*The Annals of Applied Probability*
2003, Vol. 13, No. 1, 213–252

# BOUNDING THE GENERALIZATION ERROR OF CONVEX COMBINATIONS OF CLASSIFIERS: BALANCING THE DIMENSIONALITY AND THE MARGINS

BY VLADIMIR KOLTCHINSKII,[1] DMITRIY PANCHENKO[2] AND FERNANDO LOZANO

*University of New Mexico*

A problem of bounding the generalization error of a classifier $f \in \text{conv}(\mathcal{H})$, where $\mathcal{H}$ is a "base" class of functions (classifiers), is considered. This problem frequently occurs in computer learning, where efficient algorithms that combine simple classifiers into a complex one (such as boosting and bagging) have attracted a lot of attention. Using Talagrand's concentration inequalities for empirical processes, we obtain new sharper bounds on the generalization error of combined classifiers that take into account both the empirical distribution of "classification margins" and an "approximate dimension" of the classifiers, and study the performance of these bounds in several experiments with learning algorithms.

**1. Introduction.** Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be a sample of $n$ labeled training examples that are independent identically distributed copies of a random couple $(X, Y)$, $X$ being an "instance" in a measurable space $S$ and $Y$ being a "label" taking values in $\{-1, 1\}$. Let $P$ denote the distribution of the couple $(X, Y)$. Given a measurable function $f$ from $S$ into $\mathbb{R}$, we use $\text{sign}(f(x))$ as a predictor of the unknown label of an instance $x \in S$. We will call $f$ a classifier of the examples from $S$. The quantity $\mathbb{P}\{Yf(X) \le 0\} = P\{(x, y) : yf(x) \le 0\}$ is called *the generalization error* of the classifier $f$. The goal of learning (classification) is, given a set of training examples, to find a classifier $f$ with a small generalization error.

Some of the important recent advances in statistical learning theory are related to the development of complex classifiers that are combinations of simpler ones. In so-called *voting methods* of combining classifiers (such as boosting, bagging, etc.) a complex classifier produced by a learning algorithm is a convex combination of simpler classifiers from the base class.

Received October 2000; revised May 2002.

[1]Supported in part by NSA Grant MDA904-99-1-0031.

[2]Supported in part by UNM Office of Graduate Studies RPT Grant.

*AMS 2000 subject classifications.* Primary 62G05; secondary 62G20, 60F15.

*Key words and phrases.* Generalization error, combined classifier, margin, approximate dimension, empirical process, Rademacher process, random entropies, concentration inequalities, boosting, bagging.

Let $\mathcal{H}$ be a class of functions from $S$ into $\mathbb{R}$ (base classifiers) and let $\mathcal{F} := \operatorname{conv}(\mathcal{H})$ denote the symmetric convex hull of $\mathcal{H}$:

$$\operatorname{conv}(\mathcal{H}) := \left\{ \sum_{i=1}^{N} \lambda_i h_i : N \geq 1,\ \lambda_i \in \mathbb{R}, \sum_{i=1}^{N} |\lambda_i| \leq 1,\ h_i \in \mathcal{H} \right\}.$$

Our main goal in this paper is to develop new probabilistic upper bounds on the generalization error of a classifier $f$ from the symmetric convex hull $\mathcal{F} = \operatorname{conv}(\mathcal{H})$ of the base class. The well-known approach to such a problem, developed in the pathbreaking works of Vapnik and Chervonenkis (see [27] and references therein), is based on an easy bound,

$$P\{(x, y) : yf(x) \leq 0\} \leq P_n\{(x, y) : yf(x) \leq 0\} + \sup_{C \in \mathcal{C}} [P(C) - P_n(C)],$$

where $P_n$ is the empirical distribution of the training examples; that is, for any set $C \subset S \times \{-1, 1\}$, $P_n(C)$ is the frequency of training examples in the set $C$,

$$\mathcal{C} := \{\{(x, y) : yf(x) \leq 0\} : f \in \mathcal{F}\},$$

and on further bounding of the uniform (over the class $\mathcal{C}$) deviation of the empirical distribution $P_n$ from the true distribution $P$. The methods that are used to solve this problem belong to the theory of empirical processes and the crucial role is played by the VC-dimension of the class $\mathcal{C}$ or by more sophisticated entropy characteristics of the class. For instance, if $m^{\mathcal{C}}(n)$ denotes the maximal number of subsets obtainable by intersecting a sample of size $n$ with the class $\mathcal{C}$ (the so-called shattering number), then the following bound holds (see [8], Theorem 12.6) for all $\varepsilon > 0$:

$$\mathbb{P}\{P\{(x, y) : yf(x) \leq 0\} \geq P_n\{(x, y) : yf(x) \leq 0\} + \varepsilon\} \leq 8m^{\mathcal{C}}(n)e^{-n\varepsilon^2/32}.$$

It follows from this bound and from Sauer's lemma (see [8], Theorem 13.2) that the training error measures the generalization error of a classifier $f \in \mathcal{F}$ with accuracy $O(\sqrt{(V(\mathcal{C}) \log n)/n})$, where $V(\mathcal{C})$ is the VC-dimension of the class $\mathcal{C}$ [i.e., the smallest $n$ such that $m^{\mathcal{C}}(n) < 2^n$]. In the so-called zero-error case, when there exists a classifier $\hat{f} \in \mathcal{F}$ with zero training error, we even have the bound (see [8], Theorem 12.7)

$$\mathbb{P}\{P\{(x, y) : y\hat{f}(x) \leq 0\} \geq \varepsilon\} \leq 2m^{\mathcal{C}}(2n)2^{-n\varepsilon/2},$$

which implies that the generalization error of the classifier $\hat{f}$ is $O((V(\mathcal{C}) \log n)/n)$. The above bounds, however, do not apply directly to the case of the class $\mathcal{F} = \operatorname{conv}(\mathcal{H})$, which is of interest in applications to bounding the generalization error of the voting methods, since in this case typically $V(\mathcal{C}) = +\infty$. Even when one deals with a finite number of base classifiers in a convex combination (which is the case, say, with boosting after a finite number of rounds), the VC-dimensions of the classes involved are becoming rather large, so the above bounds do not explain the

generalization ability of boosting and other voting methods observed in numerous experiments. This motivated Bartlett [2] and Schapire, Freund, Bartlett and Lee [24] (see also [1]) to develop a new class of upper bounds on the generalization error of a convex combination of classifiers, expressed in terms of the empirical distribution of margins (the role of classification margins in improving the generalization ability of learning machines was clear in earlier work on support vector machines as well; see [7]). The margin of a classifier $f$ on a training example $(X, Y)$ is defined as the product $Yf(X)$. Let $\mathcal{H}$ be a "base class" of measurable functions from $S$ into $\{-1, 1\}$. Suppose that the class of sets $\mathcal{C} := \{\{x : h(x) = +1\} : h \in \mathcal{H}\}$ is Vapnik–Chervonenkis [i.e., $V(\mathcal{C}) < +\infty$] and let $V(\mathcal{H}) := V(\mathcal{C})$. Schapire et al. [24] showed that for a given $\alpha \in (0, 1)$ with probability at least $1 - \alpha$ for all $f \in \operatorname{conv}(\mathcal{H})$,

$$P\{(x, y) : yf(x) \le 0\} \le \inf_{\delta} \Bigg[ P_n\{(x, y) : yf(x) \le \delta\} + \frac{C}{\sqrt{n}} \left( \frac{V(\mathcal{H}) \log^2(n/(V(\mathcal{H})))}{\delta^2} + \log\left(\frac{1}{\alpha}\right) \right)^{1/2} \Bigg].$$

Choosing in the above bound the value of $\delta = \hat{\delta}(f)$ that solves the equation

$$\delta P_n\{(x, y) : yf(x) \le \delta\} = \sqrt{\frac{V(\mathcal{H})}{n}}$$

(which is nearly an optimal choice), one gets (ignoring the logarithmic factors) the generalization error of a classifier $f$ from the convex hull

$$O\left( \frac{1}{\hat{\delta}(f)} \sqrt{\frac{V(\mathcal{H})}{n}} \right).$$

Koltchinskii and Panchenko [17], using the methods of the theory of empirical, Gaussian and Rademacher processes (concentration inequalities, symmetrization, comparison inequalities), generalized and refined these types of bounds. They also suggested a way to improve these bounds under certain assumptions on the growth of random entropies of a class $\mathcal{F}$ to which the classifier belongs. The new bounds are based on the notion of the $\gamma$-margin of the classifier, introduced in their paper. The $\gamma$-margins are defined for $\gamma \in (0, 1)$ (see the definitions in Section 2 below); the value of $\gamma = 1$ roughly corresponds to the case studied in [24]. The quality of the bound improves as $\gamma$ decreases to 0. However, bounds of this type are proved to hold for values of $\gamma \ge 2\alpha/(2 + \alpha)$, where $\alpha \in (0, 2)$ is the growth exponent of the random entropy of the class $\mathcal{F}$. In the case $\mathcal{F} := \operatorname{conv}(\mathcal{H})$, where $\mathcal{H}$ is a VC-class with VC-dimension $V(\mathcal{H})$, this leads to the values of $\alpha = 2(V(\mathcal{H}) - 1)/V(\mathcal{H}) < 2$, which allow one to use $\gamma$-margins with $\gamma < 1$ (but it is going to be rather close to 1 unless the VC-dimension is very small).

The experiments of Koltchinskii, Panchenko and Lozano [18] showed that, in the case of the classifiers obtained in consecutive rounds of boosting, the bounds on the generalization error in terms of $\gamma$-margins hold even for much smaller values of $\gamma$. This allows one to conjecture that such classifiers belong, in fact, to a class $\mathcal{F} \subset \operatorname{conv}(\mathcal{H})$ whose entropy might be much smaller than the entropy of the whole convex hull. The problem, though, is that it is practically impossible to identify such a class prior to experiments, leaving the question of how to choose the values of $\gamma$ for which the bounds hold open. In this paper, we develop a new approach to this problem. Namely, we suggest an adaptive bound on the generalization error of a convex combination of classifiers from a base class that is based, on the one hand, on the margins of the combined classifiers and, on the other hand, on their *approximate dimensions* (the numbers of "large enough" coefficients in the convex combinations). This adaptive bound "captures" the size of the entropy of a subset of the convex hull to which the classifier actually belongs.

The results are formulated precisely in Section 2. The proofs that heavily rely upon Talagrand's concentration and deviation inequalities for empirical processes are given in Section 3. Section 4 includes the results of several experiments with existing learning algorithms (such as boosting and bagging) for which we computed the bounds on the learning curves that follow from our results. We also discuss here some approaches to combining classifiers that attempt to minimize the margin cost function, keeping the dimension of the classifier small.

**2. Empirical margins and approximate dimensions: Main results.** Let $(S, \mathcal{A})$ be a measurable space and let $\mathcal{F}$ be a class of real-valued measurable functions on $(S, \mathcal{A})$. In this section, to shorten the notation, we suppress the labels. To apply the results in the setting of the Introduction, instead of $S$, consider the space $S \times \{-1, 1\}$ and instead of a function $f$ on $S$, consider a function $(x, y) \mapsto yf(x)$ on $S \times \{-1, 1\}$. The results also can be used in the case of multiclass problems (see Section 5 in [17]). In what follows $P$ denotes a probability measure on $(S, \mathcal{A})$, $\{X_n\}$ is a sequence of i.i.d. random variables defined on a probability space $(\Omega, \Sigma, \mathbb{P})$ and taking values in $(S, \mathcal{A})$ with distribution $P$ and $P_n$ denotes the empirical measure based on the sample $(X_1, \dots, X_n)$,

$$P_n(A) := n^{-1} \sum_{i=1}^{n} I_A(X_i), \qquad A \subset S,$$

$I_A$ being the characteristic function (the indicator) of the set $A$.

In what follows, we frequently use metric entropies to measure the complexity of function classes involved in our bounds. Given a metric space $(T, d)$, $H_d(T; \varepsilon)$ denotes the $\varepsilon$-entropy of $T$ with respect to $d$, that is,

$$H_d(T; \varepsilon) := \log N_d(T; \varepsilon),$$

where $N_d(T;\varepsilon)$ is the minimal number of balls of radius $\varepsilon$ covering $T$. If $Q$ is a probability measure on $(S;\mathcal{A})$, $d_{Q,2}$ will denote the metric of the space $L_2(S;dQ)$,

$$d_{Q,2}(f;g) := (Q|f-g|^2)^{1/2}.$$

We start by extending the bounds on the generalization error obtained by Koltchinskii and Panchenko [17] in terms of so-called $\gamma$-margins. More precisely, let $\gamma \in (0,1]$ and let

$$\varepsilon_{n,\gamma}(\delta) := \frac{1}{n^{1-\gamma/2}\delta^{\gamma}}.$$

It was shown in [17] (see also Example 1 below) that if $H_{d_{P_n,2}}(\mathcal{F};\varepsilon)$ grows as $\varepsilon^{-\alpha}$ for some $\alpha \in (0,2)$, then, with high probability for all $f \in \mathcal{F}$ and all $\delta > 0$,

$$P\{f \le 0\} \le C[P_n\{f \le \delta\} \vee \varepsilon_{n,\gamma}(\delta)], \tag{2.1}$$

where $\gamma = \frac{2\alpha}{\alpha+2}$ and $C$ is a constant. The expression in the brackets is the maximum of two functions of the margin $\delta$. The first one is the empirical distribution function of margins, $P_n\{f \le \delta\}$. It is a nondecreasing function of $\delta$. The second one, $\varepsilon_{n,\gamma}(\delta)$, is a decreasing function of $\delta$ and it depends on the parameter $\gamma$ related to the complexity of the class $\mathcal{F}$ (the growth exponent $\alpha$ of its entropy). The value of the margin $\delta$ that minimizes the expression in brackets (equivalently, the value for which the two functions in the maximum are equal) was called in [17] the $\gamma$-margin of a classifier $f$, $\hat{\delta}_n(\gamma; f)$. Clearly, $C\varepsilon_{n,\gamma}(\hat{\delta}_n(\gamma; f))$ is in this case an upper bound on the generalization error of $f$ (with high probability). The bound in [24] corresponded to the choice of $\gamma = 1$, but it is easily seen that smaller values of $\gamma$ provide sharper bounds.

Below we give a definition of what we call $\psi$-bounds that will play a major role in bounding the generalization error of classifiers. These quantities depend on a function $\psi$ that characterizes the complexity of the class $\mathcal{F}$ (more specifically, it will provide an upper bound on the so-called Dudley entropy integral for the class $\mathcal{F}$) and, therefore, determines the quality of the bounds.

Let $\psi$ be a concave nondecreasing function on $[0,+\infty)$ with $\psi(0) = 0$. For a fixed $\delta > 0$, denote by $\varepsilon_n^{\psi}(\delta)$ the smallest solution of the equation

$$\varepsilon = \frac{1}{\delta\sqrt{n}}\psi\bigl(\delta\sqrt{\varepsilon}\bigr) \tag{2.2}$$

with respect to $\varepsilon$. Similarly, for a fixed $\varepsilon > 0$, denote by $\delta_n^{\psi}(\varepsilon)$ the largest solution of (2.2) with respect to $\delta$ [if $\psi$ is strictly concave, the solutions of (2.2) are unique]. Clearly, for a concave $\psi$ the function $\varphi(x) \equiv \frac{\psi(x)}{x}$ is nonincreasing. Therefore, it is easy to see that

$$\delta_n^{\psi}(\varepsilon) = \frac{\varphi^{-1}(\sqrt{\varepsilon n})}{\sqrt{\varepsilon}}.$$

If $\psi(x) = x^{1-\alpha/2}$ with $\alpha \in (0,2)$ [which will be our choice of $\psi$ when the $L_2(P_n)$-entropy of the class $\mathcal{F}$ is on the order $\varepsilon^{-\alpha}$; see Example 1], then $\varepsilon_n^{\psi} = \varepsilon_{n,\gamma}$ (with $\gamma = \frac{2\alpha}{\alpha+2}$). Actually, we show below that for a general function $\psi$ (that bounds Dudley's entropy integral of the class $\mathcal{F}$ from above), the generalization error of all the classifiers $f \in \mathcal{F}$ is bounded from above by

$$C \inf_{\delta>0}[P_n\{f \le \delta\} \vee \varepsilon_n^{\psi}(\delta)] \tag{2.3}$$

with high probability. The $\psi$-bounds introduced below give the size of the infimum in expression (2.3).

Given a function $f$ and $t > 0$, define the quantity

$$\varepsilon_n^{\psi}(f;t) := \inf\left\{\varepsilon \ge \frac{t \vee 2\log n}{n} : P\{f \le \delta_n^{\psi}(\varepsilon)\} \le \varepsilon\right\}$$

and its empirical version

$$\hat{\varepsilon}_n^{\psi}(f;t) := \inf\left\{\varepsilon \ge \frac{t \vee 2\log n}{n} : P_n\{f \le \delta_n^{\psi}(\varepsilon)\} \le \varepsilon\right\}.$$

Since for all $\varepsilon > 0$, $\delta_n^{\psi}(\varepsilon) \ge 0$, it immediately follows from the definition that for all $f \in \mathcal{F}$,

$$P\{f \le 0\} \le \inf\{P\{f \le \delta_n^{\psi}(\varepsilon)\} : \varepsilon \ge \varepsilon_n^{\psi}(f;t)\} \le \varepsilon_n^{\psi}(f;t).$$

We will call $\varepsilon_n^{\psi}(f;t)$ and $\hat{\varepsilon}_n^{\psi}(f;t)$ *the $\psi$-bound* and *the empirical $\psi$-bound* of the classifier $f$, respectively. We show below that under a proper assumption on the random entropy of the class $\mathcal{F}$, with a high probability the empirical $\psi$-bounds $\hat{\varepsilon}_n^{\psi}(f;t)$ are, for all the functions from the class, within a multiplicative constant from the true $\psi$-bounds $\varepsilon_n^{\psi}(f;t)$. This allows one to replace $\varepsilon_n^{\psi}(f;t)$ in the above bound on $P\{f \le 0\}$ by $\hat{\varepsilon}_n^{\psi}(f;t)$ (which gives in applications a bound on the generalization errors of classifiers).

THEOREM 1. *Let $\psi$ be a concave nondecreasing function on $[0,+\infty)$ with $\psi(0) = 0$. Suppose the bound on Dudley's entropy integral holds with some $D_n > 0$,*

$$\int_0^x H_{d_{P_n,2}}^{1/2}(\mathcal{F},u)\,du \le D_n\psi(x), \qquad x > 0 \text{ a.s.}, \tag{2.4}$$

*where $D_n = D_n(X_1,\dots,X_n)$ is a function of training examples such that $\mathbb{E}D_n < \infty$. Then there exist absolute constants $A, B > 0$ such that for $\bar{A} := A(1+\mathbb{E}D_n)^2$ and for all $t > 0$,*

$$\begin{aligned}&\mathbb{P}\{\forall f \in \mathcal{F} : \bar{A}^{-1}\hat{\varepsilon}_n^{\psi}(f;t) \le \varepsilon_n^{\psi}(f;t) \le \bar{A}\hat{\varepsilon}_n^{\psi}(f;t)\}\\ &\qquad \ge 1 - B\log_2\log_2\frac{n}{t\vee 2\log n}\exp\left\{-\left(\frac{t}{2}\vee\log n\right)\right\}.\end{aligned} \tag{2.5}$$

The following corollary is immediate.

COROLLARY 1. *Under the conditions of Theorem* 1 *there exist numerical constants* $A, B > 0$ *such that for* $\bar{A} := A(1 + \mathbb{E}D_n)^2$ *and for all* $t > 0$,

$$\mathbb{P}\{\exists\, f \in \mathcal{F} : P\{f \le 0\} \ge \bar{A}\hat{\varepsilon}_n^{\psi}(f; t)\}$$
$$\le B \log_2 \log_2 \frac{n}{t \vee 2\log n} \exp\left\{-\left(\frac{t}{2} \vee \log n\right)\right\}. \tag{2.6}$$

REMARK. Because of the presence of the multiplicative constant in front of the $\psi$-bound, the last result seems to be useful only in the case of small Bayes risk (which is not unusual in modern classification problems where rather complex combinations of base classifiers are often being used). However, Koltchinskii [14] showed that if Dudley's entropy integral is $o(\psi(x))$ as $x \to 0$, then the constant in the bounds of this type becomes asymptotically close to 1 as $n \to \infty$. Thus, the above bounds might be useful even in the case of larger values of the Bayes risk.

EXAMPLE 1. Let $\alpha \in (0, 2)$ and $\psi(x) \equiv x^{1-\alpha/2}$. Let $\gamma := \frac{2\alpha}{\alpha+2}$. Koltchinskii and Panchenko [17] defined $\gamma$-margins of a function $f$ as

$$\delta_n(\gamma; f) := \sup\{\delta \in (0, 1) : \delta^{\gamma} P\{f \le \delta\} \le n^{-1+\gamma/2}\},$$
$$\hat{\delta}_n(\gamma; f) := \sup\{\delta \in (0, 1) : \delta^{\gamma} P_n\{f \le \delta\} \le n^{-1+\gamma/2}\}.$$

An easy computation shows that

$$\varepsilon_n^{\psi}(f; n^{\gamma/2}) = \frac{1}{n^{1-\gamma/2}\delta_n(\gamma; f)^{\gamma}}.$$

Corollary 1 immediately implies that if for some $\alpha \in (0, 2)$ and $D_n > 0, \mathbb{E}D_n < \infty$ and

$$H_{d_{P_n,2}}(\mathcal{F}; u) \le D_n^2 u^{-\alpha}, \qquad u > 0 \text{ a.s.},$$

then for any $\gamma \ge \frac{2\alpha}{\alpha+2}$ there exist constants $A, B > 0$ such that for $\bar{A} := A(1 + \mathbb{E}D_n)^2$,

$$\mathbb{P}\left\{\exists\, f \in \mathcal{F} : P\{f \le 0\} \ge \frac{\bar{A}}{n^{1-\gamma/2}\hat{\delta}_n(\gamma; f)^{\gamma}}\right\}$$
$$\le B \log_2 \log_2 n \exp\left\{\frac{-n^{\gamma/2}}{2}\right\} \tag{2.7}$$

(see also [17]). It is easy to see that the quantity

$$\frac{1}{n^{1-\gamma/2}\hat{\delta}_n(\gamma;\, f)^{\gamma}} \tag{2.8}$$

in the above upper bound on the generalization error *becomes smaller* as $\gamma$ decreases from 1 to 0. The Schapire–Freund–Bartlett–Lee type of bounds correspond to the worst choice of $\gamma$ ($\gamma = 1$). In the case when $\mathcal{F}$ is the symmetric convex hull of a VC-class $\mathcal{H}$ with VC-dimension $V(\mathcal{H})$, the value of $\alpha$ is equal to $\frac{2(V(\mathcal{H})-1))}{V(\mathcal{H})} < 2$ that allows us to have $\gamma < 1$, improving the previously known bound. In fact, Koltchinskii, Panchenko and Lozano [18] computed the empirical $\gamma$-margins of classifiers obtained in consecutive rounds of boosting and observed that the bounds on their generalization error in terms of $\gamma$-margins hold even for much smaller values of $\gamma$. This allows one to conjecture that such classifiers belong, in fact, to a class $\mathcal{F} \subset \operatorname{conv}(\mathcal{H})$ whose entropy might be much smaller than the entropy of the whole convex hull.

EXAMPLE 2. Consider now the case of $\psi(x) \equiv x\sqrt{\log \frac{e}{x}}$ for $x \le 1$ and $\psi(x) \equiv x$ for $x > 1$. Then, by a simple computation,

$$\delta_n^{\psi}(\varepsilon) = \frac{e^{1-n\varepsilon}}{\sqrt{\varepsilon}}, \qquad \varepsilon \ge n^{-1}.$$

If we define

$$\hat{\varepsilon}_n^{\mathrm{VC}}(f;t) := \inf\left\{\varepsilon \ge \frac{t \vee 2\log n}{n} : P_n\left\{f \le \frac{e^{1-n\varepsilon}}{\sqrt{\varepsilon}}\right\} \le \varepsilon\right\}, \tag{2.9}$$

then under the condition

$$H_{d_{P_n,2}}(\mathcal{F};u) \le D_n^2 \log\frac{1}{u} \vee 1, \qquad u > 0 \text{ a.s.},$$

with some $D_n = D_n(X_1,\dots,X_n)$, $\mathbb{E}D_n < +\infty$ (which holds, e.g., if $\mathcal{F}$ is a VC-subgraph class, that is, the class of sets

$$\bigl\{\{(x,t) \in S \times \mathbb{R} : t < f(x)\} : f \in \mathcal{H}\bigr\}$$

is Vapnik–Chervonenkis), we get from Corollary 1 that with some numerical constants $A$, $B > 0$ for all $t > 0$,

$$\begin{aligned}&\mathbb{P}\bigl\{\exists\, f \in \mathcal{F} : P\{f \le 0\} \ge \bar{A}\hat{\varepsilon}_n^{\mathrm{VC}}(f;t)\bigr\}\\ &\qquad \le B\log_2\log_2\frac{n}{t \vee 2\log n}\exp\left\{-\left(\frac{t}{2} \vee \log n\right)\right\},\end{aligned}$$

where $\bar{A} := A(1 + \mathbb{E}D_n)^2$.

The proofs of Theorems 1 and 3 are based on the following generalization of one of the results of Koltchinskii and Panchenko [17] (that itself relies heavily on the concentration inequality for empirical processes due to Talagrand; see Theorem 1.4 of [25]).

THEOREM 2. *Suppose that condition* (2.4) *holds with some concave nondecreasing* $\psi$ *such that* $\psi(0) = 0$. *Then, for all* $\delta > 0$ *and for all* $\varepsilon \geq \varepsilon_n^{\psi}(\delta) \vee \frac{2\log n}{n}$ *the bounds*

$$\mathbb{P}\left\{\exists\, f \in \mathcal{F} : P_n\{f \leq \delta\} \leq \varepsilon \text{ and } P\left\{f \leq \frac{\delta}{2}\right\} \geq \bar{A}\varepsilon\right\} \leq B \log_2 \log_2 \varepsilon^{-1} \exp\left\{-\frac{n\varepsilon}{2}\right\}$$

*and*

$$\mathbb{P}\left\{\exists\, f \in \mathcal{F} : P\{f \leq \delta\} \leq \varepsilon \text{ and } P_n\left\{f \leq \frac{\delta}{2}\right\} \geq \bar{A}\varepsilon\right\} \leq B \log_2 \log_2 \varepsilon^{-1} \exp\left\{-\frac{n\varepsilon}{2}\right\}$$

*hold, where* $\bar{A} = A(1 + \mathbb{E}D_n)^2$ *and* $A, B$ *are numerical constants.*

There are two major problems with the margin type bounds given above. First of all, the values of the constants involved in the bounds are far from being optimal and are too large at the moment (see, however, the remark after Corollary 1). Their improvement is related to a hard problem of optimizing the constants in Talagrand's concentration inequalities for empirical and Rademacher processes used in the proofs below. However, in the case when $\mathcal{F} = \mathrm{conv}(\mathcal{H})$ the constants in question depend only on the base class $\mathcal{H}$ and this allows one to use the bounds to study the behavior of the generalization error when the number of rounds of learning algorithms (such as boosting) increases. Another problem is related to the fact that there is no much prior knowledge about the subset of $\mathrm{conv}(\mathcal{H})$ to which a classifier created by boosting or another method of combining classifiers is going to belong. This makes one tend use the value of

$$\gamma = \frac{2\alpha}{\alpha + 2} = \frac{2(V(\mathcal{H}) - 1)}{2V(\mathcal{H}) - 1}, \tag{2.10}$$

which is very close to 1 unless the VC-dimension of the base is *very* small. Our major goal in the current paper is to address this problem. We do this by proving a new upper bound on the generalization error of a classifier that belongs to a convex hull of a base class. The bound includes the sum of two main terms. The first one is an "approximate dimension" of the classifier (the number of "large enough" coefficients in the convex combination) divided by the sample size. The second

term is related to the margins of the classifier. Balancing these two terms allows us to get a rather tight upper bound that "captures" the size of the entropy of a class to which the classifier actually belongs. It combines previously known bounds in terms of VC-dimension (in zero-error case) and in terms of margins, and becomes close to one of these two bounds in the extreme cases.

Let $\mathcal{H}$ be a class of measurable functions from $(S, \mathcal{A})$ into $\mathbb{R}$. Let $\mathcal{F} \subset \operatorname{conv}(\mathcal{H})$. For a function $f \in \mathcal{F}$ and a number $\Delta \in [0,1]$, we define *the approximate $\Delta$-dimension* of $f$ as the smallest integer number $d \geq 0$ such that there exist $N \geq 1$, functions $h_j \in \mathcal{H}$, $j = 1, \dots, N$, and numbers $\lambda_j \in \mathbb{R}$, $j = 1, \dots, N$, satisfying the conditions $f = \sum_{j=1}^N \lambda_j h_j$, $\sum_{j=1}^N |\lambda_j| \leq 1$ and $\sum_{j=d+1}^N |\lambda_j| \leq \Delta$. The $\Delta$-dimension of $f$ will be denoted by $d(f; \Delta)$.

In what follows we assume that for some $V > 0$ and $K > 0$ and for all probability measures $Q$ on $(S; \mathcal{A})$,

$$N_{d_{Q,2}}\bigl(\mathcal{H}; (QH^2)^{1/2}\varepsilon\bigr) \leq K\varepsilon^{-V}, \qquad \varepsilon > 0, \tag{2.11}$$

where $H$ is a measurable envelope of $\mathcal{H}$ [i.e., a nonnegative measurable function such that, for all $h \in \mathcal{H}$ and $x \in S$, $|h(x)| \leq H(x)$]. In particular, this condition holds if $\mathcal{H}$ is a VC-subgraph class. Condition (2.11) implies the bound on the entropy of the convex hull of $\mathcal{H}$,

$$H_{d_{Q,2}}\bigl(\operatorname{conv}(\mathcal{H}); (QH^2)^{1/2}\varepsilon\bigr) \leq C\varepsilon^{-2V/(V+2)}, \qquad \varepsilon > 0,$$

with $V$ from the bound (2.11) and $C := C(K; V)$ (see [26]). One can easily compute in this case that

$$\int_0^x H^{1/2}_{d_{P_n,2}}(\mathcal{F}, u)\, du \leq \tfrac{1}{2}(V+2)C^{1/2}(P_nH^2)^{V/(2(V+2))}x^{2/V+2}, \qquad x > 0 \text{ a.s.}$$

and, therefore, condition (2.4) of Theorem 1 is satisfied with $\psi(x) = x^{2/(V+2)}$ under the assumption $PH^2 < \infty$. Subsequently we will assume that one of the following two conditions holds:

1. Class $\mathcal{H}$ is uniformly bounded and $\mathcal{F} \subset \operatorname{conv}(\mathcal{H})$.
2. The envelope $H$ of the class $\mathcal{H}$ is $P$-square integrable and

$$\mathcal{F} \subset \left\{ \sum_{i=1}^N \lambda_i h_i : N \geq 1,\ h_i \in \mathcal{H},\ \lambda_i \in \mathbb{R},\ \sum_{j=1}^N |\lambda_j| = 1 \right\}.$$

Note, that under the second condition, $\mathcal{F}$ consists only of proper symmetric convex combinations.

Let $\alpha := \frac{2V}{V+2}$ and $\Delta_f = \{\Delta \in [0,1] : d(f;\Delta) \le n\}$. Define

$$\varepsilon_n(f;\delta) := \inf_{\Delta\in\Delta_f}\left[\frac{d(f;\Delta)}{n}\left(\log\frac{1}{\delta} + \log\frac{ne^2}{d(f;\Delta)}\right) + \left(\frac{\Delta}{\delta}\right)^{2\alpha/(\alpha+2)} n^{-2/(\alpha+2)}\right] \vee \frac{2\log n}{n}. \tag{2.12}$$

The function of margin $\varepsilon_n(f;\delta)$ will play exactly the same role as $\varepsilon_{n,\gamma}$ in (2.1) or $\varepsilon_n^{\psi}$ in (2.3). As before, we will essentially show that with high probability for all $f \in \mathcal{F}$ the generalization error is bounded from above by

$$C \inf_{\delta>0}[P_n\{f \le \delta\} \vee \varepsilon_n(f;\delta)]. \tag{2.13}$$

However, this time the function $\varepsilon_n(f;\delta)$ depends not only on the complexity of the class $\mathcal{F}$, but also on the complexity of a particular classifier $f \in \mathcal{F}$ for which the generalization error is to be bounded (namely, on the sparsity of the weights of $f$ reflected in the definition of the approximate $\Delta$-dimension). The value of $\delta$ for which the infimum in (2.13) is attained can be evaluated as

$$\hat{\delta}_n(f) := \sup\{\delta \in (0,1/2) : P_n\{f \le \delta\} \le \varepsilon_n(f;\delta)\}$$

and the size of the infimum becomes $\varepsilon_n(f;\hat{\delta}_n(f))$. According to the next theorem, the quantity of this type provides an upper bound on the generalization error with high probability.

THEOREM 3. *Assume that one of the above conditions* 1 *or* 2 *on the class* $\mathcal{F}$ *holds. Then there exist constants* $A, B > 0$ *such that for all* $0 < t < n^{\alpha/(2+\alpha)}$, *the following bound holds*:

$$\mathbb{P}\left\{\exists f \in \mathcal{F} : P\left\{f \le \frac{\hat{\delta}_n(f)}{4}\right\} \ge A\left(\varepsilon_n\left(f;\frac{\hat{\delta}_n(f)}{2}\right) + \frac{t}{n}\right)\right\} \le Be^{-t/4}.$$

To understand this bound, let us look again at the definition of $\varepsilon_n(f;\delta)$. First of all, if one sets $\Delta = 1$ instead of minimizing over $\Delta$, then, since $d(f,1) = 0$, the bound becomes equivalent to the previous $\gamma$-bound of Example 1 (with $\gamma = \frac{2\alpha}{\alpha+2}$), which means that the bound of the Theorem 3 improves the $\gamma$-bound. For a fixed $\Delta$, the two terms in the definition of $\varepsilon_n(f;\delta)$ correspond to two parts of the combined classifier. The first term corresponds to the sum of $d(f,\Delta)$ base classifiers with the largest weights and the form of the bound basically coincides with the standard VC-dimension based bound in the zero-error case. The second term corresponds to an "improper" convex combination of classifers with the smallest weights (the number of them is not limited) and the form of the bound is determined by the complexity of the whole convex hull, only scaled by a factor of $\Delta$. It is clear that if a voting algorithm produces a convex combination in which

there are very few classifiers with large weights, then the bound of the theorem can improve upon the $\gamma$-bound of Example 1 significantly. Another way to say it is that the faster is the weight decay in the convex combination, the smaller is the complexity of the corresponding subset of the convex hull and the sharper is the bound. This is easily demonstrated by the following example.

EXAMPLE 3. If $\mathcal{F} \subset \operatorname{conv}(\mathcal{H})$ is a class of functions such that for some $\beta > 0$,

$$\sup_{f \in \mathcal{F}} d(f; \Delta) = O(\Delta^{-\beta}), \tag{2.14}$$

then with "high probability" for any classifier $f \in \mathcal{F}$ the upper bound on its generalization error becomes on the order of

$$\frac{1}{n^{1-\gamma\beta/2(\gamma+\beta)} \hat{\delta}_n(f)^{\gamma\beta/(\gamma+\beta)}}$$

(which, of course, improves a more general bound in terms of $\gamma$-margins; the general bound corresponds to the case $\beta = +\infty$). Condition (2.14) means that the weights of the convex combination decrease polynomially fast, namely, $|\lambda_j| = O(j^{-\alpha})$, $\alpha = 1 + \beta^{-1}$. The case of exponential decrease of the weights is described by the condition

$$\sup_{f \in \mathcal{F}} d(f; \Delta) = O\left(\log \frac{1}{\Delta}\right). \tag{2.15}$$

In this case the upper bound becomes on the order of $\frac{1}{n} \log^2(n/(\hat{\delta}_n(f)))$.

## 3. Proofs of the main results.

PROOF OF THEOREM 1. We use the first bound of Theorem 2. The condition $\varepsilon \geq \varepsilon_n^{\psi}(\delta)$ is equivalent to the condition $\delta \geq \delta_n^{\psi}(\varepsilon)$. Thus, we can use this bound for $\delta = \delta_n^{\psi}(\varepsilon)$ and $\varepsilon \geq (2 \log n)/n$. We get

$$\mathbb{P}\left\{\exists\, f \in \mathcal{F} : P_n\{f \leq \delta_n^{\psi}(\varepsilon)\} \leq \varepsilon \text{ and } P\left\{f \leq \frac{\delta_n^{\psi}(\varepsilon)}{2}\right\} \geq \bar{A}\varepsilon\right\}$$
$$\leq B \log_2 \log_2 \varepsilon^{-1} \exp\left\{-\frac{n\varepsilon}{2}\right\}.$$

Next we set $\varepsilon_j := 2^{-j}$. Let $\mathcal{J} = \{j \geq 0 : \varepsilon_j \geq \frac{t \vee 2 \log n}{n}\}$ and

$$E := \left\{\exists\, j \in \mathcal{J}\ \exists\, f \in \mathcal{F} : P_n\{f \leq \delta_n^{\psi}(\varepsilon_j)\} \leq \varepsilon_j \text{ and } P\left\{f \leq \frac{\delta_n^{\psi}(\varepsilon_j)}{2}\right\} \geq \bar{A}\varepsilon_j\right\}.$$

We have

$$\mathbb{P}(E) \le B \sum_{j \in \mathcal{J}} \log_2 \log_2 \varepsilon_j^{-1} \exp\left\{-\frac{n\varepsilon_j}{2}\right\}$$

$$\le B \log_2 \log_2 \frac{n}{t \vee 2\log n} \sum_{j \ge 0} \exp\left\{-\left(\frac{t}{2} \vee \log n\right) 2^j\right\} \tag{3.1}$$

$$\le B' \log_2 \log_2 \frac{n}{t \vee 2\log n} \exp\left\{-\left(\frac{t}{2} \vee \log n\right)\right\}.$$

Suppose that for some $j$ and for some $f \in \mathcal{F}$, $\hat{\varepsilon}_n^{\psi}(t; f) \in (\varepsilon_{j+1}, \varepsilon_j]$. On the event $E^c$, the inequality $P_n\{f \le \delta_n^{\psi}(\varepsilon_j)\} \le \varepsilon_j$ implies that $P\{f \le \delta_n^{\psi}(\varepsilon_j)/2\} \le \bar{A}\varepsilon_j$. Since

$$\frac{\delta_n^{\psi}(\varepsilon_j)}{2} = \frac{\varphi^{-1}(\sqrt{\varepsilon_j n})}{2\sqrt{\varepsilon_j}} \ge \frac{\varphi^{-1}(\sqrt{4\varepsilon_j n})}{\sqrt{4\varepsilon_j}} = \delta_n^{\psi}(4\varepsilon_j),$$

we also have $P\{f \le \delta_n^{\psi}(4\varepsilon_j)\} \le \bar{A}\varepsilon_j$, which implies $P\{f \le \delta_n^{\psi}(8\hat{\varepsilon}_n^{\psi}(f; t))\} \le 2\bar{A}\hat{\varepsilon}_n^{\psi}(f; t)$. Therefore, on the event $E^c$, we get for all $f \in \mathcal{F}$, $\varepsilon_n^{\psi}(f; t) \le (2\bar{A} \vee 8)\hat{\varepsilon}_n^{\psi}(f; t)$. It follows from (3.1) that

$$\mathbb{P}\{\exists f \in \mathcal{F} : \varepsilon_n^{\psi}(f; t) \ge (2\bar{A} \vee 8)\hat{\varepsilon}_n^{\psi}(f; t)\}$$

$$\le B' \log_2 \log_2 \frac{n}{t \vee 2\log n} \exp\left\{-\left(\frac{t}{2} \vee \log n\right)\right\}.$$

Quite similarly, using the second bound of Theorem 2, one can prove that

$$\mathbb{P}\{\exists f \in \mathcal{F} : \hat{\varepsilon}_n^{\psi}(f; t) \ge (2\bar{A} \vee 8)\varepsilon_n^{\psi}(f; t)\}$$

$$\le B' \log_2 \log_2 \frac{n}{t \vee 2\log n} \exp\left\{-\left(\frac{t}{2} \vee \log n\right)\right\},$$

which implies the inequality of Theorem 1. □

To prove Theorem 2, we follow the proof of Theorem 6 in [17], which is based on an iterative application of Talagrand's concentration inequality for empirical processes (Theorem 1.4 in [25]) which allows us in some sense to localize the classifier $f$ and to evaluate better its generalization error. To implement the iterative "localization," we will choose in a rather special way (for a fixed $\delta$) a finite decreasing sequence $\delta_j$, $1 \le j \le N$, such that $\delta_j < \delta$ and $\delta_N = \delta/2$. We consider the functions $\varphi_j$, $j \ge 1$, which are defined as shown in Figure 1 and which play the role of continuous (even, Lipschitz) approximations of the indicator step functions $I(x \le \delta_j)$ (in fact, we will use even two sequences of functions like this).

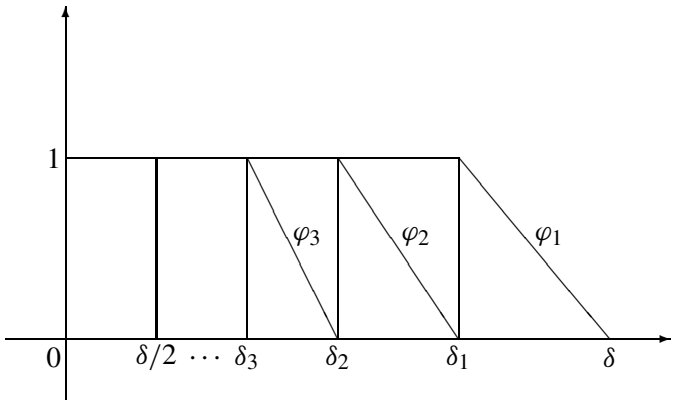


FIG. 1.

Assume that we have already bounded $P\{f \le \delta_{k-1}\}$ by a number $r_{k-1}$. The main idea is now to bound $P\{f \le \delta_k\}$ by $P(\varphi_k \circ f)$ and further by

$$P_n\{f \le \delta_{k-1}\} + \sup\{|(P_n - P)(\varphi_k \circ f)| : f \in \mathcal{F},\ P\{f \le \delta_{k-1}\} \le r_{k-1}\}.$$

Talagrand's concentration inequality allows one to replace the supremum of $|(P_n - P)(\varphi_k \circ f)|$ in the above bound by its expectation, which in turn can be bounded by Dudley's entropy integral of the class

$$\{\varphi_k \circ f : f \in \mathcal{F},\ P\{f \le \delta_{k-1}\} \le r_{k-1}\}.$$

Since $\varphi_k$ is a Lipschitz function, the $L_2(P_n)$-entropy of the above class can be easily bounded in terms of the $L_2(P_n)$-entropy of the class $\mathcal{F}$ itself. This leads to a bound $r_k$ on $P\{f \le \delta_k\}$. It happens that this procedure allows us to improve the above bounds iteratively and to end up with the bound claimed to be true in the theorem.

We now proceed to the proof.

PROOF OF THEOREM 2. Define

$$r_0 := 1, \qquad r_{k+1} = C\sqrt{r_k \varepsilon} \wedge 1,$$

where $C = c(1 + \mathbb{E}D_n)$ with a sufficiently large constant $c > 1$ (which will be chosen later). A simple induction shows that either $C\sqrt{\varepsilon} \ge 1$ and $r_k \equiv 1$, or $C\sqrt{\varepsilon} < 1$, and in the last case,

$$r_k = C^{1+2^{-1}+\cdots+2^{-(k-1)}} \varepsilon^{2^{-1}+\cdots+2^{-k}} = C^{2(1-2^{-k})} \varepsilon^{1-2^{-k}} = (C\sqrt{\varepsilon})^{2(1-2^{-k})}.$$

Let $\gamma_k := (\varepsilon/r_k)^{1/2} = C^{2^{-k}-1} \varepsilon^{2^{-k-1}}$. Then

$$\begin{aligned} \gamma_k + \gamma_{k-2} + \cdots + \gamma_0 &= C^{-1}\big[C\sqrt{\varepsilon} + (C\sqrt{\varepsilon})^{2^{-1}} + \cdots + (C\sqrt{\varepsilon})^{2^{-k}}\big] \\ &\le C^{-1}(C\sqrt{\varepsilon})^{2^{-k}}\big(1 - (C\sqrt{\varepsilon})^{2^{-k}}\big)^{-1} \le 1/2 \end{aligned} \tag{3.2}$$

for $\varepsilon \le C^{-4}$, $C > 2(2^{1/4}-1)^{-1}$ and $k \le \log_2 \log_2 \varepsilon^{-1}$ (note that $\varepsilon \le C^{-4}$ implies $C\sqrt{\varepsilon} < 1$). In what follows, we fix $\varepsilon > 0$ and use only the values of $k$ such that $k \le \log_2 \log_2 \varepsilon^{-1}$. Let $\delta > 0$. Define

$$\delta_0 = \delta, \quad \delta_k := \delta(1 - \gamma_0 - \cdots - \gamma_{k-1}), \quad \delta_{k,1/2} = \tfrac{1}{2}(\delta_k + \delta_{k+1}), \qquad k \ge 1.$$

Next we set $\mathcal{F}_0 := \mathcal{F}$ and define recursively

$$\mathcal{F}_{k+1} := \{f \in \mathcal{F}_k : P\{f \le \delta_{k,1/2}\} \le r_{k+1}/2\}.$$

For $k \ge 0$, define a continuous function $\varphi_k$ from $\mathbb{R}$ into $[0,1]$ such that $\varphi_k(u) = 1$ for $u \le \delta_{k,1/2}$, $\varphi_k(u) = 0$ for $u \ge \delta_k$ and $\varphi_k$ is linear for $\delta_{k,1/2} \le u \le \delta_k$. Also, for $k \ge 1$, let $\bar{\varphi}_k$ be a continuous function from $\mathbb{R}$ into $[0,1]$ such that $\bar{\varphi}_k(u) = 1$ for $u \le \delta_k$, $\bar{\varphi}_k(u) = 0$ for $u \ge \delta_{k-1,1/2}$, and $\bar{\varphi}_k$ is linear for $\delta_k \le u \le \delta_{k-1,1/2}$. It follows from (3.2) that $\delta_k \in (\delta/2, \delta)$ for all $k$ such that $1 \le k \le \log_2 \log_2 \varepsilon^{-1}$. Let us introduce the function classes

$$\mathcal{G}_k := \{\varphi_k \circ f : f \in \mathcal{F}_k\}, \qquad k \ge 0$$

and

$$\bar{\mathcal{G}}_k := \{\bar{\varphi}_k \circ f : f \in \mathcal{F}_k\}, \qquad k \ge 1.$$

It follows from the definitions that, for $k \ge 1$,

$$\sup_{g \in \mathcal{G}_k} Pg^2 \le \sup_{f \in \mathcal{F}_k} P\{f \le \delta_k\} \le \sup_{f \in \mathcal{F}_k} P\{f \le \delta_{k-1,1/2}\} \le r_k/2 \le r_k$$

and

$$\sup_{g \in \bar{\mathcal{G}}_k} Pg^2 \le \sup_{f \in \mathcal{F}_k} P\{f \le \delta_{k-1,1/2}\} \le r_k/2 \le r_k.$$

(For $k = 0$, the first inequality also holds since $r_0 = 1$.)

Recall a commonly used notation

$$\|Y\|_{\mathcal{G}} := \sup_{g \in \mathcal{G}} |Y(g)|, \qquad Y : \mathcal{G} \mapsto \mathbb{R},$$

$\mathcal{G}$ being a class of measurable functions on $S$, and note that in what follows we view signed measures $\nu$ on $(S, \mathcal{A})$ (such as, e.g., $P_n - P$) as mappings $g \mapsto \nu(g) = \int_S g \, d\nu$.

Consider the events

$$E^{(k)} := \{\|P_n - P\|_{\mathcal{G}_{k-1}} \le K_1 \mathbb{E}\|P_n - P\|_{\mathcal{G}_{k-1}} + K_2\sqrt{r_{k-1}\varepsilon} + K_3\varepsilon\}$$
$$\cap \{\|P_n - P\|_{\bar{\mathcal{G}}_k} \le K_1 \mathbb{E}\|P_n - P\|_{\bar{\mathcal{G}}_k} + K_2\sqrt{r_k\varepsilon} + K_3\varepsilon\}, \qquad k \ge 1.$$

By the concentration inequality of Talagrand (Theorem 1.4 in [25]; see also [22]), for some values of the numerical constants $K_1, K_2, K_3 > 0$,

$$\mathbb{P}\big((E^{(k)})^c\big) \le 2e^{-n\varepsilon/2}.$$

We set $E_0 = \Omega$,

$$E_N := \bigcap_{k=1}^{N} E^{(k)}, \qquad N \geq 1.$$

Clearly,

$$\mathbb{P}(E_N^c) \leq 2Ne^{-n\varepsilon/2}. \tag{3.3}$$

Assume, without loss of generality, that $\varepsilon < (2+C)^{-2}$, which implies $r_{k+1} < r_k$ and $\delta_k \in (\delta/2, \delta]$, $k \geq 0$. [If $\varepsilon \geq (2+C)^{-2}$, the bounds of the theorem hold with any constant $A > 2 + C$.] The rest of the proof is based on the following lemma.

LEMMA 1. *Let*

$$\mathcal{J} := \left\{ \inf_{f \in \mathcal{F}} P_n\{f \leq \delta\} \leq \varepsilon \right\}.$$

*For any $N$ such that*

$$N \leq \log_2 \log_2 \varepsilon^{-1} \quad \textit{and} \quad r_N \geq \varepsilon, \tag{3.4}$$

*we have on the event $E_N \cap \mathcal{J}$,*

(i) $$\forall f \in \mathcal{F} \qquad P_n\{f \leq \delta\} \leq \varepsilon \quad \Rightarrow \quad f \in \mathcal{F}_N$$

*and*

(ii) $$\sup_{f \in \mathcal{F}_k} P_n\{f \leq \delta_k\} \leq r_k, \qquad 0 \leq k \leq N.$$

PROOF. We will prove the lemma by induction with respect to $N$. For $N = 0$, the statement is obvious. Suppose it holds for some $N \geq 0$, such that $N+1$ still satisfies condition (3.4). Then, on the event $E_N \cap \mathcal{J}$,

$$\sup_{f \in \mathcal{F}_k} P_n\{f \leq \delta_k\} \leq r_k, \qquad 0 \leq k \leq N$$

and

$$\forall f \in \mathcal{F} \qquad P_n\{f \leq \delta\} \leq \varepsilon \quad \Rightarrow \quad f \in \mathcal{F}_N.$$

Suppose that $f \in \mathcal{F}$ is such that $P_n\{f \leq \delta\} \leq \varepsilon$. By the induction assumptions, $f \in \mathcal{F}_N$ on the event $E_N$. Hence, on the event $E_{N+1}$,

$$\begin{aligned} P\{f \leq \delta_{N,1/2}\} &\leq P_n\{f \leq \delta_N\} + \|P_n - P\|_{\mathcal{G}_N} \\ &\leq \varepsilon + K_1 \mathbb{E}\|P_n - P\|_{\mathcal{G}_N} + K_2\sqrt{r_N \varepsilon} + K_3\varepsilon. \end{aligned} \tag{3.5}$$

Given a class $\mathcal{G}$, let

$$\hat{R}_n(\mathcal{G}) := \left\| n^{-1} \sum_{i=1}^{n} \varepsilon_i \delta_{X_i} \right\|_{\mathcal{G}},$$

where $\{\varepsilon_i\}$ is a sequence of i.i.d. Rademacher random variables and $\delta_x$ denotes the probability measure concentrated at a point $x \in S$. (The random variable $\hat{R}_n(\mathcal{G})$ is called *the Rademacher complexity* of the class $\mathcal{G}$. It was used by Koltchinskii [15], Bartlett, Boucheron and Lugosi [3] and Koltchinskii and Panchenko [16] as a randomized complexity penalty in learning problems.) The symmetrization inequality

$$\mathbb{E}\|P_n - P\|_{\mathcal{G}} \le 2\mathbb{E}\hat{R}_n(\mathcal{G})$$

(see, e.g., [26], Lemma 2.3.1) yields

$$\mathbb{E}\|P_n - P\|_{\mathcal{G}_N} \le 2\mathbb{E}I_{E_N}\mathbb{E}_\varepsilon \hat{R}_n(\mathcal{G}_N) + 2\mathbb{E}I_{E_N^c}\mathbb{E}_\varepsilon \hat{R}_n(\mathcal{G}_N). \tag{3.6}$$

Using the entropy inequalities for sub-Gaussian processes (see [26], Corollary 2.2.8), we get

$$\begin{aligned}\mathbb{E}_\varepsilon \hat{R}_n(\mathcal{G}_N) &\le \inf_{g\in\mathcal{G}_N} \mathbb{E}_\varepsilon \left| n^{-1}\sum_{j=1}^n \varepsilon_j g(X_j)\right| \\ &\quad + \frac{\text{const}}{\sqrt{n}} \int_0^{(2\sup_{g\in\mathcal{G}_N} P_n g^2)^{1/2}} H^{1/2}_{d_{P_n,2}}(\mathcal{G}_N; u)\, du.\end{aligned} \tag{3.7}$$

REMARK. Here and in what follows in the proof, "const" denotes a constant; its values can be different in different places.

The induction assumption implies that on the event $E_N \cap \mathcal{J}$,

$$\begin{aligned}\inf_{g\in\mathcal{G}_N} \mathbb{E}_\varepsilon \left| n^{-1}\sum_{j=1}^n \varepsilon_j g(X_j)\right| &\le \inf_{g\in\mathcal{G}_N} \mathbb{E}_\varepsilon^{1/2} \left| n^{-1}\sum_{j=1}^n \varepsilon_j g(X_j)\right|^2 \le \frac{1}{\sqrt{n}} \inf_{g\in\mathcal{G}_N} \sqrt{P_n g^2} \\ &\le \frac{1}{\sqrt{n}} \inf_{f\in\mathcal{F}_N} \sqrt{P_n\{f \le \delta_N\}} \\ &\le \frac{1}{\sqrt{n}} \inf_{f\in\mathcal{F}_N} \sqrt{P_n\{f \le \delta\}} \le \sqrt{\frac{\varepsilon}{n}} \le \varepsilon,\end{aligned}$$

since $\varepsilon > n^{-1}$. Also, on the same event,

$$\sup_{g\in\mathcal{G}_N} P_n g^2 \le \sup_{f\in\mathcal{F}_N} P_n\{f \le \delta_N\} \le r_N.$$

The Lipschitz constants $\varphi_{k-1}$ and $\bar{\varphi}_k$ are bounded by

$$L = 2(\delta_{k-1} - \delta_k)^{-1} = 2\delta^{-1}\gamma_{k-1}^{-1} = \frac{2}{\delta}\sqrt{\frac{r_{k-1}}{\varepsilon}},$$

which yields

$$d_{P_n,2}(\varphi_N \circ f; \varphi_N \circ g) = \left( n^{-1} \sum_{j=1}^{n} |\varphi_N(f(X_j)) - \varphi_N(g(X_j))|^2 \right)^{1/2} \le \frac{2}{\delta} \sqrt{\frac{r_N}{\varepsilon}} d_{P_n,2}(f,g).$$

Note that for $\varepsilon \ge \varepsilon_n^{\psi}(\delta)$, the inequality $\psi(\delta\sqrt{\varepsilon}/2)/(\delta\sqrt{n}) \le \varepsilon$ holds. It follows that, on the event $E_N \cap \mathcal{J}$,

$$\begin{aligned} &\frac{1}{\sqrt{n}} \int_0^{(2\sup_{g\in\mathcal{G}_N} P_n g^2)^{1/2}} H^{1/2}_{d_{P_n,2}}(\mathcal{G}_N; u)\, du \\ &\quad \le \frac{1}{\sqrt{n}} \int_0^{(2r_N)^{1/2}} H^{1/2}_{d_{P_n,2}}\left(\mathcal{F}; \frac{\delta\sqrt{\varepsilon}u}{2\sqrt{r_N}}\right) du \\ &\quad \le \frac{1}{\sqrt{n}} \frac{2\sqrt{r_N}}{\delta\sqrt{\varepsilon}} \int_0^{\delta\sqrt{\varepsilon}/2} H^{1/2}_{d_{P_n,2}}(\mathcal{F}; v)\, dv \le \frac{1}{\sqrt{n}} \frac{2\sqrt{r_N}}{\delta\sqrt{\varepsilon}} D_n \psi\left(\frac{\delta\sqrt{\varepsilon}}{2}\right) \\ &\quad \le \frac{2D_n\sqrt{r_N}}{\sqrt{\varepsilon}} \varepsilon = 2D_n\sqrt{r_N\varepsilon}. \end{aligned} \tag{3.8}$$

Now (3.7) and (3.8) imply that on the same event,

$$\mathbb{E}_\varepsilon \hat{R}_n(\mathcal{G}_N) \le \text{const}(1 + D_n)\sqrt{r_N\varepsilon}. \tag{3.9}$$

Since $\mathbb{E}_\varepsilon \hat{R}_n(\mathcal{G}_{N+1}) \le 1$, we conclude from (3.3), (3.6) and (3.9) that

$$\begin{aligned} \mathbb{E}\|P_n - P\|_{\mathcal{G}_N} &\le \text{const}(1 + \mathbb{E}D_n)\sqrt{r_N\varepsilon} + 2\mathbb{P}(E_N^c) \\ &\le \text{const}(1 + \mathbb{E}D_n)\sqrt{r_N\varepsilon} + 4Ne^{-n\varepsilon/2}. \end{aligned}$$

By condition (3.4) and the fact that $\varepsilon \ge 2\log n/n$, we have $4Ne^{-n\varepsilon/2} \le \varepsilon$. Therefore,

$$\mathbb{E}\|P_n - P\|_{\mathcal{G}_N} \le \text{const}(1 + \mathbb{E}D_n)\sqrt{r_N\varepsilon}.$$

By (3.5), on the event $E_{N+1} \cap \mathcal{J}$,

$$P\{f \le \delta_{N,1/2}\} \le \text{const}(1 + \mathbb{E}D_n)(\varepsilon + \sqrt{r_N\varepsilon}). \tag{3.10}$$

Choosing a constant $c > 0$ in the recurrent relationship that defines the sequence $\{r_k\}$ properly, we ensure that on the event $E_{N+1} \cap \mathcal{J}$,

$$P\{f \le \delta_{N,1/2}\} \le \tfrac{1}{2}C\sqrt{r_N\varepsilon} = r_{N+1}/2.$$

This implies that $f \in \mathcal{F}_{N+1}$ and the induction step for (i) is proved.

To prove (ii), note that on the event $E_{N+1}$,

$$\begin{aligned}\sup_{f\in\mathcal{F}_{N+1}} P_n\{f\le\delta_{N+1}\} &\le \sup_{f\in\mathcal{F}_{N+1}} P\{f\le\delta_{N,1/2}\} + \|P_n-P\|_{\bar{\mathcal{G}}_{N+1}}\\ &\le r_{N+1}/2 + K_1\mathbb{E}\|P_n-P\|_{\bar{\mathcal{G}}_{N+1}} + K_2\sqrt{r_{N+1}\varepsilon} + K_3\varepsilon.\end{aligned} \tag{3.11}$$

Using the symmetrization inequality, we get

$$\mathbb{E}\|P_n-P\|_{\bar{\mathcal{G}}_{N+1}} \le 2\mathbb{E}I_{E_N}\mathbb{E}_\varepsilon\hat{R}_n(\bar{\mathcal{G}}_{N+1}) + 2\mathbb{E}I_{E_N^c}\mathbb{E}_\varepsilon\hat{R}_n(\bar{\mathcal{G}}_{N+1}). \tag{3.12}$$

Similarly to (3.7),

$$\begin{aligned}\mathbb{E}_\varepsilon R_n(\bar{\mathcal{G}}_{N+1}) &\le \inf_{g\in\bar{\mathcal{G}}_{N+1}} \mathbb{E}_\varepsilon\left|n^{-1}\sum_{j=1}^n \varepsilon_j g(X_j)\right|\\ &\quad + \frac{\text{const}}{\sqrt{n}}\int_0^{(2\sup_{g\in\bar{\mathcal{G}}_{N+1}} P_n g^2)^{1/2}} H^{1/2}_{d_{P_n,2}}(\bar{\mathcal{G}}_{N+1};u)\,du.\end{aligned} \tag{3.13}$$

It follows from (i) that on the event $E_{N+1}\cap\mathcal{J}$,

$$\begin{aligned}&\inf_{g\in\bar{\mathcal{G}}_{N+1}} \mathbb{E}_\varepsilon\left|n^{-1}\sum_{j=1}^n \varepsilon_j g(X_j)\right|\\ &\quad\le \inf_{g\in\bar{\mathcal{G}}_{N+1}} \mathbb{E}_\varepsilon^{1/2}\left|n^{-1}\sum_{j=1}^n \varepsilon_j g(X_j)\right|^2 \le \frac{1}{\sqrt{n}}\inf_{g\in\bar{\mathcal{G}}_{N+1}}\sqrt{P_n g^2}\\ &\quad\le \frac{1}{\sqrt{n}}\inf_{f\in\mathcal{F}_{N+1}}\sqrt{P_n\{f\le\delta_{N,1/2}\}} \le \frac{1}{\sqrt{n}}\inf_{f\in\mathcal{F}_{N+1}}\sqrt{P_n\{f\le\delta\}} \le \sqrt{\frac{\varepsilon}{n}} \le \varepsilon.\end{aligned}$$

The induction assumption implies that on the event $E_{N+1}\cap\mathcal{J}$,

$$\sup_{g\in\bar{\mathcal{G}}_{N+1}} P_n g^2 \le \sup_{f\in\mathcal{F}_N} P_n\{f\le\delta_{N,1/2}\} \le r_N.$$

Since the Lipschitz constant $\bar{\varphi}_k$ is bounded by $\frac{2}{\delta}\sqrt{r_{k-1}/\varepsilon}$, we have

$$\begin{aligned}d_{P_n,2}(\bar{\varphi}_{N+1}\circ f;\bar{\varphi}_{N+1}\circ g) &= \left(n^{-1}\sum_{j=1}^n\left|\bar{\varphi}_{N+1}\circ f(X_j)-\bar{\varphi}_{N+1}\circ g(X_j)\right|^2\right)^{1/2}\\ &\le \frac{2}{\delta}\sqrt{\frac{r_N}{\varepsilon}}d_{P_n,2}(f,g).\end{aligned}$$

Similarly to (3.8), we have on the event $E_{N+1} \cap \mathcal{J}$,

$$\begin{aligned}
&\frac{1}{\sqrt{n}} \int_0^{(2\sup_{g\in\bar{\mathcal{G}}_{N+1}} P_n g^2)^{1/2}} H^{1/2}_{d_{P_n,2}}(\bar{\mathcal{G}}_{N+1}; u)\, du \\
&\qquad \le \frac{1}{\sqrt{n}} \int_0^{(2r_N)^{1/2}} H^{1/2}_{d_{P_n,2}}\Big(\mathcal{F}; \frac{\delta\sqrt{\varepsilon}u}{2\sqrt{r_N}}\Big)\, du \\
&\qquad \le \frac{1}{\sqrt{n}} \frac{2\sqrt{r_N}}{\delta\sqrt{\varepsilon}} \int_0^{\delta\sqrt{\varepsilon}/2} H^{1/2}_{d_{P_n,2}}(\mathcal{F}; v)\, dv \le \frac{1}{\sqrt{n}} \frac{2\sqrt{r_N}}{\delta\sqrt{\varepsilon}} D_n \psi\Big(\frac{\delta\sqrt{\varepsilon}}{2}\Big) \\
&\qquad \le \frac{2D_n\sqrt{r_N}}{\sqrt{\varepsilon}}\varepsilon = 2D_n\sqrt{r_N\varepsilon}.
\end{aligned} \tag{3.14}$$

Combining all the bounds, we prove that on the same event,

$$\sup_{f\in\mathcal{F}_{N+1}} P_n\{f \le \delta_{N+1}\} \le \frac{r_{N+1}}{2} + \text{const}(1 + \mathbb{E}D_n)\sqrt{r_N\varepsilon}. \tag{3.15}$$

Choosing a constant $c > 0$ in the recurrent relationship that defines the sequence $\{r_k\}$ properly, we get on the event $E_{N+1} \cap \mathcal{J}$,

$$\sup_{f\in\mathcal{F}_{N+1}} P_n\{f \le \delta_{N+1}\} \le C\sqrt{r_N\varepsilon} = r_{N+1},$$

which completes the proof of (ii) and of the lemma. □

To complete the proof of the theorem, note that the choice of $N = [\log_2 \log_2 \varepsilon^{-1}]$ implies that $r_{N+1} \le c\varepsilon$ for some $c > 0$. Indeed, if we introduce $s_k = r_k/C$ and $\varepsilon_1 = C\varepsilon$, then $s_{k+1} = \sqrt{s_k\varepsilon}$ and $s_0 = C^{-1} \le 1$. It is easy to see that $s_N \le \varepsilon_1^{1-2^{-N}} \le 2\varepsilon_1$ for $N \ge \log_2 \log_2 \varepsilon_1^{-1}$ and, hence, $r_N \le C^2\varepsilon = \bar{A}\varepsilon$.

The proof of the second inequality is similar with minor modifications. □

To prove Theorem 3, we need the following statement, which seems to be well known, but we have not found the precise reference and give the proof here for completeness.

Let

$$\text{conv}_d(\mathcal{H}) := \Big\{ \sum_{j=1}^d \lambda_j h_j : \lambda_j \in \mathbb{R}, \sum_{j=1}^d |\lambda_j| \le 1,\ h_j \in \mathcal{H} \Big\}.$$

LEMMA 2. *Let $\mathcal{H}$ be a class of functions from $(S, \mathcal{A})$ into $\mathbb{R}$. Let $Q$ be a probability measure on $(S, \mathcal{A})$ such that*

$$\bar{H} := \sup_{h\in\mathcal{H}} (Qh^2)^{1/2} < +\infty.$$

*The following bound holds for all $d \geq 1$ and $\varepsilon > 0$:*

$$N_{d_{Q,2}}\big(\mathrm{conv}_d(\mathcal{H}), (1+\bar{H})\varepsilon\big) \leq \left(\frac{2e^2 N_{d_{Q,2}}(\mathcal{H},\varepsilon)(d'+4\varepsilon^{-2})}{d'^2}\right)^{d'},$$

*where $d' = d \wedge N_{d_{Q,2}}(\mathcal{H},\varepsilon)$.*

PROOF. First note that if $\mathcal{H}' := \mathcal{H} \cup \{h : -h \in \mathcal{H}\}$, then $\mathrm{conv}_d(\mathcal{H}') = \mathrm{conv}_d(\mathcal{H})$ and

$$N_{d_{Q,2}}(\mathcal{H}';\varepsilon) \leq 2N_{d_{Q,2}}(\mathcal{H};\varepsilon).$$

Thus, it is enough to show that for a class $\mathcal{H}$, such that $h \in \mathcal{H}$ implies $-h \in \mathcal{H}$, we have

$$N_{d_{Q,2}}\big(\mathrm{conv}_d(\mathcal{H}), (1+\bar{H})\varepsilon\big) \leq \left(\frac{e^2 N_{d_{Q,2}}(\mathcal{H},\varepsilon)(d+4\varepsilon^{-2})}{d^2}\right)^{d}.$$

For such a class we have

$$\mathrm{conv}_d(\mathcal{H}) := \left\{\sum_{j=1}^{d}\lambda_j h_j : \lambda_j \geq 0,\ \sum_{j=1}^{d}\lambda_j \leq 1,\ h_j \in \mathcal{H}\right\}.$$

Note that if $\sum_j |\lambda_j| \leq 1$, then

$$\begin{aligned} d_{Q,2}\left(\sum_j \lambda_j h_j; \sum_j \lambda_j h'_j\right) &= \left\|\sum_j \lambda_j (h_j - h'_j)\right\|_{L_2(Q)} \\ &\leq \sum_j |\lambda_j| \max_j \|h_j - h'_j\|_{L_2(Q)} \leq \max_j \|h_j - h'_j\|_{L_2(Q)}. \end{aligned}$$

It follows that if $\mathcal{H}_\varepsilon$ is an $\varepsilon$-net of $\mathcal{H}$, then a $\delta$-net of $\mathrm{conv}_d(\mathcal{H}_\varepsilon)$ is an $\varepsilon+\delta$-net of $\mathrm{conv}_d(\mathcal{H})$. This observation allows us to reduce the proof of the lemma to the case when $\mathcal{H}$ is a finite class. In this case we want to show that

$$N_{d_{Q,2}}\big(\mathrm{conv}_d(\mathcal{H}), \bar{H}\varepsilon\big) \leq \left(\frac{e^2 \mathrm{card}(\mathcal{H})(d+4\varepsilon^{-2})}{d^2}\right)^{d}.$$

To this end, we use the idea of Maurey; see [23, 26]. Let $N := \mathrm{card}(\mathcal{H})$. Consider some representation of a function $f = \sum_{i=1}^{N}\lambda_i h_i \in \mathrm{conv}_d(\mathcal{H})$. We assume that $\lambda_j \geq 0$, $\sum_j \lambda_j \leq 1$ and at most $d'$ of the coefficients are not equal to 0. Consider an i.i.d. sequence of random variables $Y_j$, $j = 1,\ldots,k$, taking values in $\mathcal{H}\cup\{0\}$ such that $P(Y_j = h_i) = \lambda_i$ for $i = 1,\ldots,N$ and $P(Y_j = 0) = 1 - \sum_{i=1}^{N}\lambda_i$. (We simply add the probabilities when the same function $h$ corresponds to several weights $\lambda_i$ with different indices.) We have

$$\begin{aligned} \mathbb{E}\left\|k^{-1}\sum_{j=1}^{k} Y_j - \sum_{i=1}^{N}\lambda_i h_i\right\|_{Q,2}^2 &= \mathbb{E}\left\|k^{-1}\sum_{j=1}^{k} Y_j - \mathbb{E}Y_1\right\|_{Q,2}^2 \\ &\leq \frac{1}{k}\mathbb{E}\|Y_1 - \mathbb{E}Y_1\|_{Q,2}^2 \leq 4\bar{H}^2 k^{-1}. \end{aligned}$$

If we set $k = 4\varepsilon^{-2}$, then with probability 1 there exists a realization $\bar{Y}_k = k^{-1}\sum_{j=1}^{k} Y_j$ such that

$$\left\| \bar{Y}_k - \sum_{i=1}^{N} \lambda_i h_i \right\|_{Q,2} \le \varepsilon \bar{H}.$$

To compute the bound for the $\bar{H}\varepsilon$-covering number we have to calculate the number of possible realizations of $k^{-1}\sum_{j=1}^{k} Y_j$. Simple combinatorics shows that this number does not exceed $\binom{N}{d'}\binom{d'+k}{k}$. Next we use the following bound, which holds for all $1 \le d \le N$:

$$\binom{N}{d}\binom{d+k}{k} \le \left(\frac{e^2 N(d+k)}{d^2}\right)^d.$$

To prove the bound, first assume that $d < N$. Then one can check using Stirling's formula that

$$\begin{aligned}\frac{N!}{d!(N-d)!}\frac{(d+k)!}{d!k!} &\le \frac{n^n}{d^d(N-d)^{N-d}}\frac{(d+k)^{d+k}}{k^k d^d} \\ &\le \left(\frac{N(d+k)}{d^2}\right)^d \left(1+\frac{d}{N-d}\right)^{N-d}\left(1+\frac{d}{k}\right)^k \\ &\le \left(\frac{e^2 N(d+k)}{d^2}\right)^d.\end{aligned}$$

The case when $d = N$ can be considered similarly. The bound immediately implies the result. □

PROOF OF THEOREM 3. Let us fix $\delta \in (0, 1/2]$. For any function $f$ we denote $d(f) := d(f, \bar{\Delta})$, where $\bar{\Delta}$ is such that the infimum in the definition (2.12) is attained at $\bar{\Delta}$. For a fixed $\delta$ we consider a partition of $\mathcal{F}$ into two classes $\mathcal{F}_1^\delta$ and $\mathcal{F}_2^\delta = \mathcal{F} \setminus \mathcal{F}_1^\delta$, where $\mathcal{F}_1^\delta := \{f : d(f) = 0\}$ [note that $d(f)$ depends on $\delta$]. The fact that $f \in \mathcal{F}_1^\delta$ means that the weights of the classifier $f$ are distributed "uniformly" and in this case the bound of the theorem does not improve upon Example 1. The family of classes that we use to localize the classifier $f$ is defined as

$$\mathcal{F}_{d,\Delta} := \{f \in \mathcal{F}_2^\delta : d(f; \Delta) \le d\}.$$

If $f \in \mathcal{F}_{d,\Delta}$ and $\Delta$ is small, then it means that the "voting power" is concentrated in the faction that consists of the first $d$ base classifiers of the convex combination. In the first four steps of the proof we will deal with $\mathcal{F}_2^\delta$ and we will assume only that the class $\mathcal{H}$ has a square integrable envelope $H$. In Step 1 we will start by estimating the complexity of the class $\mathcal{F}_{d,\Delta}$ and applying Theorem 2 to get uniform control of the margin over class $\mathcal{F}_{d,\Delta}$. In the remaining steps we will show that the bound of Step 1 is tight enough and it allows us to make this control uniform over all parameters such as $d$, $\Delta$ and $\delta$.

*Step* 1. Let $1 \le d \le n$. Denote

$$\varepsilon_n(d;\delta;\Delta) := \left[\frac{d}{n}\left(\log\frac{1}{\delta} + \log\frac{ne^2}{d}\right) + \left(\frac{\Delta}{\delta}\right)^{2\alpha/(\alpha+2)} n^{-2/(\alpha+2)}\right] \vee \frac{2\log n}{n}.$$

We start by proving (with some constants $A, B > 0$) the inequality

$$\mathbb{P}\left\{\exists\, f \in \mathcal{F}_{d,\Delta} : P_n\{f \le \delta\} \le \varepsilon_n(d;\delta;\Delta) \text{ and } P\left\{f \le \frac{\delta}{2}\right\} \ge A\varepsilon_n(d;\delta;\Delta)\right\}$$
$$\le B\left(\frac{\delta d}{n}\right)^{d/4} \exp\left\{-\frac{1}{4}\left(\sqrt{n}\frac{\Delta}{\delta}\right)^{2\alpha/(\alpha+2)}\right\}. \tag{3.16}$$

Clearly we can and do assume that $\varepsilon_n(d;\delta;\Delta) \le 1$. To prove (3.16), we bound the random entropy $H_{d_{P_n,2}}(\mathcal{F}_{d,\Delta};\varepsilon)$ of the class $\mathcal{F}_{d,\Delta}$ in the manner

$$H_{d_{P_n,2}}(\mathcal{F}_{d,\Delta};\varepsilon) \le K(1+P_nH^2)\left[d\log\frac{e}{\varepsilon} + \left(\frac{\Delta}{\varepsilon}\right)^{\alpha}\right] \qquad \text{for } \varepsilon \le 1 \tag{3.17}$$

with some constant $K > 0$. The last bound follows from the observation that each function $f \in \mathcal{F}_{d,\Delta}$ can be represented as $f = f_1 + f_2$, where

$$f_1 \in \mathcal{F}_d := \operatorname{conv}_d(\mathcal{H}) = \left\{\sum_{j=1}^{d} \lambda_j h_j : \lambda_j \in \mathbb{R}, \sum_{j=1}^{d} |\lambda_j| \le 1,\ h_j \in \mathcal{H}\right\}$$

and

$$f_2 \in \mathcal{F}_\Delta := \Delta\operatorname{conv}(\mathcal{H}).$$

Hence, by simple combining of $\varepsilon$-coverings for the classes $\mathcal{F}_d$ and $\mathcal{F}_\Delta$, we get

$$H_{d_{P_n,2}}(\mathcal{F}_{d,\Delta};\varepsilon) \le H_{d_{P_n,2}}(\mathcal{F}_d;\varepsilon/2) + H_{d_{P_n,2}}(\mathcal{F}_\Delta;\varepsilon/2).$$

Then, a routine application of Lemma 2 and (2.11) implies

$$H_{d_{P_n,2}}(\mathcal{F}_d;\varepsilon/2) \le Kd\log\frac{e(1+P_nH^2)}{\varepsilon} \qquad \text{for } \varepsilon \le 2(P_nH^2)^{1/2}.$$

[Note that for $\varepsilon > 2(P_nH^2)^{1/2}$ we easily get $H_{d_{P_n,2}}(\mathcal{F}_d;\varepsilon/2) = 0$.] For $\varepsilon \le 1$ this implies

$$H_{d_{P_n,2}}\left(\mathcal{F}_d;\frac{\varepsilon}{2}\right) \le Kd\left[\log\frac{e}{\varepsilon} + \log(1+P_nH^2)\right]$$
$$\le Kd\left[\log\frac{e}{\varepsilon} + P_nH^2\right] \le Kd(1+P_nH^2)\log\frac{e}{\varepsilon}.$$

By the bound on the entropy of the symmetric convex hull (see [26]),

$$H_{d_{P_n,2}}\left(\mathcal{F}_\Delta;\frac{\varepsilon}{2}\right) = H_{d_{P_n,2}}\left(\mathcal{F};\frac{\varepsilon}{2\Delta}\right) \le K(1+P_nH^2)^{\alpha/4}\left(\frac{\Delta}{\varepsilon}\right)^{\alpha}$$
$$\le K(1+P_nH^2)\left(\frac{\Delta}{\varepsilon}\right)^{\alpha},$$

which implies (3.17).

Next we are using margin type bounds on generalization error under random entropy conditions (see Section 2, Theorem 2). Clearly, from (3.17), we get the bound on Dudley's entropy integral,

$$\int_0^x H^{1/2}_{d_{P_n,2}}(\mathcal{F};\varepsilon)\,d\varepsilon \le K(1+P_nH^2)^{1/2}\bar{\psi}(x),$$

where $\bar{\psi}$ is a concave nondecreasing function such that for $x \in [0,1]$,

$$\bar{\psi}(x) = \left(x\left(d\log\frac{e}{x}\right)^{1/2} + \Delta^{\alpha/2}x^{1-\alpha/2}\right)$$

with some constant $K > 0$. Let

$$\psi_1(x) := x\left(d\log\frac{e}{x}\right)^{1/2}, \qquad \psi_2(x) := \Delta^{\alpha/2}x^{1-\alpha/2},$$

$$\psi(x) := \frac{\psi_1(x)+\psi_2(x)}{2}.$$

Let us first consider the equation $\varepsilon = \psi_1(\delta\sqrt{\varepsilon})/(\delta\sqrt{n})$, which can be written as $\varepsilon = \frac{d}{n}\log\frac{e}{\delta\sqrt{\varepsilon}}$. If $\varepsilon = \frac{d}{n}x^2$, then

$$xe^{x^2} = \left(\frac{n}{d}\right)^{1/2}\frac{e}{\delta}.$$

For $d \le n$ and $\delta \le 1$, it means that $xe^{x^2} \ge 1$, and, therefore,

$$e^{x^2-1} \le \left(\frac{n}{d}\right)^{1/2}\frac{e}{\delta}$$

or

$$\varepsilon = \frac{d}{n}x^2 \le \frac{d}{n}\left[1+\log\left(\left(\frac{n}{d}\right)^{1/2}\frac{e}{\delta}\right)\right] \le \frac{d}{n}\log\frac{ne^2}{d\delta} \le \varepsilon_n(d;\delta;\Delta) \le 1.$$

[Notice that in the case when $d$ becomes significantly greater than $n$, e.g., if $(nd^{-1})^{1/2}\delta^{-1} \le 1$, then $x \le 1$ and $xe^{x^2} \le ex$, which implies that $\varepsilon \ge \delta^{-2}$ and the bound of the theorem becomes useless. This explains why in the definition of $\varepsilon_n(f;\delta)$ we minimize over $d(f,\Delta) \le n$.]

The solution of the equation $\varepsilon = \psi_2(\delta\sqrt{\varepsilon})/(\delta\sqrt{n})$ is

$$\varepsilon^{(2)} := \left(\frac{\Delta}{\delta}\right)^{2\alpha/(\alpha+2)} n^{-2/(\alpha+2)}.$$

Finally, it is easy to bound the solution of the equation $\varepsilon = \psi(\delta\sqrt{\varepsilon})/(\delta\sqrt{n})$ from above by $\varepsilon^{(1)} + \varepsilon^{(2)}$. Therefore, the solution of the last equation is also bounded from above by $\varepsilon_n(d; \delta; \Delta)$. This allows us to use the bound of Theorem 2 to get the inequality

$$\mathbb{P}\left\{\exists\, f \in \mathcal{F}_{d,\Delta} : P_n\{f \le \delta\} \le \varepsilon_n(d; \delta; \Delta) \text{ and } P\left\{f \le \frac{\delta}{2}\right\} \ge A\varepsilon_n(d; \delta; \Delta)\right\}$$
$$\le B \log_2 \log_2 \varepsilon_n(d; \delta; \Delta)^{-1} \exp\left\{-\frac{n\varepsilon_n(d; \delta; \Delta)}{2}\right\}.$$

Since, for $\varepsilon := \varepsilon_n(d; \delta; \Delta)$, we have $\varepsilon \ge \frac{2\log n}{n}$, it follows that for $n \ge 3$,

$$\frac{1}{\varepsilon} \log \log_2 \log_2 \frac{1}{\varepsilon} \le \frac{n}{4},$$

which implies

$$B \log_2 \log_2 \varepsilon_n(d; \delta; \Delta)^{-1} \exp\left\{-\frac{n\varepsilon_n(d; \delta; \Delta)}{2}\right\}$$
$$\le B \exp\left\{-\frac{n\varepsilon_n(d; \delta; \Delta)}{4}\right\}. \tag{3.18}$$

A simple computation shows that

$$\exp\left\{-\frac{n\varepsilon_n(d; \delta; \Delta)}{4}\right\} \le \left(\frac{\delta d}{n}\right)^{d/4} \exp\left\{-\frac{1}{4}\left(\sqrt{n}\frac{\Delta}{\delta}\right)^{2\alpha/(\alpha+2)}\right\},$$

which implies (3.16)

It remains to eliminate the dependence of the bound on $d$, $\Delta$ and $\delta$.

*Step* 2. Next we show that with some constants $A, B \ge 1$, $\delta \le 1/2$ and $\Delta \ge \delta n^{-1/2}$,

$$\mathbb{P}\left\{\exists\, f \in \mathcal{F}_2^{\delta} : P_n\{f \le \delta\} \le \varepsilon_n\big(d(f; \Delta); \delta; \Delta\big) \text{ and } P\left\{f \le \frac{\delta}{2}\right\} \ge A\varepsilon_n\big(d(f; \Delta); \delta; \Delta\big)\right\}$$
$$\le B\delta^{1/8}\Delta^{1/8} \exp\left\{-\frac{1}{4}\left(\sqrt{n}\frac{\Delta}{\delta}\right)^{2\alpha/(\alpha+2)}\right\}, \tag{3.19}$$

where it is understood that if $d = d(f;\Delta) > n$, then $\varepsilon_n(d;\delta;\Delta) = 1$. Indeed, using (3.16), we have for $\delta \le 1/2$,

$$\begin{aligned}
&\mathbb{P}\Big\{\exists\, f \in \mathcal{F}_2^\delta : P_n\{f \le \delta\} \le \varepsilon_n\big(d(f;\Delta);\delta;\Delta\big) \text{ and } \\
&\qquad\qquad P\left\{f \le \frac{\delta}{2}\right\} \ge A\varepsilon_n\big(d(f;\Delta);\delta;\Delta\big)\Big\} \\
&\le \mathbb{P}\Big\{\exists\, d \le n\ \exists\, f \in \mathcal{F}_2^\delta : d(f;\Delta) = d,\ P_n\{f \le \delta\} \le \varepsilon_n(d;\delta;\Delta) \text{ and } \\
&\qquad\qquad P\left\{f \le \frac{\delta}{2}\right\} \ge A\varepsilon_n(d;\delta;\Delta)\Big\} \\
&\le \sum_{d=1}^n \mathbb{P}\Big\{\exists\, f \in \mathcal{F}_{d,\Delta} : P_n\{f \le \delta\} \le \varepsilon_n(d;\delta;\Delta) \text{ and } \\
&\qquad\qquad P\left\{f \le \frac{\delta}{2}\right\} \ge A\varepsilon_n(d;\delta;\Delta)\Big\} \\
&\le B\sum_{d=1}^n \left(\frac{\delta d}{n}\right)^{d/4} \exp\left\{-\frac{1}{4}\left(\sqrt{n}\frac{\Delta}{\delta}\right)^{2\alpha/(\alpha+2)}\right\}.
\end{aligned}$$

One can easily check that for $d \le n/(e\delta)$ (increasing $A$ we can assume that it holds) the expression $(\delta d/n)^{d/4}$ is decreasing in $d$ and, therefore, for any $k \le n/e$,

$$\sum_{d=1}^n \left(\frac{\delta d}{n}\right)^{d/4} \le k\left(\frac{\delta}{n}\right)^{1/4} + \sum_{d=k+1}^n \left(\frac{\delta d}{n}\right)^{d/4} \le k\left(\frac{\delta}{n}\right)^{1/4} + \delta^{k/4}.$$

Optimizing over $k$ we take $k = \log n/\log\delta^{-1} + 1$ to get

$$k\left(\frac{\delta}{n}\right)^{1/4} + \delta^{k/4} \le 2\left(\frac{\log n}{\log\delta^{-1}} + 1\right)\left(\frac{\delta}{n}\right)^{1/4} \le \delta^{1/8}\Delta^{1/8},$$

where the last inequality holds under the assumption that $\Delta \ge \delta n^{-1/2}$.

*Step* 3. Our next goal is to prove that with some constants $A, B > 1$ and for $0 < t < n^{\alpha/(2+\alpha)}$,

$$\begin{aligned}
&\mathbb{P}\Big\{\exists\, f \in \mathcal{F}_2^\delta : P_n\{f \le \delta\} \le \varepsilon_n(f;\delta) \text{ and } \\
&\qquad P\left\{f \le \frac{\delta}{2}\right\} \ge A \inf_{\Delta \ge \delta n^{-1/2} t^{1/\alpha+1/2}} \varepsilon_n\big(d(f;\Delta);\delta;\Delta\big)\Big\}. \\
&\qquad \le B\delta^{1/8}e^{-t/4}.
\end{aligned} \tag{3.20}$$

Let $\Delta_j := 2^{-j},\ j \ge 0$. Let $\mathcal{J} = \{j : \Delta_j \ge \delta n^{-1/2} t^{1/\alpha+1/2}\}$. Note that the condition $t < n^{\alpha/(2+\alpha)}$ guarantees that $\mathcal{J} \ne \varnothing$. Using (3.19), we get

$$\begin{aligned}
&\mathbb{P}\Big\{\exists\, f \in \mathcal{F}_2^\delta : P_n\{f \le \delta\} \le \varepsilon_n(f;\delta) \text{ and } \\
&\quad P\left\{f \le \frac{\delta}{2}\right\} \ge A \inf_{\mathcal{J}} \varepsilon_n\big(d(f;\Delta_j);\delta;\Delta_j\big)\Big\} \\
&\quad \le \mathbb{P}\Big\{\exists\, f \in \mathcal{F}_2^\delta\ \exists\, j \in \mathcal{J} : P_n\{f \le \delta\} \le \varepsilon_n(f;\delta) \text{ and } \\
&\qquad\qquad P\left\{f \le \frac{\delta}{2}\right\} \ge A\varepsilon_n\big(d(f;\Delta_j);\delta;\Delta_j\big)\Big\} \\
&\quad \le \sum_{\mathcal{J}} \mathbb{P}\Big\{\exists\, f \in \mathcal{F}_2^\delta : P_n\{f \le \delta\} \le \varepsilon_n(f;\delta) \text{ and } \\
&\qquad\qquad P\left\{f \le \frac{\delta}{2}\right\} \ge A\varepsilon_n\big(d(f;\Delta_j);\delta;\Delta_j\big)\Big\} \\
&\quad \le B\sum_{\mathcal{J}} \delta^{1/8}\Delta_j^{1/8} \exp\left\{-\frac{1}{4}\left(\sqrt{n}\frac{\Delta_j}{\delta}\right)^{2\alpha/(\alpha+2)}\right\} \le B'\delta^{1/8}e^{-t/4}.
\end{aligned}$$

To complete the proof of (3.20), note that for $\Delta \in (\Delta_{j+1}, \Delta_j]$ we have

$$\begin{aligned}
\frac{d(f;\Delta_j)}{n}\left(\log\frac{1}{\delta} + \log\frac{ne}{d(f;\Delta_j)}\right) &\le \frac{d(f;\Delta)}{n}\left(\log\frac{1}{\delta} + \log\frac{ne}{d(f;\Delta)}\right), \\
\left(\frac{\Delta_j}{\delta}\right)^{2\alpha/(\alpha+2)} n^{-2/(\alpha+2)} &\le 2^{2\alpha/(\alpha+2)}\left(\frac{\Delta}{\delta}\right)^{2\alpha/(\alpha+2)} n^{-2/(\alpha+2)}, \\
\log\log\frac{2}{\Delta_j} &\le \log\log\frac{2}{\Delta},
\end{aligned}$$

which implies $\varepsilon_n(f;\Delta_j;\delta) \le 2^{2\alpha/(\alpha+2)}\varepsilon_n(f;\Delta;\delta)$ and, therefore,

$$\inf_{\mathcal{J}} \varepsilon_n\big(d(f;\Delta_j);\delta;\Delta_j\big) \le 2^{2\alpha/(\alpha+2)} \inf_{\Delta \ge \delta n^{-1/2} t^{1/\alpha+1/2}} \varepsilon_n\big(d(f;\Delta);\delta;\Delta\big)$$

and (3.20) follows.

*Step* 4. Now we prove that for some constants $A, B > 1$ and for all $0 < t < n^{\alpha/2+\alpha}$,

$$\begin{aligned}
&\mathbb{P}\left\{\exists\, f \in \mathcal{F}_2^\delta : P_n\{f \le \delta\} \le \varepsilon_n(f;\delta) \text{ and } P\left\{f \le \frac{\delta}{2}\right\} \ge A\left(\varepsilon_n(f;\delta) + \frac{t}{n}\right)\right\} \\
&\qquad \le B\delta^{1/8}e^{-t/4}.
\end{aligned} \tag{3.21}$$

Because of (3.20), it is enough to show that

$$\inf_{\substack{\Delta\ge\delta n^{-1/2}t^{1/\alpha+1/2}\\ \Delta\in\Delta_f}} \varepsilon_n\big(d(f;\Delta);\delta;\Delta\big) \le \varepsilon_n(f;\delta)+\frac{t}{n}. \tag{3.22}$$

Since $d(f;\Delta)$ is a decreasing function of $\Delta$, the set $\Delta_f$ is an interval of the form $[c,1]$ for some $c\le 1$. Let $\Delta_0 := \delta n^{-1/2}t^{1/\alpha+1/2}$. If $\Delta_0\notin\Delta_f$, then (3.22) clearly holds. Otherwise, suppose that the infimum in the definition of $\varepsilon_n(f;\delta)$ is attained at $\Delta=\bar\Delta$. If $\bar\Delta\ge\Delta_0$, then (3.22) is also obvious. In the case when $\bar\Delta<\Delta_0$, note that

$$\left(\frac{\Delta_0}{\delta}\right)^{2\alpha/(\alpha+2)} n^{-2/(\alpha+2)} = \frac{t}{n}$$

and the function $\frac{d(f;\Delta)}{n}(\log\frac{1}{\delta}+\log(ne^2/(d(f;\Delta)))$ is decreasing in $\Delta$. Therefore,

$$\begin{aligned}\inf_{\substack{\Delta\ge\delta n^{-1/2}t^{1/\alpha+1/2}\\ \Delta\in\Delta_f}} \varepsilon_n\big(d(f;\Delta);\delta;\Delta\big) &\le \varepsilon_n\big(d(f;\Delta_0);\delta;\Delta_0\big)\\ &\le \frac{d(f;\bar\Delta)}{n}\left(\log\frac{1}{\delta}+\log\frac{ne^2}{d(f;\bar\Delta)}\right)+\frac{t}{n}\\ &\le \varepsilon_n\big(d(f;\bar\Delta);\delta;\bar\Delta\big)+\frac{t}{n}\le\varepsilon_n(f;\delta)+\frac{t}{n},\end{aligned}$$

which proves (3.22).

*Step* 5. To complete the proof of the theorem, define the event

$$\begin{aligned}E := \Big\{&\exists f\in\mathcal{F}\ \exists\delta\in(0,1): P_n\{f\le\delta\}\le\varepsilon_n(f;\delta)\\ &\text{and } P\left\{f\le\frac{\delta}{4}\right\}\ge A\left(\varepsilon_n\left(f;\frac{\delta}{2}\right)+\frac{t}{n}\right)\Big\}.\end{aligned}$$

Obviously, $E=E_1\cup E_2$, where

$$\begin{aligned}E_1 := \Big\{&\exists\delta\in(0,1)\ \exists f\in\mathcal{F}_1^\delta: P_n\{f\le\delta\}\le\varepsilon_n(f;\delta)\\ &\text{and } P\left\{f\le\frac{\delta}{4}\right\}\ge A\left(\varepsilon_n\left(f;\frac{\delta}{2}\right)+\frac{t}{n}\right)\Big\},\\ E_2 := \Big\{&\exists\delta\in(0,1)\ \exists f\in\mathcal{F}_2^\delta: P_n\{f\le\delta\}\le\varepsilon_n(f;\delta)\\ &\text{and } P\left\{f\le\frac{\delta}{4}\right\}\ge A\left(\varepsilon_n\left(f;\frac{\delta}{2}\right)+\frac{t}{n}\right)\Big\}.\end{aligned}$$

We set $\delta_j := 2^{-j},\ j \geq 0$, and

$$\bar{E}_2 := \Bigg\{ \exists\, j \geq 0\ \exists\, f \in \mathcal{F}_2^{\delta_j} : P_n\{f \leq \delta_j\} \leq \varepsilon_n(f;\delta_j) \\ \text{and } P\left\{ f \leq \frac{\delta_j}{2} \right\} \geq A\left( \varepsilon_n(f;\delta_j) + \frac{t}{n} \right) \Bigg\}.$$

It is easily seen that $E_2 \subset \bar{E}_2$. It follows from (3.21) that

$$\mathbb{P}(E_2) \leq \mathbb{P}(\bar{E}_2) \leq \sum_{j=0}^{\infty} \mathbb{P}\Bigg\{ \exists\, f \in \mathcal{F}_2^{\delta_j} : P_n\{f \leq \delta_j\} \leq \varepsilon_n(f;\delta_j) \\ \text{and } P\left\{ f \leq \frac{\delta_j}{2} \right\} \geq A\left( \varepsilon_n(f;\delta_j) + \frac{t}{n} \right) \Bigg\} \\ \leq \sum_{j=0}^{\infty} B\delta_j^{1/8} e^{-t/4} \leq B' e^{-t/4}.$$

If $f = \sum \lambda_i h_i \in \mathcal{F}_1^{\delta}$ for some $\delta$, then

$$\varepsilon_n(f,\delta) = \left( \frac{\Delta(f)}{\delta} \right)^{2\alpha/(2+\alpha)} n^{-2/(2+\alpha)} \vee \frac{2\log n}{n},$$

where $\Delta(f) := \sum |\lambda_i|$. Therefore, with some constant $A'$,

$$E_1 \subseteq E_1' := \Bigg\{ \exists\, \delta \in (0,1)\ \exists\, f \in \mathcal{F} : \\ P_n\{f \leq \delta\} \leq \left( \frac{2\Delta(f)}{\delta} \right)^{2\alpha/(2+\alpha)} n^{-2/(2+\alpha)} \vee \frac{2\log n}{n} \\ \text{and } P\left\{ f \leq \frac{\delta}{4} \right\} \geq A'\left( \left( \frac{\Delta(f)}{\delta} \right)^{2\alpha/(2+\alpha)} n^{-2/(2+\alpha)} \vee \frac{2\log n}{n} + \frac{t}{n} \right) \Bigg\}.$$

Let us first consider the case when the class $\mathcal{H}$ is uniformly bounded (say, by constant 1). One can observe that $\mathcal{F}' = \{f/\Delta(f) : f \in \mathcal{F}\} \subset \{f \in \operatorname{conv}(\mathcal{H}) : \Delta(f) = 1\}$. For any function $f$ and any $\delta \geq \Delta(f)$, $P(f \leq \delta) = 1$, which means that on the event $E_1'$ one has to take into account only values of $\delta \leq \Delta(f)$ or, equivalently, $\delta/\Delta(f) \leq 1$. Therefore, a simple rescaling $\delta' = \delta/\Delta(f) < 1$ shows that

$$E_1' = \Bigg\{ \exists\, \delta \in (0,1)\ \exists\, f \in \mathcal{F}' : P_n\{f \leq \delta\} \leq \left( \frac{2}{\delta} \right)^{2\alpha/(2+\alpha)} n^{-2/(2+\alpha)} \vee \frac{2\log n}{n} \\ \text{and } P\left\{ f \leq \frac{\delta}{4} \right\} \geq A\left( \left( \frac{1}{\delta} \right)^{2\alpha/(2+\alpha)} n^{-2/(2+\alpha)} \vee \frac{2\log n}{n} + \frac{t}{n} \right) \Bigg\}.$$

As to the second condition on $\mathcal{F}$, in this case, $\Delta(f) = 1$ for any $f$ by definition and the above equivalent representation of the event $E_1'$ holds automatically.

Let $\delta_j = 2^{-j}$, $j \geq 0$. Theorem 2 (see also Example 1) and a bound similar to (3.18) immediately imply that for some $A$ and $B$,

$$\mathbb{P}\Biggl\{\exists\, j\ \exists\, f \in \mathcal{F}' : P_n\{f \leq \delta_j\} \leq \biggl(\frac{1}{\delta_j}\biggr)^{2\alpha/(2+\alpha)} n^{-2/(2+\alpha)} \vee \frac{2\log n}{n}$$

$$\text{and } P\biggl\{f \leq \frac{\delta_j}{2}\biggr\} \geq A\biggl(\biggl(\frac{1}{\delta_j}\biggr)^{2\alpha/(2+\alpha)} n^{-2/(2+\alpha)} \vee \frac{2\log n}{n} + \frac{t}{n}\biggr)\Biggr\}$$

$$\leq \sum_{j\geq 0} B\exp\biggl\{-\frac{1}{4}\biggl(\frac{\sqrt{n}}{\delta_j}\biggr)^{2\alpha/(2+\alpha)}\biggr\} e^{-t/2} \leq B' e^{-t/2}.$$

The same argument as before yields $\mathbb{P}(E_1') \leq Be^{-t/2}$. Therefore, combining previous bounds, we get $\mathbb{P}(E) \leq Be^{-t/4}$, which completes the proof of the theorem. $\square$

**4. Some experiments with learning algorithms.** In this section we present some results of the experiments we conducted to test the ability of the new bounds to predict the value of the generalization error of combined classifiers. Unfortunately, the constants in the bounds of Section 2 are not known. More precisely, using the results of the recent work of Massart [22] one can calculate the constants involved in the bounds, but their current values are rather large and, most likely, not optimal. However, many important learning algorithms (such as boosting and bagging) that combine simple classifiers are iterative in nature and it is important to see whether the bounds allow one to predict the shape of the learning curves (the dependence of the generalization error on the number of iterations) correctly. To this end, we just ignore the constants and use in the experiments the quantities $(n^{1-\gamma/2}\hat{\delta}_n(\gamma; f)^\gamma)^{-1}$ (see Example 1) and $\varepsilon_n(f; \hat{\delta}_n(f))$ [see Theorem 3. Actually, the quantity $\varepsilon_n(f; \hat{\delta}_n(f)/2)$ is involved in this bound, but it is easy to see that it is within a constant from $\varepsilon_n(f; \hat{\delta}_n(f))$.] instead of the upper bounds we proved. We will refer to these quantities as the $\gamma$-bound and the $\Delta$-bound, respectively. Incidentally, these quantities did provide upper bounds on the generalization error (or on the test error) in most of our experiments. This suggests that the values of the constants involved in the bounds of Section 2 might actually be moderate (at least in the case when the bounds are applied to several well-known learning algorithms; see also the remark after Corollary 1).

4.1. *Bagging and boosting.* We begin by describing the experiments with two of the most popular techniques for combining the classifiers, namely bagging [5] and the Adaboost algorithm [10]. In both of these methods, there is access to

a learning algorithm called a *base learner*. The base learner is given a training sample $(X_i, Y_i)$, $i = 1, \ldots, n$, and it returns a classifier $h$ from a base class $\mathcal{H}$ that "approximately minimizes" the empirical error $P_n\{yh(x) \le 0\}$ (or properly weighted empirical error).

In the case of bagging, the base learner receives at each iteration $t$, $t = 1, \ldots, T$, an independent bootstrap sample $(\hat{X}_i^{(t)}, \hat{Y}_i^{(t)})$, $i = 1, \ldots, n$, and returns a classifier $h_t \in \mathcal{H}$. The output of bagging is the combined classifier $f := T^{-1} \sum_{t=1}^{T} h_t$ (in other words, bagging makes a decision by majority vote).

In the case of Adaboost, the algorithm assigns at the beginning equal weights $D_1(i) = n^{-1}$, $i = 1, \ldots, n$, to all the training examples and then updates the weights iteratively. Namely, at the $t$th iteration ($t = 1, \ldots, T$) the algorithm calls the base learner that attempts to minimize approximately the weighted training error

$$\varepsilon_t(h) := \sum_{i:h(X_i)\neq Y_i} D_t(i), \qquad h \in \mathcal{H}.$$

The base learner returns a classifier $h_t \in \mathcal{H}$ and its weighted training error $\hat{\varepsilon}_t := \varepsilon_t(h_t)$. The weights are then updated according to the formula

$$D_{t+1}(i) := \frac{D_t(i)}{Z_t}\big(1 + (\beta_t - 1) I_{\{h(X_i)=Y_i\}}\big),$$

where $\beta_t := \hat{\varepsilon}_t/(1 - \hat{\varepsilon}_t)$ and $Z_t$ is the normalizing factor such that $\sum_{i=1}^{t} D_{t+1}(i) = 1$. After $T$ iterations, Adaboost outputs a combined classifier

$$f := \left(\sum_{t=1}^{T} \log \frac{1}{\beta_t}\right)^{-1} \sum_{t=1}^{T} \log \frac{1}{\beta_t} h_t.$$

In all the experiments, we used the set of indicator functions (actually, these functions are rescaled so that they take values in $\{-1, 1\}$) of axis oriented hyperplanes (also known as decision stumps) as base classifiers. That is, $S := \mathbb{R}^d$ and

$$\mathcal{H} = \big\{I_{\{\mathbf{x}\in\mathbb{R}^d : x_i \le c\}}, c \in \mathbb{R},\ i = 1, \ldots, d\big\} \cup \big\{I_{\{\mathbf{x}\in\mathbb{R}^d : x_i \ge c\}}, c \in \mathbb{R},\ i = 1, \ldots, d\big\},$$

where $\mathbf{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d$.

4.2. *Experiments with real and simulated data.* We first describe the experiments with a "toy" problem which is simple enough to allow one to compute exactly the generalization error and other quantities such as the $\gamma$-margins. Namely, we consider a one-dimensional classification problem in which $S = [0, 1]$ and, given a set (or a concept, using the terminology of computer learning) $C_0 \subset S$ which is a finite union of disjoint intervals, the label $y$ is assigned to a point $x \in S$ according to the rule $y = f_0(x)$, where $f_0$ is equal to $+1$ on $C_0$ and to $-1$

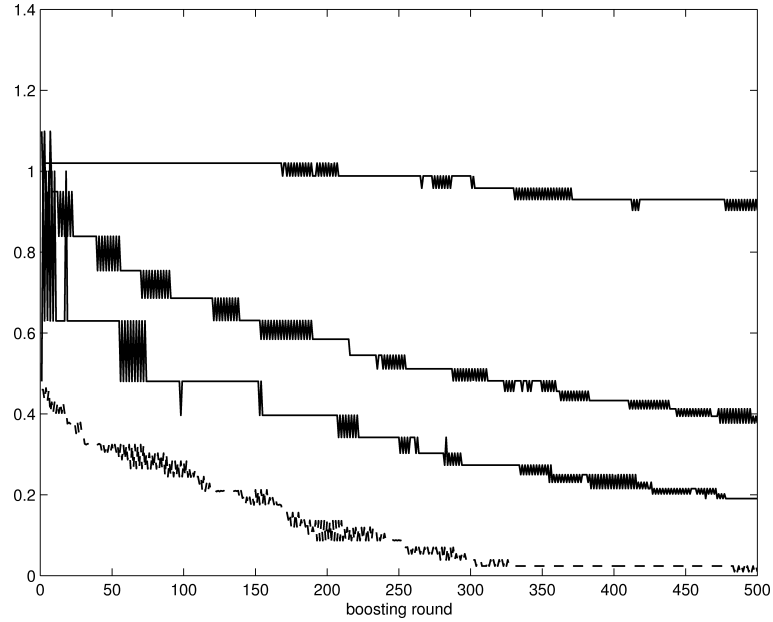


FIG. 2. *Comparison of the generalization error* (*thicker line*) *with* $(n^{1-\gamma/2}\hat{\delta}_n(\gamma; f)^{\gamma})^{-1}$ *for* $\gamma = 1, 0.8$ *and* $2/3$ (*thinner lines, top to bottom*).

on $S \setminus C_0$. We refer to this problem as the *intervals problem* (see also [13]). Note that for the class of decision stumps we have in this case $V(\mathcal{H}) = 2$ (since $\mathcal{H} = \{I_{[0,b]} : b \in [0, 1]\} \cup \{I_{[b,1]} : b \in [0, 1]\}$), and according to the results above the values of $\gamma$ in $[2/3, 1)$ provide valid bounds on the generalization error in terms of $\gamma$-margins. In our experiments, the set $C_0$ was formed by 20 equally spaced intervals and we generated a uniformly distributed sample on $[0, 1]$ of size 1000. We ran Adaboost for 500 rounds (bagging does not work well for this problem) and computed at each round the generalization error of the combined classifier and the quantity $(n^{1-\gamma/2}\hat{\delta}_n(\gamma; f)^{\gamma})^{-1}$ for different values of $\gamma$.

In Figure 2 we plot the generalization error and the bounds for $\gamma = 1,\ 0.8$ and $2/3$ against the iteration of Adaboost. As expected, for $\gamma = 1$ (which corresponds roughly to the bounds in [24]) the bound is very loose and, as $\gamma$ decreases, the bound gets closer to the generalization error. In Figure 3 we show that by reducing further the value of $\gamma$ we get a curve that is even closer to the actual generalization error (although, for $\gamma = 0.2$, it does not provide an upper bound for some of the rounds of Adaboost). This seems to support the conjecture that Adaboost actually generates combined classifiers that belong to a subset of the convex hull of $\mathcal{H}$ with a smaller random entropy than of the whole convex hull. In Figure 4 we plot the ratio $\hat{\delta_n}(\gamma; f)/\delta_n(\gamma; f)$ for $\gamma = 0.4$, $2/3$ and $0.8$ against the boosting iteration. We can see that the ratio is close to 1 in different examples (for a small number of iterations of Adaboost in the first example, the ratio is actually close to 0),

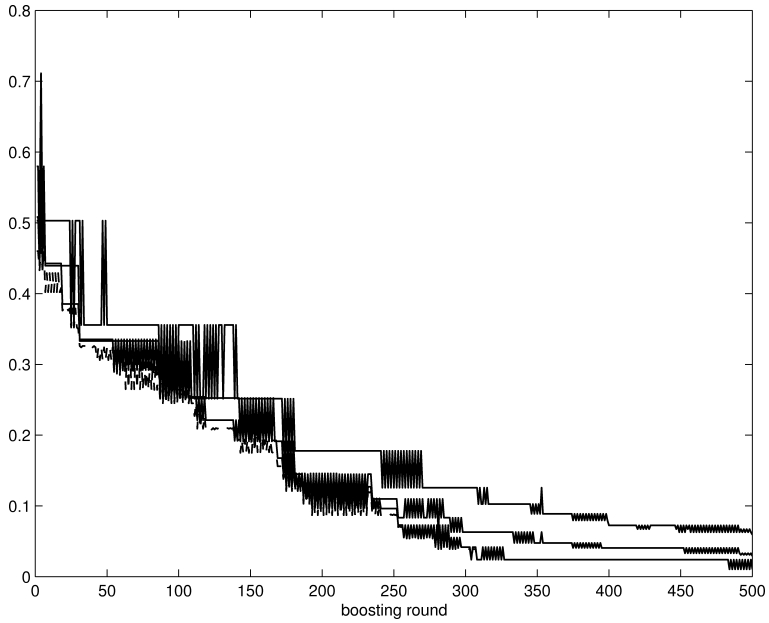


FIG. 3. *Comparison of the generalization error* (*thicker line*) *with* $(n^{1-\gamma/2}\hat{\delta}_n(\gamma; f)^{\gamma})^{-1}$ *for* $\gamma = 0.5, 0.4$ *and* $0.2$ (*thinner lines, top to bottom*).

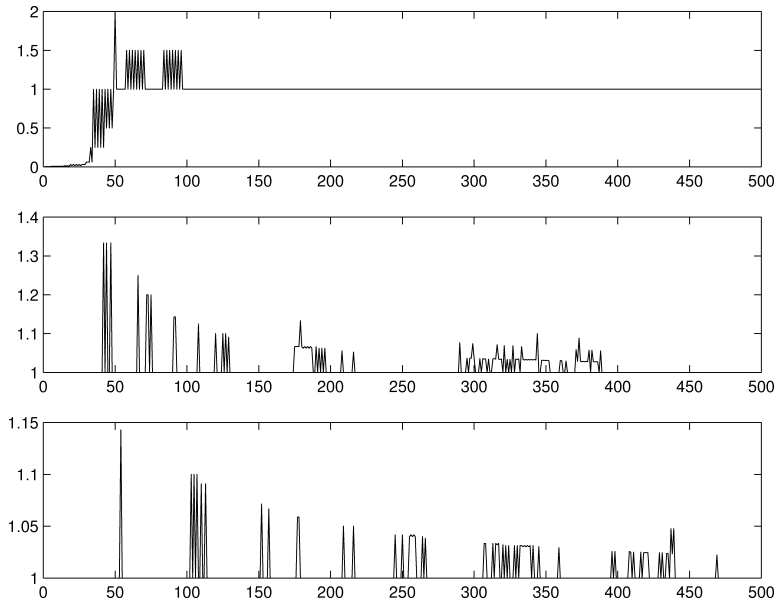


FIG. 4. *Ratio* $\hat{\delta_n}(\gamma; f)/\delta_n(\gamma; f)$ *vs. boosting round for* $\gamma = 0.4, 2/3$ *and* $0.8$ (*top to bottom*).

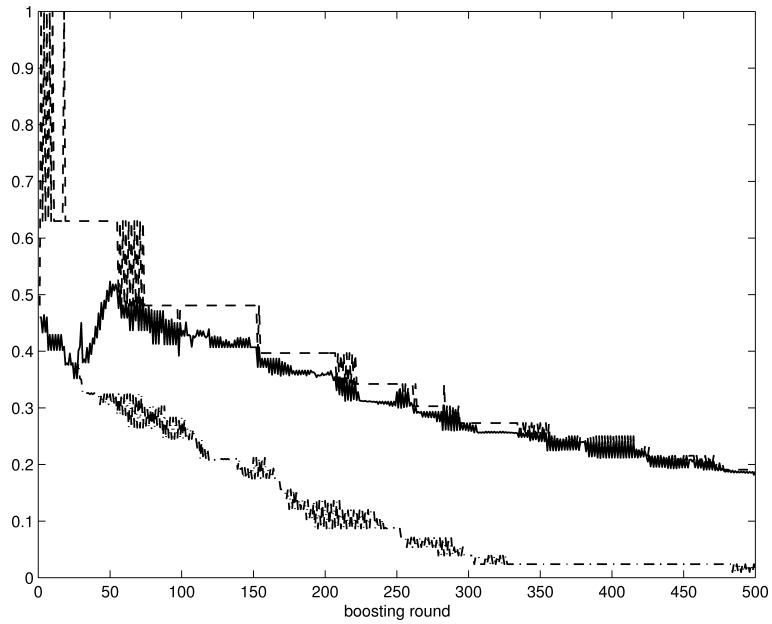


FIG. 5. *Generalization error and bounds vs. number of classifiers for the intervals problem for sample sizes of* 1000. *Test error* (*dashed lines*), $\gamma$*-margin bound with* $\gamma = 2/3$ (*dot-dashed line*) *and* $\Delta$*-bound* (*solid line*).

indicating that the value of the constant $\bar{A}$ in the bound (2.7) might be close to 1 (at least, this seems to be true in the case of classifiers produced by Adaboost for large sample sizes).

In Figure 5 we compare the $\gamma$-bound and the $\Delta$-bound obtained for this problem for a sample size of 1000. We can see that the $\Delta$-bound has two regimes. In the first regime, the effect of the $\Delta$-dimension is dominant and the bound tracks almost exactly the generalization error, giving a definite improvement over the $\gamma$-bound. In the second regime, the bound starts increasing until it reaches the curve of the $\gamma$-bound. This behavior can be explained by examining the expression being minimized in the computation of the bound:

$$\underbrace{\frac{d(f;\Delta)}{n}\left(\log\frac{1}{\delta}+\log\frac{ne^2}{d(f;\Delta)}\right)}_{\text{I}}+\underbrace{\left(\frac{\Delta}{\delta}\right)^{2\alpha/(\alpha+2)}n^{-2/(\alpha+2)}}_{\text{II}}. \tag{4.1}$$

It is easy to see that this expression will be close to the $\gamma$-bound when the second term is dominant and, in fact, becomes the $\gamma$-bound when $\Delta = 1$ (which, apparently, is the case in our experiments when the number of classifiers in the convex combination becomes large).

We also computed the bounds for more complex simulated data sets as well as for real data sets in which the same type of behavior was observed. We show the

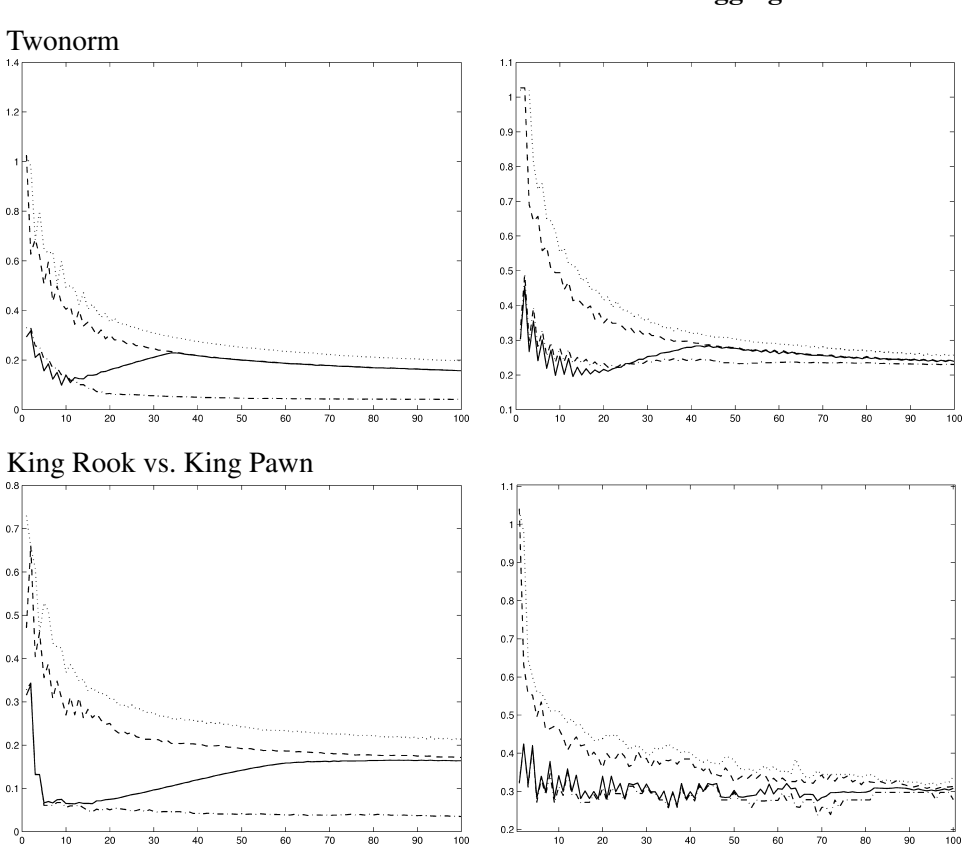


FIG. 6. *Test error and bounds vs. number of classifiers. Test error* (*dot-dashed lines*), $\gamma$*-margin bound with* $\gamma = 1$ (*dotted lines*), $\gamma = \gamma_{\min}$ (*dashed lines*) *and the* $\Delta$*-bound* (*solid lines*).

results for the so-called Twonorm data set and the King Rook vs. King Pawn data set (Figure 6), which are well-known examples in computer learning literature. The Twonorm data set (taken from [6]) is a simulated 20-dimensional data set in which positive and negative training examples are drawn from the multivariate normal distributions with unit covariance matrix centered at $(2/\sqrt{20}, \ldots, 2/\sqrt{20})$ and $(-2/\sqrt{20}, \ldots, -2/\sqrt{20})$, respectively. The King Rook vs. King Pawn data set is a real data set from the UCI Irvine repository [4]). It is a 36-dimensional data set with a sample size of 3196.

As before, we used the decision stumps as base classifiers. An upper bound on $V(\mathcal{H})$ for the class $\mathcal{H}$ of decision stumps in $\mathbb{R}^d$ is given by the smallest $n$ such that $2^{n-1} \geq (n-1)d + 1$. We computed the $\Delta$-bound and the $\gamma$-bounds for $\gamma = 1$ and for the smallest $\gamma$ allowed in Example 1 ($\gamma_{\min}$). For the Twonorm data set, we estimated the generalization error by computing the empirical error on an indepedently generated set of 20,000 observations. For the King Rook vs. King

Pawn data set, we randomly selected 90% of the data for training and used the remaining 10% to compute the test error. The experiments were averaged over 10 repetitions.

4.3. *Toward algorithms balancing the dimensionality and the margins.* The connection between increasing the margins and reducing the generalization error has led to the development of several algorithms for designing and improving combined classifiers based on optimizing margin cost functions. The examples include DOOM [20], DOOM2 [21], DOOM-LP [19], GeoLev [9] and LP-Adaboost [11]. The results in this paper motivate the development of algorithms that take into account the approximate dimensions of combined classifiers along with their margins.

We discuss below the algorithm DOOM-LP, which was designed to optimize a piecewise linear cost function of the margins by solving a sequence of linear programs. Incidentally, this algorithm also tends to reduce the dimension of the combined classifier. To describe the algorithm, define $\varphi(u) := I_{(-\infty,0]}(u) + (1-u)I_{(0,1]}(u)$ and let $\varphi_\delta(u) := \varphi(u/\delta)$. Let $\mathcal{H}$ be a base class and let $\mathcal{F} := \operatorname{conv}(\mathcal{H})$. It was proved in Koltchinskii and Panchenko [17] that with probability at least $1 - 2\exp\{-2t^2\}$ the quantity

$$\inf_{\delta\in[0,1]}\left[P_n\varphi_\delta(yf(x)) + \frac{8}{\delta}\mathbb{E}\hat{R}_n(\mathcal{H}) + \left(\frac{\log\log_2(2\delta^{-1})}{n}\right)^{1/2}\right] + \frac{t}{\sqrt{n}}$$

is an upper bound on the generalization error $P\{yf(x) \le 0\}$ of *any* classifier $f \in \mathcal{F}$. Recall that $\hat{R}_n(\mathcal{H})$ is the Rademacher complexity of the class $\mathcal{H}$. If $\mathcal{H}$ is a VC-class, then $\mathbb{E}\hat{R}_n(\mathcal{H}) \le Cn^{-1/2}$ with a constant $C$ depending on the VC-dimension of $\mathcal{H}$. The idea of the algorithm DOOM-LP is to minimize the above bound with respect to $f \in \mathcal{F}$ and $\delta \in [0,1]$ to find a classifier $\hat{f}$ with a reasonably small generalization error. More precisely, the algorithm receives a finite number of base classifiers $h_1, \dots, h_T$ along with their weights and attempts to redistribute the weights to minimize the bound.

For a fixed value of $\delta$ and fixed classifiers $h_1, \dots, h_T$, the minimization with respect to $f = \sum_{k=1}^T w_k h_k \in \mathcal{F}$ consists of finding the weights $w_k$, $\sum_{k=1}^T w_k = 1$, that minimize the quantity

$$P_n\varphi_\delta(yf(x)) = \frac{1}{n}\sum_{i=1}^n \varphi_\delta\left(Y_i\sum_{k=1}^T w_k h_k(X_i)\right). \tag{4.2}$$

For a given combined classifier $f = \sum_{k=1}^T w_k h_k \in \mathcal{F}$, define sets $S_-$, $S_l$ and $S_0$ as

$$\begin{aligned} S_- &= \{i : Y_i f(X_i) \le 0\}, \\ S_l &= \{i : 0 \le Y_i f(X_i) \le \delta\}, \\ S_0 &= \{i : Y_i f(X_i) \ge \delta\}. \end{aligned}$$

TABLE 1
*DOOM-LP algorithm*

**Require:** Initial weight vector $\mathbf{w}$, margins $\{M_i\}_{i=1}^n$
  {Initialize the partition}
  $S_- = \{i : M_i \le 0\}$
  $S_l = \{i : 0 \le M_i \le \delta\}$
  $S_0 = \{i : M_i \ge \delta\}$
  **repeat**
    $C_{\min} = \sum_{k=1}^T b_k w_k$
    **if** $|S_l| \ge 1$ **then**
      {Compute optimal solution for a new partition}
      $\mathbf{w} = \text{LPSolve}(\mathbf{w}, S_-, S_l, S_0)$
      Compute new margins $\{M_i\}_{i=1}^n$
      {Update sets}
      $S_- = S_- \cup \{i : i \in S_l, M_i = 0\} - \{i : i \in S_-, M_i = 0\}$
      $S_l = S_l \cup \{i : i \in S_-, M_i = 0\} \cup \{i : i \in S_0, M_i = \delta\}$
          $- \{i : i \in S_l, M_i = 0 \text{ or } M_i = \delta\}$
      $S_0 = S_0 \cup \{i : i \in S_l, M_i = \delta\} - \{i : i \in S_0, M_i = \delta\}$
      $C = \sum_{k=1}^T b_k w_k$
    **else**
      Terminate and return current $\mathbf{w}$
    **end if**
  **until** $C \ge C_{\min}$

Finding the weight vector that approximately minimizes $P_n\varphi_\delta(yf(x))$ for a fixed current partition $(S_-, S_l, S_0)$ can be easily posed as a linear programming problem. DOOM-LP searches for an approximate local minimum of $P_n\varphi_\delta(yf(x))$ by solving this linear program and moving to a neighboring partition by "flipping" the margins that fall in the intersection of two of the sets $S_-$, $S_l$ or $S_0$ from the set to which they currently belong to another one in the hope that with the constraints determined by the new partition the objective function can be reduced. The idea is similar in spirit to the sweeping hinge algorithm proposed by Hush and Horne [12]. The algorithm converges when the value of the minimum in two neighboring partitions is the same (see Table 1). We use the following notations in the description of the algorithm: $b_k = -\sum_{i \in S_l} Y_i h_k(X_i)$ and $M_i = Y_i f(X_i)$, where $f = \sum_k w_k h_k$.

Written in a standard form, the linear program solved by DOOM-LP at each iteration involves $T + n + |S_l| + 1$ variables ($T$ weights plus slack and surplus variables) and $n + |S_l| + 1$ equality constraints. It follows from the basic results on

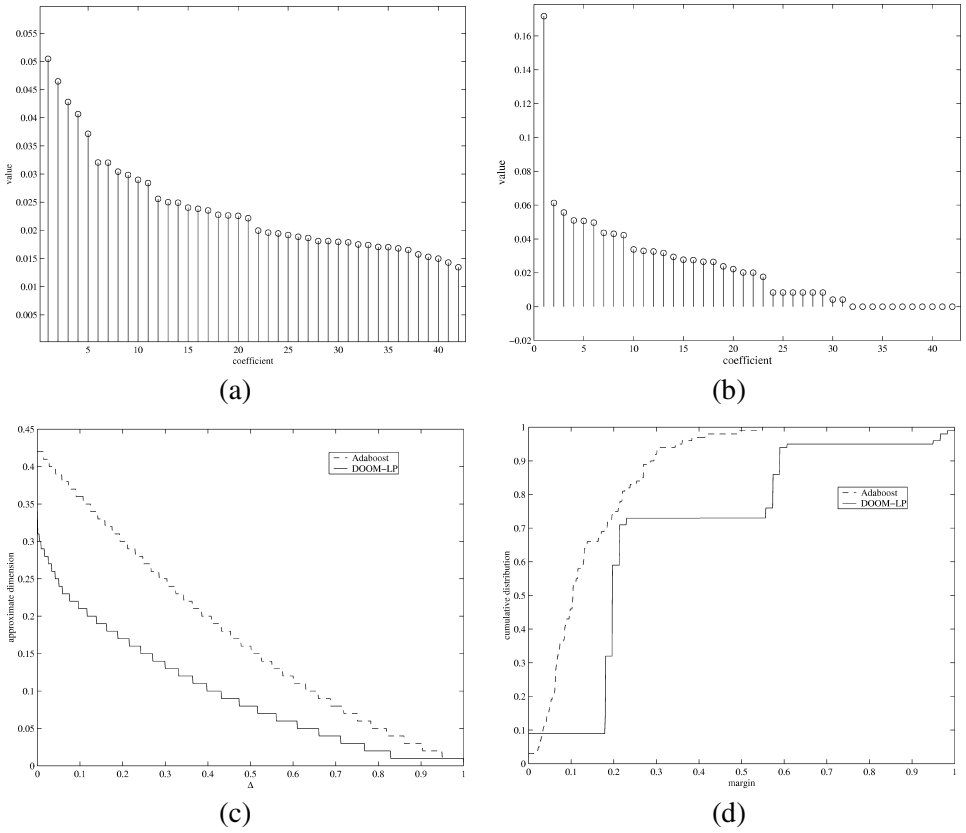


FIG. 7. *Results of running DOOM-LP on the classifier produced by Adaboost for the King Rook vs. King Pawn data set.* (a) *Adaboost sorted coefficients*; (b) *DOOM-LP sorted coefficients*; (c) *approximate* $\Delta$*-dimensions*; (d) *cumulative margin distributions.*

linear programming that if there is an optimal feasible solution and the constraint matrix is full rank, then there exists an optimal feasible solution with at most $n + |S_l| + 1$ nonzero variables. Furthermore, if the simplex method is used to solve the linear program, a solution of this type is always found. We have observed in experiments that many of the variables that are set to zero in the solution are weights and that DOOM-LP tends to reduce the $\Delta$-dimension of the classifier.

We have used DOOM-LP to improve the generalization error of combined classifiers produced by Adaboost by redistributing the weights of the base classifiers in a convex combination. An example of dimensionality reduction by DOOM-LP is illustrated in Figure 7.

It might be interesting to design new algorithms with explicit penalization for high dimensionality in the optimization procedure. For instance, assuming that the initial weights $w_t^{(0)}, t = 1, \ldots, T$, are arranged in decreasing order, one can add to the target function of the linear program a term $\sum_{t=1}^{T} a_t w_t$, where $\{a_t, t \geq 1\}$ is

an increasing sequence of positive numbers. One can also consider entropy type penalties of the form $\sum_{t=1}^{T} w_t \log(1/w_t)$ (in this case, of course, the optimization is no longer a linear programming problem).

## REFERENCES


[1] ANTHONY, M. and BARTLETT, P. (1999). *Neural Network Learning*: *Theoretical Foundations*. Cambridge Univ. Press.

[2] BARTLETT, P. (1998). The sample complexity of pattern classification with neural networks: The size of the weights is more important than the size of the network. *IEEE Trans. Inform. Theory* **44** 525–536.

[3] BARTLETT, P., BOUCHERON, S. and LUGOSI, G. (2001). Model selection and error estimation. *Machine Learning* **48** 85–113.

[4] BLAKE, C. L. and MERZ, C. J. (1998). UCI repository of machine learning databases. Available at http://www.ics.uci.edu/∼mlearn/MLRepository.html.

[5] BREIMAN, L. (1996). Bagging predictors. *Machine Learning* **26** 123–140.

[6] BREIMAN, L. (1998). Arcing classifiers. *Ann. Statist.* **26** 801–849.

[7] CORTES, C. and VAPNIK, V. (1995). Support vector networks. *Machine Learning* **24** 273–297.

[8] DEVROYE, L., GYÖRFI, L. and LUGOSI, G. (1996). *A Probabilistic Theory of Pattern Recognition*. Springer, New York.

[9] DUFFY, N. and HELMBOLD, D. (1999). A geometric approach to leveraging weak learners. *Computational Learning Theory. Lecture Notes in Comput. Sci.* 18–33. Springer, New York.

[10] FREUND, Y. and SCHAPIRE, R. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. System Sci.* **55** 119–139.

[11] GROVE, A. and SCHUURMANS, D. (1998). Boosting in the limit: Maximizing the margin of learned ensembles. In *Proceedings of the Fifteenth National Conference on Artificial Intelligence* 692–699. AAAI Press, Menlo Park, California.

[12] HUSH, D. and HORNE, B. (1998). Efficient algorithms for function approximation with piecewise linear sigmoids. *IEEE Trans. Neural Networks* **9** 1129–1141.

[13] KEARNS, M., MANSOUR, Y., NG, A. and RON, D. (1997). An experimental and theoretical comparison of model selection methods. *Machine Learning* **27** 7–50.

[14] KOLTCHINSKII, V. (2001). Bounds on margin distributions in learning problems. Preprint. Available at http://www.math.unm.edu/∼panchenk/.

[15] KOLTCHINSKII, V. (2001). Rademacher penalties and structural risk minimization. *IEEE Trans. Inform. Theory* **47** 1902–1914.

[16] KOLTCHINSKII, V. and PANCHENKO, D. (2000). Rademacher processes and bounding the risk of function learning. In *High Dimensional Probability II* (D. Mason, E. Giné and J. Wellner, eds.) 443–457. Birkhäuser, Boston.

[17] KOLTCHINSKII, V. and PANCHENKO, D. (2002). Empirical margin distribution and bounding the generalization error of combined classifiers. *Ann. Statist.* **30** 1–50.

[18] KOLTCHINSKII, V., PANCHENKO, D. and LOZANO, F. (2001). Bounding the generalization error of neural networks and combined classifiers. In *Proceedings of Thirteenth International Conference on Advances in Neural Information Processing Systems* 245–251. MIT Press.

[19] LOZANO, F. and KOLTCHINSKII, V. (2002). Direct optimization of simple cost functions of the margin. In *Proceedings of the First International NAISO Congress on Neuro Fuzzy Technologies*. Academic Press, Amsterdam.

[20] MASON, L., BARTLETT, P. and BAXTER, J. (2000). Improved generalization through explicit optimization of margins. *Machine Learning* **38** 243–255.

[21] MASON, L., BAXTER, J., BARTLETT, P. and FREAN, M. (2000). Functional gradient techniques of combining hypotheses. In *Advances in Large Margin Classifiers* (A. J. Smol, P. Bartlett, B. Schölkopf and C. Schuurmans, eds.) 221–246. MIT Press.

[22] MASSART, P. (2000). About the constants in Talagrand's concentration inequalities for empirical processes. *Ann. Probab.* **28** 863–884.

[23] PISIER, G. (1981). Remarques sur un résultat non publié de B. Maurey. In *Séminaire d'Analyse Fonctionelle 1980–1981*, Exposé 5. Ecole Polytechnique, Palaiseau.

[24] SCHAPIRE, R. E., FREUND, Y., BARTLETT, P. and LEE, W. S. (1998). Boosting the margin: A new explanation for the effectiveness of voting methods. *Ann. Statist.* **26** 1651–1687.

[25] TALAGRAND, M. (1996). New concentration inequalities in product spaces. *Invent. Math.* **126** 505–563.

[26] VAN DER VAART, A. W. and WELLNER, J. (1996). *Weak Convergence of Empirical Processes with Applications to Statistics*. Springer, New York.

[27] VAPNIK, V. (1998). *Statistical Learning Theory*. Wiley, New York.

V. KOLTCHINSKII
D. PANCHENKO
DEPARTMENT OF MATHEMATICS AND STATISTICS
UNIVERSITY OF NEW MEXICO
ALBUQUERQUE, NEW MEXICO 87131-1141
E-MAIL: vlad@math.unm.edu
panchenk@math.unm.edu

F. LOZANO
DEPARTMENT OF ELECTRONIC ENGINEERING
UNIVERSIDAD JAVERIANA
BOGOTA
COLOMBIA
E-MAIL: fernando.lozano@javeriana.edu.co