RADEMACHER PROCESSES AND BOUNDING THE RISK OF FUNCTION LEARNING

V. KOLTCHINSKII AND D. PANCHENKO

Abstract. We construct data dependent upper bounds on the risk in function learning problems. The bounds are based on the local norms of the Rademacher process indexed by the underlying function class and they do not require prior knowledge about the distribution of training examples or any specific properties of the function class. Using Talagrand's type concentration inequalities for empirical and Rademacher processes, we show that the bounds hold with high probability that decreases exponentially fast when the sample size grows. In typical situations that are frequently encountered in the theory of function learning, the bounds give nearly optimal rate of convergence of the risk to zero.

1. Local Rademacher norms and bounds on the risk: main results

Let $(S, \mathcal{A})$ be a measurable space and let $\mathcal{F}$ be a class of $\mathcal{A}$-measurable functions from S into $[0,1]$. Denote $\mathcal{P}(S)$ the set of all probability measures on $(S, \mathcal{A})$. Let $f_0 \in \mathcal{F}$ be an unknown *target function*. Given a probability measure $P \in \mathcal{P}(S)$ (also unknown), let $(X_1, \dots, X_n)$ be an i.i.d. sample in $(S, \mathcal{A})$ with common distribution P (defined on a probability space $(\Omega, \Sigma, \mathbb{P})$). In computer learning theory, the problem of estimating f_0, based on the labeled sample $(X_1, Y_1), \dots, (X_n, Y_n)$, where $Y_j := f_0(X_j)$, $j = 1, \dots, n$, is referred to as *function learning problem*. The so called *concept learning* is a special case of function learning. In this case, $\mathcal{F} := \{I_C : C \in \mathcal{C}\}$, where $\mathcal{C} \subset \mathcal{A}$ is called a class of concepts (see Vapnik (1998), Vidyasagar (1996), Devroye, Györfi and Lugosi (1996) for the account on statistical learning theory). The goal of function learning is to find an estimate $\hat{f}_n := \hat{f}_n((X_1, Y_1), \dots, (X_n, Y_n))$ of the unknown target function such that the L_1-distance between $\hat{f}_n$ and f_0 becomes small with high probability as soon as the sample size becomes large enough. The L_1-distance $P|\hat{f}_n - f_0|$ is often called *the risk* (also the generalization, or prediction error) of the estimate $\hat{f}_n$. A class $\mathcal{F}$ is called *probably approximately correctly (PAC) learnable* iff for all $\varepsilon > 0$

$$\pi_n(\mathcal{F}; \varepsilon) := \sup_{P \in \mathcal{P}(S)} \sup_{f_0 \in \mathcal{F}} \mathbb{P}\big\{P|\hat{f}_n - f_0| \geq \varepsilon\big\} \to 0 \text{ as } n \to \infty.$$

The bounds on the probability $\pi_n(\mathcal{F}; \varepsilon)$ are of importance in the theory. Such bounds allow one to determine the quantity

$$N_{\mathcal{F}}(\varepsilon; \delta) := \inf\{n : \pi_n(\mathcal{F}; \varepsilon) \leq \delta\},$$

which is called *the sample complexity of learning*. Unfortunately, a bound that is uniform in the class of all distributions $\mathcal{P}(S)$ is not necessarily tight for a particular distribution P and often such a bound does not provide a reasonable estimate of

the minimal sample size needed to achieve certain accuracy of learning in the case of a particular P.

A natural approach to the function learning problem (in the case when $f_0 \in \mathcal{F}$) is to find $\hat{f}_n \in \mathcal{F}$ such that $\hat{f}_n(X_j) = Y_j$ for all $j = 1, \dots, n$. In learning theory, such an estimate $\hat{f}_n$ is called *consistent* (this notion should not be confused with consistency in statistical sense).

We construct below a data dependent bound on the risk of a consistent estimate $\hat{f}_n$. More precisely, given $\delta > 0$, we define a quantity

$$\hat{\beta}_n(\mathcal{F}; \delta) = \hat{\beta}_n(\mathcal{F}; \delta; (X_1, Y_1), \dots, (X_n, Y_n))$$

such that for any consistent estimate $\hat{f}_n$

$$\sup_{P \in \mathcal{P}(S)} \sup_{f_0 \in \mathcal{F}} \mathbb{P}\big\{P|\hat{f}_n - f_0| \geq \hat{\beta}_n(\mathcal{F}; \delta)\big\} \leq \delta. \tag{1.1}$$

We'll consider below a couple of important examples in which the bound we suggest gives nearly optimal rate of convergence of the risk to 0 as the sample size tends to infinity.

To simplify the notations, we assume without loss of generality that $f_0 \equiv 0$ (otherwise, one can consider instead of $\mathcal{F}$ the class of functions $\{|f - f_0| : f \in \mathcal{F}\}$; note that the values of the functions from this class are known on the sample $(X_1, \dots, X_n)$). We also assume for simplicity that $\mathcal{F}$ is a countable class of functions. This condition can be easily replaced by standard measurability assumptions known in the theory of empirical processes (see, e.g., [4] or [13]; we do not make countability assumption in some of the examples below). Estimates $\hat{f}_n$ are supposed to be $\Sigma \times \mathcal{A}$-measurable. We denote by P_n the empirical measure based on the sample $(X_1, \dots, X_n)$:

$$P_n := n^{-1} \sum_{j=1}^{n} \delta_{X_j},$$

where δ_x is the probability measure concentrated at the point $x \in S$. We also use the notation $\|\cdot\|_{\mathcal{F}}$ for the sup-norm of functions from the class $\mathcal{F}$ into $\mathbb{R}$:

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

Our approach is based on the following simple idea. Denote $B(r) := \{f : P|f| \leq r\}$ and set $r_0^n = 1$. It's clear that for any consistent estimate $\hat{f}_n$ $P_n\hat{f}_n = 0$ and, hence,

$$P\hat{f}_n \leq P_n\hat{f}_n + \|P_n - P\|_{\mathcal{F}} = \|P_n - P\|_{\mathcal{F}} = \|P_n - P\|_{\mathcal{F} \cap B(r_0^n)} =: r_1^n.$$

Therefore, $\hat{f}_n \in \mathcal{F} \bigcap B(r_1^n)$. It means that actually

$$P\hat{f}_n \leq P_n\hat{f}_n + \|P_n - P\|_{\mathcal{F} \cap B_{r_1^n}} = \|P_n - P\|_{\mathcal{F} \cap B_{r_1^n}}.$$

We can repeat this recursive procedure infinitely many times. Namely, if $r_{k+1}^n := \|P_n - P\|_{\mathcal{F} \cap B(r_k^n)}$, then, by induction, $P\hat{f}_n \leq r_k^n$ for any natural k. It is also clear that the sequence $\{r_k^n\}$ is nonincreasing Indeed, by a simple induction argument, we have that $r_k^n \leq r_{k-1}^n$ implies that

$$r_{k+1}^n = \|P_n - P\|_{\mathcal{F} \cap B(r_k^n)} \leq \|P_n - P\|_{\mathcal{F} \cap B(r_{k-1}^n)} = r_k^n.$$

Thus, the following proposition holds.

Proposition 1. *The sequence* $\{r_k^n\}_{k\geq 1}$ *is nonincreasing and for any consistent estimate* $\hat{f}_n$ $P\hat{f}_n \leq \inf_{k\geq 0} r_k^n$.

The sequence $\{r_k^n\}_{k\geq 1}$ depends not only on the data; it also depends explicitly on the unknown distribution P, so it can not be used for the purposes of bounding the risk. However, there is a simple bootstrap type approach that allows one to get around this difficulty.

The Rademacher process indexed by the function class $\mathcal{F}$ is defined as

$$R_n = \frac{1}{n}\sum_{i=1}^n \varepsilon_i \delta_{X_i},$$

where $\{\varepsilon_i\}$ is a Rademacher sequence (an i.i.d. sequence of random variables taking the values $+1$ and -1 with probability $1/2$ each) independent of $\{X_i\}$. It has been used for a long time to obtain the bounds on the sup-norm of the empirical process indexed by functions (in the so called symmetrization inequalities, see [13]). Recently, Koltchinskii [6] (see also [7]) suggested to use $\|R_n\|_{\mathcal{F}}$ as data-based measure of the accuracy of empirical approximation $\|P_n - P\|_{\mathcal{F}}$ in learning problems and developed a version of structural risk minimization in which the norms of Rademacher process play the role of data-dependent penalties. Lozano [8] compared this method of penalization with the method based on VC-dimensions and the cross-validation method and found out that in the so called problem of the "intervals model selection" the Rademacher penalization performs better than other methods. Hush and Scovel (1999) used Rademacher norms to obtain posterior performance bounds for machine learning. However, the "global" norm of Rademacher process does not allow one to recover the rate of convergence of the risk to 0 in the case when $f_0 \in \mathcal{F}$ (the so called zero error case). To address this problem, we define below a sequence of localized norms of Rademacher process that majorizes the sequence $\{r_k^n\}$ defined above.

Given $\varepsilon > 0$, let $\bar{\varphi}$ be a (random) function defined by

$$\bar{\varphi}(r) := \bar{K}_1 \|R_n\|_{\mathcal{F}\cap B_{2r}^e} + \bar{K}_2\sqrt{r\varepsilon} + \bar{K}_3\varepsilon,$$

where $B_r^e = \{f \in \mathcal{F}\colon P_n f \leq r\}$ and $\bar{K}_1, \bar{K}_2, \bar{K}_3 > 0$ are numerical constants.

We introduce the following data-dependent sequence

$$\{\bar{r}_k^n\}_{k\geq 0} = \{\bar{r}_k^n(X_1,\dots,X_n;\varepsilon_1,\dots,\varepsilon_n)\}_{k\geq 0},$$

$$\bar{r}_0^n = 1,\ \bar{r}_{k+1}^n = \bar{\varphi}(\bar{r}_k^n)\wedge 1, \quad k = 0,1,2,\dots \tag{1.2}$$

Since the function $\bar{\varphi}$ is nondecreasing, a simple induction shows that the sequence $\{\bar{r}_k^n\}$ is nonincreasing.

Theorem 2. *There is a choice of numerical constants* $\bar{K}_1, \bar{K}_2, \bar{K}_3 > 0$ *such that for all* $P \in \mathcal{P}(S)$, *for all* $N \geq 1$ *and for any consistent estimate* $\hat{f}_n$

$$\mathbb{P}\{P\hat{f}_n \geq \bar{r}_N^n\} \leq 2Ne^{-\frac{n\varepsilon}{2}}.$$

Thus, if one chooses $N \geq 1$ and, for a given $\delta > 0$, $\varepsilon > (\log 2N\delta)/n$, then one can define $\hat{\beta}_n(\mathcal{F};\delta) := \bar{r}_N^n$ to get the bound (1.1). The question to be answered is how large should be the number of iterations N to achieve a reasonably good upper bound on the risk in such a way (if it is possible at all). Surprisingly, under rather general conditions the upper bound becomes sharp after very few iterations (roughly, the number of iterations N is of the order $\log_2\log_2(\frac{1}{\varepsilon})$).

In what follows, given a (pseudo)metric space $(M;d)$, we denote $N_d(M;\varepsilon)$ the minimial number of balls of radius ε, covering M, and $H_d(M;\varepsilon) := \log N_d(M;\varepsilon)$. Also, for a probability measure Q on $(S,\mathcal{A})$, $d_{Q,2}$ denotes the metric of the space $L_2(S;dQ)$.

Given a class of functions $\mathcal{F}$, assume that

$$\mathbb{E}_\varepsilon \|n^{-1/2}\sum_{i=1}^n \varepsilon_i \delta_{X_i}\|_{B^e(r)\cap\mathcal{F}} \le \hat{\psi}_n(\sqrt{r})$$

for some concave nondecreasing (random) function $\hat{\psi}_n$. Usually the role of $\hat{\psi}_n$ will be played by the random entropy integral

$$\hat{\psi}_n(r) = K\int_0^r H^{1/2}_{d_{P_n,2}}(\mathcal{F},u)du$$

or by some further upper bound on the random entropy integral. Let us denote by $\hat{\delta}_n := \hat{\delta}_n(X_1,\dots,X_n)$ the solution of the equation

$$\hat{\delta}_n = n^{-1/2}\hat{\psi}_n\big(\sqrt{\hat{\delta}_n}\big).$$

The following theorem gives the upper bound on the quantity $\bar{r}_N^n$.

Theorem 3. *If the number of iteration is equal to* $N = [\log_2\log_2\varepsilon^{-1}]+1$, *then for some numerical constant* $c>0$ *and for all* $P\in\mathcal{P}(S)$

$$\mathbb{P}\left(\bar{r}_N^n \ge c(\hat{\delta}_n\vee\varepsilon)\right) \le ([\log_2\log_2\varepsilon^{-1}]+1)e^{-\frac{n\varepsilon}{2}}.$$

Example 1. Learning a concept from a VC-class. Consider the case of the concept learning, when $\mathcal{F} := \{I_C : C\in\mathcal{C}\}$. Given a sample $(X_1,\dots,X_n)$ with unknown common distribution $P\in\mathcal{P}(S)$, we observe the labels $\{Y_j := I_{C_0}(X_j) : 1\le j\le n\}$ for an unkown target concept $C_0\in\mathcal{C}$. An estimate $\hat{C}_n = \hat{C}_n((X_1,Y_1),\dots,(X_n,Y_n))$ of the target concept C_0 is called consistent iff $I_{\hat{C}_n}(X_j) = Y_j$ for all $j=1,\dots,n$. Let

$$\Delta^{\mathcal{C}}(X_1,\dots,X_n) := \mathrm{card}\big(\{C\cap\{X_1,\dots,X_n\} : C\in\mathcal{C}\}\big).$$

Then

$$\hat{\psi}_n(r) := K(\log\Delta^{\mathcal{C}}(X_1,\dots,X_n))^{1/2}r$$

is an upper bound on the random entropy integral, which yields the value of $\hat{\delta}_n$

$$\hat{\delta}_n = K^2\frac{\log\Delta^{\mathcal{C}}(X_1,\dots,X_n)}{n}.$$

Thus, with the same choice of N we get for some numerical constant $c>0$ the bound

$$\mathbb{P}\left(\bar{r}_N^n \ge c\big(\frac{\log\Delta^{\mathcal{C}}(X_1,\dots,X_n)}{n}\vee\varepsilon\big)\right) \le ([\log_2\log_2\varepsilon^{-1}]+1)e^{-\frac{n\varepsilon}{2}}.$$

Theorem 2 implies at the same time that for any consistent estimate $\hat{C}_n$ we have $P(\hat{C}_n\triangle C_0)\le\bar{r}_N^n$ with probability at least $1-2Ne^{-n\varepsilon/2}$. This shows that for a VC-class of concepts $\mathcal{C}$ with VC-dimension $V(\mathcal{C})$ the local Rademacher norm $\bar{r}_N^n$ (which, according to Theorem 2, is an upper bound on the risk of consistent concepts $\hat{C}_n$) is bounded from above by the quantity $O(V(\mathcal{C})\log n/n)$. Up to a logarithmic factor,

this is the optimal (in a minimax sense) convergence rate of the generalization error to 0 (see, e.g., [3]).

Next we consider the conditions in terms of entropy with bracketing $H_{[\,]}(\mathcal{F},\varepsilon) := \log N_{[\,]}(\mathcal{F},\varepsilon)$. Here $N_{[\,]}(\mathcal{F},\varepsilon)$ denotes the minimal number of "brackets" $[f^-,f^+] := \{f : f^- \le f \le f^+\}$ with $d_{P,2}(f^-,f^+) \le \varepsilon$ (f^-, f^+ being two measurable functions from S into $[0,1]$, such that $f^- \le f^+$). Let

$$\psi_{[\,]}(r) = \int_0^r \left(H_{[\,]}(\mathcal{F},u)+1\right)^{1/2} du.$$

and let $\delta_{[n]} = \delta_{[n]}(P)$ be the solution of the equation

$$\delta_{[n]} = n^{-1/2}\psi_{[\,]}(\sqrt{\delta_{[n]}}).$$

Again, we set for some $\varepsilon > 0$ $N := [\log_2 \log_2 \varepsilon^{-1}] + 1$. Then the following theorem holds.

Theorem 4. *There exists a constant $c > 0$ such that for all $P \in \mathcal{P}(S)$*

$$\mathbb{P}\left(\bar{r}_N^n \ge c(\delta_{[n]}(P) \vee \varepsilon)\right) \le ([\log_2 \log_2 \varepsilon^{-1}] + 1)e^{-\frac{n\varepsilon}{2}}.$$

In particular, if $H_{[\,]}(\mathcal{F};u) = O(u^{-\gamma})$, where $\gamma < 2$, then $\psi_{[\,]}(r) \asymp r^{1-\gamma/2}$ and $\delta_{[n]} \asymp n^{-\frac{2}{2+\gamma}}$.

Example 2. Learning a concept from a d-dimensional cube. Let $S = [0,1]^d$. We consider a problem of estimation of a set (a concept) $C_0 \subset [0,1]^d$, based on the observations (X_j, Y_j), $j = 1,\dots,n$, where X_j, $j = 1,\dots,n$ are i.i.d. points in $[0,1]^d$ with common distribution P and $Y_j := I_{C_0}(X_j)$, $j = 1,\dots,n$. Such a model frequently occurs in the problems of edge estimation in image analysis (see Mammen and Tsybakov (1995)). Assume that the distribution P has a density p such that for some $B > 0$

$$B^{-1} \le p(x) \le B,\ x \in [0,1]^d.$$

Let $\mathcal{C}$ be a class of Borel subsets in $[0,1]^d$ such that $\mathcal{C} \ni C_0$. Let λ be the Lebesgue measure on $[0,1]^d$. Denote $N_I(\mathcal{C};\varepsilon)$ the minimal number of brackets $[C^-,C^+] := \{C : C^- \subset C \subset C^+\}$ with $\lambda(C^+ \setminus C^-) \le \varepsilon$ (C^-, C^+ being two measurable subsets in $[0,1]^d$ such that $C^- \subset C^+$). Let $H_I(\mathcal{C};\varepsilon) := \log N_I(\mathcal{C};\varepsilon)$. This version of entropy with bracketing is often called "entropy with inclusion". We define

$$\psi_I(r) = \int_0^r \left(H_I(\mathcal{C},u)+1\right)^{1/2} du,$$

and let $\delta_n^I = \delta_n^I(P)$ be the solution of the equation

$$\delta_n^I = n^{-1/2}\psi_I(\sqrt{\delta_n^I}).$$

If we have

$$H_I(\mathcal{C};u) = O(u^{-\gamma}),$$

then Theorems 4 easily implies that with some constant $c > 0$

$$\mathbb{P}\left(\bar{r}_N^n \ge c(\delta_n^I \vee \varepsilon)\right) \le ([\log_2 \log_2 \varepsilon^{-1}] + 1)e^{-\frac{n\varepsilon}{2}},$$

where $\delta_n^I \asymp n^{-\frac{1}{1+\gamma}}$. By Theorem 2, for any consistent estimate $\hat{C}_n$ of the set C_0 (i.e. such that $I_{\hat{C}_n}(X_j) = Y_j$, $j = 1,\dots,n$), the quantity $\bar{r}_N^n$ is an upper bound (up to a constant) on $\lambda(\hat{C}_n \triangle C_0)$.

In particular, if $\mathcal{C}$ is the class of sets with α-smooth boundary in $[0,1]^d$, then well known bounds on the bracketing entropy due to Dudley (see e.g. Dudley (1999)) imply that $\gamma = \frac{d-1}{\alpha}$ and $\delta_n^I = n^{-\frac{\alpha}{d-1+\alpha}}$. Similarly, if $\mathcal{C}$ is the class of closed convex subsets of $[0,1]^d$, the rate becomes $\delta_n^I = n^{-\frac{2}{d+1}}$. It was shown by Mammen and Tsybakov (1995) that both rates are optimal in a minimax sense.

The examples above show that the local Rademacher penalties (defined only based on the data and using neither prior information about the underlying distribution, nor the specific properties of the function class) can recover the optimal convergence rates of the estimates in function learning problems.

2. Proofs of the main results

The proofs of the results are based on a version of Talagrand's concentration inequalities for empirical processes, see [11], [12]. The version of the inequalities we are using, with explicit numerical values of the constants involved (that determine the values of the constants in our procedures, such as $\bar{K}_1, \bar{K}_2, \bar{K}_3$ above) are due to Massart (1999). It should be also mentioned that the idea to use Talagrand's concentration inequalities to bound the risk in nonparametric estimation and, especially, in model selection problems goes back to Birgé and Massart (see [2], [1] and references therein).

We formulate now Massart's inequality in a form convenient for our purposes.

Theorem 5. *Let $\mathcal{F}$ be some countable family of real valued measurable functions, such that $\|f\|_\infty \le b < \infty$ for every $f \in \mathcal{F}$. Let Z denote either $\|P_n - P\|_{\mathcal{F}}$ or $\|R_n\|_{\mathcal{F}}$. Let $\sigma^2 = n \sup \mathrm{Var}(f(X_1))$. Then for any positive real number x and $0 < \gamma < 1$*

$$\mathbb{P}(Z \ge (1+\gamma)\mathbb{E}Z + [\sigma\sqrt{2kx} + k(\gamma)bx]/n) \le e^{-x}, \tag{2.1}$$

where k and $k(\gamma)$ can be taken equal to $k = 4$ and $k(\gamma) = 3.5 + 32\gamma^{-1}$. Moreover, one also has

$$\mathbb{P}(Z \le (1-\gamma)\mathbb{E}Z - [\sigma\sqrt{2k'x} - k'(\gamma)bx]/n) \le e^{-x}, \tag{2.2}$$

where $k' = 5.4$ and $k'(\gamma) = 3.5 + 43.2\gamma^{-1}$.

Proof of Theorem 2. Let for any fixed real positive number r

$$\varphi_1(r) = \|P_n - P\|_{\mathcal{F}\cap B(r)}$$

$$\varphi_2(r) = (1+\gamma)\mathbb{E}\|P_n - P\|_{\mathcal{F}\cap B(r)} + 2\sqrt{r\varepsilon} + (1.75 + 16\gamma^{-1})\varepsilon.$$

$$\varphi_3(r) = \frac{2(1+\gamma)}{1-\gamma'}\left[\|R_n\|_{\mathcal{F}\cap B(r)} + \sqrt{5.4r\varepsilon} + (1.75 + 21.6\gamma'^{-1})\varepsilon\right] + 2\sqrt{r\varepsilon} + (1.75 + 16\gamma^{-1})\varepsilon.$$

Then, for any $r > 0$

$$\mathbb{P}\Big(\varphi_1(r) \le \varphi_2(r) \le \varphi_3(r)\Big) \ge 1 - 2e^{-\frac{n\varepsilon}{2}}. \tag{2.3}$$

Indeed, in order to apply inequalities (2.1) and (2.2), we notice that for every $f \in \mathcal{F}\bigcap B(r)$ the sup-norm $\|f\|_\infty \le b = 1$ and

$$\sigma^2 = \sup_{\mathcal{F}\cap B_r} n\mathrm{Var}(f(X)) \le \sup_{\mathcal{F}\cap B(r)} nPf^2 \le \sup_{\mathcal{F}\cap B(r)} nPf \le nr.$$

Moreover, if we set $x = n\varepsilon/2$, then (2.1) implies

$$\mathbb{P}\Big(\|P_n - P\|_{\mathcal{F}\cap B(r)} \geq (1+\gamma)\mathbb{E}\|P_n - P\|_{\mathcal{F}\cap B(r)} + 2\sqrt{r\varepsilon} + (1.75 + 16\gamma^{-1})\varepsilon\Big) \leq e^{-\frac{n\varepsilon}{2}},$$

and (2.2) implies

$$\mathbb{P}\Big(\mathbb{E}\|R_n\|_{\mathcal{F}\cap B(r)} \geq (1-\gamma')^{-1}[\|R_n\|_{\mathcal{F}\cap B(r)} + \sqrt{5.4r\varepsilon} + (1.75 + 21.6\gamma'^{-1})\varepsilon]\Big) \leq e^{-\frac{n\varepsilon}{2}}.$$

Taking into account the symmetrization inequality

$$\mathbb{E}\|P_n - P\|_{\mathcal{F}\cap B(r)} \leq 2\mathbb{E}\|R_n\|_{\mathcal{F}\cap B(r)},$$

we get (2.3).

We set

$$\bar{K}_1 := \frac{2(1+\gamma)}{1-\gamma'},\ \bar{K}_2 := \frac{2\sqrt{5.4}(1+\gamma)}{1-\gamma'} + 2,$$
$$\bar{K}_3 := \frac{2(1+\gamma)}{1-\gamma'}(1.75 + 21.6\gamma'^{-1}) + (1.75 + 16\gamma^{-1}).$$

Let us introduce the following sequence: $\hat{r}_0^n := 1$ and $\hat{r}_{k+1}^n = \varphi_2(\hat{r}_k^n) \wedge 1$ for $k = 0, 1, 2, \ldots$. Since φ_2 is nondecreasing, it's easy to prove by induction that the sequence $\{\hat{r}_k^n\}$ is nonincreasing.

We will also prove by induction that for all $k \geq 0$

$$\mathbb{P}\Big\{r_i^n \leq \hat{r}_i^n \leq \bar{r}_i^n,\ i \leq k\Big\} \geq 1 - 2ke^{-\frac{n\varepsilon}{2}}. \tag{2.4}$$

For $k = 0$ (2.4) is trivial since $r_0^n = \hat{r}_0^n = \bar{r}_0^n = 1$. We proceed by the induction argument. Let us introduce the events

$$\mathcal{A}_k = \{r_i^n \leq \hat{r}_i^n \leq \bar{r}_i^n,\ i \leq k\} \quad \text{and } \mathcal{B}_k = \{\varphi_1(\hat{r}_k^n) \leq \varphi_2(\hat{r}_k^n) \leq \varphi_3(\hat{r}_k^n)\}.$$

To make the induction step, let us assume that we have already proven that

$$\mathbb{P}(\mathcal{A}_k) \geq 1 - 2ke^{-\frac{n\varepsilon}{2}}.$$

Then (2.3) implies

$$\mathbb{P}(\mathcal{B}_k) \geq 1 - 2e^{-\frac{n\varepsilon}{2}}.$$

On the event $\mathcal{A}_k \bigcap \mathcal{B}_k$,

$$\mathcal{F}\cap B(\hat{r}_k^n) \subseteq \mathcal{F}\cap B^e(2\hat{r}_k^n),$$

since for $f \in \mathcal{F}\bigcap B(\hat{r}_k^n)$

$$\begin{aligned} P_n f &\leq Pf + \|P_n - P\|_{\mathcal{F}\cap B(\hat{r}_k^n)} \leq \hat{r}_k^n + \|P_n - P\|_{\mathcal{F}\cap B(\hat{r}_k^n)} \\ &= \hat{r}_k^n + \varphi_1(\hat{r}_k^n) \leq \hat{r}_k^n + \varphi_2(\hat{r}_k^n) = \hat{r}_k^n + \hat{r}_{k+1}^n \leq 2\hat{r}_k^n, \end{aligned}$$

which implies that the inequalities $\varphi_3(\hat{r}_k^n) \leq \bar{\varphi}(\hat{r}_k^n) \leq \bar{\varphi}(\bar{r}_k^n) = \bar{r}_{k+1}^n$ hold. Therefore, on the event $\mathcal{A}_k \bigcap \mathcal{B}_k$,

$$r_{k+1}^n = \varphi_1(r_k^n) \leq \varphi_1(\hat{r}_k^n) \leq \varphi_2(\hat{r}_k^n) = \hat{r}_{k+1}^n \leq \varphi_3(\hat{r}_k^n) \leq \bar{r}_{k+1}^n.$$

So, $\mathcal{A}_k \bigcap \mathcal{B}_k \subseteq \mathcal{A}_{k+1}$, that completes the proof of the induction step

$$\mathbb{P}(\mathcal{A}_{k+1}) \geq 1 - 2(k+1)e^{-\frac{n\varepsilon}{2}}.$$

It follows that

$$\mathbb{P}(r_N^n > \bar{r}_N^n) \leq 2Ne^{-\frac{n\varepsilon}{2}},$$

and since, by Proposition 1, $P\hat{f}_n \le r_N^n$, we conclude that

$$\mathbb{P}\{P\hat{f}_n > \bar{r}_N^n\} \le 2Ne^{-\frac{n\varepsilon}{2}}.$$

Proof of Theorem 3. Let $(\Omega_\varepsilon, \Sigma_\varepsilon, \mathbb{P}_\varepsilon)$ denote the probability space on which the Rademacher sequence $\varepsilon_1, \dots, \varepsilon_n, \dots$ is defined, $\mathbb{E}_\varepsilon$ being the expectation with respect to $\mathbb{P}_\varepsilon$. We introduce the function

$$\varphi_4(r) = \frac{2(1+\gamma)}{1-\gamma'}\left[(1+\gamma''^{-1})\mathbb{E}_\varepsilon\|R_n\|_{\mathcal{F}\cap B^e(2r)} + 2\sqrt{r\varepsilon}\right.$$
$$\left.+(1.75+16\gamma''^{-1})\varepsilon + \sqrt{5.4r\varepsilon} + (1.75+21.6\gamma'^{-1})\varepsilon\right]$$
$$+2\sqrt{r\varepsilon} + (1.75+16\gamma^{-1})\varepsilon, \tag{2.5}$$

where $\gamma'' > 0$. The inequalities (2.1) and (2.2) also hold for the conditional probability $\mathbb{P}_\varepsilon$ and the process $Z = R_n$ with fixed $X_1, \dots, X_n$. Therefore, for any $r > 0$

$$\mathbb{P}_\varepsilon(\bar{\varphi}(r) \le \varphi_4(r)) \ge 1 - e^{-\frac{n\varepsilon}{2}}.$$

Define a sequence

$$\check{r}_0^n = \varphi_4(1),\ \ \check{r}_{k+1}^n = \varphi_4(\check{r}_k^n) \wedge 1, \quad k = 0, 1, 2, \dots$$

By the induction argument, similar to the one we used in the proof of theorem 2, we get

$$\mathbb{P}_\varepsilon\left(\bigcap_{i=1}^N \{\bar{r}_i^n \le \check{r}_i^n\}\right) \ge 1 - Ne^{\frac{n\varepsilon}{2}}.$$

If we prove that $\check{r}_k^n \le a_k$ for a sequence a_k, independent of $\varepsilon_1, \dots, \varepsilon_n$, then the unconditional probability

$$\mathbb{P}\left(\bigcap_{i=1}^N \{\bar{r}_i^n \le a_i\}\right) \ge 1 - Ne^{\frac{n\varepsilon}{2}}.$$

By the assumption we have

$$\mathbb{E}_\varepsilon\|n^{-1}\sum_{i=1}^n \varepsilon_i\delta_{X_i}\|_{B^e(r)\cap\mathcal{F}} \le \hat{\psi}_n(\sqrt{r}). \tag{2.6}$$

Hence, we can choose $c \ge 1$, depending on the parameters $\gamma, \gamma', \gamma''$ in the definition (2.5) of the function φ_4, in such a way that

$$\check{r}_{k+1}^n = \varphi_4(\check{r}_k^n) \le c\left(\varepsilon + (\check{r}_k^n\varepsilon)^{1/2} + n^{-1/2}\hat{\psi}_n\left(\sqrt{\check{r}_k^n}\right)\right).$$

The above inequality implies by induction that the sequence

$$r_0 = 1,\ \ r_{k+1} = c\left(\varepsilon + (r_k\varepsilon)^{1/2} + n^{-1/2}\hat{\psi}_n\left(\sqrt{r_k}\right)\right) \wedge 1,$$

majorizes the sequence $\check{r}_k^n$.

It's clear that in the case when $r_1 < 1$ the sequence r_k is decreasing and it converges to the solution δ of the equation

$$\delta = c\left(\varepsilon + (\delta\varepsilon)^{1/2} + n^{-1/2}\psi\left(\sqrt{\delta}\right)\right).$$

Let us study the behaviour of the difference $d_k := r_k - \delta$. Since the function $\hat{\psi}_n$ is concave, we have

$$\hat{\psi}_n'(\sqrt{\delta}) \le \hat{\psi}_n(\sqrt{\delta})/\sqrt{\delta}.$$

The definition of δ implies that

$$c\left(n^{-1/2}\hat{\psi}_n(\sqrt{\delta})+\sqrt{\delta\varepsilon}\right)\le\delta.$$

Therefore

$$\begin{aligned}d_{k+1}&=r_{k+1}-\delta=c\left(n^{-1/2}\hat{\psi}_n(\sqrt{r_k})-n^{-1/2}\hat{\psi}_n(\sqrt{\delta})+\sqrt{r_k\varepsilon}-\sqrt{\delta\varepsilon}\right)\\&\le c\left(n^{-1/2}\hat{\psi}_n'(\delta)+\sqrt{\varepsilon}\right)\sqrt{r_k-\delta}\le c\left(n^{-1/2}\hat{\psi}_n(\sqrt{\delta})+\sqrt{\delta\varepsilon}\right)/\sqrt{\delta}\sqrt{d_k}\\&\le\sqrt{\delta d_k}.\end{aligned}$$

We have proven that the sequence d_k satisfies the following inequality

$$d_{k+1}\le\sqrt{\delta d_k},\ k\ge 0.$$

Now it's easy to show by induction that

$$d_N\le\delta^{2^{-1}+\ldots+2^{-N}}=\delta^{1-2^{-N}}.$$

Going back to the sequence r_k, we get that

$$r_N=\delta+d_N\le\delta\left(1+\delta^{-2^{-N}}\right).$$

Since the definition of δ implies that $\delta^{-1}<\varepsilon^{-1}$, then the choice of

$$N=\left[\log_2\log_2\varepsilon^{-1}\right]+1$$

guarantees that $\delta^{-2^{-N}}\le 2$ and, hence, $r_N\le(1+2)\delta=3\delta$. What remains to do in order to finish the proof of the theorem, is to bound δ by the maximum of ε and the solution $\hat{\delta}_n$ of the equation $\hat{\delta}_n=n^{-1/2}\hat{\psi}_n(\sqrt{\hat{\delta}_n})$. Actually, we will prove that δ is bounded dy $\delta'':=(3c)^2\delta'$, where $\delta'=\left(\hat{\delta}_n\vee\varepsilon.\right)$ First of all let us notice that the fact that $\hat{\psi}_n$ is concave and $\hat{\psi}_n(0)=0$ implies that for $c\ge 1$ $\hat{\psi}_n(cx)\le c\hat{\psi}_n(x)$. Also note that, since $\delta'\ge\hat{\delta}_n$, the concavity of $\hat{\psi}_n$ and the definition of $\hat{\delta}_n$ imply

$$n^{-1/2}\hat{\psi}_n\left(\sqrt{\delta'}\right)\le\frac{n^{-1/2}\hat{\psi}_n(\sqrt{\hat{\delta}_n})}{\sqrt{\hat{\delta}_n}}\sqrt{\delta'}=\sqrt{\hat{\delta}_n}\sqrt{\delta'}\le\delta'.$$

Combining these properties, we get

$$c\left(\varepsilon+(9c^2\delta'\varepsilon)^{1/2}+n^{-1/2}\hat{\psi}_n\left(3c\sqrt{\delta'}\right)\right)\le c\left(2\sqrt{(3c)^2}\delta'+\delta'\right)\le 9c^2\delta'=\delta''.$$

With necessity it means that $\delta\le\delta''=9c^2(\hat{\delta}_n\vee\varepsilon)$. And, hence, $\bar{r}_N^n\le\delta''\le 27c^2(\hat{\delta}_n\vee\varepsilon)$.

The theorem is proven.

Proof of Theorem 4. In order to bound $\bar{r}_k$, we first construct the bound on $\|R_n\|_{\mathcal{F}\cap B^e(2\bar{r}_k)}$ in terms of $\mathbb{E}\|P_n-P\|_{\mathcal{F}\cap B(\check{r}_k)}$ for properly defined sequence $\check{r}_k$. Afterwards, the expectation can be majorized by the bracketing entropy integral. We will show that the sequence $\check{r}_k$ can be chosen as follows

$$\check{r}_0=1,\ \ \check{r}_{k+1}=\left(\tilde{c}_1\mathbb{E}\|P_n-P\|_{\mathcal{F}\cap B(3\check{r}_k)}+\tilde{c}_2\sqrt{\varepsilon\check{r}_k}+\tilde{c}_3\right)\wedge 1,$$

for some large enough constants $\tilde{c}_1,\tilde{c}_2,\tilde{c}_3>0$. One can argue similarly to the proof of Theorem 3 to show that the following bound holds:

$$\mathbb{P}\Big(\bigcap_{k\le i}\{\bar{r}_k\le\check{r}_k\}\Big)\ge 1-2ie^{-\frac{n\varepsilon}{2}}. \tag{2.7}$$

We will prove even a stronger assertion that for the event

$$\mathcal{A}_i = \bigcap_{k\le i}\Big(\{\bar r_k \le \check r_k\}\cap\{\mathcal{F}\cap B^e(2\bar r_k)\subseteq \mathcal{F}\cap B(3\check r_k)\}\Big)$$

we have

$$\mathbb{P}(\mathcal{A}_i)\ge 1-2ie^{-\frac{n\varepsilon}{2}}. \tag{2.8}$$

Let us choose the constants $c_1', c_2', c_3' > 0$ and $\tilde c_1, \tilde c_2, \tilde c_3 > 0$ in such a way that for the functions

$$\varphi_5(r) = \left(c_1'\|P_n - P\|_{\mathcal{F}\cap B(r)} + c_2'\sqrt{\varepsilon r} + c_3'\varepsilon\right)$$

and

$$\varphi_6(r) = \left(\tilde c_1\mathbb{E}\|P_n - P\|_{\mathcal{F}\cap B(r)} + \tilde c_2\sqrt{\varepsilon r} + \tilde c_3\varepsilon\right),$$

the inequalities of Massart (see Theorem 5) would imply that for any fixed $r > 0$

$$\varphi_3(r)\le\varphi_5(r)\le\varphi_6(r)$$

with probability at least $1-2e^{-\frac{n\varepsilon}{2}}$ (the function φ_3 was defined in the proof of Theorem 2). Clearly, we have $\check r_{k+1} = \varphi_6(\check r_k)\wedge 1$.

First observe that (2.8) holds for $i = 0$ (since $\bar r_0 = \check r_0 = 1$). Define

$$\mathcal{B}_i := \{\varphi_3(3\check r_i)\le\varphi_5(3\check r_i)\le\varphi_6(\check r_i)\}.$$

Then

$$\mathbb{P}(\mathcal{B}_i)\ge 1-2e^{-\frac{n\varepsilon}{2}}.$$

To make an induction step, we first of all notice that on the event $\mathcal{A}_i\cap\mathcal{B}_i$, we have

$$\bar r_{i+1} = \bar\varphi(\bar r_i)\wedge 1\le\varphi_3(3\check r_i)\wedge 1\le\varphi_5(3\check r_i)\wedge 1\le\varphi_6(3\check r_i)\wedge 1 = \check r_{i+1}.$$

Also, on the event $\mathcal{A}_i\cap\mathcal{B}_i$, we have $\mathcal{F}\cap B^e(2\bar r_{i+1})\subseteq\mathcal{F}\cap B(3\check r_{i+1})$. Indeed, if $f\in\mathcal{F}\cap B^e(2\bar r_{i+1})$, then

$$\begin{aligned}
Pf &\le 2\bar r_{i+1} + \|P_n - P\|_{\mathcal{F}\cap B^e(2\bar r_{i+1})}\le 2\bar r_{i+1} + \|P_n - P\|_{\mathcal{F}\cap B^e(2\bar r_i)}\\
&\le 2\bar r_{i+1} + \|P_n - P\|_{\mathcal{F}\cap B(3\check r_i)}\le 2\bar r_{i+1} + \varphi_5(3\check r_i)\wedge 1\\
&\le 2\bar r_{i+1} + \varphi_6(3\check r_i)\wedge 1 = 2\bar r_{i+1} + \check r_{i+1}\le 3\check r_{i+1}
\end{aligned}$$

(to show that $\|P_n - P\|_{\mathcal{F}\cap B(3\check r_i)}\le\varphi_5(3\check r_i)\wedge 1$ we used the fact that the costant c_1' in the definition of φ_5 is larger than 1). Thus, $\mathcal{A}_i\cap\mathcal{B}_i\subset\mathcal{A}_{i+1}$ and

$$\mathbb{P}(\mathcal{A}_{i+1})\ge 1-2(i+1)e^{-\frac{n\varepsilon}{2}}.$$

The proof of the induction step and of the bounds (2.8) and (2.7) is complete.

To finish the proof of the theorem one has to bound $\mathbb{E}\|P_n - P\|_{\mathcal{F}\cap B(r)}$. Since for all $g\in\mathcal{F}\bigcap B(r)$ we have $\|g\|_{P,2}\le(Pg)^{1/2}\le\sqrt{r}$ and $|g|\le 1$ then by Theorem 2.14.2 in [13]

$$\mathbb{E}\|P_n - P\|_{\mathcal{F}\cap B(r)}\le c\left(n^{-1/2}\psi_{[\,]}\left(\sqrt{r}\right) + I\{1 > \sqrt{n}a(\sqrt{r})\}\right),$$

where

$$a(\sqrt{r}) = \sqrt{r}/\sqrt{1 + H_{[\,]}(\mathcal{F},\sqrt{r})}.$$

We can assume that $\check{r}_N \geq \delta_{[n]}$, otherwise, bound (2.7) immediately implies the assertion of the theorem. Therefore, $\check{r}_k \geq \delta_{[n]}$ for all $k \leq N$, which implies that $1 \leq \sqrt{n}a(\sqrt{3\check{r}_k})$. Indeed, using concavity of $\psi_{[\,]}$ and the definition of $\delta_{[n]}$, we have

$$\frac{\psi_{[\,]}(\sqrt{3\check{r}_k})}{\sqrt{3\check{r}_k}} \leq \frac{\psi_{[\,]}(\sqrt{\delta_{[n]}})}{\sqrt{\delta_{[n]}}} = \sqrt{n}\sqrt{\delta_{[n]}} \leq \sqrt{n}\sqrt{3\check{r}_k},$$

which implies

$$3\check{r}_k \geq n^{-1/2}\psi_{[\,]}\left(\sqrt{3\check{r}_k}\right) \geq n^{-1/2}\left(3\check{r}_k\right)^{1/2}\left(1 + H_{[\,]}\left(\mathcal{F}, \sqrt{3\check{r}_k}\right)\right)^{1/2}.$$

Hence, $1 \leq \sqrt{n}a(\sqrt{3\check{r}_k})$ and

$$\mathbb{E}\|P_n - P\|_{\mathcal{F}\cap B(3\check{r}_k)} \leq cn^{-1/2}\psi_{[\,]}\left(\sqrt{3\check{r}_k}\right).$$

Finally, with some constant $c > 0$

$$\check{r}_{k+1} \leq c\left(n^{-1/2}\psi_{[\,]}(\sqrt{3\check{r}_k}) + \varepsilon + \sqrt{\varepsilon\check{r}_k}\right).$$

The proof can be completed by the argument we used in Theorem 3.

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] 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.

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

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

[5] Hush, D. and Scovel, C. (1999) Posterior Performance Bounds for Machine Learning. Preprint, Los Alamos National Laboratory.

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

[7] Koltchinskii, V., Abdallah, C.T., Ariola, M., Dorato, P. and Panchenko, D. (1999) Statistical Learning Control of Uncertain Systems: It is better than it seems. Preprint, UNM.

[8] Lozano, F. (1999) Model Selection Using Rademacher Penalization. Preprint.

[9] Massart, P. (1998) About the constants in Talagrand's concentration inequalities for empirical processes. Preprint, Université Paris-Sud.

[10] Mammen, E. and Tsybakov, A. (1995) Asymptotical minimax recovery of sets with smooth boundaries. *Ann. Statist.*, 23, 502–524.

[11] Talagrand, M. A new look at independence. *Ann. Probab.* 24, 1-34.

[12] Talagrand, M. New concentration inequalities in product spaces *Invent. Math.* 126, 505-563.

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

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

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