The Annals of Statistics
2002, Vol. 30, No. 1, 1–50

EMPIRICAL MARGIN DISTRIBUTIONS AND BOUNDING THE GENERALIZATION ERROR OF COMBINED CLASSIFIERS

BY V. KOLTCHINSKII[1] AND D. PANCHENKO[2]

University of New Mexico

Dedicated to A.V. Skorohod on his seventieth birthday

We prove new probabilistic upper bounds on generalization error of complex classifiers that are combinations of simple classifiers. Such combinations could be implemented by neural networks or by voting methods of combining the classifiers, such as boosting and bagging. The bounds are in terms of the empirical distribution of the margin of the combined classifier. They are based on the methods of the theory of Gaussian and empirical processes (comparison inequalities, symmetrization method, concentration inequalities) and they improve previous results of Bartlett (1998) on bounding the generalization error of neural networks in terms of ℓ_1-norms of the weights of neurons and of Schapire, Freund, Bartlett and Lee (1998) on bounding the generalization error of boosting. We also obtain rates of convergence in Lévy distance of empirical margin distribution to the true margin distribution uniformly over the classes of classifiers and prove the optimality of these rates.

1. Introduction. Let (X, Y) be a random couple, where X is an instance in a space S and $Y \in \{-1, 1\}$ is a label. Let $\mathcal{G}$ be a set of functions from S into $\mathbb{R}$. For $g \in \mathcal{G}$, $\operatorname{sign}(g(X))$ will be used as a predictor (a classifier) of the unknown label Y. If the distribution of (X, Y) is unknown, then the choice of the predictor is based on the training data $(X_1, Y_1), \ldots, (X_n, Y_n)$ that consists of n i.i.d. copies of (X, Y). The goal of learning is to find a predictor $\hat{g} \in \mathcal{G}$ (based on the training data) whose *generalization* (*classification*) error $\mathbb{P}\{Y\hat{g}(X) \le 0\}$ is small enough. In this paper, our main concern is to find reasonably good probabilistic upper bounds on the generalization error. The standard approach to this problem was developed in seminal papers of Vapnik and Chervonenkis in the 70s and 80s [see Vapnik (1998), Devroye, Györfi and Lugosi (1996), Vidyasagar (1997)] and it is based on bounding the difference between the generalization error $\mathbb{P}\{Yg(X) \le 0\}$ and the training error

$$n^{-1} \sum_{j=1}^{n} I_{\{Y_j g(X_j) \le 0\}}$$

Received February 2001; revised June 2001.

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

[2]Supported in part by Boeing Computer Services Grant 3-48181.

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

Key words and phrases. Generalization error, combined classifier, margin, empirical process, Rademacher process, Gaussian process, neural network, boosting, concentration inequalities.

uniformly over the whole class $\mathcal{G}$ of classifiers g. These bounds are expressed in terms of data dependent entropy characteristics of the class of sets $\{\{(x, y) : yg(x) \le 0\} : g \in \mathcal{G}\}$ or, frequently, in terms of the so called VC-dimension of the class. It happened, however, that in many important examples (for instance, in neural network learning) the VC-dimension of the class can be very large, or even infinite, and that makes impossible the direct application of Vapnik–Chervonenkis type of bounds. Recently, several authors [see Bartlett (1998), Schapire, Freund, Bartlett and Lee (1998), Anthony and Bartlett (1999)] suggested another class of upper bounds on generalization error that are expressed in terms of the empirical distribution of *the margin* of the predictor (the classifier). The margin is defined as the product $Y\hat{g}(X)$. The bounds in question are especially useful in the case of the classifiers that are combinations of simpler classifiers (that belong, say, to a class $\mathcal{H}$). One of the examples of such classifiers is provided by neural networks. Other examples are given by the classifiers obtained by boosting, bagging and other voting methods of combining the classifiers. The bounds in terms of margins are also of interest in application to generalization performance of support vector machines, Cortes and Vapnik (1995), Vapnik (1998), Bartlett and Shawe-Taylor (1999). The upper bounds have the form (up to some extra terms)

$$\inf_{\delta>0}\left[n^{-1}\sum_{j=1}^{n} I_{\{Y_j\hat{g}(X_j)\le\delta\}} + C(\mathcal{G})\phi(\delta)\frac{C(\mathcal{H})}{\sqrt{n}}\right],$$

where $C(\mathcal{G})$ is a constant depending on the class $\mathcal{G}$ (in other words, on the method of combining the simple classifiers), ϕ is a decreasing function such that $\phi(\delta) \to \infty$ as $\delta \to 0$ [often, for instance, $\phi(\delta) = \frac{1}{\delta}$], and $C(\mathcal{H})$ is a constant depending on the class $\mathcal{H}$ (in particular, on the VC-dimension, or some type of entropy characteristics of the class).

It was observed in experiments that classifiers produced by such methods as boosting tend to have rather *large margin* of correctly classified examples. This allows one to choose a relatively large value of δ in the above bound without increasing substantially the value of the empirical distribution function of the margin (which is the first term of the bound) comparing with the training error. For large enough δ, the second term becomes small, which ensures a reasonably small value of the infimum. This allowed the above mentioned authors to explain partially (at least at qualitative level) a very good generalization performance of voting and some other methods of combining simple classifiers observed in many experiments. This also motivated the development of the methods of combining the classifiers based on explicit optimization of the penalized average cost function of the margins; see Mason, Bartlett and Baxter (1999) and Mason, Baxter, Bartlett and Frean (2000).

Despite the fact that previously developed bounds provide some explanations of the generalization performance of complex classifiers, it was actually acknowledged by Bartlett (1998) and Schapire, Freund, Bartlett and Lee (1998) that the

bounds in question have not reached their final form yet and more research is needed to understand better the probabilistic nature of these bounds. This becomes especially important because of the growing number of boosting type methods [see Friedman, Hastie and Tibshirani (2000) and Friedman (1999)] for which a comprehensive theory is yet to be developed. The methods of proof developed by Bartlett (1998) are based on the so called fat-shattering dimensions of function classes and on the extension of Vapnik–Chervonenkis type inequalities to such dimensions. The method of Schapire, Freund, Bartlett and Lee (1998) exploits the fact that the complex classifiers are *convex combinations* of base classifiers (these authors suggest also an extension of their method to the classes of functions for which there exist so called ε-sloppy θ-covering). The use of these methods in the case of general cost functions of the margins poses some difficulties [see Mason, Bartlett and Baxter (1999)].

In this paper, we develop a new approach that allows us to improve and better understand some of the previously known bounds. Our method is based on the general results of the theory of Gaussian, Rademacher and empirical processes [such as comparison inequalities, e.g., Slepian's Lemma, symmetrization and random multipliers inequalities, concentration inequalities, see Ledoux and Talagrand (1991), van der Vaart and Wellner (1996) and Dudley (1999)]. We give the bounds in terms of general functions of the margins, satisfying a Lipschitz condition. They can be readily applied to the classifiers based on explicit optimization of margin cost functions [such as in the paper of Mason, Bartlett and Baxter (1999)]. In the case of Bartlett's bounds for feedforward neural networks in terms of the ℓ_1-norms of the weights of the neurons [see Bartlett (1998) and also Fine (1999)], the improvement we got is substantial. In Bartlett's bounds the constant $C(\mathcal{G})$ is of the order $(AL)^{l(l+1)/2}$, where A is an upper bound on the ℓ_1-norms of the weights of neurons, L is the Lipschitz constant of the sigmoids, and l is the number of layers of the network. Also, in his bound $\phi(\delta) = \frac{1}{\delta^l}$. We obtained in a similar context $C(\mathcal{G})$ of the order $(AL)^l$ with $\phi(\delta) = \frac{1}{\delta}$.

Based on our bounds, we developed a method of complexity penalization of the training error of neural network learning with penalties defined as functionals of the weights of neurons and prove oracle inequalities showing some form of optimality of this method.

We also obtained general rates of convergence of the empirical margin distributions to the theoretical one in the Lévy distance. Namely, we proved that the empirical margin distribution converges to the true margin distribution with probability 1 uniformly over the class $\mathcal{G}$ of classifiers if and only if the class $\mathcal{G}$ is Glivenko–Cantelli. Moreover, if $\mathcal{G}$ is a Donsker class, then the rate of convergence in Lévy distance is $O(n^{-1/4})$. Faster rates [up to $O(n^{-1/2})$] are possible under some assumptions on random entropies of the class $\mathcal{G}$. We give some examples, showing the optimality of these rates.

We improved previously known bounds on generalization error of convex combinations of classifiers. In particular, our results in Section 3 imply that

if the random ε-entropy of the class $\mathcal{G}$ grows as $\varepsilon^{-\alpha}$ for $\alpha \in (0,2)$, then the generalization error of any classifier from $\mathcal{G}$ with zero training error is bounded from above with very high probability by the quantity

$$\frac{C}{n^{2/(2+\alpha)}\hat{\delta}^{2\alpha/(2+\alpha)}},$$

where $\hat{\delta}$ is the minimal classification margin of the training examples and C is a constant. The previously known result of Schapire, Freund, Bartlett and Lee (1998) gives [up to logarithmic factors, for $\mathcal{G} = \operatorname{conv}(\mathcal{H})$, $\mathcal{H}$ being a VC-class] the bound $O(\frac{1}{n^{1/2}\hat{\delta}})$ which corresponds to the worst choice of α ($\alpha = 2$). We introduce in Section 3 more subtle notions of γ-margin $\delta_n(\gamma; g)$ and empirical γ-margin $\hat{\delta}_n(\gamma; g)$ (parametrized by $\gamma \in (0,1]$) of a classifier g. These quantities allow us to obtain similar upper bounds on generalization error of the form

$$\frac{C_\gamma}{n^{1-\gamma/2}\hat{\delta}_n^{\gamma}(\gamma; g)},$$

in the case when the training error of the classifier g is not necessarily equal to 0. We call the quantity

$$\frac{1}{n^{1-\gamma/2}\hat{\delta}_n^{\gamma}(\gamma; g)}$$

the γ-bound of g. It follows from the definitions given in Section 3 that the γ-bounds decrease when γ decreases from 1 to 0. We prove that for any $\gamma \geq \frac{2\alpha}{2+\alpha}$ with very high probability the γ-bounds are indeed upper bounds on the generalization error (up to a multiplicative constant C_γ).

The proof of the bounds of this type is based on the powerful concentration inequalities of Talagrand (1996a, b). For small α, the bound may become arbitrarily close to the rate $O(n^{-1})$, which is known to be the best possible convergence rate in the zero error case. In the case of convex combinations of classifiers from a VC-class $\mathcal{H}$, one can choose $\alpha = 2(V-1)/V$, where V is the VC-dimension of the class $\mathcal{H}$, which improves the previously known bounds for convex combinations of classifiers. We believe that these results can be of importance in some other learning problems [such as support vector learning, see Vapnik (1998)].

Koltchinskii, Panchenko and Lozano (2000a, b) studied the behavior of the γ-bounds and some other bounds of similar type in a number of experiments with AdaBoost and other methods of combining classifiers. We have run AdaBoost for a number of rounds with a weak learner that outputs simple classifiers (e.g., decision stumps) from a small VC-class. In some of the experiments, we dealt with a toy learning problem ("intervals problem") for which it was easy to compute the generalization error precisely. In other cases, we dealt with real data from UCI Irvine repository [see Blake and Merz (1998)] and we estimated the generalization

error based on test samples. In both cases, we computed the γ-margins and the corresponding γ-bounds based on the training data and compared the bounds with the generalization error (or with the test error). We give here only a short summary of the results of these experiments (and some related theoretical results). The details are given in Koltchinskii, Panchenko and Lozano (2000a, b).

One of the goals of the experiments was to determine the value of the constant C_γ involved in the γ-margin bounds on generalization error. The results of Section 3 of this paper show that such a constant exists. Its size, however, is related to a hard problem of optimizing the constants involved in Talagrand's concentration inequality for empirical processes that was used in the proofs. Our experiments showed that the choice $C_\gamma = 1$ worked rather well in the bounds of this type. They also showed that the γ-bounds did improve the previously known bounds on generalization error of AdaBoost. The improvement was significant when the VC-dimension of the base class was small and, hence, the parameter γ could be chosen much smaller than 1. Figure 1 shows a typical result of the experiments.

We also observed that the ratios $\frac{\hat{\delta}_n(\gamma;g)}{\delta_n(\gamma;g)}$ of the empirical γ-margins to the true γ-margins of classifiers g produced by AdaBoost were surprisingly close to 1 (at least for large sample sizes). The results of Section 3 imply that, with high

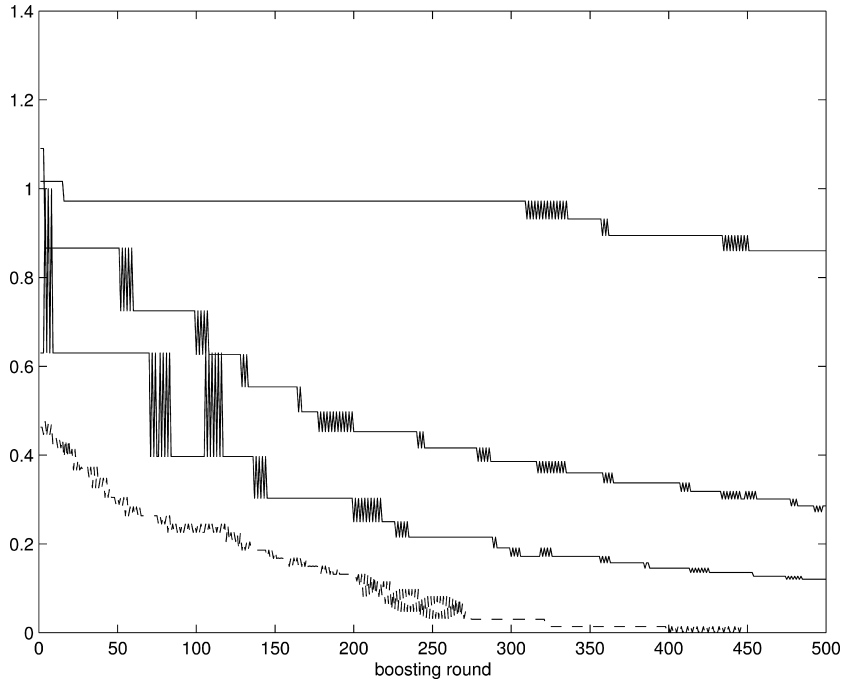

FIG. 1. *Comparison of the generalization error* (*dashed line*) *with the* γ*-bounds for* $\gamma = 1, 0.8$ *and* $2/3$ (*solid lines, top to bottom*).

probability, these ratios are bounded away from 0 and from ∞ uniformly in $g \in \mathcal{G}$ for any $\gamma \geq \frac{2\alpha}{2+\alpha}$. Recently, the first author proved that the ratios do converge to 1 uniformly in $g \in \mathcal{G}$ a.s. as $n \to \infty$ for $\gamma > \frac{2\alpha}{2+\alpha}$ (the example was also given showing that for $\gamma = \frac{2\alpha}{2+\alpha}$ the ratios do not necessarily converge to 1 and for $\gamma < \frac{2\alpha}{2+\alpha}$ they can tend to ∞). The closeness of the ratios to 1 explains why the γ-bounds are valid with $C_\gamma = 1$.

In the case of the classifiers obtained in consecutive rounds of AdaBoost, the γ-bounds hold even for values of γ that are substantially smaller than the threshold $\frac{2\alpha}{2+\alpha}$ given by the theory. It might be related to the fact that the threshold is based on the bounds on the entropy of the *whole* convex hull of the base class $\mathcal{H}$. On the other hand, AdaBoost and other algorithms of this type output classifiers that belong to a subset $\mathcal{G} \subset \mathrm{conv}(\mathcal{H})$ whose entropy might be much smaller than the entropy of the whole convex hull. Because of this, it is important to develop adaptive versions of the margin type bounds on generalization error that take into account the complexity of the classifiers output by learning algorithms as well as their empirical margins. A possible approach to this problem was developed in Koltchinskii, Panchenko and Lozano (2000a).

It should be mentioned that this paper describes only one of a number of growing areas of applications of probability to computer learning problems. Some other important examples of such applications are given in Yukich, Stinchcombe and White (1995), Barron (1991a, b), Barron, Birgé and Massart (1999), Talagrand (1998) and Freund (1995, 1999).

2. Probabilistic bounds for general function classes in terms of Gaussian and Rademacher complexities. Let $(S, \mathcal{A}, P)$ be a probability space and let $\mathcal{F}$ be a class of measurable functions from $(S, \mathcal{A})$ into $\mathbb{R}$. [Later, in Sections 5, 6 we will replace S by $S \times \{-1, 1\}$, considering labeled observations; at this point, it is not important.] Let $\{X_k\}$ be a sequence of i.i.d. random variables taking values in $(S, \mathcal{A})$ with common distribution P. We assume that this sequence is defined on a probability space $(\Omega, \Sigma, \mathbb{P})$. Let P_n be the empirical measure based on the sample $(X_1, \ldots, X_n)$,

$$P_n := n^{-1} \sum_{i=1}^{n} \delta_{X_i},$$

where δ_x denotes the probability distribution concentrated at the point x. We will denote $Pf := \int_S f \, dP$, $P_n f := \int_S f \, dP_n$, etc.

In what follows, $\ell^\infty(\mathcal{F})$ denotes the Banach space of uniformly bounded real valued functions on $\mathcal{F}$ with the norm

$$\|Y\|_{\mathcal{F}} := \sup_{f \in \mathcal{F}} |Y(f)|.$$

We assume throughout the paper that $\mathcal{F}$ satisfies standard measurability assumptions of the theory of empirical processes [see Dudley (1999) and van

der Vaart and Wellner (1996)] (for simplicity, one can assume that $\mathcal{F}$ is countable, but this, of course, is not necessary).

Our goal in this section is to construct data dependent upper bounds on the probability $P\{f \le 0\}$ and on the difference $|P_n\{f \le 0\} - P\{f \le 0\}|$ that hold for all $f \in \mathcal{F}$ with high probability. These inequalities will be used in the next sections to upper bound the generalization error of combined classifiers. The bounds will depend on some measures of "complexity" of the class $\mathcal{F}$ which will be introduced next.

Define

$$G_n(\mathcal{F}) := \mathbb{E}\left\| n^{-1}\sum_{i=1}^n g_i\delta_{X_i}\right\|_{\mathcal{F}},$$

where $\{g_i\}$ is a sequence of i.i.d. standard normal random variables, independent of $\{X_i\}$. [Actually, it is common to assume that $\{g_i\}$ is defined on a separate probability space $(\Omega_g, \Sigma_g, \mathbb{P}_g)$ and that the basic probability space is now $(\Omega \times \Omega_g, \Sigma \times \Sigma_g, \mathbb{P} \times \mathbb{P}_g)$.] We will call $n \mapsto G_n(\mathcal{F})$ *the Gaussian complexity function* of the class $\mathcal{F}$.

Similarly, we define

$$R_n(\mathcal{F}) := \mathbb{E}\left\| n^{-1}\sum_{i=1}^n \varepsilon_i\delta_{X_i}\right\|_{\mathcal{F}},$$

where $\{\varepsilon_i\}$ is a sequence of i.i.d. Rademacher (taking values $+1$ and -1 with probability $1/2$ each) random variables, independent of $\{X_i\}$. We will call $n \mapsto R_n(\mathcal{F})$ *the Rademacher complexity function* of the class $\mathcal{F}$.

One can find in the literature [see, e.g., van der Vaart and Wellner (1996)] various upper bounds on such quantities as $G_n(\mathcal{F})$ and $R_n(\mathcal{F})$ in terms of entropies, VC-dimensions, etc.

First, we give bounds on $P\{f \le 0\}$ in terms of a class of so called *margin cost functions*. These bounds will be used in Section 5 in the context of classification problems to improve recent results of Mason, Bartlett and Baxter (1999).

Consider a countable family of Lipschitz functions $\Phi = \{\varphi_k : k \ge 1\}$, where $\varphi_k : \mathbb{R} \to \mathbb{R}$ are such that $I_{(-\infty,0]}(x) \le \varphi_k(x)$ for all k. For each $\varphi \in \Phi$, $L(\varphi)$ will denote its Lipschitz constant.

We assume that for any $x \in S$ the set of real numbers $\{f(x) : f \in \mathcal{F}\}$ is bounded.

THEOREM 1. *For all $t > 0$,*

$$\mathbb{P}\left\{\exists f \in \mathcal{F} : P\{f \le 0\} > \inf_{k\ge 1}\left[P_n\varphi_k(f) + 4L(\varphi_k)R_n(\mathcal{F}) + \left(\frac{\log k}{n}\right)^{1/2}\right] + \frac{t}{\sqrt{n}}\right\}$$

$$\le 2\exp\{-2t^2\}$$

and

$$\mathbb{P}\Biggl\{\exists f\in\mathcal{F}: P\{f\le 0\} > \inf_{k\ge 1}\Biggl[P_n\varphi_k(f)+\sqrt{2\pi}L(\varphi_k)G_n(\mathcal{F})+\biggl(\frac{\log k}{n}\biggr)^{1/2}\Biggr] + \frac{t+2}{\sqrt{n}}\Biggr\} \le 2\exp\bigl\{-2t^2\bigr\}.$$

PROOF. Without loss of generality we can and do assume that each $\varphi\in\Phi$ takes its values in $[0,1]$ (otherwise it can be redefined as $\varphi\wedge 1$). Clearly, in this case $\varphi(x)=1$ for $x\le 0$. For a fixed $\varphi\in\Phi$ and for all $f\in\mathcal{F}$ we have

$$P\{f\le 0\}\le P\varphi(f)\le P_n\varphi(f)+\|P_n-P\|_{\mathcal{G}_\varphi}, \tag{2.1}$$

where

$$\mathcal{G}_\varphi := \{\varphi\circ f-1 : f\in\mathcal{F}\}.$$

By the exponential inequalities for martingale difference sequences [see Devroye, Györfi and Lugosi (1996), pages 135–136], we have

$$\mathbb{P}\Biggl\{\|P_n-P\|_{\mathcal{G}_\varphi}\ge \mathbb{E}\|P_n-P\|_{\mathcal{G}_\varphi}+\frac{t}{\sqrt{n}}\Biggr\}\le\exp\bigl\{-2t^2\bigr\}.$$

Thus, with probablity at least $1-\exp\{-2t^2\}$ for all $f\in\mathcal{F}$

$$P\{f\le 0\}\le P_n\varphi(f)+\mathbb{E}\|P_n-P\|_{\mathcal{G}_\varphi}+\frac{t}{\sqrt{n}}. \tag{2.2}$$

The symmetrization inequality gives [van der Vaart and Wellner (1996)]

$$\mathbb{E}\|P_n-P\|_{\mathcal{G}_\varphi}\le 2\mathbb{E}\Biggl\|n^{-1}\sum_{i=1}^n\varepsilon_i\delta_{X_i}\Biggr\|_{\mathcal{G}_\varphi}. \tag{2.3}$$

Since a function $(\varphi-1)/L(\varphi)$ is a contraction and $\varphi(0)-1=0$, the Rademacher comparison inequality [Ledoux and Talagrand (1991), Theorem 4.12, page 112] implies

$$\mathbb{E}_\varepsilon\Biggl\|n^{-1}\sum_{i=1}^n\varepsilon_i\delta_{X_i}\Biggr\|_{\mathcal{G}_\varphi}\le 2L(\varphi)\mathbb{E}_\varepsilon\Biggl\|n^{-1}\sum_{i=1}^n\varepsilon_i\delta_{X_i}\Biggr\|_{\mathcal{F}}.$$

It now follows from (2.2), (2.3) that with probability at least $1-e^{-2t^2}$ we have for all $f\in\mathcal{F}$

$$P\{f\le 0\}\le P_n\varphi(f)+4L(\varphi)R_n(\mathcal{F})+\frac{t}{\sqrt{n}}. \tag{2.4}$$

We use now (2.4) with $\varphi = \varphi_k$ and t replaced by $t + \sqrt{\log k}$ to obtain

$$\mathbb{P}\Biggl\{\exists f \in \mathcal{F} : P\{f \le 0\} > \inf_{k\ge 1}\Biggl[P_n\varphi_k(f) + 4L(\varphi_k)R_n(\mathcal{F}) + \biggl(\frac{\log k}{n}\biggr)^{1/2}\Biggr] + \frac{t}{\sqrt{n}}\Biggr\} \tag{2.5}$$

$$\le \sum_{k\ge 1}\exp\Bigl\{-2\bigl(t+\sqrt{\log k}\bigr)^2\Bigr\} \le \sum_{k\ge 1} k^{-2}e^{-2t^2} = \frac{\pi^2}{6}e^{-2t^2} \le 2e^{-2t^2}.$$

The proof of the second bound is quite similar with the following changes. The class $\mathcal{G}_\varphi$ is defined in this case as $\{\varphi \circ f : f \in \mathcal{F}\}$. Instead of (2.3), we have in this case, by the symmetrization inequality and the Gaussian multiplier inequality [see van der Vaart and Wellner (1996), pages 108–109, 177–179], that

$$\mathbb{E}\|P_n - P\|_{\mathcal{G}_\varphi} \le 2\mathbb{E}\biggl\|n^{-1}\sum_{i=1}^n \varepsilon_i\delta_{X_i}\biggr\|_{\mathcal{G}_\varphi} \le \sqrt{2\pi}\,\mathbb{E}\biggl\|n^{-1}\sum_{i=1}^n g_i\delta_{X_i}\biggr\|_{\mathcal{G}_\varphi}. \tag{2.6}$$

Define Gaussian processes

$$Z_1(f,\sigma) := \sigma n^{-1/2}\sum_{i=1}^n g_i(\varphi\circ f)(X_i)$$

and

$$Z_2(f,\sigma) := L(\varphi)n^{-1/2}\sum_{i=1}^n g_i f(X_i) + \sigma g,$$

where $\sigma = \pm 1$ and g is standard normal independent of the sequence $\{g_i\}$. If we denote by $\mathbb{E}_g$ the expectation on the probability space $(\Omega_g, \Sigma_g, \mathbb{P}_g)$ on which the sequence $\{g_i\}$ and g are defined then we have

$$\mathbb{E}_g\bigl|Z_1(f,\sigma) - Z_1\bigl(h,\sigma'\bigr)\bigr|^2 \le \mathbb{E}_g\bigl|Z_2(f,\sigma) - Z_2\bigl(h,\sigma'\bigr)\bigr|^2, \tag{2.7}$$

which is easy to observe if we consider separately the cases when $\sigma\sigma'$ is equal to 1 and to -1. Indeed, if $\sigma\sigma' = 1$ then (2.7) is equivalent to

$$n^{-1}\sum_{i=1}^n \bigl|\varphi(f(X_i)) - \varphi(h(X_i))\bigr|^2 \le L(\varphi)^2 n^{-1}\sum_{i=1}^n \bigl[f(X_i) - h(X_i)\bigr]^2$$

which holds since φ satisfies the Lipschitz condition with constant $L(\varphi)$. If $\sigma\sigma' = -1$ then since $0 \le \varphi \le 1$ we have

$$\begin{aligned}\mathbb{E}_g\bigl|Z_1(f,\sigma) - Z_1\bigl(h,\sigma'\bigr)\bigr|^2 &\le 2n^{-1}\sum_{i=1}^n \varphi^2(f(X_i)) + 2n^{-1}\sum_{i=1}^n \varphi^2(h(X_i))\\ &\le \mathbb{E}(2g)^2 \le \mathbb{E}_g\bigl|Z_2(f,\sigma) - Z_2\bigl(h,\sigma'\bigr)\bigr|^2.\end{aligned}$$

A version of Slepian's Lemma [see Ledoux and Talagrand (1991), pages 76–77] implies that

$$\mathbb{E}_g \sup\bigl\{Z_1(f,\sigma) : f \in \mathcal{F},\ \sigma = \pm 1\bigr\} \le \mathbb{E}_g \sup\bigl\{Z_2(f,\sigma) : f \in \mathcal{F},\ \sigma = \pm 1\bigr\}.$$

We have

$$\begin{aligned}\mathbb{E}_g \left\| n^{-1/2} \sum_{i=1}^n g_i \delta_{X_i} \right\|_{\mathcal{G}_\varphi} &= \mathbb{E}_g \sup_{h \in \bar{\mathcal{G}}_\varphi} \left[n^{-1/2} \sum_{i=1}^n g_i h(X_i) \right] \\ &= \mathbb{E}_g \sup\bigl\{Z_1(f,\sigma) : f \in \mathcal{F}, \sigma = \pm 1\bigr\},\end{aligned}$$

where $\bar{\mathcal{G}}_\varphi := \{\varphi(f), -\varphi(f) : f \in \mathcal{F}\}$, and similarly

$$L(\varphi)\mathbb{E}_g \left\| n^{-1/2} \sum_{i=1}^n g_i \delta_{X_i} \right\|_{\mathcal{F}} + \mathbb{E}|g| \ge \mathbb{E}_g \sup\bigl\{Z_2(f,\sigma) : f \in \mathcal{F},\ \sigma = \pm 1\bigr\}.$$

This immediately gives us

$$\mathbb{E}_g \left\| n^{-1} \sum_{i=1}^n g_i \delta_{X_i} \right\|_{\mathcal{G}_\varphi} \le L(\varphi)\mathbb{E}_g \left\| n^{-1} \sum_{i=1}^n g_i \delta_{X_i} \right\|_{\mathcal{F}} + n^{-1/2}\mathbb{E}|g|. \tag{2.8}$$

It follows from (2.2), (2.6) and (2.8) that with probability at least $1 - e^{-2t^2}$

$$P\{f \le 0\} \le P_n\varphi(f) + \sqrt{2\pi}\, L(\varphi) G_n(\mathcal{F}) + \frac{t+2}{\sqrt{n}}. \tag{2.9}$$

The proof now can be completed the same way as in the case of the first bound. □

Let us consider a special family of cost functions. Assume that φ is a fixed *nonincreasing* function such that $\varphi(x) \ge I_{(-\infty,0]}(x)$ for $x \in \mathbb{R}$ and φ satisfies the Lipschitz condition with constant $L(\varphi)$. Let

$$\Phi_0 := \bigl\{\varphi(\cdot/\delta) : \delta \in (0,1]\bigr\}.$$

One can easily observe that $L(\varphi(\cdot/\delta)) \le L(\varphi)\delta^{-1}$. For this family, Theorem 1 easily implies the following statement, which, in turn, implies the result of Schapire, Freund, Bartlett and Lee (1998) for VC-classes of base classifiers (see Section 5).

THEOREM 2. *For all $t > 0$,*

$$\begin{aligned}\mathbb{P}\Biggl\{\exists f \in \mathcal{F} : P\{f \le 0\} > \inf_{\delta \in (0,1]} \Biggl[P_n\varphi\Bigl(\frac{f}{\delta}\Bigr) + \frac{8L(\varphi)}{\delta} R_n(\mathcal{F}) \\ + \Bigl(\frac{\log\log_2(2\delta^{-1})}{n}\Bigr)^{1/2} \Biggr] + \frac{t}{\sqrt{n}} \Biggr\} \\ \le 2\exp\{-2t^2\}\end{aligned}$$

and

$$\mathbb{P}\Biggl\{\exists f\in\mathcal{F}: P\{f\le 0\} > \inf_{\delta\in(0,1]}\Biggl[P_n\varphi\Bigl(\frac{f}{\delta}\Bigr)+\frac{2\sqrt{2\pi}L(\varphi)}{\delta}G_n(\mathcal{F}) + \Bigl(\frac{\log\log_2(2\delta^{-1})}{n}\Bigr)^{1/2}\Biggr]+\frac{t+2}{\sqrt{n}}\Biggr\} \le 2\exp\{-2t^2\}.$$

PROOF. One has to apply the bounds of Theorem 1 for the sequence $\varphi_k(\cdot) := \varphi(\cdot/\delta_k)$, where $\delta_k = 2^{-k}$, and then notice that for $\delta\in(\delta_k,\delta_{k-1}]$, we have

$$\frac{1}{\delta_k}\le\frac{2}{\delta},\qquad P_n\varphi\Bigl(\frac{f}{\delta_k}\Bigr)\le P_n\varphi\Bigl(\frac{f}{\delta}\Bigr)$$

and

$$\sqrt{\log k}=\sqrt{\log\log_2\frac{1}{\delta_k}}\le\sqrt{\log\log_2\frac{2}{\delta}}.$$ □

REMARK. The constant 8 in front of the Rademacher complexity and the constant $2\sqrt{2\pi}$ in front of the Gaussian complexity can be replaced by $4c$ and $\sqrt{2\pi}c$, respectively, for any $c>1$ (with minor changes in the logarithmic term). Also, one can choose $c=c(\delta)$, where $c(\delta)=1+o(1)$ as $\delta\to 0$.

In the next statements we use the Rademacher complexities, but Gaussian complexities can be used similarly.

Assuming now that φ is a function from $\mathbb{R}$ into $\mathbb{R}$ such that $\varphi(x)\le I_{(-\infty,0]}(x)$ for all $x\in\mathbb{R}$ and φ still satisfies the Lipschitz condition with constant $L(\varphi)$, one can prove the following statement.

THEOREM 3. *For all* $t>0$,

$$\mathbb{P}\Biggl\{\exists f\in\mathcal{F}: P\{f\le 0\} < \sup_{\delta\in(0,1]}\Biggl(P_n\varphi\Bigl(\frac{f}{\delta}\Bigr)-\frac{8L(\varphi)}{\delta}R_n(\mathcal{F}) - \Bigl(\frac{\log\log_2(2\delta^{-1})}{n}\Bigr)^{1/2}\Biggr)-\frac{t}{\sqrt{n}}\Biggr\} \le 2\exp\{-2t^2\}.$$

Denote

$$\Delta_n(\mathcal{F};\delta) := \frac{8}{\delta}R_n(\mathcal{F})+\Bigl(\frac{\log\log_2(2\delta^{-1})}{n}\Bigr)^{1/2}.$$

The bounds of Theorems 2 and 3 easily imply that for all $t > 0$

$$\mathbb{P}\Big\{\exists f \in \mathcal{F} : P\{f \le 0\} > P_n\{f \le 0\} + \inf_{\delta\in(0,1]}\big[P_n\{0 < f \le \delta\} + \Delta_n(\mathcal{F};\delta)\big] + \frac{t}{\sqrt{n}}\Big\} \le 2\exp\{-2t^2\}$$

and

$$\mathbb{P}\Big\{\exists f \in \mathcal{F} : P\{f \le 0\} < P_n\{f \le 0\} - \inf_{\delta\in(0,1]}\big[P_n\{-\delta < f \le 0\} + \Delta_n(\mathcal{F};\delta)\big] - \frac{t}{\sqrt{n}}\Big\} \le 2\exp\{-2t^2\}.$$

To prove this it's enough to take φ equal to 1 for $x \le 0$, 0 for $x \ge 1$ and linear in between in the case of the first bound; in the case of the second bound, the choice of φ is 1 for $x \le -1$, 0 for $x \ge 0$ and linear in between. Similarly, it can be shown that

$$\mathbb{P}\Big\{\exists f \in \mathcal{F} : P_n\{f \le 0\} > P\{f \le 0\} + \inf_{\delta\in(0,1]}\big[P\{0 < f \le \delta\} + \Delta_n(\mathcal{F};\delta)\big] + \frac{t}{\sqrt{n}}\Big\} \le 2\exp\{-2t^2\}$$

and

$$\mathbb{P}\Big\{\exists f \in \mathcal{F} : P_n\{f \le 0\} < P\{f \le 0\} - \inf_{\delta\in(0,1]}\big[P\{-\delta < f \le 0\} + \Delta_n(\mathcal{F};\delta)\big] - \frac{t}{\sqrt{n}}\Big\} \le 2\exp\{-2t^2\}.$$

Combining the last bounds, we get the following result:

THEOREM 4. *For all $t > 0$,*

$$\mathbb{P}\Big\{\exists f \in \mathcal{F} : \big|P_n\{f \le 0\} - P\{f \le 0\}\big| > \inf_{\delta\in(0,1]}\big[P_n\{|f| \le \delta\} + \Delta_n(\mathcal{F};\delta)\big] + \frac{t}{\sqrt{n}}\Big\} \le 4\exp\{-2t^2\}$$

and

$$\mathbb{P}\Big\{\exists f \in \mathcal{F} : \big|P_n\{f \le 0\} - P\{f \le 0\}\big| > \inf_{\delta\in(0,1]}\big[P\{|f| \le \delta\} + \Delta_n(\mathcal{F};\delta)\big] + \frac{t}{\sqrt{n}}\Big\} \le 4\exp\{-2t^2\}.$$

Denote

$$H_f(\delta) := \delta P\{|f| \le \delta\},\ H_{n,f}(\delta) := \delta P_n\{|f| \le \delta\}.$$

Plugging in the second bound of Theorem 4 $\delta := H_f^{-1}(R_n(\mathcal{F})) \wedge 1$ (we use the notation $a \wedge b := \min(a, b)$) easily gives us the following upper bound that holds for any $t > 0$ with probability at least $1 - 4e^{-2t^2}$:

$$\forall f \in \mathcal{F}, \qquad \bigl|P_n\{f \le 0\} - P\{f \le 0\}\bigr| \le \frac{9R_n(\mathcal{F})}{\delta} + \biggl(\frac{\log\log_2(2\delta^{-1})}{n}\biggr)^{1/2} + \frac{t}{\sqrt{n}}.$$

Similarly, the first bound of Theorem 4 gives that for any $t > 0$ with probability at least $1 - 4e^{-2t^2}$:

$$\forall f \in \mathcal{F}, \qquad |P_n\{f \le 0\} - P\{f \le 0\}| \le \frac{9R_n(\mathcal{F})}{\delta} + \biggl(\frac{\log\log_2(2\delta^{-1})}{n}\biggr)^{1/2} + \frac{t}{\sqrt{n}}$$

with $\delta := H_{n,f}^{-1}(R_n(\mathcal{F})) \wedge 1$.

The next example shows that, in general, the term $\frac{1}{\delta}R_n(\mathcal{F})$ of the bound of Theorem 2 (and other similar results, in particular, Theorem 4) can not be improved.

Let us consider a sequence $\{X_n\}$ of independent identically distributed random variables in l_∞ defined by

$$X_n = \Bigl\{\varepsilon_k^n \bigl(2\log(k+1)\bigr)^{-\frac{1}{2}}\Bigr\}_{k\ge 1}, \qquad n \ge 1,$$

where ε_k^n are i.i.d. Rademacher random variables ($P(\varepsilon_k^n = \pm 1) = 1/2$). We consider a class of functions that consists of canonical projections on each coordinate

$$\mathcal{F} = \{f_k : f_k(x) = x_k\}.$$

Let $\phi(x)$ be an increasing function such that $\phi(0) = 0$. Then the following proposition holds.

PROPOSITION 1.

$$\mathbb{P}\biggl\{\exists f \in \mathcal{F} : P\{f \le 0\} \ge \inf_{\delta\in(0,1]}\biggl[P_n\{f \le \delta\} + \frac{8}{\phi(\delta)}R_n(\mathcal{F})\biggr] + \frac{t}{\sqrt{n}}\biggr\} \to 1$$

when $n \to \infty$ uniformly for all $t \le 2^{-1}n^{1/2}\phi((4n)^{-1/2}) - c$, where $c > 0$ is some fixed constant.

PROOF. It is well known that $\mathcal{F}$ is a bounded CLT class for the distribution P of the sequence $\{X_n\}$ [see Ledoux and Talagrand (1991), pages 276–277]. Notice that $P(f_k \le 0) = 1/2$ for all k and $\mathbb{E}\|n^{-1}\sum \varepsilon_i \delta_{X_i}\|_{\mathcal{F}} \le cn^{-1/2}$ for some constant

$c > 0$. Let us denote by $t' = t + 2\sqrt{2\pi}c$. The infimum inside the probability is less than or equal to the value of the expression at any fixed point. Therefore, for each k we will choose δ to be equal to a $\delta_k > (2\log(k+1))^{-1/2}$. It's easy to see that for this value of δ,

$$P_n\{f_k \le \delta_k\} = \frac{1}{n}\sum_{i=1}^{n} I(\varepsilon_k^i = -1).$$

Combining these estimates we get that the probability defined in the statement of the proposition is greater than or equal to

$$\mathbb{P}\left\{\exists k : \frac{1}{2} \ge \frac{1}{n}\sum_{i\le n} I(\varepsilon_k^i = -1) + \frac{t'}{\phi(\delta_k)\sqrt{n}}\right\}$$
$$= 1 - \prod_k \mathbb{P}\left\{\frac{1}{2} < \frac{1}{n}\sum_{i\le n} I(\varepsilon_k^i = -1) + \frac{t'}{\phi(\delta_k)\sqrt{n}}\right\}.$$

In the product above factors are possibly not equal to 1 only for k in the set of indices

$$\mathcal{K} = \left\{k : \gamma_k = \frac{t'}{\phi(\delta_k)\sqrt{n}} \le \frac{1}{2}\right\}.$$

Clearly,

$$\mathbb{P}\left\{1/2 < n^{-1}\sum_{i\le n} I(\varepsilon_1^i = -1) + \delta\right\} \le 1 - \binom{n}{k_0}2^{-n},$$

where $k_0 = [n/2 - \delta n] - 1$. For simplicity of calculations we will set $k_0 = n/2 - \delta n$. Utilizing the following estimates in Stirling's formula for the factorial [see Feller (1950)],

$$(2\pi)^{\frac{1}{2}} n^{n+\frac{1}{2}} e^{-n+1/(12n+1)} < n! < (2\pi)^{\frac{1}{2}} n^{n+\frac{1}{2}} e^{-n+1/12n}, \tag{2.10}$$

it is straightforward to check that for some constant $c > 0$,

$$\binom{n}{k_0}2^{-n} \ge cn^{-\frac{1}{2}}\left((1-2\delta)^{1-2\delta}(1+2\delta)^{1+2\delta}\right)^{-\frac{n}{2}} \ge cn^{-\frac{1}{2}}\exp\left(-4n\delta^2\right). \tag{2.11}$$

The last inequality is due to the fact that

$$\exp(x^2) \le (1-x)^{1-x}(1+x)^{1+x} \le \exp(2x^2)$$

for $x < 2^{-1/2}$. It follows from (2.11) that

$$\mathbb{P}\left\{\tfrac{1}{2} < \frac{1}{n}\sum_{i\le n} I(\varepsilon_k^i = -1) + \gamma_k\right\} \le 1 - cn^{-1/2}\exp(-4n\gamma_k^2).$$

Since $\gamma_k \le 1/2$ for $k \in \mathcal{K}$, we can continue and come to the lower bound

$$1 - \prod_{k\in\mathcal{K}} \bigl(1 - cn^{-1/2}\exp(-4n\gamma_k^2)\bigr) \ge 1 - \exp\Bigl(-\sum_{k\in\mathcal{K}} cn^{-1/2}\exp(-4n\gamma_k^2)\Bigr)$$
$$\ge 1 - \exp\bigl(-\operatorname{card}(\mathcal{K})cn^{-1/2}e^{-n}\bigr) \to 1,$$

uniformly in t', if we check that $\operatorname{card}(\mathcal{K})cn^{-1/2}e^{-n} \to \infty$. Indeed, if

$$t' \le 2^{-1}n^{1/2}\phi\bigl((4n)^{-1/2}\bigr)$$

then for n large enough,

$$t' \le 2^{-1}n^{1/2}\phi\bigl((4n)^{-1/2}\bigr) \le 2^{-1}n^{1/2}\phi\bigl(\bigl(2\log\bigl([cne^n]+1\bigr)\bigr)^{-1/2}\bigr).$$

This means that $[cne^n] \in \mathcal{K}$, and, therefore,

$$\operatorname{card}(\mathcal{K})cn^{-1/2}e^{-n} \ge n^{1/2} - \frac{1}{cn^{1/2}e^n} \to \infty.$$

The proposition is proven. □

REMARKS. If $\phi(x) = x^{1-\alpha}$ for some positive α then the convergence in the proposition holds for $t \le cn^{\alpha/2}$. Also, if $\frac{\phi(\delta)}{\delta} \to \infty$ as $\delta \to 0$, then the convergence in the proposition holds uniformly in $t \in [0, T]$ for any $T > 0$. It means that the bound of Theorem 2 does not hold with $\frac{1}{\delta}R_n(\mathcal{F})$ replaced by $\frac{1}{\phi(\delta)}R_n(\mathcal{F})$. Similarly, one can show that

$$\mathbb{P}\Biggl\{\exists f \in \mathcal{F} : \bigl|P_n\{f \le 0\} - P\{f \le 0\}\bigr| \ge \inf_{\delta\in(0,1]}\Biggl[P_n\{|f| \le \delta\} + \frac{8}{\phi(\delta)}R_n(\mathcal{F})\Biggr]$$
$$+ \frac{t}{\sqrt{n}}\Biggr\} \to 1$$

when $n \to \infty$ uniformly for all $t \le 2^{-1}n^{1/2}\phi((4n)^{-1/2}) - c$.

3. Conditions on random entropies and γ-margins. Given a metric space (T, d), we denote by $H_d(T;\varepsilon)$ the ε-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 ε covering T. Let $d_{P_n,2}$ denote the metric of the space $L_2(S; dP_n)$:

$$d_{P_n,2}(f,g) := \bigl(P_n|f-g|^2\bigr)^{1/2}.$$

The next theorems improve the bounds of the previous section under some assumptions on the growth of random entropies $H_{d_{P_n,2}}(\mathcal{F};\cdot)$. We will use these results in Section 5 to obtain an improvement of the bound of Schapire, Freund,

Bartlett and Lee (1998) on generalization error of boosting. The method of proof is similar to the one developed in Koltchinskii and Panchenko (2000) and is based on powerful concentration inequalities of Talagrand (1996) [see also Massart (2000)].

Define for $\gamma \in (0, 1]$

$$\delta_n(\gamma; f) := \sup\Big\{\delta \in (0,1) : \delta^\gamma P\{f \le \delta\} \le n^{-1+\frac{\gamma}{2}}\Big\}$$

and

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

We call $\delta_n(\gamma; f)$ and $\hat{\delta}_n(\gamma; f)$, respectively, the *$\gamma$-margin* and the *empirical γ-margin* of f.

The main result of this section is Theorem 5 which gives the condition on the random entropy $H_{d_{P_n,2}}(\mathcal{F}; \cdot)$ under which the true γ-margin of any $f \in \mathcal{F}$ is with probability very close to 1 within a multiplicative constant from its empirical γ-margin. This implies that with high probability for all $f \in \mathcal{F}$,

$$P\{f \le 0\} \le \frac{\text{const}}{n^{1-\gamma/2}\hat{\delta}_n(\gamma; f)^\gamma}.$$

The bounds of the previous section correspond to the case of $\gamma = 1$. It is easy to see from the definitions of $\gamma-$margins that the quantity $(n^{1-\gamma/2}\hat{\delta}_n(\gamma; f)^\gamma)^{-1}$ (called in the introduction the γ-bound) increases in $\gamma \in (0, 1]$. This shows that the bound in the case of $\gamma < 1$ is tighter than the bounds of Section 2.

THEOREM 5. *Suppose that for some $\alpha \in (0, 2)$ and for some constant $D > 0$,*

$$H_{d_{P_n,2}}(\mathcal{F}; u) \le Du^{-\alpha}, \qquad u > 0 \text{ a.s.} \tag{3.1}$$

Then for any $\gamma \ge \frac{2\alpha}{2+\alpha}$, for some constants $A, B > 0$ and for all large enough n

$$\mathbb{P}\big\{\forall f \in \mathcal{F} : A^{-1}\hat{\delta}_n(\gamma; f) \le \delta_n(\gamma; f) \le A\hat{\delta}_n(\gamma; f)\big\}$$
$$\ge 1 - B\log_2\log_2 n \exp\{-n^{\frac{\gamma}{2}}/2\}.$$

The proof is based on the following result.

THEOREM 6. *Suppose that for some $\alpha \in (0, 2)$ and for some constant $D > 0$ condition* (3.1) *holds. Then for some constants $A, B > 0$, for all $\delta \ge 0$ and*

$$\varepsilon \ge \left(\frac{1}{n\delta^\alpha}\right)^{\frac{2}{2+\alpha}} \vee \frac{2\log n}{n}, \tag{3.2}$$

and for all large enough n, the following bounds hold:

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

and

$$\mathbb{P}\left\{\exists f \in \mathcal{F}\ P\{f \le \delta\} \le \varepsilon \text{ and } P_n\left\{f \le \frac{\delta}{2}\right\} \ge A\varepsilon\right\} \le B \log_2 \log_2 \varepsilon^{-1} \exp\left\{-\frac{n\varepsilon}{2}\right\}.$$

PROOF. Define recursively

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

with some sufficiently large constant $C > 1$ (the choice of C will be explained later). By a simple induction argument we have either $C\sqrt{\varepsilon} \ge 1$ and $r_k \equiv 1$, or $C\sqrt{\varepsilon} < 1$ and in this case

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

Without loss of generality we can assume that $C\sqrt{\varepsilon} < 1$. Let

$$\gamma_k := \sqrt{\frac{\varepsilon}{r_k}} = C^{2^{-k}-1} \varepsilon^{2^{-k-1}}.$$

For a fixed $\delta > 0$, define

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

WARNING. In what follows in the proof "c" denotes a constant; its values can be different in different places.

Define $\mathcal{F}_0 := \mathcal{F}$, and further recursively

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

For $k \ge 0$, let φ_k be a continuous function from $\mathbb{R}$ into $[0, 1]$ such that $\varphi_k(u) = 1$ for $u \le \delta_{k,\frac{1}{2}}$, $\varphi_k(u) = 0$ for $u \ge \delta_k$, and linear for $\delta_{k,\frac{1}{2}} \le u \le \delta_k$. For $k \ge 1$ let φ_k' be a continuous function from $\mathbb{R}$ into $[0, 1]$ such that $\varphi_k'(u) = 1$ for $u \le \delta_k$, $\varphi_k'(u) = 0$ for $u \ge \delta_{k-1,\frac{1}{2}}$, and linear for $\delta_k \le u \le \delta_{k-1,\frac{1}{2}}$. We have

$$\begin{aligned}\sum_{i=0}^{k} \gamma_i &= C^{-1}\left[C\sqrt{\varepsilon} + \left(C\sqrt{\varepsilon}\right)^{2^{-1}} + \cdots + \left(C\sqrt{\varepsilon}\right)^{2^{-k}}\right] \\ &\le C^{-1}\left(C\sqrt{\varepsilon}\right)^{2^{-k}} \left(1 - \left(C\sqrt{\varepsilon}\right)^{2^{-k}}\right)^{-1} \le 1/2,\end{aligned}$$

for $\varepsilon \le C^{-4}$, $C > 2(2^{1/4} - 1)^{-1}$ and $k \le \log_2 \log_2 \varepsilon^{-1}$. Hence, for small enough ε (note that our choice of $\varepsilon \le C^{-4}$ implies $C\sqrt{\varepsilon} < 1$), we have

$$\gamma_0 + \cdots + \gamma_k \le \tfrac{1}{2}, \qquad k \ge 1.$$

Therefore, for all $k \geq 1$, we get $\delta_k \in (\delta/2, \delta)$. Note also that below our choice of k will be such that the restriction $k \leq \log_2 \log_2 \varepsilon^{-1}$ for any fixed $\varepsilon > 0$ will always be fulfilled.

Define

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

and

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

Clearly, by these definitions, for $k \geq 1$,

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

and

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

Since $r_0 = 1$, for $k = 0$ the first inequality becomes trivial. If now we introduce the following events:

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

then it follows from the concentration inequalities of Talagrand (1996a, b) [see also Massart (2000)] that with some numerical constants $K_1, K_2, K_3 > 0$,

$$\mathbb{P}((E^{(k)})^c) \leq 2e^{-\frac{n\varepsilon}{2}}.$$

Denote $E_0 = \Omega$,

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

Then

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

In what follows we can and do assume without loss of generality that $\varepsilon < C^{-4}$ and therefore, $r_{k+1} < r_k$ and $\delta_k \in (\delta/2, \delta]$, $k \leq \log_2 \log_2 \varepsilon^{-1}$. (If $\varepsilon \geq C^{-4}$, then the bounds of the theorem obviously hold with any constant $A > C^4$.) The following lemma holds.

LEMMA 1. *Let N be such that*

$$N \leq \log_2 \log_2 \varepsilon^{-1} \quad \textit{and} \quad r_N \geq \varepsilon. \tag{3.3}$$

Let $\mathcal{J} = \{\inf_{f\in\mathcal{F}} P_n\{f \le \delta\} \le \varepsilon\}$. *Then the following properties hold on the event* $E_N \bigcap \mathcal{J}$:

$$\text{(i)} \quad \forall f \in \mathcal{F} \quad P_n\{f \le \delta\} \le \varepsilon \Longrightarrow f \in \mathcal{F}_N$$

and

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

PROOF. We will use induction with respect to N. For $N = 0$, the statement is obvious. Suppose it holds for some $N \ge 0$, such that $N + 1$ still satisfies condition (3.3) of the lemma. Then on the event $E_N \cap \mathcal{J}$ we have

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

and

$$\forall f \in \mathcal{F} \ P_n\{f \le \delta\} \le \varepsilon \Longrightarrow f \in \mathcal{F}_N.$$

Suppose now that $f \in \mathcal{F}$ is such that $P_n\{f \le \delta\} \le \varepsilon$. By the induction assumptions, on the event E_N, we have $f \in \mathcal{F}_N$. Because of this, we obtain on the event E_{N+1}

$$\begin{aligned} P\{f \le \delta_{N,\frac{1}{2}}\} &\le P_n\{f \le \delta_N\} + \|P_n - P\|_{\mathcal{G}_N} \\ &\le \varepsilon + K_1\mathbb{E}\|P_n - P\|_{\mathcal{G}_N} + K_2\sqrt{r_N\varepsilon} + K_3\varepsilon. \end{aligned} \tag{3.4}$$

For a class $\mathcal{G}$, define

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

where $\{\varepsilon_i\}$ is a sequence of i.i.d. Rademacher random variables. By the symmetrization inequality,

$$\mathbb{E}\|P_n - P\|_{\mathcal{G}_N} \le 2\mathbb{E}I_{E_N}\mathbb{E}_\varepsilon\widehat{R}_n(\mathcal{G}_N) + 2\mathbb{E}I_{E_N^c}\mathbb{E}_\varepsilon\widehat{R}_n(\mathcal{G}_N). \tag{3.5}$$

Next, by the well known entropy inequalities for subgaussian processes [see van der Vaart and Wellner (1996), Corollary 2.2.8], we have

$$\begin{aligned} \mathbb{E}_\varepsilon\widehat{R}_n(\mathcal{G}_N) &\le \inf_{g\in\mathcal{G}_N} \mathbb{E}_\varepsilon\left| n^{-1}\sum_{j=1}^n \varepsilon_j g(X_j)\right| \\ &\quad + \frac{c}{\sqrt{n}}\int_0^{(2\sup_{g\in\mathcal{G}_N} P_n g^2)^{1/2}} H^{1/2}_{d_{P_n,2}}(\mathcal{G}_N; u)\,du. \end{aligned} \tag{3.6}$$

By the induction assumption, on the event $E_N \cap \mathcal{J}$,

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

We also have on the event $E_N \cap \mathcal{J}$,

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

The Lipschitz norm of φ_{k-1} and φ_k' is 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 implies the following bound on the distance:

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

Therefore, on the event $E_N\cap\mathcal{J}$,

$$\frac{1}{\sqrt{n}}\int_0^{(2\sup_{g\in\mathcal{G}_N} P_n g^2)^{1/2}} H^{1/2}_{d_{P_n,2}}(\mathcal{G}_N;u)\,du$$
$$\le \frac{1}{\sqrt{n}}\int_0^{(2r_N)^{1/2}} H^{1/2}_{d_{P_n,2}}\left(\mathcal{F};\frac{\delta\sqrt{\varepsilon}u}{2\sqrt{r_N}}\right)du \tag{3.7}$$
$$\le c\left(\frac{r_N}{\varepsilon}\right)^{\alpha/4} \frac{r_N^{1/2-\alpha/4}}{\sqrt{n}\delta^{\alpha/2}} \le c\frac{r_N^{1/2}}{\varepsilon^{\alpha/4}}\varepsilon^{\frac{2+\alpha}{4}} = c\sqrt{r_N\varepsilon},$$

where we used the fact that condition (3.2) of the theorem implies

$$\frac{1}{n^{1/2}\delta^{\alpha/2}} \le \varepsilon^{\frac{2+\alpha}{4}}.$$

It follows from (3.6), (3.7) that on the event $E_{N+1}\cap\mathcal{J}$,

$$\mathbb{E}_\varepsilon \widehat{R}_n(\mathcal{G}_N) \le c\sqrt{r_N\varepsilon}. \tag{3.8}$$

Since we also have

$$\mathbb{E}_\varepsilon \widehat{R}_n(\mathcal{G}_{N+1}) \le 1,$$

(3.5) and (3.8) yield

$$\mathbb{E}\|P_n - P\|_{\mathcal{G}_N} \le c\sqrt{r_N\varepsilon} + 2\mathbb{P}(E_N^c) \le c\sqrt{r_N\varepsilon} + 4Ne^{-n\varepsilon/2}.$$

Since $4Ne^{-n\varepsilon/2} \le \varepsilon$ [it holds due to the conditions (3.2) and (3.3), for all large enough n] we conclude that with some constant $c > 0$,

$$\mathbb{E}\|P_n - P\|_{\mathcal{G}_N} \le c\sqrt{r_N\varepsilon}.$$

Now we use (3.4) and see that on the event $E_{N+1} \cap \mathcal{J}$

$$P\{f \le \delta_{N,\frac{1}{2}}\} \le c\big(\varepsilon + \sqrt{r_N\varepsilon}\big). \tag{3.9}$$

Therefore, it follows that with a proper choice of constant $C > 0$ in the recurrence relationship defining the sequence $\{r_k\}$, we have on the event $E_{N+1} \cap \mathcal{J}$

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

This means that $f \in \mathcal{F}_{N+1}$ and the induction step for (i) is proved. This will now imply (ii). We have on the event E_{N+1}

$$\begin{aligned}\sup_{f\in\mathcal{F}_{N+1}} P_n\{f \le \delta_{N+1}\} &\le \sup_{f\in\mathcal{F}_{N+1}} P\{f \le \delta_{N,\frac{1}{2}}\} + \|P_n - P\|_{\mathcal{G}'_{N+1}} \\ &\le r_{N+1}/2 + K_1\mathbb{E}\|P_n - P\|_{\mathcal{G}'_{N+1}} \\ &\quad + K_2\sqrt{r_{N+1}\varepsilon} + K_3\varepsilon.\end{aligned} \tag{3.10}$$

By the symmetrization inequality,

$$\mathbb{E}\|P_n - P\|_{\mathcal{G}'_{N+1}} \le 2\mathbb{E}I_{E_N}\mathbb{E}_\varepsilon \widehat{R}_n\big(\mathcal{G}'_{N+1}\big) + 2\mathbb{E}I_{E_N^c}\mathbb{E}_\varepsilon \widehat{R}_n\big(\mathcal{G}'_{N+1}\big). \tag{3.11}$$

As above, we have

$$\begin{aligned}\mathbb{E}_\varepsilon R_n(\mathcal{G}'_{N+1}) &\le \inf_{g\in\mathcal{G}'_{N+1}} \mathbb{E}_\varepsilon\left|n^{-1}\sum_{j=1}^n \varepsilon_j g(X_j)\right| \\ &\quad + \frac{c}{\sqrt{n}}\int_0^{(2\sup_{g\in\mathcal{G}'_{N+1}} P_n g^2)^{1/2}} H^{1/2}_{d_{P_n,2}}\big(\mathcal{G}'_{N+1}; u\big)\,du.\end{aligned} \tag{3.12}$$

Since we already proved (i) it implies that on the event $E_{N+1} \cap \mathcal{J}$

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

$$\le \frac{1}{\sqrt{n}} \inf_{f\in\mathcal{F}_{N+1}} \sqrt{P_n\{f \le \delta_{N,\frac{1}{2}}\}}$$
$$\le \frac{1}{\sqrt{n}} \inf_{f\in\mathcal{F}_{N+1}} \sqrt{P_n\{f \le \delta\}}$$
$$\le \sqrt{\frac{\varepsilon}{n}} \le \varepsilon.$$

By the induction assumption, we also have on the event $E_{N+1}\cap\mathcal{J}$,

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

The bound for the Lipschitz norm of φ'_k gives the following bound on the distance

$$d^2_{P_n,2}(\varphi'_{N+1}\circ f;\varphi'_{N+1}\circ g) = n^{-1}\sum_{j=1}^n |\varphi'_{N+1}\circ f(X_j) - \varphi'_{N+1}\circ g(X_j)|^2$$
$$\le \left(\frac{2}{\delta}\sqrt{\frac{r_N}{\varepsilon}}\right)^2 d^2_{P_n,2}(f,g).$$

Therefore, on the event $E_{N+1}\cap\mathcal{J}$, we get quite similarly to (3.7)

$$\frac{1}{\sqrt{n}}\int_0^{(2\sup_{g\in\mathcal{G}'_{N+1}} P_n g^2)^{1/2}} H^{1/2}_{d_{P_n,2}}(\mathcal{G}'_{N+1};u)\,du$$
$$\le \frac{1}{\sqrt{n}}\int_0^{(2r_N)^{1/2}} H^{1/2}_{d_{P_n,2}}\left(\mathcal{F};\frac{\delta\sqrt{\varepsilon}u}{2\sqrt{r_N}}\right)du \tag{3.13}$$
$$\le c\left(\frac{r_N}{\varepsilon}\right)^{\alpha/4} \frac{r_N^{1/2-\alpha/4}}{\sqrt{n}\delta^{\alpha/2}} \le c\sqrt{r_N\varepsilon}.$$

We collect all bounds to see that on the event $E_{N+1}\cap\mathcal{J}$,

$$\sup_{f\in\mathcal{F}_{N+1}} P_n\{f\le\delta_{N+1}\} \le \frac{r_{N+1}}{2} + c\sqrt{r_N\varepsilon}. \tag{3.14}$$

Therefore, it follows that with a proper choice of constant $C>0$ in the recurrence relationship defining the sequence $\{r_k\}$, we have 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 proves the induction step for (ii) and, therefore, the lemma is proved. □

To complete the proof of the theorem, we have to note that the choice of $N=[\log_2\log_2\varepsilon^{-1}]$ implies that $r_{N+1}\le c\varepsilon$ for some $c>0$. The second inequality of the theorem can be proved similarly with some minor modifications. □

PROOF OF THEOREM 5. Consider sequences $\delta_j := 2^{-j\frac{2}{\gamma}}$,

$$\varepsilon_j := \left(\frac{1}{n\delta_j^{\alpha'}}\right)^{\frac{2}{2+\alpha'}}, \qquad j \geq 0,$$

where $\alpha' := \frac{2\gamma}{2-\gamma} \geq \alpha$. The first inequality of Theorem 6 implies

$$\mathbb{P}\{\exists j \geq 0\, \exists f \in \mathcal{F}\ P_n\{f \leq \delta_j\} \leq \varepsilon_j \text{ and } P\{f \leq \delta_j/2\} \geq A'\varepsilon_j\}$$

$$\leq B' \log_2 \log_2 n \sum_{j\geq 0} \exp\left\{-\frac{n^{\gamma/2}}{2} 2^{2j}\right\} \leq B \log_2 \log_2 n \exp\left\{-\frac{n^{\gamma/2}}{2}\right\} \tag{3.15}$$

with some $B, B', A' > 0$. If for some $j \geq 1$, we have

$$\hat{\delta}_n(\gamma; f) \in (\delta_j, \delta_{j-1}],$$

then by the definition of $\hat{\delta}_n(\gamma; f)$,

$$P_n\{f \leq \delta_j\} \leq \varepsilon_j.$$

Suppose that for some $f \in \mathcal{F}$ the inequality $A^{-1}\hat{\delta}_n(\gamma; f) \leq \delta_n(\gamma; f)$ fails. Then, it follows from the definition of $\delta_n(\gamma; f)$ that

$$P\{f \leq \delta_j/2\} \geq P\left\{f \leq \frac{\delta_{j-1}}{A}\right\} \geq \left(\frac{1}{n\delta_{j-1}^{\alpha'}}\right)^{\frac{2}{2+\alpha'}} A^{\frac{2\alpha'}{2+\alpha'}} \geq A'\varepsilon_j,$$

where the last inequality holds for the proper choice of a constant A. Hence, (3.15) guarantees the probability bound for the left side inequality of the theorem. The right side inequality is proved similarly utilizing the second inequality of Theorem 6. □

4. Convergence rates of empirical margin distributions. As we defined in Section 2, $\mathcal{F}$ is a class of measurable functions from S into $\mathbb{R}$. For $f \in \mathcal{F}$, let

$$F_f(y) := P\{f \leq y\},\ F_{n,f}(y) := P_n\{f \leq y\}, \qquad y \in \mathbb{R}.$$

Let L denote the Lévy distance between the distribution functions in $\mathbb{R}$:

$$L(F, G) := \inf\big\{\delta > 0 : F(t) \leq G(t+\delta) + \delta \\ \text{and } G(t) \leq F(t+\delta) + \delta, \text{ for all } t \in \mathbb{R}\big\}.$$

In what follows, for a function f from S into $\mathbb{R}$ and $M > 0$, we denote by f_M the function that is equal to f if $|f| \leq M$, is equal to M if $f > M$ and is equal to $-M$ if $f < -M$. We set

$$\mathcal{F}_M := \{f_M : f \in \mathcal{F}\}.$$

As always, a function F from S into $[0,+\infty)$ is called an envelope of $\mathcal{F}$ iff $|f(x)| \le F(x)$ for all $f \in \mathcal{F}$ and all $x \in S$.

We write $\mathcal{F} \in GC(P)$ iff $\mathcal{F}$ is a Glivenko–Cantelli class with respect to P (i.e., $\|P_n - P\|_{\mathcal{F}} \to 0$ as $n \to \infty$ a.s.). We write $\mathcal{F} \in BCLT(P)$ and say that $\mathcal{F}$ satisfies the Bounded Central Limit Theorem for P iff

$$\mathbb{E}\|P_n - P\|_{\mathcal{F}} = O\big(n^{-1/2}\big).$$

In particular, this holds if $\mathcal{F}$ is a P-Donsker class [see Dudley (1999) and van der Vaart and Wellner (1996) for precise definitions].

Our main goal in this section is to prove the following results.

THEOREM 7. *Suppose that*

$$\sup_{f\in\mathcal{F}} P\{|f| \ge M\} \to 0 \qquad \text{as } M \to \infty. \tag{4.1}$$

Then, the following two statements are equivalent:

(i) $$\mathcal{F}_M \in GC(P) \qquad \text{for all } M > 0$$

and

(ii) $$\sup_{f\in\mathcal{F}} L(F_{n,f}, F_f) \to 0 \qquad \text{a.s. as } n \to \infty.$$

THEOREM 8. *The following two statements are equivalent*:

(i) $\mathcal{F} \in GC(P)$;

(ii) *there exists a P-integrable envelope for the class $\mathcal{F}^{(c)} = \{f - Pf : f \in \mathcal{F}\}$ and*

$$\sup_{f\in\mathcal{F}} L(F_{n,f}, F_f) \to 0 \qquad \text{a.s. as } n \to \infty.$$

THEOREM 9. *Suppose that the class $\mathcal{F}$ is uniformly bounded. If $\mathcal{F} \in BCLT(P)$, then*

$$\sup_{f\in\mathcal{F}} L(F_{n,f}, F_f) = O_P(n^{-1/4}) \qquad \text{as } n \to \infty.$$

Moreover, if for some $\alpha \in (0,2)$ and for some $D > 0$

$$H_{d_{P_n,2}}(\mathcal{F}; u) \le Du^{-\alpha}, \qquad u > 0 \text{ a.s.}, \tag{4.2}$$

then

$$\sup_{f\in\mathcal{F}} L(F_{n,f}, F_f) = O\big(n^{-\frac{1}{2+\alpha}}\big) \qquad \text{as } n \to \infty \text{ a.s.}$$

The following theorem gives the bound that plays an important role in the proofs.

THEOREM 10. *Let $M > 0$ and let $\mathcal{F}$ be a class of measurable functions from S into $[-M, M]$. For all $t > 0$,*

$$\mathbb{P}\left\{ \sup_{f\in\mathcal{F}} L(F_{n,f}, F_f) \geq 2\left(\mathbb{E}\left\|n^{-1}\sum_{i=1}^{n}\varepsilon_i\delta_{X_i}\right\|_{\mathcal{F}} + \frac{M}{\sqrt{n}}\right)^{1/2} + \frac{t}{\sqrt{n}}\right\} \leq \exp\{-2t^2\}.$$

PROOF. Let $\delta > 0$. Let $\varphi(x)$ be equal to 1 for $x \leq 0$, 0 for $x \geq 1$ and linear in between. One can get the following bounds:

$$\begin{aligned} F_f(y) = P\{f \leq y\} &\leq P\varphi\left(\frac{f-y}{\delta}\right) \leq P_n\varphi\left(\frac{f-y}{\delta}\right) + \|P_n - P\|_{\widetilde{\mathcal{G}}_\delta} \\ &\leq F_{n,f}(y+\delta) + \|P_n - P\|_{\widetilde{\mathcal{G}}_\delta} \end{aligned}$$

and

$$\begin{aligned} F_{n,f}(y) = P_n\{f \leq y\} &\leq P_n\varphi\left(\frac{f-y}{\delta}\right) \leq P\varphi\left(\frac{f-y}{\delta}\right) + \|P_n - P\|_{\widetilde{\mathcal{G}}_\delta} \\ &\leq F_f(y+\delta) + \|P_n - P\|_{\widetilde{\mathcal{G}}_\delta}, \end{aligned}$$

where

$$\widetilde{\mathcal{G}}_\delta := \left\{\varphi\circ\left(\frac{f-y}{\delta}\right) - 1 : f \in \mathcal{F}, y \in [-M, M]\right\}.$$

Similarly to the proof of Theorem 1 we get that with probability at least $1 - 2e^{-2t^2}$,

$$\|P_n - P\|_{\widetilde{\mathcal{G}}_\delta} \leq \frac{4}{\delta}\left[\mathbb{E}\left\|n^{-1}\sum_{i=1}^{n}\varepsilon_i\delta_{X_i}\right\|_{\mathcal{F}} + Mn^{-1/2}\right] + \frac{t}{\sqrt{n}}. \tag{4.3}$$

Setting

$$\delta := 2\left(\mathbb{E}\left\|n^{-1}\sum_{i=1}^{n}\varepsilon_i\delta_{X_i}\right\|_{\mathcal{F}} + Mn^{-1/2}\right)^{1/2},$$

we get that with probablity at least $1 - \exp\{-2t^2\}$,

$$\sup_{f\in\mathcal{F}} L(F_{n,f}, F_f) \leq 2\left(\mathbb{E}\left\|n^{-1}\sum_{i=1}^{n}\varepsilon_i\delta_{X_i}\right\|_{\mathcal{F}} + Mn^{-1/2}\right)^{1/2} + \frac{t}{\sqrt{n}},$$

which completes the proof. □

PROOF OF THEOREM 7. First we prove that (i) implies (ii). Since $\mathcal{F}_M \in GC(P)$, we have

$$\mathbb{E}\|P_n - P\|_{\mathcal{F}_M} \to 0 \qquad \text{as } n \to \infty,$$

which, by symmetrization inequality, implies

$$\mathbb{E}\|n^{-1}\sum_{i=1}^{n}\varepsilon_i\delta_{X_i}\|_{\mathcal{F}_M}\to 0 \qquad \text{as } n\to\infty.$$

Plugging in the bound of Theorem 10 $t=\log n$ and using the Borel–Cantelli Lemma proves that for all $M>0$

$$\sup_{f\in\mathcal{F}} L(F_{n,f_M},F_{f_M}) = \sup_{f\in\mathcal{F}_M} L(F_{n,f},F_f)\to 0 \qquad \text{as } n\to\infty \text{ a.s.}$$

The following bounds easily follow from the definition of Lévy distance:

$$\sup_{f\in\mathcal{F}} L(F_f,F_{f_M}) \le \sup_{f\in\mathcal{F}} P\{|f|\ge M\}$$

and

$$\sup_{f\in\mathcal{F}} L(F_{n,f},F_{n,f_M}) \le \sup_{f\in\mathcal{F}} P_n\{|f|\ge M\}.$$

By condition (4.1) of the theorem,

$$\sup_{f\in\mathcal{F}} L(F_f,F_{f_M})\to 0 \qquad \text{as } M\to\infty.$$

To prove that also

$$\lim_{M\to\infty}\limsup_{n\to\infty}\sup_{f\in\mathcal{F}} L(F_{n,f},F_{n,f_M})=0 \qquad \text{a.s.},$$

it is enough to show that

$$\lim_{M\to\infty}\limsup_{n\to\infty}\sup_{f\in\mathcal{F}} P_n\{|f|\ge M\}=0 \qquad \text{a.s.} \tag{4.4}$$

To this end, consider the function φ from $\mathbb{R}$ into $[0,1]$ that is equal to 0 for $|u|\le M-1$, is equal to 1 for $|u|>M$ and is linear in between. We have

$$\begin{aligned}
\sup_{f\in\mathcal{F}} P_n\{|f|\ge M\} &= \sup_{f\in\mathcal{F}_M} P_n\{|f|\ge M\}\\
&\le \sup_{f\in\mathcal{F}_M} P_n\varphi(|f|)\\
&\le \sup_{f\in\mathcal{F}_M} P\varphi(|f|)+\|P_n-P\|_{\mathcal{G}}\\
&\le \sup_{f\in\mathcal{F}_M} P\{|f|\ge M-1\}+\|P_n-P\|_{\mathcal{G}},
\end{aligned} \tag{4.5}$$

where

$$\mathcal{G}:=\{\varphi\circ f : f\in\mathcal{F}_M\}.$$

Since φ satisfies the Lipschitz condition with constant 1, the argument based on symmetrization inequality and comparison inequalities (see the proofs above) allows one to show that the condition (i) implies that

$$\mathbb{E}\|P_n - P\|_{\mathcal{G}} \to 0 \qquad \text{as } n \to \infty.$$

Then, the standard use of concentration inequality implies that

$$\|P_n - P\|_{\mathcal{G}} \to 0 \qquad \text{as } n \to \infty \text{ a.s.}$$

Therefore, (4.4) immediately follows from condition (4.1) and (4.5). Now, the triangle inequality for the Lévy distance allows one easily to complete the proof of (ii).

To prove that (ii) implies (i), we use the bound

$$\left| \int_{-M}^{M} t d(F - G)(t) \right| \le cL(F, G),$$

which holds with some constant $c = c(M)$ for any two distribution functions on $[-M, M]$. The bound implies that

$$\begin{aligned} \|P_n - P\|_{\mathcal{F}_M} &= \sup_{f \in \mathcal{F}_M} |P_n f - Pf| = \sup_{f \in \mathcal{F}_M} \left| \int_{-M}^{M} t d(F_{n,f} - F_f)(t) \right| \\ &\le c \sup_{f \in \mathcal{F}_M} L(F_{n,f}; F_f). \end{aligned} \tag{4.6}$$

Since for all $M > 0$ and for all $f \in \mathcal{F}$ it is easily proved that

$$L(F_{n,f_M}, F_{f_M}) \le L(F_{n,f}, F_f), \tag{4.7}$$

the bound (4.6) and condition (ii) imply (i), which completes the proof of the second statement. □

PROOF OF THEOREM 8. Since centering does not change Lévy distance and does not change Glivenko–Cantelli property we can start by assuming that $\mathcal{F}$ is centered, that is, $\mathcal{F} = \mathcal{F}^{(c)}$. To prove that (i) implies (ii), note first of all that the condition $\mathcal{F} \in GC(P)$ yields that $\mathcal{F} = \mathcal{F}^{(c)}$ has a P-integrable envelope [see van der Vaart and Wellner (1996), page 125]. Also, the existence of a P-integrable envelope implies (4.1). Finally, if $\mathcal{F} \in GC(P)$, then for all $M > 0$ $\mathcal{F}_M \in GC(P)$. [To prove this claim note that $f_M = \varphi_M \circ f$, where φ_M is the function from $\mathbb{R}$ into $[-M, M]$ that is equal to u for $|u| \le M$, is equal to M for $u > M$ and is equal to $-M$ for $u < -M$. The function φ_M is Lipschitz with constant 1 which allows one to prove the claim by the argument based on the comparison inequality and used many times above.] We can use Theorem 7 to conclude that (i) implies (ii). On the other hand, if (ii) holds then by the inequality (4.7) we get that

$$\sup_{f \in \mathcal{F}_M} L(F_{n,f}, F_f) \to 0 \qquad \text{as } n \to \infty \text{ a.s.}$$

As we pointed out above, (4.1) holds so, by Theorem 7, we have $\mathcal{F}_M \in GC(P)$ for all $M > 0$. The integrability of the envelope of the class $\mathcal{F}$ allows us to conclude the proof of (i) by a standard truncation argument. $\square$

PROOF OF THEOREM 9. Since $\mathcal{F}$ is uniformly bounded, we can choose $M > 0$ such that $\mathcal{F}_M = \mathcal{F}$. To prove the first statement note that $\mathcal{F} \in BCLT(P)$ means that

$$\mathbb{E}\|P_n - P\|_{\mathcal{F}} = O\big(n^{-1/2}\big),$$

which implies

$$\mathbb{E}\left\|n^{-1}\sum_{i=1}^{n}\varepsilon_i\delta_{X_i}\right\|_{\mathcal{F}} = O\big(n^{-1/2}\big).$$

Thus, the bound of Theorem 10 implies that with some constant $C > 0$,

$$\mathbb{P}\left\{\sup_{f\in\mathcal{F}} L(F_{n,f}, F_f) \ge \left(\frac{C}{\sqrt{n}} + \frac{4M}{\sqrt{n}}\right)^{1/2} + \frac{t}{\sqrt{n}}\right\} \le \exp\{-2t^2\}.$$

It follows that

$$\lim_{u\to\infty}\limsup_{n\to\infty}\mathbb{P}\left\{n^{1/4}\sup_{f\in\mathcal{F}} L(F_{n,f}, F_f) \ge u\right\} = 0.$$

To prove the second statement, we follow the proof of Theorem 10. We use Rademacher symmetrization inequality to get the bound

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

and then use the entropy inequalities for subgaussian processes [see van der Vaart and Wellner (1996), Corollary 2.2.8] to show that

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

To bound the random entropy $H_{d_{P_n,2}}$, we use the Lipschitz condition for the function φ. It yields (via a standard argument based on constructing minimal covering of the class $\mathcal{F}$ with respect to the metric $d_{P_n,2}$ and of the interval $[-M, M]$ with respect to the usual distance on the real line and "combining" the coverings properly) the following bound:

$$H_{d_{P_n,2}}(\tilde{\mathcal{G}}_\delta; u) \le H_{d_{P_n,2}}(\mathcal{F}; \delta u/2) + \log\frac{4M}{u\delta}.$$

Therefore, we get (with a proper constant $c > 0$)

$$\mathbb{E}_\varepsilon \widehat{R}_n(\widetilde{\mathcal{G}}_\delta) \le \frac{c}{\sqrt{n}}\left[\int_0^{\sqrt{2}} H^{1/2}_{d_{P_n,2}}(\mathcal{F}; \delta u)\, du + \sqrt{\log \frac{4M}{\delta}} + 1\right],$$

which, under condition (4.2), is bounded from above by $\frac{c}{\sqrt{n}\delta^{\alpha/2}}$. Thus, we have proved the bound

$$\mathbb{E}\|P_n - P\|_{\widetilde{\mathcal{G}}_\delta} \le \frac{c}{\sqrt{n}\delta^{\alpha/2}}.$$

Arguing now the same way as in the proof of Theorem 10, we can show that with probability at least $1 - \exp\{-2t^2\}$,

$$\sup_{f\in\mathcal{F}} L(F_{n,f}, F_f) \le \delta \vee \frac{c}{\sqrt{n}\delta^{\alpha/2}} + \frac{t}{\sqrt{n}}.$$

Plugging in the last inequality

$$\delta := \frac{c}{n^{\frac{1}{2+\alpha}}},$$

we get

$$\mathbb{P}\left\{\sup_{f\in\mathcal{F}} L(F_{n,f}, F_f) \ge \frac{c}{n^{\frac{1}{2+\alpha}}} + \frac{t}{\sqrt{n}}\right\} \le \exp\{-2t^2\}.$$

By choosing $t := \log n$ and using the Borel–Cantelli Lemma, we complete the proof of the second statement. □

REMARK. It is interesting to mention that the condition $\mathcal{F} \in GC(P)$ *does not* imply that

$$\sup_{f\in\mathcal{F}} \sup_{t\in\mathbb{R}} \left|F_{n,f}(t) - F_f(t)\right| \to 0$$

with probability 1, which is equivalent to saying that the class of sets $\{I(f \le t) : f \in \mathcal{F}, t \in \mathbb{R}\}$ is $GC(P)$. As an example, consider the case where S is a unit ball in an infinite-dimensional separable Banach space. Let $\mathcal{F}$ be the restriction of the unit ball in the dual space on S. For i.i.d. random variables $\{X_n\}$ in S, we have, by the LLN in separable Banach spaces,

$$\|P_n - P\|_{\mathcal{F}} := \left\|n^{-1}\sum_{j=1}^n (X_j - \mathbb{E}X)\right\| \to 0 \qquad \text{a.s.},$$

so $\mathcal{F} \in GC(P)$. On the other hand, there exists an example of a distribution P such that $\mathcal{H} \notin GC(P)$, where $\mathcal{H}$ is the class of all halfspaces [see Sazonov (1963) and also Topsøe, Dudley and Hoffmann-Jørgensen (1976)]. Hence,

$$\sup_{f\in\mathcal{F}} \sup_{t\in\mathbb{R}} \left|F_{n,f}(t) - F_f(t)\right| = \|P_n - P\|_{\mathcal{H}}$$

does not converge to 0 a.s.

In the next proposition, we are again considering the class $\mathcal{F}$ used already in Proposition 1 and the sequence of observations $\{X_n\}$ defined by

$$X_n = \{\varepsilon_k^n (2\log(k+1))^{-\frac{1}{2}-\beta}\}_{k\geq 1}, \qquad n \geq 1,$$

where $\beta := \frac{1}{\alpha} - \frac{1}{2}$, $\alpha \in (0,2]$ and ε_k^n are i.i.d. Rademacher random variables. The proposition shows the optimality of the rates of convergence obtained in Theorem 9.

PROPOSITION 2. *Consider the sequence δ_n such that*

$$\sup_{f\in\mathcal{F}} L(F_{n,f}, F_f) = O_P(\delta_n).$$

Then

$$\delta_n \geq cn^{-\frac{1}{2+\alpha}}$$

(*when $\alpha = 2$, we have $\delta_n \geq cn^{-1/4}$*). *On the other hand, for $\alpha \in (0,2)$, we have*

$$H_{d_{P_n,2}}(\mathcal{F}; u) \leq Du^{-\alpha}, \qquad u > 0$$

and

$$\sup_{f\in\mathcal{F}} L(F_{n,f}, F_f) = O(n^{-\frac{1}{2+\alpha}}) \qquad a.s.;$$

for $\alpha = 2$ we have $\mathcal{F} \in BCLT(P)$ and

$$\sup_{f\in\mathcal{F}} L(F_{n,f}, F_f) = O_P\big(n^{-\frac{1}{4}}\big).$$

PROOF. We can assume without loss of generality that with probability more than $1/2$ for all $k \geq 1$, $y \in [-1,1]$ and n large enough we have

$$P(f_k \leq y) \leq P_n(f_k \leq y + \delta) + \delta. \tag{4.8}$$

If we take $y = 0$ and consider only such k that satisfy the inequality $(2\log(k+1))^{\beta+1/2} < \delta^{-1}$ then (4.8) becomes equivalent to

$$1/2 \leq n^{-1}\sum_{i\leq n} I(\varepsilon_k^i = -1) + \delta.$$

The inequality $(2\log(k+1))^{\beta+1/2} < \delta^{-1}$ holds for $k \leq \psi_1(\delta) = 1/2\exp(\delta^{-\frac{2}{1+2\beta}}/2)$. Therefore, for large n

$$\begin{aligned} 1/2 &\leq \mathbb{P}\Big\{\bigcap_{k\leq\psi_1(\delta)} \Big\{1/2 \leq n^{-1}\sum_{i\leq n} I(\varepsilon_k^i = -1) + \delta\Big\}\Big\} \\ &= \mathbb{P}\Big\{1/2 \leq n^{-1}\sum_{i\leq n} I(\varepsilon_1^i = -1) + \delta\Big\}^{\psi_1(\delta)} \leq \Big(1 - \binom{n}{k_0}2^{-n}\Big)^{\psi_1(\delta)}, \end{aligned} \tag{4.9}$$

where $k_0 = [n/2 - \delta n] - 1$. Using (2.11), we get

$$2^{-\frac{1}{\psi_1(\delta)}} \le 1 - cn^{-\frac{1}{2}} \exp\bigl(-4n\delta^2\bigr).$$

Taking logarithms of both sides and taking into account that $\log(1-x) \le -x$ we get [recall that $\psi_1(\delta) = 1/2\exp(\delta^{-\frac{2}{1+2\beta}}/2)$]

$$\exp\bigl(-2^{-1}\delta^{-\frac{2}{1+2\beta}}\bigr) \ge cn^{-\frac{1}{2}} \exp\bigl(-4n\delta^2\bigr).$$

Therefore,

$$1/\bigl(2\delta^{2/(1+2\beta)}\bigr) \le 4n\delta^2 + c\log n$$

and

$$1/2 \le 4n\delta^{4(1+\beta)/(1+2\beta)} + c\delta^{2/(1+2\beta)}\log n.$$

This finally implies that

$$\delta \ge cn^{-\frac{1+2\beta}{4(1+\beta)}} = cn^{-\frac{1}{2+\alpha}}.$$

The second statement follows from Theorem 9. To check condition (4.2), note that in this case, as soon as $2\log N \ge (u/2)^{-\alpha}$, we have $|f_k(X_n)| \le u/2$ for all $k \ge N$ and $n \ge 1$. Hence,

$$d_{P_n,2}(f_k, f_N) \le u, \qquad k \ge N$$

and we have

$$H_{d_{P_n,2}}(\mathcal{F}; u) \le \log N,$$

which implies (4.2). For $\alpha = 2$, we also have $\mathcal{F} \in BCLT(P)$ [see Ledoux and Talagrand (1991), pages 276–277]. Theorem 9 allows one to complete the proof. $\square$

5. Bounding the generalization error of convex combinations of classifiers. In this and in the next section we consider applications of the bounds of Section 2 to various learning (classification) problems. We start with an application of the inequalities of Section 2 to bounding the generalization error in general multiclass problems. Namely, we assume that the labels take values in a finite set $\mathcal{Y}$ with $\operatorname{card}(\mathcal{Y}) = M$. Consider a class $\widetilde{\mathcal{F}}$ of functions from $\widetilde{S} := S \times \mathcal{Y}$ into $\mathbb{R}$. A function $f \in \widetilde{\mathcal{F}}$ predicts a label $y \in \mathcal{Y}$ for an example $x \in S$ iff

$$f(x, y) > \max_{y' \ne y} f\bigl(x, y'\bigr).$$

The margin of a labeled example (x, y) is defined as

$$m_f(x, y) := f(x, y) - \max_{y' \ne y} f\bigl(x, y'\bigr),$$

so f misclassifies the labeled example (x, y) iff $m_f(x, y) \le 0$. Let

$$\mathcal{F} := \bigl\{f(\cdot, y) : y \in \mathcal{Y},\ f \in \tilde{\mathcal{F}}\bigr\}.$$

The proof of the next result is based on the application of Theorem 2.

THEOREM 11. *For all* $t > 0$,

$$\mathbb{P}\Biggl\{\exists f \in \widetilde{\mathcal{F}} : P\{m_f \le 0\} > \inf_{\delta\in(0,1]}\Biggl[P_n\{m_f \le \delta\} + \frac{8M(2M-1)}{\delta}R_n(\mathcal{F}) + \biggl(\frac{\log\log_2(2\delta^{-1})}{n}\biggr)^{1/2}\Biggr] + \frac{t}{\sqrt{n}}\Biggr\}$$
$$\le 2\exp\{-2t^2\}.$$

To prove the theorem, we use the following lemma.
For a class of functions $\mathcal{H}$, we will denote

$$\mathcal{H}^{(l)} = \{\max(h_1,\dots,h_l) : h_1,\dots,h_l \in \mathcal{H}\}.$$

LEMMA 2. *The following bound holds*:

$$\mathbb{E}\biggl\|\sum_{i=1}^n \varepsilon_i\delta_{X_i}\biggr\|_{\mathcal{H}^{(l)}} \le 2l\,\mathbb{E}\biggl\|\sum_{i=1}^n \varepsilon_i\delta_{X_i}\biggr\|_{\mathcal{H}}.$$

PROOF. Let $x^+ := x \vee 0$. Obviously $x \mapsto x^+$ is a nondecreasing convex function such that $(a+b)^+ \le a^+ + b^+$. We will first prove that

$$\mathbb{E}\biggl(\sup_{\mathcal{H}^{(l)}}\sum_{i=1}^n \varepsilon_i h(X_i)\biggr)^+ \le l\,\mathbb{E}\biggl(\sup_{\mathcal{H}}\sum_{i=1}^n \varepsilon_i h(X_i)\biggr)^+. \tag{5.1}$$

Let us consider classes of functions $\mathcal{F}_1$, $\mathcal{F}_2$ and

$$\mathcal{F} = \{\max(f_1, f_2) : f_1 \in \mathcal{F}_1,\ f_2 \in \mathcal{F}_2\}.$$

Since

$$\max(f_1, f_2) = \tfrac{1}{2}\bigl((f_1+f_2) + |f_1 - f_2|\bigr),$$

we have

$$\mathbb{E}\biggl(\sup_{\mathcal{F}}\sum_{i=1}^n \varepsilon_i f(X_i)\biggr)^+$$
$$\le \mathbb{E}\biggl(\sup_{\mathcal{F}_1,\mathcal{F}_2}\sum_{i=1}^n \varepsilon_i \tfrac{1}{2}\bigl(f_1(X_i)+f_2(X_i)\bigr) + \sup_{\mathcal{F}_1,\mathcal{F}_2}\sum_{i=1}^n \varepsilon_i \tfrac{1}{2}|f_1(X_i)-f_2(X_i)|\biggr)^+$$
$$\le \tfrac{1}{2}\mathbb{E}\biggl(\sup_{\mathcal{F}_1,\mathcal{F}_2}\sum_{i=1}^n \varepsilon_i\bigl(f_1(X_i)+f_2(X_i)\bigr)\biggr)^+ + \tfrac{1}{2}\mathbb{E}\biggl(\sup_{\mathcal{F}_1,\mathcal{F}_2}\sum_{i=1}^n \varepsilon_i|f_1(X_i)-f_2(X_i)|\biggr)^+$$

$$\le \tfrac{1}{2}\mathbb{E}\left(\sup_{\mathcal{F}_1}\sum_{i=1}^n \varepsilon_i f_1(X_i)\right)^+ + \tfrac{1}{2}\mathbb{E}\left(\sup_{\mathcal{F}_2}\sum_{i=1}^n \varepsilon_i f_2(X_i)\right)^+ + \tfrac{1}{2}\mathbb{E}\left(\sup_{\mathcal{F}_1,\mathcal{F}_2}\sum_{i=1}^n \varepsilon_i \big|f_1(X_i) - f_2(X_i)\big|\right)^+.$$

The proof of Theorem 4.12 in Ledoux and Talagrand (1991) contains the following statement. If T is a bounded subset of $\mathbb{R}^n$, functions φ_i, $i = 1, \dots, n$, are contractions such that $\varphi_i(0) = 0$ and a function $G : \mathbb{R} \to \mathbb{R}$ is convex and nondecreasing then

$$\mathbb{E}G\left(\sup_{t\in T}\sum_{i=1}^n \varepsilon_i \varphi_i(t_i)\right) \le \mathbb{E}G\left(\sup_{t\in T}\sum_{i=1}^n \varepsilon_i t_i\right).$$

If we take $G(x) = x^+$, $\varphi_i(x) = |x|$ and $T = \{(f_1(X_i) - f_2(X_i))_{i=1}^n : f_1 \in \mathcal{F}_1, f_2 \in \mathcal{F}_2\}$ we get [first conditionally on $(X_i)_{i=1}^n$ and then taking expectations]

$$\begin{aligned}\mathbb{E}\left(\sup_{\mathcal{F}_1,\mathcal{F}_2}\sum_{i=1}^n \varepsilon_i \big|f_1(X_i) - f_2(X_i)\big|\right)^+ &\le \mathbb{E}\left(\sup_{\mathcal{F}_1,\mathcal{F}_2}\sum_{i=1}^n \varepsilon_i \big(f_1(X_i) - f_2(X_i)\big)\right)^+ \\ &\le \mathbb{E}\left(\sup_{\mathcal{F}_1}\sum_{i=1}^n \varepsilon_i f_1(X_i)\right)^+ \\ &\quad + \mathbb{E}\left(\sup_{\mathcal{F}_2}\sum_{i=1}^n \varepsilon_i f_2(X_i)\right)^+,\end{aligned}$$

where in the last inequality we used the fact that the sequence $(-\varepsilon_i)_{i=1}^n$ is equal in distribution to $(\varepsilon_i)_{i=1}^n$. Combining the bounds gives

$$\mathbb{E}\left(\sup_{\mathcal{F}}\sum_{i=1}^n \varepsilon_i f(X_i)\right)^+ \le \mathbb{E}\left(\sup_{\mathcal{F}_1}\sum_{i=1}^n \varepsilon_i f_1(X_i)\right)^+ + \mathbb{E}\left(\sup_{\mathcal{F}_2}\sum_{i=1}^n \varepsilon_i f_2(X_i)\right)^+.$$

Now by induction we easily get (5.1). Finally, again using the fact that $(-\varepsilon_i)_{i=1}^n$ is equal in distribution to $(\varepsilon_i)_{i=1}^n$, we conclude the proof:

$$\begin{aligned}\mathbb{E}\left\|\sum_{i=1}^n \varepsilon_i \delta_{X_i}\right\|_{\mathcal{H}^{(l)}} &\le \mathbb{E}\left(\sup_{\mathcal{H}^{(l)}}\sum_{i=1}^n \varepsilon_i h(X_i)\right)^+ + \mathbb{E}\left(-\sup_{\mathcal{H}^{(l)}}\sum_{i=1}^n \varepsilon_i h(X_i)\right)^+ \\ &= 2\mathbb{E}\left(\sup_{\mathcal{H}^{(l)}}\sum_{i=1}^n \varepsilon_i h(X_i)\right)^+ \le 2l\,\mathbb{E}\left(\sup_{\mathcal{H}}\sum_{i=1}^n \varepsilon_i h(X_i)\right)^+ \\ &\le 2l\,\mathbb{E}\left\|\sum_{i=1}^n \varepsilon_i \delta_{X_i}\right\|_{\mathcal{H}}.\end{aligned}$$

□

PROOF OF THEOREM 11. We have the following bounds:

$$\begin{aligned}
&\mathbb{E}\sup_{f\in\widetilde{\mathcal{F}}}\left|n^{-1}\sum_{j=1}^{n}\varepsilon_j m_f(X_j,Y_j)\right| \\
&\quad=\mathbb{E}\sup_{f\in\widetilde{\mathcal{F}}}\left|n^{-1}\sum_{j=1}^{n}\varepsilon_j\sum_{y\in\mathcal{Y}}m_f(X_j,y)I_{\{Y_j=y\}}\right| \\
&\quad\le\sum_{y\in\mathcal{Y}}\mathbb{E}\sup_{f\in\widetilde{\mathcal{F}}}\left|n^{-1}\sum_{j=1}^{n}\varepsilon_j m_f(X_j,y)I_{\{Y_j=y\}}\right| \\
&\quad\le\tfrac{1}{2}\sum_{y\in\mathcal{Y}}\mathbb{E}\sup_{f\in\widetilde{\mathcal{F}}}\left|n^{-1}\sum_{j=1}^{n}\varepsilon_j m_f(X_j,y)(2I_{\{Y_j=y\}}-1)\right| \\
&\qquad+\tfrac{1}{2}\sum_{y\in\mathcal{Y}}\mathbb{E}\sup_{f\in\widetilde{\mathcal{F}}}\left|n^{-1}\sum_{j=1}^{n}\varepsilon_j m_f(X_j,y)\right|.
\end{aligned}$$

Denote $\sigma_j(y):=2I_{\{Y_j=y\}}-1$. Given $\{(X_j,Y_j):1\le j\le n\}$, the random variables $\{\varepsilon_j\sigma_j(y):1\le j\le n\}$ are i.i.d. Rademacher. Hence, we have

$$\begin{aligned}
&\mathbb{E}\sup_{f\in\widetilde{\mathcal{F}}}\left|n^{-1}\sum_{j=1}^{n}\varepsilon_j m_f(X_j,y)(2I_{\{Y_j=y\}}-1)\right| \\
&\quad=\mathbb{E}\sup_{f\in\widetilde{\mathcal{F}}}\left|n^{-1}\sum_{j=1}^{n}\varepsilon_j\sigma_j(y)m_f(X_j,y)\right| \\
&\quad=\mathbb{E}\mathbb{E}_\varepsilon\sup_{f\in\widetilde{\mathcal{F}}}\left|n^{-1}\sum_{j=1}^{n}\varepsilon_j\sigma_j(y)m_f(X_j,y)\right| \\
&\quad=\mathbb{E}\mathbb{E}_\varepsilon\sup_{f\in\widetilde{\mathcal{F}}}\left|n^{-1}\sum_{j=1}^{n}\varepsilon_j m_f(X_j,y)\right| \\
&\quad=\mathbb{E}\sup_{f\in\widetilde{\mathcal{F}}}\left|n^{-1}\sum_{j=1}^{n}\varepsilon_j m_f(X_j,y)\right|.
\end{aligned}$$

Therefore, we have

$$\mathbb{E}\sup_{f\in\widetilde{\mathcal{F}}}\left|n^{-1}\sum_{j=1}^{n}\varepsilon_j m_f(X_j,Y_j)\right|\le\sum_{y\in\mathcal{Y}}\mathbb{E}\sup_{f\in\widetilde{\mathcal{F}}}\left|n^{-1}\sum_{j=1}^{n}\varepsilon_j m_f(X_j,y)\right|.$$

Next, using Lemma 2, we get for all $y \in \mathcal{Y}$,

$$\begin{aligned}
\mathbb{E} \sup_{f \in \widetilde{\mathcal{F}}} \left| n^{-1} \sum_{j=1}^{n} \varepsilon_j m_f(X_j, y) \right| &\le \mathbb{E} \sup_{f \in \widetilde{\mathcal{F}}} \left| n^{-1} \sum_{j=1}^{n} \varepsilon_j f(X_j, y) \right| \\
&\quad + \mathbb{E} \sup_{f \in \widetilde{\mathcal{F}}} \left| n^{-1} \sum_{j=1}^{n} \varepsilon_j \max_{y' \neq y} f(X_j, y') \right| \\
&\le \mathbb{E} \sup_{f \in \mathcal{F}} \left| n^{-1} \sum_{j=1}^{n} \varepsilon_j f(X_j) \right| \\
&\quad + \mathbb{E} \sup_{f \in \mathcal{F}^{(M-1)}} \left| n^{-1} \sum_{j=1}^{n} \varepsilon_j f(X_j) \right| \\
&\le (2M-1) \mathbb{E} \sup_{f \in \mathcal{F}} \left| n^{-1} \sum_{j=1}^{n} \varepsilon_j f(X_j) \right|.
\end{aligned}$$

This implies

$$\begin{aligned}
\mathbb{E} \sup_{f \in \widetilde{\mathcal{F}}} \left| n^{-1} \sum_{j=1}^{n} \varepsilon_j m_f(X_j, Y_j) \right| &\le \sum_{y \in \mathcal{Y}} (2M-1) \mathbb{E} \sup_{f \in \mathcal{F}} \left| n^{-1} \sum_{j=1}^{n} \varepsilon_j f(X_j) \right| \\
&= M(2M-1) \mathbb{E} \sup_{f \in \mathcal{F}} \left| n^{-1} \sum_{j=1}^{n} \varepsilon_j f(X_j) \right|,
\end{aligned}$$

and the result follows from Theorem 2 (one can use in this theorem the continuous function φ that is equal to 1 on $(-\infty, 0]$, is equal to 0 on $[1, +\infty)$ and is linear in between). $\square$

In the rest of the paper, we assume that the set of labels is $\{-1, 1\}$, so that $\tilde{S} := S \times \{-1, 1\}$ and $\widetilde{\mathcal{F}} := \{\tilde{f} : f \in \mathcal{F}\}$, where $\tilde{f}(x, y) := yf(x)$. P will denote the distribution of (X, Y), P_n the empirical distribution based on the observations $((X_1, Y_1), \ldots, (X_n, Y_n))$. Clearly, we have

$$R_n(\widetilde{\mathcal{F}}) = \mathbb{E} \sup_{f \in \mathcal{F}} \left| n^{-1} \sum_{i=1}^{n} \varepsilon_i Y_i f(X_i) \right| = \mathbb{E}\mathbb{E}_\varepsilon \sup_{f \in \mathcal{F}} \left| n^{-1} \sum_{i=1}^{n} \tilde{\varepsilon}_i f(X_i) \right|,$$

where $\tilde{\varepsilon}_i := Y_i \varepsilon_i$. Since, for given $\{(X_i, Y_i)\}$, $\{\tilde{\varepsilon}_i\}$ and $\{\varepsilon_i\}$ have the same distribution, we get

$$\mathbb{E}_\varepsilon \sup_{f \in \mathcal{F}} \left| n^{-1} \sum_{i=1}^{n} \tilde{\varepsilon}_i f(X_i) \right| = \mathbb{E}_\varepsilon \sup_{f \in \mathcal{F}} \left| n^{-1} \sum_{i=1}^{n} \varepsilon_i f(X_i) \right|,$$

which immediately implies $R_n(\widetilde{\mathcal{F}}) = R_n(\mathcal{F})$.

The results of Section 2 now give some useful bounds for boosting and other methods of combining the classifiers. Namely, we get in this case the following theorem [compare with the recent result of Schapire, Freund, Bartlett and Lee (1998)].

Given a class $\mathcal{H}$ of measurable functions from S into $\mathbb{R}$, we denote by $\operatorname{conv}(\mathcal{H})$ the closed convex hull of $\mathcal{H}$, that is, $\operatorname{conv}(\mathcal{H})$ consists of all functions on S that are pointwise limits of convex combinations of functions from $\mathcal{H}$:

$$\operatorname{conv}(\mathcal{H}) := \Biggl\{ f : \forall x \in S\ f(x) = \lim f_N(x), \\ f_N = \sum_{j=1}^N w_j^N h_j^N, w_j^N \geq 0,\ \sum_{j=1}^N w_j^N = 1,\ h_j^N \in \mathcal{H},\ N \geq 1 \Biggr\}.$$

Let φ be a function such that $\varphi(x) \geq I_{(-\infty,0]}(x)$ for all $x \in \mathbb{R}$ and φ satisfies the Lipschitz condition with constant $L(\varphi)$.

THEOREM 12. *Let $\mathcal{F} := \operatorname{conv}(\mathcal{H})$, where $\mathcal{H}$ is a class of measurable functions from $(S, \mathcal{A})$ into $\mathbb{R}$. For all $t > 0$,*

$$\mathbb{P}\Biggl\{ \exists f \in \mathcal{F} : P\{\tilde{f} \leq 0\} > \inf_{\delta \in (0,1]} \Biggl[P_n \varphi\Bigl(\frac{\tilde{f}}{\delta}\Bigr) + \frac{8L(\varphi)}{\delta} R_n(\mathcal{H}) \\ + \Biggl(\frac{\log \log_2(2\delta^{-1})}{n} \Biggr)^{1/2} \Biggr] + \frac{t}{\sqrt{n}} \Biggr\} \\ \leq 2\exp\{-2t^2\}.$$

PROOF. Since $\mathcal{F} := \operatorname{conv}(\mathcal{H})$, where $\mathcal{H}$ is a class of measurable functions from $(S, \mathcal{A})$ into $\mathbb{R}$, we have

$$R_n(\mathcal{F}) = \mathbb{E} \Biggl\| n^{-1} \sum_{i=1}^n \varepsilon_i \delta_{X_i} \Biggr\|_{\mathcal{F}} \\ = \mathbb{E} \sup \Biggl\{ \Biggl| n^{-1} \sum_{i=1}^n \varepsilon_i f_N(X_i) \Biggr| : f_N = \sum_{j=1}^N w_j^N h_j^N, w_j^N \geq 0, \\ \sum_{j=1}^N w_j^N = 1,\ h_j^N \in \mathcal{H},\ N \geq 1 \Biggr\} \\ = \mathbb{E} \Biggl| n^{-1} \sum_{i=1}^n \varepsilon_i \delta_{X_i} \Biggr|_{\mathcal{H}} = R_n(\mathcal{H}).$$

It follows that $R_n(\widetilde{\mathcal{F}}) = R_n(\mathcal{H})$, and Theorem 2 implies the result. $\square$

In the voting methods of combining the classifiers (such as boosting, bagging [Breiman (1996)], etc.), a classifier produced at each iteration is a convex combination $f_{\mathcal{S}} \in \operatorname{conv}(\mathcal{H})$ of simple base classifiers from the class $\mathcal{H}$ [$f_{\mathcal{S}}$ depends on the training sample $\mathcal{S} := ((X_1, Y_1), \dots, (X_n, Y_n))$]. The bound of Theorem 12 implies that for a given $\alpha \in (0, 1)$ with probability at least $1 - \alpha$,

$$P\{\tilde{f}_{\mathcal{S}} \le 0\} \le \inf_{\delta \in (0,1]} \left[P_n\{\tilde{f}_{\mathcal{S}} \le \delta\} + \frac{8}{\delta} R_n(\mathcal{H}) + \left(\frac{\log \log_2(2\delta^{-1})}{n} \right)^{1/2} \right] + \frac{t_\alpha}{\sqrt{n}},$$

where $t_\alpha := \sqrt{\frac{1}{2} \log \frac{2}{\alpha}}$. In particular, if $\mathcal{H}$ is a VC-class of classifiers $h : S \mapsto \{-1, 1\}$ (which means that the class of sets $\{\{x : h(x) = +1\} : h \in \mathcal{H}\}$ is a Vapnik–Chervonenkis class) with VC-dimension $V(\mathcal{H})$, we have, with some constant $C > 0$,

$$R_n(\mathcal{H}) \le C \sqrt{\frac{V(\mathcal{H})}{n}}.$$

This implies that with probability at least $1 - \alpha$

$$P\{\tilde{f}_{\mathcal{S}} \le 0\} \le \inf_{\delta \in (0,1]} \left[P_n\{\tilde{f}_{\mathcal{S}} \le \delta\} + \frac{C}{\delta} \sqrt{\frac{V(\mathcal{H})}{n}} + \left(\frac{\log \log_2(2\delta^{-1})}{n} \right)^{1/2} \right] + \sqrt{\frac{1}{2n} \log \frac{2}{\alpha}},$$

which slightly improves the main bound of the paper of Schapire, Freund, Bartlett and Lee (1998), which has a factor $\log(n/V(\mathcal{H}))$ in front of the term $C\delta^{-1}(V(\mathcal{H})/n)^{1/2}$.

EXAMPLE. In this example we consider a popular boosting algorithm called AdaBoost. At the beginning (at the first iteration) AdaBoost assigns uniform weights $w_j^{(1)} = n^{-1}$ to the labeled observations $(X_1, Y_1), \dots, (X_n, Y_n)$. At each iteration the algorithm updates the weights. Let $w^{(k)} = (w_1^{(k)}, \dots, w_n^{(k)})$ denote the vector of weights at kth iteration. Let $P_{n,w^{(k)}}$ be the weighted empirical measure on the kth iteration:

$$P_{n,w^{(k)}} := \sum_{i=1}^{n} w_i^{(k)} \delta_{(X_i, Y_i)}.$$

AdaBoost calls iteratively a base learning algorithm (called "weak learner") that returns at kth iteration a classifier $h_k \in \mathcal{H}$ and computes the weighted training error of h_k:

$$e_k := P_{n,w^{(k)}}\{y \ne h_k\}.$$

(In fact, the weak learner attempts to find a classifier with small enough weighted training error, at least such that $e_k \le 1/2$.) Then the weights are updated according to the rule

$$w_j^{(k+1)} := \frac{w_j^{(k)} \exp\{-Y_j \alpha_k h_k(X_j)\}}{Z_k},$$

where

$$Z_k := \sum_{j=1}^{n} w_j^{(k)} \exp\{-Y_j \alpha_k h_k(X_j)\}$$

and

$$\alpha_k := \frac{1}{2} \log \frac{1-e_k}{e_k}.$$

After N iterations AdaBoost outputs a classifier

$$f_{\mathcal{S}}(x) := \frac{\sum_{k=1}^{N} \alpha_k h_k(x)}{\sum_{k=1}^{N} \alpha_k}.$$

The above bounds, of course, apply to this classifier since $f_{\mathcal{S}} \in \operatorname{conv}(\mathcal{H})$. Another way to use Theorem 12 in the case of this example is to choose a decreasing function φ, satisfying all the conditions of Theorem 12 with $L(\varphi) = 1$ and such that $\varphi(u) \le e^{-u}$ for all $u \in \mathbb{R}$. It is easy to see that such a choice is possible. Let us also set

$$\delta := \frac{1}{\sum_1^N \alpha_k} \wedge 1.$$

Then it is not hard to check that

$$\varphi\left(\frac{y \sum_1^N \alpha_k h_k(x)}{\delta \sum_1^N \alpha_k}\right) \le \varphi\left(y \sum_1^N \alpha_k h_k(x)\right) \le \exp\left\{-y \sum_1^N \alpha_k h_k(x)\right\}.$$

Therefore

$$P_n \varphi\left(\frac{\tilde{f}_{\mathcal{S}}}{\delta}\right) \le P_n \exp\left\{-y \sum_1^N \alpha_k h_k(x)\right\}.$$

A simple [and well known in the literature on boosting; see, e.g., Schapire, Freund, Bartlett and Lee (1998)] computation shows that

$$P_n \exp\left\{-y \sum_1^N \alpha_k h_k(x)\right\} = \prod_{k=1}^{N} 2\sqrt{e_k(1-e_k)}.$$

We also have

$$\sum_{k=1}^{N} \alpha_k = \log \prod_{k=1}^{N} \sqrt{\frac{1-e_k}{e_k}}.$$

It follows now from the bound of Theorem 12 that with probability at least $1-\alpha$

$$P\{\tilde{f}_{\mathcal{S}} \le 0\} \le \prod_{k=1}^{N} 2\sqrt{e_k(1-e_k)} + 8\left(\log \prod_{k=1}^{N} \sqrt{\frac{1-e_k}{e_k}} \vee 1\right) R_n(\mathcal{H})$$
$$+ \left(\frac{\log \log_2 (2(\log \prod_{k=1}^{N} \sqrt{\frac{1-e_k}{e_k}} \vee 1))}{n}\right)^{1/2} + \sqrt{\frac{1}{2n} \log \frac{2}{\alpha}}.$$

The results of Section 3 provide some improvements of the above bounds on generalization error of convex combinations of base classifiers. To be specific, consider the case when $\mathcal{H}$ is a VC-class of classifiers. Let $V := V(\mathcal{H})$ be its VC-dimension. A well known bound on the entropy of the convex hull of a VC-class [see van der Vaart and Wellner (1996), page 142] implies that

$$H_{d_{P_n,2}}(\mathrm{conv}(\mathcal{H}); u) \le \sup_{Q \in \mathcal{P}(S)} H_{d_{Q,2}}\big(\mathrm{conv}(\mathcal{H}); u\big) \le D u^{-\frac{2(V-1)}{V}}.$$

[The bound on the entropy of a convex hull goes back to Dudley; the precise value of the exponent was given by Ball and Pajor, van der Vaart and Wellner, Carl; in the case of the convex hull of a VC-class, the above bound relies also on Haussler's improvement of Dudley's original bound on the entropy of a VC-class. See the discussion in the books of van der Vaart and Wellner (1996) and Dudley (1999) and references therein.] It immediately follows from Theorem 5 that for all $\gamma \ge \frac{2(V-1)}{2V-1}$ and for some constants C, B,

$$\mathbb{P}\left\{\exists f \in \mathrm{conv}(\mathcal{H}) : P\{\tilde{f} \le 0\} > \frac{C}{n^{1-\gamma/2} \hat{\delta}_n(\gamma; f)^{\gamma}}\right\}$$
$$\le B \log_2 \log_2 n \exp\left\{-\frac{1}{2} n^{\frac{\gamma}{2}}\right\},$$

where

$$\hat{\delta}_n(\gamma; f) := \sup\left\{\delta \in (0,1) : \delta^{\gamma} P_n\{(x,y) : yf(x) \le \delta\} \le n^{-1+\frac{\gamma}{2}}\right\}.$$

This shows that in the case when the VC-dimension of the base is relatively small the generalization error of boosting and some other convex combinations of simple classifiers obtained by various versions of voting methods becomes better than was suggested by the bounds of Schapire, Freund, Bartlett and Lee (1998). One can also conjecture, based on the bounds of Section 3, that outstanding generalization ability of these methods observed in numerous experiments is related not only

to the fact that they produce large margin classifiers, but also to the fact that the combined classifier belongs to a subset of the whole convex hull for which the random entropy $H_{d_{P_n,2}}$ is much smaller than for the whole convex hull.

Finally, it is worth mentioning that the bounds in terms of the so called margin cost functions [see, e.g., Mason, Bartlett and Baxter (1999), Mason, Baxter, Bartlett and Frean (2000)] easily follow from Theorem 1. Namely, Theorem 1 implies that with probability at least $1-\alpha$,

$$P\{\tilde{f}_{\mathcal{S}} \le 0\} \le \inf_{N\ge 1}\left[P_n\varphi_N(\tilde{f}_{\mathcal{S}}) + CL_N\sqrt{\frac{V(\mathcal{H})}{n}} + \left(\frac{\log N}{n}\right)^{1/2}\right] + \sqrt{\frac{1}{2n}\log\frac{2}{\alpha}},$$

where $\{\varphi_N\}$ is any sequence of Lipschitz cost functions such that $\varphi_N(x) \ge I_{(-\infty,0]}(x)$ for all $x \in \mathbb{R}$, $N \ge 1$ and L_N is a Lipschitz constant of φ_N.

6. Bounding the generalization error in neural network learning. We turn now to the applications of the bounds of the previous section in neural network learning. We start with the description of the class of feedforward neural networks for which the bounds on the generalization error will be proved. Let $\mathcal{H}$ be a class of measurable functions from $(S, \mathcal{A})$ into $\mathbb{R}$ (base functions). Consider an acyclic directed graph G. Suppose that G has a unique vertex v_i (input) that has no incoming edges and a unique vertex v_o (output) that has one outcoming edge. The vertices (nodes) of the graph will be called *neurons*. Suppose the set V of all the neurons is divided into layers

$$V = \{v_i\} \cup \bigcup_{j=0}^{l} V_j,$$

where $l \ge 0$ and $V_l = \{v_o\}$. The neurons v_i, v_o are called the input and the output neurons, respectively. The neurons of the layer V_0 will be called the base neurons. Suppose also that the inputs of the base neurons are the outputs of the input neuron. Suppose also that the inputs of the neurons of the layer V_j, $j \ge 1$ are the outputs of the neurons from the set $\bigcup_{k=0}^{j-1} V_k$. To define the network, we will assign the labels to the neurons in the following way. Each of the base neurons is labeled by a function from the base class $\mathcal{H}$. Each neuron of the jth layer V_j, where $j \ge 1$, is labeled by a vector $w := (w_1, \dots, w_n) \in \mathbb{R}^n$, where n is the number of inputs of the neuron. w will be called the vector of weights of the neuron.

Given a Borel function σ from $\mathbb{R}$ into $[-1, 1]$ (a sigmoid) and a vector $w := (w_1, \dots, w_n) \in \mathbb{R}^n$, let

$$N_{\sigma,w}\colon \mathbb{R}^n \mapsto \mathbb{R},\ N_{\sigma,w}(u_1, \dots, u_n) := \sigma\left(\sum_{i=1}^{n} w_j u_j\right).$$

For $w \in \mathbb{R}^n$,

$$\|w\|_{\ell_1} := \sum_{i=1}^{n} |w_i|.$$

Let $\sigma_j : j \geq 1$ be functions from $\mathbb{R}$ into $[-1, 1]$, satisfying the Lipschitz conditions

$$\left|\sigma_j(u) - \sigma_j(v)\right| \leq L_j|u - v|, \qquad u, v \in \mathbb{R}.$$

The network works the following way. The input neuron inputs an instance $x \in S$. A base neuron computes the value of the base function (it is labeled with) on this instance and outputs the value through its output edges. A neuron in the jth layer ($j \geq 1$) computes and outputs through its output edges the value $N_{\sigma_j,w}(u_1, \ldots, u_n)$ (where $u_1, \ldots, u_n$ are the values of the inputs of the neuron). The network outputs the value $f(x)$ (of a function f it computes) through the output edge.

We denote by $\mathcal{N}_l$ the set of all such networks. We call $\mathcal{N}_l$ the class of feedforward neural networks with base $\mathcal{H}$ and l layers of neurons (and with sigmoids $\{\sigma_j\}$). Let $\mathcal{N}_\infty := \bigcup_{j=0}^\infty \mathcal{N}_j$. Define $\mathcal{H}_0 := \mathcal{H}$, and then recursively

$$\mathcal{H}_j := \left\{N_{\sigma_j,w}(h_1, \ldots, h_n) : n \geq 0, h_i \in \mathcal{H}_{j-1},\ w \in \mathbb{R}^n\right\} \cup \mathcal{H}_{j-1}.$$

Denote $\mathcal{H}_\infty := \bigcup_{j=0}^\infty \mathcal{H}_j$. Clearly, $\mathcal{H}_\infty$ includes all the functions computable by feedforward neural networks with base $\mathcal{H}$.

Let $\{A_j\}$ be a sequence of positive numbers. We also define recursively classes of functions computable by feedforward neural networks with restrictions on the weights of neurons:

$$\begin{aligned}\mathcal{H}_j(A_1, \ldots, A_j)&\\ &:= \left\{N_{\sigma_j,w}(h_1, \ldots, h_n) : n \geq 0, h_i \in \mathcal{H}_{j-1}(A_1, \ldots, A_{j-1}),\right.\\ &\qquad\qquad \left. w \in \mathbb{R}^n, \|w\|_{\ell_1} \leq A_j\right\}\\ &\quad \cup \mathcal{H}_{j-1}(A_1, \ldots, A_{j-1}).\end{aligned}$$

Clearly,

$$\mathcal{H}_j := \bigcup\left\{\mathcal{H}_j(A_1, \ldots, A_j) : A_1, \ldots, A_j < +\infty\right\}.$$

As in the previous section, let φ be a function such that $\varphi(x) \geq I_{(-\infty,0]}(x)$ for all $x \in \mathbb{R}$ and φ satisfies the Lipschitz condition with constant $L(\varphi)$.

We start with the following result.

THEOREM 13. *For all $t > 0$ and for all $l \geq 1$,*

$$\begin{aligned}\mathbb{P}\Bigg\{&\exists f \in \mathcal{H}_l(A_1, \ldots, A_l) : P\{\tilde{f} \leq 0\}\\ &> \inf_{\delta\in(0,1]}\left[P_n\varphi\left(\frac{\tilde{f}}{\delta}\right) + \frac{2\sqrt{2\pi}L(\varphi)}{\delta}\prod_{j=1}^{l}(2L_jA_j + 1)G_n(\mathcal{H})\right] + \frac{t+2}{\sqrt{n}}\Bigg\}\\ &\leq 2\exp\{-2t^2\}.\end{aligned}$$

PROOF. We apply Theorem 2 to the class $\mathcal{F} = \mathcal{H}_l(A_1, \dots, A_l) =: \mathcal{H}'_l$, which gives for all $t > 0$,

$$\mathbb{P}\Biggl\{\exists f \in \mathcal{H}'_l : P\{\tilde{f} \le 0\} > \inf_{\delta\in[0,1]}\Biggl[P_n\varphi\Bigl(\frac{\tilde{f}}{\delta}\Bigr) + \frac{2\sqrt{2\pi}L(\varphi)}{\delta}G_n(\mathcal{H}'_l) + \Bigl(\frac{\log\log_2(2\delta^{-1})}{n}\Bigr)^{1/2}\Biggr] + \frac{t+2}{\sqrt{n}}\Biggr\} \le 2\exp\{-2t^2\}.$$

Thus, it's enough to show that

$$G_n(\mathcal{H}'_l) = \mathbb{E}\Bigl\|n^{-1}\sum_{i=1}^n g_i\delta_{X_i}\Bigr\|_{\mathcal{H}'_l} \le \prod_{j=1}^l (2L_jA_j+1)\mathbb{E}\Bigl\|n^{-1}\sum_{i=1}^n g_i\delta_{X_i}\Bigr\|_{\mathcal{H}}.$$

To this end, note that

$$\mathbb{E}\Bigl\|n^{-1}\sum_{i=1}^n g_i\delta_{X_i}\Bigr\|_{\mathcal{H}'_l} \le \mathbb{E}\Bigl\|n^{-1}\sum_{i=1}^n g_i\delta_{X_i}\Bigr\|_{\mathcal{G}_l} + \mathbb{E}\Bigl\|n^{-1}\sum_{i=1}^n g_i\delta_{X_i}\Bigr\|_{\mathcal{H}'_{l-1}}, \tag{6.1}$$

where

$$\mathcal{G}_l := \{N_{\sigma_l,w}(h_1,\dots,h_n) : n \ge 0, h_i \in \mathcal{H}_{l-1}(A_1,\dots,A_{l-1}), w \in \mathbb{R}^n, \|w\|_{\ell_1} \le A_l\}.$$

Consider two Gaussian processes,

$$Z_1(f) := n^{-1/2}\sum_{i=1}^n g_i(\sigma_l \circ f)(X_i)$$

and

$$Z_2(f) := L_l n^{-1/2}\sum_{i=1}^n g_i f(X_i),$$

where

$$f \in \Biggl\{\sum_{i=1}^n w_i h_i : n \ge 0, h_i \in \mathcal{H}'_{l-1},\ w \in \mathbb{R}^n,\ \|w\|_{\ell_1} \le A_l\Biggr\} =: \mathcal{G}'_l.$$

We have

$$\mathbb{E}_g|Z_1(f)-Z_1(h)|^2 = n^{-1}\sum_{i=1}^n |\sigma_l(f(X_i)-\sigma_l(h(X_i))|^2$$
$$\le L_l^2 n^{-1}\sum_{i=1}^n |f(X_i)-h(X_i)|^2 = \mathbb{E}_g|Z_2(f)-Z_2(h)|^2.$$

By Slepian's Lemma [see Ledoux and Talagrand (1991)], we get

$$\mathbb{E}_g\left\|n^{-1}\sum_{i=1}^n g_i\delta_{X_i}\right\|_{\mathcal{G}_l} = n^{-1/2}\mathbb{E}_g\|Z_1\|_{\mathcal{G}_l'} \le 2n^{-1/2}\mathbb{E}_g\|Z_2\|_{\mathcal{G}_l'}$$
$$= 2L_l\mathbb{E}_g\left\|n^{-1}\sum_{i=1}^n g_i\delta_{X_i}\right\|_{\mathcal{G}_l'}. \tag{6.2}$$

Since $\mathcal{G}_l' = A_l\,\mathrm{conv}_s(\mathcal{H}_{l-1})$ [here $\mathrm{conv}_s(\mathcal{G})$ denotes closed symmetric convex hull of a class $\mathcal{G}$, that is, closed convex hull of the class $\mathcal{H}\cup -\mathcal{H}$], it is easy to get that

$$\mathbb{E}\left\|n^{-1}\sum_{i=1}^n g_i\delta_{X_i}\right\|_{\mathcal{G}_l'} = A_l\mathbb{E}\left\|n^{-1}\sum_{i=1}^n g_i\delta_{X_i}\right\|_{\mathcal{H}_{l-1}}. \tag{6.3}$$

It follows from the bounds (6.1)–(6.3) that

$$\mathbb{E}\left\|n^{-1}\sum_{i=1}^n g_i\delta_{X_i}\right\|_{\mathcal{H}_l} \le (2L_lA_l+1)\mathbb{E}\left\|n^{-1}\sum_{i=1}^n g_i\delta_{X_i}\right\|_{\mathcal{H}_{l-1}}.$$

The result now follows by induction. □

REMARK. It can be shown that in the case of multilayer perceptrons (in which the neurons in each layer are linked only to the neurons in the previous layer) the factor $\prod_{j=1}^l(2L_jA_j+1)$ in the bound of the theorem can be replaced by $\prod_{j=1}^l(2L_jA_j)$. If the sigmoids are odd functions, the same factor in the case of general feedforward architecture of the network becomes $\prod_{j=1}^l(L_jA_j+1)$, and in the case of multilayer perceptrons $\prod_{j=1}^l L_jA_j$. Bartlett (1998) obtained a bound similar to the first inequality of Theorem 13 for a more special class $\mathcal{H}$ and with larger constants. In the case when $A_j\equiv A$, $L_j\equiv L$ (the case considered by Bartlett) the expression in the right-hand side of his bound includes $\frac{(AL)^{l(l+1)/2}}{\delta^l}$, which is replaced in our bound by $\frac{(AL)^l}{\delta}$. These improvement can be substantial in applications, since the above quantities play the role of complexity penalties.

Given a neural network $f \in \mathcal{N}_\infty$, let

$$\ell(f) := \min\{j \geq 1 : f \in \mathcal{N}_j\}.$$

Let $\{b_k\}$ be a sequence of nonnegative numbers. For a number k, $1 \leq k \leq \ell(f)$, let $V_k(f)$ denote the set of all neurons of layer k in the graph representing f. Denote

$$W_k(f) := \max_{N \in V_k(f)} \|w^{(N)}\|_{\ell_1} \vee b_k, \qquad k = 1, 2, \ldots, \ell(f),$$

and let

$$\Lambda(f) := \prod_{k=1}^{\ell(f)} (4L_k W_k(f) + 1),$$

$$\Gamma_\alpha(f) := \sum_{k=1}^{\ell(f)} \sqrt{\frac{\alpha}{2} \log(2 + |\log_2 W_k(f)|)},$$

where $\alpha > 0$ is a number such that $\zeta(\alpha) < 3/2$, ζ being the Riemann zeta-function:

$$\zeta(\alpha) := \sum_{k=1}^{\infty} k^{-\alpha}.$$

THEOREM 14. *For all $t > 0$ and for all $\alpha > 0$ such that $\zeta(\alpha) < 3/2$, the following bounds hold*:

$$\mathbb{P}\left\{\exists f \in \mathcal{H}_\infty : P\{\tilde{f} \leq 0\} > \inf_{\delta \in (0,1)} \left[P_n \varphi\left(\frac{\tilde{f}}{\delta}\right) + \frac{2\sqrt{2\pi} L(\varphi)}{\delta} \Lambda(f) G_n(\mathcal{H}) + \left(\frac{\log\log_2(2\delta^{-1})}{n}\right)^{1/2} \right] + \frac{\Gamma_\alpha(f) + t + 2}{\sqrt{n}} \right\}$$
$$\leq 2(3 - 2\zeta(\alpha))^{-1} \exp\{-2t^2\}.$$

PROOF. With a little abuse of notation, we write f for both the neural network and the function it computes. Denote

$$\Delta_k := \begin{cases} [2^{k-1}, 2^k), & \text{for } k \in \mathbb{Z},\ k \neq 0, 1, \\ [1/2, 2), & \text{for } k = 1. \end{cases}$$

The conditions $\ell(f) = l$ and

$$W_j(f) \in \Delta_{k_j}, \qquad k_j \in \mathbb{Z} \setminus \{0\},\ j = 1, \ldots, l,$$

easily imply that

$$\Lambda(f) \geq \prod_{j=1}^{l} (2L_j 2^{k_j} + 1), \qquad \Gamma_\alpha(f) \geq \sum_{j=1}^{l} \sqrt{\frac{\alpha}{2} \log(|k_j| + 1)}$$

and also that $f \in \mathcal{H}_l(2^{k_1}, \ldots, 2^{k_l})$. Therefore, the following bounds hold:

$$\begin{aligned}
&\mathbb{P}\Bigg\{\exists f \in \mathcal{H}_\infty : P\{\tilde{f} \leq 0\} \\
&\quad > \inf_{\delta \in (0,1)} \Bigg[P_n \varphi\left(\frac{\tilde{f}}{\delta}\right) + \frac{2\sqrt{2\pi} L(\varphi)}{\delta} \Lambda(f) G_n(\mathcal{H}) + \left(\frac{\log \log_2(2\delta^{-1})}{n}\right)^{1/2} \Bigg] \\
&\qquad\qquad + \frac{\Gamma_\alpha(f) + t + 2}{\sqrt{n}} \Bigg\} \\
&\quad \leq \sum_{l=0}^{\infty} \sum_{k_1 \in \mathbb{Z}\setminus\{0\}} \cdots \sum_{k_l \in \mathbb{Z}\setminus\{0\}} \\
&\qquad \times \mathbb{P}\Bigg\{\exists f \in \mathcal{H}_\infty \cap \{f : \ell(f) = l,\ W_j(f) \in \Delta_{k_j},\ j = 1, \ldots, l\} : P\{\tilde{f} \leq 0\} \\
&\qquad\quad > \inf_{\delta \in (0,1)} \Bigg[P_n \varphi\left(\frac{\tilde{f}}{\delta}\right) + \frac{2\sqrt{2\pi} L(\varphi)}{\delta} \Lambda(f) G_n(\mathcal{H}) \\
&\qquad\qquad\qquad + \left(\frac{\log \log_2(2\delta^{-1})}{n}\right)^{1/2} \Bigg] + \frac{\Gamma_\alpha(f) + t + 2}{\sqrt{n}} \Bigg\} \\
&\quad \leq \sum_{l=0}^{\infty} \sum_{k_1 \in \mathbb{Z}\setminus\{0\}} \cdots \sum_{k_l \in \mathbb{Z}\setminus\{0\}} \mathbb{P}\Bigg\{\exists f \in \mathcal{H}_l(2^{k_1}, \ldots, 2^{k_l}) : P\{\tilde{f} \leq 0\} \\
&\qquad\qquad > \inf_{\delta \in (0,1)} \Bigg[P_n \varphi\left(\frac{\tilde{f}}{\delta}\right) + \frac{2\sqrt{2\pi} L(\varphi)}{\delta} \\
&\qquad\qquad\qquad \times \prod_{j=1}^{l} (2L_j 2^{k_j} + 1) G_n(\mathcal{H}) \\
&\qquad\qquad\qquad + \left(\frac{\log \log_2(2\delta^{-1})}{n}\right)^{1/2} \Bigg] \\
&\qquad\qquad + \frac{\sum_{j=1}^{l} \sqrt{\frac{\alpha}{2} \log(|k_j| + 1)} + t + 2}{\sqrt{n}} \Bigg\}.
\end{aligned}$$

Using the bound of Theorem 13, we obtain

$$
\begin{aligned}
&\mathbb{P}\Biggl\{\exists f\in\mathcal{H}_\infty : P\{\tilde f\le 0\}\\
&\quad> \inf_{\delta\in(0,1)}\Biggl[P_n\varphi\Bigl(\frac{\tilde f}{\delta}\Bigr)+\frac{2\sqrt{2\pi}L(\varphi)}{\delta}\Lambda(f)G_n(\mathcal{H})\\
&\qquad\qquad+\Bigl(\frac{\log\log_2(2\delta^{-1})}{n}\Bigr)^{1/2}\Biggr]+\frac{\Gamma_\alpha(f)+t+2}{\sqrt{n}}\Biggr\}\\
&\quad\le\sum_{l=0}^{\infty}\sum_{k_1\in\mathbb{Z}\setminus\{0\}}\cdots\sum_{k_l\in\mathbb{Z}\setminus\{0\}}2\exp\Biggl\{-2\Biggl(\sum_{j=1}^{l}\sqrt{\frac{\alpha}{2}\log(|k_j|+1)}+t\Biggr)^2\Biggr\}\\
&\quad\le\sum_{l=0}^{\infty}\sum_{k_1\in\mathbb{Z}\setminus\{0\}}\cdots\sum_{k_l\in\mathbb{Z}\setminus\{0\}}2\exp\Biggl\{-\sum_{j=1}^{l}\alpha\log(|k_j|+1)-2t^2\Biggr\}\\
&\quad=2\sum_{l=0}^{\infty}\sum_{k_1\in\mathbb{Z}\setminus\{0\}}\cdots\sum_{k_l\in\mathbb{Z}\setminus\{0\}}\prod_{j=1}^{l}\bigl(|k_j|+1\bigr)^{-\alpha}\exp\{-2t^2\}\\
&\quad=2\sum_{l=0}^{\infty}\prod_{j=1}^{l}\Biggl(2\sum_{k=2}^{\infty}k^{-\alpha}\Biggr)\exp\{-2t^2\}=2\sum_{l=0}^{\infty}\bigl[2(\zeta(\alpha)-1)\bigr]^l\exp\{-2t^2\}\\
&\quad=2\bigl(3-2\zeta(\alpha)\bigr)^{-1}\exp\{-2t^2\},
\end{aligned}
$$

which yields the bound of the theorem. □

It follows, in particular, that for any classifier $f_{\mathcal{S}}\in\mathcal{H}_\infty$, based on the training data $\mathcal{S}:=((X_1,Y_1),\dots,(X_n,Y_n))$, we have

$$
\begin{aligned}
&\mathbb{P}\Biggl\{P\{\tilde f_{\mathcal{S}}\le 0\}> \inf_{\delta\in(0,1)}\Biggl[P_n\varphi\Bigl(\frac{\tilde f_{\mathcal{S}}}{\delta}\Bigr)+\frac{2\sqrt{2\pi}L(\varphi)}{\delta}\Lambda(f_{\mathcal{S}})G_n(\mathcal{H})\\
&\qquad\qquad+\Bigl(\frac{\log\log_2(2\delta^{-1})}{n}\Bigr)^{1/2}\Biggr]+\frac{\Gamma_\alpha(f_{\mathcal{S}})+t+2}{\sqrt{n}}\Biggr\}\\
&\quad\le 2\bigl(3-2\zeta(\alpha)\bigr)^{-1}\exp\{-2t^2\}.
\end{aligned}
$$

Next we consider a method of complexity penalization in neural network learning based on the penalties that depend on ℓ_1-norms of the vectors of weights of the neurons. Suppose that $f_{\mathcal{S}}$ is the neural network from $\mathcal{F}\subset\mathcal{H}_\infty$ that

minimizes the penalized training error

$$f_{\hat{s}} := \operatorname*{argmin}_{f\in\mathcal{F}} \inf_{\delta\in(0,1]}\left[P_n(\{\tilde{f}\le\delta\}) + \frac{2\sqrt{2\pi}}{\delta}\Lambda(f)G_n(\mathcal{H}) + \left(\frac{\log\log_2(2\delta^{-1})}{n}\right)^{1/2}\right] + \frac{\Gamma_\alpha(f)}{\sqrt{n}}$$
$$= \operatorname*{argmin}_{f\in\mathcal{F}}\left[P_n(\{\tilde{f}\le 0\}) + \inf_{\delta\in(0,1]}\hat{\pi}_n(f;\delta)\right],$$

where the quantity $\inf_{\delta\in(0,1]}\hat{\pi}_n(f;\delta)$ plays the role of the complexity penalty,

$$\hat{\pi}_n(f;\delta) := P_n\big(\{0<\tilde{f}\le\delta\}\big) + \Psi_n(f;\delta),$$
$$\Psi_n(f;\delta) := \frac{2\sqrt{2\pi}}{\delta}\Lambda(f)G_n(\mathcal{H}) + \left(\frac{\log\log_2(2\delta^{-1})}{n}\right)^{1/2} + \frac{\Gamma_\alpha(f)}{\sqrt{n}}.$$

We define a distribution dependent version of this data dependent penalty as $\inf_{\delta\in(0,1]}\pi_n(f;\delta)$, where

$$\pi_n(f;\delta) := P\big(\{0<\tilde{f}\le 2\delta\}\big) + 2\Psi_n(f;\delta).$$

The first inequality of the next theorem provides an upper confidence bound on the generalization error of the classifier $f_{\hat{s}}$. The second bound is an "oracle inequality" that shows that the estimate $f_{\hat{s}}$ obtained by the above method possesses some optimality property [see Johnstone (1998) and Barron, Birgé and Massart (1999) for a general approach to penalization and oracle inequalities in nonparametric statistics].

THEOREM 15. *For all $t>0$ and for all $\alpha>0$ with $\zeta(\alpha)<3/2$, the following bounds hold*:

$$\mathbb{P}\left\{P\{\tilde{f}_{\hat{s}}\le 0\} > \inf_{f\in\mathcal{F}}\left[P_n\{\tilde{f}\le 0\} + \inf_{\delta\in(0,1]}\hat{\pi}_n(f;\delta)\right] + \frac{t+2}{\sqrt{n}}\right\}$$
$$\le 2\big(3-2\zeta(\alpha)\big)^{-1}\exp\{-2t^2\}$$

and

$$\mathbb{P}\left\{P\{\tilde{f}_{\hat{s}}\le 0\} - \inf_{g\in\mathcal{F}}P\{\tilde{g}\le 0\}\right.$$
$$\left.> \inf_{f\in\mathcal{F}}\left[P\{\tilde{f}\le 0\} - \inf_{g\in\mathcal{F}}P\{\tilde{g}\le 0\} + \inf_{\delta\in(0,1]}\pi_n(f;\delta)\right] + \frac{2t+4}{\sqrt{n}}\right\}$$
$$\le 4\big(3-2\zeta(\alpha)\big)^{-1}\exp\{-2t^2\}.$$

PROOF. The first bound follows from Theorem 14 and the definition of the estimate $\tilde{f}_{\mathcal{S}}$. To prove the second bound, we repeat the proof of Theorems 1, 2 to show that for any class $\mathcal{F}'$,

$$\begin{aligned}\mathbb{P}\Biggl\{&\exists f\in\mathcal{F}'\ \exists\delta\in(0,1]: P_n\bigl\{\tilde{f}\le\delta\bigr\}\\ &>\biggl[P\varphi\Bigl(\frac{\tilde{f}-\delta}{\delta}\Bigr)+\frac{2\sqrt{2\pi}}{\delta}G_n(\mathcal{F}')+\Bigl(\frac{\log\log_2(2\delta^{-1})}{n}\Bigr)^{1/2}\biggr]+\frac{t+2}{\sqrt{n}}\Biggr\}\\ &\le 2\exp\bigl\{-2t^2\bigr\}.\end{aligned}$$

The argument that led to Theorems 13 and 14 shows that

$$\begin{aligned}\mathbb{P}\Biggl\{&\exists f\in\mathcal{F}\ \exists\delta\in(0,1]: P_n\bigl\{\tilde{f}\le\delta\bigr\}\\ &>\biggl[P\bigl\{\tilde{f}\le 2\delta\bigr\}+\frac{2\sqrt{2\pi}}{\delta}\Lambda(f)G_n(\mathcal{H})\\ &\quad+\Bigl(\frac{\log\log_2(2\delta^{-1})}{n}\Bigr)^{1/2}+\frac{\Gamma_\alpha(f)}{\sqrt{n}}+\frac{t+2}{\sqrt{n}}\biggr]\Biggr\}\\ &\le 2\bigl(3-2\zeta(\alpha)\bigr)^{-1}\exp\bigl\{-2t^2\bigr\}.\end{aligned}$$

If now

$$\begin{aligned}&\inf_{f\in\mathcal{F}}\inf_{\delta\in(0,1]}\bigl[P_n\bigl(\bigl\{\tilde{f}\le\delta\bigr\}\bigr)+\Psi_n(f;\delta)\bigr]+\frac{t+2}{\sqrt{n}}\\ &\quad>\inf_{f\in\mathcal{F}}\inf_{\delta\in(0,1]}\bigl[P\bigl\{\tilde{f}\le 2\delta\bigr\}+2\Psi_n(f;\delta)\bigr]+\frac{2t+4}{\sqrt{n}},\end{aligned}$$

then

$$\exists\, f\in\mathcal{F}\quad\exists\,\delta\in(0,1]: P_n\bigl\{\tilde{f}\le\delta\bigr\}>\bigl[P\bigl\{\tilde{f}\le 2\delta\bigr\}+\Psi_n(f;\delta)\bigr]+\frac{t+2}{\sqrt{n}}.$$

Combining this with the first bound gives

$$\begin{aligned}\mathbb{P}\Biggl\{&P\bigl\{\tilde{f}_{\mathcal{S}}\le 0\bigr\}>\inf_{f\in\mathcal{F}}\inf_{\delta\in(0,1)}\bigl[P\bigl\{\tilde{f}\le 2\delta\bigr\}+2\Psi_n(f;\delta)\bigr]+\frac{2t+4}{\sqrt{n}}\Biggr\}\\ &\le 4\bigl(3-2\zeta(\alpha)\bigr)^{-1}\exp\bigl\{-2t^2\bigr\},\end{aligned}$$

which implies the result. □

Acknowledgments. The authors are very thankful to Jon Wellner for reading the manuscript and making a number of comments and suggestions that improved the paper. We want to thank Evarist Giné for pointing out the current formulation of Theorem 8. We also want to thank Fernando Lozano for providing the results of the experiment presented in Figure 1.

REFERENCES

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

BARRON, A. (1991a). Complexity regularization with applications to artificial neural networks. In *Nonparametric Functional Estimation and Related Topics* (G. Roussas, ed.) 561–576. Kluwer, Dordrecht.

BARRON, A. (1991b). Approximation and estimation bounds for artificial neural networks. In *Proceedings of the Fourth Annual Workshop on Computational Learning Theory* 243–249. Morgan Kaufmann, San Francisco.

BARRON, A., BIRGÉ, L. and MASSART, P. (1999). Risk bounds for model selection via penalization. *Probab. Theory Related Fields* **113** 301–413.

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.

BARTLETT, P. and SHAWE-TAYLOR, J. (1999). Generalization performance of support vector machines and other pattern classifiers. In *Advances in Kernel Methods*: *Support Vector Learning* (B. Schölkopf, C. Burges and A. J. Smola, eds.) 43–54. MIT Press.

BIRGÉ, L. and MASSART, P. (1997). From model selection to adaptive estimation. In *Festschrift for Lucien Le Cam*: *Research Papers in Probability and Statistics* (D. Pollard, E. Torgersen and G. Yang, eds.) 55–87. Springer, New York.

BLAKE, C. and MERZ, C. (1998). UCI repository of machine learning databases. Available at www.ics.uci.edu/∼mlearn/MLRepository.html.

BREIMAN, L. (1996). Bagging predictors. *Machine Learning* **24** 123–140.

CORTES, C. and VAPNIK, V. (1995). Support vector networks. *Machine Learning* **20** 273–297.

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

DUDLEY, R. M. (1999). *Uniform Central Limit Theorems*. Cambridge Univ. Press.

FELLER, W. (1950). *An Introduction to Probability Theory and Its Applications* **1**. Wiley, New York.

FINE, T. L. (1999). *Feedforward Neural Network Methodology*. Springer, New York.

FREUND, Y. (1995). Boosting a weak learning algorithm by majority. *Inform. and Comput.* **121** 256–285.

FREUND, Y. (1999). An adaptive version of the boost by majority algorithm. Preprint.

FRIEDMAN, J. (1999). Greedy function approximation: a gradient boosting machine. Technical report, Dept. Statistics, Stanford Univ.

FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2000). Additive logistic regression: a statistical view of boosting. *Ann. Statist.* **28** 337–374.

JOHNSTONE, I. M. (1998). Oracle inequalities and nonparametric function estimation. In *Doc. Math.* (*Extra Volume*) **3** 267–278.

KOLTCHINSKII, V. (2001). Rademacher penalties and structural risk minimization. *IEEE Trans. Inform. Theory*. To appear.

KOLTCHINSKII, V. and PANCHENKO, D. (2000). Rademacher processes and bounding the risk of function learning. In *High Dimensional Probability II* (E. Giné, D. Mason and J. Wellner, eds.) 444–459. Birkhäuser, Boston.

KOLTCHINSKII, V., PANCHENKO, D. and LOZANO, F. (2000a). Bounding the generalization error of convex combinations of classifiers: balancing the dimensionality and the margins. Preprint. Available at www.boosting.org/.

KOLTCHINSKII, V., PANCHENKO, D. and LOZANO, F. (2000b). Some new bounds on the generalization error of combined classifiers. *Advances in Neural Information Processing Systems*. Available at www.boosting.org/.

LEDOUX, M. and TALAGRAND, M. (1991). *Probability in Banach Spaces.* Springer, New York.

MASON, L., BARTLETT, P. and BAXTER, J. (1999). Improved generalization through explicit optimization of margins. *Machine Learning* **38** 243–255.

MASON, L., BAXTER, J., BARTLETT, P. and FREAN, M. (2000). Functional gradient techniques for combining hypotheses. In *Advances in Large Margin Classifiers* (A. J. Smola, P. Bartlett, B. Schölkopf and D. Schurmans, eds.) MIT Press.

MASSART, P. (2000). About the constants in Talagrand's concentration inequalities for empirical processes. *Ann. Probab.* **28** 863–885.

SAZONOV, V. V. (1963). On the Glivenko-Cantelli theorem. *Theory Probab. Appl.* **8** 282–285.

SCHAPIRE, R., FREUND, Y., BARTLETT, P. and LEE, W. S. (1998). Boosting the margin: A new explanation for the effectiveness of voting methods. *Ann. Statist.* **26** 1651–1686.

TALAGRAND, M. (1996a). A new look at independence. *Ann. Probab.* **24** 1–34.

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

TALAGRAND, M. (1998). Rigorous results for the Hopfield model with many patterns. *Probab. Theory Related Fields* **110** 177–276.

TOPSØE, F., DUDLEY, R. and HOFFMANN-JØRGENSEN, J. (1976). Two examples concerning uniform convergence of measures w.r.t. balls in Banach spaces. *Empirical Distributions and Processes. Lecture Notes in Math.* **566** 141–146. Springer, Berlin.

VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak convergence and Empirical Processes. With Applications to Statistics.* Springer, New York.

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

VIDYASAGAR, M. (1997). *A Theory of Learning and Generalization.* Springer, New York.

YUKICH, J., STINCHCOMBE, H. and WHITE, H. (1995). Sup-norm approximation bounds for networks through probabilistic methods. *IEEE Trans. Inform. Theory* **41** 1021–1027.

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