Ann. Inst. Statist. Math.
Vol. 55, No. 2, 371–389 (2003)

NEW APPROACHES TO STATISTICAL LEARNING THEORY

OLIVIER BOUSQUET

Max Planck Institute for Biological Cybernetics, Spemannstr. 38, D-72076 Tübingen, Germany,
e-mail: olivier.bousquet@tubingen.mpg.de

(Received May 10, 2002; revised September 3, 2002)

Abstract. We present new tools from probability theory that can be applied to the analysis of learning algorithms. These tools allow to derive new bounds on the generalization performance of learning algorithms and to propose alternative measures of the complexity of the learning task, which in turn can be used to derive new learning algorithms.

Key words and phrases: Statistical learning theory, concentration inequalities, Rademacher averages, error bounds.

1. Introduction

Recently, there has been a large increase of the interest for theoretical issues in the Machine Learning community. This is mainly due to the fact that Statistical Learning Theory has demonstrated its usefulness by providing the ground for developping successful and well-founded learning algorithms such as Support Vector Machines (SVM). This has shown that elegant and powerful mathematical objects such as reproducing kernels can be effectively used in practice. This renewed interest for theory naturally boosted the development of performance bounds. This, in turn, has raised concerns about the relevance of these bounds, in particular because of their observed looseness on real-world problems.

We would like to show here that these concerns do not impair the relevance of the initial motivation of statistical learning theory, neither should they restrain the development of new approaches in this theory. We hope to convince the reader that this theory, if used appropriately, can yield new advances in the understanding of learning algorithms as well as in the development of new techniques with practical applications.

The problem we shall focus on is the following. Given a set of data consisting of labelled objects, the goal is to find a function that assigns labels to objects such that, if new objects are given, this function will label them correctly. Of course, if we do not assume any relationship between the data at hand and future unseen data, there is no way to solve this problem. The classical assumption is that all the data (both observed and unobserved) is generated by the same process, which is formalized by saying that the data is sampled independently from a fixed probability distribution.

Now the issue is that this probability distribution is unkown and that the only knowledge we have about it comes from a sample of observations. Typically, one chooses a class of possible functions (hypotheses) that correspond to the possible ways in which the labels can be related to the objects, and this choice is made a priori (before seeing the data). Then one chooses a function in that class which agrees as much as possible with the given data.

The questions that a theory of learning should address are thus how to choose an appropriate class of functions and how to choose a function in that class. Since we assume that the data is sampled independently from a certain distribution, it is possible to relate the observed behavior of a function on the data to its expected behavior on future data, by means of probability theory. More precisely, for each given function in the class of interest, one can obtain confidence intervals for the expected misclassification error (expected number of labelling mistakes of this function when computed on objects from the distribution) in terms of the observed misclassification error (also called empirical error).

However, this is not enough to guarantee that, in a given class, a function which has a small empirical error will have a small expected error. Indeed, if the class if large (think for example of the class of all possible functions defined on the objects at hand), then it is likely that, on a given dataset, many functions will perfectly predict the labels of the objects (i.e. classify the objects) but these functions may have very different values on the other objects and thus may have diverse expected errors.

Considering our sample as a random variable, we see that the empirical error of each function in our class is a random variable, which means we have a collection of random variables, i.e. a *stochastic process* which may not be independent. In this context, if one wants to have a bound on the expected error of the learning algorithm, since the particular function that the algorithm will pick after seeing the data is not known in advance, one has to bound the error of all the functions in the class. In other words, one has to bound uniformly the deviations of the stochastic process, which in this case is called an *empirical process* since each element of this process is distributed as a sum of independent and identically distributed random variables.

According to the above view, the question of bounding the error of a learning algorithm boils down to bounding the maximum of an empirical process, that is of a collection of random variables. It is clear that the larger is this collection, the larger will the fluctuations of the maximum be. However, the notion of size here is not clear yet. Indeed, what matters is not necessarily the number of functions in the class but more how the errors of the functions are correlated. As an example, if all the errors are highly correlated, it is likely that their maximum will be smaller (on average) than if they are independent.

It should thus be understood that the notion of *size* of a collection of random variables (indexed by a class of functions) is crucial in statistical learning theory. Several such notions have been proposed in the past. For example, the notion of Vapnik-Chervonenkis (VC) dimension which applies to classes of boolean functions, measures how many points a class can *shatter*, i.e. can separate in all possible ways. Other quantities like covering numbers, which measure the number of balls of a given radius are need to cover the space of functions, have been introduced to capture, on a finer scale, the "size" of the function class. However, since we are interested in the behaviour of random variables indexed by the class and not the class itself, it is important that the notion of size we use takes this fact into account. In order to do so, it is possible to use specific metrics on the function space that are related to the covariance structure of the errors. For example, the empirical ℓ_2 metric (defined below) is directly related to the covariance of two functions. In this paper we will make use of a specific notion of size, called the *Rademacher average* (Bartlett *et al.* (2002*a*)), which is directly related to the behaviour of the maximum of the empirical process, thus capturing precisely the quantity we are interested in, without the need for introducing metrics on the function class.

Even with a good measure of the size of a functions class, the obtained bounds might be loose. Indeed, we said before that we want to bound the maximum of the empirical process, which means that we want to bound the worst deviation of the empirical error from the expected error for functions in our class. It is clear that a particular learning algorithm may very well pick a function whose corresponding error deviation is much smaller than the maximum over the class. More precisely, most learning algorithms will be inclined towards choosing functions which have a small empirical error and thus are likely to have a small expected error as well. For such functions, the error deviation may also be small since the error being a non-negative (bounded) quantity, its variance is related to its expectation, so that functions with small expected error will also have a small error deviation.

From this reasoning, we deduce that what really matters is not the size of the entire class of functions but rather the size of the subclass of functions with small error. In other words, we now want to bound the maximum of an empirical process which is indexed by a subclass of the initial class. It turns out that the notion of Rademacher average can be naturally modified to take this into account, yielding the so-called *local Rademacher average* (Koltchinskii and Panchenko (2000)).

Our claim is that this notion of size is capturing, in a natural and sharp way the complexity of the learning problem, in the sense that it really measures the magnitude of the error deviation of functions with a small errors, which are the ones that are likely to be picked by the learning algorithm. We say that we consider functions with "small" expected error, but we have to make this more precise. It turns out that one can actually use a circular definition: the expected error will be said to be small if it is of the order of the maximum error deviation in the class of functions with small error. This may seem meaningless but we will see how to formulate this mathematically as a fixed point equation. The solution of this equation which we call the *complexity radius* will be precisely, both the maximum expected error of the functions which we call "small", and the maximum error deviation of those functions.

What are the implications of all this? The simple answer is that we obtain new and sharper bounds. However, we would like to emphasize that the sharpness of the bounds should not be the main drive. Indeed, if one looks for example at bounds for maximum margin classifiers, the important aspect of such bounds is not so much that the margin or the square root of the margin or the squared of the margin is the correct term. It is rather the fact that the margin enters the bound at all. In other words, one should not be concerned about the quantitative value of the bound or even about its functional form but rather about the terms that appear in the bound. In that respect, a useful bound is one which allows to understand which quantities are involved in the learning process.

As a result, performance bounds should be used for what they are good for. They should not be used to actually predict the value of the expected error. Indeed, they usually contain prohibitive constants or extra terms that are mostly mathematical artifacts. They should not be used directly as a criterion to optimize since their precise functional form (e.g. whether we have the square or the square root of the margin) may also be a mathematical artifact. However, they should be used to modify the design of the learning algorithms or to build new algorithms.

In this paper we introduce new tools from probability theory known as *concentration inequalities* that allow to obtain error bounds in terms of local Rademacher averages.

It should be understood that the use of these new tools not only allows to obtain new proofs of earlier results (see Section 3) or refined error bounds (as in Section 4), but more importantly, provides new insights on the performance of learning algorithms.

The outline of the paper is as follows. In Section 1, we will introduce the mathematical framework and the notation used throughout the paper and describe in more details the ideas behind the proofs of performance bounds. Section 2 will present applications of concentration inequalities to quantities that are relevant in statistical learning theory and explain their relevance and implications. Then, in Section 3 we will discuss previous results and in particular present a simple application of the concentration inequalities we introduced. In Section 4 we will state and discuss our main result and comment about its consequences.

2. Preliminaries

2.1 *Notation and definitions*

As usual in the framework of statistical learning theory, we consider a space $\mathcal{X}$ of possible inputs (instance space) and a space $\mathcal{Y}$ of possible outputs (label set). The product space $\mathcal{X} \times \mathcal{Y}$ is assumed to be measurable and is endowed with an *unknown* probability measure denoted P.

Input-output pairs (X, Y) are sampled according to P. A sample of size n is created by sampling n independent and identically distributed (i.i.d.) pairs denoted $(X_1, Y_1), \ldots, (X_n, Y_n)$.

We will denote by $\mathbb{P}[A]$ the probability of the random event A, where the probability is taken over all the random variables occuring in the definition of A. Similarly $\mathbb{E}[B]$ will denote the expectation taken with respect to all the random variables occuring in the quantity B. When considering conditional expectations, we will use an index to specify over which random variables the integration runs. For example if X and σ are two random variables, $\mathbb{E}_\sigma[f(X,\sigma)]$ will denote the following function of X, $\mathbb{E}[f(X,\sigma) \mid X]$ (where σ is integrated out).

In order to measure the size of a class of functions we will use the notion of covering numbers. Given a sample $(X_1, Y_1), \ldots, (X_n, Y_n)$, we will denote by $N(\mathcal{F}, \varepsilon)$ the ε-covering number of $\mathcal{F}$, i.e. the minimal number of balls of radius ε with centers in $\mathcal{F}$ needed to cover $\mathcal{F}$ in the empirical pseudo-metric given by

$$d(f, f') := \left(\frac{1}{n} \sum_{i=1}^{n} (f(X_i, Y_i) - f'(X_i, Y_i))^2 \right)^{1/2} .$$

Notice that the sample is drawn according to the product distribution $P \otimes P \otimes \cdots \otimes P = P^{\otimes n}$. We introduce the *empirical measure* P_n of the sample which is the discrete random measure that puts mass $1/n$ at each observation, and can be written as the linear combination $P_n = n^{-1} \sum_{i=1}^{n} \delta_{X_i, Y_i}$ of the dirac measures at the observations.

If Q is a measure and f is a Q-measurable function, it is common to use the notation $Qf = \int f dQ$ (see e.g. van der Vaart and Wellner (1996)). We will thus denote $Pf = \mathbb{E}[f(X, Y)]$, and

$$P_n f = \frac{1}{n} \sum_{i=1}^{n} f(X_i, Y_i).$$

Notice that $P_n f$ is a random variable that depends on the sample $(X_i, Y_i)_{i=1,\ldots,n}$.

Similarly, we introduce the random variable

$$R_n f = \frac{1}{n}\sum_{i=1}^{n} \sigma_i f(X_i, Y_i),$$

where $\sigma_1, \ldots, \sigma_n$ are n Rademacher random variables, i.e. n i.i.d. random variables such that $\mathbb{P}[\sigma_i = 1] = \mathbb{P}[\sigma_i = -1] = 1/2$. Notice that $R_n f$ is a random variable that depends both on the sample and on the σ_i. The notation σ will denote the random vector whose coordinates are $\sigma_1, \ldots, \sigma_n$.

This will allow us to define the so-called *Rademacher average* of a class $\mathcal{F}$ of functions as

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} R_n f\right],$$

where the expectation is taken with respect to both the sample $((X_i, Y_i))$ and the Rademacher variables.

Moreover we will denote by $\mathbb{E}_\sigma[\ ]$ the expectation with respect to the Rademacher variables only (conditional to the sample).

We will also use the so-called *local Rademacher average* defined as

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}:Pf\leq r} R_n f\right],$$

which measures (in a sense to be precised later on) the "size" of the subclass with error smaller than r.

2.2 *Obtaining bounds in statistical learning theory*

2.2.1 *The goal of learning*

The goal of a learning algorithm is to pick a function g in a space $\mathcal{G}$ of functions from $\mathcal{X}$ to $\mathcal{Y}$ in such a way that this function should capture as much as possible the relationship (which may not be of a functional nature) between the random variables X and Y. In order to measure how well this is captured, a cost function c from $\mathcal{Y} \times \mathcal{Y}$ to $\mathbb{R}_+$ is defined that measures the cost of predicting a wrong label. The *risk* (or expected risk, or expected error) of a function $g \in \mathcal{G}$ is then the expected cost, defined as

$$L(g) = \mathbb{E}\left[c(g(X), Y)\right],$$

and the empirical risk (or empirical error) is defined as

$$L_n(g) = \frac{1}{n}\sum_{i=1}^{n} c(g(X_i), Y_i).$$

For convenience we will denote by ℓ the function from $\mathcal{G}$ to $\mathbb{R}^{\mathcal{X}\times\mathcal{Y}}$ which maps a function g to its associated *loss function* $\ell(g)$ defined as

$$\begin{aligned} \ell(g) : \mathcal{X} \times \mathcal{Y} &\to \mathbb{R}_+ \\ (x, y) &\mapsto c(g(x), y). \end{aligned}$$

We then define the *loss class* $\mathcal{F}$ associated to $\mathcal{G}$ as

$$\mathcal{F} = \{\ell(g) : g \in \mathcal{G}\}.$$

In the remainder we will only consider classes $\mathcal{G}$ and cost functions c that yield loss classes $\mathcal{F}$ of bounded non-negative functions. In other words, we will assume that $c(\cdot, \cdot) \in [0, b]$ for some $b > 0$.

2.2.2 *The empirical processes view*

We now explain how to formulate the problem of obtaining error bounds in terms of bounding the maximum of an empirical process.

An *error bound* (or performance bound, or generalization bound) is a probabilistic bound on the quantity $L(g_n) - L_n(g_n)$, where g_n is the function choosen by the learning algorithm. Since g_n is not known before we see the data, it is convenient to bound the quantity

$$\sup_{g \in \mathcal{G}} L(g) - L_n(g),$$

since this will directly yield an error bound. If we consider the class $\mathcal{F}$ defined above, this quantity can be written as

$$\sup_{f \in \mathcal{F}} \mathbb{E}\left[f(X,Y)\right] - \frac{1}{n} f(X_i, Y_i),$$

which is called the supremum of the empirical process indexed by the class of functions $\mathcal{F}$, empirical process meaning that we consider a collection of random variables $\mathbb{E}\left[f(X,Y)\right] - \frac{1}{n} f(X_i, Y_i)$, each one distributed as a sum of n independent and identically distributed random variables.

Using the notations introduced earlier, we can rewrite the supremum we want to bound in a shorter way as $\sup_{f \in \mathcal{F}} Pf - P_n f$. Note that this is a random variable depending on the sample. Our goal is thus to bound the probability that this quantity goes above a certain value.

2.2.3 *Measuring the size of a class of functions*

We have introduced above the notion of covering numbers. The metric we proposed is the so-called *empirical ℓ_2 metric* since it is a (pseudo) metric based on the data (and thus random).

This metric is not the one typically used in learning theory. Other possible metrics are the ℓ_∞ metric defined as $d_\infty(f, f') = \sup_{i=1,\dots,n} |f(X_i) - f'(X_i)|$ or the ℓ_1 metric defined as $d_1(f, f') = n^{-1} \sum_{i=1}^{n} |f(X_i) - f'(X_i)|$. Notice that the ℓ_1 distance is the smallest one and the ℓ_∞ the largest one. As a result the ℓ_1 covering numbers are smaller than the ℓ_2 ones which are smaller than the ℓ_∞ ones.

However, we will not use the covering numbers as our measure of size of the function space but rather to compare the notion we introduce to classical ones. Indeed, since the quantity we want to bound is $\sup_{f \in \mathcal{F}} Pf - P_n f$, we will see that the deviations of this quantity about its expectations do not depend on the size of $\mathcal{F}$ and as a consequence, we will claim that the "ideal" notion of size of the space $\mathcal{F}$ (with respect to the distribution at hand) is given precisely by $\mathbb{E}\left[\sup_{f \in \mathcal{F}} Pf - P_n f\right]$.

It turns out that there is a tight connection between such a quantity and the expectation of the Rademacher average via the well-known symmetrization inequality (see Appendix A)

$$\mathbb{E}\left[\sup_{f \in \mathcal{F}} Pf - P_n f\right] \leq 2\mathbb{E}\left[\sup_{f \in \mathcal{F}} R_n f\right].$$

Moreover, there exist corresponding lower bounds (see e.g. Bartlett and Mendelson (2002)). In view of these inequalities, the Rademacher average is a natural and sharp notion of size of a function class. We will see later that it can be upper bounded by a function of the covering numbers.

3. Concentration inequalities for empirical processes

Recently, new tools have appeared in probability theory. These tools called concentration inequalities allow to bound the deviation of a random function from its expectation, just knowing the sensitivity of this function to the change or removal of one of its arguments. We give a list of some of these inequalities in Appendix A and we give in this section applications of these inequalities to quantities that are of interest in statistical learning theory. Next section will comment on the consequences of such results in terms of measuring the size of function classes and obtaining data-dependent bounds.

3.1 *Results*

We now present applications of the above inequalities to quantities that are relevant in the analysis of learning algorithms.

The first result we give is a straightforward application of McDiarmid's inequality to a class of bounded functions.

THEOREM 3.1. *Let $\mathcal{F}$ be a countable set of functions from $\mathcal{X} \times \mathcal{Y}$ to $\mathbb{R}$ and assume that all functions f in $\mathcal{F}$ satisfy $\sup_{x,y} |Pf - f(x,y)| \leq b$. Then, defining Z as*

$$Z = \sup_{f \in \mathcal{F}} Pf - P_n f,$$

for all $\varepsilon \geq 0$, we have

$$\mathbb{P}\left[Z \geq \mathbb{E}[Z] + b\sqrt{\frac{2\varepsilon}{n}}\right] \leq e^{-\varepsilon}.$$

Next we give deviation bounds for quantities similar to the Rademacher average. In particular these bounds relate the Rademacher average computed on a particular sample to its expectation.

THEOREM 3.2. (Boucheron *et al.* (2002)) *Let $\mathcal{F}$ be a countable set of functions from $\mathcal{X} \times \mathcal{Y}$ to $\mathbb{R}$ and assume that all functions f in $\mathcal{F}$ satisfy $\sup_{x,y} |f(x,y)| \leq b$. Then defining Z as*

$$Z = \mathbb{E}_\sigma\left[\sup_{f \in \mathcal{F}} R_n f\right].$$

We have for all $\varepsilon > 0$,

$$\mathbb{P}\left[Z \geq \inf_{\alpha > 0}\left((1+\alpha)\mathbb{E}[Z] + \left(\frac{1}{3} + \frac{1}{2\alpha}\right)\frac{b\varepsilon}{n}\right)\right] \leq e^{-\varepsilon},$$

and

$$\mathbb{P}\left[\mathbb{E}[Z] \geq \inf_{\alpha \in (0,1)}\left(\frac{Z}{1-\alpha} + \frac{b\varepsilon}{2n(1-\alpha)\alpha}\right)\right] \leq e^{-\varepsilon}.$$

Finally we provide an inequality for the supremum of an empirical process indexed by a class of functions that are simply upper bounded. The fact that this result does not require the lower boundedness of the functions will be crucial in the derivation of our main result presented in a subsequent section.

THEOREM 3.3. (Bousquet (2002*a*)) *Let $\mathcal{F}$ be a countable set of functions from $\mathcal{X} \times \mathcal{Y}$ to $\mathbb{R}$ and assume that all functions f in $\mathcal{F}$ satisfy $\sup_{x,y} Pf - f(x,y) \leq b$. Then defining Z as*

$$Z = \sup_{f \in \mathcal{F}} Pf - P_n f,$$

and $v = \sup_{f \in \mathcal{F}} P(f^2)$, then for all $\varepsilon \geq 0$, we have

$$\mathbb{P}\left[Z \geq \inf_{\alpha>0}\left((1+\alpha)\mathbb{E}[Z] + \sqrt{\frac{2v}{n}} + \left(\frac{1}{3}+\frac{1}{\alpha}\right)\frac{b\varepsilon}{n}\right)\right] \leq e^{-\varepsilon}.$$

3.2 *Consequences*

We will now discuss the relevance of the above results for statistical learning theory.

Recall that the goal of a learning algorithm is to minimize the expected risk, that is to find a function $f \in \mathcal{F}$ which minimizes

$$Pf = \mathbb{E}[f(X,Y)].$$

However, since P is unknown to the algorithm, an estimate of this quantity has to be used instead, e.g.

$$P_n f = \frac{1}{n}\sum_{i=1}^{n} f(X_i, Y_i).$$

This is called the empirical risk of the function f. Typically, learning algorithms use $P_n f$ to select their output, it is thus crucial to be able to control the quality of the estimation of Pf by $P_n f$. A key question is to provide bounds on $Pf - P_n f$ *uniformly* over the class $\mathcal{F}$, since the particular f that will be chosen by the algorithm is not known in advance.

3.2.1 *Complexity of function classes*

Let's consider a class $\mathcal{F}$ of functions satisfying $0 \leq f \leq b$ for all $f \in \mathcal{F}$ and let $\varepsilon > 0$. By Theorem 3.3 we have with probability at least $1 - e^{-\varepsilon}$,

$$\forall f \in \mathcal{F}, Pf \leq P_n f + \inf_{\alpha>0}\left((1+\alpha)\mathbb{E}\left[\sup_{f\in\mathcal{F}} Pf - P_n f\right] + \sqrt{\frac{2v}{n}} + \left(\frac{1}{3}+\frac{1}{\alpha}\right)\frac{b\varepsilon}{n}\right),$$

with $v = \sup_{f \in \mathcal{F}} P(f^2) \leq b^2$. We are thus able to upper bound, uniformly over $\mathcal{F}$, the difference between Pf and $P_n f$ with a quantity which depends on the size of $\mathcal{F}$ only through the term

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

This is really important: it tells that the random variable $\sup_{f \in \mathcal{F}} Pf - P_n f$ is always sharply concentrated around its expectation, no matter how big the class $\mathcal{F}$ is. Also it tells that the *complexity* (for learning) of a function class can be measured by the quantity $\mathbb{E}[\sup_{f \in \mathcal{F}} Pf - P_n f]$ only.

In this sense, concentration inequalities allow to decompose the study of the random quantity $\sup_{f \in \mathcal{F}} Pf - P_n f$ into the study of its random fluctuations and its average size.

3.2.2 *Stability*

It can be argued that considering $\sup_{f \in \mathcal{F}} Pf - P_n f$ is too crude and does not really capture the fact that learning algorithms typically select functions that may be far from achieving this supremum. Indeed the quantity one should study is rather $Pf_n - P_n f_n$ where f_n is the function actually selected by the learning algorithm.

It is possible to prove (see Vapnik and Chervonenkis (1991)) that those two quantities are essentially the same when the algorithm minimizes the empirical risk (i.e. $f_n = \arg\min_{f \in \mathcal{F}} P_n f$).

However, most successful learning algorithms do not simply minimze the empirical risk, but rather use some form of regularization. Under certain conditions (studied for example in Bousquet and Elisseeff (2002)) it can be proven that $Pf_n - P_n f_n$ is a function of the training sample that satisfies the bounded increments condition of Theorem A.1.

One can thus apply concentration inequalities directly to the quantity $Pf_n - P_n f_n$. This approach is called the *stability approach.*

3.2.3 *Reweighting*

Another way to take into account the specificity of the learning algorithm to be studied is to consider a reweighted version of the empirical process. The idea is that since f_n, the function chosen by the algorithm, typically has small empirical error, it should also have small expected error and thus if we consider the quantity

$$\sup_{f \in \mathcal{F}} \frac{Pf - P_n f}{\sqrt{Pf}},$$

it is likely that the supremum will be reached at a function which has small expected error.

Studying the above quantity can thus be justified in two ways:

- When studying algorithms which minimize the empirical error, one is more interested in the deviation $Pf - P_n f$ for functions f that have small empirical error (and thus also small expected error).
- For non-negative bounded functions, Pf is an upper bound (up to a fixed constant) on the variance $\operatorname{var} f(X)$ and reweighting by the variance allows to 'uniformize' the deviation $Pf - P_n f$.

In the remainder we will focus on this quantity and try to provide bounds that involve Rademacher averages computed locally, i.e. for functions with small error.

4. Previous work

4.1 *Vapnik-Chervonenkis inequalities*

First we recall a classical result by Vapnik and Chervonenkis (Vapnik and Chervonenkis (1971)) for a class $\mathcal{F}$ of $\{0,1\}$-valued functions.

$$\mathbb{P}\left[\sup_{f \in \mathcal{F}} |Pf - P_n f| > t\right] \leq 4\mathbb{E}[N^2(\mathcal{F}, n^{-1/2})]e^{-nt^2/8}.$$

This immediately gives the following inequality relating the expected and empirical risk of all the functions in $\mathcal{F}$. For all $\varepsilon > 0$, with probability at least $1 - e^{-\varepsilon}$,

$$\forall f \in \mathcal{F}, Pf \leq P_n f + \sqrt{\frac{8(\varepsilon + \log 4\mathbb{E}[N^2(\mathcal{F}, n^{-1/2})])}{n}}. \tag{4.1}$$

Notice that for $\{0,1\}$-valued functions the quantity $N^2(\mathcal{F}, n^{-1/2})$ is an upper bound on the so-called shatter coefficient on a double sample, which is the quantity initially used by Vapnik and Chervonenkis.

Vapnik and Chervonenkis also obtained the following result for classes of $\{0,1\}$-valued functions

$$\mathbb{P}\left[\sup_{f\in\mathcal{F}} \frac{Pf - P_n f}{\sqrt{Pf}} \geq t\right] \leq 4\mathbb{E}[N^2(\mathcal{F}, n^{-1/2})]e^{-nt^2/4}, \tag{4.2}$$

where the constants above were actually achieved by Anthony and Shawe-Taylor (1993).

Combining this inequality with Lemma B.1 gives the following inequality relating the expected and empirical risk of all the functions in $\mathcal{F}$. For all $\varepsilon > 0$, with probability at least $1 - e^{-\varepsilon}$, for all $f \in \mathcal{F}$,

$$Pf \leq \inf_{\alpha>0}\left[(1+\alpha)P_n f + \frac{4}{n}\left(1 + \frac{1}{4\alpha}\right)(\log 4\mathbb{E}[N^2(\mathcal{F}, n^{-1/2})] + \varepsilon)\right]. \tag{4.3}$$

4.2 *A simple application of concentration*

Here we provide a simple result which is an adaptation of a result obtained in (Bartlett *et al.* (2002*a*)). We slightly modify the original result in order to obtain a Rademacher average without absolute values. This quantity has some advantages (see e.g. Bartlett *et al.* (2002*b*)).

The inequality below can be considered as an analogue of (4.1) where the complexity which was measured by empirical covering numbers is now measured by the empirical Rademacher average.

THEOREM 4.1. *Let $\mathcal{F}$ be a class of functions such that $0 \leq f \leq b$ for all $f \in \mathcal{F}$. Then for all $n \geq 2$, all $\varepsilon > 0$ with probability at least $1 - 2e^{-\varepsilon}$,*

$$\forall f \in \mathcal{F}, Pf \leq P_n f + 4\mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}} R_n f\right] + 5b\sqrt{\frac{\varepsilon}{n}}.$$

We shall briefly discuss about the above result. If we inspect the proof (see Appendix B) we can see that the result is quite sharp. Indeed, in the first step, we separate the random fluctuations of the variable of interest from its expectation. Its expectation is, as said before, what measures the complexity of the class $\mathcal{F}$. Next we symmetrize this expectation. The symmetrization inequality actually works both ways so that the loss in this step is simply a (universal) constant factor in front of the complexity term. Finally, we replace the expectation of the Rademacher average by its value on the sample and again the concentration is very sharp since it does not depend on the size of the function class. The bound is thus, up to small constants, very sharp.

An interesting application of such a bound is to classes $\mathcal{F}$ that are the convex hull of some class $\mathcal{H}$. In this case, it can be proven that the Rademacher average of $\mathcal{F}$ is equal to that of $\mathcal{H}$, which is a property that is not shared by other measures of complexity such as covering numbers. Indeed, the covering numbers of the convex hull of a small class can be quite large (see e.g. Mendelson (2001)).

It is possible to relate the Rademacher average to the empirical covering numbers using the following bound (see e.g. Ledoux and Talagrand (1991))

$$\mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}} R_n f\right] \leq \frac{4\sqrt{2}}{n}\int_0^D \sqrt{\log N(\mathcal{F},\varepsilon)}d\varepsilon, \tag{4.4}$$

where D is the diameter of the class $\mathcal{F}$, $D = \sup_{f,f'\in\mathcal{F}} d(f,f')$.

5. Main result

We have shown how to derive an analogue of (4.1) with concentration inequalities. Our main result is a generalization of (4.3) which uses a concentration approach.

5.1 *Statement*

First we present a generalization of (4.2).

THEOREM 5.1. *Let $\mathcal{F}$ be a class of functions such that $0 \leq f \leq b$ for all $f \in \mathcal{F}$. Let ϕ be a non-negative, non-decreasing function such that $\phi(r)/\sqrt{r}$ is non-increasing for $r > 0$ and such that*

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}:Pf\leq r} |Pf - P_n f|\right] \leq \phi(r),$$

and let r^* be the largest solution of $\phi(r) = r$.

Then, for all $\varepsilon > 0$, the following inequality holds with probability at least $1 - e^{-\varepsilon}$, for all $f \in \mathcal{F}$,

$$\frac{Pf - P_n f}{\sqrt{Pf}} \leq \inf_{\alpha>0}\left((1+\alpha)\sqrt{r^*}\left(1 + \frac{e}{2}\log\frac{eb}{r^*}\right) + \sqrt{\frac{2b\varepsilon}{n}} + \frac{(3+\alpha)\varepsilon\sqrt{b}}{3\alpha n}\right).$$

The following corollary gives an error bound that can be compared to (4.3).

COROLLARY 5.1. *Let $\mathcal{F}$ be a class of non-negative functions bounded by b almost surely. Let ϕ be a non-negative, non-decreasing function such that $\phi(r)/\sqrt{r}$ is non-increasing for $r > 0$ and such that*

$$2\mathbb{E}\left[\sup_{f\in\mathcal{F}:Pf\leq r} |R_n f|\right] \leq \phi(r),$$

and let r^* be the largest solution of $\phi(r) = r$.

There exist a universal constant K such that for all $\varepsilon > 0$, with probability at least $1 - e^{-\varepsilon}$, for all $f \in \mathcal{F}$,

$$Pf \leq \inf_{\alpha>0}\left[(1+\alpha)P_n f + \left(1 + \frac{1}{4\alpha}\right)\left(31r^*\log^2\frac{b}{r^*} + 50\frac{b\varepsilon}{n}\right)\right].$$

We thus have fulfilled our goal: we now have a bound which is similar to (4.3) but where the complexity term now depends on the local Rademacher average, that is the complexity of the subset of functions with small error.

5.2 *Discussion*

It is important to understand the role of the function ϕ and the real r^*. The conditions required on ϕ, which are here mainly for mathematical reasons, may seem restrictive but there are several easy ways to satisfy them. The first way is to use the entropy integral (4.4) in combination with an upper bound on the empirical covering numbers.

Let's give an example. For Vapnik-Chervonenkis classes of functions with VC-dimension d, it can be proved (see e.g. Koltchinskii and Panchenko (2000)) that

$$r^* \leq K\frac{d}{n}\log n.$$

This shows that the above bound is tight up to logarithmic factors. Indeed the optimal order for such a bound is $\frac{d}{n}$ and we would obtain $\frac{d}{n}\log^3 n$.

However, we can use a different approach. It is possible to slightly enlarge the class of functions in order to get a function ϕ with appropriate properties. Indeed, if we consider the star-hull of $\mathcal{F}$ around 0, defined as the following set

$$\star(\mathcal{F}) := \{\alpha f : f \in \mathcal{F}, \alpha \in [0,1]\},$$

and if we set

$$\phi(r) = \mathbb{E}\left[\sup_{f\in\star(\mathcal{F}):Pf\leq r} |R_n f|\right],$$

then ϕ is non-negative, non-decreasing and $\phi(r)/\sqrt{r}$ is non-increasing (see e.g. Bartlett *et al.* (2002*c*) for a proof). Moreover, it can be shown that the covering numbers of $\star(\mathcal{F})$ are not much larger than those of $\mathcal{F}$.

Now the quantity r^* which can be called the *complexity radius* of the class $\mathcal{F}$ is really what measures the complexity of a class of non-negative functions. Indeed, if the (global) Rademacher average has the correct order for general classes of functions, for non-negative bounded functions, the complexity is better measured by r^*. The main reason is the relationship that exists for such classes between the variance and the error, $P(f^2) \leq bPf$. The meaning of this inequality is that small error functions have small variance, so that the random fluctuations of the empirical error of the functions in $\mathcal{F}$ have a size controlled by the expected error itself.

In order to make our main result actually usable in practice (in learning algorithms) it is necessary to replace the expected local Rademacher average by its empirical counterpart. This is studied in Bousquet (2002*b*) or Bartlett *et al.* (2002*b*). Indeed, the bound of Theorem 4.1 can be computed from the data only while the bound in the above corollary depends on the distribution.

An equivalent result with empirical local Rademacher averages may have direct consequences on the design of learning algorithms. In particular, if one is able to bound r^* by a function of some parameters of the algorithm, it is possible to allow the algorithm to automatically tune these parameters.

5.3 *Sketch of the proof*

We try here to give a flavor of the ideas behind the proof of Theorem 5.1 in order to explain why the complexity radius arises. For details we refer the reader to Appendix B.

Examining Theorem 3.3, we see that an important role is played by the variance term v. It is thus clear that if one wants to sharpen the results, one has to use carefully this term. Indeed if one applies the theorem to the initial class the variance term will be bounded only by the maximum expectation in the class which is a constant term. However, keeping in mind the fact that the algorithm may pick functions with small error, we should try to apply Theorem 3.3 to a class of functions on which we can control the maximum variance.

The trick is thus to modify the class of functions on which we apply this theorem in order to make it sharper. This is done by simply reweighting the class of functions by dividing each of them by the square root of its expectation.

Since the functions in $\mathcal{F}$ are non-negative, so will be the functions $f/\sqrt{Pf}$. However they are no longer lower bounded (which is not an issue for applying Theorem 3.3).

Once we have obtained a deviation inequality for this reweighted function class, we have to bound the expectation of the supremum. The main idea is to decompose the function class in successive balls (in the Pf seminorm) and to apply, on each of them the bound given by the function ϕ. On each of these subclasses, there will be a weighting factor coming from the minimum value of Pf over that subclass. By balancing the size of the subclasses with the growth of their radius, one can obtain a bound that involves the complexity radius.

Indeed, the weight of each ball is the inverse of its radius r while its size is given by $\phi(r)$. Thus the right balanced will be achieved when the radius is of the order of the size.

The symmetrization inequality will then yield the corollary.

6. Conclusion

We have quoted several recent concentration inequalities and proved that they may have applications in the analysis of learning algorithms.

In particular we have derived a generalization of Vapnik and Chervonenkis' relative error inequality which uses a novel measure of the size of function classes, namely the local Rademacher average. For non-negative bounded function classes, we have proved that a sharp measure of the size is given by the so-called complexity radius.

Further research should focus on obtaining bounds that depend on an empirical counterpart of the complexity radius that can be computed from the data only and which can be related to various parameters of the learning algorithm of interest.

Appendix

A. Concentration inequalities

Concentration inequalities are tools from probability theory that allow to bound the deviation from its expectation of a random variable. Here we will only consider concentration inequalities defined on a product measure spaces. In other words, the random variables we are interested in are functions of n i.i.d. random variables defined on an arbitrary measurable space (here the product $\mathcal{X} \times \mathcal{Y}$). The inequalities we will present give deviation bounds for such functions simply from conditions on their increments. More precisely, if one is able to bound the variation of the function when one coordinate is modified or removed, then one can apply one of the following inequalities.

Let's denote by T the pair of random variables (X, Y) and similarly, let $T_i = (X_i, Y_i)$. The results in this section apply to any type of independent random variables T_i, i.e. not only pairs and not only identically distributed but we will apply them to the random variables introduced in previous section.

As mentioned earlier we will denote by $\mathbb{E}[\cdot]$ the expectation with respect to $T_1, \ldots, T_n$, and we introduce the notation $\mathbb{E}_n^k[\cdot]$ for the expectation with respect to T_k only (i.e. the expectation conditional to $T_1, \ldots, T_{k-1}, T_{k+1}, \ldots, T_n$).

We consider an arbitrary random variable Z which is a function of $T_1, \ldots, T_n$, i.e. $Z := f(T_1, \ldots, T_n)$. We want to bound the difference $|Z - \mathbb{E}[Z]|$. In order to do so, we introduce auxiliary functions (and corresponding random variables) as (for all $k = 1, \ldots, n$)

$$Z_k := f_k(T_1, \ldots, T_{k-1}, T_{k+1}, \ldots, T_n),$$

and

$$Z'_k := g_k(T_1, \ldots, T_n).$$

These functions will be used as comparison functions. For example, if one defines f_k as the function f computed on all but the k-th random variable, then the difference $Z - Z_k$ will measure the sensitivity of Z to the removal of the k-th of its arguments. The results we present below rely on various types of assumptions on the behavior of the differences $Z - Z_k$, also called the *increments* of the random function.

The first result we quote is the so called McDiarmid's inequality.

THEOREM A.1. (McDiarmid (1989)) *Let Z and $(Z_k)_{k=1,\ldots,n}$ be as defined above. Assume that for all $k = 1, \ldots, n$ the following inequality is satisfied*

$$|Z - Z_k| \leq b,$$

then for all $\varepsilon \geq 0$,

$$\mathbb{P}\left[Z \geq \mathbb{E}[Z] + \sqrt{\frac{nb^2\varepsilon}{2}}\right] \leq e^{-\varepsilon},$$

and

$$\mathbb{P}\left[Z \leq \mathbb{E}[Z] - \sqrt{\frac{nb^2\varepsilon}{2}}\right] \leq e^{-\varepsilon}.$$

Although very simple to use, this result has the disadvantage of providing a bound that depends on the dimension n of the sample. As the next inequality shows, it is possible to obtain dimension-free bounds when more restrictive conditions are satisfied by the increments $Z - Z_k$.

THEOREM A.2. (Boucheron *et al.* (2000)) *Let Z and $(Z_k)_{k=1,\ldots,n}$ be as defined above. Assume that for all $k = 1, \ldots, n$ the following inequality is satisfied*

$$0 \leq Z - Z_k \leq 1,$$

and

$$\sum_{k=1}^{n} Z - Z_k \leq Z,$$

then for all $\varepsilon \geq 0$,

$$\mathbb{P}\left[Z \geq \mathbb{E}\left[Z\right] + \sqrt{2\varepsilon\mathbb{E}\left[Z\right]} + \frac{\varepsilon}{3}\right] \leq e^{-\varepsilon},$$

and

$$\mathbb{P}[Z \leq \mathbb{E}\left[Z\right] - \sqrt{2\varepsilon\mathbb{E}\left[Z\right]}] \leq e^{-\varepsilon}.$$

The bound for deviation above the expectation can actually be generalized as shows the next result (inspired by the work of Rio (2001)). The non-negativity of the increment can be replaced by a weaker condition, at the expense of paying for the variance of the increments.

THEOREM A.3. (Bousquet (2002*a*)) *Let* Z, $(Z_k)_{k=1,\dots,n}$ *and* $(Z'_k)_{k=1,\dots,n}$ *be as defined above. Assume that there is a constant* $u > 0$ *such that for all* $k = 1,\dots,n$, *the inequalities below are satisfied*

$$Z'_k \leq Z - Z_k \leq 1, \quad \mathbb{E}^k_n\left[Z'_k\right] \geq 0 \quad \textit{and} \quad Z'_k \leq u,$$

and

$$v^2 \geq \frac{1}{n}\sum_{k=1}^{n}\mathbb{E}^k_n\left[(Z'_k)^2\right].$$

Then let $\gamma = (1+u)\mathbb{E}\left[Z\right] + nv^2$. *We obtain for all* $\varepsilon > 0$,

$$\mathbb{P}\left[Z \geq \mathbb{E}\left[Z\right] + \sqrt{2\gamma\varepsilon} + \frac{\varepsilon}{3}\right] \leq e^{-\varepsilon}.$$

Although the difference between Theorems A.2 and A.3 does not seem significant at first glance (it may look like if one could obtain the latter by combining the former with Jensen's inequality), the latter allows to deal with more general random functions. In particular, inspecting the hypotheses of Theorem A.2, one notices that they imply that $Z \geq 0$ so that this theorem will only apply to non-negative functions. Moreover, the requirement of non-negative increments prevents to apply this theorem to even simple functions like sums of independent upper bounded random variables.

The importance of Theorem A.3 is demonstrated in Subsection 3.1 where it is shown to provide a sharp bound on the deviation of the supremum of an empirical process, which is precisely the object of interest here.

B. Proofs

PROOF OF THEOREM 4.1. The proof consists of three steps.

1. First concentration

We apply Theorem 3.1 to $Z = \sup_{f\in\mathcal{F}} Pf - P_nf$. With probability at least $1 - e^{-\varepsilon}$, we have

$$\sup_{f\in\mathcal{F}} Pf - P_nf \leq \mathbb{E}\left[\sup_{f\in\mathcal{F}} Pf - P_nf\right] + b\sqrt{\frac{2\varepsilon}{n}}.$$

2. Symmetrization

We introduce an i.i.d. sample $((X'_i, Y'_i))_{i=1,\dots,n}$ independent from the sample $((X_i, Y_i))_{i=1,\dots,n}$ and denote by P'_n the corresponding empirical measure. We have the following

$$\begin{aligned}
&\mathbb{E}\left[\sup_{f\in\mathcal{F}} Pf - P_n f\right] \\
&\quad = \mathbb{E}\left[\sup_{f\in\mathcal{F}} \mathbb{E}\left[P'_n f\right] - P_n f\right] \\
&\quad \leq \mathbb{E}\left[\sup_{f\in\mathcal{F}} P'_n f - P_n f\right] \qquad \text{(by Jensen's inequality)} \\
&\quad = \mathbb{E}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i(f(X'_i, Y'_i) - f(X_i, Y_i))\right] \qquad \text{(introducing random signs)} \\
&\quad \leq 2\mathbb{E}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i f(X_i, Y_i)\right] \qquad (\text{since } \sup a + b \leq \sup a + \sup b) \\
&\quad = 2\mathbb{E}\left[\sup_{f\in\mathcal{F}} R_n f\right].
\end{aligned}$$

3. Second concentration
We apply Theorem 3.2 and obtain with probability at least $1 - e^{-\varepsilon}$,

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} R_n f\right] \leq 2\mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}} R_n f\right] + \frac{2b\varepsilon}{n}.$$

Combining the above inequalities yields the result. □

We now prove our main result (Theorem 5.1). We first give a simple lemma which is used to relate the deviation to the reweighted deviation.

LEMMA B.1. *Let $\mathcal{F}$ a class of non-negative functions and let*

$$V = \sup_{f\in\mathcal{F}} \frac{Pf - P_n f}{\sqrt{Pf}}.$$

We have

$$\forall f \in \mathcal{F}, Pf \leq P_n f + V\sqrt{P_n f} + V^2,$$

and also,

$$\forall f \in \mathcal{F}, Pf \leq \inf_{\alpha>0}\left[(1+\alpha)P_n f + \left(1 + \frac{1}{4\alpha}\right)V^2\right].$$

PROOF. Notice that

$$Pf \leq P_n f + \sqrt{Pf}\left(\sup_{f'\in\mathcal{F}} \frac{Pf - P_n f}{\sqrt{Pf}}\right).$$

Moreover, it is easy to check that

$$x \leq A\sqrt{x} + B,$$

implies

$$x \leq B + A\sqrt{B} + A^2.$$

Finally, we use the simple fact that $\sqrt{ab} = \inf_{\alpha>0}(\alpha a + b/4\alpha)$. □

Now we will prove Theorem 5.1 in several steps. The first step consists as before in applying concentration.

LEMMA B.2. *Let $\mathcal{F}$ a class of functions such that $0 \leq f \leq b$ for all $f \in \mathcal{F}$ and let*

$$V = \sup_{f\in\mathcal{F}} \frac{Pf - P_n f}{\sqrt{Pf}}.$$

Let ε be a non-negative real number. We have with probability at least $1 - e^{-\varepsilon}$,

$$V \leq \inf_{\alpha>0}\left((1+\alpha)\mathbb{E}[V] + \sqrt{\frac{2\varepsilon b}{n}} + \left(\frac{1}{3} + \frac{1}{\alpha}\right)\frac{\varepsilon\sqrt{b}}{n}\right).$$

PROOF. We apply Theorem 3.3 to the class of functions $\mathcal{H} := \{h = \frac{f}{\sqrt{f}} : f \in \mathcal{F}\}$. We have for all $h \in \mathcal{H}$,

$$Ph - h(\cdot) = \frac{Pf - f(\cdot)}{\sqrt{Pf}} \leq \sqrt{Pf} \leq \sqrt{b}.$$

Also

$$P(h^2) = P\left(\frac{f^2}{Pf}\right) = \frac{P(f^2)}{Pf} \leq b. \qquad \square$$

Next we use an idea from Massart (2000) to bound the expectation.

LEMMA B.3. *Let $\mathcal{F}$ be a class of functions such that $0 \leq f \leq b$ for all $f \in \mathcal{F}$. Let ϕ be a non-negative, non-decreasing function such that $\phi(r)/\sqrt{r}$ is non-increasing for $r > 0$ and such that*

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}:Pf\leq r} |Pf - P_n f|\right] \leq \phi(r).$$

Then for all $r > 0$ we have

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} \frac{Pf - P_n f}{\sqrt{Pf}}\right] \leq \sqrt{r} + \frac{e\phi(r)}{2\sqrt{r}}\left(1 + \log\frac{b}{r}\right).$$

PROOF. We choose some $x > 1$ and we define $\mathcal{F}(a,b) = \{f \in \mathcal{F} : a \leq Pf \leq b\}$. Notice that $\sup_{f\in\mathcal{F}} Pf - P_n f \leq b$ and let $N = \lfloor \log(b/r)/\log x \rfloor$. We have

$$\begin{aligned}
\sup_{f\in\mathcal{F}} \frac{Pf - P_n f}{\sqrt{Pf}} &\leq \sup_{f\in\mathcal{F}(0,r)} \frac{Pf - P_n f}{\sqrt{Pf}} + \sum_{k=0}^{N} \sup_{f\in\mathcal{F}(rx^k, rx^{k+1})} \frac{|Pf - P_n f|}{\sqrt{Pf}} \\
&\leq \sup_{f\in\mathcal{F}(0,r)} \frac{Pf - P_n f}{\sqrt{Pf}} + \sum_{k=0}^{N} \sup_{f\in\mathcal{F}(rx^k, rx^{k+1})} \frac{|Pf - P_n f|}{r^{1/2} x^{k/2}} \\
&\leq \sup_{f\in\mathcal{F}(0,r)} \frac{Pf - P_n f}{\sqrt{Pf}} + \frac{1}{\sqrt{r}} \sum_{k=0}^{N} \sup_{f\in\mathcal{F}(0, rx^{k+1})} \frac{|Pf - P_n f|}{x^{k/2}}.
\end{aligned}$$

For $Pf \leq r$ we have

$$\frac{Pf - P_n f}{\sqrt{Pf}} \leq \sqrt{Pf} \leq \sqrt{r}.$$

Taking the expectation we obtain

$$\begin{aligned}
\mathbb{E}\left[\sup_{f\in\mathcal{F}} \frac{|Pf - P_n f|}{\sqrt{Pf}}\right] &\leq \sqrt{r} + \frac{1}{\sqrt{r}} \sum_{k=0}^{N} \frac{\phi(rx^{k+1})}{x^{k/2}} \\
&\leq \sqrt{r} + \frac{\phi(r)}{\sqrt{r}} (N+1) x^{1/2} \\
&\leq \sqrt{r} + \frac{\phi(r)}{\sqrt{r}} (1 + \log(b/r)/\log x) x^{1/2},
\end{aligned}$$

and we get the result by taking $x = e^2$. □

Now, if we choose in previous lemma $r = r^*$, the largest solution of the equation $\phi(r) = r$, we obtain

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} \frac{Pf - P_n f}{\sqrt{Pf}}\right] \leq \sqrt{r^*}\left(1 + \frac{e}{2}\left(1 + \log\frac{b}{r^*}\right)\right).$$

This bound combined with the concentration result above gives Theorem 5.1.

Now if we take $\alpha = .4$ in Theorem 5.1 we obtain the upper bound

$$\sqrt{2r^*}e\left(1 + \log\frac{b}{r^*}\right) + \sqrt{b}(\sqrt{2\varepsilon/n} + 3\varepsilon/n),$$

which, for $\varepsilon \leq n$ is upper bounded by

$$\sqrt{2r^*}e\left(1 + \log\frac{b}{r^*}\right) + 5\sqrt{\frac{b\varepsilon}{n}}.$$

Using Lemma B.1 and the fact that $(a+b)^2 \leq 2a^2 + 2b^2$ we obtain the result stated in the corollary.

References

Anthony, M. and Shawe-Taylor, J. (1993). A result of Vapnik with applications, *Discrete Appl. Math.*, **47**, 207–217.

Bartlett, P. and Mendelson, S. (2002). Rademacher and gaussian complexities: Risk bounds and structural results, *Journal of Machine Learning Research*, **3**, 463–482.

Bartlett, P., Boucheron, S. and Lugosi, G. (2002*a*). Model selection and error estimation, *Machine Learning*, **48**, 85–113.

Bartlett, P., Bousquet, O. and Mendelson, S. (2002*b*). Local rademacher complexities (preprint).

Bartlett, P., Bousquet, O. and Mendelson, S. (2002*c*). Localized rademacher complexity, *Proceedings of the 15th Annual Conference on Computational Learning Theory*, Lecture Notes in Comput. Sci., 44–58, Springer, Berlin.

Boucheron, S., Lugosi, G. and Massart, P. (2000). A sharp concentration inequality with applications, *Random Structures Algorithms*, **16**(3), 277–292.

Boucheron, S., Lugosi, G. and Massart, P. (2002). Concentration inequalities using the entropy method, *Ann. Probab.* (to appear).

Bousquet, O. (2002*a*). A Bennett concentration inequality and its application to suprema of empirical processes, *Computes Rendus Mathématique Academie des Sciences. Paris*, **334**, 495–500.

Bousquet, O. (2002*b*). Concentration inequalities and empirical processes theory applied to the analysis of learning algorithms, Ph.D. thesis, Centre de Mathématiques Appliquées, Ecole Polytechnique (preprint).

Bousquet, O. and Elisseeff, A. (2002). Stability and generalization, *Journal of Machine Learning Research*, **2**, 499–526.

Koltchinskii, V. and Panchenko, D. (2000). Rademacher processes and bounding the risk of function learning, *High Dimensional Probability II* (eds. E. Gine, D. Mason and J. Wellner), 443–459.

Ledoux, M. and Talagrand, M. (1991). *Probability in Banach Spaces*, Springer, Berlin.

Massart, P. (2000). Some applications of concentration inequalities to statistics, *Ann. Fac. Sci. Toulouse Math. (6)*, **9**(2), 245–303.

McDiarmid, C. (1989). On the method of bounded differences, *Surveys in Combinatorics*, London Math. Soc. Lecture Note Ser., **141**, 148–188, Cambridge University Press, Cambridge.

Mendelson, S. (2001). On the size of convex hulls of small sets, *Journal of Machine Learning Research*, **2**, 1–18.

van der Vaart, A. and Wellner, J. (1996). *Weak Convergence and Empirical Processes with Applications to Statistics*, Wiley, New York.

Vapnik, V. and Chervonenkis, A. (1971). On the uniform convergence of relative frequencies of events to their probabilities, *Theory Probab. Appl.*, **16**, 264–280.

Vapnik, V. and Chervonenkis, A. (1991). The necessary and sufficient conditions for consistency of the method of empirical risk minimization, *Pattern Recognition and Image Analysis*, **1**(3), 284–305.