Ann. I. H. Poincaré – PR 39, 6 (2003) 943–978
© 2003 Éditions scientifiques et médicales Elsevier SAS. All rights reserved
10.1016/S0246-0203(03)00023-2/FLA

BOUNDS ON MARGIN DISTRIBUTIONS IN LEARNING PROBLEMS

Vladimir KOLTCHINSKII [1]

Department of Mathematics and Statistics, The University of New Mexico, Albuquerque, NM 87131-1141, USA

Received 5 November 2001, accepted 18 February 2003

ABSTRACT. – Let $(S, \mathcal{A}, P)$ be a probability space and let P_n be the empirical measure based on i.i.d. sample $(X_1, \ldots, X_n)$ from P. Let $\mathcal{F}$ be a class of measurable real valued functions on $(S, \mathcal{A})$. For $f \in \mathcal{F}$, define $F_f(t) := P\{f \leqslant t\}$ and $F_{n,f}(t) := P_n\{f \leqslant t\}$. Given $\gamma \in (0, 1]$, define $\varepsilon_{n,\gamma}(\delta) := 1/(n^{1-\gamma/2}\delta^{\gamma})$. We show that if the $L_2(P_n)$-entropy of the class $\mathcal{F}$ grows as $\varepsilon^{-\alpha}$ for some $\alpha \in (0, 2)$, then, for all $f \in \mathcal{F}$ and all $\delta \in (0, \Delta_n)$, $\Delta_n = \mathrm{O}(n^{1/2})$,

$$F_f\left(\frac{\delta}{c(\sigma)}\right) \leqslant c(\sigma)\left[F_{n,f}(\delta) \vee \frac{1}{\sigma}\varepsilon_{n,\gamma}(\delta)\right]$$

and

$$F_{n,f}\left(\frac{\delta}{c(\sigma)}\right) \leqslant c(\sigma)\left[F_f(\delta) \vee \frac{1}{\sigma}\varepsilon_{n,\gamma}(\delta)\right],$$

where $\gamma = \frac{2\alpha}{2+\alpha}$ and $c(\sigma) \downarrow 1$ as $\sigma \downarrow 0$ (the above inequalities hold for any fixed $\sigma \in (0, 1]$ with a high probability). Also, define

$$\delta_n(\gamma; f) := \sup\{\delta\colon\ F_f(\delta) \leqslant \varepsilon_{n,\gamma}(\delta)\} \quad \text{and} \quad \hat{\delta}_n(\gamma; f) := \sup\{\delta\colon\ F_{n,f}(\delta) \leqslant \varepsilon_{n,\gamma}(\delta)\}.$$

Then for all $\gamma > \frac{2\alpha}{2+\alpha}$

$$\frac{\hat{\delta}_n(\gamma; f)}{\delta_n(\gamma; f)} \to 1 \quad \text{as } n \to \infty$$

uniformly in $\mathcal{F}$ and with probability 1 (for $\gamma = \frac{2\alpha}{2+\alpha}$ the above ratio is bounded away from 0 and from ∞). The results are motivated by recent developments in machine learning, where they are used to bound the generalization error of learning algorithms. We also prove some more general results of similar nature, show the sharpness of the conditions and discuss the applications in learning theory.

RÉSUMÉ. – Soient $(S, \mathcal{A}, P)$ un espace probabilisé et P_n la mesure empirique supportée par l'échantillon $(X_1, \ldots, X_n)$ de n variables aléatoires i.i.d. tirées selon P. Soit $\mathcal{F}$ une classe de

E-mail address: vlad@math.unm.edu (V. Koltchinskii).

[1] Partially supported by NSA Grant MDA904-02-1-0075.

fonctions à valeurs réelles, mesurables sur $(S,\mathcal{A})$. Pour $f\in\mathcal{F}$, notons $F_f(t):=P\{f\leqslant t\}$ et $F_{n,f}(t):=P_n\{f\leqslant t\}$. Etant donné $\gamma\in(0,1]$, définissons $\varepsilon_{n,\gamma}(\delta):=1/(n^{1-\gamma/2}\delta^\gamma)$. Nous montrons que si la $L_2(P_n)$-entropie de la classe $\mathcal{F}$ croit en $\varepsilon^{-\alpha}$ avec $\alpha\in(0,2)$, alors, pour toute fonction $f\in\mathcal{F}$ et tout réel $\delta\in(0,\Delta_n)$, $\Delta_n=\mathrm{O}(n^{1/2})$,

$$F_f\left(\frac{\delta}{c(\sigma)}\right)\leqslant c(\sigma)\left[F_{n,f}(\delta)\vee\frac{1}{\sigma}\varepsilon_{n,\gamma}(\delta)\right]$$

et

$$F_{n,f}\left(\frac{\delta}{c(\sigma)}\right)\leqslant c(\sigma)\left[F_f(\delta)\vee\frac{1}{\sigma}\varepsilon_{n,\gamma}(\delta)\right]$$

où $\gamma=\frac{2\alpha}{2+\alpha}$ et $c(\sigma)\downarrow 1$ quand $\sigma\downarrow 0$ (les inégalités ci-dessus sont valides avec forte probabilité pour tout $\sigma\in(0,1]$). De plus, si l'on pose

$$\delta_n(\gamma;f):=\sup\{\delta\colon\ F_f(\delta)\leqslant\varepsilon_{n,\gamma}(\delta)\}\quad\text{et}\quad\hat{\delta}_n(\gamma;f):=\sup\{\delta\colon\ F_{n,f}(\delta)\leqslant\varepsilon_{n,\gamma}(\delta)\},$$

alors pour tout réel $\gamma>\frac{2\alpha}{2+\alpha}$

$$\frac{\hat{\delta}_n(\gamma;f)}{\delta_n(\gamma;f)}\to 1\quad\text{quand } n\to\infty$$

uniformément sur $\mathcal{F}$ et avec probabilité 1 (pour $\gamma=\frac{2\alpha}{2+\alpha}$, le rapport ci-dessus reste strictement positif et borné). Ces résultats sont motivés par des développements récents en apprentissage automatique où ils sont utilisés pour borner l'erreur en généralisation des algorithmes d'apprentissage. De plus, nous prouvons d'autres résultats généraux du même genre, nous montrons que les conditions imposées sont précises et nous discutons de possibles applications en théorie de l'apprentissage.

1. Introduction

Consider a measurable space $(S,\mathcal{A})$ and let $\mathcal{F}$ be a class of real valued measurable functions on $(S,\mathcal{A})$. Let $\{X_n\}$ be a sequence of i.i.d. random variables, defined on a probability space $(\Omega,\Sigma,\mathbb{P})$ and taking values in $(S,\mathcal{A})$ with common distribution P. In what follows, P_n denote the empirical measure based on the sample $(X_1,\ldots,X_n)$:

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

Given a real valued measurable function f on $(S,\mathcal{A})$, let

$$F_f(\delta):=P\{f\leqslant\delta\},\qquad F_{n,f}(\delta):=P_n\{f\leqslant\delta\}.$$

In this paper, we prove upper and lower bounds on F_f in terms of $F_{n,f}$ uniformly in $f\in\mathcal{F}$ under suitable conditions on the metric entropy of the class $\mathcal{F}$.

It is well known that, even if $\mathcal{F}$ is a P-Donsker class, the class of sets $\mathcal{C} := \{\{f \leqslant t\}: f \in \mathcal{F},\ t \in \mathbb{R}\}$ does not have to be P-Glivenko–Cantelli, i.e., the supremum

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

does not necessarily converge to 0. Koltchinskii and Panchenko [9] studied the convergence of $F_{n,f}$ to F_f in Lévy distance (uniformly over $\mathcal{F}$) and proved that this convergence is equivalent to $\mathcal{F}$ being a P-Glivenko–Cantelli class with the rate of convergence depending on the complexity of the class. The Lévy distance measures the closeness of two distribution functions not at the same point, but at two different points (close to each other): if the Lévy distance between $F_{n,f}$ and F_f is smaller than ε, then for all t

$$F_f(t) \leqslant F_{n,f}(t+\varepsilon) + \varepsilon \quad \text{and} \quad F_{n,f}(t) \leqslant F_f(t+\varepsilon) + \varepsilon.$$

However, the closeness of the distributions in Lévy distance tells almost nothing about bounding $F_f(\delta)$ in terms of $F_{n,f}$ for those $f \in \mathcal{F}$ and $\delta > 0$ for which $F_{n,f}(\delta)$ is small, so, one should try to control *the ratio* of F_f and $F_{n,f}$ rather than their *difference*. There exists an important circle of problems in learning theory (related to the development of so-called *large margin classification algorithms*, see the discussion below) where this question is crucial since the large margin algorithms tend to output functions f for which $F_{n,f}(\delta)$ remains small for large enough values of δ. In such cases, it is rather natural to measure the closeness of $F_{n,f}$ to F_f in a different way (that can be viewed as a "multiplicative" version of Lévy distance). Namely, it is important to know that for $c > 1$ that is sufficiently close to 1 with a high probability

$$F_f(\delta) \leqslant cF_{n,f}(c\delta) \quad \text{and} \quad F_{n,f}(\delta) \leqslant cF_f(c\delta)$$

for all $f \in \mathcal{F}$ and for all δ in a broad enough interval (unfortunately, it is impossible to have this type of bounds for all δ). To prove these bounds will be our goal and we give below more precise description of the main results.

Note that Koltchinskii and Panchenko [9] dealt with a problem of bounding $F_f(0)$ by an expression involving $cF_{n,f}(\delta)$ (bounding the generalization error in the context of learning theory) which can be viewed as a special case of the above problem; the constant c involved in their bounds was large. It might be also of some interest to obtain bounds of the above type that take into account both the translations and the dilations of the real line (i.e., combine the closeness in standard "additive" and in "multiplicative" Lévy distances), but there seems to be no obvious application of such more general bounds at the moment.

For each $\gamma \in (0,1]$, define

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

In particular, we show that if the $L_2(P_n)$-entropy of the class $\mathcal{F}$ grows as $\varepsilon^{-\alpha}$ with some $\alpha \in (0,2)$, then for $\gamma \geqslant \frac{2\alpha}{2+\alpha}$

$$\mathbb{P}\biggl\{\exists f \in \mathcal{F}\ \exists \delta \leqslant \frac{n^{1/2}}{t^{1/\gamma}}\colon\ F_f\biggl(\frac{\delta}{c(\sigma)}\biggr) \geqslant c(\sigma)\biggl[F_{n,f}(\delta) \vee \frac{1}{\sigma}\varepsilon_{n,\gamma}(\delta)\biggr]\biggr\} \leqslant A(\sigma)\exp\{-\theta t\} \tag{1.1}$$

and

$$\mathbb{P}\biggl\{\exists f \in \mathcal{F}\ \exists \delta \leqslant \frac{n^{1/2}}{t^{1/\gamma}}\colon\ F_{n,f}\biggl(\frac{\delta}{c(\sigma)}\biggr) \geqslant c(\sigma)\biggl[F_f(\delta) \vee \frac{1}{\sigma}\varepsilon_{n,\gamma}(\delta)\biggr]\biggr\} \leqslant A(\sigma)\exp\{-\theta t\}, \tag{1.2}$$

where $A(\sigma) < +\infty,\ \sigma \in (0,1]$ and $c(\sigma) \downarrow 1$ as $\sigma \downarrow 0$ (the bounds hold for all $t > 0$ and $\sigma \in (0,1];\ \theta$ is a numerical constant).

Let now

$$\delta_n(\gamma; f) := \sup\{\delta\colon\ F_f(\delta) \leqslant \varepsilon_{n,\gamma}(\delta)\} \tag{1.3}$$

and

$$\hat{\delta}_n(\gamma; f) := \sup\{\delta\colon\ F_{n,f}(\delta) \leqslant \varepsilon_{n,\gamma}(\delta)\}. \tag{1.4}$$

These quantities provide the size of δ for which $F_{n,f}(\delta)$ or $F_f(\delta)$ in bounds (1.1), (1.2) start exceeding the term $\varepsilon_{n,\gamma}(\delta)$ related to the behavior of the entropy of the class $\mathcal{F}$. Thus, for all $\delta \geqslant \hat{\delta}_n(\gamma; f),\ F_f(\delta/c)$ is bounded from above by $cF_{n,f}(\delta)$ and, for all $\delta \geqslant \delta_n(\gamma; f),\ F_{n,f}(\delta/c)$ is bounded from above by $cF_f(\delta)$ (with high probability and with some $c > 1$). We also show that, for $\gamma > \frac{2\alpha}{2+\alpha}$, such bounds hold for all $c > 1$ (for all large enough n).

We also study the asymptotic behavior of the ratios $\hat{\delta}_n(\gamma; f)/\delta_n(\gamma; f)$ uniformly in $f \in \mathcal{F}$. We show that, for all $\gamma > \frac{2\alpha}{2+\alpha}$ (where α is again the exponent of the entropy), the ratios converge to 1 uniformly in $\mathcal{F}$ with probability 1. For $\gamma = \frac{2\alpha}{2+\alpha}$, the ratios are known to be bounded away both from 0 and from infinity uniformly in $\mathcal{F}$ with probability 1 (see Koltchinskii and Panchenko [9]). We give examples showing that for $\gamma < \frac{2\alpha}{2+\alpha}$, the ratios can tend to 0 or to infinity and also showing the optimality of the bounds (1.1), (1.2).

The proofs of the main results are based on Talagrand's concentration inequalities for empirical processes (see Talagrand [17,18] and also Massart [14], where the inequalities are given in the form we are using them here) along with entropy type bounds (see Dudley [5], van der Vaart and Wellner [20]). The method is close to the one used in Koltchinskii and Panchenko [8,9]. It is based on iterative localization of complexity of function classes with application of the concentration inequalities and the entropy bounds at each iteration. The method can be of independent interest in the problems related to bounding the ratios of empirical distributions to true distributions [6] as well as in nonparametric statistics (see Massart [15] for some close ideas).

Since the class $\mathcal{I} \circ \mathcal{F} := \{I_{(-\infty,t]} \circ f\colon\ f \in \mathcal{F},\ t \in \mathbb{R}\}$ can have large complexity (e.g., it is not necessarily Donsker) and it is hard to relate its entropy directly to the entropy of the class $\mathcal{F}$, we had to approximate this class by the classes $\Phi \circ \mathcal{F} := \{\varphi \circ f\colon\ f \in \mathcal{F},\ \varphi \in \Phi\}$ with a properly choosen family of Lipschitz functions Φ approximating the indicators of the intervals $(-\infty, t]$. Such a smoothing allows us then to estimate the entropy of

$\Phi \circ \mathcal{F}$ in terms of the entropy of $\mathcal{F}$, but the "price" of this approximation is the need to compare F_f and $F_{n,f}$ at different points.

The problems of this nature are motivated by some recent developments in machine learning. More precisely, we deal with so-called *binary classification* problem, described below (see also Devroye, Györfi and Lugosi [4]). Suppose that the space S is replaced by $S \times \{-1, 1\}$. Functions f in the previous definitions will be now replaced by $(x, y) \mapsto yf(x)$. In a couple $(x, y) \in S \times \{-1, 1\}$, x is interpreted as an "instance" and y as a "label" assigned to this instance (we consider binary classification only for simplicity, all the results apply also to multiclass problems the same way as it is done in Koltchinskii and Panchenko [9]). Let (X, Y) be a random couple in $S \times \{-1, 1\}$ with unknown distribution P. It is supposed that the instance X is observable, but the label Y is not, and it is to be predicted based on the observation of X. We will call a function $f : S \mapsto \mathbb{R}$ *a classifier*. A classifier f predicts the label $+1$ if $f(X) > 0$ and the label -1 if $f(X) < 0$ (if $f(X) = 0$, f does not return any label). With this conventions, the probability that f either misclassifies, or does not return the label, is $P\{(x, y): yf(x) \leqslant 0\}$. In machine learning literature, this quantity is referred to as *generalization error*. The goal of learning is to find a classifier (in a given class $\mathcal{F}$) with a small generalization error. Since P is unknown, it is replaced by the empirical distribution P_n based on n i.i.d. training examples $(X_1, Y_1), \ldots, (X_n, Y_n)$ (independent copies of (X, Y)). An important problem is to develop sharp probabilistic bounds on the generalization error of classifiers $\hat{f} \in \mathcal{F}$ based on the training data. The quantity $Yf(X)$ is often called *classification margin* of f. Correspondingly,

$$F_f(t) := P\{(x, y): yf(x) \leqslant t\}$$

is called the *margin distribution* of f and

$$F_{n,f}(t) := P_n\{(x, y): yf(x) \leqslant t\}$$

is called the *empirical margin distribution* (clearly, the generalization error is equal to $F_f(0)$ and *the training error* is $F_{n,f}(0)$). There has been a lot of attention to so called *large margin classification methods* (voting methods, support vector machines) in which learning algorithms output classifiers with the empirical margin distribution that is shifted to the right so that often $F_{n,f}(t) = 0$ for positive (large enough) values of t. The algorithms search for classifiers of this type in rather large function classes $\mathcal{F}$ that often consist of "combinations" of functions from a simpler base class $\mathcal{H}$ (the "combinations" are convex combinations in the case of such methods as boosting, or they might be implemented by large neural networks, etc.) The success of this type of methods has not been understood to the end yet, but it is clear that it has something to do with their ability to produce classifiers with large margin. We refer to Vapnik [21], Anthony and Bartlett [1], Cortes and Vapnik [3], Bartlett [2], Schapire et al. [16], Koltchinskii and Panchenko [9] and references therein for the discussion of various aspects of this problem.

One of the important results in this area is due to Schapire et al. [16] (see also Bartlett [2] who proved similar results in the context of neural network learning and Koltchinskii and Panchenko [9], Koltchinskii, Panchenko and Lozano [11] who refined and generalized these results using the methods of Gaussian and empirical processes). Schapire et al. [16] considered the class $\mathcal{F} := \mathrm{conv}(\mathcal{H})$, where $\mathcal{H}$ is a Vapnik–Chervonenkis class with VC-dimension $V(\mathcal{H})$ and showed that for a given $\alpha \in (0,1)$ with probability at least $1-\alpha$ for all $f \in \mathrm{conv}(\mathcal{H})$

$$F_f(0) \leqslant \inf_{\delta}\left[F_{n,f}(\delta) + \frac{C}{\sqrt{n}}\left(\frac{V(\mathcal{H})\log^2(n/V(\mathcal{H}))}{\delta^2} + \log\left(\frac{1}{\alpha}\right)\right)^{1/2}\right].$$

Let $\hat{\delta}(f)$ denote the solution of the equation $\delta F_{n,f}(\delta) = \sqrt{V(\mathcal{H})/n}$. Plugging in the above bound $\delta = \hat{\delta}(f)$, one gets (up to logarithmic factors) the generalization error of a classifier f from the convex hull of the order $\mathrm{O}((1/\hat{\delta}(f))\sqrt{V(\mathcal{H})/n})$. Boosting and other large margin classification methods tend to produce classifiers with large value of $\hat{\delta}(f)$, so the above bound provides a partial explanation of their success.

The quantities $\delta_n(\gamma; f)$ and $\hat{\delta}_n(\gamma; f)$ were introduced by Koltchinskii and Panchenko [9]. They were called *the γ-margin* and *the empirical γ-margin*, respectively, and they can be used to bound the generalization error of large margin classifiers. Indeed, define

$$\varepsilon_n(\gamma; f) := \varepsilon_{n,\gamma}\big(\delta_n(\gamma; f)\big) \quad \text{and} \quad \hat{\varepsilon}_n(\gamma; f) := \varepsilon_{n,\gamma}\big(\hat{\delta}_n(\gamma; f)\big). \tag{1.5}$$

We clearly have

$$\begin{aligned} &\varepsilon_n(\gamma; f) \in \big[F_f\big(\delta_n(\gamma; f) - 0\big), F_f\big(\delta_n(\gamma; f)\big)\big], \\ &\hat{\varepsilon}_n(\gamma; f) \in \big[F_{n,f}\big(\hat{\delta}_n(\gamma; f) - 0\big), F_{n,f}\big(\hat{\delta}_n(\gamma; f)\big)\big]. \end{aligned} \tag{1.6}$$

Then, by the bounds (1.1), (1.2) on F_f, one gets that with high probability the generalization error $F_f(0)$ is bounded by $c\hat{\varepsilon}_n(\gamma; f)$ for all $f \in \mathcal{F}$, where $\gamma \geqslant \frac{2\alpha}{2+\alpha}$ (if $\gamma > \frac{2\alpha}{2+\alpha}$, then the result is true with any $c > 1$ for all large enough n).

The closeness of the ratios of γ-margins to 1 (which is equivalent to the closeness of the ratio $\hat{\varepsilon}_n(\gamma, f)/\varepsilon_n(\gamma, f)$ to 1 and which allows one to use $\hat{\varepsilon}_n(\gamma; f)$ as an estimate of $\varepsilon_n(\gamma; f)$) was first observed in the experiments of Koltchinskii, Panchenko and Lozano [10–12] in the case of classifiers f produced by a well known learning algorithm AdaBoost. On the other hand, it was proved by Koltchinskii and Panchenko [9] (and it follows from Theorem 1 below) that for all $\gamma \geqslant \frac{2\alpha}{2+\alpha}$ we have

$$\mathbb{P}\left\{\forall f \in \mathcal{F}\colon\ A^{-1} \leqslant \frac{\hat{\delta}_n(\gamma; f)}{\delta_n(\gamma; f)} \leqslant A\right\} \geqslant 1 - B\log_2\log_2 n \exp\left\{-\frac{n^{\gamma/2}}{2}\right\}$$

(with some constants $A, B > 0$).

It is easy to see that the quantity

$$\hat{\varepsilon}_n(\gamma; f) = \frac{1}{n^{1-\gamma/2}\hat{\delta}_n(\gamma; f)^{\gamma}} \tag{1.7}$$

involved in the upper bound on the generalization error *becomes smaller* as γ decreases from 1 to 0. The Schapire–Freund–Bartlett–Lee type of bounds correspond to the worst choice of γ ($\gamma = 1$). In the case when $\mathcal{F}$ is the symmetric convex hull of a VC-class $\mathcal{H}$ with VC-dimension $V(\mathcal{H})$ the value of α is equal to $2(V(\mathcal{H}) - 1)/V(\mathcal{H}) < 2$ that allows us to have $\gamma < 1$, improving the previously known bound. In fact, Koltchinskii, Panchenko and Lozano [10–12] computed the empirical γ-margins of classifiers obtained in consecutive rounds of boosting and observed that the bounds on their generalization error in terms of γ-margins hold even for much smaller values of γ, which leads to a conjecture that such classifiers belong, in fact, to a class $\mathcal{F} \subset \mathrm{conv}(\mathcal{H})$ of a smaller entropy than the entropy of the whole convex hull. Koltchinskii, Panchenko and Lozano [11,12] consider the problem of adapting margin type quantities to the complexity of the classifier. Recently, Kégl, Linder and Lugosi [7] suggested some other interesting margin type bounds on generalization error in which the shattering dimension of the class is used instead of its $L_2(P_n)$-entropy.

Our main focus in this paper is bounding not only the generalization error, but also the true margin distribution function F_f. This might be essential in the development of large margin classification methods since in many cases the goal may be to find a classifier $f \in \mathcal{F}$ that not only has a small generalization error, but also has a large true margin (i.e., such that $F_f(\delta)$ remains small for large enough values of the margin $\delta > 0$). Recent work of Tsybakov [19] shows that if $\eta(x) := \mathbb{E}(Y|x = x)$ (this regression function defines the optimal Bayes classifier), then the convergence rates of empirical risk minimizers to the Bayes risk crucially depend on the behavior of the distribution function of $|\eta|$. Estimation of this margin type distribution function might be an important step in the development of adaptive classification algorithms (for which optimal convergence rates to the Bayes risk are attained), and bounding the true margin distribution by the empirical one might be very useful in the analysis of such methods.

In the current paper, we attempt to address these problems and we get the bounds outlined at the beginning of the Introduction, but under more general assumptions on the entropy of the class $\mathcal{F}$. Our results also clarify the meaning of γ-margins and give a mathematical explanation of some of their intriguing properties observed earlier in the experiments (such as the closeness to 1 of the ratio of the empirical γ-margin to the true γ-margin).

2. Main results

In this section we introduce some more general margin type quantities whose behavior is related to the growth of the entropy of the class $\mathcal{F}$.

Given a metric space (T, d), let $H_d(T; \varepsilon)$ be the ε-entropy of T with respect to d, i.e.,

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

where $N_d(T; \varepsilon)$ is the minimal number of balls of radius ε covering T. For a probability measure Q on $(S; \mathcal{A})$, $d_{Q,2}$ will denote the metric of the space $L_2(S; dQ)$: $d_{Q,2}(f; g) := (Q|f - g|^2)^{1/2}$.

Let Ψ be the class of strictly concave nondecreasing functions ψ on $[0,+\infty)$ with $\psi(0)=0$ and such that

$$\frac{\psi(x)}{x}\to 0 \quad \text{as } x\to\infty, \qquad \frac{\psi(x)}{x}\to +\infty \quad \text{as } x\to 0.$$

Ψ_0 will denote the class of functions $\psi\in\Psi$ such that

$$\psi(xy)\leqslant \psi(x)\psi(y), \quad x,y\geqslant 0.$$

Suppose the following bound on Dudley's entropy integral holds with some $D_n = D_n(X_1,\dots,X_n)>0$ such that $\mathbb{E}D_n<\infty$ and with $\psi\in\Psi$:

$$\int\limits_0^x H^{1/2}_{d_{P_n,2}}(\mathcal{F},u)\,du \leqslant D_n\psi(x), \quad x>0 \text{ a.s.} \tag{2.1}$$

The function ψ characterizes the complexity of the class $\mathcal{F}$ and it will be involved in the definition of ψ-bounds and ψ-margins below.

Assuming that $\psi\in\Psi$ and given $\varepsilon>0$, denote by $\delta_n^\psi(\varepsilon)$ the solution of the equation

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

with respect to δ. Similarly, for a fixed $\delta>0$, $\varepsilon_n^\psi(\delta)$ denotes the solution of (2.2) with respect to ε (since ψ is strictly concave, the solutions are unique in both cases). Concavity of ψ implies that the function $\varphi(x):=\frac{\psi(x)}{x}$ is nonincreasing (also, since $\psi\in\Psi$, $\varphi((0,+\infty))=(0,+\infty)$), and we have

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

Given a function f, we define *the ψ-bound* as

$$\varepsilon_n^\psi(f):=\inf\big\{\varepsilon\geqslant 0\colon\ F_f\big(\delta_n^\psi(\varepsilon)\big)\leqslant\varepsilon\big\}$$

and *the empirical ψ-bound* as

$$\hat{\varepsilon}_n^\psi(f):=\inf\big\{\varepsilon\geqslant 0\colon\ F_{n,f}\big(\delta_n^\psi(\varepsilon)\big)\leqslant\varepsilon\big\}.$$

The “dual” quantities will be called *the ψ-margins*:

$$\delta_n^\psi(f):=\sup\big\{\delta\geqslant 0\colon\ F_f(\delta)\leqslant\varepsilon_n^\psi(\delta)\big\}$$

and

$$\hat{\delta}_n^\psi(f):=\sup\big\{\delta\geqslant 0\colon\ F_{n,f}(\delta)\leqslant\varepsilon_n^\psi(\delta)\big\}.$$

An easy consequence of these definitions is that

$$\varepsilon_n^{\psi}(f) = \varepsilon_n^{\psi}\bigl(\delta_n^{\psi}(f)\bigr) \quad \text{and} \quad \hat{\varepsilon}_n^{\psi}(f) = \varepsilon_n^{\psi}\bigl(\hat{\delta}_n^{\psi}(f)\bigr). \tag{2.3}$$

Clearly, we also have

$$\begin{aligned} &\varepsilon_n^{\psi}(f) \in \bigl[F_f\bigl(\delta_n^{\psi}(f) - 0\bigr), F_f\bigl(\delta_n^{\psi}(f)\bigr)\bigr] \quad \text{and} \\ &\hat{\varepsilon}_n^{\psi}(f) \in \bigl[F_{n,f}\bigl(\hat{\delta}_n^{\psi}(f) - 0\bigr), F_{n,f}\bigl(\hat{\delta}_n^{\psi}(f)\bigr)\bigr]. \end{aligned} \tag{2.4}$$

In some of the statements and in the proofs below, we will need the truncated versions of these quantities. Namely, given a function f and $t > 0$, we define *the truncated ψ-bound* and *the truncated empirical ψ-bound* as

$$\varepsilon_n^{\psi}(f;t) := \inf\left\{\varepsilon \geqslant \frac{t}{n}\colon\ F_f\bigl(\delta_n^{\psi}(\varepsilon)\bigr) \leqslant \varepsilon\right\} \quad \text{and}$$
$$\hat{\varepsilon}_n^{\psi}(f;t) := \inf\left\{\varepsilon \geqslant \frac{t}{n}\colon\ F_{n,f}\bigl(\delta_n^{\psi}(\varepsilon)\bigr) \leqslant \varepsilon\right\}.$$

The *truncated ψ-margins* are defined as follows:

$$\delta_n^{\psi}(f;t) := \sup\left\{\delta \leqslant \delta_n^{\psi}\left(\frac{t}{n}\right)\colon\ F_f(\delta) \leqslant \varepsilon_n^{\psi}(\delta)\right\} \quad \text{and}$$
$$\hat{\delta}_n^{\psi}(f;t) := \sup\left\{\delta \leqslant \delta_n^{\psi}\left(\frac{t}{n}\right)\colon\ F_{n,f}(\delta) \leqslant \varepsilon_n^{\psi}(\delta)\right\}.$$

The properties similar to (2.3), (2.4) hold in the truncated case as well.

The result below was proved in Koltchinskii, Panchenko and Lozano [12].

THEOREM 1. – *Suppose that the condition* (2.1) *holds. Then there exist absolute constants $A, B > 0$ such that for $\bar{A} := A(1 + \mathbb{E}D_n)^2$ and for all $t \geqslant 2\log n$*

$$\begin{aligned} &\mathbb{P}\bigl\{\forall f \in \mathcal{F}\colon\ \bar{A}^{-1}\hat{\varepsilon}_n^{\psi}(f;t) \leqslant \varepsilon_n^{\psi}(f;t) \leqslant \bar{A}\hat{\varepsilon}_n^{\psi}(f;t)\bigr\} \\ &\quad \geqslant 1 - B\log_2\log_2\frac{n}{t}\exp\left\{-\frac{t}{2}\right\}. \end{aligned} \tag{2.5}$$

The next theorem and its corollary describes the asymptotic behavior of the ratios of ψ-bounds and ψ-margins.

THEOREM 2. – *Suppose the condition* (2.1) *holds with some $\psi \in \Psi$ such that*

$$\psi(x) \geqslant 2\sqrt{2}x\sqrt{\log\frac{\mathrm{e}}{x}}, \quad x \leqslant 1,$$

and with

$$\sup_n \mathbb{E}D_n < +\infty.$$

Suppose also that

$$\sup_{f \in \mathcal{F}} P\{f \geqslant u\} \to 0 \quad \textit{as } u \to \infty. \tag{2.6}$$

Then

$$\mathbb{P}\left\{0 < \liminf_{n} \inf_{f\in\mathcal{F}} \frac{\hat{\varepsilon}_n^{\psi}(f)}{\varepsilon_n^{\psi}(f)} \leqslant \limsup_{n} \sup_{f\in\mathcal{F}} \frac{\hat{\varepsilon}_n^{\psi}(f)}{\varepsilon_n^{\psi}(f)} < +\infty\right\} = 1. \tag{2.7}$$

Moreover,

$$\mathbb{P}\left\{\sup_{f\in\mathcal{F}} \left|\frac{\hat{\varepsilon}_n^{\phi}(f)}{\varepsilon_n^{\phi}(f)} - 1\right| \to 0 \text{ as } n \to \infty\right\} = 1 \tag{2.8}$$

for any $\phi \in \Psi$ such that

$$\frac{\phi(x)}{\psi(x)} \to +\infty \quad \text{as } x \to 0. \tag{2.9}$$

COROLLARY 1. – *Suppose that the conditions of Theorem* 2 *hold with $\psi \in \Psi_0$. Then*

$$\mathbb{P}\left\{0 < \liminf_{n} \inf_{f\in\mathcal{F}} \frac{\hat{\delta}_n^{\psi}(f)}{\delta_n^{\psi}(f)} \leqslant \limsup_{n} \sup_{f\in\mathcal{F}} \frac{\hat{\delta}_n^{\psi}(f)}{\delta_n^{\psi}(f)} < +\infty\right\} = 1. \tag{2.10}$$

If now $\phi \in \Psi_0,\ \phi(1) = 1$ and

$$\frac{\phi(x)}{\psi(x)} \to +\infty \quad \text{as } x \to 0, \tag{2.11}$$

then

$$\mathbb{P}\left\{\sup_{f\in\mathcal{F}} \left|\frac{\hat{\delta}_n^{\phi}(f)}{\delta_n^{\phi}(f)} - 1\right| \to 0 \text{ as } n \to \infty\right\} = 1. \tag{2.12}$$

The following theorem provides upper bounds on F_f in terms of $F_{n,f}$ and on $F_{n,f}$ in terms of F_f uniformly over the class $\mathcal{F}$ satisfying the entropy condition (2.1).

THEOREM 3. – *Suppose that condition* (2.1) *holds with some $\psi \in \Psi$ such that*

$$\psi(x) \geqslant x\sqrt{\log\frac{\mathrm{e}}{x}}, \quad x \leqslant 1$$

and with

$$\sup_{n} \mathbb{E} D_n < +\infty.$$

Then there exist $\theta > 0$ and for any $\sigma \in (0,1]$ $A(\sigma) < +\infty$ and $c := c(\sigma) \geqslant 1,\ c(\sigma) \downarrow 1$ as $\sigma \downarrow 0$ such that for all $\sigma \in (0,1]$ and all $t \geqslant 2\log n$

$$\mathbb{P}\left\{\exists f \in \mathcal{F}\ \exists \delta \leqslant \delta_n^{\psi}\left(\frac{t}{n}\right)\colon\ F_f\left(\frac{\delta}{c(\sigma)}\right) \geqslant c(\sigma)\left[F_{n,f}(\delta) \vee \frac{1}{\sigma}\varepsilon_n^{\psi}(\delta)\right]\right\} \leqslant A(\sigma)\exp\{-\theta t\} \tag{2.13}$$

and

$$\mathbb{P}\left\{\exists f \in \mathcal{F}\ \exists \delta \leqslant \delta_n^{\psi}\left(\frac{t}{n}\right)\colon\ F_{n,f}\left(\frac{\delta}{c(\sigma)}\right) \geqslant c(\sigma)\left[F_f(\delta) \vee \frac{1}{\sigma}\varepsilon_n^{\psi}(\delta)\right]\right\} \leqslant A(\sigma)\exp\{-\theta t\}. \tag{2.14}$$

Remark. – It follows from the proofs below that $c(\sigma)$ in Theorem 3 is of the order $1 + \mathrm{O}(\sigma^{\lambda})$ for some $\lambda > 0$ and that $A(\sigma)$ grows as a power of σ as $\sigma \to 0$. It would be interesting to determine the best (the largest) possible value of the exponent λ for a given exponent of $A(\sigma)$.

The next statement follows almost immediately.

THEOREM 4. – *Suppose that condition* (2.1) *holds with some* $\psi \in \Psi$ *such that*

$$\psi(x) \geqslant \sqrt{a}x\sqrt{\log\frac{\mathrm{e}}{x}}, \quad x \leqslant 1$$

(*where* $a > 2/\theta,$ θ *being the constant in* (2.13), (2.14)). *Suppose also that*

$$\sup_n \mathbb{E}D_n < +\infty$$

and condition (2.6) *holds. Then, for for some* $c > 1,$ *with probability* 1

$$\exists N \geqslant 1\ \forall n \geqslant N\ \forall f \in \mathcal{F}\ \forall \delta \geqslant \hat{\delta}_n^{\psi}(f)\colon\ F_f\left(\frac{\delta}{c}\right) \leqslant cF_{n,f}(\delta) \tag{2.15}$$

and

$$\exists N \geqslant 1\ \forall n \geqslant N\ \forall f \in \mathcal{F}\ \forall \delta \geqslant \delta_n^{\psi}(f)\colon\ F_{n,f}\left(\frac{\delta}{c}\right) \leqslant cF_f(\delta). \tag{2.16}$$

Moreover, let $\phi \in \Psi$ *be such that condition* (2.9) *holds. Then with probability* 1

$$\forall c > 1\ \exists N \geqslant 1\ \forall n \geqslant N\ \forall f \in \mathcal{F}\ \forall \delta \geqslant \hat{\delta}_n^{\phi}(f)\colon\ F_f\left(\frac{\delta}{c}\right) \leqslant cF_{n,f}(\delta). \tag{2.17}$$

and

$$\forall c > 1\ \exists N \geqslant 1\ \forall n \geqslant N\ \forall f \in \mathcal{F}\ \forall \delta \geqslant \delta_n^{\phi}(f)\colon\ F_{n,f}\left(\frac{\delta}{c}\right) \leqslant cF_f(\delta). \tag{2.18}$$

The proofs of the results are based on the following theorem that refines previous bounds of this type obtained by Koltchinskii and Panchenko [9], Koltchinskii, Panchenko and Lozano [12].

THEOREM 5. – *Suppose that condition* (2.1) *holds with some* $\psi \in \Psi$. *Then, for all* $\delta > 0$ *and for all* $\varepsilon > 0,$ $\sigma \in (0, 1]$ *such that* $\varepsilon\sigma \geqslant \varepsilon_n^{\psi}(\delta) \vee \frac{2\log n}{n}$, *the following bounds hold*

$$\begin{aligned}&\mathbb{P}\bigl\{\exists f \in \mathcal{F}\ F_{n,f}(\delta) \leqslant \varepsilon \text{ and } F_f\bigl(A_{\nu}(\sigma)\delta\bigr) \geqslant B_{\lambda}(\sigma)\varepsilon\bigr\}\\ &\quad\leqslant D\left(\log_2\frac{\log_2(\varepsilon\sigma)^{-1}}{1+\log_2\sigma^{-1}} \vee 1\right)\exp\left\{-\frac{n\sigma\varepsilon}{2}\right\}\end{aligned}$$

and

$$\mathbb{P}\big\{\exists f\in\mathcal{F}\ F_f(\delta)\leqslant\varepsilon \text{ and } F_{n,f}\big(A_\nu(\sigma)\delta\big)\geqslant B_\lambda(\sigma)\varepsilon\big\}$$
$$\leqslant D\left(\log_2\frac{\log_2(\varepsilon\sigma)^{-1}}{1+\log_2\sigma^{-1}}\vee 1\right)\exp\left\{-\frac{n\sigma\varepsilon}{2}\right\},$$

where $A_\nu(\sigma):=(1-A\sigma^\nu)\vee\frac{1}{2},\ B_\lambda(\sigma):=1+B\sigma^\lambda,\ \nu,\lambda>0,\ \lambda\leqslant 1/2,\ \lambda+4\nu\leqslant 1,$ $A=\bar{A}(1+\mathbb{E}D_n)^2,\ B=\overline{B}(1+\mathbb{E}D_n)^2$*, and* $\bar{A},\ \overline{B},\ D$ *are numerical constants.*

3. Proofs of the main results

Proof of Theorem 2. – First we show, following Koltchinskii and Panchenko [9], that conditions (2.6) and (2.1) imply that with probability 1

$$\lim_{M\to\infty}\limsup_{n\to\infty}\sup_{f\in\mathcal{F}}P_n\{f\geqslant M\}=0. \tag{3.1}$$

Indeed, let g be the function from $\mathbb{R}$ into [0, 1] that is equal to 0 for $u\leqslant M-1$, equal to 1 for $u>M$ and is linear in between. Then

$$\sup_{f\in\mathcal{F}}P_n\{f\geqslant M\}\leqslant\sup_{f\in\mathcal{F}}P_n g(f)$$
$$\leqslant\sup_{f\in\mathcal{F}}Pg(f)+\|P_n-P\|_{\mathcal{G}}\leqslant\sup_{f\in\mathcal{F}}P\{f\geqslant M-1\}+\|P_n-P\|_{\mathcal{G}},$$

where

$$\mathcal{G}:=\{g\circ f\colon\ f\in\mathcal{F}\}\cup\{0\}.$$

Since the first term tends to 0 as $M\to\infty$, it is enough to show that $\|P_n-P\|_{\mathcal{G}}\to 0$ as $n\to\infty$ a.s. By concentration inequalities, this reduces to showing that $\mathbb{E}\|P_n-P\|_{\mathcal{G}}\to 0$ as $n\to\infty$, which in turn would follow (by a standard symmetrization argument) from

$$\mathbb{E}\left\|n^{-1}\sum_{i=1}^n\varepsilon_i\delta_{X_i}\right\|_{\mathcal{G}}\to 0,$$

where $\{\varepsilon_i\}$ are i.i.d. Rademacher random variables independent of $\{X_i\}$. The entropy bound for the Rademacher process yields

$$\mathbb{E}_\varepsilon\left\|n^{-1}\sum_{i=1}^n\varepsilon_i\delta_{X_i}\right\|_{\mathcal{G}}\leqslant\frac{\text{const}}{\sqrt{n}}\int_0^{\sqrt{2}}H^{1/2}_{d_{P_n,2}}(\mathcal{G};u)\,du.$$

Since g is Lipschitz with constant 1, we have

$$d_{P_n,2}(g\circ f_1,g\circ f_2)\leqslant d_{P_n,2}(f_1,f_2),\qquad f_1,f_2\in\mathcal{F}.$$

This easily gives (see (3.20) in the proof of Theorem 5)

$$H^{1/2}_{d_{P_n,2}}(\mathcal{G};u)\leqslant H^{1/2}_{d_{P_n,2}}(\mathcal{F};u)+1.$$

Therefore, under the condition (2.1) and the condition $\sup_n \mathbb{E}D_n < +\infty$

$$\mathbb{E}\left\|n^{-1}\sum_{i=1}^{n}\varepsilon_i\delta_{X_i}\right\|_{\mathcal{G}} \leqslant \frac{\text{const}}{\sqrt{n}}\big(\mathbb{E}D_n\psi(\sqrt{2}) + \sqrt{2}\big) \to 0,$$

which completes the proof of (3.1).

We will prove only (2.8). The proof of (2.7) is very similar (but somewhat easier).

Let $\sigma \in (0,1]$ be an arbitrary (small) number. Denote $t_n := \frac{2+\gamma}{\sigma}\log n$ (with $\gamma > 0$). Since $\frac{\phi(x)}{\psi(x)} \to +\infty$ as $x \to 0$, for any $\sigma \in (0,1]$ there exists $\kappa := \kappa(\sigma)$ such that for all $x \leqslant \kappa$ $\psi(x) \leqslant \sigma\phi(x)$. Suppose that $\delta_n^{\psi}(\varepsilon\sigma)\sqrt{\varepsilon\sigma} \leqslant \kappa$. Since ϕ is nondecreasing, this implies for $\sigma \in (0,1)$ that

$$\frac{\phi(\delta_n^{\psi}(\varepsilon\sigma)\sqrt{\varepsilon})}{\delta_n^{\psi}(\varepsilon\sigma)\sqrt{n}} \geqslant \frac{\phi(\delta_n^{\psi}(\varepsilon\sigma)\sqrt{\varepsilon\sigma})}{\delta_n^{\psi}(\varepsilon\sigma)\sqrt{n}} \geqslant \frac{1}{\sigma}\frac{\psi(\delta_n^{\psi}(\varepsilon\sigma)\sqrt{\varepsilon\sigma})}{\delta_n^{\psi}(\varepsilon\sigma)\sqrt{n}} = \frac{1}{\sigma}\varepsilon\sigma = \varepsilon.$$

Using the fact that the concavity of ϕ implies that the function $\delta \mapsto \phi(\delta\sqrt{\varepsilon})/\delta\sqrt{n}$ is nonincreasing, we easily conclude from the definitions of δ_n^{ψ}, δ_n^{ϕ} that

$$\delta_n^{\psi}(\varepsilon\sigma) \leqslant \delta_n^{\phi}(\varepsilon). \tag{3.2}$$

On the other hand, if $\delta_n^{\psi}(\varepsilon\sigma)\sqrt{\varepsilon\sigma} > \kappa$, we have

$$\sigma\varepsilon = \frac{\psi(\delta_n^{\psi}(\varepsilon\sigma)\sqrt{\varepsilon\sigma})}{\delta_n^{\psi}(\varepsilon\sigma)\sqrt{n}} \leqslant \frac{\psi(\kappa/\sqrt{\varepsilon\sigma}\sqrt{\varepsilon\sigma})}{\kappa/\sqrt{\varepsilon\sigma}\sqrt{n}} \leqslant \frac{\psi(\kappa)}{\kappa}\sqrt{\frac{\varepsilon\sigma}{n}},$$

which implies

$$\varepsilon \leqslant \left(\frac{\psi(\kappa)}{\kappa}\right)^2\frac{1}{\sigma n}.$$

Since $t_n \to \infty$ as $n \to \infty$, for all large enough n we have

$$t_n > \left(\frac{\psi(\kappa)}{\kappa}\right)^2\frac{1}{\sigma}.$$

Therefore, for all $\varepsilon > t_n/n$, we have $\delta_n^{\psi}(\varepsilon\sigma)\sqrt{\varepsilon\sigma} \leqslant \kappa$, which, as we proved, implies (3.2).

Next note that the condition $\psi(x) \geqslant 2\sqrt{2}x\sqrt{\log(\mathrm{e}/x)}$ for $x \leqslant 1$ easily implies that $\varphi^{-1}(\sqrt{y}) \geqslant \mathrm{e}^{1-y/8}$ for $y \geqslant 8$ (recall that $\varphi(x) = \psi(x)/x$). Therefore, for large n, we have (if $\gamma < 2$)

$$\delta_n^{\phi}\left(\frac{t_n}{n}\right) \geqslant \delta_n^{\psi}\left(\frac{t_n\sigma}{n}\right) = \frac{\varphi^{-1}(\sqrt{t_n\sigma})}{\sqrt{t_n\sigma}}\sqrt{n} \geqslant \frac{\mathrm{e}^{1-t_n\sigma/8}}{\sqrt{t_n\sigma}}\sqrt{n} = \frac{\mathrm{e}}{\sqrt{t_n\sigma}}n^{-(2+\gamma)/8+1/2} \to \infty. \tag{3.3}$$

Hence, using (2.6) and (3.1),

$$\inf_{f\in\mathcal{F}} F_f\left(\delta_n^{\phi}\left(\frac{t_n}{n}\right)\right) \to 1$$

and a.s.

$$\inf_{f\in\mathcal{F}} F_{n,f}\left(\delta_n^{\phi}\left(\frac{t_n}{n}\right)\right)\to 1,$$

which implies that a.s. for large enough n and for all $f\in\mathcal{F}$

$$\varepsilon_n^{\phi}(f)=\varepsilon_n^{\phi}(f;t_n)\quad\text{and}\quad\hat{\varepsilon}_n^{\phi}(f)=\hat{\varepsilon}_n^{\phi}(f;t_n).$$

Therefore, in the rest of the proof we can replace the ϕ-bounds by the truncated ϕ-bounds.

By the definition of t_n, for all large enough n we have

$$t_n>\left(\frac{\psi(\kappa)}{\kappa}\right)^2\frac{1}{\sigma}\vee\frac{2}{\sigma}\log n\vee\frac{1}{\sigma^2}.$$

Note that the condition $\varepsilon\sigma\geqslant\varepsilon_n^{\psi}(\delta)$ is equivalent to the condition $\delta\geqslant\delta_n^{\psi}(\varepsilon\sigma)$. Therefore, for $\varepsilon\geqslant t_n/n$ and $\delta=\delta_n^{\phi}(\varepsilon)$ we have $\varepsilon\sigma\geqslant\varepsilon_n^{\psi}(\delta)\vee\frac{2\log n}{n}$, which means that the bounds of Theorem 5 hold for $\varepsilon\geqslant t_n/n$ and $\delta=\delta_n^{\phi}(\varepsilon)$.

Recall that $A_\nu(\sigma)\leqslant 1\leqslant B_\lambda(\sigma)$. Applying the first bound, we get

$$\begin{aligned}&\mathbb{P}\bigl\{\exists f\in\mathcal{F}\ F_{n,f}\bigl(\delta_n^{\phi}(\varepsilon)\bigr)\leqslant\varepsilon\text{ and }F_f\bigl(A_\nu(\sigma)\delta_n^{\phi}(\varepsilon)\bigr)\geqslant B_\lambda(\sigma)\varepsilon\bigr\}\\&\quad\leqslant B\log_2\frac{\log_2(\varepsilon\sigma)^{-1}}{1+\log_2\sigma^{-1}}\exp\left\{-\frac{n\varepsilon\sigma}{2}\right\}.\end{aligned}$$

Next we set $\varepsilon_j:=B_\lambda(\sigma)^{-j}$. Let $\mathcal{J}=\{j\geqslant 0\colon\ \varepsilon_j\geqslant t_n/n\}$ and

$$E_n:=\bigl\{\exists j\in\mathcal{J}\ \exists f\in\mathcal{F}: F_{n,f}\bigl(\delta_n^{\phi}(\varepsilon_j)\bigr)\leqslant\varepsilon_j\text{ and }F_f\bigl(A_\nu(\sigma)\delta_n^{\phi}(\varepsilon_j)\bigr)\geqslant B_\lambda(\sigma)\varepsilon_j\bigr\}.$$

Note that for all $j\in\mathcal{J}$

$$\varepsilon_j\geqslant B_\lambda(\sigma)^{j_0-j}\frac{t_n}{n},$$

where $j_0:=\inf\mathcal{J}$. Hence, we have

$$\begin{aligned}\mathbb{P}(E_n)&\leqslant B\sum_{j\in\mathcal{J}}\log_2\frac{\log_2(\varepsilon_j\sigma)^{-1}}{1+\log_2\sigma^{-1}}\exp\left\{-\frac{n\varepsilon_j\sigma}{2}\right\}\\&\leqslant B\log_2\left(\frac{\log_2(n/t_n)+\log_2\sigma^{-1}}{1+\log_2\sigma^{-1}}\right)\sum_{j\geqslant 0}\exp\left\{-\frac{t_n\sigma}{2}B_\lambda(\sigma)^j\right\}\\&\leqslant B'(\sigma)\log_2\left(\frac{\log_2(n/t_n)+\log_2\sigma^{-1}}{1+\log_2\sigma^{-1}}\right)\exp\left\{-\frac{t_n\sigma}{2}\right\}.\end{aligned}\tag{3.4}$$

Suppose that for some j and for some $f\in\mathcal{F}$, $\hat{\varepsilon}_n^{\phi}(f;t_n)\in(\varepsilon_{j+1},\varepsilon_j]$. On the event E_n^c, the inequality $F_{n,f}(\delta_n^{\phi}(\varepsilon_j))\leqslant\varepsilon_j$ implies that $F_f(A_\nu(\sigma)\delta_n^{\phi}(\varepsilon_j))\leqslant B_\lambda(\sigma)\varepsilon_j$. Since, for $\bar{\varphi}(x):=\frac{\phi(x)}{x}$,

$$A_\nu(\sigma)\delta_n^{\phi}(\varepsilon_j)=\frac{A_\nu(\sigma)\bar{\varphi}^{-1}(\sqrt{\varepsilon_j n})}{\sqrt{\varepsilon_j}}\geqslant\frac{\bar{\varphi}^{-1}\bigl(\sqrt{A_\nu(\sigma)^{-2}\varepsilon_j n}\bigr)}{\sqrt{A_\nu(\sigma)^{-2}\varepsilon_j}}=\delta_n^{\phi}\bigl(A_\nu(\sigma)^{-2}\varepsilon_j\bigr),$$

we also have $F_f(\delta_n^\phi(A_\nu(\sigma)^{-2}\varepsilon_j)) \leqslant B_\lambda(\sigma)\varepsilon_j$, which implies

$$F_f\big(\delta_n^\phi\big(A_\nu(\sigma)^{-2}B_\lambda(\sigma)\hat{\varepsilon}_n^\phi(f;t_n)\big)\big) \leqslant B_\lambda(\sigma)^2\hat{\varepsilon}_n^\phi(f;t_n).$$

Therefore, on the event E_n^c, we get for all $f \in \mathcal{F}$,

$$\varepsilon_n^\phi(f;t_n) \leqslant \big(A_\nu(\sigma)^{-2}B_\lambda(\sigma) \vee B_\lambda(\sigma)^2\big)\hat{\varepsilon}_n^\phi(f;t_n).$$

It follows from (3.4) that

$$\mathbb{P}\big\{\exists f \in \mathcal{F}\colon\ \varepsilon_n^\phi(f;t_n) \geqslant \big(A_\nu(\sigma)^{-2}B_\lambda(\sigma) \vee B_\lambda(\sigma)^2\big)\hat{\varepsilon}_n^\phi(f;t_n)\big\}$$
$$\leqslant B'(\sigma)\log_2\left(\frac{\log_2(n/t_n)+\log_2\sigma^{-1}}{1+\log_2\sigma^{-1}}\right)\exp\left\{-\frac{t_n\sigma}{2}\right\}.$$

Quite similarly, using the second bound of Theorem 5, one can prove that

$$\mathbb{P}\big\{\exists f \in \mathcal{F}\colon\ \hat{\varepsilon}_n^\phi(f;t_n) \geqslant \big(A_\nu(\sigma)^{-2}B_\lambda(\sigma) \vee B_\lambda(\sigma)^2\big)\varepsilon_n^\phi(f;t_n)\big\}$$
$$\leqslant B'(\sigma)\log_2\left(\frac{\log_2(n/t_n)+\log_2\sigma^{-1}}{1+\log_2\sigma^{-1}}\right)\exp\left\{-\frac{t_n\sigma}{2}\right\}.$$

By the definition of t_n we have for all $\sigma > 0$

$$\sum_n B'(\sigma)\log_2\left(\frac{\log_2(n/t_n)+\log_2\sigma^{-1}}{1+\log_2\sigma^{-1}}\right)\exp\left\{-\frac{t_n\sigma}{2}\right\} < +\infty.$$

By Borel–Cantelli lemma, we conclude that with probability 1, eventually (for all large n), we have for all $f \in \mathcal{F}$

$$\hat{\varepsilon}_n^\phi(f;t_n) < \big(A_\nu(\sigma)^{-2}B_\lambda(\sigma) \vee B_\lambda(\sigma)^2\big)\varepsilon_n^\phi(f;t_n)$$

and

$$\varepsilon_n^\phi(f;t_n) < \big(A_\nu(\sigma)^{-2}B_\lambda(\sigma) \vee B_\lambda(\sigma)^2\big)\hat{\varepsilon}_n^\phi(f;t_n).$$

Since the above bounds hold for all $\sigma > 0$ and $A_\nu(\sigma) \to 1$, $B_\lambda(\sigma) \to 1$, the result follows. □

Proof of Corollary 1. – The condition $\psi(xy) \leqslant \psi(x)\psi(y)$ easily implies that $\varphi(xy) \leqslant \varphi(x)\varphi(y)$ and $\varphi^{-1}(xy) \leqslant \varphi^{-1}(x)\varphi^{-1}(y)$. Hence, for all $\varepsilon > 0$

$$\delta_n^\psi(\varepsilon) = \frac{\varphi^{-1}(\sqrt{\varepsilon n})}{\sqrt{\varepsilon}} = \sqrt{c}\frac{\varphi^{-1}(\sqrt{c^{-1}c\varepsilon n})}{\sqrt{c\varepsilon}} \leqslant \sqrt{c}\varphi^{-1}\big(c^{-1/2}\big)\delta_n^\psi(c\varepsilon),$$

which implies for all δ

$$c\varepsilon_n^\psi(\delta) \leqslant \varepsilon_n^\psi\left(\frac{\delta}{\sqrt{c}\varphi^{-1}(c^{-1/2})}\right).$$

Since $c \mapsto c^{1/2}\varphi^{-1}(c^{-1/2})$ is an increasing continuous function and

$$c^{1/2}\varphi^{-1}\big(c^{-1/2}\big) \to \infty \quad \text{as } c \to \infty, \qquad c^{1/2}\varphi^{-1}\big(c^{-1/2}\big) \to 0 \quad \text{as } c \to 0,$$

for all $A > 0$ there exists $c := C(A) > 0$ such that

$$c^{1/2}\varphi^{-1}\bigl(c^{-1/2}\bigr) = A.$$

Moreover, $C(A) \to \infty$ as $A \to \infty$. Therefore, we have the following bound for all δ:

$$C(A)\varepsilon_n^{\psi}(\delta) \leqslant \varepsilon_n^{\psi}\left(\frac{\delta}{A}\right).$$

Now, the assumption $\hat{\delta}_n^{\psi}(f) \leqslant \frac{1}{A}\delta_n^{\psi}(f)$ implies that

$$\hat{\varepsilon}_n^{\psi}(f) = \varepsilon_n^{\psi}\bigl(\hat{\delta}_n^{\psi}(f)\bigr) \geqslant \varepsilon_n^{\psi}\bigl(A^{-1}\delta_n^{\psi}(f)\bigr) \geqslant C(A)\varepsilon_n^{\psi}\bigl(\delta_n^{\psi}(f)\bigr) = C(A)\varepsilon_n^{\psi}(f).$$

Therefore,

$$\limsup_n \sup_{f\in\mathcal{F}} \frac{\hat{\varepsilon}_n^{\psi}(f)}{\varepsilon_n^{\psi}(f)} < C(A)$$

implies

$$\liminf_n \inf_{f\in\mathcal{F}} \frac{\hat{\delta}_n^{\psi}(f)}{\delta_n^{\psi}(f)} > \frac{1}{A}.$$

Quite similarly,

$$\liminf_n \inf_{f\in\mathcal{F}} \frac{\hat{\varepsilon}_n^{\psi}(f)}{\varepsilon_n^{\psi}(f)} > \frac{1}{C(A)}$$

implies

$$\limsup_n \sup_{f\in\mathcal{F}} \frac{\hat{\delta}_n^{\psi}(f)}{\delta_n^{\psi}(f)} < A.$$

Since $C(A) \to \infty$ as $A \to \infty$, the first statement of the corollary follows from Theorem 2.

The proof of the second statement is very similar if one takes into account that the condition $\phi(1) = 1$ implies

$$c^{1/2}\bar{\varphi}^{-1}\bigl(c^{-1/2}\bigr) \downarrow 1 \quad \text{as } c \downarrow 1,$$

where $\bar{\varphi}(x) = \phi(x)/x$, which allows us to show that

$$C(A)\varepsilon_n^{\phi}(\delta) \leqslant \varepsilon_n^{\phi}\left(\frac{\delta}{A}\right)$$

with $C(A) \downarrow 1$ as $A \downarrow 1$. □

Proof of Theorem 3. – First note that it is enough to prove the bounds for sufficiently large n (since it is assumed that $t \geqslant 2\log n$, for small enough n the right-hand sides of the inequalities can be made larger than 1 by a proper choice of $A(\sigma)$ which makes them trivial). Also, we are assuming in what follows in the proof that, for given $t \geqslant 2\log n$ and σ, ε and δ are such that

$$\varepsilon\sigma \geqslant \varepsilon_n^{\psi}(\delta) \geqslant \frac{t}{n}.$$

Now, note that for all large n the condition

$$\varepsilon\sigma \geqslant \frac{t}{n} \geqslant \frac{2\log n}{n}$$

implies that

$$\log_2 \frac{\log_2(\varepsilon\sigma)^{-1}}{1+\log_2\sigma^{-1}} \leqslant \log_2\log_2(\varepsilon\sigma)^{-1} \leqslant \mathrm{e}^{n\varepsilon\sigma/6}.$$

Therefore, it follows from Theorem 5 that with some $c = c(\sigma) \downarrow 1$ as $\sigma \downarrow 0$ we have (under the assumption $\varepsilon\sigma \geqslant \varepsilon_n^{\psi}(\delta)$)

$$\mathbb{P}\left\{\exists f \in \mathcal{F}\colon F_{n,f}(\delta) \leqslant \varepsilon \text{ and } F_f\left(\frac{\delta}{c}\right) \geqslant c\varepsilon\right\} \leqslant D\,\mathrm{e}^{-n\varepsilon\sigma/3}.$$

Note also that $\varepsilon_n^{\psi}(\delta) \geqslant \frac{t}{n}$ iff $\delta \leqslant \delta_n^{\psi}(\frac{t}{n})$ and $\varepsilon\sigma \geqslant \varepsilon_n^{\psi}(\delta)$ iff $\delta \geqslant \delta_n^{\psi}(\varepsilon\sigma)$. Denote

$$\delta_j := \delta_n^{\psi}\left(\frac{t}{n}\right)(1+\sigma)^{-j}, \quad j = 0, 1, 2, \ldots,$$

and let

$$J := \left\{j\colon \delta_n^{\psi}(\varepsilon\sigma) \leqslant \delta_j \leqslant \delta_n^{\psi}\left(\frac{t}{n}\right)\right\}.$$

Denote also

$$E := \left\{\exists f \in \mathcal{F}\ \exists j \in J\colon F_{n,f}(\delta_j) \leqslant \varepsilon \text{ and } F_f\left(\frac{\delta_j}{c}\right) \geqslant c\varepsilon\right\}.$$

Then

$$\mathbb{P}(E) \leqslant D\frac{\log(\delta_n^{\psi}(t/n)/\delta_n^{\psi}(\varepsilon\sigma))}{\log(1+\sigma)}\exp\left\{-\frac{n\varepsilon\sigma}{3}\right\}.$$

Since

$$\frac{\delta_n^{\psi}(t/n)}{\delta_n^{\psi}(\varepsilon\sigma)} = \frac{\varphi^{-1}(\sqrt{t})}{\sqrt{t}}\frac{\sqrt{n\varepsilon\sigma}}{\varphi^{-1}(\sqrt{n\varepsilon\sigma})},$$

we get, using that $\varphi(x) \geqslant \sqrt{\log(\mathrm{e}/x)}$, $x \leqslant 1$, that for all large n

$$\frac{1}{\varphi^{-1}(\sqrt{n\varepsilon\sigma})} \leqslant \mathrm{e}^{-1+n\varepsilon\sigma},$$

which implies

$$\log\frac{\delta_n^{\psi}(t/n)}{\delta_n^{\psi}(\varepsilon\sigma)} \leqslant \frac{1}{2}\log(n\varepsilon\sigma) + n\varepsilon\sigma - 1 + \log\frac{\varphi^{-1}(\sqrt{t})}{\sqrt{t}} \leqslant 2n\varepsilon\sigma.$$

Then, for large enough n and $n\varepsilon\sigma \geqslant t \geqslant 2\log n$, we have

$$\mathbb{P}(E) \leqslant D\frac{2n\varepsilon\sigma}{\log(1+\sigma)}\exp\left\{-\frac{n\varepsilon\sigma}{3}\right\} \leqslant \frac{4D}{\sigma}\exp\left\{-\frac{n\varepsilon\sigma}{4}\right\}.$$

On the event E^c, for any $j \in J$ and for any $\delta \in (\delta_j, \delta_{j-1}]$, the condition $F_{n,f}(\delta) \leqslant \varepsilon$ implies $F_{n,f}(\delta_j) \leqslant \varepsilon$, which in turn implies $F_f(\delta_j/c) \leqslant c\varepsilon$ and hence $F_f(\delta/c(1+\sigma)) \leqslant c\varepsilon$. Replacing c by $c(1+\sigma)$ and $4D$ by D, this yields

$$\mathbb{P}\left\{\exists f \in \mathcal{F}\ \exists\delta \in \left[\delta_n^{\psi}(\varepsilon\sigma), \delta_n^{\psi}\left(\frac{t}{n}\right)\right]:\ F_{n,f}(\delta) \leqslant \varepsilon \text{ and } F_f\left(\frac{\delta}{c}\right) \geqslant c\varepsilon\right\}$$
$$\leqslant \frac{D}{\sigma}\exp\left\{-\frac{n\varepsilon\sigma}{4}\right\}.$$

Next we set

$$\varepsilon_j := (1+\sigma)^j\frac{1}{\sigma}\frac{t}{n}$$

and introduce the event

$$F := \left\{\exists f \in \mathcal{F}\ \exists j \geqslant 0\ \exists\delta \in \left[\delta_n^{\psi}(\varepsilon_j\sigma), \delta_n^{\psi}\left(\frac{t}{n}\right)\right]:\ F_{n,f}(\delta) \leqslant \varepsilon_j \text{ and } F_f\left(\frac{\delta}{c}\right) \geqslant c\varepsilon_j\right\}.$$

We get

$$\mathbb{P}(F) \leqslant \frac{D}{\sigma}\sum_{j=0}^{\infty}\mathrm{e}^{-n\varepsilon_j\sigma/4} = \frac{D}{\sigma}\sum_{j=0}^{\infty}\exp\left\{-\frac{t(1+\sigma)^j}{4}\right\}$$
$$\leqslant \frac{D}{\sigma}\mathrm{e}^{-t/4}\sum_{j=0}^{\infty}\mathrm{e}^{-tj\sigma/4} = \frac{D}{\sigma}\mathrm{e}^{-t/4}\left(1-\mathrm{e}^{-t\sigma/4}\right)^{-1},$$

where we used a simple inequality $(1+\sigma)^j - 1 \geqslant j\sigma$. Note that, for $\sigma < 1/2$,

$$1-\mathrm{e}^{-t\sigma/4} \geqslant \left(\frac{t\sigma}{4}\right)\mathrm{e}^{-t\sigma/4} \geqslant \left(\frac{t\sigma}{4}\right)\mathrm{e}^{-t/8}$$

and for $\sigma \geqslant 1/2$

$$1-\mathrm{e}^{-t\sigma/4} \geqslant 1-\mathrm{e}^{-t/8} \geqslant \frac{1}{2},$$

since $t \geqslant 2\log n$ and n can be assumed large enough. This yields (with some D) the bound

$$\mathbb{P}(F) \leqslant \frac{D}{\sigma^2}\mathrm{e}^{-t/8}.$$

On the event F^c, for any j, for any $\varepsilon \in (\varepsilon_{j-1}, \varepsilon_j]$, for all $f \in \mathcal{F}$ and for any

$$\delta \in \left[\delta_n^{\psi}(\varepsilon\sigma), \delta_n^{\psi}\left(\frac{t}{n}\right)\right] \subset \left[\delta_n^{\psi}(\varepsilon_j\sigma), \delta_n^{\psi}\left(\frac{t}{n}\right)\right],$$

the condition $F_{n,f}(\delta) \leqslant \varepsilon$ implies $F_{n,f}(\delta) \leqslant \varepsilon_j$, which implies that $F_f(\delta/c) \leqslant c\varepsilon_j$ and hence $F_f(\delta/c) \leqslant c(1+\sigma)\varepsilon$. If we replace c by $c(1+\sigma)$, the above remarks allow us to show that

$$\begin{aligned}&\mathbb{P}\left\{\exists f \in \mathcal{F}\ \exists \varepsilon \geqslant \frac{t}{\sigma n}\ \exists \delta \in \left[\delta_n^{\psi}(\varepsilon\sigma), \delta_n^{\psi}\left(\frac{t}{n}\right)\right]\colon F_{n,f}(\delta) \leqslant \varepsilon \text{ and } F_f\left(\frac{\delta}{c}\right) \geqslant c\varepsilon\right\}\\ &\quad\leqslant \frac{D}{\sigma^2}\exp\left\{-\frac{t}{8}\right\}.\end{aligned}$$

If $\delta \leqslant \delta_n^{\psi}(t/n)$ and

$$\varepsilon := F_{n,f}(\delta) \vee \frac{\varepsilon_n^{\psi}(\delta)}{\sigma},$$

we have $\varepsilon_n^{\psi}(\delta) \geqslant t/n$ and hence $\varepsilon \geqslant t/(\sigma n)$. At the same time, we have $\varepsilon\sigma \geqslant \varepsilon_n^{\psi}(\delta)$, which implies $\delta \geqslant \delta_n^{\psi}(\varepsilon\sigma)$. Therefore, we obtain

$$\mathbb{P}\left\{\exists f \in \mathcal{F}\ \exists \delta \leqslant \delta_n^{\psi}\left(\frac{t}{n}\right)\colon F_f\left(\frac{\delta}{c}\right) \geqslant c\left[F_{n,f}(\delta) \vee \frac{\varepsilon_n^{\psi}(\delta)}{\sigma}\right]\right\} \leqslant \frac{D}{\sigma^2}\exp\left\{-\frac{t}{8}\right\}.$$

The proof of the second inequality is similar. □

Proof of Theorem 4. – We set $t_n := \frac{1+\gamma}{\theta}\log n$ with $\gamma \in (0, (a\theta)/2 - 1)$. Theorem 3 implies that the event

$$\left\{\exists f \in \mathcal{F}\ \exists \delta \in \left[\hat{\delta}_n^{\psi}(f), \delta_n^{\psi}\left(\frac{t_n}{n}\right)\right]\colon F_f\left(\frac{\delta}{c(\sigma)}\right) \geqslant \frac{c(\sigma)}{\sigma}F_{n,f}(\delta)\right\}$$

occurs with probability at most

$$A(\sigma)\,\mathrm{e}^{-\theta t_n} = A(\sigma)n^{-1-\gamma}.$$

By Borel–Cantelli Lemma, this implies that with probability 1

$$\exists N \geqslant 1\ \forall n \geqslant N\ \forall f \in \mathcal{F}\ \forall \delta \in \left[\hat{\delta}_n^{\psi}(f), \delta_n^{\psi}\left(\frac{t_n}{n}\right)\right]\colon F_f\left(\frac{\delta}{c(\sigma)}\right) \leqslant \frac{c(\sigma)}{\sigma}F_{n,f}(\delta). \tag{3.5}$$

The condition $\psi(x) \geqslant \sqrt{a}x\sqrt{\log(\mathrm{e}/x)}$ for $x \leqslant 1$ implies that $\varphi^{-1}(\sqrt{y}) \geqslant \mathrm{e}^{1-y/a}$ for $y \geqslant a$, which in turn implies that

$$\delta_n^{\psi}\left(\frac{t_n}{n}\right) \geqslant \frac{\mathrm{e}^{-t_n/a}}{\sqrt{t_n}}n^{1/2} = t_n^{-1/2}n^{1/2-(1+\gamma)/(a\theta)} \to \infty.$$

Since

$$\sup_{f\in\mathcal{F}} P\{f \geqslant u\} \to 0 \quad \text{as } u \to \infty$$

and a.s.

$$\lim_{u\to\infty}\limsup_{n\to\infty}\sup_{f\in\mathcal{F}} P_n\{f\geqslant u\}=0$$

(see the proof of Theorem 2), we get

$$\inf_{f\in\mathcal{F}} F_{n,f}\left(\delta_n^{\psi}\left(\frac{t_n}{n}\right)\right)\to 1 \quad \text{a.s.}$$

This implies that with probability 1

$$\exists N\geqslant 1\ \forall n\geqslant N\ \forall f\in\mathcal{F}\ \forall\delta\geqslant\delta_n^{\psi}\left(\frac{t_n}{n}\right):\ F_f\left(\frac{\delta}{c(\sigma)}\right)\leqslant 1\leqslant\frac{c(\sigma)}{\sigma}F_{n,f}(\delta), \tag{3.6}$$

(provided that $c(\sigma)/\sigma>1$). Together with (3.5) this proves (2.15).

The proof of (2.16) is exactly the same.

The proof of (2.17) and (2.18) is also similar, if one takes into account the following observations made in the proof of Theorem 2 (specifically, see the derivation of (3.2) and (3.3)). First, by (3.3), for large enough n

$$\delta_n^{\phi}\left(\frac{t_n}{n}\right)\geqslant\delta_n^{\psi}\left(\frac{t_n\sigma}{n}\right)\geqslant\delta_n^{\psi}\left(\frac{t_n}{n}\right).$$

Therefore, $\delta\leqslant\delta_n^{\psi}(t_n/n)$ implies $\delta\leqslant\delta_n^{\phi}(t_n/n)$, which is equivalent to $\varepsilon:=\varepsilon_n^{\phi}(\delta)\geqslant t_n/n$. Because of this (see the proof of (3.2)), we have $\delta_n^{\psi}(\varepsilon\sigma)\sqrt{\varepsilon\sigma}\leqslant\kappa$, which implies that $\delta_n^{\psi}(\varepsilon\sigma)\leqslant\delta_n^{\phi}(\varepsilon)=\delta$, or, equivalently, $\varepsilon_n^{\phi}(\delta)\geqslant\frac{1}{\sigma}\varepsilon_n^{\psi}(\delta)$. This allows us to rewrite the first bound of Theorem 3 the following way

$$\mathbb{P}\left\{\exists f\in\mathcal{F}\ \exists\delta\leqslant\delta_n^{\psi}\left(\frac{t_n}{n}\right):\ F_f\left(\frac{\delta}{c(\sigma)}\right)\geqslant c(\sigma)\left[F_{n,f}(\delta)\vee\varepsilon_n^{\phi}(\delta)\right]\right\}\leqslant A(\sigma)\exp\{-\theta t_n\}. \tag{3.7}$$

By the definition of $\hat{\delta}_n^{\phi}(f)$, this implies that the event

$$\left\{\exists f\in\mathcal{F}\ \exists\delta\in\left[\hat{\delta}_n^{\phi}(f),\delta_n^{\psi}\left(\frac{t_n}{n}\right)\right]:\ F_f\left(\frac{\delta}{c(\sigma)}\right)\geqslant c(\sigma)F_{n,f}(\delta)\right\}$$

occurs with probability at most $A(\sigma)n^{-1-\gamma}$, and, by Borel–Cantelli Lemma (as in the proof of (3.5)), with probability 1

$$\exists N\geqslant 1\ \forall n\geqslant N\ \forall f\in\mathcal{F}\ \forall\delta\in\left[\hat{\delta}_n^{\phi}(f),\delta_n^{\psi}\left(\frac{t_n}{n}\right)\right]:\ F_f\left(\frac{\delta}{c(\sigma)}\right)\leqslant c(\sigma)F_{n,f}(\delta).$$

Since we can assume that $c(\sigma)>1$ for all $\sigma\in(0,1]$, we can get rid of the restriction $\delta\leqslant\delta_n^{\psi}(t_n/n)$ by exactly the same argument as before (see (3.6)). Given $c>1$, we can choose σ small enough so that $c(\sigma)<c$ and conclude that with probability 1

$$\exists N\geqslant 1\ \forall n\geqslant N\ \forall f\in\mathcal{F}\ \forall\delta\geqslant\hat{\delta}_n^{\phi}(f):\ F_f\left(\frac{\delta}{c}\right)\leqslant cF_{n,f}(\delta),$$

and since the last events are monotone with respect to c, this completes the proof of (2.17). Similarly, the second bound of Theorem 3 leads to (2.18). □

Proof of Theorem 5. – The method of the proof was developed in Koltchinskii and Panchenko [8,9]. Throughout the proof "const" denotes a constant; its values can be different in different places. Define

$$r_0 := 1, \quad r_{k+1} = \bigl(\varepsilon + C\sqrt{r_k\varepsilon\sigma}\,\bigr) \wedge 1 \tag{3.8}$$

where $C = c(1+\mathbb{E}D_n)$ with a sufficiently large constant $c > 1$ (which will be determined later in the process of the proof). A simple induction shows that either $\varepsilon + C\sqrt{\varepsilon\sigma} \geqslant 1$ and $r_k \equiv 1$, or $\varepsilon + C\sqrt{\varepsilon\sigma} < 1$, and in the last case $\{r_k\}$ is a nonincreasing sequence that converges to the solution $\bar{r}$ of the equation

$$\bar{r} = \varepsilon + C\sqrt{\bar{r}\varepsilon\sigma}. \tag{3.9}$$

A simple computation shows that $\bar{r}$ is bounded from above by $\varepsilon(1 + b\sqrt{\sigma}\,)$ with some constant $b > 0$. Let $d_k := r_k - \bar{r}$. Then

$$d_{k+1} = r_{k+1} - \bar{r} = C\sqrt{\varepsilon\sigma}\bigl(\sqrt{r_k} - \sqrt{\bar{r}}\,\bigr) \leqslant C\sqrt{\varepsilon\sigma}\sqrt{r_k - \bar{r}} = C\sqrt{\varepsilon\sigma}\sqrt{d_k}.$$

By induction, this implies that

$$d_k \leqslant C^{1+2^{-1}+\cdots+2^{-(k-1)}}(\varepsilon\sigma)^{2^{-1}+\cdots+2^{-k}} = C^{2(1-2^{-k})}(\varepsilon\sigma)^{1-2^{-k}} = \bigl(C\sqrt{\varepsilon\sigma}\,\bigr)^{2(1-2^{-k})}.$$

As soon as

$$2^k \geqslant \frac{\log_2(\varepsilon\sigma)^{-1}}{1 + (1-\lambda)\log_2\sigma^{-1}}, \tag{3.10}$$

we have $d_k \leqslant 2C^2\varepsilon\sigma^{\lambda}$. If $\lambda \leqslant 1/2$, we also have in this case (with some constant b)

$$r_k \leqslant \varepsilon\bigl(1 + b\sigma^{\lambda}\bigr).$$

Next we define

$$\tilde{r}_0 := 1, \quad \tilde{r}_{k+1} = C\sqrt{\tilde{r}_k\varepsilon\sigma} \wedge 1.$$

Clearly $\tilde{r}_k \leqslant r_k$ for all $k \geqslant 0$. We also have (in the case when $C\sqrt{\varepsilon\sigma} < 1$)

$$\tilde{r}_k = C^{1+2^{-1}+\cdots+2^{-(k-1)}}(\varepsilon\sigma)^{2^{-1}+\cdots+2^{-k}} = C^{2(1-2^{-k})}(\varepsilon\sigma)^{1-2^{-k}} = \bigl(C\sqrt{\varepsilon\sigma}\,\bigr)^{2(1-2^{-k})}.$$

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

$$\begin{aligned}\gamma_{k-1} + \gamma_{k-2} + \cdots + \gamma_0 &= C^{-1}\bigl[C\sqrt{\varepsilon\sigma} + \bigl(C\sqrt{\varepsilon\sigma}\,\bigr)^{2^{-1}} + \cdots + \bigl(C\sqrt{\varepsilon\sigma}\,\bigr)^{2^{-k}}\bigr]\\ &\leqslant C^{-1}\bigl(C\sqrt{\varepsilon\sigma}\,\bigr)^{2^{-k}}\bigl(1 - \bigl(C\sqrt{\varepsilon\sigma}\,\bigr)^{2^{-k}}\bigr)^{-1}.\end{aligned} \tag{3.11}$$

As far as

$$2^{k+1} \leqslant \nu^{-1} \frac{\log_2(\varepsilon\sigma)^{-1}}{\log_2 \sigma^{-1} + (4\nu)^{-1}}, \tag{3.12}$$

we have

$$\left(C\sqrt{\varepsilon\sigma}\right)^{2^{-k}} \leqslant 2^{-1/4}\sqrt{C}\sigma^{\nu}$$

and hence

$$\gamma_{k-1} + \gamma_{k-2} + \cdots + \gamma_0 \leqslant 2^{-1/4}C^{-1/2}\sigma^{\nu}\left(1 - 2^{-1/4}C^{1/2}\sigma^{\nu}\right)^{-1} \leqslant 2C^{-1/2}\sigma^{\nu}, \tag{3.13}$$

provided that $\sigma \leqslant 2^{-1/\nu}C^{-1/(2\nu)}$. Note also that if $\varepsilon < C^{-4}$, then, for $\sigma \in (0,1]$, $C\sqrt{\varepsilon\sigma} \leqslant (\varepsilon\sigma)^{1/4}$, which implies

$$\left(C\sqrt{\varepsilon\sigma}\right)^{2^{-k}} \leqslant 2^{-1/8}\sigma^{\nu/2} \leqslant 2^{-1/8}$$

and

$$\gamma_{k-1} + \gamma_{k-2} + \cdots + \gamma_0 \leqslant C^{-1}\frac{2^{-1/8}}{1 - 2^{-1/8}} \leqslant \frac{1}{2}$$

(for a sufficiently large C). If $\sigma > 2^{-1/\nu}C^{-1/(2\nu)}$ and $\varepsilon \geqslant C^{-4}$, then the inequalities of the theorem are trivially satisfied by choosing the constant B large enough (so that $B_\lambda(\sigma)\varepsilon > 1$). With an exception of this trivial case, we have (assuming that $2C^{-1/2} \leqslant C^{1/2}$)

$$\gamma_{k-1} + \gamma_{k-2} + \cdots + \gamma_0 \leqslant \sqrt{C}\sigma^{\nu} \wedge 2^{-1}. \tag{3.14}$$

Note that if $\lambda + 4\nu \leqslant 1$, then both (3.10) and (3.12) are satisfied for some k.

Let $\delta > 0$. Define

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

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

$$\mathcal{F}_{k+1} := \left\{ f \in \mathcal{F}_k \colon\ F_f(\delta_{k,1/2}) \leqslant \varepsilon + \frac{C}{2}\sqrt{r_k\varepsilon\sigma} \wedge 1 \right\}.$$

For $k \geqslant 0$, consider a continuous function φ_k from $\mathbb{R}$ into $[0,1]$ such that $\varphi_k(u) = 1$ for $u \leqslant \delta_{k,1/2}$, $\varphi_k(u) = 0$ for $u \geqslant \delta_k$, and φ_k is linear for $\delta_{k,1/2} \leqslant u \leqslant \delta_k$. Also, for $k \geqslant 1$, let $\bar{\varphi}_k$ be a continuous function from $\mathbb{R}$ into $[0,1]$ such that $\bar{\varphi}_k(u) = 1$ for $u \leqslant \delta_k$, $\bar{\varphi}_k(u) = 0$ for $u \geqslant \delta_{k-1,1/2}$, and $\bar{\varphi}_k$ is linear for $\delta_k \leqslant u \leqslant \delta_{k-1,1/2}$. It follows from (3.14) that $\delta_k \in (\delta(1 - a\sigma^{\nu} \wedge 2^{-1}), \delta)$ for all k such that (3.12) holds (with some $a > 0$). Let us introduce the following function classes:

$$\mathcal{G}_k := \{\varphi_k \circ f \colon\ f \in \mathcal{F}_k\} \cup \{0\}, \quad k \geqslant 0,$$

and

$$\overline{\mathcal{G}}_k := \{\bar{\varphi}_k \circ f \colon\ f \in \mathcal{F}_k\} \cup \{0\}, \quad k \geqslant 1.$$

Then, for $k \geqslant 1$,

$$\sup_{g\in\mathcal{G}_k} Pg^2 \leqslant \sup_{f\in\mathcal{F}_k} F_f(\delta_k) \leqslant \sup_{f\in\mathcal{F}_k} F_f(\delta_{k-1,1/2}) \leqslant \varepsilon + \frac{C}{2}\sqrt{r_{k-1}\varepsilon\sigma} \wedge 1 \leqslant r_k$$

and

$$\sup_{g\in\bar{\mathcal{G}}_k} Pg^2 \leqslant \sup_{f\in\mathcal{F}_k} F_f(\delta_{k-1,1/2}) \leqslant \varepsilon + \frac{C}{2}\sqrt{r_{k-1}\varepsilon\sigma} \wedge 1 \leqslant r_k.$$

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

$$\begin{aligned} E^{(k)} := {}& \bigl\{\|P_n - P\|_{\mathcal{G}_{k-1}} \leqslant K_1\mathbb{E}\|P_n - P\|_{\mathcal{G}_{k-1}} + K_2\sqrt{r_{k-1}\varepsilon\sigma} + K_3\varepsilon\sigma\bigr\} \\ & \cap \bigl\{\|P_n - P\|_{\bar{\mathcal{G}}_k} \leqslant K_1\mathbb{E}\|P_n - P\|_{\bar{\mathcal{G}}_k} + K_2\sqrt{r_k\varepsilon\sigma} + K_3\varepsilon\sigma\bigr\}, \quad k \geqslant 1, \end{aligned}$$

By concentration inequalities of Talagrand [17,18] (see also Massart [14]), for some values of the numerical constants $K_1, K_2, K_3 > 0$,

$$\mathbb{P}\bigl(\bigl(E^{(k)}\bigr)^c\bigr) \leqslant 2\,\mathrm{e}^{-n\varepsilon\sigma/2}.$$

We set $E_0 = \Omega$,

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

Clearly,

$$\mathbb{P}\bigl(E_N^c\bigr) \leqslant 2N\,\mathrm{e}^{-n\varepsilon\sigma/2}. \tag{3.15}$$

We will prove by induction with respect to N the following statement:
For any N such that

$$N + 1 \leqslant \log_2 \frac{1}{\nu}\frac{\log_2(\varepsilon\sigma)^{-1}}{\log_2\sigma^{-1} + (4\nu)^{-1}}, \tag{3.16}$$

we have on the event E_N:

$$\text{(i)} \quad \forall f \in \mathcal{F}\ F_{n,f}(\delta) \leqslant \varepsilon \quad \Longrightarrow \quad f \in \mathcal{F}_N$$

and

$$\text{(ii)} \quad \sup_{f\in\mathcal{F}_k} F_{n,f}(\delta_k) \leqslant r_k, \quad 0 \leqslant k \leqslant N.$$

The statement holds for $N = 0$. We assume that it holds for some $N \geqslant 0$, such that $N + 1$ still satisfies condition (3.16). Then, on the event E_N,

$$\sup_{f\in\mathcal{F}_k} F_{n,f}(\delta_k) \leqslant r_k, \quad 0 \leqslant k \leqslant N,$$

and

$$\forall f \in \mathcal{F}\ F_{n,f}(\delta) \leqslant \varepsilon \quad \Longrightarrow \quad f \in \mathcal{F}_N.$$

Suppose that $F_{n,f}(\delta) \leqslant \varepsilon$ for some $f \in \mathcal{F}$. The induction assumptions imply that $f \in \mathcal{F}_N$ on the event E_N. Hence, on the event E_{N+1},

$$\begin{aligned} F_f(\delta_{N,1/2}) &\leqslant F_{n,f}(\delta_N) + \|P_n - P\|_{\mathcal{G}_N} \\ &\leqslant \varepsilon + K_1 \mathbb{E}\|P_n - P\|_{\mathcal{G}_N} + K_2\sqrt{r_N \varepsilon\sigma} + K_3 \varepsilon\sigma. \end{aligned} \tag{3.17}$$

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

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

where $\{\varepsilon_i\}$ is a sequence of i.i.d. Rademacher random variables. The symmetrization inequality yields

$$\mathbb{E}\|P_n - P\|_{\mathcal{G}_N} \leqslant 2\mathbb{E} I_{E_N} \mathbb{E}_\varepsilon \widehat{R}_n(\mathcal{G}_N) + 2\mathbb{E} I_{E_N^c} \mathbb{E}_\varepsilon \widehat{R}_n(\mathcal{G}_N). \tag{3.18}$$

Using the entropy inequalities for subgaussian processes (see van der Vaart and Wellner [20], Corollary 2.2.8), we get

$$\mathbb{E}_\varepsilon \widehat{R}_n(\mathcal{G}_N) \leqslant \frac{\text{const}}{\sqrt{n}} \int_0^{(2\sup_{g\in\mathcal{G}_N} P_n g^2)^{1/2}} H^{1/2}_{d_{P_n,2}}(\mathcal{G}_N; u)\, du. \tag{3.19}$$

By the induction assumptions, on the same event E_N

$$\sup_{g\in\mathcal{G}_N} P_n g^2 \leqslant \sup_{f\in\mathcal{F}_N} F_{n,f}(\delta_N) \leqslant r_N.$$

We use the bound on the Lipschitz constants of φ_{k-1} and $\bar{\varphi}_k$

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

to get

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

By the definition of the class $\mathcal{G}_N$,

$$H^{1/2}_{d_{P_n,2}}(\mathcal{G}_N; u) \leqslant \sqrt{\log\left(N_{d_{P_n,2}}\left(\mathcal{F}; \frac{\delta\sqrt{\varepsilon\sigma}u}{2\sqrt{r_N}}\right) + 1\right)}.$$

Since for $N \geqslant 1$,

$$\sqrt{\log(N+1)} \leqslant \sqrt{\log N + \log\left(1 + \frac{1}{N}\right)} \leqslant \sqrt{\log N + \frac{1}{N}} \leqslant \sqrt{\log N} + 1,$$

we get

$$H^{1/2}_{d_{P_n,2}}(\mathcal{G}_N; u) \leqslant H^{1/2}_{d_{P_n,2}}\left(\mathcal{F}; \frac{\delta\sqrt{\varepsilon\sigma}u}{2\sqrt{r_N}}\right) + 1. \tag{3.20}$$

Note that for $\varepsilon\sigma \geqslant \varepsilon_n^{\psi}(\delta)$ the inequality $\psi(\delta\sqrt{\varepsilon\sigma}/2)/(\delta\sqrt{n}) \leqslant \varepsilon\sigma$ holds. Recall also that $\varepsilon\sigma \geqslant \frac{2\log n}{n}$. It follows that, on the event E_N,

$$\begin{aligned}
\frac{1}{\sqrt{n}} \int\limits_0^{(2\sup_{g\in\mathcal{G}_N} P_n g^2)^{1/2}} H^{1/2}_{d_{P_n,2}}(\mathcal{G}_N; u)\, du &\leqslant \frac{1}{\sqrt{n}} \int\limits_0^{(2r_N)^{1/2}} \left[H^{1/2}_{d_{P_n,2}}\left(\mathcal{F}; \frac{\delta\sqrt{\varepsilon\sigma}u}{2\sqrt{r_N}}\right) + 1\right] du \\
&\leqslant \frac{1}{\sqrt{n}} \frac{2\sqrt{r_N}}{\delta\sqrt{\varepsilon\sigma}} \int\limits_0^{\delta\sqrt{\varepsilon\sigma/2}} H^{1/2}_{d_{P_n,2}}(\mathcal{F}; v)\, dv + \sqrt{\frac{2r_N}{n}} \leqslant \frac{1}{\sqrt{n}} \frac{2\sqrt{r_N}}{\delta\sqrt{\varepsilon\sigma}} D_n \psi\left(\frac{\delta\sqrt{\varepsilon\sigma}}{\sqrt{2}}\right) \\
&\quad + \sqrt{2r_N\varepsilon\sigma} \leqslant \frac{2D_n\sqrt{r_N}}{\sqrt{\varepsilon\sigma}}\varepsilon\sigma + \sqrt{2r_N\varepsilon\sigma} \leqslant 2(D_n+1)\sqrt{r_N\varepsilon\sigma}.
\end{aligned} \tag{3.21}$$

Now (3.19) and (3.21) imply that on the same event

$$\mathbb{E}_\varepsilon \widehat{R}_n(\mathcal{G}_N) \leqslant \mathrm{const}(1+D_n)\sqrt{r_N\varepsilon\sigma}. \tag{3.22}$$

Since $\mathbb{E}_\varepsilon \widehat{R}_n(\mathcal{G}_{N+1}) \leqslant 1$, we conclude from (3.15), (3.18) and (3.22) that

$$\begin{aligned}
\mathbb{E}\|P_n - P\|_{\mathcal{G}_N} &\leqslant \mathrm{const}(1+\mathbb{E}D_n)\sqrt{r_N\varepsilon\sigma} + 2\mathbb{P}(E_N^c) \\
&\leqslant \mathrm{const}(1+\mathbb{E}D_n)\sqrt{r_N\varepsilon\sigma} + 4N\,\mathrm{e}^{-n\varepsilon\sigma/2}.
\end{aligned}$$

By condition (3.16) and the fact that $\varepsilon\sigma \geqslant 2\log n/n$, we have

$$4N\,\mathrm{e}^{-n\varepsilon\sigma/2} \leqslant 4\log_2\bigl(4\log_2(\varepsilon\sigma)^{-1}\bigr)\,\mathrm{e}^{-n\varepsilon\sigma/2} \leqslant \mathrm{const}\cdot\varepsilon\sigma$$

(we assume here that $\varepsilon\sigma \leqslant \kappa$ for some $\kappa \in (0,1)$; note that if $\varepsilon\sigma > \kappa$, then also $\varepsilon\sigma^\lambda > \kappa$ and the bounds of the theorem become trivial with sufficiently large B so that $B_\lambda(\sigma)\varepsilon > 1$). Therefore (note that $r_N \geqslant \varepsilon\sigma$),

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

By (3.17), on the event E_{N+1}

$$F_f(\delta_{N,1/2}) \leqslant \varepsilon + \mathrm{const}(1+\mathbb{E}D_n)\sqrt{r_N\varepsilon\sigma}. \tag{3.23}$$

Choosing the constant c and thus also the constant $C = c(1+\mathbb{E}D_n)$ in the recurrent relationship (3.8) properly, we ensure that on the event E_{N+1}

$$F_f(\delta_{N,1/2}) \leqslant \varepsilon + \frac{C}{2}\sqrt{r_N\varepsilon\sigma}.$$

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

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

$$\begin{aligned}\sup_{f\in\mathcal{F}_{N+1}} F_{n,f}(\delta_{N+1}) &\leqslant \sup_{f\in\mathcal{F}_{N+1}} F_f(\delta_{N,1/2}) + \|P_n - P\|_{\overline{\mathcal{G}}_{N+1}}\\ &\leqslant \varepsilon + \frac{C}{2}\sqrt{r_N\varepsilon\sigma} + K_1\mathbb{E}\|P_n - P\|_{\overline{\mathcal{G}}_{N+1}} + K_2\sqrt{r_{N+1}\varepsilon\sigma} + K_3\varepsilon\sigma. \qquad (3.24)\end{aligned}$$

Using the symmetrization inequality, we get

$$\mathbb{E}\|P_n - P\|_{\overline{\mathcal{G}}_{N+1}} \leqslant 2\mathbb{E}I_{E_N}\mathbb{E}_\varepsilon \widehat{R}_n(\overline{\mathcal{G}}_{N+1}) + 2\mathbb{E}I_{E_N^c}\mathbb{E}_\varepsilon \widehat{R}_n(\overline{\mathcal{G}}_{N+1}). \qquad (3.25)$$

Similarly to (3.19)

$$\mathbb{E}_\varepsilon R_n(\overline{\mathcal{G}}_{N+1}) \leqslant \frac{\text{const}}{\sqrt{n}} \int_0^{(2\sup_{g\in\overline{\mathcal{G}}_{N+1}} P_n g^2)^{1/2}} H^{1/2}_{d_{P_n,2}}(\overline{\mathcal{G}}_{N+1}; u)\,du. \qquad (3.26)$$

The induction assumption implies that on the event E_{N+1}

$$\sup_{g\in\overline{\mathcal{G}}_{N+1}} P_n g^2 \leqslant \sup_{f\in\mathcal{F}_N} F_{n,f}(\delta_{N,1/2}) \leqslant r_N.$$

Since the Lipschitz constant of $\bar\varphi_k$ is bounded by $(2/\delta)\sqrt{r_{k-1}/(\varepsilon\sigma)}$, we have

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

Similarly to (3.21), we have on the event E_{N+1},

$$\begin{aligned} \frac{1}{\sqrt{n}} \int_0^{(2\sup_{g\in\overline{\mathcal{G}}_{N+1}} P_n g^2)^{1/2}} H^{1/2}_{d_{P_n,2}}(\overline{\mathcal{G}}_{N+1}; u)\,du &\leqslant \frac{1}{\sqrt{n}} \int_0^{(2r_N)^{1/2}} \left[H^{1/2}_{d_{P_n,2}}\left(\mathcal{F}; \frac{\delta\sqrt{\varepsilon\sigma}u}{2\sqrt{r_N}}\right) + 1\right] du\\ &\leqslant \frac{1}{\sqrt{n}}\frac{2\sqrt{r_N}}{\delta\sqrt{\varepsilon\sigma}} \int_0^{\delta\sqrt{\varepsilon\sigma/2}} H^{1/2}_{d_{P_n,2}}(\mathcal{F}; v)\,dv + \sqrt{\frac{2r_N}{n}}\\ &\leqslant \frac{1}{\sqrt{n}}\frac{2\sqrt{r_N}}{\delta\sqrt{\varepsilon\sigma}} D_n\psi\left(\frac{\delta\sqrt{\varepsilon\sigma}}{\sqrt{2}}\right) + \sqrt{2r_N\varepsilon\sigma}\\ &\leqslant \frac{2D_n\sqrt{r_N}}{\sqrt{\varepsilon\sigma}}\varepsilon\sigma + \sqrt{2r_N\varepsilon\sigma} \leqslant 2(D_n+1)\sqrt{r_N\varepsilon\sigma}. \qquad (3.27)\end{aligned}$$

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

$$\sup_{f\in\mathcal{F}_{N+1}} F_{n,f}(\delta_{N+1}) \leqslant \varepsilon + \frac{C}{2}\sqrt{r_N\varepsilon\sigma} + \text{const}(1 + \mathbb{E}D_n)\sqrt{r_N\varepsilon\sigma}. \qquad (3.28)$$

Properly choosing the constant $c > 0$ (and, thus, C in the recurrent relationship (3.8)), we get on the event E_{N+1}

$$\sup_{f\in\mathcal{F}_{N+1}} F_{n,f}(\delta_{N+1}) \leqslant \bigl(\varepsilon + C\sqrt{r_N\varepsilon\sigma}\,\bigr)\vee 1 = r_{N+1},$$

which completes the proof of (ii) and of the induction step. Recall that, in particular, it means the following: on the event E_N, the assumption $F_{n,f}(\delta)\leqslant\varepsilon$ implies that $f\in\mathcal{F}_N$, and hence $F_f(\delta_N)\leqslant r_N$.

To complete the proof of the first bound of the theorem, it's enough to recall that N can be choosen to satisfy the inequalities

$$\log_2\frac{\log_2(\varepsilon\sigma)^{-1}}{1+(1-\lambda)\log_2\sigma^{-1}} - 1 \leqslant N \leqslant \log_2\frac{1}{\nu}\frac{\log_2(\varepsilon\sigma)^{-1}}{\log_2\sigma^{-1}+(4\nu)^{-1}} - 2,$$

which implies that $r_{N+1}\leqslant\varepsilon(1+a\sigma^{\lambda})$ for some constant a. We also have (since $4\nu < 1$)

$$\log_2\frac{1}{\nu}\frac{\log_2(\varepsilon\sigma)^{-1}}{\log_2\sigma^{-1}+(4\nu)^{-1}} - 2 \leqslant \log_2\frac{1}{\nu}\frac{\log_2(\varepsilon\sigma)^{-1}}{\log_2\sigma^{-1}+1},$$

which is bounded by

$$D\left(\log_2\frac{\log_2(\varepsilon\sigma)^{-1}}{\log_2\sigma^{-1}+1}\vee 1\right)$$

with some constant D.

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

4. Applications to learning problems and examples

In this section, we deal with a binary classification problem, i.e., S is replaced by $S\times\{-1,1\}$ and f is replaced by $S\times\{-1,1\}\ni(x,y)\mapsto yf(x)$ (see the introduction). Theorems 1 and 2 of Section 2 immediately imply the following result about the behavior of the generalization error.

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

$$\mathbb{P}\bigl\{\exists f\in\mathcal{F}\colon\ P\bigl\{(x,y)\colon\ yf(x)\leqslant 0\bigr\}\geqslant \bar A\hat\varepsilon_n^{\psi}(f;t)\bigr\}\leqslant B\log_2\log_2\frac{n}{t}\exp\left\{-\frac{t}{2}\right\}. \tag{4.1}$$

Moreover, if (2.6) *holds, then with probability* 1

$$\limsup_{n\to\infty}\sup_{f\in\mathcal{F}}\frac{P\{(x,y)\colon\ yf(x)\leqslant 0\}}{\hat\varepsilon_n^{\psi}(f)} < +\infty. \tag{4.2}$$

Under the conditions of Theorem 2,

$$\limsup_{n\to\infty}\sup_{f\in\mathcal{F}}\frac{P\{(x,y)\colon\ yf(x)\leqslant 0\}}{\hat\varepsilon_n^{\phi}(f)}\leqslant 1. \tag{4.3}$$

If $\sup_{f\in\mathcal{F}} P\{(x,y)\colon\ yf(x)\leqslant 0\}>0$, *then the* $\limsup$ *in the last equation is equal to* 1.

Thus, for any classifier $\hat f\in\mathcal{F}$, with probability 1

$$P\{y\hat f(x)\leqslant 0\}\leqslant (1+\mathrm{o}(1))\hat\varepsilon_n^{\phi}(\hat f).$$

Let $\alpha\in(0,2)$ and $\psi(x)\equiv x^{1-\alpha/2}$. Let $\gamma:=\frac{2\alpha}{\alpha+2}$. As in the introduction, we define γ-margins of a function f as follows:

$$\delta_n(\gamma;f):=\sup\{\delta>0\colon\ \delta^{\gamma}F_f(\delta)\leqslant n^{-1+\gamma/2}\},$$

$$\hat\delta_n(\gamma;f):=\sup\{\delta>0\colon\ \delta^{\gamma}F_{n,f}(\delta)\leqslant n^{-1+\gamma/2}\}.$$

Note that Koltchinskii and Panchenko [9] used slightly different (truncated) quantities: the suprema there was over the set $\delta\in(0,1)$. We will use for these quantities the notations $\delta_n^t(\gamma;f)$ and $\hat\delta_n^t(\gamma;f)$. It's easy to see that

$$\delta_n^t(\gamma;f)=\delta_n^{\psi}(f;n^{\gamma/2}),\qquad \hat\delta_n^t(\gamma;f)=\hat\delta_n^{\psi}(f;n^{\gamma/2}).$$

Theorem 1 immediately implies (recall (2.3) and the definition of $\varepsilon_{n,\gamma}$ from the introduction) that if for some $\alpha\in(0,2)$ and $D_n>0$, $\mathbb{E}D_n<\infty$

$$H_{d_{P_n,2}}(\mathcal{F};u)\leqslant D_n^2u^{-\alpha},\quad u>0\text{ a.s.,}\tag{4.4}$$

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

$$\mathbb{P}\left\{\exists f\in\mathcal{F}\colon\ P\{(x,y)\colon\ yf(x)\leqslant 0\}\geqslant\frac{\bar A}{n^{1-\gamma/2}\hat\delta_n^t(\gamma;f)^{\gamma}}\right\}$$
$$\leqslant B\log_2\log_2 n\exp\left\{-\frac{n^{\gamma/2}}{2}\right\}\tag{4.5}$$

(Koltchinskii and Panchenko [9]). Theorem 2 shows that as soon as $\gamma>\frac{2\alpha}{2+\alpha}$

$$\limsup_{n\to\infty}\sup_{f\in\mathcal{F}} n^{1-\gamma/2}\hat\delta_n(\gamma;f)^{\gamma}P\{(x,y)\colon\ yf(x)\leqslant 0\}\leqslant 1\quad\text{a.s.}\tag{4.6}$$

In fact, Theorems 1 and 2 imply the following corollary about the asymptotic behavior of γ-margins (and the same is true for their truncated versions).

COROLLARY 3. – *If the condition* (4.4) *holds with some* $\alpha\in(0,2)$, *then for all* $\gamma\geqslant\frac{2\alpha}{2+\alpha}$

$$\limsup_{n}\sup_{f\in\mathcal{F}}\frac{\hat\delta_n(\gamma;f)}{\delta_n(\gamma,f)}<+\infty\tag{4.7}$$

and

$$\limsup_{n}\sup_{f\in\mathcal{F}}\frac{\delta_n(\gamma;f)}{\hat\delta_n(\gamma,f)}<+\infty.\tag{4.8}$$

Moreover, for all $\gamma > \frac{2\alpha}{2+\alpha}$

$$\mathbb{P}\left\{\sup_{f\in\mathcal{F}}\left|\frac{\hat{\delta}_n(\gamma;f)}{\delta_n(\gamma;f)}-1\right|\to 0 \textit{ as } n\to\infty\right\}=1. \tag{4.9}$$

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

$$\delta_n^{\psi}(\varepsilon)=\frac{\mathrm{e}^{1-n\varepsilon}}{\sqrt{\varepsilon}},\quad \varepsilon\geqslant n^{-1}.$$

If we define

$$\hat{\varepsilon}_n^{VC}(f;t):=\inf\left\{\varepsilon\geqslant\frac{t}{n}\colon\ P_n\left\{f\leqslant\frac{\mathrm{e}^{1-n\varepsilon}}{\sqrt{\varepsilon}}\right\}\leqslant\varepsilon\right\}, \tag{4.10}$$

then under the condition

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

with some $D_n=D_n(X_1,\dots,X_n), \mathbb{E}D_n<+\infty$ (which holds, for instance, if $\mathcal{F}$ is a VC-subgraph class), we get from Theorem 1 that with some numerical constants $A,B>0$ for all $t\geqslant 2\log n$

$$\mathbb{P}\left\{\exists f\in\mathcal{F}\colon\ P\{(x,y)\colon\ yf(x)\leqslant 0\}\geqslant\bar{A}\hat{\varepsilon}_n^{VC}(f;t)\right\}\leqslant B\log_2\log_2\frac{n}{t}\exp\left\{-\frac{t}{2}\right\},$$

where $\bar{A}:=A(1+\mathbb{E}D_n)^2$. Now Theorem 2 adds to this that as soon as

$$\frac{\phi(x)}{x\sqrt{\log(x^{-1})}}\to\infty\quad\text{as } x\to 0$$

we have $\hat{\varepsilon}_n^{\phi}(f)/\varepsilon_n^{\phi}(f)\to 1$ uniformly in $f\in\mathcal{F}$ a.s.

We construct below examples that show the sharpness of our main results. They are close to some examples in Koltchinskii and Panchenko [9] (earlier, similar examples in the context of CLT in Banach spaces were looked at by Ledoux and Talagrand [13]). Let S be the space of all sequences converging to 0 and let $\mathcal{F}$ be the set of all coordinate functions: $\mathcal{F}:=\{f_k\colon\ k\geqslant 1\},\ f_k(x)=x_k,\ x=\{x_k\}\in S$. Let

$$X_n:=\left\{\frac{\varepsilon_{k,n}}{\lambda_k}\right\}_{k\geqslant 1},$$

where $\{\varepsilon_{k,n}\colon\ k\geqslant 1,\ n\geqslant 1\}$ are i.i.d. Rademacher random variables and

$$\lambda_k:=\frac{1}{\varphi^{-1}(\sqrt{\log(k+1)})},\quad k\geqslant 1,$$

φ being a nonincreasing positive function with $\varphi(\delta) \to +\infty$ as $\delta \to 0$. We assume that with some constant $K > 0$

$$\int_0^x \varphi(u)\,du \leqslant Kx\varphi(x), \quad x \geqslant 0,$$

and that for any $\varepsilon \in (0,1)$ and, for all large enough x, $\varphi^{-1}(x) \leqslant \varepsilon\varphi^{-1}(\varepsilon x)$ (for instance, $\varphi(x) = x^{-\alpha/2}$ for $\alpha \in (0,2)$, or $\varphi(x) = \sqrt{\log(\mathrm{e}/x)} \vee 1$ are functions of this type).

Finally, assume that the sequence of labels $\{Y_n\}$ is a Rademacher sequence independent of $\{X_n\}$. Clearly, in this case the generalization error of any classifier $f \in \mathcal{F}$ is equal to $1/2$.

PROPOSITION 1. – *The condition* (2.1) *holds for the sequence* $\{X_n\}$ *with* $D_n = D$, D *being a numerical constant and* $\psi(x) := x\varphi(x)$, $\psi \in \Psi$. *The condition* (2.6) *also holds for the class* $\mathcal{F}$. *For any* $\phi \in \Psi$ *such that*

$$\phi(x) = \mathrm{o}\big(\psi(x)\big) \quad \text{as } x \to 0,$$

we have

$$\limsup_n \sup_{f\in\mathcal{F}} \frac{\varepsilon_n^{\phi}(f)}{\hat{\varepsilon}_n^{\phi}(f)} = +\infty \tag{4.11}$$

and

$$\limsup_n \sup_{f\in\mathcal{F}} \frac{\hat{\varepsilon}_n^{\phi}(f)}{\varepsilon_n^{\phi}(f)} = 2. \tag{4.12}$$

In addition,

$$\limsup_n \sup_{f\in\mathcal{F}} \frac{P\{(x,y):\ yf(x) \leqslant 0\}}{\hat{\varepsilon}_n^{\phi}(f)} = +\infty. \tag{4.13}$$

Moreover, there exists $A > 1$ *such that*

$$\liminf_n \sup_{f\in\mathcal{F}} \frac{\varepsilon_n^{\psi}(f)}{\hat{\varepsilon}_n^{\psi}(f)} \geqslant A \tag{4.14}$$

and

$$\liminf_n \sup_{f\in\mathcal{F}} \frac{\hat{\varepsilon}_n^{\psi}(f)}{\varepsilon_n^{\psi}(f)} \geqslant A. \tag{4.15}$$

It follows that

$$\liminf_n \sup_{f\in\mathcal{F}} \frac{P\{(x,y):\ yf(x) \leqslant 0\}}{\hat{\varepsilon}_n^{\psi}(f)} \geqslant A. \tag{4.16}$$

In particular, it means that if $\hat{f}$ is a classifier that minimizes the bound $\hat{\varepsilon}_n^{\phi}(f)$ on the class $\mathcal{F}$ (a natural choice from the point of view of "large margin" approach) and

$\frac{\phi(x)}{\psi(x)} \to 0$ as $x \to 0$, then

$$\frac{P\{(x,y)\colon\ y\hat{f}(x) \leqslant 0\}}{\hat{\varepsilon}_n^{\phi}(\hat{f})} \to \infty \quad \text{a.s.},$$

i.e., in this case the margin type bound $\hat{\varepsilon}_n^{\phi}(\hat{f})$ can become way too optimistic. To avoid this, the definition of the margin type bounds is to be related to the complexity of the class (the condition $\frac{\phi(x)}{\psi(x)} \to \infty$ as $x \to 0$ guarantees this).

Proof of Proposition 1. – First note that since the condition (2.6) holds, the ψ- and ϕ-bounds can and will be replaced by their truncated versions with $t_n \asymp \log n$ (see the argument at the beginning of the proof of Theorem 2). Next, for

$$k \geqslant \exp\{\varphi^2(\varepsilon)\} - 1 =: N(\varepsilon)$$

we have $\|f_k\|_\infty \leqslant \varepsilon$. This immediately implies $\|f_k\|_{L_2(P_n)} \leqslant \varepsilon$, which means

$$N_{d_{P_n},2}(\mathcal{F};\varepsilon) \leqslant \exp\{\varphi^2(\varepsilon)\},$$

and the condition (2.1) follows. Next, it's easy to see that for $\delta < \lambda_k^{-1}$, $F_{f_k}(\delta) = 1/2$. It means that for all $k < N(\delta_n^{\phi}(1/2))$ we have $F_{f_k}(\delta_n^{\phi}(1/2)) = 1/2$, which implies $\varepsilon_n^{\phi}(f_k;t_n) = 1/2$. For $k < N(\delta_n^{\phi}(1/2))$, this yields

$$\mathbb{P}\{\varepsilon_n^{\phi}(f_k;t_n) < A\hat{\varepsilon}_n^{\phi}(f_k,t_n)\} = \mathbb{P}\left\{\hat{\varepsilon}_n^{\phi}(f_k,t_n) > \frac{1}{2A}\right\},$$

which for $k < N(\delta_n^{\phi}(\frac{1}{2A}))$ (or, equivalently, $\lambda_k^{-1} > \delta_n^{\phi}(\frac{1}{2A})$) is equal to

$$\mathbb{P}\left\{F_{n,f_k}\left(\delta_n^{\phi}\left(\frac{1}{2A}\right)\right) > \frac{1}{2A}\right\} = \mathbb{P}\left\{\sum_{j=1}^{n} I_{\{\varepsilon_{k,j}=-1\}} > \left(\frac{1}{2}-\delta\right)n\right\},$$

where $\delta = 0.5(1-A^{-1})$. Using well known computations for binomial probabilities (based on Stirling's formula), the last probability can be bounded from above by

$$1 - cn^{-1/2}\exp\{-4n\delta^2\}$$

(see Koltchinskii and Panchenko [9]). This implies that

$$\mathbb{P}\{\varepsilon_n^{\phi}(f_k;t_n) < A\hat{\varepsilon}_n^{\phi}(f_k,t_n)\} \leqslant 1 - cn^{-1/2}\exp\{-n(1-A^{-1})^2\}$$

for all $k < N(\delta_n^{\phi}(\frac{1}{2A}))$.

Let $\bar{\varphi}(x) := \frac{\phi(x)}{x}$. If $\phi(x) = \mathrm{o}(\psi(x))$ as $x \to 0$, we have

$$\bar{\varphi}(x) \leqslant \sigma\varphi(x)$$

for all $\sigma > 0$ and small enough $x > 0$. If $\phi \equiv \psi$, then the above bound holds with $\sigma = 1$. Then a simple argument shows that for all large enough n

$$\bar{\varphi}^{-1}\left(\sqrt{\frac{n}{2A}}\right) \leqslant \varphi^{-1}\left(\frac{1}{\sigma}\sqrt{\frac{n}{2A}}\right).$$

By the assumptions on φ, we get

$$\delta_n^{\phi}\left(\frac{1}{2A}\right) = \frac{\bar{\varphi}^{-1}(\sqrt{n/(2A)})}{\sqrt{1/(2A)}} \leqslant \frac{\varphi^{-1}((1/\sigma)\sqrt{n/(2A)})}{\sqrt{1/(2A)}} \leqslant \varphi^{-1}\left(\frac{1}{2A\sigma}\sqrt{n}\right).$$

Therefore

$$N\left(\delta_n^{\phi}\left(\frac{1}{2A}\right)\right) \geqslant \exp\left\{\varphi^2\left(\varphi^{-1}\left(\frac{1}{2A\sigma}\sqrt{n}\right)\right)\right\} = \exp\left\{\frac{n}{4A^2\sigma^2}\right\} := K_n.$$

By independence of the components of X_n, we get

$$\begin{aligned}
&\mathbb{P}\{\forall k \leqslant K_n\colon\ \varepsilon_n^{\phi}(f_k; t_n) < A\hat{\varepsilon}_n^{\phi}(f_k, t_n)\}\\
&\quad = \prod_{k=1}^{K_n} \mathbb{P}\{\varepsilon_n^{\phi}(f_k; t_n) < A\hat{\varepsilon}_n^{\phi}(f_k, t_n)\} \leqslant \left(1 - cn^{-1/2}\exp\{-n(1-A^{-1})^2\}\right)^{K_n}\\
&\quad \leqslant \exp\left\{-cn^{-1/2}\exp\left\{-n(1-A^{-1})^2 + \frac{n}{4A^2\sigma^2}\right\}\right\} = \mathrm{o}(n^{-2}) \quad \text{as } n \to \infty,
\end{aligned}$$

provided that

$$\frac{1}{4A^2\sigma^2} > \left(1 - \frac{1}{A}\right)^2.$$

If $\sigma < 1/2$, it is satisfied for all $1 < A < \frac{1}{2\sigma}$, and for $\sigma = 1$ it's true if A is close enough to 1. In the case when $\phi(x) = \mathrm{o}(\psi(x))$ σ can be taken arbitrarily small. Borel–Cantelli lemma shows in this case that for any A with probability 1

$$\sup_{f\in\mathcal{F}} \frac{\varepsilon_n^{\phi}(f; t_n)}{\hat{\varepsilon}_n^{\phi}(f; t_n)}$$

is eventually (for all large n) larger than A. In the case when $\phi = \psi$ the same conclusion holds for some $A > 1$.

The proof of the remaining statements is quite similar. One just has to take into account that For $k < N(\delta_n^{\phi}(1/2))$,

$$\begin{aligned}
&\mathbb{P}\{\hat{\varepsilon}_n^{\phi}(f_k; t_n) < A\varepsilon_n^{\phi}(f_k, t_n)\} = \mathbb{P}\left\{\hat{\varepsilon}_n^{\phi}(f_k, t_n) < \frac{A}{2}\right\} = \mathbb{P}\left\{F_{n,f_k}\left(\delta_n^{\phi}\left(\frac{A}{2}\right)\right) < \frac{A}{2}\right\}\\
&\quad = \mathbb{P}\left\{\sum_{j=1}^{n} I_{\{\varepsilon_{k,j}=-1\}} < \left(\frac{A}{2}\right)n\right\} = \mathbb{P}\left\{\sum_{j=1}^{n} I_{\{\varepsilon_{k,j}=+1\}} \geqslant \left(\frac{1}{2} - \delta\right)n\right\},
\end{aligned}$$

where $\delta = 0.5(A-1)$ and $A \in (1,2)$ and continue as in the previous part of the proof (one should also take into account that

$$\frac{\hat{\varepsilon}_n^{\phi}(f)}{\varepsilon_n^{\phi}(f)} \leqslant 2$$

since $\varepsilon_n^{\phi}(f) = 1/2$ and $\hat{\varepsilon}_n^{\phi}(f) \leqslant 1$). □

The following result is a special case of Proposition 1 (and its proof is quite similar to the proof of this proposition; alternatively, the result of this type can be deduced directly from Proposition 1 using (1.5) and (1.7)). It shows that the condition $\gamma > \frac{2\alpha}{2+\alpha}$ is sharp for the uniform convergence of the ratios of γ margins to 1 while the condition $\gamma \geqslant \frac{2\alpha}{2+\alpha}$ is sharp for the boundedness of the ratios. Namely, let

$$X_n := \left\{\varepsilon_{k,n}\left(2\log(k+1)\right)^{-1/\alpha}\right\}_{k\geqslant 1},$$

where $\alpha \in (0,2)$ and $\{\varepsilon_{k,n}\colon\ k \geqslant 1,\ n \geqslant 1\}$ are i.i.d. Rademacher random variables. As before, S is the space of all sequences converging to 0 and $\mathcal{F} := \{f_k\colon\ k \geqslant 1\},\ f_k(x) = x_k,\ x = \{x_k\} \in S$. The sequence $\{Y_n\}$ of labels is also the same as before, so the generalization error of any classifier $f \in \mathcal{F}$ is equal to $1/2$.

PROPOSITION 2. – *The condition* (4.4) *holds for the sequence* $\{X_n\}$ *with* $D_n = D$, D *being a numerical constant. For all* $\gamma < \frac{2\alpha}{2+\alpha}$,

$$\lim_n \sup_{f\in\mathcal{F}} \frac{\hat{\delta}_n(\gamma; f)}{\delta_n(\gamma, f)} = +\infty \tag{4.17}$$

and

$$\lim_n \sup_{f\in\mathcal{F}} \frac{\delta_n(\gamma; f)}{\hat{\delta}_n(\gamma, f)} = 2^{1/\gamma}. \tag{4.18}$$

It implies that

$$\lim_n \sup_{f\in\mathcal{F}} n^{1-\gamma/2}\hat{\delta}_n(\gamma; f)^{\gamma} P\left\{(x,y)\colon\ yf(x) \leqslant 0\right\} = +\infty. \tag{4.19}$$

Let z_α *denote the solution of the equation*

$$2^{-\alpha/4}z^{\alpha/2} + z^{2\alpha/(\alpha+2)} = 1.$$

Then $z_\alpha < 1$ *and for* $\gamma = \frac{2\alpha}{2+\alpha}$

$$\liminf_n \sup_{f\in\mathcal{F}} \frac{\hat{\delta}_n(\gamma; f)}{\delta_n(\gamma, f)} \geqslant \frac{1}{z_\alpha} \tag{4.20}$$

and

$$\liminf_n \sup_{f\in\mathcal{F}} n^{1-\gamma/2}\hat{\delta}_n(\gamma; f)^{\gamma} P\left\{(x,y)\colon\ yf(x) \leqslant 0\right\} \geqslant \frac{1}{z_\alpha}. \tag{4.21}$$

We also have

$$\liminf_n \sup_{f\in\mathcal{F}} \frac{\delta_n(\gamma;f)}{\hat{\delta}_n(\gamma,f)} \geqslant \bigl(1-2^{-1-\alpha/4}\bigr)^{-(2+\alpha)/(2\alpha)}. \tag{4.22}$$

Finally, we present a proposition that shows the sharpness of the bound of Theorem 3.

PROPOSITION 3. – *Under the conditions of Proposition* 1, *there exist* $c(\sigma)>1$, $\beta_1>0$ *and* $\beta_2(\sigma)>0$, $\sigma\in(0,1]$ *sucht that* $c(\sigma)\downarrow 1$, $\beta_2(\sigma)\downarrow 0$ *as* $\sigma\downarrow 0$, *and for large enough n*

$$\mathbb{P}\biggl\{\exists f\in\mathcal{F}\ \exists\delta\leqslant\delta_n^\psi\biggl(\frac{t}{n}\biggr):\ F_f\biggl(\frac{\delta}{c(\sigma)}\biggr)\geqslant c(\sigma)\biggl[F_{n,f}(\delta)\vee\frac{1}{\sigma}\varepsilon_n^\psi(\delta)\biggr]\biggr\}$$
$$\geqslant 1-\exp\bigl\{-\beta_1\,\mathrm{e}^{n\beta_2(\sigma)}\bigr\} \tag{4.23}$$

and

$$\mathbb{P}\biggl\{\exists f\in\mathcal{F}\ \exists\delta\leqslant\delta_n^\psi\biggl(\frac{t}{n}\biggr):\ F_{n,f}\biggl(\frac{\delta}{c(\sigma)}\biggr)\geqslant c(\sigma)\biggl[F_f(\delta)\vee\frac{1}{\sigma}\varepsilon_n^\psi(\delta)\biggr]\biggr\}$$
$$\geqslant 1-\exp\bigl\{-\beta_1\,\mathrm{e}^{n\beta_2(\sigma)}\bigr\}. \tag{4.24}$$

Proof. – We use the notations of Proposition 1. If

$$k<N(\delta):=\exp\bigl\{\varphi^2(\delta)\bigr\}-1,$$

then $F_{f_k}(\delta)=1/2$ and $F_{f_k}(\delta/c)=1/2$ for all $c>1$. Therefore we have

$$\mathbb{P}\biggl\{F_{n,f_k}(\delta)\geqslant\frac{1}{c}F_{f_k}\biggl(\frac{\delta}{c}\biggr)\biggr\}=\mathbb{P}\biggl\{F_{n,f_k}(\delta)\geqslant\frac{1}{2c}\biggr\}$$
$$=\mathbb{P}\biggl\{\sum_{j=1}^n I_{\{\varepsilon_{k,j}=-1\}}>\biggl(\frac{1}{2}-\tau\biggr)n\biggr\}\leqslant 1-\beta_1 n^{-1/2}\exp\bigl\{-4n\tau^2\bigr\},$$

where $\tau=\frac{1}{2}(1-c^{-1})$ and $\beta_1>0$ is a constant. Hence for $K<N(\delta)$,

$$\mathbb{P}\biggl\{F_{n,f_k}(\delta)\geqslant\frac{1}{c}F_{f_k}\biggl(\frac{\delta}{c}\biggr),k=1,\dots,K\biggr\}\leqslant\bigl(1-\beta_1 n^{-1/2}\exp\bigl\{-4n\tau^2\bigr\}\bigr)^K$$
$$\leqslant\exp\bigl\{-\beta_1 n^{-1/2}\exp\bigl\{-n\bigl(1-c^{-1}\bigr)^2+\log K\bigr\}\bigr\}.$$

If $\frac{1}{\sigma}\varepsilon_n^\psi(\delta)<\frac{1}{2c}$, or equivalently $\delta>\delta_n^\psi(\sigma/(2c))$, then

$$\mathbb{P}\biggl\{F_{n,f_k}(\delta)\vee\frac{1}{\sigma}\varepsilon_n^\psi(\delta)\geqslant\frac{1}{c}F_{f_k}\biggl(\frac{\delta}{c}\biggr),\ k=1,\dots,K\biggr\}$$
$$\leqslant\exp\bigl\{-\beta_1 n^{-1/2}\exp\bigl\{-n\bigl(1-c^{-1}\bigr)^2+\log K\bigr\}\bigr\}.$$

By the definition of δ_n^ψ,

$$\delta_n^\psi\bigl(\sigma/(2c)\bigr)=\sqrt{(2c)/\sigma}\,\varphi^{-1}\bigl(\sqrt{\sigma/(2c)n}\,\bigr).$$

The condition on φ^{-1} implies (since $\sqrt{\sigma/(2c)}\leqslant 1$) that

$$\sqrt{(2c)/\sigma}\,\varphi^{-1}\bigl(\sqrt{\sigma/(2c)n}\,\bigr)\leqslant\varphi^{-1}\bigl(\sigma/(2c)\sqrt{n}\,\bigr).$$

Hence

$$\varphi^2\bigl(\delta_n^{\psi}(\sigma/(2c))\bigr) = \varphi^2\bigl(\sqrt{(2c)/\sigma}\,\varphi^{-1}\bigl(\sqrt{\sigma/(2c)n}\,\bigr)\bigr) \geqslant \varphi\bigl(\varphi^{-1}\bigl(\sigma/(2c)\sqrt{n}\,\bigr)\bigr)^2$$
$$= \sigma^2/\bigl(4c^2\bigr)n.$$

Then it is easy to see that, for large enough n and for some $\delta \in (\delta_n^{\psi}(\sigma/(2c)), \delta_n^{\psi}(t/n))$,

$$\log N(\delta) \geqslant \frac{\sigma^2}{8c^2}n.$$

Therefore, we can choose $\log K$ of the order $\sigma^2/(8c^2)n$ to get

$$\mathbb{P}\biggl\{F_{n,f_k}(\delta) \vee \frac{1}{\sigma}\varepsilon_n^{\psi}(\delta) \geqslant \frac{1}{c}F_{f_k}\biggl(\frac{\delta}{c}\biggr),\ k = 1, \ldots, K\biggr\}$$
$$\leqslant \exp\bigl\{-\beta_1 n^{-1/2}\exp\bigl\{-n\bigl(1-c^{-1}\bigr)^2 + \sigma^2/\bigl(8c^2\bigr)n\bigr\}\bigr\}.$$

If now $c = c(\sigma) \downarrow 1, \sigma \downarrow 0$ is such that

$$\frac{\sigma^2}{8c^2} > \biggl(1 - \frac{1}{c}\biggr)^2,$$

then with some choice of $\beta_2(\sigma),\ \beta_2(\sigma) \downarrow 0$ as $\sigma \downarrow 0$, we have

$$\mathbb{P}\biggl\{\forall f \in \mathcal{F}\ \forall\delta \leqslant \delta_n^{\psi}\biggl(\frac{t}{n}\biggr):\ F_f\biggl(\frac{\delta}{c(\sigma)}\biggr) \geqslant c(\sigma)\biggl[F_{n,f}(\delta) \vee \frac{1}{\sigma}\varepsilon_n^{\psi}(\delta)\biggr]\biggr\}$$
$$\leqslant \exp\bigl\{-\beta_1\,\mathrm{e}^{n\beta_2(\sigma)}\bigr\}, \tag{4.25}$$

which implies the first inequality of the proposition.

To prove the second inequality, note that

$$\mathbb{P}\biggl\{F_{f_k}(\delta) \geqslant \frac{1}{c}F_{n,f_k}\biggl(\frac{\delta}{c}\biggr)\biggr\} = \mathbb{P}\biggl\{F_{n,f_k}\biggl(\frac{\delta}{c}\biggr) \leqslant \frac{c}{2}\biggr\} = \mathbb{P}\biggl\{\sum_{j=1}^{n} I_{\{\varepsilon_{k,j}=-1\}} \leqslant \frac{nc}{2}\biggr\}$$
$$= \mathbb{P}\biggl\{\sum_{j=1}^{n} I_{\{\varepsilon_{k,j}=+1\}} > \biggl(\frac{1}{2} - \tau\biggr)n\biggr\} \leqslant 1 - \beta_1 n^{-1/2}\exp\bigl\{-4n\tau^2\bigr\},$$

where $\tau = (c-1)/2$. The rest of the proof is quite similar. □

Acknowledgement

This paper was written when the author was visiting the Department of Statistics, University of Washington and H. Poincaré Institute in Paris. The author is very thankful to these institutions for their hospitality and also to Jon Wellner for numerous discussions on the problems closely related to the paper and to Olivier Bousquet for help with some parts of the manuscript. The author is especially thankful to the anonymous referee for an exceptionally careful reading of the manuscript, noticing numerous mistakes and making a number of suggestions that improved the paper.

REFERENCES

[1] M. Anthony, P. Bartlett, Neural Network Learning: Theoretical Foundations, Cambridge University Press, 1999.

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

[3] C. Cortes, V. Vapnik, Support vector networks, Machine Learning 20 (1995) 273–297.

[4] L. Devroye, L. Györfi, G. Lugosi, A Probabilistic Theory of Pattern Recognition, Springer-Verlag, New York, 1996.

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

[6] E. Giné, V. Koltchinskii, J. Wellner, Ratio limit theorems for empirical processes, Preprint, 2003.

[7] B. Kégl, T. Linder, G. Lugosi, Data-dependent margin-based generalization bounds for classification, in: D. Helmbold, B. Williamson (Eds.), Proc. of 14th Annual Conference on Computational Learning Theory, COLT2001, Lecture Notes in Artificial Intelligence, Springer, New York, 2001, pp. 368–384.

[8] V. Koltchinskii, D. Panchenko, Rademacher processes and bounding the risk of function learning, in: E. Giné, D. Mason, J. Wellner (Eds.), High Dimensional Probability II, Birkhäuser, Boston, 2000, pp. 444–459.

[9] V. Koltchinskii, D. Panchenko, Empirical margin distributions and bounding the generalization error of combined classifiers, Ann. Statist. 30 (2002) 1–50.

[10] V. Koltchinskii, D. Panchenko, F. Lozano, Some new bounds on the generalization error of combined classifiers, in: T.K. Leen, T.G. Dietterich, V. Tresp (Eds.), Proc. of NIPS'2000, in: Advances in Neural Information Processing Systems, Vol. 13, MIT Press, 2001, pp. 245–251. URL:http://www.boosting.org/.

[11] V. Koltchinskii, D. Panchenko, F. Lozano, Further explanation of the effectiveness of voting methods: the game between margins and weights, in: D. Helmbold, B. Williamson (Eds.), Proc. of 14th Annual Conference on Computational Learning Theory, COLT2001, in: Lecture Notes in Artif. Intell., Springer, New York, 2001, pp. 241–255.

[12] V. Koltchinskii, D. Panchenko, F. Lozano, Bounding the generalization error of convex combinations of classifiers: balancing the dimensionality and the margins, Ann. Appl. Probab. 13 (1) (2003) 213–252.

[13] M. Ledoux, M. Talagrand, Probability in Banach Spaces, Springer-Verlag, New York, 1991.

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

[15] P. Massart, Some applications of concentration inequalities to statistics, Ann. Fac. Sci. Tolouse (IX) (2000) 245–303.

[16] R. Schapire, Y. Freund, P. Bartlett, W.S. Lee, Boosting the margin: a new explanation of effectiveness of voting methods, Ann. Statist. 26 (1998) 1651–1687.

[17] M. Talagrand, A new look at independence, Ann. Probab. 24 (1996) 1–34.

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

[19] A. Tsybakov, Optimal aggregation of classifiers in statistical learning, Preprint, 2002.

[20] A.W. van der Vaart, J.A. Wellner, Weak Convergence and Empirical Processes. With Applications to Statistics, Springer-Verlag, New York, 1996.

[21] V. Vapnik, Statistical Learning Theory, Wiley, New York, 1998.