mathcal{K}(\epsilon, \Theta, D_\infty) + \log \max_k \mathbb{E}_k \exp[S_n^{(k)}]$$

Proof:

$$\begin{aligned}
e^{R_n} &= \int d\nu(y^n) \sup_{\theta\in\Theta} p_\theta(y^n) \quad \text{by Equation (2)} \\
&\le \int d\nu(y^n) \sum_{k=1}^{\mathcal{D}(\epsilon,\Theta,D_\infty)} \sup_{\theta\in\Theta_k} p_\theta(y^n) \\
&\le \mathcal{D}(\epsilon,\Theta,D_\infty) \max_k \int d\nu(y^n) \sup_{\theta\in\Theta_k} p_\theta(y^n) \\
&= \mathcal{D}(\epsilon,\Theta,D_\infty) \max_k \mathbb{E}_k \sup_{\theta\in\Theta_k} \frac{p_\theta(y^n)}{p_{\theta_k}(y^n)} \\
&= \mathcal{D}(\epsilon,\Theta,D_\infty) \max_k \mathbb{E}_k \exp\left[\sup_{\theta\in\Theta_k} \sum_{i=1}^n \log\frac{p_\theta(y_i)}{p_{\theta_k}(y_i)}\right] \\
&\le \mathcal{D}(\epsilon,\Theta,D_\infty) \max_k \mathbb{E}_k \exp\left[\sup_{\theta\in\Theta_k} \sum_{i=1}^n \left(\log\frac{p_\theta(y_i)}{p_{\theta_k}(y_i)} - \mathbb{E}_k \log\frac{p_\theta(y_i)}{p_{\theta_k}(y_i)}\right)\right] \\
&\le \mathcal{D}(\epsilon,\Theta,D_\infty) \max_k \mathbb{E}_k \exp[S_n^{(k)}].
\end{aligned}$$

The penultimate line follows from the positivity of the KL-divergence, since for each $i$

$$\mathbb{E}_k \log\frac{p_\theta(y_i)}{p_{\theta_k}(y_i)} = \int d\nu(y) p_{\theta_k}(y) \log\frac{p_\theta(y)}{p_{\theta_k}(y)} \le 0.$$

□

Now let us fix the $k$th set in the partition, omitting the index $k$ from now on when it is clear from the context. Note that $S_n(y^n)$ is the $L_\infty$ norm of the collection $\{Z_{n,\theta} : \theta \in \Theta_k\}$ of random variables. When $y^n$ is chosen randomly according to the distribution $p_{\theta_k}$, then for each fixed $\theta$, the random variable $Z_{n,\theta}$ is actually a sum of $n$ i.i.d. zero mean random variables, with distribution depending on $\theta$. This is the type of quantity that we can use the techniques of empirical process theory to bound. The first step is to relate $\mathbb{E}\exp[S_n]$ to $\mathbb{E}S_n$.

LEMMA 3. *Let $X_\theta$ be a family of functions $X_\theta : Y \to R$, indexed by $\theta \in \Theta$, such that for all $y \in Y$, $\sup_{\theta\in\Theta} |X_\theta(y)| \le C/2$. Let $y_1, \ldots, y_n$ be a collection of i.i.d. random variables and let $T_n(y^n) = \sup_{\theta\in\Theta} |\sum_{i=1}^n X_\theta(y_i)|$. Then*

$$\mathbb{E}e^{T_n} \le \exp[\frac{1}{2}nC^2]e^{\mathbb{E}T_n}.$$

The lemma is proved in the appendix using Lemma 6.16 of [11].

To apply this lemma, let $\Theta = \Theta_k$ and

$$X_\theta(y) = \log\frac{p_\theta(y)}{p_{\theta_k}(y)} - \mathbb{E}_k \log\frac{p_\theta(y)}{p_{\theta_k}(y)},$$

so that $T_n = S_n$. Note that

$$X_\theta(y) \leq 2 \sup_y |\log p_\theta(y) - \log p_{\theta_k}(y)| = 2D_\infty(\theta, \theta_k) \leq 2\epsilon \tag{3}$$

since the diameter in each set of the partition is at most $\epsilon$. Hence from Lemma 2 and Lemma 3 we have

$$R_n \leq \mathcal{K}(\epsilon, \Theta, D_\infty) + 8n\epsilon^2 + \max_k \mathbb{E}_k[S_n^{(k)}]. \tag{4}$$

To bound $\mathbb{E}S_n$, we need

DEFINITION 3. *A collection of zero mean random variables* $\{Z_\theta : \theta \in \Theta\}$ *is called a* sub-Gaussian *process with respect to the seminorm* $D$ *on* $\Theta$, *if for any* $\theta, \theta' \in \Theta$,

$$\Pr(|Z_\theta - Z_{\theta'}| > t) \leq 2e^{-\frac{1}{2}t^2/D^2(\theta,\theta')}.$$

The following lemma easily follows from Corollary 2.2.8 on page 101 of [10]

LEMMA 4. *Let* $\{Z_\theta : \theta \in \Theta\}$ *be a sub-Gaussian process under the norm* $D$ *with finite packing numbers* $\mathcal{M}(\epsilon, \Theta, D)$ *for all* $\epsilon > 0$. *Then there exists a positive universal constant* $A$ *such that for every* $\epsilon > 0$ *and for each* $\theta^* \in \Theta$

$$\mathbb{E} \sup_{\theta : D(\theta^*, \theta) \leq \epsilon} |Z_\theta| \leq \mathbb{E}|Z_{\theta^*}| + A \int_0^\epsilon \sqrt{\log \mathcal{M}(\delta, \Theta, D)} d\delta.$$

To apply this lemma, we choose $\theta^* = \theta_k$ for each set in the partition and (omitting the dependence on $k$ for convenience) set $Z_\theta = Z_{n,\theta}$. Let the density on $y^n$ be $d\nu(y^n)p_{\theta_k}(y^n)$ and $\mathbb{E}$ denote expectation under this measure, as above.

Now fix some $\theta$ and $\theta'$ in $\Theta$. Let $U(y^n) = \sum_{i=1}^n U_i$, where $U_i = \log \frac{p_\theta(y_i)}{p_{\theta'}(y_i)} - \mathbb{E} \log \frac{p_\theta(y_i)}{p_{\theta'}(y_i)}$. Then $U(y^n) = Z_\theta(y^n) - Z_{\theta'}(y^n)$. As in Equation 3, it is clear that $|U_i| \leq 2D_\infty(\theta, \theta')$. Thus $U$ is a sum of $n$ bounded i.i.d. random variables. Hence, we may apply Hoeffding's inequality [15] to obtain

$$\Pr(|U| > t) \leq 2 \exp\left[-t^2/(2nD_\infty^2(\theta, \theta'))\right],$$

Since $U = Z_\theta - Z_{\theta'}$, this shows that $Z_\theta$ is sub-Gaussian with respect to $D = \sqrt{n}D_\infty$.

Since $\theta^* = \theta_k$, and $Z_{\theta^*} = Z_{n,\theta_k} = 0$, it follows from Lemma 4 and Equation (4) that

$$R_n \leq \mathcal{K}(\epsilon, \Theta, D_\infty) + 8n\epsilon^2 + A \int_0^\epsilon \sqrt{\log \mathcal{M}(\delta, \Theta, D)} d\delta. \tag{5}$$

Since $D = \sqrt{n}D_\infty$, the theorem follows. □

**5. Lower Bound.** A lower bound on $R_n$ is provided in terms of the metric entropy of $\Theta$ with respect to the so called Hellinger distance, which is defined as

$$D_H(\theta,\theta') = \left\{ \int d\nu(y) \left( \sqrt{p_\theta(y)} - \sqrt{p_{\theta'}(y)} \right)^2 \right\}^{\frac{1}{2}} .$$

The bound is established from the simple fact that $R_n$ is not smaller than the minimax risk in the framework where the data are generated at random from a distribution in $\Theta$, that is, from equation (1)

$$R_n \geq \inf_q \sup_\theta \int d\nu(y^n) p_\theta(y^n) \left\{ -\log q(y^n) + \log p_\theta(y^n) \right\} . \tag{6}$$

A general lower bound on the latter quantity for product distributions $p_\theta(y^n)$ was recently obtained in [12]. From Lemma 7, part 1 of [12] and Equation (6) above, we get

LEMMA 5. *Assume* $(\Theta, D_H)$ *is totally bounded. Then for all* $n \geq 1$,

$$R_n \geq \sup_{\epsilon \geq 0} \min\{\mathcal{K}(\epsilon,\Theta,D_H), \frac{n\epsilon^2}{8}\} - \log 2$$

Thus $R_n$ is bounded below in terms of the Hellinger metric entropy and above in terms of the $D_\infty$ metric entropy. When these entropies are close, the resulting bounds can sometimes be used to characterize the growth rate of $R_n$, as seen in the following.

**6. Example: A Nonparametric Family of Densities.** Many interesting nonparametric families of densities have metric entropies which scale as $\mathcal{K}(\delta,\Theta,D_\infty) \leq const(\frac{1}{\delta})^\alpha$ as $\delta \to 0$. Assuming that $\alpha < 2$, we can show that in the same limit

$$\int_0^\epsilon \sqrt{\mathcal{K}(\delta,\Theta,D_\infty)} d\delta \leq const\ \epsilon^{\frac{2-\alpha}{2}} .$$

In such a case, $R_n$ in theorem (1) is easily bounded by setting $\epsilon \propto n^{-\frac{1}{2+\alpha}}$ for $\alpha < 2$. Then one has for large $n$,

$$R_n \leq const\ n^{\frac{\alpha}{2+\alpha}} \tag{7}$$

for $\alpha < 2$.

A common example is given by the Lipschitz class $\Theta$ of densities on a real interval which have all derivatives $|p_\theta^{(i)}(y)| \leq C_i$, for $i = 0, 1, \ldots, r$, and $|p_\theta^{(r)}(y) - p_\theta^{(r)}(y')| \leq C|y-y'|^\gamma$, $(0 < \gamma \leq 1)$. If we further assume that all densities are uniformly bounded away from zero, we can use a result of [13] to show that the metric entropy behaves like $\mathcal{K}(\epsilon,\Theta,D_\infty) = const \left(\frac{1}{\epsilon}\right)^{\frac{1}{r+\gamma}}$, for $\epsilon \to 0$, which yields

$$R_n \leq const\ n^{\frac{1}{2(r+\gamma)+1}}$$

for large $n$ and $r+\gamma > \frac{1}{2}$. As can be shown [12] for this example, the lower bound (5) yields the same exponent for increase of $R_n$ with $n$ as the upper bound. Since the lower bound is related to the statistical risk of the random sequence framework, the latter result also shows, that (at least for the present example) the more pessimistic assumption of the worst sequence framework does not lead to much higher extra losses than those of the random sequence framework. A similar result was obtained for parametric families in [4, 6, 7, 9]. Whether this will be true in significantly more general settings is a problem for further research.

## 7. Appendix.

Proof of lemma (3): Write $T_n - \mathbb{E}T_n$ as a sum of Martingale–differences $d_j$, i.e. $T_n - \mathbb{E}T_n = \sum_{j=1}^n d_j$ where

$$d_j = \mathbb{E}^{\mathcal{A}_j} T_n - \mathbb{E}^{\mathcal{A}_{j-1}} T_n. \tag{8}$$

Here, for any $k$, and any function $F(y^n)$, $\mathbb{E}^{\mathcal{A}_k} F \doteq \mathbb{E}(F(y^n)|y_1,\ldots,y_k)$ denotes the conditional expectation given $y_1,\ldots,y_k$.

The proof is based on the following inequality

$$|d_j| \leq \sup_\theta |X_\theta(y_j)| + \mathbb{E}\sup_\theta |X_\theta(y)| \tag{9}$$

which is due to V. Yurinskii and is proved in Lemma 6.16 on page 163 of [11]. For completeness, we give a sketch of the proof here. With $T_n = \sup_{\theta\in\Theta} |\sum_{i=1}^n X_\theta(y_i)|$ and the definition $T_{n\setminus j} = \sup_\theta |\sum_{i\neq j} X_\theta(y_i)|$, we get from the triangular inequality for the sup norm

$$|T_n - T_{n\setminus j}| \leq \sup_\theta |X_\theta(y_j)|. \tag{10}$$

Further, we can write

$$d_j = \mathbb{E}^{\mathcal{A}_j} T_n - \mathbb{E}^{\mathcal{A}_{j-1}} T_n - \left\{\mathbb{E}^{\mathcal{A}_j} T_{n\setminus j} - \mathbb{E}^{\mathcal{A}_{j-1}} T_{n\setminus j}\right\}. \tag{11}$$

This is true, because by the independence of $\sum_{i\neq j} X_\theta(y_i)$ and $y_j$, the terms in the curly brackets give zero. Hence, using (10) and (11) we get

$$\begin{align} |d_j| \leq\ & \mathbb{E}^{\mathcal{A}_j} \sup_\theta |X_\theta(y_j)| + \mathbb{E}^{\mathcal{A}_{j-1}} \sup_\theta |X_\theta(y_j)| \tag{12}\\ =\ & \sup_\theta |X_\theta(y_j)| + \mathbb{E}\sup_\theta |X_\theta(y)|, \tag{13} \end{align}$$

which proves (9).

We now use the properties of conditional expectations to bound

$$\begin{aligned} \mathbb{E}e^{T_n - \mathbb{E}T_n} &= \mathbb{E}e^{\sum_{j=1}^n d_j} \\ &= \mathbb{E}e^{\sum_{j=1}^{n-1} d_j} \mathbb{E}^{\mathcal{A}_{n-1}} e^{d_n} \\ &\leq \mathbb{E}e^{\sum_{j=1}^{n-1} d_j} \exp[\frac{1}{2}(\sup_\theta |X_\theta(y_n)| + \mathbb{E}\sup_\theta |X_\theta(y_n)|)^2] \end{aligned}$$

$$\leq \exp\left[\frac{n}{2}(\sup_\theta |X_\theta(y_n)| + \mathbb{E}\sup_\theta |X_\theta(y_n)|)^2\right]$$
$$\leq \exp[\frac{1}{2}nC^2].$$

In the first inequality, the Martingale property $\mathbb{E}^{\mathcal{A}_{n-1}}d_n = 0$ is used together with (9) and the fact that any bounded random variable $V$ with $|V| \leq A$ and $\mathbb{E}V = 0$ satifies $\mathbb{E}e^V \leq e^{\frac{1}{2}A^2}$. The second inequality is obtained by iterating the first one. □

**8. Acknowledgements.** We acknowledge inspiring discussions with Peter Auer, Andrew Barron, Yoav Freund, Nick Hengartner, Manfred Warmuth and Jon A. Wellner.

## REFERENCES

[1] Kolmogorov, A. N. and V. M. Tihomirov, *$\epsilon$-Entropy and $\epsilon$-Capacity of Sets in Functional Spaces*, Amer. Math. Soc. Translations (Ser. 2), 17, 277-364 (1961).

[2] V.G. Vovk, *Aggregating strategies*, Proceedings of the 1990 conference on computational learning theory, Morgan Kaufmann, 371–381 (1990).

[3] N. Cesa-Bianchi, Y. Freund, D.H. Helmbold, D. Haussler, R.E. Schapire, and M.K. Warmuth, *How to use expert advice*, in 25th Annual ACM Symposium on Theory of Computing, 382–392, San Diego, CA (1993).

[4] Y. Freund, *Predicting a binary sequence as well as the optimal biased coin*, Proceedings of the ninth annual conference on computational learning theory, ACM Press, (1996).

[5] J. Shtarkov, *Coding of discrete sources with unknown statistics*, In: *Topics in Information Theory, 559–574*, I. Csiszar and P. Elias, editors, North Holland, Amsterdam, 1975.

[6] A.R. Barron and Q. Xie, *Asymptotic minimax loss for data compression, gambling, and prediction*, Proceedings of the ninth annual conference on computational learning theory, ACM Press, (1996).

[7] J. Rissanen, *Fisher Information and Stochastic Complexity*, IEEE Trans. on Inf. Theory 42, 40–47, (1996).

[8] T. Cover and Joy A. Thomas, *Elements of Information Theory*, Wiley Series in Telecommunications, New York, 1991.

[9] T.M Cover and E. Ordentlich, *Universal portfolios with side information*, IEEE Transactions on Information Theory 42(2), 348-363, (1996).

[10] Aad W. van der Vaart and Jon A. Wellner, *Weak Convergence and Empirical Processes*, Springer Series in Statistics, 1996.

[11] M. Ledoux and M. Talagrand, *Probability in Banach Spaces: Isoperimetry and Processes*, Springer Verlag, Berlin (1991).

[12] D. Haussler and M. Opper, *Mutual Information, Metric Entropy, and Risk in Estimation of Probability Distributions*, Annals of Statistics 25 (6), (December, 1997).

[13] G.F. Clements, *Entropy of several sets of real valued functions*, Pacific J. Math. 13, 1085 (1963).

[14] M.J. Weinberger, N. Merhav and M. Feder, *Optimal Sequential Probability Assignment for Individual Sequences*, IEEE Trans. on Inf. Theory 40, 384–396 (1994).

[15] W. Hoeffding, *Probability Inequalities for Sums of Bounded Random Variables*, American Statistical Association Journal 58, 13–30, (1963).