EMPIRICAL MARGIN DISTRIBUTIONS AND BOUNDING THE GENERALIZATION ERROR OF COMBINED CLASSIFIERS

V. Koltchinskii* and D. Panchenko†
Department of Mathematics and Statistics
The University of New Mexico

March 23, 2000

Abstract

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 Levy distance of empirical margin distribution to the true margin distribution uniformly over the classes of classifiers and prove the optimality of these rates.

*Partially supported by NSA Grant MDA904-99-1-0031
†Partially supported by Boeing Computer Services Grant 3-48181

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 $g \in \mathcal{G}$, $\mathrm{sign}(g(X))$ 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 $(X_1,Y_1),\dots,(X_n,Y_n)$ 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) and Devroye, Györfi and Lugosi (1996)) and it is based on bounding the difference between the generalization error $\mathbb{P}\{Yg(X) \le 0\}$ and the training error

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

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) 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)) 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 the 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 upper bounds have the following form (up to some extra terms)

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

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 $\delta \to 0$ (for instance, there could be $\phi(\delta) = \frac{1}{\delta}$), $C(\mathcal{H})$ is a constant depending on the class $\mathcal{H}$ (in particular, on the VC-dimension, or some type of entropy characteristics of the class).

We develop a new approach that allows us to improve some of the previously known bounds. In the case of the Bartlett's bounds for neural networks in terms of the ℓ_1-norms of the weights of the neurons, the improvement 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 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 $\phi(\delta) = \frac{1}{\delta}$. The methods of the proofs are also different. Bartlett uses the so called fat-shattering dimensions of function classes and the extension of Vapnik–Chervonenkis type inequalities to such dimensions. Our method is based on the general results of the theory of Gaussian and empirical processes (such as comparison inequalities, e.g. Slepian's Lemma, symmetrization and random multipliers inequalities, concentration inequalities). 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 Levy 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 Levy distance is $O(n^{-1/4})$. We gave some examples, showing the optimality of these rates.

2 Probabilistic bounds for general function classes

Let $(S, \mathcal{A}, P)$ be a probability space and let $\mathcal{F}$ be a class of measurable functions from $(S, \mathcal{A})$ into $\mathbb{R}$. 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 $(X_1, \dots, X_n)$,

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

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

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

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 section to upper bound the generalization error of combined classifiers. The bounds will depend on a measure of "complexity" of the class $\mathcal{F}$ which will be introduced next. Define

$$G_n(\mathcal{F}) := 2\sqrt{\pi}\mathbb{E}\|n^{-1} \sum_{i=1}^{n} g_i \delta_{X_i}\|_{\mathcal{F}},$$

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 $\mathcal{F}$. One can find in the literature (see, e.g., van

der Vaart and Wellner (1996)) various upper bounds on such quantities as $G_n(\mathcal{F})$ in terms of entropies, VC-dimensions, etc.

Let

$$\Delta_n(\mathcal{F};t) := G_n(\mathcal{F}) + \frac{t+4\sqrt{2}}{\sqrt{n}}.$$

Let φ be a function from $\mathbb{R}$ into $[0,1]$ such that $\varphi(u) = 1$ for $u \le 0$, $\varphi(u) = 0$ for $u \ge 1$ and φ satisfies the Lipschitz condition: $|\varphi(t) - \varphi(s)| \le |t-s|$.

First we will prove the following fact.

Theorem 1 *For all $t > 0$,*

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

Proof. For each $\delta \in (0,1]$ and for all $f \in \mathcal{F}$ we have

$$P\{f \le 0\} \le P\varphi(\frac{f}{\delta}) \le P_n\varphi(\frac{f}{\delta}) + \frac{1}{\delta}\|P_n - P\|_{\mathcal{G}}, \tag{1}$$

where

$$\mathcal{G} := \Big\{t\varphi \circ \frac{f}{t} : f \in \mathcal{F}, t \in (0,1]\Big\}.$$

By the exponential inequalities for martingale difference sequences (see Devroye, Györfi and Lugosi (1996)), we have

$$\mathbb{P}\Big\{\|P_n - P\|_{\mathcal{G}} \ge \mathbb{E}\|P_n - P\|_{\mathcal{G}} + \varepsilon\Big\} \le \exp\{-2\varepsilon^2 n\}.$$

Thus, with probablity at least $1 - \exp\{-2\varepsilon^2 n\}$

$$P\{f \le 0\} \le P_n\varphi(\frac{f}{\delta}) + \frac{1}{\delta}\mathbb{E}\|P_n - P\|_{\mathcal{G}} + \frac{\varepsilon}{\delta}. \tag{2}$$

Next using the Symmetrization Inequality and Gaussian Multiplier Inequality (see van der Vaart and Wellner (1996)), we get

$$\mathbb{E}\|P_n - P\|_{\mathcal{G}} \le 2\mathbb{E}\|n^{-1}\sum_{i=1}^n \varepsilon_i\delta_{X_i}\|_{\mathcal{G}} \le \sqrt{2\pi}\mathbb{E}\|n^{-1}\sum_{i=1}^n g_i\delta_{X_i}\|_{\mathcal{G}}. \tag{3}$$

We also have for all $t, s \in \mathbb{R}\setminus\{0\}$ the following bound

$$|t\varphi(\frac{u}{t}) - s\varphi(\frac{v}{s})| \le |u-v| + 2|t-s|. \tag{4}$$

Indeed, for all $t, s > 0$

$$|t\varphi(\frac{u}{t}) - s\varphi(\frac{v}{s})| \le |t\varphi(\frac{u}{t}) - t\varphi(\frac{v}{t})| + |t\varphi(\frac{v}{t}) - s\varphi(\frac{v}{s})| \le$$

$$\le t|\frac{u}{t} - \frac{v}{t}| + |t\varphi(\frac{v}{t}) - s\varphi(\frac{v}{s})| \le$$

$$\leq |u-v| + |t\varphi(\frac{v}{t}) - s\varphi(\frac{v}{t})| + |s\varphi(\frac{v}{t}) - s\varphi(\frac{v}{s})| \leq$$
$$\leq |u-v| + |t-s| + s|\varphi(\frac{v}{t}) - \varphi(\frac{v}{s})|.$$

Assume that $t \geq s$. If $v \geq t$, or $v \leq 0$, we have

$$|t\varphi(\frac{u}{t}) - s\varphi(\frac{v}{s})| \leq |u-v| + |t-s|.$$

Otherwise, we get

$$s|\varphi(\frac{v}{t}) - \varphi(\frac{v}{s})| \leq \frac{sv}{ts}|t-s| \leq |t-s|,$$

and the bound (4) follows. Similarly, it can be proved for all $t, s < 0$. In the case $ts < 0$, we get

$$|t\varphi(\frac{u}{t}) - s\varphi(\frac{v}{s})| \leq |t| + |s| = |t-s|,$$

so (4) also holds.

Let $d_{P_n,2}$ denote the metric of the space $L_2(S; dP_n)$:

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

Now we use (4) to get for all $t, s > 0$

$$d^2_{P_n,2}(t\varphi \circ \frac{f}{t}; s\varphi \circ \frac{h}{s}) = n^{-1}\sum_{i=1}^n |t\varphi(\frac{f(X_i)}{t}) - s\varphi(\frac{h(X_i)}{s})|^2 \leq$$

$$\leq n^{-1}\sum_{i=1}^n [|f(X_i) - h(X_i)| + 2|t-s|]^2 \leq n^{-1}\sum_{i=1}^n [2|f(X_i) - h(X_i)|^2 + 8|t-s|^2]^2 \leq$$
$$\leq 2d^2_{P_n,2}(f,h) + 8|t-s|^2. \tag{5}$$

Define Gaussian processes

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

and

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

where g is a standard normal random variable on $(\Omega_g, \Sigma_g, \mathbb{P}_g)$ independent of $\{g_i\}$. Then, for all $t, s > 0$, we can rewrite (5) as

$$\mathbb{E}_g|Z_1(f,t) - Z_1(h,s)|^2 \leq \mathbb{E}_g|Z_2(f,t) - Z_2(h,s)|^2$$

and we also have similarly

$$\mathbb{E}_g|Z_1(-f,-t) - Z_1(-h,-s)|^2 \leq \mathbb{E}_g|Z_2(-f,-t) - Z_2(-h,-s)|^2.$$

Here $\mathbb{E}_g$ denotes the expectation on the probability space $(\Omega_g, \Sigma_g, \mathbb{P}_g)$ (on which the sequence $\{g_i\}$ and the random variable g were defined). A version of Slepian's Lemma (see Ledoux and Talagrand (1991)) implies that

$$\mathbb{E}_g \sup\{Z_1(f,t) : f \in \mathcal{F}, t \in (0,1] \text{ or } -f \in \mathcal{F}, -t \in (0,1]\} \leq$$

$$\leq \mathbb{E}_g \sup\{Z_2(f,t) : f \in \mathcal{F}, t \in (0,1] \text{ or } -f \in \mathcal{F}, -t \in (0,1]\}.$$

We have

$$\mathbb{E}_g \| n^{-1} \sum_{i=1}^n g_i \delta_{X_i} \|_{\mathcal{G}} =$$

$$= \mathbb{E}_g \sup_{h \in \mathcal{G}} [n^{-1} \sum_{i=1}^n g_i h(X_i)] \bigvee \sup_{h \in \mathcal{G}} [n^{-1} \sum_{i=1}^n g_i (-h)(X_i)] = \mathbb{E}_g \sup_{h \in \bar{\mathcal{G}}} [n^{-1} \sum_{i=1}^n g_i h(X_i)] =$$

$$= \mathbb{E}_g \sup\{Z_1(f,t) : f \in \mathcal{F}, t \in (0,1] \text{ or } -f \in \mathcal{F}, -t \in (0,1]\},$$

where $\bar{\mathcal{G}} := \{t\varphi(\frac{f}{t}), -t\varphi(\frac{-f}{-t}) : f \in \mathcal{F}, t \in (0,1]\}$, and similarly it can be proved that

$$\sqrt{2}\mathbb{E}_g \| n^{-1} \sum_{i=1}^n g_i \delta_{X_i} \|_{\mathcal{F}} = \mathbb{E}_g \sup\{Z_2(f,t) : f \in \mathcal{F}, t \in (0,1] \text{ or } -f \in \mathcal{F}, -t \in (0,1]\}.$$

This immediately gives us

$$\mathbb{E}_g \| n^{-1} \sum_{i=1}^n g_i \delta_{X_i} \|_{\mathcal{G}} \leq \sqrt{2}\mathbb{E}_g \| n^{-1} \sum_{i=1}^n g_i \delta_{X_i} \|_{\mathcal{F}} + \frac{2\sqrt{2}\mathbb{E}|g|}{\sqrt{n}}. \tag{6}$$

Combining the bounds (2), (3), (6), we prove that with probablity at least $1 - \exp\{-2\varepsilon^2 n\}$

$$P\{f \leq 0\} \leq P_n \varphi(\frac{f}{\delta}) + 2\sqrt{\pi}\frac{1}{\delta}\mathbb{E}\| n^{-1} \sum_{i=1}^n g_i \delta_{X_i} \|_{\mathcal{F}} + \frac{4\sqrt{2}}{\delta} n^{-1/2} + \frac{\varepsilon}{\delta}.$$

Setting $\varepsilon := tn^{-1/2}$ completes the proof.

□

Quite similarly, assuming now that φ is a function from $\mathbb{R}$ into $[0,1]$ such that $\varphi(u) = 1$ for $u \leq -1$, $\varphi(u) = 0$ for $u \geq 0$ and φ still satisfies the Lipschitz condition $|\varphi(t) - \varphi(s)| \leq |t - s|$, one can prove the following statement.

Theorem 2 *For all $t > 0$,*

$$\mathbb{P}\Big\{\exists f \in \mathcal{F} : P\{f \leq 0\} < \sup_{\delta \in (0,1)} \Big[P_n \varphi(\frac{f}{\delta}) - \frac{1}{\delta}\Delta_n(\mathcal{F}; t)\Big]\Big\} \leq \exp\{-2t^2\}.$$

The bounds of theorems 1 and 2 easily imply that for all $t > 0$

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

and

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

Similarly, it can be shown that

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

and

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

Combining the last bounds, we get the following result:

Theorem 3 *For all $t > 0$,*

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

and

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

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 3 $\delta := H_f^{-1}(\Delta_n(\mathcal{F};t)) \bigwedge 1$ easily gives us the following upper bound that holds for any $t > 0$ with probability at least $1 - 2e^{-2t^2}$:

$$\forall f \in \mathcal{F}\ |P_n\{f \le 0\} - P\{f \le 0\}| \le \frac{\Delta_n(\mathcal{F};t)}{H_f^{-1}(\Delta_n(\mathcal{F};t))} \bigvee \Delta_n(\mathcal{F};t).$$

Similarly, the first bound of Theorem 3 gives that for any $t > 0$ with probability at least $1 - 2e^{-2t^2}$:

$$\forall f \in \mathcal{F}\ |P_n\{f \le 0\} - P\{f \le 0\}| \le \frac{\Delta_n(\mathcal{F};t)}{H_{n,f}^{-1}(\Delta_n(\mathcal{F};t))} \bigvee \Delta_n(\mathcal{F};t).$$

The next example shows that, in general, the term $\frac{1}{\delta}\Delta_n(\mathcal{F};t)$ of the bound of Theorem 1 (and other similar results, in particular, Theorem 3) can not be improved.

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

$$X_n = \left(\varepsilon_k^n (2\log(k+1))^{-\frac{1}{2}}\right)_{k \ge 1},$$

where ε_k^n are i.i.d. Rademacher random variables. 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

$$\mathbb{P}\left(\exists f \in \mathcal{F} : P\{f \le 0\} \ge \inf_{\delta \in (0,1]} [P_n\{f \le \delta\} + \frac{1}{\phi(\delta)} \Delta_n(\mathcal{F}; t)]\right) \to 1$$

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's well known that $\mathcal{F}$ is a bounded CLT class for the distribution P of the sequence $\{X_n\}$ (see Ledoux and Talagrand (1991)). Notice that $P(f_k \le 0) = 1/2$ for all k and $\mathbb{E}\|n^{-1} \sum g_i \delta_{X_i}\|_{\mathcal{F}} \le c n^{-1/2}$ for some constant $c > 0$. Let us denote by $t' = t + 4\sqrt{2} + 2\sqrt{\pi} c$. The infimum inside the probability is less then 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,

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

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)

$$(2\pi)^{\frac{1}{2}} n^{n+\frac{1}{2}} e^{-n+1/(12n+1)} < n! < (2\pi)^{\frac{1}{2}} n^{n+\frac{1}{2}} e^{-n+1/12n} \tag{7}$$

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

$$\binom{n}{k_0} 2^{-n} \ge c n^{-\frac{1}{2}} \left((1-2\delta)^{1-2\delta}(1+2\delta)^{1+2\delta}\right)^{-\frac{n}{2}} \ge c n^{-\frac{1}{2}} \exp(-4n\delta^2). \tag{8}$$

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 (??) that

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

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

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

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

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

then for n large enough

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

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

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

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 1 does not hold with $\frac{1}{\delta}\Delta_n(\mathcal{F};t)$ replaced by $\frac{1}{\phi(\delta)}\Delta_n(\mathcal{F};t)$. Similarly, one can show that

$$\mathbb{P}\left(\exists f \in \mathcal{F} : |P_n\{f \le 0\} - P\{f \le 0\}| \ge \inf_{\delta\in(0,1]}[P_n\{|f| \le \delta\} + \frac{1}{\phi(\delta)}\Delta_n(\mathcal{F};t)]\right) \to 1$$

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

3 Convergence rates of empirical margin distributions

We are again in the setting of Section 2 with the class $\mathcal{F}$ of measurable functions from S into $[-1,1]$. For $f \in \mathcal{F}$, let

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

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

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

Theorem 4 *For all $t > 0$,*

$$\mathbb{P}\Big\{\sup_{f\in\mathcal{F}} L(F_{n,f}, F_f) \geq (\sqrt{6\pi}\mathbb{E}\|n^{-1}\sum_{i=1}^{n} g_i\delta_{X_i}\|_{\mathcal{F}} + \frac{t+6\sqrt{3}}{\sqrt{n}})^{1/2}\Big\} \leq \exp\{-2t^2\}.$$

Proof. Similarly to (1), we get the following bounds:

$$F_f(y) = P\{f \leq y\} \leq P\varphi(\frac{f-y}{\delta}) \leq P_n\varphi(\frac{f-y}{\delta}) + \frac{1}{\delta}\|P_n - P\|_{\tilde{\mathcal{G}}} \leq F_{n,f}(y+\delta) +$$

$$+\frac{1}{\delta}\|P_n - P\|_{\tilde{\mathcal{G}}}$$

and

$$F_{n,f}(y) = P_n\{f \leq y\} \leq P_n\varphi(\frac{f-y}{\delta}) \leq P\varphi(\frac{f-y}{\delta}) + \frac{1}{\delta}\|P_n - P\|_{\tilde{\mathcal{G}}} \leq F_f(y+\delta) +$$

$$+\frac{1}{\delta}\|P_n - P\|_{\tilde{\mathcal{G}}},$$

where

$$\tilde{\mathcal{G}} := \Big\{t\varphi\circ(\frac{f-y}{t}) : f\in\mathcal{F}, t\in(0,1), y\in[-1,1]\Big\}.$$

It follows that

$$\sup_{f\in\mathcal{F}} L(F_{n,f}, F_f) \leq \delta \bigvee \frac{1}{\delta}\|P_n - P\|_{\tilde{\mathcal{G}}}.$$

Minimizing the righthand side over δ, we get

$$\sup_{f\in\mathcal{F}} L(F_{n,f}, F_f) \leq (\|P_n - P\|_{\tilde{\mathcal{G}}})^{1/2}. \tag{9}$$

Similarly to (4),

$$|t\varphi(\frac{u-y}{t}) - s\varphi(\frac{v-z}{s})| \leq |u-v| + 2|t-s| + |y-z|,$$

which implies

$$d^2_{P_n,2}(t\varphi\circ(\frac{f-y}{t}); s\varphi\circ(\frac{g-z}{s})) \leq 3d^2_{P_n,2}(f,g) + 12|t-s|^2 + 3|y-z|^2.$$

Defining

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

and

$$Z_2(f,t,y) := \sqrt{3}n^{-1/2}\sum_{i=1}^{n} g_i f(X_i) + 2\sqrt{3}tg + \sqrt{3}y\tilde{g},$$

we have

$$\mathbb{E}_g|Z_1(f,t,y) - Z_1(g,s,z)|^2 \leq \mathbb{E}_g|Z_2(f,t,y) - Z_2(g,s,z)|^2.$$

The same way as in the proof of Theorem 1, Slepian's Lemma implies that

$$\mathbb{E}\sup\{Z_1(f,t,y) : y \in [-1,1], f \in \mathcal{F}, t \in (0,1] \text{ or } -f \in \mathcal{F}, -t \in (0,1]\} \leq$$

$$\leq \mathbb{E}\sup\{Z_2(f,t,y)| : y \in [-1,1], f \in \mathcal{F}, t \in (0,1] \text{ or } -f \in \mathcal{F}, -t \in (0,1]\}.$$

Again, arguing as in the proof of Theorem 1, we get

$$\mathbb{E}_g\|n^{-1}\sum_{i=1}^n g_i\delta_{X_i}\|_{\tilde{\mathcal{G}}} \leq \sqrt{3}\mathbb{E}_g\|n^{-1}\sum_{i=1}^n g_i\delta_{X_i}\|_{\mathcal{F}} + \frac{3\sqrt{3}\mathbb{E}|g|}{\sqrt{n}}. \tag{10}$$

Next, by the exponential inequalities for martingale difference sequences, we have

$$\mathbb{P}\Big\{\|P_n - P\|_{\tilde{\mathcal{G}}} \geq \mathbb{E}\|P_n - P\|_{\tilde{\mathcal{G}}} + \varepsilon\Big\} \leq \exp\{-2\varepsilon^2 n\} \tag{11}$$

and, similarly to (3),

$$\mathbb{E}\|P_n - P\|_{\tilde{\mathcal{G}}} \leq \sqrt{2\pi}\mathbb{E}\|n^{-1}\sum_{i=1}^n g_i\delta_{X_i}\|_{\tilde{\mathcal{G}}}. \tag{12}$$

Combining the bounds (9)–(112) and setting $\varepsilon := tn^{-1/2}$, we get that with probablity at least $1 - \exp\{-2t^2\}$

$$\sup_{f\in\mathcal{F}} L(F_{n,f}, F_f) \leq (\sqrt{6\pi}\mathbb{E}\|n^{-1}\sum_{i=1}^n g_i\delta_{X_i}\|_{\mathcal{F}} + \frac{t+6\sqrt{3}}{\sqrt{n}})^{1/2},$$

which completes the proof.

□

Theorem 5 *The following two statements are equivalent:*

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

and

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

Moreover, if for some $\alpha \in [1/2, 1)$

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

then

$$\sup_{f\in\mathcal{F}} L(F_{n,f}, F_f) = O_P(n^{-\alpha/2}) \text{ as } n \to \infty. \tag{13}$$

In particular, if $\mathcal{F} \in CLT(P)$, *then*

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

Proof. To prove that (i) implies (ii) we use Gaussian Multiplier Inequality (see van der Vaart and Wellner (1996)):

$$\mathbb{E}\|n^{-1}\sum_{i=1}^{n} g_i\delta_{X_i}\|_{\mathcal{F}} \le 2(n_0-1)\mathbb{E}\max_{1\le i\le n}\frac{|g_i|}{n}+$$

$$+\|g\|_{2,1}n^{-1/2}\max_{n_0\le k\le n}\mathbb{E}\|k^{-1/2}\sum_{i=n_0}^{k}\varepsilon_i\delta_{X_i}\|_{\mathcal{F}}.$$

For $n_0 = 1$, we get

$$\mathbb{E}\|n^{-1}\sum_{i=1}^{n} g_i\delta_{X_i}\|_{\mathcal{F}} \le \|g\|_{2,1}n^{-1/2}\max_{1\le k\le n}\mathbb{E}\|k^{-1/2}\sum_{i=1}^{k}\varepsilon_i\delta_{X_i}\|_{\mathcal{F}} \le$$

$$\le \|g\|_{2,1}n^{-1/2}\max_{1\le k\le m}\mathbb{E}\|k^{-1/2}\sum_{i=1}^{k}\varepsilon_i\delta_{X_i}\|_{\mathcal{F}} + \|g\|_{2,1}\sup_{k\ge m}\mathbb{E}\|k^{-1}\sum_{i=1}^{k}\varepsilon_i\delta_{X_i}\|_{\mathcal{F}}.$$

Since $\mathcal{F}\in GC(P)$, we have

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

which, by symmetrization inequality, implies

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

Therefore,

$$\sup_{k\ge m}\mathbb{E}\|k^{-1}\sum_{i=1}^{k}\varepsilon_i\delta_{X_i}\|_{\mathcal{F}} \to 0 \text{ as } m\to\infty.$$

On the other hand, for a fixed m,

$$n^{-1/2}\max_{1\le k\le m}\mathbb{E}\|k^{-1/2}\sum_{i=1}^{k}\varepsilon_i\delta_{X_i}\|_{\mathcal{F}} \le (m/n)^{1/2} \to 0 \text{ as } n\to\infty.$$

Therefore,

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

Plugging in the bound of Theorem 4 $t = \log n$ and using Borel-Cantelli Lemma proves (ii).

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

$$|\int_{-1}^{1} td(F-G)(t)| \le 2L(F,G),$$

which holds for any two distribution functions on $[-1,1]$. The bound implies that

$$\|P_n - P\|_{\mathcal{F}} = \sup_{f\in\mathcal{F}}|P_nf - Pf| = \sup_{f\in\mathcal{F}}|\int_{-1}^{1} td(F_{n,f}-F_f)(t)| \le 2\sup_{f\in\mathcal{F}}L(F_{n,f};F_f),$$

which implies the statement.

If

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

then the Gaussian Multiplier Inequality and the symmetrization inequality imply that

$$\mathbb{E}\|n^{-1}\sum_{i=1}^{n} g_i \delta_{X_i}\|_{\mathcal{F}} = O(n^{-\alpha}).$$

Thus, the bound of Theorem 9 implies that with some constant $C > 0$

$$\mathbb{P}\{\sup_{f\in\mathcal{F}} L(F_{n,f}, F_f) \geq (\frac{C}{n^{\alpha}} + \frac{t+6\sqrt{3}}{n^{1/2}})^{1/2}\} \leq \exp\{-2t^2\}.$$

It follows that

$$\lim_{u\to\infty} \limsup_{n\to\infty} \mathbb{P}\{n^{\alpha/2} \sup_{f\in\mathcal{F}} L(F_{n,f}, F_f) \geq u\} = 0.$$

□

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

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

where $\alpha \geq 0$ and ε_k^n are i.i.d. Rademacher random variables.

Proposition 2 *Consider the sequence $\delta = \delta(n)$ such that*

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

Then

$$\delta \geq cn^{-\frac{1+2\alpha}{4(1+\alpha)}}$$

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

$$\sup_{f\in\mathcal{F}} L(F_{n,f}, F_f) = O_P(n^{-\frac{1+2\alpha}{4(1+\alpha)}}).$$

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

$$P(f_k \leq y) \leq P_n(f_k \leq y + \delta) + \delta. \tag{14}$$

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

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

Inequality $(2\log(k+1))^{\alpha+1/2} < \delta^{-1}$ holds for $k \le \psi_1(\delta) = 1/2\exp(\delta^{-\frac{2}{1+2\alpha}}/2)$. Therefore, for large n

$$\begin{aligned} 1/2 &\le \mathbb{P}\left(\bigcap_{k\le\psi_1(\delta)}\left\{1/2 \le n^{-1}\sum_{i\le n} I(\varepsilon_k^i = -1) + \delta\right\}\right) \\ &= \mathbb{P}\left(1/2 \le n^{-1}\sum_{i\le n} I(\varepsilon_1^i = -1) + \delta\right)^{\psi_1(\delta)} \le \left(1 - \binom{n}{k_0}2^{-n}\right)^{\psi_1(\delta)}, \end{aligned} \tag{15}$$

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

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

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

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

Therefore,

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

and

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

This finally implies that

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

To prove the second statement note, first of all, that in the supremum $\sup_{f\in\mathcal{F}} L(F_{n,f}, F_f)$ it's enough to consider only those k that satisfy the inequality

$$2\delta^{-1} \ge (2\log(k+1))^{1/2+\alpha},$$

because, for all other k we automatically have

$$P(f_k \le y) \le P_n(f_k \le y+\delta) + \delta, \quad P_n(f_k \le y) \le P(f_k \le y+\delta) + \delta \tag{16}$$

for all y. The above condition on k is equivalent to

$$k \le \psi_2(\delta) = \exp\left(\frac{1}{2}\left(\frac{2}{\delta}\right)^{\frac{2}{1+2\alpha}}\right).$$

Let us notice that probability $P(f_k \le y)$ can take one of three values $0, 1/2$ or 1. When it's equal to 0 or 1, conditions (**??**) become trivial. On the other hand, if $P(f_k \le y) = 1/2$ then, obviously, (**??**) hold when $|n^{-1}\sum I(\varepsilon_k^i = -1) - 1/2| \le \delta$.

This observations imply that

$$\begin{aligned} &\mathbb{P}\left(L(F_{n,f}, F_f) \le \delta\right) \ge \mathbb{P}\left(\bigcap_{k\le\psi_2(\delta)}\left\{\left|\frac{1}{n}\sum_{i\le n} I(\varepsilon_k^i = -1) - \frac{1}{2}\right| \le \delta\right\}\right) \\ &= \mathbb{P}\left(\left|\frac{1}{n}\sum_{i\le n} I(\varepsilon_1^i = -1) - \frac{1}{2}\right| \le \delta\right)^{\psi_2(\delta)} = \left(1 - \mathbb{P}\left(\left|\frac{1}{n}\sum_{i\le n} I(\varepsilon_1^i = -1) - \frac{1}{2}\right| \le \delta\right)\right)^{\psi_2(\delta)} \\ &\ge (1 - 2e^{-2n\delta^2})^{\psi_2(\delta)} \to 1, \end{aligned}$$

when

$$\psi_2(\delta)e^{-2n\delta^2} = \exp\left(\frac{1}{2}\left(\frac{2}{\delta}\right)^{\frac{2}{1+2\alpha}} - 2n\delta^2\right) \to 0.$$

This holds, for instance, when $\delta = n^{-(1+2\alpha)/4(1+\alpha)}$.

4 Bounding the generalization error of neural networks and other combined classifiers

In this section, we assume that $\tilde{S} := S \times \{-1, 1\}$ and $\tilde{\mathcal{F}} := \{\tilde{f} : f \in \mathcal{F}\}$, 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

$$G_n(\tilde{\mathcal{F}}) = 2\sqrt{\pi}\mathbb{E}\sup_{f\in\mathcal{F}}|n^{-1}\sum_{i=1}^n g_i Y_i f(X_i)| = 2\sqrt{\pi}\mathbb{E}\mathbb{E}_g\sup_{f\in\mathcal{F}}|n^{-1}\sum_{i=1}^n \tilde{g}_i f(X_i)|,$$

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

$$\mathbb{E}_g\sup_{f\in\mathcal{F}}|n^{-1}\sum_{i=1}^n \tilde{g}_i f(X_i)| = \mathbb{E}_g\sup_{f\in\mathcal{F}}|n^{-1}\sum_{i=1}^n g_i f(X_i)|,$$

which immediately implies $G_n(\tilde{\mathcal{F}}) = G_n(\mathcal{F})$.

Theorem 1 now implies some useful bounds for boosting and other methods of combining the classifiers. Namely, we get in this case the following theorem that implies (and slightly improves) the recent bound of Schapire, Freund, Bartlett and Lee (1998).

Theorem 6 *Let $\mathcal{F} := \mathrm{conv}(\mathcal{H})$, where $\mathcal{H}$ is a class of measurable functions from $(S, \mathcal{A})$ into $[-1, 1]$. For all $t > 0$,*

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

and

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

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

Proof. Since $\mathcal{F} := \mathrm{conv}(\mathcal{H})$, where $\mathcal{H}$ is a class of measurable functions from $(S, \mathcal{A})$ into $[-1, 1]$, we have

$$G_n(\mathcal{F}) = 2\sqrt{\pi}\mathbb{E}\|n^{-1}\sum_{i=1}^n g_i\delta_{X_i}\|_{\mathcal{F}} \le 2\sqrt{\pi}\mathbb{E}\|n^{-1}\sum_{i=1}^n g_i\delta_{X_i}\|_{\mathcal{H}} = G_n(\mathcal{H}).$$

It follows that $G_n(\tilde{\mathcal{F}}) \le G_n(\mathcal{H})$, and theorems 1 and 3 imply the result.

□

In voting methods of combining the classifiers (such as boosting, bagging, etc.), a classifier produced at each iteration is a convex combination $\hat{f} \in \mathrm{conv}(\mathcal{H})$ of simple base classifiers from the class $\mathcal{H}$. The first bound of Theorem 4 implies that for a given $\alpha \in (0,1)$ with probability at least $1-\alpha$

$$P\{\tilde{\hat{f}} \le 0\} \le \inf_{\delta \in (0,1]} [P_n\{\tilde{\hat{f}} \le \delta\} + \frac{1}{\delta}\Delta_n(\mathcal{H}; \sqrt{\frac{1}{2}\log\frac{1}{\alpha}})].$$

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 $C > 0$

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

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

$$P\{\tilde{\hat{f}} \le 0\} \le \inf_{\delta \in (0,1]} [P_n\{\tilde{\hat{f}} \le \delta\} + \frac{1}{\delta}(C\sqrt{\frac{V(\mathcal{H})}{n}} + \frac{\sqrt{\frac{1}{2}\log\frac{1}{\alpha}} + 4\sqrt{2}}{\sqrt{n}})],$$

which slightly improves the bound obtained previously by Schapire, Freund, Bartlett and Lee (1998).

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 k-th iteration. Let $P_{n,w^{(k)}}$ be the weighted empirical measure on k-th 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 k-th 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

$$\hat{f}(x) := \frac{\sum_{k=1}^N \alpha_k h_k(x)}{\sum_{k=1}^N \alpha_k}.$$

The above bounds, of course, apply to this classifier since $\hat{f} \in \mathrm{conv}(\mathcal{H})$. Another way to use Theorem 4 in this example to choose a decreasing function φ, satisfying all the conditions of the Theorem 4 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

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

Then it is not hard to check that

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

Therefore

$$P_n\varphi(\frac{\tilde{\hat{f}}}{\delta}) \le P_n \exp\{-y\sum_1^N \alpha_k h_k(x)\}.$$

A simple (and well known in the literature on boosting) computation shows that

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

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 first bound of Theorem 4 that with probability $1-e^{-2t^2}$

$$P\{\tilde{f} \le 0\} \le \prod_{k=1}^N 2\sqrt{e_k(1-e_k)} + \log\prod_{k=1}^N \sqrt{\frac{1-e_k}{e_k}}\Delta_n(\mathcal{H};t).$$

We turn now to the applications of the bounds of previous section in neural network learning. Let $\mathcal{H}$ be a class of measurable functions from $(S,\mathcal{A})$ into $\mathbb{R}$. Given a Borel function σ from $\mathbb{R}$ into $[-1,1]$ and a vector $w := (w_1,\dots,w_n) \in \mathbb{R}^n$, let

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

We call the function $N_{\sigma,w}$ *a neuron* with weights w and sigmoid σ. Given a neuron N, we denote $\sigma^{(N)}$ and $w^{(N)}$ its sigmoid and its vector of weights, respectively. 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:

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

Define $\mathcal{H}_0 := \mathcal{H}$, and then recursively

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

We call $\mathcal{H}_j$ the class of feedforward neural networks with base $\mathcal{H}$ and j layers of neurons. Denote

$$\mathcal{H}_\infty := \bigcup_{j=0}^{\infty} \mathcal{H}_j.$$

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

$$\mathcal{H}_j(A_1,\ldots,A_j) :=$$

$$:= \Big\{N_{\sigma_j,w}(h_1,\ldots,h_n) : n \geq 0, h_i \in \mathcal{H}_{j-1}(A_1,\ldots,A_{j-1}),\ w \in \mathbb{R}^n, \|w\|_{\ell_1} \leq A_j\Big\} \bigcup$$

$$\bigcup \mathcal{H}_{j-1}(A_1,\ldots,A_{j-1}).$$

Clearly,

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

We start with the following result.

Theorem 7 *For all $t > 0$ and for all $l \geq 1$*

$$\mathbb{P}\Big\{\exists f \in \mathcal{H}_l(A_1,\ldots,A_l) : P\{\tilde{f} \leq 0\} > \inf_{\delta\in(0,1]}\Big[P_n\varphi(\frac{\tilde{f}}{\delta}) +$$

$$+\frac{1}{\delta}(\prod_{k=1}^{l}(2L_jA_j+1)G_n(\mathcal{H}) + \frac{t+4\sqrt{2}}{\sqrt{n}})\Big]\Big\} \leq \exp\{-2t^2\}$$

and

$$\mathbb{P}\Big\{\exists f \in \mathcal{H}_l(A_1,\ldots,A_l) : |P_n\{\tilde{f} \leq 0\} - P\{\tilde{f} \leq 0\}| > \inf_{\delta\in(0,1]}\Big[P_n\{|\tilde{f}| \leq \delta\} +$$

$$+\frac{1}{\delta}(\prod_{k=1}^{l}(2L_jA_j+1)G_n(\mathcal{H}) + \frac{t+4\sqrt{2}}{\sqrt{n}})\Big]\Big\} \leq 2\exp\{-2t^2\},$$

$$\mathbb{P}\Big\{\exists f \in \mathcal{H}_l(A_1,\ldots,A_l) : |P_n\{\tilde{f} \leq 0\} - P\{\tilde{f} \leq 0\}| > \inf_{\delta\in(0,1]}\Big[P\{|\tilde{f}| \leq \delta\}$$

$$+\frac{1}{\delta}(\prod_{k=1}^{l}(2L_jA_j+1)G_n(\mathcal{H}) + \frac{t+4\sqrt{2}}{\sqrt{n}})\Big]\Big\} \leq 2\exp\{-2t^2\}.$$

Proof. We apply Theorem 1 and Theorem 3 to the class $\mathcal{F} = \mathcal{H}_l(A_1, \ldots, A_l) =: \mathcal{H}'_l$, which gives for all $t > 0$

$$\mathbb{P}\Big\{\exists f \in \mathcal{H}'_l : P\{\tilde{f} \le 0\} > \inf_{\delta \in (0,1)} \Big[P_n\varphi(\frac{\tilde{f}}{\delta}) + \frac{1}{\delta}(G_n(\mathcal{H}'_l) + \frac{t + 4\sqrt{2}}{\sqrt{n}})\Big]\Big\} \le \exp\{-2t^2\}.$$

Thus, it's enough to show that

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

To this end, note that

$$\mathbb{E}\|n^{-1}\sum_{i=1}^n g_i\delta_{X_i}\|_{\mathcal{H}'_l} \le \mathbb{E}\|n^{-1}\sum_{i=1}^n g_i\delta_{X_i}\|_{\mathcal{G}_l} + \mathbb{E}\|n^{-1}\sum_{i=1}^n g_i\delta_{X_i}\|_{\mathcal{H}'_{l-1}}, \tag{17}$$

where

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

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

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

We have

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

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

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

$$\begin{aligned}\mathbb{E}_g\|n^{-1}\sum_{i=1}^n g_i\delta_{X_i}\|_{\mathcal{G}_l} &= n^{-1/2}\mathbb{E}_g\|Z_1\|_{\mathcal{G}'_l} \\ &\le 2n^{-1/2}\mathbb{E}_g\|Z_2\|_{\mathcal{G}'_l} = 2L_l\mathbb{E}_g\|n^{-1}\sum_{i=1}^n g_i\delta_{X_i}\|_{\mathcal{G}'_l}. \end{aligned}\tag{18}$$

Since $\mathcal{G}'_l = A_l\mathrm{conv}(\mathcal{H}_{l-1})$, we get

$$\mathbb{E}\|n^{-1}\sum_{i=1}^n g_i\delta_{X_i}\|_{\mathcal{G}'_l} = A_l\mathbb{E}\|n^{-1}\sum_{i=1}^n g_i\delta_{X_i}\|_{\mathcal{H}_{l-1}}. \tag{19}$$

It follows from the bounds (17)–(19) that

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

The result now follows by induction.

□

Remark. Bartlett (1998) obtained a bound similar to the first inequality of Theorem 7 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 $f \in \mathcal{H}_\infty$, let

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

Let $\{b_k\}$ be a sequence of nonnegative numbers. For a number $k, 1 \le k \le \ell(f)$, let $\mathcal{N}_k(f)$ denote the set of all neurons of layer k (with sigmoid σ_k) in the representation of f. Denote

$$W_k(f) := \max_{N\in\mathcal{N}_k(f)} \|w^{(N)}\|_{\ell_1} \bigvee b_k,\ \ k = 1, 2, \dots, \ell(f),$$

and let

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

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

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

$$\mathbb{P}\Big\{\exists f \in \mathcal{H}_\infty : P\{\tilde{f} \le 0\} > \inf_{\delta\in(0,1)}\Big[P_n\varphi(\frac{\tilde{f}}{\delta})+$$

$$+\frac{1}{\delta}(\Lambda(f)G_n(\mathcal{H}) + \frac{\Gamma_\alpha(f)+t+4\sqrt{2}}{\sqrt{n}})\Big]\Big\} \le (3-2\zeta(\alpha))^{-1}\exp\{-2t^2\}$$

and

$$\mathbb{P}\Big\{\exists f \in \mathcal{H}_\infty : |P_n\{\tilde{f} \le 0\} - P\{\tilde{f} \le 0\}| > \inf_{\delta\in(0,1]}\Big[P_n\{|\tilde{f}| \le \delta\}+$$

$$+\frac{1}{\delta}(\Lambda(f)G_n(\mathcal{H}) + \frac{\Gamma_\alpha(f)+t+4\sqrt{2}}{\sqrt{n}})\Big]\Big\} \le 2(3-2\zeta(\alpha))^{-1}\exp\{-2t^2\},$$

$$\mathbb{P}\Big\{\exists f\in\mathcal{H}_\infty : |P_n\{\tilde f\le 0\}-P\{\tilde f\le 0\}| > \inf_{\delta\in(0,1]}\Big[P\{|\tilde f|\le\delta\}+$$

$$+\frac{1}{\delta}(\Lambda(f)G_n(\mathcal{H})+\frac{\Gamma_\alpha(f)+t+4\sqrt{2}}{\sqrt{n}})\Big]\Big\}\le 2(3-2\zeta(\alpha))^{-1}\exp\{-2t^2\}.$$

Proof. 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

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

easily imply that

$$\Lambda(f)\ge\prod_{j=1}^{l}(2L_j2^{k_j}+1),$$

$$\Gamma_\alpha(f)\ge\sum_{j=1}^{l}\sqrt{\frac{\alpha}{2}\log(|k_j|+1)}.$$

and also that

$$f\in\mathcal{H}_l(2^{k_1},\dots,2^{k_l}).$$

Therefore, the following bounds hold:

$$\mathbb{P}\Big\{\exists f\in\mathcal{H}_\infty : P\{\tilde f\le 0\} > \inf_{\delta\in(0,1)}\Big[P_n\varphi(\frac{\tilde f}{\delta})+\frac{1}{\delta}(\Lambda(f)G_n(\mathcal{H})+\frac{\Gamma_\alpha(f)+t+4\sqrt{2}}{\sqrt{n}})\Big]\Big\}\le$$

$$\le\sum_{l=0}^{\infty}\sum_{k_1\in\mathbb{Z}\setminus\{0\}}\dots\sum_{k_l\in\mathbb{Z}\setminus\{0\}}\mathbb{P}\Big\{\exists f\in\mathcal{H}_\infty\bigcap\{f:\ell(f)=l,\ W_j(f)\in\Delta_{k_j},\ j=1,\dots,l\}:$$

$$P\{\tilde f\le 0\} > \inf_{\delta\in(0,1)}\Big[P_n\varphi(\frac{\tilde f}{\delta})+\frac{1}{\delta}(\Lambda(f)G_n(\mathcal{H})+\frac{\Gamma_\alpha(f)+t+4\sqrt{2}}{\sqrt{n}})\Big]\Big\}\le$$

$$\le\sum_{l=0}^{\infty}\sum_{k_1\in\mathbb{Z}\setminus\{0\}}\dots\sum_{k_l\in\mathbb{Z}\setminus\{0\}}\mathbb{P}\Big\{\exists f\in\mathcal{H}_l(2^{k_1},\dots,2^{k_l}) : P\{\tilde f\le 0\} > \inf_{\delta\in(0,1)}\Big[P_n\varphi(\frac{\tilde f}{\delta})+$$

$$+\frac{1}{\delta}(\prod_{j=1}^{l}(2L_j2^{k_j}+1)G_n(\mathcal{H})+\frac{\sum_{j=1}^{l}\sqrt{\frac{\alpha}{2}\log(|k_j|+1)}+t+4\sqrt{2}}{\sqrt{n}})\Big]\Big\}.$$

Using the first bound of Theorem 7, we obtain

$$\mathbb{P}\Big\{\exists f\in\mathcal{H}_l(2^{k_1},\dots,2^{k_l}) : P\{\tilde f\le 0\} > \inf_{\delta\in(0,1)}\Big[P_n\varphi(\frac{\tilde f}{\delta})+$$

$$+\frac{1}{\delta}(\Lambda(f)G_n(\mathcal{H})+\frac{\Gamma_\alpha(f)+t+4\sqrt{2}}{\sqrt{n}})\Big]\Big\}\le$$

$$\leq \sum_{l=0}^{\infty} \sum_{k_1 \in \mathbb{Z}\setminus\{0\}} \dots \sum_{k_l \in \mathbb{Z}\setminus\{0\}} \exp\{-2(\sum_{j=1}^{l} \sqrt{\frac{\alpha}{2}\log(|k_j|+1)} + t)^2\} \leq$$

$$\leq \sum_{l=0}^{\infty} \sum_{k_1 \in \mathbb{Z}\setminus\{0\}} \dots \sum_{k_l \in \mathbb{Z}\setminus\{0\}} \exp\{-\sum_{j=1}^{l} \alpha \log(|k_j|+1) - 2t^2\} =$$

$$= \sum_{l=0}^{\infty} \sum_{k_1 \in \mathbb{Z}\setminus\{0\}} \dots \sum_{k_l \in \mathbb{Z}\setminus\{0\}} \prod_{j=1}^{l} (|k_j|+1)^{-\alpha} \exp\{-2t^2\} =$$

$$= \sum_{l=0}^{\infty} \prod_{j=1}^{l} (2\sum_{k=2}^{\infty} k^{-\alpha}) \exp\{-2t^2\} = \sum_{l=0}^{\infty} [2(\zeta(\alpha)-1)]^l \exp\{-2t^2\} =$$

$$= (3-2\zeta(\alpha))^{-1} \exp\{-2t^2\}$$

which yields the first bound of the theorem. Two other bounds are proved quite similarly.

□

It follows, in particular, that for any classifier $\hat{f} \in \mathcal{H}_\infty$, based on the data $(X_1, \dots, X_n)$, we have

$$\mathbb{P}\Big\{P\{\tilde{\hat{f}} \leq 0\} > \inf_{\delta \in (0,1]} \Big[P_n\varphi(\frac{\tilde{\hat{f}}}{\delta}) +$$

$$+\frac{1}{\delta}(\Lambda(\tilde{\hat{f}})G_n(\mathcal{H}) + \frac{\Gamma_\alpha(\hat{f}) + t + 4\sqrt{2}}{\sqrt{n}})\Big]\Big\} \leq (3-2\zeta(\alpha))^{-1}\exp\{-2t^2\},\ t>0.$$

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 $\check{f}$ is the neural network from $\mathcal{F} \subset \mathcal{H}_\infty$ that minimizes the penalized empirical distribution of the margin:

$$\check{f} := \mathrm{argmin}_{f\in\mathcal{F}} \inf_{\delta\in(0,1]} \Big[P_n(\{f \leq \delta\}) + \frac{1}{\delta}(\Lambda(f)G_n(\mathcal{H}) + \frac{\Gamma_\alpha(f)}{\sqrt{n}})\Big] =$$

$$= \mathrm{argmin}_{f\in\mathcal{F}} \Big[P_n(\{f \leq 0\}) + \inf_{\delta\in(0,1]} \hat{\pi}_n(f;\delta)\Big],$$

where the quantity

$$\hat{\pi}_n(f;\delta) := P_n(\{0 < f \leq \delta\}) + \frac{1}{\delta}(\Lambda(f)G_n(\mathcal{H}) + \frac{\Gamma_\alpha(f)}{\sqrt{n}})$$

plays the role of complexity penalty. We define a distribution dependent version of this data dependent penalty as

$$\pi_n(f;\delta) := P(\{0 < f \leq 2\delta\}) + \frac{2}{\delta}(\Lambda(f)G_n(\mathcal{H}) + \frac{\Gamma_\alpha(f)}{\sqrt{n}}).$$

The next result is a "oracle inequality" that shows that the estimate $\check{f}$ obtained by the above method posses some optimality property (see Barron, Birgé and Massart (1999) for a general approach to penalization and oracle inequalities in nonparametric statistics).

Theorem 9 *For all $t > 0$ and for all $\alpha > 0$ with $\zeta(\alpha) < 3/2$, the following bounds hold:*

$$\mathbb{P}\Big\{P\{\tilde{\tilde{f}} \le 0\} > \inf_{f\in\mathcal{F}}[P_n\{f \le 0\} + \inf_{\delta\in(0,1]}(\hat{\pi}_n(f;\delta) + \frac{1}{\delta}\frac{t+4\sqrt{2}}{\sqrt{n}})]\Big\} \le (3-2\zeta(\alpha))^{-1}\exp\{-2t^2\}$$

and

$$\mathbb{P}\Big\{P\{\tilde{\tilde{f}} \le 0\} - \inf_{g\in\mathcal{F}} P\{\tilde{g} \le 0\} > \inf_{f\in\mathcal{F}}\Big[P\{\tilde{f} \le 0\} - \inf_{g\in\mathcal{F}} P\{\tilde{g} \le 0\} +$$

$$+ \inf_{\delta\in(0,1]}(\hat{\pi}(f;\delta) + \frac{2}{\delta}\frac{t+4\sqrt{2}}{\sqrt{n}})\Big]\Big\} \le 2(3-2\zeta(\alpha))^{-1}\exp\{-2t^2\}.$$

Proof. The first bound follows from Theorem 8 and the definition of the estimate $\tilde{f}$. To prove the second bound, we repeat the proof of Theorem 1 to show that for any class $\mathcal{F}'$

$$\mathbb{P}\Big\{\exists f \in \mathcal{F}'\ \exists\delta \in (0,1] : P_n\{\tilde{f} \le \delta\} > \Big[P\varphi(\frac{\tilde{f}-\delta}{\delta}) + \frac{1}{\delta}(G_n(\mathcal{F}') + \frac{t+4\sqrt{2}}{\sqrt{n}})\Big]\Big\} \le$$

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

The class $\mathcal{G}$ in this proof is now defined as

$$\mathcal{G} := \Big\{t\varphi(\frac{\tilde{f}-t}{t}) : f \in \mathcal{F}', t \in (0,1]\Big\}.$$

The argument that led to Theorems 7 and 8 shows that

$$\mathbb{P}\Big\{\exists f \in \mathcal{F}\ \exists\delta \in (0,1] : P_n\{\tilde{f} \le \delta\} > \Big[P\{\tilde{f} \le 2\delta\} +$$

$$+\frac{1}{\delta}(\Lambda(f)G_n(\mathcal{H}) + \frac{\Gamma_\alpha(f)}{\sqrt{n}} + \frac{t+4\sqrt{2}}{\sqrt{n}})\Big]\Big\} \le (3-2\zeta(\alpha))^{-1}\exp\{-2t^2\}.$$

If now

$$\inf_{f\in\mathcal{F}}\inf_{\delta\in(0,1]}\Big[P_n(\{\tilde{f} \le \delta\}) + \frac{1}{\delta}(\Lambda(f)G_n(\mathcal{H}) + \frac{\Gamma_\alpha(f)+t+4\sqrt{2}}{\sqrt{n}})\Big] >$$

$$> \inf_{f\in\mathcal{F}}\inf_{\delta\in(0,1]}\Big[P\{\tilde{f} \le 2\delta\} + \frac{2}{\delta}(\Lambda(f)G_n(\mathcal{H}) + \frac{\Gamma_\alpha(f)+t+4\sqrt{2}}{\sqrt{n}})\Big],$$

then

$$\exists f \in \mathcal{F}\exists\delta \in (0,1] : P_n\{\tilde{f} \le \delta\} > \Big[P\{\tilde{f} \le 2\delta\} + \frac{1}{\delta}(\Lambda(f)G_n(\mathcal{H}) + \frac{\Gamma_\alpha(f)+t+4\sqrt{2}}{\sqrt{n}})\Big].$$

Combining this with the first bound gives

$$\mathbb{P}\Big\{P\{\tilde{\tilde{f}} \le 0\} > \inf_{f\in\mathcal{F}}\inf_{\delta\in(0,1)}\Big[P\{\tilde{f} \le 2\delta\} +$$

$$+\frac{2}{\delta}(\Lambda(f)G_n(\mathcal{H}) + \frac{\Gamma_\alpha(f)+t+4\sqrt{2}}{\sqrt{n}})\Big]\Big\} \le 2(3-2\zeta(\alpha))^{-1}\exp\{-2t^2\},$$

which implies the result.

□

Finally, it is worth mentioning that the theorems of Section 1 can be applied also to bounding the generalization error in multiclass problems. Namely, we assume that the labels take values in a finite set $\mathcal{Y}$ with $\mathrm{card}(\mathcal{Y}) =: L$. Consider a class $\tilde{\mathcal{F}}$ of functions from $\tilde{S} := S \times \mathcal{Y}$ into $\mathbb{R}$. A function $f \in \tilde{\mathcal{F}}$ predicts a label $y \in \mathcal{Y}$ for an example $x \in S$ iff

$$f(x,y) > \max_{y' \neq y} f(x,y').$$

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

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

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

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

The next result follows from Theorem 1.

Theorem 10 *For all $t > 0$,*

$$\mathbb{P}\Big\{\exists f \in \tilde{\mathcal{F}} : P\{m_f \leq 0\} > \inf_{\delta \in (0,1]} \Big[P_n\{m_f \leq \delta\} + \frac{1}{\delta}(L^2 G_n(\mathcal{F}) + \frac{t+4\sqrt{2}}{\sqrt{n}})\Big]\Big\} \leq \exp\{-2t^2\}.$$

To prove the theorem, we use the following easy lemma.

For a class of functions $\mathcal{H}$ we will denote by

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

Lemma 1 *The following bound holds:*

$$\mathbb{E}\|\sum_{i=1}^n g_i \delta_{X_i}\|_{\mathcal{H}^{(l)}} \leq l\mathbb{E}\|\sum_{i=1}^n g_i \delta_{X_i}\|_{\mathcal{H}}.$$

Proof. 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

$$\max(f_1,f_2) = \frac{1}{2}\left(|f_1+f_2| + |f_1-f_2|\right)$$

we have

$$\mathbb{E}\|\sum_{i=1}^n g_i \delta_{X_i}\|_{\mathcal{F}} \leq \frac{1}{2}\mathbb{E}\sup_{\mathcal{F}_1,\mathcal{F}_2} |\sum_{i=1}^n g_i |f_1(X_i)+f_2(X_i)|| +$$

$$\frac{1}{2}\mathbb{E}\sup_{\mathcal{F}_1,\mathcal{F}_2} |\sum_{i=1}^n g_i |f_1(X_i)-f_2(X_i)|| \leq \frac{1}{2}\mathbb{E}\sup_{\mathcal{F}_1,\mathcal{F}_2} |\sum_{i=1}^n g_i (f_1(X_i)+f_2(X_i))|$$

$$+\frac{1}{2}\mathbb{E}\sup_{\mathcal{F}_1,\mathcal{F}_2} |\sum_{i=1}^n g_i (f_1(X_i)-f_2(X_i))| \leq \mathbb{E}\|\sum_{i=1}^n g_i \delta_{X_i}\|_{\mathcal{F}_1} + \mathbb{E}\|\sum_{i=1}^n g_i \delta_{X_i}\|_{\mathcal{F}_2}.$$

The statement of lemma follows by induction over l.

□

Proof of Theorem 10. We have the following bounds:

$$\mathbb{E}\sup_{f\in\tilde{\mathcal{F}}}\Big|n^{-1}\sum_{j=1}^{n}g_j m_f(X_j,Y_j)\Big| = \mathbb{E}\sup_{f\in\tilde{\mathcal{F}}}\Big|n^{-1}\sum_{j=1}^{n}g_j\sum_{y\in\mathcal{Y}}m_f(X_j,y)I_{\{Y_j=y\}}\Big| \leq$$

$$\leq \sum_{y\in\mathcal{Y}}\mathbb{E}\sup_{f\in\tilde{\mathcal{F}}}\Big|n^{-1}\sum_{j=1}^{n}g_j m_f(X_j,y)I_{\{Y_j=y\}}\Big| \leq$$

$$\leq \frac{1}{2}\sum_{y\in\mathcal{Y}}\mathbb{E}\sup_{f\in\tilde{\mathcal{F}}}\Big|n^{-1}\sum_{j=1}^{n}g_j m_f(X_j,y)(2I_{\{Y_j=y\}}-1)\Big| + \frac{1}{2}\sum_{y\in\mathcal{Y}}\mathbb{E}\sup_{f\in\tilde{\mathcal{F}}}\Big|n^{-1}\sum_{j=1}^{n}g_j m_f(X_j,y)\Big|.$$

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

$$\mathbb{E}\sup_{f\in\tilde{\mathcal{F}}}\Big|n^{-1}\sum_{j=1}^{n}g_j m_f(X_j,y)(2I_{Y_j=y}-1)\Big| = \mathbb{E}\sup_{f\in\tilde{\mathcal{F}}}\Big|n^{-1}\sum_{j=1}^{n}g_j\sigma_j(y)m_f(X_j,y)\Big| =$$

$$= \mathbb{E}\mathbb{E}_g\sup_{f\in\tilde{\mathcal{F}}}\Big|n^{-1}\sum_{j=1}^{n}g_j\sigma_j(y)m_f(X_j,y)\Big| = \mathbb{E}\mathbb{E}_g\sup_{f\in\tilde{\mathcal{F}}}\Big|n^{-1}\sum_{j=1}^{n}g_j m_f(X_j,y)\Big| =$$

$$= \mathbb{E}\sup_{f\in\tilde{\mathcal{F}}}\Big|n^{-1}\sum_{j=1}^{n}g_j m_f(X_j,y)\Big|$$

Therefore, we have

$$\mathbb{E}\sup_{f\in\tilde{\mathcal{F}}}\Big|n^{-1}\sum_{j=1}^{n}g_j m_f(X_j,Y_j)\Big| \leq \sum_{y\in\mathcal{Y}}\mathbb{E}\sup_{f\in\tilde{\mathcal{F}}}\Big|n^{-1}\sum_{j=1}^{n}g_j m_f(X_j,y)\Big|.$$

Next, using Lemma 1, we get for all $y\in\mathcal{Y}$

$$\mathbb{E}\sup_{f\in\tilde{\mathcal{F}}}\Big|n^{-1}\sum_{j=1}^{n}g_j m_f(X_j,y)\Big| \leq \mathbb{E}\sup_{f\in\tilde{\mathcal{F}}}\Big|n^{-1}\sum_{j=1}^{n}g_j f(X_j,y)\Big| + \mathbb{E}\sup_{f\in\tilde{\mathcal{F}}}\Big|n^{-1}\sum_{j=1}^{n}g_j\max_{y'\neq y}f(X_j,y')\Big| \leq$$

$$\leq \mathbb{E}\sup_{f\in\mathcal{F}}\Big|n^{-1}\sum_{j=1}^{n}g_j f(X_j)\Big| + \mathbb{E}\sup_{f\in\mathcal{F}^{(L-1)}}\Big|n^{-1}\sum_{j=1}^{n}g_j f(X_j)\Big| \leq$$

$$\leq L\mathbb{E}\sup_{f\in\mathcal{F}}\Big|n^{-1}\sum_{j=1}^{n}g_j f(X_j)\Big|,$$

and the result follows from the above bounds and from Theorem 1.

□

References

[1] Barron, A., Birgé, L. and Massart, P. (1999) Risk Bounds for Model Selection via Penalization. *Probability Theory and Related Fields*, to appear.

[2] Bartlett, P. (1998) The Sample Complexity of Pattern Classification with Neural Networks: The Size of the Weights is More Important than the Size of the Network. *IEEE Transactions on Information Theory*, 44, 525-536.

[3] Birgé, L. and Massart, P. (1997) From Model Selection to Adaptive Estimation. In: Festschrift for L. Le Cam. Research Papers in Probability and Statistics. D. Pollard, E. Torgersen and G. Yang (Eds.), 55-87. Springer, New York.

[4] Devroye, L., Györfi, L. and Lugosi, G. (1996) A probabilistic theory of pattern recognition. Springer-Verlag, New York.

[5] Dudley, R.M. (1999) Uniform Central Limit Theorems. Cambridge University Press.

[6] Koltchinskii, V. (1999) Rademacher penalties and structural risk minimization, preprint.

[7] Ledoux, M. and Talagrand, M. (1991) Probability in Banach Spaces. Springer-Verlag, New York.

[8] Schapire, R., Freund, Y., Bartlett, P. and Lee, W. S. (1998) Boosting the Margin: A New Explanation of Effectiveness of Voting Methods. *Ann. Statist.*, to appear.

[9] van der Vaart, A. and Wellner, J. (1996) Weak convergence and Empirical Processes. With Applications to Statistics. Springer-Verlag, New York.

[10] Vapnik, V. (1998) Statistical Learning Theory. John Wiley & Sons, New York.

[11] Vidyasagar, M. (1997) A theory of learning and generalization. Springer-Verlag, New York.