Journal of Machine Learning Research 8 (2007) 863-889 Submitted 6/05; Revised 10/06; Published 4/07

Combining PAC-Bayesian and Generic Chaining Bounds

Jean-Yves Audibert[*] AUDIBERT@CERMICS.ENPC.FR
CERTIS, Ecole Nationale des Ponts et Chausses
19, rue Alfred Nobel - Cité Descartes
F-77455 Marne-la-Vallée cedex 2, France

Olivier Bousquet[†] O.BOUSQUET@PERTINENCE.COM
Pertinence
32, rue des Jeûneurs
F-75002 Paris, France

Editor: John Shawe-Taylor

Abstract

There exist many different generalization error bounds in statistical learning theory. Each of these bounds contains an improvement over the others for certain situations or algorithms. Our goal is, first, to underline the links between these bounds, and second, to combine the different improvements into a single bound. In particular we combine the PAC-Bayes approach introduced by McAllester (1998), which is interesting for randomized predictions, with the optimal union bound provided by the generic chaining technique developed by Fernique and Talagrand (see Talagrand, 1996), in a way that also takes into account the variance of the combined functions. We also show how this connects to Rademacher based bounds.

Keywords: statistical learning theory, PAC-Bayes theorems, generalization error bounds

1. Introduction

Since the first results of Vapnik and Chervonenkis on uniform laws of large numbers for classes of $\{0;1\}$-valued functions, there has been a considerable amount of work aiming at obtaining generalizations and refinements of these bounds. This work has been carried out by different communities. On the one hand, people developing empirical processes theory like Dudley and Talagrand (among others) obtained very interesting results concerning the behavior of the suprema of empirical processes. On the other hand, people exploring learning theory tried to obtain refinements for specific algorithms with an emphasis on data-dependent bounds.

The goal of a generalization error bound is to control the behavior of the function that is returned by the algorithm. This function is data-dependent and thus unknown before seeing the data. As a consequence, if one wants to make statements about its error, one has to be able to *predict* which function is likely to be chosen by the algorithm. Since this cannot be done exactly, there is a need to provide guarantees that hold simultaneously for several candidate functions. This is known as the union bound. The way to perform this union bound optimally is now well mastered in the empirical

∗. Part of this work was done while this author was at laboratoire de Probabilités et Modèles Aléatoires, Université Paris 6, Paris

†. Part of this work was done while this author was at the Max Planck Institute for Biological Cybernetics, Tübingen

processes community. In particular, the role of the metric structure of the space of functions in the deviations of the empirical process has been thoroughly studied (see, for example, Talagrand, 2005).

In the learning theory setting, one is interested in bounds that are as algorithm and data dependent as possible. This particular focus has made concentration inequalities (see, for example, Boucheron et al., 2000) popular as they allow to obtain data-dependent results in an effortless way. Another aspect that is of interest for learning is the case where the classifiers are randomized or averaged. McAllester (1998, 1999) has proposed a new type of bound that takes the randomization into account in a clever way.

Another direction in which the error bounds can be improved is by using the variance of the functions in order to control the fluctuations of the empirical error. This idea originated in Huber's peeling device (here peeling refers to the fact that the class of function is "peeled off" into layers according to the variance of the functions) and is often referred to as "localization" (see, for example, van der Vaart and Wellner, 1996; van de Geer, 2000; Massart, 2000). It allows to get bounds with optimal rates of convergence, for example in the case of empirical error minimization (see, for example, Bartlett et al., 2005).

Our goal is to combine several of these improvements, bringing together the power of the majorizing measures as an optimal union bound technique and the power of the PAC-Bayesian bounds to handle randomized predictions efficiently in a way that is sensitive to the variance (localization) effect.

The paper is structured as follows. Next section introduces the notation and gives an overview of the existing bounds. Section 3 then presents our main result while Section 4 discusses its applications, showing in particular how to recover previously known results.

2. A Survey of Previous Results

In this section, after having introduced some notation and the setup of generalization bounds, we state and compare existing generalization bounds. By doing so, we hope to give a better and more global view on the various approaches that have been developed to obtain error bounds. In particular, we want to emphasize the differences and the complementarity of the approaches. Our goal is not to entirely cover the topic of error bounds and we thus refer the interested reader to Boucheron et al. (2005), Koltchinskii (2006), Bartlett et al. (2004), Bartlett and Mendelson (2006), Massart (2006), Massart (2000) and references therein for more information.

2.1 Notation

Random Variables and Distributions. We consider an input space $\mathcal{X}$, an output space $\mathcal{Y}$ and a probability distribution P on the product space $\mathcal{Z} := \mathcal{X} \times \mathcal{Y}$. Let $Z := (X,Y)$ ($Z \in \mathcal{Z}$) denote a pair of random variables distributed according to P and for a given integer n, let $Z_1, \ldots, Z_n$ and $Z'_1, \ldots, Z'_n$ be two independent samples of n independent copies of Z. We denote by P_n, P'_n and P_{2n} the empirical measures associated respectively to the first, the second and the union of both samples. $\mathbb{E}^n$, $\mathbb{E}'^n$ and $\mathbb{E}^{2n}$ denote the expectation with respect to the first, second and union of both training samples, while $\mathbb{P}^n, \mathbb{P}'^n$ and $\mathbb{P}^{2n}$ denote the distribution of these samples (i.e., $\mathbb{P}^n$ is the n-fold product distribution whose marginals are P).

Regret Functions. To each function $g : \mathcal{X} \to \mathcal{Y}$ and each function $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, we associate the function $f : \mathcal{Z} \to \mathbb{R}$ defined as

$$f(z) = L(g(x), y).$$

Such functions g, L and f will be respectively called prediction function, loss function and regret function. In classification, the loss function is $L = \mathbb{1}_{g(x) \neq y}$ where $\mathbb{1}$ denotes the indicator function. To each pair of prediction functions g_1 and g_2, we define the relative regret function $f : \mathcal{Z} \to \mathbb{R}$ defined as

$$f(z) = L(g_1(x), y) - L(g_2(x), y).$$

To each measurable real-valued function f defined on $\mathcal{Z}$, we denote their expectation under P by Pf and their empirical expectation by $P_n f$ $\left(\text{i.e., } P_n f = n^{-1} \sum_{i=1}^{n} f(Z_i)\right)$. For (relative) regret functions, Pf is often called the (relative) risk.

Geometry of the Regret Class. Let $\mathcal{F}$ denote a set of measurable real-valued functions defined on $\mathcal{Z}$. Typical examples of such a class are based on regret function or relative regret functions and derived from a set of prediction functions. By slight extension, any set of measurable real-valued functions defined on $\mathcal{Z}$ will be called a regret class. On the set $\mathcal{F}$, we consider the pseudo-distance

$$d(f_1, f_2) = \sqrt{P(f_1 - f_2)^2},$$

and define similarly d_n, d_n' and d_{2n}. We define the covering number $N(\mathcal{F}, \varepsilon, d)$ as the minimum number of balls of radius ε needed to cover $\mathcal{F}$ in the pseudo-distance d.

An important task for obtaining error bounds is to define sieves, that is, subsets of $\mathcal{F}$ that approximate $\mathcal{F}$ to a certain distance. The simplest way is to use minimum covers at various scales, but it is also possible to use sequences of nested partitions as defined below.

Definition 1 *A sequence $(\mathcal{A}_j)_{j \in \mathbb{N}}$ is a* sequence of nested partitions *of $\mathcal{F}$ if*

- *$\mathcal{A}_j$ is a partition of $\mathcal{F}$ either countable or equal to the set of all singletons of $\mathcal{F}$*
- *The $\mathcal{A}_j$ are nested: each element of $\mathcal{A}_{j+1}$ is contained in an element of $\mathcal{A}_j$, and $\mathcal{A}_0 = \{\mathcal{F}\}$*

For a partition $\mathcal{A}$, we denote by $A(f)$ the unique element of $\mathcal{A}$ containing f.

Given a sequence of nested partitions $(\mathcal{A}_j)$, we can build a collection $(S_j)_{j \in \mathbb{N}}$ of approximating subsets of $\mathcal{F}$ in the following way: for each $j \in \mathbb{N}$, for each element A of $\mathcal{A}_j$, choose a unique element of $\mathcal{F}$ contained in A and define S_j as the set of all chosen elements. We have $|S_0| = 1$ and denoting by $p_j(f)$ the unique element of S_j contained in $A_j(f)$ we have

$$\begin{cases} p_j(f) = f \text{ for any } f \in S_j, \\ p_{j-1} \circ p_j = p_{j-1}. \end{cases}$$

Measures on the Regret Class. We denote by ρ and π two probability measures on the space $\mathcal{F}$, so that ρPf will actually mean the expectation of Pf when f is sampled according to the probability measure ρ. We will denote by $\mathcal{M}_+^1(\mathcal{F})$ the set of all probability measures on $\mathcal{F}$. For two such measures, $K(\rho, \pi)$ will denote their Kullback-Leibler divergence defined as

$$K(\rho, \pi) = \rho \log \frac{d\rho}{d\pi} := \int_{\mathcal{F}} \log \left(\frac{d\rho}{d\pi}(f) \right) \rho(df),$$

when ρ is absolutely continuous with respect to π and $K(\rho, \pi) = +\infty$ otherwise.

Other Notation. The notation x_+ and x_- refers to the positive and negative parts respectively, that is, $x_+ := \max(x,0)$ and $x_- := \max(-x,0)$. β will denote some positive real number while C will be some positive constant (whose value may differ from line to line).

2.2 The Setting for Error Bounds

Generalization error bounds give upper bounds on the true (i.e., under P) error of the function returned by a learning algorithm. These upper bound typically involve the empirical error of this function. A learning algorithm (or learning rule), is a map from sequences of pairs $(X_1,Y_1),\ldots,(X_n,Y_n)$ to functions $g : \mathcal{X} \to \mathbb{R}$. We will denote by g_n the function returned by the algorithm under investigation.

2.2.1 ABSOLUTE AND RELATIVE RISK BOUNDS

The goal is thus to formulate a statement of the following form: for each (small enough) $\beta > 0$, with probability at least $1-\beta$ with respect to the random draw of the sample,

$$\mathbb{E}\left[L(g_n(X),Y)|(X_1,Y_1),\ldots,(X_n,Y_n)\right] \leq \frac{1}{n}\sum_{i=1}^{n} L(g_n(X_i),Y_i) + B(n,\beta)\,, \tag{1}$$

where the expectation in the left-hand side is taken with respect to the distribution of (X,Y), meaning: conditionally to the training sample distribution.

Another type of result that can be obtained is a relative risk bound where one compares the risk of the algorithm to the risk of a fixed prediction function $\tilde{g}$. The type of inequality that could be obtained looks like this

$$\begin{aligned}\mathbb{E}\left[L(g_n(X),Y)|(X_1,Y_1),\ldots,(X_n,Y_n)\right] - \mathbb{E}\left[L(\tilde{g}(X),Y)\right] \\ \leq \tfrac{1}{n}\textstyle\sum_{i=1}^{n}\left[L(g_n(X_i),Y_i) - L(\tilde{g}(X_i),Y_i)\right] + B(n,\beta)\,.\end{aligned} \tag{2}$$

With the notation introduced above, denoting by f_n the (relative) regret function associated to the predicition g_n we can rewrite both previous inequalities as follows:

$$Pf_n - P_n f_n \leq B(n,\beta)\,. \tag{3}$$

This shows that there is no essential difference between the techniques used to get bounds of the form (1) and (2).

2.2.2 AVERAGED BOUNDS

Most algorithms simply pick a candidate function g_n from a fixed set $\mathcal{G}$. However, Bayesian algorithms usually aggregate several such functions by taking their weighted average. The corresponding regret function is not the average of the regrets of the individual functions which makes the analysis difficult. In order to avoid this caveat, it is common to first study randomized estimators and then relate the randomized to the aggregate ones. Randomized estimators are built by replacing the weighted average by a randomized choice (where the probability of choosing a specific function is proportional to its weight). The advantage of such a procedure is that it is relatively simple to analyze. Indeed, if one uses weights given by a probability distribution ρ on $\mathcal{G}$, the error of the randomized estimator is simply the average of the errors of the combined estimators ρPf.

Hence, bounds for randomized estimators will have the form

$$\rho_n P f \leq \rho_n P_n f + B(n, \beta)\,, \tag{4}$$

where ρ_n is the specific distribution chosen by the algorithm based on the data.

2.2.3 ALGORITHM AND DATA DEPENDENT BOUNDS

In the form stated in (3), the quantity B only depends on the confidence level β and the sample size n. The only way in which B depends on the algorithm is usually by incorporating terms that depend on the class of prediction functions used by the algorithm. For example if the algorithm picks functions in a class $\mathcal{G}$ whose associated regret class is $\mathcal{F}$, statement (3) can be deduced from

$$\sup_{f \in \mathcal{F}} \left\{ P f - P_n f \right\} \leq B(n, \beta)\,.$$

This is a *supremum bound* and is the most common type found in the literature. Unfortunately, this usually yields quantitatively loose bounds which do not tell much about the algorithm's behavior.

Ideally, B should depend on the sample (data-dependent bound) and on the specific function f_n or distribution ρ_n chosen by the algorithm (algorithm-dependent bound).

Hence we can aim at obtaining bounds of the form (algorithm-dependent)

$$\forall f \in \mathcal{F}\,, P f \leq P_n f + B(n, \beta, f)\,,$$

or even (algorithm and data-dependent)

$$\forall f \in \mathcal{F}\,, P f \leq P_n f + B(n, \beta, f, Z_1, \ldots, Z_n)\,.$$

2.3 Previous Error Bounds

We are now in a position to state and compare some of the previously known error bounds. From now on, the functions in the regret class $\mathcal{F}$ are assumed to be bounded. Without loss of generality, we may assume that they take their values in $[0;1]$. We do not always state the bounds in their original form in order to allow an easier comparison and we thus include the proofs and the explicit values of the constant C in Section B.3.

2.3.1 SUPREMUM BOUNDS

We start with the most common bounds involving the supremum over a class of the difference between true and empirical error.

Single function. The starting point is to consider a class containing only one function f. By Hoeffding's inequality one easily gets that with probability at least $1-\beta$,

$$P f - P_n f \leq C \sqrt{\frac{\log(\beta^{-1})}{n}}\,. \tag{5}$$

Unfortunately, when $\mathcal{F}$ has more than one element, this statement can only be made separately for each function. [Proof in Section B.3.1]

Finite union bound. It is easy to convert the above statement into one which is valid simultaneously for a finite set of functions $\mathcal{F}$. The simplest form of the union bound gives that with probability at least $1-\beta$,

$$\sup_{f\in\mathcal{F}}\left\{Pf-P_nf\right\}\leq C\sqrt{\frac{\log|\mathcal{F}|+\log(\beta^{-1})}{n}}. \tag{6}$$

The term $\log|\mathcal{F}|$ represents the extra price to pay to obtain a uniform statement. It can be considered as a measure of the complexity of the class $\mathcal{F}$. [Proof in Section B.3.2]

Symmetrization. When $\mathcal{F}$ is infinite this cannot work directly. The trick is to introduce a second sample $Z'_1,\ldots,Z'_n$ (see the notation section for definitions) and to consider the set of vectors formed by the values of each function in $\mathcal{F}$ on the double sample. When the functions have values in $\{0;1\}$, this is a finite set and the above union bound applies. This idea was first used in Vapnik and Chervonenkis (1971) to obtain that with probability at least $1-\beta$,

$$\sup_{f\in\mathcal{F}}\left\{Pf-P_nf\right\}\leq C\sqrt{\frac{\log\mathbb{E}^{2n}N(\mathcal{F},1/2n,d_{2n})+\log(2\beta^{-1})}{n}}. \tag{7}$$

The capacity is here better estimated than in (6) since the term $N(\mathcal{F},1/2n,d_{2n})$ does not count twice functions classifying in the same way the training and virtual samples. However the quantity $\mathbb{E}^{2n}N(\mathcal{F},1/2n,d_{2n})$ cannot be computed in general since P is unknown, but upper bounds can be obtained in terms of combinatorial parameters such as the VC dimension.

We now see that the complexity of $\mathcal{F}$ has to be measured in a way that involves the metric induced by the distribution. What matters is thus not how many functions there are in $\mathcal{F}$ but how they span the space of possible vectors $(f(Z_1),\ldots,f(Z_n),f(Z'_1),\ldots,f(Z'_n))$, which can be measured by the minimum number of balls required to get an approximation at scale $1/2n$. [Proof in Section B.3.3]

Chaining. One limitation of the above result is that it applies only to $\{0;1\}$-valued functions and it measures the size only at the smallest scale which is known to be suboptimal in general. The union bound can be refined by considering finite covers of the set of function at different scales. This is called the *chaining* technique, pioneered by Dudley (1984) since one constructs a chain of functions that approximate a given function more and more closely. The results involve the Koltchinskii-Pollard entropy integral as, for example in Devroye and Lugosi (2001). One has with probability at least $1-\beta$,

$$\sup_{f\in\mathcal{F}}\left\{Pf-P_nf\right\}\leq C\left(\frac{1}{\sqrt{n}}\mathbb{E}^n\int_0^\infty\sqrt{\log N(\mathcal{F},\varepsilon,d_n)}d\varepsilon+\sqrt{\frac{\log(\beta^{-1})}{n}}\right). \tag{8}$$

[Proof in Section B.3.8]. Chained bounds are particularly useful when one has a tight control of the entropy numbers of the set $\mathcal{F}$ (see, for example, van der Vaart 1998 and van de Geer 2000).

Generic chaining. It has been noticed by Fernique and Talagrand that it is possible to capture the complexity in a better way than using minimal covers by considering majorizing measures or generic chaining. This is essentially optimal for Gaussian processes and optimal up to logarithmic factors for empirical processes. Let $r>1$ and $(\mathcal{A}_j)_{j\geq1}$ be a sequence of nested partitions of $\mathcal{F}$ such that all the elements of $\mathcal{A}_j$ have diameter at most r^{-j} w.r.t. the distance d_n. Let also $(\pi^{(j)})$ be

a sequence of probability distributions defined on $\mathcal{F}$. We can prove that with probability at least $1-\beta$,

$$\sup_{f\in\mathcal{F}}\left\{Pf-P_nf\right\}\leq C\left(\frac{1}{\sqrt{n}}\mathbb{E}^n\sup_{f\in\mathcal{F}}\sum_{j=1}^{\infty}r^{-j}\sqrt{\log\{1/\pi^{(j)}[A_j(f)]\}}+\sqrt{\frac{\log(\beta^{-1})}{n}}\right). \tag{9}$$

[Proof in Section B.3.9]. The interpretation of the complexity term is a bit harder now.[1] If one takes partitions induced by minimal covers of $\mathcal{F}$ at radii r^{-j}, and uniform measures concentrated at the centers of the balls, one has $1/\pi^{(j)}[A_j(f)]=N(\mathcal{F},r^{-j},d_n)$ so that (9) leads to a sum of terms of the form $r^{-j}\sqrt{\log N(\mathcal{F},r^{-j},d_n)}$, which allows to recover (8). This means that the complexity term of (9) is at least as sharp as the one of (8), but has more flexibility since the partitions can be built in a better way than via uniform covers.

As in Section 2.1, from the nested partitions $(\mathcal{A}_j)$, we can build a collection $(S_j)_{j\in\mathbb{N}}$ of approximating subsets of $\mathcal{F}$ in the following way: for each $j\in\mathbb{N}$, for each element A_j of $\mathcal{A}_j$, choose a unique element of $\mathcal{F}$ contained in A_j and define S_j as the set of all chosen elements. Then consider $p_j(f)$ the unique element of S_j contained in $A_j(f)$. The $p_j(f)$ define successive approximations of f. Without loss of generality, consider that the null function belongs to $\mathcal{A}_0$ and that S_0 contains this function. The sum in (9) (as the integral in (8)) comes from the core of the chaining idea which is to decompose the function f (which is a difference between two functions in the case of relative regret classes, or a difference between itself and the null function in the case of a regret class) into $\sum_{j>0}p_j(f)-p_{j-1}(f)$ and to separately control the deviation of each term. These terms form a *chain* of finer and finer approximations to f that give the name to the method (chaining). This is essentially the idea that we use in the proof of our main results.

Let us try to give some insights about how to construct the partitions $(\mathcal{A}_j)$ (for more details, see Talagrand, 2005). First of all, one should understand that the measures $\pi^{(j)}$ play no essential role here (as was noticed by Talagrand, 2001, , they can be taken as finitely supported uniform measures on appropriate subsets) The only reason why we state the result in the above form is to emphasize the connection with our main result (to be presented in Section 3).

It turns out that there are many ways to state the majorizing measure/generic chaining bound. Each way involves a different geometric construction on the regret class $\mathcal{F}$ and a different notion of size, but most can be shown to be equivalent up to constant factors. The general form of such bounds is

$$\sup_{f\in\mathcal{F}}\sum_{i\geq i_0}F(f,i)G(f,i)$$

where $F(f,i)$ is a measure of the scale of the geometric object of order i containing f, while $G(f,i)$ is a measure of the size of this object. For example, if one uses balls, $F(f,i)$ is the radius of the ball $\big($e.g., $F(f,i)=2^{-i}\big)$ and $G(f,i)$ is the "mass" (or rather a function of the mass) of the ball centered at f and of radius 2^{-i} $\big($e.g., $G(f,i)=\sqrt{\log 1/\pi^{(i)}[B(f,2^{-i})]}\big)$.

1. It is important to mention that there are various possible ways to measure the complexity of the space $\mathcal{F}$ that all lead to an essentially optimal result. For example, one could replace the set $A_j(f)$ by a ball centered at f and with radius r^{-j}. This would give the standard majorizing measure bound (where all the $\pi^{(j)}$ are the same). Also, one can use partitions whose elements are allowed to have different diameters and replace the r^{-j} term by the diameter of $A_j(f)$. Using such an approach actually allows to get rid of the measures $\pi^{(j)}$, the sum in the complexity term becoming $\sum d(A_j(f))\sqrt{\log|\mathcal{A}_j|}$. We did not choose this formulation as we wish to obtain a result that explicitly uses the probability π so that it can be more directly related to the PAC-Bayesian bounds. The interested reader is referred to Talagrand (2005) for more information about the variants of the generic chaining bounds.

Coming back to nested partitions (which really are the key ingredient here), let us say a few words about their construction. Talagrand proposed a partitioning scheme that runs as follows. The general idea is that one starts with a size function (that measures how big a subset of $\mathcal{F}$ is). This function has to satisfy certain compatibility conditions in order to ensure that the resulting construction will be optimal (up to constants). Then one starts with $\mathcal{A}_0 = \{\mathcal{F}\}$ and, at each stage, the partition is refined by splitting each of its element in a greedy way: one selects separated enough points in the element to be split and builds balls arounds these points. Starting with the "biggest" ball (as measured by the size function), one removes the balls until nothing is left. The sub-partition elements are thus the subsets that are removed with each ball.

Rademacher averages. As we said before, the generic chaining bound is optimal up to constant factors. More precisely, it is a tight upper bound for the quantity

$$\mathbb{E}^n \left[\sup_{f \in \mathcal{F}} \left\{ Pf - P_n f \right\} \right].$$

It is also a lower bound for this quantity but up to a $\log n$ factor. Hence the generic chaining bound is as good (up to this log) as another well-known quantity in the learning theory community, the Rademacher average of $\mathcal{F}$:

$$\mathbb{E}^n \left[\mathbb{E}_\sigma \sup_{f \in \mathcal{F}} \frac{1}{n} \sum \sigma_i f(Z_i) \right],$$

where the σ_i are independent random signs ($+1, -1$ with probability $1/2$). Using this quantity as a measure of complexity, it is possible to obtain the following statement. One has with probability at least $1-\beta$,

$$\sup_{f \in \mathcal{F}} Pf - P_n f \leq C \left(\frac{1}{n} \mathbb{E}^n \mathbb{E}_\sigma \sup_{f \in \mathcal{F}} \sum_{i=1}^n \sigma_i f(Z_i) + \sqrt{\frac{\log(\beta^{-1})}{n}} \right). \tag{10}$$

[Proof in Section B.3.7]

2.3.2 VARIANCE (LOCALIZED) BOUNDS

It has been noticed that the supremum bounds do not give optimal rates of convergence for algorithms based on the minimization of the empirical risk. The main reason is that the size of the deviations between Pf and $P_n f$ depends on the variance $\operatorname{Var} f$. A lot of work has been carried out recently in order to take this into account in the bounds presented above, for example by restricting the complexity term to functions with small variance.

Variance for Single Functions. Since the deviations between Pf and $P_n f$ for a given function f actually depend on its variance, one can refine (5) into

$$Pf - P_n f \leq C \left(\sqrt{\frac{\operatorname{Var} f \log(\beta^{-1})}{n}} + \frac{\log(\beta^{-1})}{n} \right). \tag{11}$$

[Proof in Section B.3.5]

Variance with Symmetrization. The above inequality can be combined with the symmetrization trick. This was done by Vapnik and Chervonenkis (1974) (for functions in $\{0;1\}$). Their result is that with probability at least $1-\beta$,

$$\forall f \in \mathcal{F}, Pf - P_n f \leq C\sqrt{Pf}\sqrt{\frac{\log \mathbb{E}^{2n} N(\mathcal{F}, 1/2n, d_{2n}) + \log(4\beta^{-1})}{n}}, \tag{12}$$

which also gives with probability at least $1-\beta$,

$$\forall f \in \mathcal{F}, Pf - P_n f \leq C\Bigg(\sqrt{P_n f}\sqrt{\frac{\log \mathbb{E}^{2n} N(\mathcal{F}, 1/2n, d_{2n}) + \log(4\beta^{-1})}{n}} + \frac{\log \mathbb{E}^{2n} N(\mathcal{F}, 1/2n, d_{2n}) + \log(4\beta^{-1})}{n}\Bigg).$$

One can thus consider that the capacity term is weighted by the risk (or equivalently here by the empirical risk).

Localized Rademacher Averages. It is also possible to combine (11) with (10) as was done for example by Bartlett et al. (2005). This gives a complexity term which involves Rademacher averages that are computed on subsets of functions with small variance.

How to Use Variance Bounds. In order to better explain how variance bounds can be used efficiently, we should notice that the quantitative gain occurs only when Var f is small for the function chosen by the algorithm.

For a loss function L taking its values $[0;1]$, the variance of a regret function f can be bounded successively by Pf^2 and Pf. Consequently, in low noise setting (i.e., when there exists a prediction function such that its associated risk Pf is small), one can expect that the deviation of the risk of an algorithm based on the minimization of the empirical error is much smaller than in noisy situations.

For relative regret function the situation is different. We are still interested in low (relative) risk but in general we will not have any particular relation between the variance Var f and the relative risk Pf. Consequently variance localized bounds will not yield any significant improvement. However in some situations as in classification under Mammen and Tsybakov noise condition (Tsybakov, 2004) or in least square regression (see, for example, Audibert, 2004a), one has the inequality $\mathrm{Var}\, f \leq (Pf)^{\alpha}$ for some positive α. This inequality is well exploited by variance localized bounds.

2.3.3 ALGORITHM DEPENDENT COMPLEXITY

Another direction in which the bounds can be improved is by making the bound depend more directly of the function chosen by the algorithm. The variance bounds already have this property but the bound only depends on Var f but not on where the function is in the space $\mathcal{F}$. We now present results where the complexity term depends directly on the selected function.

Weighted union bound and algorithm dependence. The finite union bound can be directly extended to the countable case by introducing a probability distribution π over $\mathcal{F}$ which weights each function (McAllester, 1998) and gives, with probability at least $1-\beta$,

$$\forall f \in \mathcal{F}, Pf - P_n f \leq C\sqrt{\frac{\log 1/\pi(f) + \log(\beta^{-1})}{n}}. \tag{13}$$

The complexity term now depends on f and of course, taking a uniform distribution on a finite set we recover (6).

Surprisingly, the capacity term could be arbitrarily small if π were chosen appropriately. However, π has to be chosen before seeing the data, so there is no way to ensure that the bound will be always small.

It is important to understand that the choice of π is completely arbitrary and need not reflect any prior belief in what is the true target function. The distribution π is just a "technical" prior which is used to formulate the bound.

Inequality (13) can be read as follows: given a fixed π, if one samples data repeatedly, with high probability, the error of any function f in the class will be upper bounded by a function of $\pi(f)$. This function of $\pi(f)$ is basically a bound on the deviation one would expect for each individual function, plus an extra term coming from the fact that the function f to be considered is unknown prior to observing the sample so that we need to have a statement which holds simultaneously for all functions in the class. The interesting point is that in making this statement uniform, there is a freedom in the choice of how we distribute the 'extra cost'. The distribution of the uniformization cost is thus represented by π. Any probability distribution will do, and if one is lucky, the functions of interest (the ones returned by the classification algorithm when given typical samples from the problem) will have large prior and the bound for them will be relatively small.

Thus, if one wants to obtain a bound which has small values, one has to "guess" how likely each function $f \in \mathcal{F}$ is to be chosen by the algorithm.

Averaging. We now come to the case of randomized predictions which give rise (as explained above) to averaged bounds of the form (4). Consider a probability distribution ρ_n defined on a countable $\mathcal{F}$, take the expectation of (13) with respect to ρ_n and use Jensen's inequality. This gives with probability at least $1-\beta$,

$$\rho_n(Pf - P_n f) \leq C\sqrt{\frac{K(\rho_n,\pi) + H(\rho_n) + \log(\beta^{-1})}{n}},$$

where ρ_n is still the specific distribution chosen by the algorithm based on the data and $H(\rho_n)$ is its Shannon entropy. The l.h.s. is the difference between true and empirical error of a randomized classifier which uses ρ_n as weights for choosing the decision function (independently of the data). The following PAC-Bayes bound (McAllester, 1999) refines the above bound since it has the form (for possibly uncountable $\mathcal{F}$)

$$\rho_n(Pf - P_n f) \leq C\sqrt{\frac{K(\rho_n,\pi) + \log n + \log(\beta^{-1})}{n}}. \tag{14}$$

This in particular shows that the entropy term is unnecessary. To some extent, one can consider that the PAC-Bayes bound is a refined union bound where the gain happens when ρ_n is not concentrated on a single function (or more precisely ρ_n has entropy larger than $\log n$).

The complexity now depends upon the arbitrary choice of π and one may notice that it is modulated by the "spread" of ρ_n. Indeed, if ρ_n is concentrated, this term can be big, but if ρ_n is similar to π it becomes very small. As a special case, if the randomizing distribution is concentrated at a single function (corresponding to classical algorithms that simply pick a function), the bound has the form (13) which is not suited for large (e.g., uncountable) sets of functions.

2.3.4 DATA-DEPENDENT BOUNDS

We now consider ways to obtain bounds where the complexity term itself depend on the sample (hence can be computed from the data only).

Transductive priors. It is actually possible to combine the symmetrization and weighting ideas (Catoni, 2003). For example, if one defines a function $\Pi : \mathcal{Z}^{2n} \to \mathcal{M}_+^1(\mathcal{F})$ that is *almost exchangeable* in the sense that if we exchange the value of z_i and z_{i+n} we do not change the value of the function, then one gets, with probability at least $1-\beta$ (over the random choice of a double sample),

$$\forall f \in \mathcal{F}, P'_n f - P_n f \leq C\sqrt{\frac{\log 1/\Pi(Z_1,\ldots,Z_n,Z'_1,\ldots,Z'_n)(f) + \log(\beta^{-1})}{n}}.$$

Note that we no longer require $\mathcal{F}$ to be countable but $\Pi(.)$ should have countable support for each value of the double sample. This type of result is interesting in the transduction framework, where the instances (or inputs) are known in advance and where the randomness is in the way the data is split into a training set (with known labels) and a testing set (with unknown labels). Converting this into an induction statement (comparing the empirical with the expected errors) gives with probability $1-\beta$,

$$\forall f \in \mathcal{F}, Pf - P_n f \leq C\mathbb{E}'^n\sqrt{\frac{\log 1/\Pi(Z_1,\ldots,Z_n,Z'_1,\ldots,Z'_n)(f) + \log(\beta^{-1})}{n}}.$$

This is useful provided one can upper bound the first logarithmic factor in the r.h.s. either with a data-independent quantity or with an observable function of the first sample.

Concentration. Using concentration inequalities as in Boucheron et al. (2000) for example, one can get rid of the expectation appearing in the r.h.s. of (7), (8), (10) or (9) and thus obtain a bound that can be computed from the data.

In particular, a data-dependent and localized version of (10) is given in Bartlett et al. (2005). However it has not been combined with the PAC-Bayes improvement for randomized predictions.

2.3.5 SUMMARY

Our goal in this work is to see how to attempt to combine the different approaches that have been used before, in the hope to obtain a bound that has the combined properties

1. structural (metric structure effect)
2. statistical (variance effect)
3. PAC-Bayesian (averaging effect).

The connection between 1 and 2 already exists, so does the connection between 2 and 3. Our main result aims at dealing simultaneously with 1, 2 and 3.

The main difficulty in connecting 1 and 2 with 3 is that if one has a non-countable infinite set of functions, even using symmetrization, if the prior π is non-atomic, then $\pi(\{f\}) = 0$ for all f. Hence the complexity term $K(\rho_n, \pi)$ blows up when ρ_n is concentrated on a single function. The result we present below is, to our knowledge, the first one which actually bridges this gap since the complexity term does not blow up when ρ_n is concentrated at one function.

3. Main results

We now state and comment our main result. We recall that $\mathcal{F}$ denotes a set of functions defined on a measurable space $\mathcal{Z}$ and taking their values in $[0;1]$. In the theorem we present, one has to

- choose a sequence of nested partitions $(\mathcal{A}_j)_{j\in\mathbb{N}}$ of the set $\mathcal{F}$,
- build an associated sequence of approximating sets $(S_j)_{j\in\mathbb{N}}$ (see Section 2.1),
- for each f and $j\in\mathbb{N}$, define its approximating functions $p_j(f)$ as the unique element of S_j contained in $A_j(f)$, where $A_j(f)$ is the set of the partition $\mathcal{A}_j$ containing f
- for each $j\in\mathbb{N}$, choose the distribution $\pi^{(j)}$ on $\mathcal{F}$.

The quantities $\pi^{(j)}$, S_j, and p_j are allowed to depend on the sample $Z_1,\dots,Z_n$ provided that it also depends on a double sample $Z'_1,\dots,Z'_n$ in an almost exchangeable way (i.e., exchanging Z_i and Z'_i does not affect their value). Denote δ_f the Dirac measure on f. For a probability distribution ρ on $\mathcal{F}$, define its j-th projection as

$$[\rho]_j = \sum_{f\in S_j} \rho[A_j(f)]\delta_f,$$

when S_j is countable and $[\rho]_j=\rho$ otherwise. When S_j is countable, $[\rho]_j$ is a probability distribution on $\mathcal{F}$ supported by S_j and can be viewed as the projection of ρ on S_j. Let ρ_n be a randomized estimator, that is, a data-dependent distribution on the set $\mathcal{F}$. To shorten the notation, the average (w.r.t. the probabiblity distribution ρ_n) distance between two successive approximations is denoted by

$$\rho_n d_j^2 := \rho_n d_{2n}^2[p_j(f), p_{j-1}(f)],$$

where we recall that d_{2n} is the empirical pseudo-distance $d_{2n}(f_1,f_2)=\sqrt{P_{2n}(f_1-f_2)^2}$. Finally, let

$$\Delta_{n,j}(f) := P'_n[f-p_j(f)] - P_n[f-p_j(f)],$$

$$\chi(x) := \sqrt{x}\log\log\left(4e^2/x\right),$$

and for $\beta>0$ introduce

$$K_j := K([\rho_n]_j, [\pi^{(j)}]_j) + \log[j(j+1)\beta^{-1}] \tag{15}$$

in which the leading term is the Kullback-Leibler divergence between the j-th projections of the randomized distribution ρ_n and the (j-th prior) distribution $\pi^{(j)}$.

The previous definitions are common to both transduction and induction setting. However the double sample is a totally virtual one in the induction setting.

We consider first the so-called transduction setting, or rather, a version of it. In this setting, we recall that the two independent samples $Z_1,\dots,Z_n$ and $Z'_1,\dots,Z'_n$ are drawn i.i.d. according to the unknown probability P.

The learning algorithm is allowed to use the instances of both samples (training and testing samples) but has access to the labels of the training instances only. Its goal is to correctly predict the instances of the testing samples. In this context, the quantity of interest is the difference between the average misclassification error obtained on the testing sample and the one on the training sample.

Theorem 2 (Transduction) *Let $0<\beta\le 2e^{-1}$. If the following condition holds*

$$\lim_{j\to+\infty} \sup_{f\in\mathcal{F}} \Delta_{n,j}(f) = 0, \qquad \mathbb{P}^{2n}\text{-a.s.} \tag{16}$$

then, with $\mathbb{P}^{2n}$-probability at least $1-\beta$, for any distribution ρ_n, we have

$$\rho_n P'_n f - P'_n f_0 \leq \rho_n P_n f - P_n f_0 + \frac{2\sqrt{2}e^{\frac{1}{4}}}{\sqrt{n}} \sum_{j=1}^{+\infty} \sqrt{K_j \rho_n d_j^2} + \frac{2\sqrt{2}e^{\frac{1}{4}}}{\sqrt{n}} \sum_{j=1}^{+\infty} \chi\Big(\frac{\rho_n d_j^2}{K_j}\Big).$$

In the induction setting, the learning algorithm is only allowed to use the first sample $Z_1, \ldots, Z_n$. Nevertheless the proof technique uses an independent and i.i.d. virtual sample $Z'_1, \ldots, Z'_n$ that is at the origin of the following expectations $\mathbb{E}'^n$.

Theorem 3 (Induction) *With the above notation, if the following condition holds*

$$\limsup_{j \to +\infty} \sup_{f \in \mathcal{F}} \mathbb{E}'^n \Delta_{n,j}(f) \leq 0, \qquad \mathbb{P}^n\text{-a.s.} \tag{17}$$

then for any $0 < \beta \leq 0.73$, with $\mathbb{P}^n$-probability at least $1-\beta$, we have

$$\rho_n P f - P f_0 \leq \rho_n P_n f - P_n f_0 + \tfrac{3.7}{\sqrt{n}} \textstyle\sum_{j=1}^{+\infty} \sqrt{\mathbb{E}'^n K_j \, \mathbb{E}'^n \rho_n d_j^2} + \tfrac{3.7}{\sqrt{n}} \textstyle\sum_{j=1}^{+\infty} \chi\Big(\frac{\mathbb{E}'^n \rho_n d_j^2}{\mathbb{E}'^n K_j}\Big).$$

Remark 1 *The second sum in the bound is in general negligible w.r.t. the first one*[2] *and has at worse the same order.*

Remark 2 *Let $\mathcal{G}$ be a model (i.e., a set of prediction functions). Let $\tilde{g}$ be a reference function (not necessarily in $\mathcal{G}$ and possibly depending on the data in an exchangeable way). Consider the class of regret functions $\mathcal{F} = \{z \mapsto L[g(x), y] : g \in \mathcal{G} \cup \{\tilde{g}\}\}$. Define $f_0 = L[\tilde{g}(x), y]$. The induction (resp. transduction) theorem compares the risk (resp. the risk on the second sample) of any (randomized) estimator with the risk (resp. the risk on the second sample) of the reference function $\tilde{g}$.*

Remark 3 *Assumption* (16) *is not very restrictive. For instance, it is satisfied when one of the following condition holds:*

- *there exists $J \in \mathbb{N}^*$ such that $S_J = \mathcal{F}$,*
- *almost surely* $\lim_{j \to +\infty} \sup_{f \in \mathcal{F}} \|f - p_j(f)\|_\infty = 0$ *(it is in particular the case when the bracketing entropy of the set $\mathcal{F}$ is finite for any radius and when the S_j's and p_j's are appropriately built on the bracketing nets of radius going to 0 when $j \to +\infty$),*
- *almost surely* $\lim_{j \to +\infty} \sup_{f \in \mathcal{F}} d_{2n}(f, p_j(f)) = 0$.[3]

Finally, by Lebesgue's dominated convergence theorem and standard probabilistic arguments, one may prove that (16) *implies* (17).

Remark 4 *Note that $\mathbb{E}'^n \rho_n d_j^2 = \frac{1}{2}\big\{\rho_n d_n^2[p_j(f), p_{j-1}(f)] + \rho_n d^2[p_j(f), p_{j-1}(f)]\big\}$. Besides, when the quantities $\pi^{(j)}$, S_j, f_0 and p_j do not depend on the data, we have $\mathbb{E}'^n K_j = K_j$, so that the conditional expectation $\mathbb{E}'^n$ disappears in Theorem 3.*

Remark 5 *A slightly different version of Theorem 3 can be obtained by using Theorem 2 and Lemma 19.*

2. Since K_j is greater than or equal to 1 and the function χ is an upper bounded function which behaves as the square root near 0.
3. Here we give a typical construction for which this assumption is satisfied when the functions in $\mathcal{F}$ take their values in a common finite set (e.g., classification setting). Take the sequence S_j as embedded nets w.r.t. the pseudo-distance d_{2n} of radius tending to 0 when j goes to infinity. Then there exists J satisfying $\big|\big(f(X_1), \ldots, f(X_n), f(X'_1), \ldots, f(X'_n)\big) : f \in S_J\big| = \big|\big(f(X_1), \ldots, f(X_n), f(X'_1), \ldots, f(X'_n)\big) : f \in \mathcal{F}\big|$ and $S_j = S_J$ for any $j > J$. Consider projections p_j consistent with d_{2n} to the extent that, if $d_{2n}(f, S_j) = 0$, then $d_{2n}(f, p_j(f)) = 0$. Then the assumption holds.

4. Discussion

We now present in which sense the result presented above combines several previous improvements in a single bound. The discussion will also clarify how to choose the priors $\pi^{(j)}$, and the sets S_j by giving more explicit bounds for some particular choices.

4.1 Supremum Bounds

We first show that we can derive from Theorem 3 a result similar to the generic chaining bound (9).

Corollary 4 *Under Assumption* (17), *we have with* $\mathbb{P}^n$*-probability at least* $1-\beta$,

$$\forall f \in \mathcal{F}, Pf - P_n f - Pf_0 + P_n f_0 \\ \leq C\left(\frac{1}{\sqrt{n}}\sum_{j\geq j(f)} 2^{-j}\sqrt{\mathbb{E}'^n \log\{1/\pi^{(j)}[A_j(f)]\}} + \sqrt{\frac{\log 2j(f)+\log(\beta^{-1})}{2^{j(f)}n}}\right),$$

where $j(f) := \min\{j \in \mathbb{N}^* : p_j(f) \neq f_0\}$.

Proof Choose for ρ_n a distribution concentrated at the single function f_n. Then we have $[\rho_n]_j = \delta_{p_j(f_n)}$ and K_j, defined in (15), reduces to $\log\{1/\pi^{(j)}[A_j(f_n)]\} + \log[j(j+1)\beta^{-1}]$. We can take the partitions $\mathcal{A}_j$ to have diameter 2^{-j} so that $\sup_{f\in\mathcal{F}} d_{2n}[f, p_j(f)] \leq 2^{-j}$ and thus $d_j \leq 2^{-j+1}$. ■

The "closer" the function f is from f_0, the bigger the integer $j(f)$ is. This result improves on (9) since it is algorithm dependent (the l.h.s. depends on f_n) and takes into account the variance. Since the series starts from $j(f)$, our bound is better when f is "close" to f_0.

Since we essentially have a result that is as powerful as generic chaining, it should be possible to recover Rademacher averages bounds. However, there is a difficulty coming from the fact that the result we have is not "symmetric" in the sense that it involves taking expectations over the second sample. Taking care of this remains a topic for further research.

4.2 Variance Bounds

We now show how the variance of the function of interest can be obtained explicitly in the upper bound. In particular, we have the following corollary.

Corollary 5 *Under Assumption* (17), *if the functions in* $\mathcal{F}$ *have values in* $\{0;1\}$, *we have with* $\mathbb{P}^n$*-probability at least* $1-\beta$,

$$\forall f \in \mathcal{F}, Pf - P_n f - Pf_0 + P_n f_0 \\ \leq \frac{C}{\sqrt{n}}\sqrt{\mathbb{E}'^n P_{2n}(f-f_0)^2 (\mathbb{E}'^n N(\mathcal{F}, 1/2n, d_{2n}) + \log(\beta^{-1}))}.$$

Proof As before, choose for ρ_n a distribution concentrated at a single function f_n, then $[\rho_n]_j = \delta_{p_j(f_n)}$ and thus $K_j = \log 1/\pi^{(j)}[A_j(f_n)]$. Now since the functions are binary valued, $\mathcal{F}$ is a finite metric space under the metric d_{2n}. We can thus take $\mathcal{A}_1$ as a minimal cover at radius $1/2n$, $\mathcal{A}_2 = \mathcal{F}$ and we will get $d_2(f) = 0$ and $d_1(f) = P_{2n}(f-f_0)^2$. Taking $\pi^{(j)}$ to be uniform on the centers of the cover, we will obtain $K_1 = \log N(\mathcal{F}, 1/2n, d_{2n}) + \log(\beta^{-1})$ which gives the result. ■

We thus essentially recover (12) and a fortiori standard VC bounds.

4.3 Averaging

Theorem 2 also includes the PAC-Bayesian improvement for averaging classifiers since if one considers the set $S_1 = \mathcal{F}$ one recovers a result similar to McAllester's (14). More precisely, we obtain:

Corollary 6 *Let $\mathcal{F}$ be a set of functions taking their values in $[-1;1]$. Let $0 < \beta \leq 0.73$ and $\mathcal{K} := K(\rho_n, \pi) + \log(2\beta^{-1})$. With $\mathbb{P}^{2n}$-probability at least $1-\beta$, we have*

$$\rho_n P'_n f - \rho_n P_n f \leq \frac{3.7}{\sqrt{n}} \sqrt{\mathcal{K} \rho_n P_{2n} f^2} + \frac{3.7}{\sqrt{n}} \chi\left(\frac{\rho_n P_{2n} f^2}{\mathcal{K}}\right).$$

As a consequence, for any $\beta > 0$, with $\mathbb{P}^n$-probability at least $1-\beta$, we have

$$\rho_n P f - \rho_n P_n f \leq C \sqrt{\frac{K(\rho_n, \pi) + \log(2\beta^{-1})}{n}}.$$

Note that this last inequality is slightly better than (14) since there is no extra logarithmic term. However the cost of removing the logarithmic factor is to have a slightly bigger constant C.

Proof Even if it means expanding $\mathcal{F}$, we may assume that the function f_0 identically equal to 0 is in $\mathcal{F}$. Then we can take $S_0 := \{f_0\}$ and $S_1 := \mathcal{F}$. We have for any $j \geq 1$, $S_j = \mathcal{F}$ and $p_j = \mathrm{Id}_{\mathcal{F}}$. The first assertion is then a direct application of Theorem 2.

From the first result of the corollary, for any $0 < \beta \leq 0.73$, with $\mathbb{P}^{2n}$-probability at least $1-\beta$, for any distribution ρ_n, we have

$$\rho_n P'_n f - \rho_n P_n f \leq 3.7 \sqrt{\frac{K(\rho_n, \pi) + \log(2\beta^{-1})}{n}} + \frac{3.7 \max_{[0;1]} \chi}{\sqrt{n}}.$$

The second assertion of the corollary then follows. ■

4.4 Data-dependent Bounds

The obtained bound is not completely empirical since it involves the expectation with respect to an extra sample. In the transduction setting, this is not an issue, it is even an advantage as one can use the unlabelled data in the computation of the bound. However, in the inductive setting, this is a drawback. Future work will focus on using concentration inequalities to give a fully empirical bound.

5. Conclusion and Perspectives

We have obtained a generalization error bound for randomized classifiers which combines several previous improvements. It contains an optimal union bound, both in the sense of optimally taking into account the metric structure of the set of functions (via the majorizing measure approach) and in the sense of taking into account the averaging distribution. It also is sensitive to the variance of the functions and is thus "localized".

In particular it is the first PAC-Bayesian bound that remains finite when the averaging distribution is concentrated at a point.

There still remains work in order to get a fully empirical bound and to better understand the connection with Rademacher averages. In particular, the way the approximating sets S_j should be constructed in practical cases has to be investigated.

Acknowledgments

We would like to thank Gilles Blanchard and the referees for their numerous useful comments and suggestions that have greatly improved the first version of this work.

Appendix A. Proof of Our Main Results

The proof of our main results is inspired by previous work on the PAC-Bayesian bounds (Catoni, 2003), Audibert (2004b) and on the generic chaining (Talagrand, 1996). Classical PAC-Bayesian bounds are mainly based on the combination of three ingredients:

1. the duality property of the entropy (Lemma 7),
2. Markov's inequality (which leads to Lemma 8),
3. and upper bounds on the exponential moment of a bounded random variable (which is the underlying idea of Lemma 9).

We reused these three main ingredients and slightly extended them to fit our needs. The additional extra step that is required to obtain the generic chaining aspect is to decompose the functions into a *chain* (as explained in Section 2.3.1) and to combine together all the individual PAC-Bayesian bounds obtained for each element of the chain. This chaining part is actually done in two steps: in the first one (Section A.1.4) we perform a non uniform union bound with appropriately chosen weights in order to make the bound independent on λ, while in the second one (Section A.1.5), we actually chain the inequalities to obtain the final result.

A.1 Proof of Theorem 2

In order to emphasize the various techniques that are used, we decompose its proof in a series of short steps.

A.1.1 Legendre Transform of the Kullback-Leibler Divergence

The first step consists in using a duality property of the relative entropy (see, for example, Dembo and Zeitouni 1998, or page 10 of Catoni 2003). Namely, one has the following

Lemma 7 (Legendre transform of the KL-divergence) *For any π-measurable function $h : \mathcal{F} \to \mathbb{R}$ and any probability distribution ρ on $\mathcal{F}$,*

$$\rho h \leq \log \pi e^h + K(\rho, \pi).$$

Proof We have $\rho h - K(\rho,\pi) - \log \pi e^h = -K\left(\rho, \frac{e^h}{\pi e^h} \cdot \pi\right) \leq 0$ with equality when $\rho = \frac{e^h}{\pi e^h} \cdot \pi$. ■

Lemma 8 *Consider a function $h : \mathcal{F} \to \mathbb{R}$ and probability distributions π and ρ on $\mathcal{F}$ that depend on $Z_1, \ldots, Z_n, Z'_1, \ldots, Z'_n$ in a measurable way, we have for any $\beta > 0$, with probability at least $1-\beta$ with respect to the samples distribution*

$$\rho h \leq \log \mathbb{E}^{2n} \pi e^h + K(\rho, \pi) + \log(\beta^{-1}).$$

Also, with probability at least $1-\beta$ *with respect to the first sample distribution,*

$$\mathbb{E}'^n \rho h \leq \log \mathbb{E}^{2n} \pi e^h + \mathbb{E}'^n K(\rho, \pi) + \log(\beta^{-1}).$$

Proof From Markov's inequality applied to the non-negative random variable πe^h, we obtain that for any $t > 0$ $\mathbb{P}\left(\pi e^h > t\right) \leq \frac{1}{t}\mathbb{E}^{2n}\pi e^h$, hence for any $\beta > 0$, with probability at least $1-\beta$ with respect to the samples distribution,

$$\log \pi e^h \leq \log \mathbb{E}^{2n} \pi e^h + \log(\beta^{-1}).$$

Then the proof of the first result follows from Lemma 7. The second assertion can be proved in a similar way (applying Markov's inequality conditionally to the second sample) and using the inequality $\mathbb{E}'^n \log \mathbb{E}^n .. \leq \log \mathbb{E}^{2n} ...$ ■

A.1.2 DEVIATION INEQUALITY

Lemma 9 *For any* $\lambda > 0$*, any function* $\mathcal{W} : \mathcal{F} \times \mathcal{Z} \to \mathbb{R}$ *and any exchangeable function* $\pi : \mathcal{X}^{2n} \to \mathcal{M}_+^1(\mathcal{F})$*, we have*

$$\mathbb{E}^{2n} \pi e^{\lambda P_n' \mathcal{W} - \lambda P_n \mathcal{W} - \frac{2\lambda^2}{n} P_{2n} \mathcal{W}^2} \leq 1.$$

Proof Denote $\Delta_i := \mathcal{W}(\cdot, Z_i') - \mathcal{W}(\cdot, Z_i)$ and

$$h := \lambda P_n' \mathcal{W} - \lambda P_n \mathcal{W} - \frac{2\lambda^2}{n} P_{2n} \mathcal{W}^2.$$

By the exchangeability of π, for any $\sigma \in \{-1;+1\}^n$, we have

$$\begin{aligned} \mathbb{E}^{2n} \pi e^h &= \mathbb{E}^{2n} \pi e^{-\frac{2\lambda^2}{n} P_{2n} \mathcal{W}^2 + \frac{\lambda}{n} \sum_{i=1}^n \Delta_i} \\ &= \mathbb{E}^{2n} \pi e^{-\frac{2\lambda^2}{n} P_{2n} \mathcal{W}^2 + \frac{\lambda}{n} \sum_{i=1}^n \sigma_i \Delta_i}. \end{aligned}$$

Now taking the expectation w.r.t. σ, where σ is a n-dimensional vector of Rademacher variables. We obtain

$$\begin{aligned} \mathbb{E}^{2n} \pi e^h &= \mathbb{E}^{2n} \pi \left[e^{-\frac{2\lambda^2}{n} P_{2n} \mathcal{W}^2} \prod_{i=1}^n \cosh\left(\frac{\lambda}{n} \Delta_i\right) \right] \\ &\leq \mathbb{E}^{2n} \pi \left[e^{-\frac{2\lambda^2}{n} P_{2n} \mathcal{W}^2} e^{\sum_{i=1}^n \frac{\lambda^2}{2n^2} \Delta_i^2} \right] \end{aligned}$$

where at the last step we use that $\cosh s \leq e^{\frac{s^2}{2}}$. The result follows from the inequality $\Delta_i^2 \leq 2\mathcal{W}^2(\cdot, Z_i') + 2\mathcal{W}^2(\cdot, Z_i)$. ■

A.1.3 CONSEQUENCES

We first prove the following lemma.

Lemma 10 *For any* $\beta > 0$*,* $\lambda > 0$*, any function* $\mathcal{W} : \mathcal{F} \times \mathcal{Z} \to \mathbb{R}$ *and any exchangeable function* $\pi : \mathcal{X}^{2n} \to \mathcal{M}_+^1(\mathcal{F})$*, with* $\mathbb{P}^n$*-probability at least* $1-\beta$*, for any probability distribution* $\rho_n \in \mathcal{M}_+^1(\mathcal{F})$*, we have*

$$\rho_n P_n' \mathcal{W} - \rho_n P_n \mathcal{W} \leq \frac{2\lambda}{n} \rho_n P_{2n} \mathcal{W}^2 + \frac{K(\rho_n, \pi) + \log(\beta^{-1})}{\lambda}.$$

Proof The result follows from Lemmas 8 and 9 applied to $h := \lambda P_n' \mathcal{W} - \lambda P_n \mathcal{W} - \frac{2\lambda^2}{n} P_{2n} \mathcal{W}^2$ and $\rho = \rho_n$. ■

Now let us apply this result to the projected measures $[\pi^{(j)}]_j$ and $[\rho_n]_j$ and for $\mathcal{W}(f,Z) = p_j(f)(Z) - p_{j-1}(f)(Z)$. Since, by definition, $\pi^{(j)}$, S_j and p_j are exchangeable, $[\pi^{(j)}]_j$ is also exchangeable. With $\mathbb{P}^{2n}$-probability at least $1-\beta$, uniformly in ρ_n, we have

$$[\rho_n]_j \Big\{ P_n'[p_j(f) - p_{j-1}(f)] - P_n[p_j(f) - p_{j-1}(f)] \Big\} \leq \frac{2\lambda}{n} [\rho_n]_j d_{2n}^2 [p_j(f), p_{j-1}(f)] + \frac{K_j'}{\lambda},$$

where $K_j' := K([\rho_n]_j, [\pi^{(j)}]_j) + \log(\beta^{-1})$. By definition of $[\rho_n]_j$, it implies that

$$\rho_n \Big\{ P_n'[p_j(f) - p_{j-1}(f)] - P_n[p_j(f) - p_{j-1}(f)] \Big\} \leq \frac{2\lambda}{n} \rho_n d_{2n}^2 [p_j(f), p_{j-1}(f)] + \frac{K_j'}{\lambda},$$

which, by using the notation introduced on page 874, can be shortened into

$$\rho_n \Delta_{n,j} \leq \frac{2\lambda}{n} \rho_n d_j^2 + \frac{K_j'}{\lambda}, \tag{18}$$

The parameter $\lambda = \sqrt{\frac{nK_j'}{2\rho_n d_j^2}}$ minimizing the r.h.s. of the previous inequality depends on ρ_n. To use this data-dependent parameter, we need that (18) holds uniformly in λ.

A.1.4 WEIGHTED UNION BOUND ON THE PARAMETER λ

To get a uniform version of (18), introduce a grid $(\lambda_k)_{k\in\mathbb{N}^*}$ of $\mathbb{R}_+^*$. Let $(w_k)_{k\in\mathbb{N}^*}$ denote some positive real numbers such that $\sum_{k\geq 1} w_k = 1$. Define $\mathcal{B} := \inf_{k\geq 1} \left\{ \frac{2\rho_n d_j^2}{n} \lambda_k + \frac{K_j' + \log w_k^{-1}}{\lambda_k} \right\}$. Using a weighted union bound of (18), precisely using (18) for $(\lambda, \beta) = (\lambda_k, w_k \beta)$ for $k \in \mathbb{N}^*$, we get that, with probability at least $1-\beta$, we have $\rho_n \Delta_{n,j} \leq \mathcal{B}$.

Our goal is then to choose the λ_k's and the w_k's such that $\mathcal{B}$ is the smallest possible. Ideally, we want to obtain a bound close to $a_\lambda := \min_{\lambda\in\mathbb{R}_+^*} \left\{ \frac{2\rho_n d_j^2}{n} \lambda + \frac{K_j'}{\lambda} \right\}$.

Let $m := e^{-\frac{1}{4}} \sqrt{\frac{n}{8}}$. Taking $\lambda_k = m e^{\frac{k}{2}}$ and $w_k = \frac{1}{k(k+1)}$, we are going to prove that we achieve this target up to a multiplicative constant and an additive $\log\log$ term.

First, since we have $\rho_n \Delta_{n,j} = 0$ when $\rho_n d_j^2 = 0$, we only focus on the case when $\rho_n d_j^2 > 0$. Define $\lambda^* := 2m \sqrt{\frac{K_j'}{\rho_n d_j^2}}$. We have $\rho_n d_j^2 \leq 4$ and $K_j' \geq 1$ for $\beta \leq e^{-1}$, hence $\lambda^* \geq m$. So there exists $k^* \in \mathbb{N}^*$ such that $\lambda_{k^*} e^{-\frac{1}{2}} \leq \lambda^* < \lambda_{k^*}$. We have

$$\begin{aligned}
\mathcal{B} &\leq \frac{2\rho_n d_j^2}{n} \lambda_{k^*} + \frac{K_j' + \log w_{k^*}^{-1}}{\lambda_{k^*}} \\
&\leq \frac{2e^{\frac{1}{2}} \rho_n d_j^2}{n} \lambda^* + \frac{K_j' + \log[k^*(k^*+1)]}{\lambda^*} \\
&\leq 2\sqrt{2} e^{\frac{1}{4}} \sqrt{\frac{\rho_n d_j^2 K_j'}{n}} + \frac{\log[k^*(k^*+1)]}{\lambda^*}.
\end{aligned}$$

The inequality $\lambda_{k^*} e^{-\frac{1}{2}} \le \lambda^*$ implies $k^* - 1 \le 2\log\left(\frac{\lambda^*}{m}\right)$, hence $k^* + 1 \le \log\left(4e^2 \frac{K'_j}{\rho_n d_j^2}\right)$. Finally we have proved that with probability at least $1-\beta$,

$$\rho_n \Delta_{n,j} \le 2\sqrt{2} e^{\frac{1}{4}} \sqrt{\frac{\rho_n d_j^2 K'_j}{n}} + \frac{2\sqrt{2} e^{\frac{1}{4}}}{\sqrt{n}} \chi\Big(\frac{\rho_n d_j^2}{K'_j}\Big).$$

We recall that $K'_j := K([\rho_n]_j, [\pi^{(j)}]_j) + \log(\beta^{-1})$ and $\chi(x) := \sqrt{x}\log\log\left(4e^2/x\right)$.

A.1.5 CHAINING THE INEQUALITIES

By simply using a union bound with weights equal to $\frac{1}{j(j+1)}$, the previous inequality holds[4] uniformly in $j \in \mathbb{N}^*$ provided that β is replaced with $\beta/[j(j+1)]$, hence to apply the result of the previous section we need that $\beta/2 \le e^{-1}$.

Since $p_{j-1} = p_{j-1} \circ p_j$, we have

$$\begin{aligned}
&\rho_n\big[P'_n f - P'_n f_0 + P_n f_0 - P_n f\big] \\
&\qquad = \rho_n \Delta_{n,J}(f) + \rho_n\Big\{\textstyle\sum_{j=1}^J \big[(P'_n - P_n)p_j(f) - (P'_n - P_n)p_{j-1}(f)\big]\Big\} \\
&\qquad = \rho_n \Delta_{n,J}(f) + \textstyle\sum_{j=1}^J \rho_n\big[(P'_n - P_n)p_j(f) - (P'_n - P_n)p_{j-1}(f)\big] \\
&\qquad = \rho_n \Delta_{n,J}(f) + \textstyle\sum_{j=1}^J [\rho_n]_j\big[(P'_n - P_n)f - (P'_n - P_n)p_{j-1}(f)\big]
\end{aligned}$$

Setting $K_j := K([\rho_n]_j, [\pi^{(j)}]_j) + \log[j(j+1)\beta^{-1}]$, with $\mathbb{P}^n$-probability at least $1-\beta$, for any distribution ρ_n, we have

$$\begin{aligned}
\rho_n\big[P'_n f - P'_n f_0 + P_n f_0 - P_n f\big] \le \sup_{\mathcal{F}} \Delta_{n,J} + 2\sqrt{2} e^{\frac{1}{4}} \textstyle\sum_{j=1}^J \sqrt{\frac{\rho_n d_j^2 K j}{n}} \\
+ \frac{2\sqrt{2} e^{\frac{1}{4}}}{\sqrt{n}} \textstyle\sum_{j=1}^J \chi\Big(\frac{\rho_n d_j^2}{K_j}\Big).
\end{aligned}$$

Making $J \to +\infty$, we obtain Theorem 2.

A.2 Proof of Theorem 3

It suffices to modify Lemma 10 in the proof of Theorem 2. Indeed, using the second part of Lemma 8 instead of the first one we get

$$\rho_n P \mathcal{W} - \rho_n P_n \mathcal{W} \le \frac{2\lambda}{n} \mathbb{E}'^n \rho_n P_{2n} \mathcal{W}^2 + \frac{\mathbb{E}'^n K(\rho_n, \pi) + \log(\beta^{-1})}{\lambda}.$$

The remaining parts of the proof (i.e., the union bound and the chaining) are similar.

Appendix B. Additional Material

Lemma 11 *For any random variable Z_f which is $\pi \otimes P$ measurable we have*

$$\log \pi e^{\mathbb{E}[Z_f]} \le \mathbb{E}\left[\log \pi e^{Z_f}\right] \le \log \pi \mathbb{E}\left[e^{Z_f}\right].$$

4. This is because $\sum_{j \in \mathbb{N}^*} \frac{1}{j(j+1)} = 1$.

Proof By duality (Lemma 7), we have

$$\mathbb{E}\left[\log \pi e^{Z_f}\right] = \mathbb{E}\left[\sup_{\rho_n}\left\{\rho_n Z_f - K(\rho_n,\pi)\right\}\right] \geq \sup_{\rho_n} \rho_n \mathbb{E}\left[Z_f\right] - K(\rho_n,\pi) = \log \pi e^{\mathbb{E}\left[Z_f\right]}.$$

This gives the first inequality. The second inequality follows from Jensen's inequality (applied to the convex function $-\log$) and Fubini's theorem. ■

B.1 Concentration Inequalities

In this section, we recall some concentration inequalities whose proofs can be found in Lugosi (2003). We start with Markov's inequality

Theorem 12 (Markov's inequality) *For any real-valued random variable X,*

$$\mathbb{P}(X >= t) \leq e^{-t}\mathbb{E}\left[e^X\right].$$

Theorem 13 (Hoeffding's inequality, 1963) *For any centered random variable X such that $a \leq X \leq b$, and any $\lambda > 0$, we have $\mathbb{E}e^{\lambda X} \leq e^{\frac{\lambda^2(b-a)^2}{8}}$. As a consequence, for any i.i.d. random variables $X_1,\ldots,X_n$ such that $a \leq X_i \leq b$, we have*

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n X_i > t\right) \leq e^{-\frac{2nt^2}{(b-a)^2}}.$$

Theorem 14 (Bernstein's inequality) *Let $X_1,\ldots,X_n$ be n i.i.d. centered random variables such that $a \leq X_i \leq b$. We have*

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n X_i > t\right) \leq e^{-\frac{nt^2}{2\operatorname{Var}X + \frac{2(b-a)t}{3}}}.$$

We say that a function has the bounded differences property if for some constant $c > 0$

$$\sup_{x_1,\ldots,x_n,x'\in\mathcal{X},i\in\{1,\ldots,n\}} \left|g(x_1,\ldots,x_n) - g(x_1,\ldots,x_{i-1},x',x_{i+1},\ldots,x_n)\right| \leq c.$$

Theorem 15 (McDiarmid's inequality) *Let g satisfy the bounded differences assumption with constant c. Then we have*

$$\mathbb{P}\left[g(X_1,\ldots,X_n) - \mathbb{E}g(X_1,\ldots,X_n) \geq t\right] \leq e^{-\frac{2t^2}{nc^2}}.$$

Note that McDiarmid's result generalizes Hoeffding's inequality.

B.2 Symmetrization Inequalities

Lemma 16 (Symmetrization in probability, Vapnik and Chervonenkis 1971) *Assume the functions in $\mathcal{F}$ have range in $[a;b]$. For any $t > 0$ such that $nt^2 \geq 2(b-a)^2$,*

$$\mathbb{P}^n\left(\sup_{f\in\mathcal{F}}\left\{Pf - P_n f\right\} \geq t\right) \leq 2\mathbb{P}^{2n}\left(\sup_{f\in\mathcal{F}}\left\{P'_n f - P_n f\right\} \geq \frac{t}{2}\right).$$

Proof Let f_n be the function (depending on the first sample) achieving the supremum of $(P-P_n)f$ over $\mathcal{F}$ (if it does not exist, one can use a limiting argument). We have $\mathbb{1}_{(P-P_n)f_n>t}\mathbb{1}_{(P-P'_n)f_n<\frac{t}{2}} \leq \mathbb{1}_{(P'_n-P_n)f>\frac{t}{2}}$. Taking expectations w.r.t. the second sample gives $\mathbb{1}_{(P-P_n)f_n>t}\mathbb{P}'^n[(P-P'_n)f_n<\frac{t}{2}] \leq \mathbb{P}'^n[(P'_n-P_n)f>\frac{t}{2}]$. Now by Chebyshev's inequality, we have $\mathbb{P}'^n[(P-P'_n)f_n \geq \frac{t}{2}] \leq \frac{4\operatorname{Var} f_n}{nt^2} \leq \frac{(b-a)^2}{nt^2} \leq \frac{1}{2}$. We obtain $\mathbb{1}_{(P-P_n)f_n>t} \leq 2\mathbb{P}'^n[(P'_n-P_n)f>\frac{t}{2}]$ and conclude by taking expectations w.r.t. the first sample. ■

Remark 6 *By replacing Chebyshev's inequality with Bernstein's inequality, we can improve the previous result to take into account t of order smaller than $1/\sqrt{n}$. One can also slightly generalize the previous result to obtain: for any positive reals η and t,*

$$\mathbb{P}^n\left(\sup_{f\in\mathcal{F}}\left\{Pf-P_nf\right\}\geq t\right) \leq \frac{\mathbb{P}^{2n}\left(\sup_{f\in\mathcal{F}}\left\{P'_nf-P_nf\right\}\geq(1-\eta)t\right)}{1-e^{-\frac{n\eta^2t^2}{2\operatorname{Var}f+2(b-a)\eta t/3}}}.$$

Lemma 17 (Symmetrization for expectations, Giné and Zinn 1984) *For any set $\mathcal{F}$ of functions,*

$$\mathbb{E}^n\sup_{f\in\mathcal{F}}\left\{Pf-P_nf\right\} \leq \frac{2}{n}\mathbb{E}^n\mathbb{E}_\sigma\sup_{f\in\mathcal{F}}\sum_{i=1}^n\sigma_i f(Z_i).$$

Proof We have

$$\begin{aligned}
\mathbb{E}^n\sup_{f\in\mathcal{F}}\left\{Pf-P_nf\right\} &= \mathbb{E}^n\sup_{f\in\mathcal{F}}\left\{\mathbb{E}'^n[P'_nf]-P_nf\right\}\\
&\leq \mathbb{E}^{2n}\sup_{f\in\mathcal{F}}\left\{P'_nf-P_nf\right\}\\
&= \mathbb{E}^{2n}\mathbb{E}_\sigma\sup_{f\in\mathcal{F}}\tfrac{1}{n}\textstyle\sum\sigma_i(f(Z'_i)-f(Z_i))\\
&\leq \tfrac{2}{n}\mathbb{E}^n\mathbb{E}_\sigma\sup_{f\in\mathcal{F}}\textstyle\sum\sigma_i f(Z_i).
\end{aligned}$$

■

The following lemmas, due to Panchenko, allow to convert transductive bounds into inductive ones.

Lemma 18 *For any function $B:\mathcal{Z}^{2n}\to\mathbb{R}$, if for any $\beta>0$, with $\mathbb{P}^{2n}$-probability at least $1-\beta$, we have $B\leq\log(\beta^{-1})$, then for any $\beta>0$, with $\mathbb{P}^n$-probability $1-e\beta$, we have*

$$\mathbb{E}'^nB\leq\log(\beta^{-1}).$$

Proof It directly comes from Lemma 1 in Panchenko (2003). ■

Lemma 19 *Let B_1, B_2 and B_3 be three functions of $Z_1,\ldots,Z_n$ and $Z'_1,\ldots,Z'_n$ with $B_2\geq0$ and $B_3\geq0$. If for any $\beta>0$, with $\mathbb{P}^{2n}$-probability at least $1-\beta$, we have*

$$B_1\leq\sqrt{B_2(B_3+\log(\beta^{-1}))}$$

then for all $\beta > 0$, *with* $\mathbb{P}^n$*-probability* $1 - e\beta$,

$$\mathbb{E}'^n B_1 \leq \sqrt{\mathbb{E}'^n B_2 \left[\mathbb{E}'^n B_3 + \log(\beta^{-1})\right]}.$$

Proof It suffices to modify slightly the proof of Corollary 1 in Panchenko (2003). Specifically, we apply Lemma 18 to the quantity $B := \sup_{\lambda>0} \left\{4\lambda(B_1 - \lambda B_3) - B_2\right\}$ since simple computations show that $\left\{B \geq \log(\beta^{-1})\right\} = \left\{B_1 \geq \sqrt{B_3(B_2 + \log(\beta^{-1}))}\right\}$.

■

B.3 Proof of Known Results

Here we prove the results presented in the survey section (see Section 2).

B.3.1 PROOF OF INEQUALITY (5)

For any $t \in \mathbb{R}$, Hoeffding's inequality (see Section B.1) implies $\mathbb{P}^n \left[Pf - P_n f > t\right] \leq e^{-2nt^2}$. Choosing $\beta = e^{-2nt^2}$, we obtain

$$Pf - P_n f \leq \frac{1}{\sqrt{2}} \sqrt{\frac{\log(\beta^{-1})}{n}}.$$

B.3.2 PROOF OF INEQUALITY (6)

Statement (5) holds uniformly over $|\mathcal{F}|$ functions with probability at least $1 - |\mathcal{F}|\beta$. Setting $\beta' = |\mathcal{F}|\beta$, we obtain the desired result.

B.3.3 PROOF OF INEQUALITY (7) (VAPNIK AND CHERVONENKIS, 1971)

First, use the symmetrization lemma 16. Denote $\mathcal{F}_{Z,Z'}$ the set of vectors formed by the values of each function in $\mathcal{F}$ on the double sample. Let σ_i be independent random signs ($+1, -1$ with probability $1/2$). We have

$$\begin{aligned}
\mathbb{P}^{2n}\left(\sup_{f\in\mathcal{F}} \left\{P'_n f - P_n f\right\} \geq \tfrac{t}{2}\right) &= \mathbb{P}^{2n}\left(\sup_{f\in\mathcal{F}_{Z,Z'}} \left\{P'_n f - P_n f\right\} \geq \tfrac{t}{2}\right) \\
&= \mathbb{P}_\sigma \mathbb{P}^{2n}\left(\sup_{f\in\mathcal{F}_{Z,Z'}} \tfrac{1}{n}\textstyle\sum_{i=1}^n \sigma_i [f(Z_i) - f(Z'_i)] \geq \tfrac{t}{2}\right) \\
&= \mathbb{E}^{2n}\mathbb{E}_\sigma \sup_{f\in\mathcal{F}_{Z,Z'}} \mathbb{1}_{\frac{1}{n}\sum_{i=1}^n \sigma_i [f(Z_i) - f(Z'_i)] \geq \frac{t}{2}} \\
&\leq \mathbb{E}^{2n}\mathbb{E}_\sigma \sum_{f\in\mathcal{F}_{Z,Z'}} \mathbb{1}_{\frac{1}{n}\sum_{i=1}^n \sigma_i [f(Z_i) - f(Z'_i)] \geq \frac{t}{2}} \\
&\leq \mathbb{E}^{2n}|\mathcal{F}_{Z,Z'}| \sup_{b\in\{-1;0;+1\}^n} \mathbb{P}_\sigma\left(\tfrac{1}{n}\textstyle\sum_{i=1}^n b_i\sigma_i \geq \tfrac{t}{2}\right) \\
&\leq e^{-\frac{nt^2}{8}} \mathbb{E}^{2n}|\mathcal{F}_{Z,Z'}|,
\end{aligned}$$

where the last step comes from Hoeffding's inequality. Putting $\beta = 2\mathbb{E}^{2n}|\mathcal{F}_{Z,Z'}| e^{-\frac{nt^2}{8}}$, we get with probability at least $1 - \beta$,

$$\forall f \in \mathcal{F}, Pf - P_n f \leq \sqrt{8} \sqrt{\frac{\log \mathbb{P}^{2n} N(\mathcal{F}, 1/n, d_{2n}) + \log(2\beta^{-1})}{n}}.$$

B.3.4 PROOF OF INEQUALITY (13) (MCALLESTER, 1998)

We use the same C as for statement (5). The probability that (13) does not hold is upper bounded with

$$\sum_{f\in\mathcal{F}} \mathbb{P}^n\left[Pf - P_n f > C\sqrt{\frac{\log 1/[\beta\pi(f)]}{n}}\right] \leq \sum_{f\in\mathcal{F}} \pi(f)\beta = \beta.$$

B.3.5 PROOF OF INEQUALITY (11)

To prove (11), we replace Hoeffding's inequality with Bernstein's inequality (see Section B.1): $\mathbb{P}^n\left[Pf - P_n f > t\right] \leq e^{-\frac{nt^2}{2\mathrm{Var}\, f + 2t/3}}$. Then choosing β equal to the right-hand side gives the result.

B.3.6 PROOF OF INEQUALITY (14) (MCALLESTER, 1999; SEEGER, 2003)

Let $c \geq 1/2$. Here we propose a slightly different proof in order to get rid of the logarithmic term in (14). Using Jensen's inequality and Lemma 7, we get

$$\left[\rho_n(Pf - P_n f)_+\right]^2 \leq \rho_n(Pf - P_n f)_+^2 \leq \frac{c}{n}\left(\log \pi e^{\frac{n}{c}(Pf-P_n f)_+^2} + K(\rho_n,\pi)\right).$$

Moreover, by Markov's inequality and Fubini's Theorem, we have

$$\mathbb{P}^n\left[\log \pi e^{\frac{n}{c}(Pf-P_n f)_+^2} > t\right] \leq e^{-t}\pi\mathbb{E}^n\left[e^{\frac{n}{c}(Pf-P_n f)_+^2}\right],$$

and

$$\begin{aligned}
\mathbb{E}^n e^{\frac{n}{c}(Pf-P_n f)_+^2} &= \int_0^{+\infty} \mathbb{P}^n\left(e^{\frac{n}{c}(Pf-P_n f)_+^2} \geq u\right) du \\
&= 1 + \int_1^{e^{n/c}} \mathbb{P}^n\left(Pf - P_n f \geq \sqrt{\tfrac{c\log u}{n}}\right) du \\
&\leq 1 + \int_1^{e^{n/c}} \frac{du}{u^{-2c}} \\
&= 1 + \frac{1-e^{-n(2c-1)/c}}{2c-1},
\end{aligned}$$

where the inequality comes from the proof of (5). Taking $\beta = 1 + \frac{1-e^{-n(2c-1)/c}}{2c-1}e^{-t}$, we obtain for $c = 1/2$, with probability at least $1-\beta$,

$$\rho_n(Pf - P_n f) \leq \frac{1}{\sqrt{2}}\sqrt{\frac{K(\rho_n,\pi) + \log(2n+1) + \log 1/\beta}{n}}$$

and for any $c > \frac{1}{2}$, with probability at least $1-\beta$,

$$\rho_n(Pf - P_n f) \leq \sqrt{c}\sqrt{\frac{K(\rho_n,\pi) + \log(\frac{2c}{2c-1}) + \log 1/\beta}{n}}.$$

B.3.7 PROOF OF INEQUALITY (10)

From Theorem 15 and Lemma 17, we obtain that with probability at least $1-\beta$,

$$\sup_{f\in\mathcal{F}}(Pf - P_n f) \leq \frac{2}{n}\mathbb{E}^n\left[\mathbb{E}_\sigma \sup_{f\in\mathcal{F}} \sum \sigma_i f(Z_i)\right] + \frac{1}{\sqrt{2}}\sqrt{\frac{\log(\beta^{-1})}{n}}. \tag{19}$$

B.3.8 PROOF OF INEQUALITY (8)

The starting point is Inequality (19). Let f_n be the function achieving the supremum. (To shorten the proof, we assume its existence.) Introduce the vectors h^* and $h^{(0)}$ such that $h_i^* = f_n(Z_i)$ and $h_i^{(0)} = 0$. Consider the canonical distance on $\mathbb{R}^n$: $\|x\| := \sqrt{\sum x_i^2}$, and take the minimal covering nets $\mathcal{N}_k$ of the set of vectors $\{[f(Z_i)]_{i=1,\dots,n} : f \in \mathcal{F}\}$ of respective radius 2^{-k} for $k = 1,\dots,K$ where $K := \lfloor \log_2 \sqrt{n} \rfloor + 1$. Let $h^{(k)}$ be a nearest neighbour of h^* in the net $\mathcal{N}_k$. Let $e_k := \mathbb{E}_\sigma \max_{(h',h'') \in \overset{=}{\mathcal{N}}_k} \sum_i \sigma_i (f_i - g_i)$ and

$$\overset{=}{\mathcal{N}}_k := \left\{ h' \in \mathcal{N}_k, h'' \in \mathcal{N}_{k-1} : \|h' - h''\| \le 32^{-k} \right\}.$$

We have

$$\begin{aligned}
\mathbb{E}_\sigma \sup_{f \in \mathcal{F}} \sum \sigma_i f(Z_i) &= \mathbb{E}_\sigma \sum \sigma_i h_i^* \\
&= \mathbb{E}_\sigma \sum \sigma_i \left(h_i^* - h_i^{(K)} \right) + \mathbb{E}_\sigma \sum_{i,k} \sigma_i \left(h_i^{(k)} - h_i^{(k-1)} \right) \\
&\le 1 + \mathbb{E}_\sigma \sum_{i,k} \sigma_i \left(h_i^{(k)} - h_i^{(k-1)} \right) \\
&\le 1 + \sum_k \mathbb{E}_\sigma \sum_i \sigma_i \left(h_i^{(k)} - h_i^{(k-1)} \right) \\
&\le 1 + \sum_k e_k.
\end{aligned}$$

From the following lemma, using that for any $(h',h'') \in \overset{=}{\mathcal{N}}_k$, $\|h' - h''\| \le 3 \times 2^{-k}$ and $|\overset{=}{\mathcal{N}}_k| \le |\mathcal{N}_k|^2$, we get $e_k \le 3 \times 2^{-k} \sqrt{4n \log N(\mathcal{F}, 2^{-k}, d_n)}$.

Lemma 20 (Pisier 1986) *Let $\sigma_i, i = 1,\dots,n$ be Rademacher variables and $c_j, j = 1,\dots,J$ be $\mathbb{R}^n$-vectors such that for any $j \in \{1,\dots,J\}$, $\|c_j\| \le c$. Then we have*

$$\mathbb{E} \max_{j \in \{1,\dots,J\}} \sum_i \sigma_i c_{j,i} \le c\sqrt{2n \log J}.$$

Proof For any $\lambda > 0$, we have

$$\mathbb{E} \max_{j \in \{1,\dots,J\}} \sum_i \sigma_i c_{j,i} \le \frac{1}{\lambda} \log \sum_j \mathbb{E} e^{\sum_i \lambda \sigma_i c_{j,i}} \le \frac{\log J}{\lambda} + \frac{\lambda n c^2}{2}.$$

Optimizing the parameter λ, the upper bound becomes $c\sqrt{2n \log J}$. ■

Then we have

$$\begin{aligned}
\sum_k e_k &\le 12\sqrt{n} \sum 2^{-(k+1)} \sqrt{\log N(\mathcal{F}, 2^{-k}, d_n)} \\
&\le 12\sqrt{n} \int_0^1 \sqrt{\log N(\mathcal{F}, r, d_n)} dr.
\end{aligned}$$

To conclude, the chaining trick showed that Rademacher averages are bounded by the Koltchinskii-Pollard integral.

B.3.9 PROOF OF INEQUALITY (9)

The proof is inspired from Talagrand (1996). Let $\left(\pi^{(j)}\right)_{j\in\mathbb{N}}$ be a family of probability distributions on $\mathcal{F}$. We want to prove that the Rademacher averages are bounded with $C\sup_{f\in\mathcal{F}}\sum_{j=1}^{\infty} r^{-j}\sqrt{\log\{1/\pi^{(j)}[A_j(f)]\}}$. For any $j\in\mathbb{N}$ and any $A\in\mathcal{A}_j$, we choose an arbitrary point $x_A\in A$. Let us define $p_j(f):=x_{A_j(f)}$ and $S_j:=\{p_j(f):f\in\mathcal{F}\}$. We have $\mathcal{A}_0=\{\mathcal{F}\}$. Let $f_0=x_{\mathcal{F}}$. Define $X_f:=\sum_i\sigma_i f(Z_i)$.

From Cauchy-Schwarz inequality, we have $\sup_{\sigma,Z_1,\ldots,Z_n,f}|X_f-X_{p_j(f)}|\leq\sqrt{n}r^{-j}$, hence $\sum_{j\geq1}[X_{p_j(f)}-X_{p_{j-1}(f)}]$ converges uniformly towards $X_f-X_{f_0}$. Introduce a probability distribution π' such that $\pi'(\{x_A\})\geq 2^{-j-1}\pi^{(j)}(A)$ for any $A\in\mathcal{A}_j$. Define the quantities $a_j(f):=r^{-j+1}\sqrt{2n\log[2/\pi'(f)]}$ and $M:=\sup_{f\in\mathcal{F}}\sum_{j\geq1}a_j[p_j(f)]$. We have

$$\begin{aligned}\mathbb{E}_\sigma\exp\left\{\lambda(X_f-X_g)\right\} &= \prod_{i=1}^n\mathbb{E}_\sigma\exp\left\{\sigma_i\lambda[f(Z_i)-g(Z_i)]\right\}\\ &= \prod_{i=1}^n\cosh\{\lambda[f(Z_i)-g(Z_i)]\}\\ &\leq \prod_{i=1}^n\exp\left\{\tfrac{\lambda^2[f(Z_i)-g(Z_i)]^2}{2}\right\}\\ &= e^{\frac{n\lambda^2d_n^2(f,g)}{2}},\end{aligned}$$

hence for any $u>0$, $\mathbb{P}_\sigma(X_f-X_g\geq u)\leq e^{-\frac{u^2}{2nd_n^2(f,g)}}$. Then for any $u\geq1$, we get

$$\begin{aligned}\mathbb{P}_\sigma\left(\sup_{f\in\mathcal{F}}\left\{X_f-X_{f_0}\right\}\geq uM\right) &\leq \sum_{j\geq1,v\in S_j}\mathbb{P}_\sigma\left[X_v-X_{p_{j-1}(v)}\geq ua_j(v)\right]\\ &\leq \sum_{j\geq1,v\in S_j}e^{-\frac{u^2a_j^2(v)}{2n\mathrm{Diam}^2A_{j-1}(v)}}\\ &\leq \sum_{j\geq1,v\in S_j}e^{-u^2\log\frac{2}{\pi'(v)}}\\ &\leq 2^{1-u^2},\end{aligned}$$

since $\pi'(v)^{u^2}\leq\pi'(v)$. We obtain

$$\mathbb{E}_\sigma\sup_{f\in\mathcal{F}}X_f=\mathbb{E}_\sigma\sup_{f\in\mathcal{F}}\left\{X_f-X_{f_0}\right\}\leq\int_0^{+\infty}2^{1-u^2}duM\leq 2.2M.$$

By plugging the definitions of X_f and M, we obtain

$$\mathbb{E}_\sigma\sup_{f\in\mathcal{F}}\sum_i\sigma_i f(Z_i)\leq 4\sqrt{n}\sup_{f\in\mathcal{F}}\sum_{j\geq1}r^{-j+1}\sqrt{\log\{2^{j+2}/\pi^{(j)}[A_j(f)]\}}.$$

From (19) and by using that $\sqrt{\log\{2^{j+2}/\pi^{(j)}[A_j(f)]\}}\leq\sqrt{j+2}+\sqrt{\log\{1/\pi^{(j)}[A_j(f)]\}}$, we get the desired result.

References

J.-Y. Audibert. Aggregated estimators and empirical complexity for least square regression. *Ann. Inst. Henri Poincaré, Probab. Stat.*, 40(6):685–736, 2004a.

J.-Y. Audibert. A better variance control for PAC-Bayesian classification. Technical report n.905, `http://www.proba.jussieu.fr/mathdoc/textes/PMA-905Bis.pdf`, Laboratoire de Probabilités et Modèles Aléatoires, Universités Paris 6 and Paris 7, 2004b.

P. L. Bartlett, O. Bousquet, and S. Mendelson. Local Rademacher complexities. *Ann. Stat.*, 33(4): 1497–1537, 2005.

P. L. Bartlett and S. Mendelson. Empirical minimization. *Probability Theory and Related Fields*, 135(3):311 – 334, 2006.

P. L. Bartlett, S. Mendelson, and P. Philips. Local complexities for empirical risk minimization. In J. Shawe-Taylor, editor, *17th Annual Conference on Learning Theory, COLT 2004*, LNCS-3120, berlin, 2004. Springer-Verlag.

S. Boucheron, O. Bousquet, and G. Lugosi. Theory of classification: a survey of some recent advances. *ESAIM: Probability and Statistics*, 9:323–375, 2005.

S. Boucheron, G. Lugosi, and S. Massart. A sharp concentration inequality with applications. *Random Structures and Algorithms*, 16:277–292, 2000.

O. Catoni. A PAC-Bayesian approach to adaptive classification. Technical report n.840, http://www.proba.jussieu.fr/mathdoc/preprints/, Laboratoire de Probabilités et Modèles Aléatoires, Universités Paris 6 and Paris 7, 2003.

A. Dembo and O. Zeitouni. *Large Deviation Techniques and Applications*. Springer, 1998.

L. Devroye and G. Lugosi. *Combinatorial Methods in Density Estimation*. Springer Series in Statistics. Springer Verlag, New York, 2001.

R. M. Dudley. A course on empirical processes. *Lecture Notes in Mathematics*, 1097:2–142, 1984.

E. Giné and J. Zinn. Some limit theorems for empirical processes. *Ann. Probab.*, 12(4):929–989, 1984.

W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58:13–30, 1963.

V. Koltchinskii. Local Rademacher complexities and oracle inequalities in risk minimization. *Annals of Statistics*, 34(6), 2006.

G. Lugosi. Concentration of measure inequalities. *Lecture notes*, pages 1–62, 2003. available from http://www.econ.upf.es/ lugosi/anu.ps.

P. Massart. Some applications of concentration inequalities to statistics. *Ann. Fac. Sci. Toulouse*, Math. 9(2):245–303, 2000.

P. Massart. *Concentration Inequalities and Model Selection: Ecole d'été de Probabilités de Saint-Flour XXXIII - 2003*. Lecture Notes in Mathematics. Springer, 2006.

D. A. McAllester. Some PAC-Bayesian theorems. In *Proceedings of the 11th Annual Conference on Computational Learning Theory*, pages 230–234. ACM Press, 1998.

D. A. McAllester. PAC-Bayesian model averaging. In *Proceedings of the 12th Annual Conference on Computational Learning Theory*. ACM Press, 1999.

D. Panchenko. Symmetrization approach to concentration inequalities for empirical processes. *Annals of Probability*, 31(4):2068–2081, 2003.

G. Pisier. Probabilistic methods in the geometry of banach spaces. In *Probability and analysis, Lect. Sess. C.I.M.E Varena, Italy 1985, Lecture Notes in Mathematics*, volume 1206, pages 167–241, Berlin, 1986. Springer.

M. Seeger. Bayesian gaussian process models: PAC-bayesian generalisation error bounds and sparse approximations. *PhD Thesis, University of Edinburgh*, December 2003.

M. Talagrand. Majorizing measures: The generic chaining. *Annals of Probability*, 24(3):1049–1103, 1996.

M. Talagrand. Majorizing measures without measures. *Annals of Probability*, 29(1):411–417, 2001.

M. Talagrand. *The Generic Chaining: Upper and Lower Bounds for Stochastic Processes*. Springer, 2005.

A. B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *Ann. Stat.*, 32(1):135–166, 2004.

S. van de Geer. *Empirical Processes in M-Estimation*. Cambridge University Press, Cambridge, UK, 2000.

A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998.

A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes*. Springer, 1996.

V. Vapnik and A. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2):264–280, 1971.

V. Vapnik and A. Chervonenkis. *Theory of Pattern Recognition [in Russian]*. Nauka, Moscow, 1974. (German Translation: W. Wapnik & A. Tscherwonenkis, *Theorie der Zeichenerkennung*, Akademie–Verlag, Berlin, 1979).