References

[1] J. Borges, C. Fernández, and K. Phelps, "Quaternary Reed-Muller codes," *IEEE Trans. Inf. Theory*, vol. 52, no. 7, Jul. 2005.

[2] J. Borges, K. P. Phelps, J. Rifá, and V. Zinoviev, "On $\mathbb{Z}_4$-linear preparata-like and kerdock-like codes," *IEEE Trans. Inf. Theory*, vol. 50, no. 11, pp. 2834–2843, Nov. 2003.

[3] A. Hammons, P. V. Kumar, A. R. Calderbank, N. J. A. Sloane, and P. Solé, "The $\mathbb{Z}_4$-linearity of kerdock, preparata, goethals and related codes," *IEEE Trans. Inf. Theory*, vol. 41, pp. 301–319, 1994.

[4] X.-D. Hou, J. T. Lahtonen, and S. Koponen, "The Reed-Muller Code $R(r, m)$ is not $\mathbb{Z}_4$-linear for $3 \leq r \leq m-2$," *IEEE Trans. Inf. Theory*, vol. 45, pp. 798–799, 1998.

[5] F. J. MacWilliams and N. J. A. Sloane, *The Theory of Error-Correcting Codes*. Amsterdam, The Netherlands: North-Holland, 1977.

[6] V. S. Pless, W. C. Huffman, and R. A. Brualdi, *Handbook of Coding Theory: Volume I*. Amsterdam, The Netherlands: North-Holland, 1998.

[7] K. Phelps and J. Rifà, "On binary additive 1-perfect codes," *IEEE Trans. Inf. Theory*, vol. 49, no. 9, pp. 2587–2591, 2002.

[8] J. Pujol and J. Rifà, "Translation-invariant propelinear codes," *IEEE Trans. Inf. Theory*, vol. 44, no. 2, pp. 590–598, 1997.

[9] Z.-X. Wan, *Quaternary Codes*. Singapore: World Scientific, 1997.

Improved Risk Tail Bounds for On-Line Algorithms

Nicolò Cesa-Bianchi and Claudio Gentile

***Abstract*—Tight bounds are derived on the risk of models in the ensemble generated by incremental training of an arbitrary learning algorithm. The result is based on proof techniques that are remarkably different from the standard risk analysis based on uniform convergence arguments, and improves on previous bounds published by the same authors.**

***Index Terms*—Martingales, on-line learning, risk bounds, statistical learning theory.**

I. Introduction

In this correspondence, we analyze the risk of models selected from the ensemble produced by training a learning algorithm incrementally on a sequence of independent and identically distributed (i.i.d.) data. A learner trained in an incremental fashion receives the examples one by one in a sequence of trials. At each new trial the learner's performance is measured by evaluating the loss of its current model on the next example of the sequence. The example is then used by the learner to adjust its parameters. Afterwards, the resulting updated model is added to the ensemble. The sum of the losses measured in this way is called on-line (or cumulative) loss.

Since any learning algorithm can be trained this way (for example, by retraining at each new trial on the entire set of examples observed so far), we call *online* any algorithm that is trained incrementally (irrespective of efficiency). For certain algorithms, such as the well-known Perceptron algorithm, this type of training is a basic and efficient mode of operation. Such algorithms are thus the most natural target of our theory.

Let M_n denote the per-trial on-line loss incurred by the learner on a training sequence of length n. We study the ensemble of models generated by the sequence of adjustments during the training process. We describe a deterministic procedure that uses upper confidence bounds for selecting a model from the ensemble. This procedure implies that the risk of the chosen hypothesis is, with high probability, at most

$$M_n + O\left(\frac{(\ln n)^2}{n} + \sqrt{M_n \frac{\ln n}{n}}\right).$$

Hence the per-trial on-line loss is almost always close to the risk of a certain model that—as we will see—is not necessarily the last one in the ensemble. In particular, the risk converges to M_n at rate $O(\sqrt{(\ln n)/n})$ and vanishes at rate $(\ln n)^2/n$ whenever the on-line loss nM_n is $O(1)$.

This result is proven through a refinement of the techniques used in [4] to prove the substantially weaker bound $M_n + O(\sqrt{(\ln n)/n})$. As in the proof of this older result, we analyze the empirical process associated with a run of the on-line learner using exponential inequalities for martingales. However, this time we control the large deviations of the on-line process using Bernstein's inequality rather than the Azuma–Hoeffding inequality. This provides a much tighter bound on the average risk of the ensemble. Finally, we relate the risk of a specific model within the ensemble to the average risk. As in [4], we select this hypothesis using a deterministic sequential testing procedure, but the use of Bernstein's inequality makes the analysis of this procedure significantly more involved.

The study of the statistical risk of models generated by on-line algorithms, initiated by Littlestone [8], uses tools that are sharply different from those used for uniform convergence analysis, a popular approach based on the manipulation of suprema of empirical processes (see, e.g., [1], [2], [5], [7], [9]–[11]). Unlike uniform convergence, which is tailored to empirical risk minimization, our bounds apply to *any* learning algorithm, since we obtain an ensemble of hypotheses from any learner via incremental training. Finally note that our bounds are naturally data-dependent, as the per-trial on-line loss M_n depends on the empirical behavior of the on-line algorithm on the realized training sequence.

Notation: An *example* is a pair (x, y), where $x \in \mathcal{X}$ (which we call *instance*) is a data element and $y \in \mathcal{Y}$ is the *label* associated with it. In our setup, $\mathcal{X}$ and $\mathcal{Y}$ are generic sets such that $\mathcal{X} \times \mathcal{Y}$ is measurable. In typical applications instances x are tuples of numerical and/or symbolic attributes, and $\mathcal{Y}$ is a finite set of symbols (the class elements) or an interval of the real line, depending on whether the task is classification or regression. Following a standard terminology in learning theory we call *hypothesis* the classifier or regressor generated by a learning algorithm after training. We allow a learning algorithm to output hypotheses of the form $H : \mathcal{X} \to \mathcal{D}$, where $\mathcal{D}$ is a decision space not necessarily equal to $\mathcal{Y}$. The predictive performance of hypothesis H on example (x, y) is measured by the quantity $\ell(H(x), y)$, where $\ell : \mathcal{D} \times \mathcal{Y} \to \mathbb{R}$ is a nonnegative and bounded *loss function*.

II. A Bound on the Average Risk

We start by defining on-line algorithms, that is, generic learners that are trained incrementally. An on-line algorithm $\mathtt{A}$ works in a sequence

Manuscript received April 20, 2006; revised May 7, 2007. This work was supported in part by the PASCAL Network of Excellence under EC Grant 506778. The material in this correspondence was presented in part at the 18th Conference on Neural Information Processing Systems (NIPS 2005), Vancouver, BC, Canada, December 2005. A preliminary version of this paper appeared in the *Proc. 18th Conf. Neural Information Processing Systems (NIPS 2005)*, Cambridge, MA: MIT Press, 2006. This publication only reflects the authors' views.

N. Cesa-Bianchi is with Dipartimento di Scienze dell'Informazione, Università di Milano, 20135 Milano, Italy (e-mail: cesa-bianchi@dsi.unimi.it).

C. Gentile is with Dipartimento di Informatica e Comunicazione, Università dell'Insubria, 21100 Varese, Italy (e-mail: claudio.gentile@uninsubria.it).

Communicated by P. L. Bartlett, Associate Editor for Pattern Recognition, Statistical Learning, and Inference.

Digital Object Identifier 10.1109/TIT.2007.911292

of trials. In each trial $t = 1, 2, \ldots$ the algorithm takes in input a hypothesis H_{t-1} and an example $Z_t = (X_t, Y_t)$, and returns a new hypothesis H_t to be used in the next trial. (The initial hypothesis H_0 is an arbitrary function from $\mathcal{X}$ to $\mathcal{D}$).

We follow the standard assumptions in statistical learning: the sequence of examples $Z^n = ((X_1, Y_1), \ldots, (X_n, Y_n))$ is drawn i.i.d. according to an unknown distribution over $\mathcal{X} \times \mathcal{Y}$. We also assume that the loss function ℓ satisfies $0 \leq \ell \leq 1$. The success of a hypothesis H is measured by the *risk*, denoted by $R(H)$. This is the expected loss of H on an example (X, Y) drawn from the underlying distribution, $R(h) = \mathbb{E}\ell(H(X), Y)$. Define also $\hat{R}(H)$ to be the empirical risk of H on a sample Z^n,

$$\hat{R}(H) = \frac{1}{n}\sum_{t=1}^{n} \ell(H(X_t), Y_t).$$

Given a sample Z^n and an on-line algorithm $\mathtt{A}$, we use $H_0, H_1, \ldots, H_{n-1}$ to denote the *ensemble of hypotheses generated by* $\mathtt{A}$. Note that the ensemble is a function of the random training sample Z^n. Our bounds hinge on the per-trial on-line loss

$$M_n = M_n(Z^n) = \frac{1}{n}\sum_{t=1}^{n} \ell(H_{t-1}(X_t), Y_t)$$

a sample statistic that can be easily computed as the on-line algorithm is run on Z^n.

The following bound, a consequence of Bernstein's inequality for martingales (see, e.g., Freedman [6]), is of primary importance for proving our results.

Lemma 1: Let $L_1, L_2, \ldots$ be a sequence of random variables, $0 \leq L_t \leq 1$. Define the bounded martingale difference sequence $V_t = \mathbb{E}[L_t \,|\, L_1, \ldots, L_{t-1}] - L_t$ and the associated martingale $S_n = V_1 + \cdots + V_n$ with conditional variance

$$K_n = \sum_{t=1}^{n} \mathrm{Var}[L_t \,|\, L_1, \ldots, L_{t-1}].$$

Then, for all $s, k \geq 0$

$$\mathbb{P}(S_n \geq s, K_n \leq k) \leq \exp\left(-\frac{s^2}{2k + 2s/3}\right).$$

The next proposition, derived from Lemma 1, establishes a bound on the average risk of the ensemble of hypotheses.

Proposition 2: Let $H_0, \ldots, H_{n-1}$ be the ensemble of hypotheses generated by an arbitrary on-line algorithm $\mathtt{A}$. Then, for any $0 < \delta \leq 1$,

$$\mathbb{P}\left(\frac{1}{n}\sum_{t=1}^{n} R(H_{t-1}) \geq M_n + \frac{36}{n}\ln\left(\frac{nM_n + 3}{\delta}\right) + 2\sqrt{\frac{M_n}{n}\ln\left(\frac{nM_n + 3}{\delta}\right)}\right) \leq \delta.$$

In independent work [13], Zhang derived a new martingale inequality which he used to prove a bound on the average risk of the ensemble of the form

$$M_n + \frac{3(1 - M_n)}{n}\ln\left(\frac{nM_n + 2}{\delta}\right) + 2\sqrt{(1 - M_n)\frac{M_n}{n}\ln\left(\frac{nM_n + 2}{\delta}\right)}.$$

Compared to Proposition 2, Zhang's bound has a better leading constant on the second term; moreover, the bound does not become vacuous (i.e., larger than 1) when M_n is close to 1(i.e., when the learner is performing badly).

Proof: Let

$$\mu_n = \frac{1}{n}\sum_{t=1}^{n} R(H_{t-1})$$

and $V_{t-1} = R(H_{t-1}) - \ell(H_{t-1}(X_t), Y_t)$ for $t \geq 1$.

Let κ_t be the conditional variance

$$\kappa_t = \mathrm{Var}(\ell(H_{t-1}(X_t), Y_t) \,|\, Z_1, \ldots, Z_{t-1}).$$

Also, set for brevity $K_n = \sum_{t=1}^{n} \kappa_t$, $K_n' = \lfloor \sum_{t=1}^{n} \kappa_t \rfloor$, and introduce the function

$$A(x) = 2\ln\frac{(x+1)(x+3)}{\delta},\ x \geq 0.$$

We find upper and lower bounds on the probability

$$\mathbb{P}\left(\sum_{t=1}^{n} V_{t-1} \geq A(K_n) + \sqrt{A(K_n)K_n}\right). \tag{1}$$

The upper bound is derived through a simple stratification argument over Lemma 1. We can write

$$\begin{aligned}
&\mathbb{P}\left(\sum_{t=1}^{n} V_{t-1} \geq A(K_n) + \sqrt{A(K_n)K_n}\right)\\
&\quad\leq \mathbb{P}\left(\sum_{t=1}^{n} V_{t-1} \geq A(K_n') + \sqrt{A(K_n')K_n'}\right)\\
&\quad\leq \sum_{s=0}^{n}\mathbb{P}\left(\sum_{t=1}^{n} V_{t-1} \geq A(s) + \sqrt{A(s)s}, K_n' = s\right)\\
&\quad\leq \sum_{s=0}^{n}\mathbb{P}\left(\sum_{t=1}^{n} V_{t-1} \geq A(s) + \sqrt{A(s)s}, K_n \leq s + 1\right)\\
&\quad\leq \sum_{s=0}^{n}\exp\left(-\frac{(A(s) + \sqrt{A(s)s})^2}{\frac{2}{3}(A(s) + \sqrt{A(s)s}) + 2(s+1)}\right)
\end{aligned}$$

where we used Lemma 1 in the last step. Since

$$\frac{(A(s) + \sqrt{A(s)s})^2}{\frac{2}{3}(A(s) + \sqrt{A(s)s}) + 2(s+1)} \geq A(s)/2$$

for all $s \geq 0$, we obtain

$$\begin{aligned}
(1) &\leq \sum_{s=0}^{n} e^{-A(s)/2}\\
&= \sum_{s=0}^{n}\frac{\delta}{(s+1)(s+3)} < \delta.
\end{aligned} \tag{2}$$

As far as the lower bound on (1) is concerned, we note that our assumption $0 \leq \ell \leq 1$ implies $\kappa_t \leq R(H_{t-1})$ for all t which, in turn, gives $K_n \leq n\mu_n$. Thus

$$\begin{aligned}
(1) &= \mathbb{P}(n\mu_n - nM_n \geq A(K_n) + \sqrt{A(K_n)K_n})\\
&\geq \mathbb{P}(n\mu_n - nM_n \geq A(n\mu_n) + \sqrt{A(n\mu_n)n\mu_n})\\
&= \mathbb{P}(2n\mu_n \geq 2nM_n + 3A(n\mu_n)\\
&\quad + \sqrt{4nM_nA(n\mu_n) + 5A(n\mu_n)^2})\\
&= \mathbb{P}\left(x \geq B + \frac{3}{2}A(x) + \sqrt{BA(x) + \frac{5}{4}A^2(x)}\right),
\end{aligned}$$

where we set for brevity $x = n\mu_n$, $B = nM_n$, and divided by two inside the probability. We would like to solve the inequality

$$x \geq B + \frac{3}{2}A(x) + \sqrt{BA(x) + \frac{5}{4}A^2(x)} \tag{3}$$

w.r.t. x. More precisely, we would like to find a suitable upper bound on the (unique) x^* such that the above is satisfied as an equality.

A (tedious) derivative argument along with the upper bound $A(x) \le 4\ln(\frac{x+3}{\delta})$ show that

$$x' = B + 2\sqrt{B\ln\left(\frac{B+3}{\delta}\right)} + 36\ln\left(\frac{B+3}{\delta}\right)$$

makes the left-hand side of (3) larger than its right-hand side. Thus x' is an upper bound on x^*, and we conclude that

$$(1) \ge \mathbb{P}\left(x \ge B + 2\sqrt{B\ln\left(\frac{B+3}{\delta}\right)} + 36\ln\left(\frac{B+3}{\delta}\right)\right)$$

which, recalling the definitions of x and B, and combining with (2), proves the bound. □

III. Selecting A Good Hypothesis From the Ensemble

If the decision space $\mathcal{D}$ of $\mathtt{A}$ is a convex set and the loss function ℓ is convex in its first argument, then via Jensen's inequality we can directly apply the bound of Proposition 2 to the risk of the *average hypothesis* $\bar{H} = \frac{1}{n}\sum_{t=1}^{n} H_{t-1}$. This yields

$$\mathbb{P}\left(R(\bar{H}) \ge M_n + \frac{36}{n}\ln\left(\frac{nM_n+3}{\delta}\right) + 2\sqrt{\frac{M_n}{n}\ln\left(\frac{nM_n+3}{\delta}\right)}\right) \le \delta. \quad (4)$$

Observe that this is a $O(1/n)$ bound whenever the cumulative loss nM_n is $O(1)$.

If the convexity assumptions do not hold (as in the case of classification problems), then the bound in (4) applies to a random hypothesis in the ensemble (this was investigated in [3] though with different goals).

In this section, we show how to *deterministically* pick from the ensemble a hypothesis with a good risk bound. Although based on Proposition 2, the bound we prove for this hypothesis is not directly comparable to bound (4) for the average or random hypothesis (see the discussion before the proof of Theorem 4).

To see how this deterministic choice could be made, let us first introduce the two functions

$$\mathcal{E}_\delta(r,t) = \frac{8B}{3(n-t)} + \sqrt{\frac{2Br}{n-t}}$$
$$c_\delta(r,t) = \mathcal{E}_\delta\left(r + \sqrt{\frac{2Br}{n-t}}, t\right)$$

with $B = \ln\frac{n(n+2)}{\delta}$.

Let $\hat{R}(H_t, t+1) + \mathcal{E}_\delta(\hat{R}(H_t, t+1), t)$ be the *penalized empirical risk* of hypothesis H_t, where

$$\hat{R}(H_t, t+1) = \frac{1}{n-t}\sum_{i=t+1}^{n} \ell(H_t(X_i), Y_i)$$

is the empirical risk of H_t on the remaining sample $Z_{t+1}, \ldots, Z_n$. We now analyze the performance of the learning algorithm that returns the hypothesis $\hat{H}$ minimizing the penalized risk estimate over all hypotheses in the ensemble, i.e.,[1]

$$\hat{H} = \arg\min_{0\le t<n}\left(\hat{R}(H_t, t+1) + \mathcal{E}_\delta(\hat{R}(H_t, t+1), t)\right). \quad (5)$$

It is worth stressing that the actual index t of $\hat{H}$ depends on how the training set size n compares to δ. Indeed, if n is "too small," then t tends to be close to 1, whereas if n is "large enough" then t tends to be close to n. Note that this behavior is different from what one expects when the main quantity of interest is the expectation of the risk (see the discussion in Section IV).

Lemma 3: Let $H_0, \ldots, H_{n-1}$ be the ensemble of hypotheses generated by an arbitrary on-line algorithm $\mathtt{A}$ working with a loss ℓ satisfying $0 \le \ell \le 1$. Then, for any $0 < \delta \le 1$, the hypothesis $\hat{H}$ satisfies

$$\mathbb{P}\left(R(\hat{H}) > \min_{0\le t<n}\left(R(H_t) + 2c_\delta(R(H_t), t)\right)\right) \le \delta.$$

Proof: We introduce the following short-hand notation

$$\rho_t = \hat{R}(H_t, t+1),$$
$$\hat{T} = \operatorname*{argmin}_{0\le t<n}(\rho_t + \mathcal{E}_\delta(\rho_t, t))$$
$$T^* = \operatorname*{argmin}_{0\le t<n}(R(H_t) + 2c_\delta(R(H_t), t)).$$

Also, let $H^* = H_{T^*}$ and $\rho^* = \hat{R}(H_{T^*}, T^*+1) = \rho_{T^*}$. Note that $\hat{H}$ defined in (5) coincides with $H_{\hat{T}}$. Finally, let

$$Q(r,t) = \frac{\sqrt{2B(2B + 9r(n-t))} - 2B}{3(n-t)}.$$

With this notation we can write

$$\begin{aligned}
&\mathbb{P}(R(\hat{H}) > R(H^*) + 2c_\delta(R(H^*), T^*)) \\
&\quad \le \mathbb{P}(R(\hat{H}) > R(H^*) + 2c_\delta(\rho^* - Q(\rho^*, T^*), T^*)) \\
&\qquad + \mathbb{P}(R(H^*) < \rho^* - Q(\rho^*, T^*)) \\
&\quad \le \mathbb{P}(R(\hat{H}) > R(H^*) + 2c_\delta(\rho^* - Q(\rho^*, T^*), T^*)) \\
&\qquad + \sum_{t=0}^{n-1}\mathbb{P}(R(H_t) < \rho_t - Q(\rho_t, t)).
\end{aligned}$$

Applying the standard Bernstein's inequality (see, e.g., ([5, Ch. 8])) to the random variables ρ_t with $|\rho_t| \le 1$ and expected value $R(H_t)$, and upper bounding the variance of ρ_t with $R(H_t)$, yields

$$\mathbb{P}\left(R(H_t) < \rho_t - \frac{B + \sqrt{B(B + 18(n-t)R(H_t))}}{3(n-t)}\right) \le e^{-B}.$$

With a little algebra, it is easy to show that

$$R(H_t) < \rho_t - \frac{B + \sqrt{B(B + 18(n-t)R(H_t))}}{3(n-t)}$$

is equivalent to $R(H_t) < \rho_t - Q(\rho_t, t)$. Hence, we get

$$\begin{aligned}
&\mathbb{P}(R(\hat{H}) > R(H^*) + 2c_\delta(R(H^*), T^*)) \\
&\le \mathbb{P}(R(\hat{H}) > R(H^*) + 2c_\delta(\rho^* - Q(\rho^*, T^*), T^*)) + ne^{-B} \\
&\le \mathbb{P}(R(\hat{H}) > R(H^*) + 2\mathcal{E}_\delta(\rho^*, T^*)) + ne^{-B} \qquad (6)
\end{aligned}$$

where in the last step we used

$$Q(r,t) \le \sqrt{\frac{2Br}{n-t}} \quad \text{and} \quad c_\delta\left(r - \sqrt{\frac{2Br}{n-t}}, t\right) = \mathcal{E}_\delta(r,t).$$

[1] Note that, from an algorithmic point of view, this hypothesis is fairly easy to compute. In particular, if the underlying on-line algorithm is a standard kernel-based algorithm, $\hat{H}$ can be calculated via a single sweep through the example sequence.

Now, we focus on the probability term in (6), and set for brevity $\mathcal{E} = \mathcal{E}_\delta(\rho^*, T^*)$. We have

$$\begin{aligned}
&\mathbb{P}(R(\hat{H}) > R(H^*) + 2\mathcal{E}) \\
&\quad = \mathbb{P}(R(\hat{H}) > R(H^*) + 2\mathcal{E}, \rho_{\hat{T}} + \mathcal{E}_\delta(\rho_{\hat{T}}, \hat{T}) \le \rho^* + \mathcal{E}) \\
&\quad (\text{since } \rho_{\hat{T}} + \mathcal{E}_\delta(\rho_{\hat{T}}, \hat{T}) \le \rho^* + \mathcal{E} \text{ holds with certainty}) \\
&\quad \le \sum_{t=0}^{n-1} \mathbb{P}(\rho_t + \mathcal{E}_\delta(\rho_t, t) \le \rho^* + \mathcal{E},\ R(H_t) > R(H^*) + 2\mathcal{E}). \quad (7)
\end{aligned}$$

Now, if $\rho_t + \mathcal{E}_\delta(\rho_t, t) \le \rho^* + \mathcal{E}$ holds, then at least one of the following three conditions:

$$\begin{aligned}
\rho_t &\le R(H_t) - \mathcal{E}_\delta(\rho_t, t), \\
\rho^* &> R(H^*) + \mathcal{E}, \\
2\mathcal{E} &> R(H_t) - R(H^*)
\end{aligned}$$

must hold. Hence, for any fixed t, we can decompose into three probability terms. We can write

$$\begin{aligned}
&\mathbb{P}(\rho_t + \mathcal{E}_\delta(\rho_t, t) \le \rho^* + \mathcal{E}, R(H_t) > R(H^*) + 2\mathcal{E}) \\
&\quad \le \mathbb{P}(\rho_t \le R(H_t) - \mathcal{E}_\delta(\rho_t, t), \\
&\qquad R(H_t) > R(H^*) + 2\mathcal{E}) \\
&\qquad + \mathbb{P}(\rho^* > R(H^*) + \mathcal{E}, R(H_t) > R(H^*) + 2\mathcal{E}) \\
&\qquad + \mathbb{P}(R(H_t) - R(H^*) < 2\mathcal{E} \\
&\qquad R(H_t) > R(H^*) + 2\mathcal{E}) \\
&\quad (\text{note that the last probability term turns out to be } 0). \\
&\quad \le \mathbb{P}(\rho_t \le R(H_t) - \mathcal{E}_\delta(\rho_t, t)) + \mathbb{P}(\rho^* > R(H^*) + \mathcal{E}) \qquad (8)
\end{aligned}$$

Plugging (8) into (7) we have

$$\begin{aligned}
&\mathbb{P}(R(\hat{H}) > R(H^*) + 2\mathcal{E}) \\
&\quad \le \sum_{t=0}^{n-1} \mathbb{P}(\rho_t \le R(H_t) - \mathcal{E}_\delta(\rho_t, t)) + n\,\mathbb{P}(\rho^* > R(H^*) + \mathcal{E}) \\
&\quad \le n e^{-B} + n \sum_{t=0}^{n-1} \mathbb{P}(\rho_t \ge R(H_t) + \mathcal{E}_\delta(\rho_t, t)) \\
&\quad \le\ n e^{-B} + n^2 e^{-B}
\end{aligned}$$

where in the last two inequalities we applied again Bernstein's inequality to the random variables ρ_t with mean $R(H_t)$. Putting together we obtain

$$\mathbb{P}(R(\hat{H}) > R(H^*) + 2c_\delta(R(H^*), T^*)) \le (2n + n^2)e^{-B}$$

which, recalling that $B = \ln \frac{n(n+2)}{\delta}$, implies the thesis. □

Fix $n \ge 1$ and $\delta \in (0, 1]$. For each $t = 0, \dots, n-1$, introduce the function

$$f_t(x) = x + \frac{11C}{3}\frac{\ln(n-t) + 1}{n-t} + 2\sqrt{\frac{2Cx}{n-t}}, \quad x \ge 0$$

where $C = \ln \frac{2n(n+2)}{\delta}$. Note that each f_t is monotonically increasing. We are now ready to state and prove the main result of this correspondence.

Theorem 4: Fix any loss function ℓ satisfying $0 \le \ell \le 1$. Let $H_0, \dots, H_{n-1}$ be the ensemble of hypotheses generated by an arbitrary on-line algorithm $\mathtt{A}$ and let $\hat{H}$ be the hypothesis minimizing the penalized empirical risk expression obtained by replacing δ with $\delta/2$ in (5). Then, for any $0 < \delta \le 1$, $\hat{H}$ satisfies

$$\mathbb{P}\left(R(\hat{H}) \ge \min_{0 \le t < n} f_t\left(M_{t,n} + \frac{36}{n-t}\ln\frac{2n(n+3)}{\delta} + 2\sqrt{\frac{M_{t,n}\ln\frac{2n(n+3)}{\delta}}{n-t}}\right)\right) \le \delta$$

where $M_{t,n} = \frac{1}{n-t}\sum_{i=t+1}^{n} \ell(H_{i-1}(X_i), Y_i)$ is the average loss of the online algorithm on the suffix $\{t+1, \dots, n\}$.

Note that, due to its dependence on the *best* penalized average loss over suffixes $t+1, \dots, n$, this bound is generally incomparable to the corresponding bound (4) for the average or random hypothesis. In order to force a comparison between the two bounds, we can weaken Theorem 4 by upper bounding the minimum over t with $t = 0$. This gives

$$\mathbb{P}\left(R(\hat{H}) \ge f_0\left(M_n + \frac{36}{n}\ln\frac{2n(n+3)}{\delta} + 2\sqrt{\frac{M_n \ln\frac{2n(n+3)}{\delta}}{n}}\right)\right) \le \delta. \quad (9)$$

For $n \to \infty$, (9) shows that $R(\hat{H})$ is bounded with high probability by

$$M_n + O\left(\frac{\ln^2 n}{n} + \sqrt{\frac{M_n \ln n}{n}}\right).$$

If the empirical cumulative loss nM_n is small (say, $M_n \le c/n$, where c is constant with n), then the above bound has rate $O((\ln^2 n)/n)$. In this case, the average or random hypothesis of inequality (2) achieves the sharper bound $O(1/n)$.

Proof [Theorem 4]: Let

$$\mu_{t,n} = \frac{1}{n-t}\sum_{i=t}^{n-1} R(H_i).$$

Applying Lemma 3 with $c_{\delta/2}$ we obtain

$$\mathbb{P}\left(R(\hat{H}) > \min_{0 \le t < n}(R(H_t) + c_{\delta/2}(R(H_t), t))\right) \le \frac{\delta}{2}. \quad (10)$$

We then observe that

$$\begin{aligned}
&\min_{0 \le t < n}(R(H_t) + c_{\delta/2}(R(H_t), t)) \\
&\quad = \min_{0 \le t < n}\ \min_{t \le i < n}(R(H_i) + c_{\delta/2}(R(H_i), i)) \\
&\quad \le \min_{0 \le t < n} \frac{1}{n-t}\sum_{i=t}^{n-1}(R(H_i) + c_{\delta/2}(R(H_i), i)) \\
&\quad \le \min_{0 \le t < n}\left(\mu_{t,n} + \frac{1}{n-t}\sum_{i=t}^{n-1}\frac{8}{3}\frac{C}{n-i} + \frac{1}{n-t}\sum_{i=t}^{n-1}\left(\sqrt{\frac{2CR(H_i)}{n-i}} + \frac{C}{n-i}\right)\right) \\
&\qquad \left(\text{using the inequality } \sqrt{x+y} \le \sqrt{x} + \tfrac{y}{2\sqrt{x}}\right) \\
&\quad = \min_{0 \le t < n}\left(\mu_{t,n} + \frac{1}{n-t}\sum_{i=t}^{n-1}\frac{11}{3}\frac{C}{n-i} + \frac{1}{n-t}\sum_{i=t}^{n-1}\sqrt{\frac{2CR(H_i)}{n-i}}\right)
\end{aligned}$$

$$\leq \min_{0\leq t<n}\left(\mu_{t,n}+\frac{11C}{3}\frac{\ln(n-t)+1}{n-t}+2\sqrt{\frac{2C\mu_{t,n}}{n-t}}\right)$$

(using $\sum_{i=1}^{k} 1/i \leq 1+\ln k$ and the concavity of the square root)

$$= \min_{0\leq t<n} f_t(\mu_{t,n}).$$

Now, it is clear that Proposition 2 can be immediately generalized to imply the following set of inequalities, one for each $t = 0,\dots,n-1$,

$$\mathbb{P}\left(\mu_{t,n}\geq M_{t,n}+\frac{36A}{n-t}+2\sqrt{\frac{M_{t,n}A}{n-t}}\right)\leq \frac{\delta}{2n} \tag{11}$$

where $A = \ln(2n(n+3))/(\delta)$.

Introduce the random variables $K_0,\dots,K_{n-1}$ to be defined later. We can write

$$\mathbb{P}\left(\min_{0\leq t<n}(R(H_t)+c_{\delta/2}(R(H_t),t))\geq \min_{0\leq t<n} K_t\right)$$
$$\leq \mathbb{P}\left(\min_{0\leq t<n} f_t(\mu_{t,n})\geq \min_{0\leq t<n} K_t\right)$$
$$\leq \sum_{t=0}^{n-1}\mathbb{P}(f_t(\mu_{t,n})\geq K_t).$$

Now, for each $t = 0,\dots,n-1$, define

$$K_t = f_t\left(M_{t,n}+\frac{36A}{n-t}+2\sqrt{\frac{M_{t,n}A}{n-t}}\right).$$

Then (11) and the monotonicity of $f_0,\dots,f_{n-1}$ allow us to obtain

$$\mathbb{P}\left(\min_{0\leq t<n}(R(H_t)+c_{\delta/2}(R(H_t),t))\geq \min_{0\leq t<n} K_t\right)$$
$$\leq \sum_{t=0}^{n-1}\mathbb{P}\left(f_t(\mu_{t,n})\geq f_t\left(M_{t,n}+\tfrac{36A}{n-t}+2\sqrt{\tfrac{M_{t,n}A}{n-t}}\right)\right)$$
$$= \sum_{t=0}^{n-1}\mathbb{P}\left(\mu_{t,n}\geq M_{t,n}+\frac{36A}{n-t}+2\sqrt{\frac{M_{t,n}A}{n-t}}\right)\leq \delta/2.$$

Combining with (10) concludes the proof. □

IV. Discussion, Conclusions, and Open Problems

In this correspondence, we have shown tail risk bounds for specific hypotheses selected from the ensemble generated by training incrementally an arbitrary learning algorithm. Proposition 2, our simplest bound, is proven via an easy application of Bernstein's inequality for martingales, a quite basic result in probability theory. The analysis of Theorem 4 is also centered on the same martingale inequality.

Our technique of deriving data-dependent risk tail bounds is based on a penalized risk estimate to control the variance of the risk of the selected ensemble hypothesis. As discussed in Section III, this hypothesis could in principle be very different from the *last* hypothesis H_n in the ensemble. Indeed, the variance of $R(H_n)$ could be high, since we make no assumptions on the behavior of the learning algorithm. On the other hand, if one is just interested in the *expectation* of the risk w.r.t. the training sample draw, then well-known results in stochastic approximation provide bounds on the rate of convergence of $\mathbb{E}[R(H_n)]$ to 0 as $n \to \infty$ (see, e.g., [12] for recent work on this subject).

As an open problem, we would like to simplify the analysis in Section III, possibly obtaining a more readable bound. Also, the bound shown in Theorem 4 contains $\ln n$ terms. We do not know whether these logarithmic terms can be improved to $\ln(M_n n)$, similarly to Proposition 2. A further open problem is to prove lower bounds, even in the special case when nM_n is bounded by a constant.

Acknowledgment

The authors would like to thank the anonymous reviewers for their comments that helped improve the presentation of this paper.

References

[1] P. Bartlett, "The sample complexity of pattern classification with neural networks," *IEEE Trans. Inf. Theory*, vol. 44, no. 2, pp. 525–536, 1998.

[2] P. Bartlett and S. Mendelson, "Rademacher and Gaussian complexities: Risk bounds and structural results," *J. Machine Learning Res.*, vol. 3, pp. 463–482, 2002.

[3] A. Blum, A. Kalai, and J. Langford, "Beating the hold-out: Bounds for k-fold and progressive cross-validation," in *Proc. 12th Ann. Conf. Comput. Learning Theory*, 1999, pp. 203–208.

[4] N. Cesa-Bianchi, A. Conconi, and C. Gentile, "On the generalization ability of on-line learning algorithms," *IEEE Trans. Inf. Theory*, vol. 50, no. 9, pp. 2050–2057, 2004.

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

[6] D. A. Freedman, "On tail probabilities for martingales," *Ann. Prob.*, vol. 3, pp. 100–118, 1975.

[7] V. Koltchinskii and D. Panchenko, "Empirical margin distributions and bounding the generalization error of combined classifiers," *Ann. Stat.*, vol. 30, no. 1, pp. 1–50, 2002.

[8] N. Littlestone, "From on-line to batch learning," in *Proc. 2nd Ann. Workshop Comput. Learning Theory*, 1989, pp. 269–284, Morgan Kaufmann.

[9] J. Shawe-Taylor, P. L. Bartlett, R. C. Williamson, and M. Anthony, "Structural risk minimization over data-dependent hierarchies," *IEEE Trans. Inf. Theory*, vol. 44, no. 5, pp. 1926–1940, 1998.

[10] V. Vapnik and A. Chervonenkis, "On the uniform convergence of relative frequencies of events to their probabilities," *Theory Prob. Its Appl.*, vol. 16, no. 2, pp. 264–280, 1971.

[11] V. Vapnik, *The Nature of Statistical Learning Theory*, 2nd ed. New York: Springer Verlag, 1999.

[12] Y. Ying and D. X. Zhou, "Online regularized classification algorithms," *IEEE Trans. Inf. Theory*, vol. 52, no. 11, pp. 4775–4788, 2006.

[13] T. Zhang, "Data dependent concentration bounds for sequential prediction algorithms," in *Proc. 18th Ann. Conf. Learning Theory*, 2005, pp. 173–187.