Localized Rademacher Complexities

Peter L. Bartlett[1,3], Olivier Bousquet[1,2], and Shahar Mendelson[3]

[1] BIOwulf Technologies
Berkeley, CA, USA
Peter.Bartlett@anu.edu.au
[2] Centre de Mathématiques Appliquées
Ecole Polytechnique
91128 Palaiseau, France
bousquet@cmapx.polytechnique.fr
[3] Research School of Information Sciences and Engineering
The Australian National University
Canberra, ACT 0200, Australia
shahar.mendelson@anu.edu.au

Abstract. We investigate the behaviour of global and local Rademacher averages. We present new error bounds which are based on the local averages and indicate how data-dependent local averages can be estimated without *a priori* knowledge of the class at hand.

1 Introduction

In this article we investigate the role that Rademacher averages have in formulating error bounds. Given a class $\mathcal{F}$ of functions on a probability space $(\mathcal{X}, P)$, the (global) Rademacher averages associated with the class and with P are

$$\mathbb{E}\left[\frac{1}{n}\sup_{f\in F}\left|\sum_{i=1}^{n}\sigma_i f(X_i)\right|\right],$$

where $(X_i)_{i=1}^n$ are independent random variables distributed according to P and $(\sigma_i)_{i=1}^n$ are independent Rademacher (that is, symmetric $\{-1,1\}$-valued) random variables. The expectation is taken with respect to both (X_i) and (σ_i). Recent results have shown [5, 1, 6, 9, 2] that the Rademacher averages can be used to measure the sample complexity of a learning problem or as a complexity term in error bounds.

At the heart of our discussion is a concentration inequality which estimates the deviation of $\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n f(X_i)\right|$ (or $\mathbb{E}_\sigma\left[\frac{1}{n}\left|\sum_{i=1}^n \sigma_i f(X_i)\right|\right]$) from its mean value. The result we use is a version of Talagrand's concentration inequality for empirical processes [4]. The benefit of this result is that it enables one to control the deviation in terms of the Rademacher averages and the largest variance of a class member.

As an application of this result we obtain error bounds for loss classes using the Rademacher averages of the entire class (see Section 5). Moreover, we show

that the important quantity is not the Rademacher averages associated with the entire class, but rather, the local Rademacher averages, which are defined for every $r > 0$ as

$$\mathbb{E}\left[\frac{1}{n}\sup_{f\in\mathcal{F},\ Pf^2\le r}\left|\sum_{i=1}^{n}\sigma_i f(X_i)\right|\right],$$

where Pf^2 denotes the expectation of f^2 with respect to P.

Thus, instead of computing the supremum over the entire class, the supremum is taken with respect to the intersection of the class and a ball in $L_2(P)$ of radius $\sqrt{r}$, which leads to significantly better bounds.

Our results are based on two structural assumptions on the class. The first one is that the class consists of uniformly bounded functions. The other assumption is that for every member of the class, it is possible to control its variance using its expectation. In other words, we assume that there is some constant B such that for every $f \in \mathcal{F}$, $Pf^2 \le BPf$. The classes we are interested in are loss classes naturally appearing in learning theory. Although this structural assumption seems restrictive, in the learning setup it is satisfied by many loss classes. For example, in proper learning, where the target is assumed to be a member of the class, each loss function is nonnegative and uniformly bounded. Thus, $Pf^2 \le BPf$ for every loss function. The case of improper learning is less trivial. It is possible to show that if the original class is convex and $\mathcal{F}$ is the class of excess squared loss (so that each $f \in \mathcal{F}$ is the difference between expected squared loss and minimal expected squared loss), then for every $f \in \mathcal{F}$, $Pf^2 \le 16Pf$ [7]. Related results are known for other loss classes, such as those defined using p-norms with $p > 2$ [9], and for certain non-convex classes [10].

This article is organized as follows; first, we present some definitions and notation. Then, we present the basic properties of the Rademacher (both local and global) complexities. In Section 4 we recall the basic concentration result and prove a deviation bound based on the Rademacher complexities of the class. Then, in Section 5 we present error bounds using the global averages. The error bounds based on the local Rademacher complexities are presented in Section 6. Finally, we establish several results which can enable one to compute the Rademacher complexities using empirical data.

2 Definitions and Notation

Let $\mathcal{X}$ be the input space and $\mathcal{Y} \subset \mathbb{R}$. Fix a probability measure P on the product space $\mathcal{X} \times \mathcal{Y}$ and consider n independent random pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$ distributed according to P. Denote by $\mathcal{G}$ a class of functions mapping $\mathcal{X}$ into $\mathcal{Y}$, and by D_n the data $(X_i, Y_i)_{i=1}^n$. We use P to denote both the probability distribution on $\mathcal{X} \times \mathcal{Y}$ and the marginal distribution on $\mathcal{X}$. The distinction should always be clear from the context.

Let ℓ be a loss function and define the loss class

$$\mathcal{F} = \{f(x, y) = \ell(g(x), y) : g \in \mathcal{G}\}. \tag{1}$$

For every $f \in \mathcal{F}$, denote by Pf the expectation of f, $\mathbb{E}[f(X,Y)]$, and in our case, this would be called the expected loss. Given a sample D_n, set $P_n f = n^{-1}\sum_{i=1}^n f(X_i,Y_i)$ to be the empirical loss.

For every $f \in \mathcal{F}$, let $R_n f = n^{-1}\sum_{i=1}^n \sigma_i f(X_i,Y_i)$. Hence, $R_n f$ is a function of both the Rademacher random variables and the sample D_n, and the expectation of its magnitude is the Rademacher average of $\mathcal{F}$. We use $\mathbb{E}_\sigma[R_n f]$ to denote $\mathbb{E}[R_n f|D_n]$.

A class $\mathcal{F}$ is called *star-shaped* around f_0 if for every $f \in \mathcal{F}$ and every $0 \le t \le 1$, $tf + (1-t)f_0 \in \mathcal{F}$. We denote by $\mathrm{star}(\mathcal{F}, f_0)$ the set of all functions $tf + (1-t)f_0$, with $0 \le t \le 1$ and $f \in \mathcal{F}$. We call this set the *star hull* of $\mathcal{F}$ around f_0.

Finally, recall that $a \vee b = \max(a,b)$ and $a \wedge b = \min(a,b)$.

3 Measuring the Complexity Locally

We are interested in the local Rademacher average, defined for some $r > 0$ as

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}:Pf^2\le r} |R_n f|\right].$$

This is a function of r which measures the so-called Rademacher complexity of subspaces of the function class defined by a bound on the variance. In order to see why this is important for bounding the generalization error, suppose that the function chosen by a learning algorithm has small expected error. For classes where there is a relationship between the variance and the expectation, this function also has small variance. In this case, the algorithm chooses elements with small variance and the generalization error bound should thus take into account the complexity of the subclass of elements with small variance only.

However, since the local Rademacher average is a function of r, we have to specify some r_0 at which we compute it. We shall see in Section 6 that the complexity term (the only term in the bound that depends on the 'size' of the class of functions) is proportional to the value r_0 of r which satisifies the following fixed point equation,

$$\psi(r) = r\,,$$

where $\psi(r)$ is an upper bound on the local Rademacher average at radius r.

Of course, in order for this fixed point equation to have a (unique) solution, ψ should satisfy certain regularity conditions (which may not be satisfied by the local Rademacher average itself).

The relevance of the quantity r_0 defined above has been pointed out by Koltchinskii and Panchenko [6] and Massart [8]. Typically one upper bounds the local Rademacher average by an entropy integral [6]

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}:Pf^2\le r} |R_n f|\right] \le \mathbb{E}\left[\frac{K}{\sqrt{n}}\int_0^{\sqrt{r}} \sqrt{\log N(\mathcal{F},u,d_n)}\,du\right],$$

where $N(\mathcal{F}, u, d_n)$ is the covering number of $\mathcal{F}$ at radius u for the empirical L_2 metric. Another possibility is to upper bound it using VC-dimension or VC-entropies as in [8].

Both of these approaches are rather indirect, and introduce some looseness. Here we show that it is possible, by slightly enlarging the class on which the local Rademacher average is computed, to ensure that this average satisfies the required conditions for the existence of a (unique) fixed point r_0.

Lemma 1. *Let $\mathcal{F}$ be a class of functions. We have*

$$\mathbb{E}\left[\sup_{f \in star(\mathcal{F},0)} |R_n f|\right] = \mathbb{E}\left[\sup_{f \in \mathcal{F}} |R_n f|\right],$$

and defining for all $r > 0$,

$$F(r) = \mathbb{E}\left[\sup_{f \in star(\mathcal{F},0): Pf^2 \le r} |R_n f|\right],$$

then $F(r)$ is an upper bound on the local Rademacher average of $\mathcal{F}$ at radius r and $F(r)/\sqrt{r}$ is non-increasing.

Proof. The first claim is true since the averages of a class and of its symmetric convex hull coincide. Therefore, they both equal the averages of $\mathrm{star}(\mathcal{F}, 0)$ because the latter contains $\mathcal{F}$ and is contained in the symmetric convex hull of $\mathcal{F}$.

For the second claim, we will show that for any $r_1 \le r_2$, $F(r_1) \ge \sqrt{r_1/r_2} \cdot F(r_2)$. Indeed, fix any sample and any realization of the Rademacher random variables, and set f to be a function for which $\sup_{f \in F,\ Pf^2 \le r_2} |\sum_{i=1}^n \sigma_i f(x_i)|$ is attained (if the supremum is not attained only a slight modification is required). Since $Pf^2 \le r_2$, we have $P(\sqrt{r_1/r_2} \cdot f)^2 \le r_1$. Furthermore, the function $\sqrt{r_1/r_2} f$ is in $\mathcal{F}$ because $\mathcal{F}$ is star-shaped around zero, and satisfies $P(\sqrt{r_1/r_2} f)^2 \le r_1$. Hence,

$$\sup_{f \in F: Pf^2 \le r_1} \Big|\sum_{i=1}^n \sigma_i f(x_i)\Big| \ge \Big|\sum_{i=1}^n \sigma_i \sqrt{\frac{r_1}{r_2}} \cdot f(x_i)\Big| = \sqrt{\frac{r_1}{r_2}} \sup_{f \in F: Pf^2 \le r_2} \Big|\sum_{i=1}^n \sigma_i f(x_i)\Big|,$$

and the proof follows by taking expectations with respect to the Rademacher random variables. □

We observe that the global Rademacher average is not affected by taking the star hull of a class, while the local Rademacher average *is* affected.

Also, we notice that taking the star hull makes the complexity uniform over the balls $\{f : Pf^2 \le r\}$, in a sense. In other words, computing a local average in a star-shaped class corresponds to taking into account the complexity of the whole class, scaled appropriately. One could think that we thus lose the interest of looking locally, but this is not the case since we gain the fact that we have a scaling parameter (the radius of the ball) to adjust.

To see why this is interesting, consider again the entropy bound. Assume that we have a Vapnik-Chervonenkis class of functions of VC dimension V. Then it is well known [12] that one has the bound $N(\mathcal{F}, \epsilon, d_n) \leq K\epsilon^{-V}$. Moreover, taking the star hull of the class does not affect significantly the covering numbers (see e.g. [9]) and we obtain

$$N(\mathrm{star}(\mathcal{F}, 0), \epsilon, d_n) \leq K\epsilon^{-V-1} .$$

As a result, we obtain by the entropy bound

$$\mathbb{E}\left[\sup_{f \in \mathrm{star}(\mathcal{F},0): Pf^2 \leq r} |R_n f|\right] \leq K\sqrt{\frac{Vr}{n}\log\frac{n}{r}} ,$$

so that the complexity r_0 of the class (or its star-hull) will be of order

$$r_0 = O\left(\frac{V}{n}\log\frac{n}{V}\right) ,$$

which is optimal for such classes. It is possible to check that, for other rates of growth of the covering numbers, the value of r_0 obtained by this computation gives the optimal rate of convergence of the generalization error (see [6]).

Thus, in considering upper bounds F on the local Rademacher average of the class, we can say that taking the local Rademacher average of the star-hull gives the smallest (at least, in order of magnitude) upper bound that satisfies the condition that $F(r)/\sqrt{r}$ is non-increasing.

4 Concentration Results

This section is devoted to the main concentration results we require. The following is an improvement of Rio's [11] version of Talagrand's concentration inequality and is due to Bousquet [4].

Theorem 1. *Assume the X_i are identically distributed according to P. Let $\mathcal{F}$ be a countable set of functions from $\mathcal{X}$ to $\mathbb{R}$ and assume that all the functions in $\mathcal{F}$ are P-measurable, square-integrable and satisfy $\mathbb{E}[f] = 0$. If $\sup_{f\in\mathcal{F}} ess\ sup f \leq 1$, denote $Z = \sup_{f\in\mathcal{F}} \sum_{i=1}^n f(X_i)$ and if $\sup_{f\in\mathcal{F}} \|f\|_\infty \leq 1$ denote $Z = \sup_{f\in\mathcal{F}} |\sum_{i=1}^n f(X_i)|$.*

Let σ be a positive real number such that $\sigma^2 \geq \sup_{f\in\mathcal{F}} \mathsf{Var}[f(X_1)]$. Then, for any $x \geq 0$, we have

$$\mathbb{P}\left[Z \geq \mathbb{E}[Z] + x\right] \leq \exp\left(-vh\left(\frac{x}{v}\right)\right) ,$$

where $h(x) = (1+x)\log(1+x) - x$ and $v = n\sigma^2 + 2\mathbb{E}[Z]$. Also,

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

In a similar way one can obtain a concentration result for the Rademacher averages of a class (see e.g. [3]).

Theorem 2. *Assume* $|f(x)| \leq 1$. *Let*

$$Z := \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} \sum_{i=1}^n \sigma_i f(X_i) \right] \quad or \quad Z := \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} \left| \sum_{i=1}^n \sigma_i f(X_i) \right| \right] ,$$

then for all $x \geq 0$,

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

and

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

A standard fact we shall use is that the expected deviation of the empirical means from the actual ones can be controlled by the Rademacher averages of the class.

Lemma 2. *[12] For any class of functions* $\mathcal{F}$,

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

Using the concentration results and the symmetrization lemma one can estimate the probability of the deviation of the empirical means from the actual ones in terms of the Rademacher averages. Note that the bound in the following result improves as the largest variance of a class member decreases.

Corollary 1. *Let* $\mathcal{F}$ *be a class of functions which maps* $\mathcal{X} \times \mathcal{Y}$ *into* $[a, a+1]$ *for some* $a \in \mathbb{R}$. *Assume that there is some* $r > 0$ *such that for every* $f \in \mathcal{F}$, $\mathsf{Var}[f(X_i)] \leq r$ *and set*

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

Then, for every $x > 0$ *and every* $\alpha > 0$ *there is a set of probability larger than* $1 - e^{-x}$ *on which*

$$V \leq (1+\alpha)\mathbb{E}[V] + \sqrt{\frac{2rx}{n}} + \left(\frac{1}{3} + \frac{1}{\alpha} \right) \frac{x}{n}.$$

Moreover, for all $0 < \alpha < 1$, *with probability at least* $1 - 2e^{-x}$,

$$V \leq 2\frac{1+\alpha}{1-\alpha}\mathbb{E}_\sigma \left[\sup_{f \in F} |R_n f| \right] + \sqrt{\frac{2rx}{n}} + \left(\frac{1}{3} + \frac{1}{\alpha} + \frac{1}{2\alpha(1-\alpha)} \right) \frac{x}{n}.$$

The proof requires two additional preliminary results. The first is easy to verify.

Lemma 3. *For $u, v \geq 0$,*

$$\sqrt{u+v} \leq \sqrt{u} + \sqrt{v},$$

and for any $\alpha > 0$,

$$2\sqrt{uv} \leq \alpha u + \frac{v}{\alpha}$$

Lemma 4. *Using the notation of Corollary 1, with probability at least $1 - e^{-x}$,*

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} |R_n f|\right] \leq \frac{1}{1-\alpha}\mathbb{E}_\sigma\left[\sup_{f\in F} |R_n f|\right] + \frac{x}{2n\alpha(1-\alpha)}\,.$$

Proof. The last inequality of Theorem 2 and Lemma 3 imply that

$$\begin{aligned}\mathbb{E}\left[\sup_{f\in F} |R_n f|\right] &\leq \mathbb{E}_\sigma\left[\sup_{f\in F} |R_n f|\right] + \sqrt{\frac{2x}{n}\mathbb{E}\left[\sup_{f\in F} |R_n f|\right]} \\ &\leq \mathbb{E}_\sigma\left[\sup_{f\in F} |R_n f|\right] + \alpha\mathbb{E}\left[\sup_{f\in F} |R_n f|\right] + \frac{x}{2n\alpha},\end{aligned}$$

hence, our claim follows. □

Proof. (of Corollary 1) The proof of the first part follows from Theorem 1. Using its notation,

$$\sigma^2 = \sup_{f\in\mathcal{F}} \mathsf{Var}[f(X_i)] \leq r.$$

Thus, with probability at least $1 - e^{-x}$, we have

$$V \leq \mathbb{E}\,[V] + \sqrt{\frac{2xr}{n} + \frac{4x\mathbb{E}\,[V]}{n}} + \frac{x}{3n}\,.$$

The first part of the corollary follows from Lemma 3. The second part follows by combining this inequality with Lemmas 2 and 4. □

5 Error Bounds and Global Averages

Using the various concentration results presented in the previous section, we are now ready to present the error bounds we promised. The bounds are based on the Rademacher averages of the entire class.

Simply applying uniform convergence results to the entire class leads to unsatisfactory error bounds. Better results can be obtained by using various normalization or scaling schemes, in which more weight is assigned to functions that are likely to be the best ones in the class (that is, closer to 0), thus allowing one to "zoom in" on the best region of the space.

When the functions in the class are nonnegative and bounded one can normalize by dividing class members by Pf. Otherwise one normalizes by dividing by Pf^2. In both cases, one needs to be careful when getting too close to zero. Hence, the normalization is conducted only for functions which are "sufficiently far" from 0.

5.1 Normalizing by Pf

Let us introduce some additional notation. Given the class $\mathcal{F}$ and some $r > 0$ set

$$\mathcal{G}_r = \left\{ \frac{r}{r \vee Pf} f : f \in \mathcal{F} \right\} .$$

It is easy to see that $\mathcal{G}_r \subset \{\alpha f : f \in \mathcal{F},\ \alpha \in [0,1]\} \subset \mathrm{star}(\mathcal{F}, 0)$. It is also easy to check that every $g \in G_r$ satisfies $Pg \le r$. Define

$$V_r = \sup_{g \in \mathcal{G}_r} Pg - P_n g.$$

Lemma 5. *For any $r > 0$ and $K > 1$, if $V_r \le \frac{r}{K}$ then every $f \in \mathcal{F}$ satisfies*

$$Pf \le \frac{K}{K-1} P_n f + \frac{r}{K}.$$

Proof. Observe that for any $g \in \mathcal{G}_r$, $Pg \le P_n g + V_r$. Let $f \in \mathcal{F}$ and thus $g = rf/(r \vee Pf)$.

First, assume that $Pf \le r$. By the definition of $\mathcal{G}_r$, $g = f$, hence $Pf \le P_n f + V_r \le (K/(K-1))P_n f + r/K$.

Otherwise, if $Pf > r$, then $g = rf/Pf$. Therefore, since $Pg \le P_n g + V_r$ then

$$r \le r \frac{P_n f}{Pf} + V_r.$$

In that case, if $V_r \le r/K < r$, we obtain

$$Pf \le \frac{P_n f}{1 - V_r/r} \le \frac{K}{K-1} P_n f \le \frac{K}{K-1} P_n f + \frac{r}{K} .$$

□

Combining this deterministic result and the probabilistic estimates of the previous section gives the main result of this section.

Lemma 6. *Let $\mathcal{F}$ be a class of nonnegative functions which are bounded by 1. For any $K > 1$ and $x > 0$ there is a set of probability larger than $1 - 2e^{-x}$, such that for any $f \in \mathcal{F}$,*

$$Pf \le \frac{K}{K-1} P_n f + 12 \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} |R_n f| \right] + \frac{(2K+17)x}{n} .$$

Although this result suggests that the error decreases at the rate $1/n$, this is misleading because one can show that if $\mathcal{F}$ has even a single function for which $Pf^2 \ge c$ then $\mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} |R_n f| \right] \ge c'/\sqrt{n}$, where c' depends only on c. Thus, when applied to the whole class, this bound will typically be dominated by the Rademacher term. However, when we consider subsets of the class, this result will enable us to obtain error bounds using local averages.

Proof. We begin with the observation that functions in $\mathcal{G}_r$ satisfy that $Pg^2 \leq r$. Indeed, if $Pf \leq r$ then $g = f$ and $Pg^2 = Pf^2 \leq Pf \leq r$ (where we used the fact that f maps to $[0,1]$, and hence $f^2 \leq f$). Otherwise, if $Pf > r$ then $g = rf/Pf$ and $Pg^2 = r^2Pf^2/(Pf)^2 < rPf^2/Pf \leq r$.

Applying Corollary 1 and Lemma 2 to V_r it follows that for any $0 < \alpha < 1$,

$$V_r \leq 2(1+\alpha)\mathbb{E}\left[\sup_{g\in\mathcal{G}}|R_n g|\right] + \sqrt{\frac{2rx}{n}} + \left(\frac{1}{3}+\frac{1}{\alpha}\right)\frac{x}{n}$$

with probability larger than $1 - e^{-x}$.

Since $\mathcal{G}_r \subset \mathrm{star}(\mathcal{F}, 0)$ and by Lemma 1,

$$\mathbb{E}\left[\sup_{g\in\mathcal{G}}|R_n g|\right] \leq \mathbb{E}\left[\sup_{f\in\mathrm{star}(\mathcal{F},0)}|R_n f|\right] = \mathbb{E}\left[\sup_{f\in\mathcal{F}}|R_n f|\right].$$

Therefore, with probability at least $1 - e^{-x}$,

$$V_r \leq 2(1+\alpha)\mathbb{E}\left[\sup_{f\in\mathcal{F}}|R_n f|\right] + \sqrt{\frac{2rx}{n}} + \left(\frac{1}{3}+\frac{1}{\alpha}\right)\frac{x}{n}.$$

Now, we can select a specific value of r (which is denoted by r^*). Fix some $K > 1$ and set r to be such that the term on the right hand side is equal to r/K. (Such an r always exists, since the right hand side is of the form $C + A\sqrt{r}$.) Note that if $r/K = C + A\sqrt{r}$ then $r \leq K^2A^2 + 2KC$, and thus r^* may be selected as

$$\begin{aligned} r^* &\leq 4K(1+\alpha)\mathbb{E}\left[\sup_{f\in\mathcal{F}}|R_n f|\right] + \left(2K^2 + 2K\left(\frac{1}{3}+\frac{1}{\alpha}\right)\right)\frac{x}{n} \\ &\leq 4K\frac{1+\alpha}{1-\alpha}\mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}}|R_n f|\right] + \frac{2K(1+\alpha)x}{n\alpha(1-\alpha)} + \left(2K^2 + 2K\left(\frac{1}{3}+\frac{1}{\alpha}\right)\right)\frac{x}{n}, \end{aligned}$$

where the second inequality holds with probability of at least $1-e^{-x}$ by Lemma 4. By Lemma 5 there is a set of probability at least $1 - 2e^{-x}$ such that for every $f \in \mathcal{F}$,

$$Pf \leq \frac{K}{K-1}P_n f + 4\frac{1+\alpha}{1-\alpha}\mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}}|R_n f|\right] + \frac{2(1+\alpha)x}{n\alpha(1-\alpha)} + 2\left(K + \left(\frac{1}{3}+\frac{1}{\alpha}\right)\right)\frac{x}{n},$$

and our result follows by taking $\alpha = 1/2$. □

5.2 Normalizing by Pf^2

In this section, we present error bounds in the more general case, where the elements of $\mathcal{F}$ are not assumed to be nonnegative. The results are analogous to those of the previous section, but here we must impose the assumption that the

variance of class members is small if their expectation is small. In particular, we assume that there is some $B > 0$ such that for every $f \in \mathcal{F}$, $Pf^2 \leq BPf$.

Given a class $\mathcal{F}$ let

$$\mathcal{G}_r = \left\{ \frac{r}{r \vee Pf^2} f : f \in \mathcal{F} \right\} ,$$

and note that $\mathcal{G}_r \subset \{\alpha f : f \in \mathcal{F},\ \alpha \in [0,1]\} \subset \mathrm{star}(\mathcal{F}, 0)$. Define

$$V_r = \sup_{g \in \mathcal{G}_r} Pg - P_n g.$$

Lemma 7. *Let $\mathcal{F}$ be a class of functions with ranges in $[-1,1]$. Assume that there is a constant $B > 0$ such that for every $f \in \mathcal{F}$, $Pf^2 \leq BPf$. For every $r > 0$ and $K > 1$ which satisfy that $V_r \leq \frac{r}{BK}$, and any $f \in \mathcal{F}$,*

$$Pf \leq \frac{K}{K-1} P_n f + \frac{r}{BK} .$$

The proof is similar to that of Lemma 5, we omit it due to lack of space.

Lemma 8. *Let $\mathcal{F}$ be a class of functions with ranges in $[-1,1]$. Assume that there is a constant $B > 0$ such that for every $f \in \mathcal{F}$, $Pf^2 \leq BPf$. For any $K > 1$ and $x > 0$ there is a set of probability larger than $1 - 2e^{-x}$, on which for any $f \in \mathcal{F}$,*

$$Pf \leq \frac{K}{K-1} P_n f + 12\mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} |R_n f| \right] + \frac{(2BK + 22)x}{n} .$$

Again, the proof is similar to the one used in the previous section, and we omit it.

6 Error Bounds and Local Averages

Thus far, the Rademacher averages of the entire class were used as the complexity measure in the error bounds. In this section our aim is to show that the *local* Rademacher averages can serve as a complexity measure. The advantage in using the local version of the averages is that they can be considerably smaller than the global ones.

6.1 Distribution Dependent Complexity

Our analysis is connected to the idea of *peeling* and was inspired by the work of Massart [8].

Theorem 3. *Let $\mathcal{F}$ be a class of functions with ranges in $[0,1]$. Let ψ be a function such that $\psi(r)/\sqrt{r}$ is non-increasing for $r>0$ and for all $r \geq 0$*

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

For any $K>1$ and every $x>0$, there is a set of probability larger than $1-e^{-x}$, on which every $f\in\mathcal{F}$ satisfies

$$Pf \leq \frac{K}{K-1}P_nf + 300Kr_0 + \frac{x(5+4K)}{n}\,,$$

where r_0 is the largest solution of the equation $r_0=\psi(r_0)$.

Proof. Let $\mathcal{G}_r$ be defined as in section 5.1. Fix any $r>0$ and recall that for all $x>0$, with probability $1-e^{-x}$,

$$V_r \leq 2(1+\alpha)\mathbb{E}\left[\sup_{g\in\mathcal{G}_r}|R_ng|\right] + \sqrt{\frac{2rx}{n}} + \left(\frac{1}{3}+\frac{1}{\alpha}\right)\frac{x}{n}\,.$$

Let $\mathcal{F}(x,y) := \{f : x\leq Pf\leq y,\, f\in\mathcal{F}\}$. Let $\lambda>1$ and let k be the smallest integer such that $r\lambda^{k+1}\geq 1$. We have

$$\begin{aligned}
\mathbb{E}\left[\sup_{g\in\mathcal{G}_r}|R_ng|\right] &\leq \mathbb{E}\left[\sup_{\mathcal{F}(0,r)}|R_nf|\right] + \mathbb{E}\left[\sup_{\mathcal{F}(r,1)}\frac{r}{Pf^2}|R_nf|\right] \\
&\leq \mathbb{E}\left[\sup_{\mathcal{F}(0,r)}|R_nf|\right] + \sum_{j=0}^{k}\mathbb{E}\left[\sup_{\mathcal{F}(r\lambda^j, r\lambda^{j+1})}\frac{r}{Pf^2}|R_nf|\right] \\
&\leq \mathbb{E}\left[\sup_{\mathcal{F}(0,r)}|R_nf|\right] + \sum_{j=0}^{k}\lambda^{-j}\mathbb{E}\left[\sup_{\mathcal{F}(r\lambda^j, r\lambda^{j+1})}|R_nf|\right] \\
&\leq \psi(r) + \sum_{j=0}^{k}\lambda^{-j}\psi(r\lambda^{j+1})\,.
\end{aligned}$$

By our assumption it follows that for $\alpha\geq 1$

$$\psi(\alpha r) \leq \sqrt{\alpha}\psi(r)\,,$$

so that

$$\mathbb{E}\left[\sup_{g\in\mathcal{G}_r}|R_ng|\right] \leq \psi(r)\left(1+\sqrt{\lambda}\sum_{j=0}^{k}\lambda^{-j/2}\right)\,,$$

and taking $\lambda=4$, the right-hand side is upper bounded by $5\psi(r)$. Moreover, for $r\geq r_0$, we have $\psi(r)\leq\sqrt{r/r_0}\psi(r_0)=\sqrt{rr_0}$, and thus

$$V_r \leq 10(1+\alpha)\sqrt{rr_0} + \sqrt{\frac{2rx}{n}} + \left(\frac{1}{3}+\frac{1}{\alpha}\right)\frac{x}{n}\,.$$

Let r^* be the value of r for which the right hand side is equal to r^*/K. It follows that

$$r^* \leq K^2 \left(10(1+\alpha)\sqrt{r_0} + \sqrt{\frac{2x}{n}} \right)^2 + 2K \left(\frac{1}{3} + \frac{1}{\alpha} \right) \frac{x}{n} \,,$$

(we use the fact that $r/K = C + A\sqrt{r}$ implies $r \leq K^2 A^2 + 2KC$). The same reasoning as in previous proofs gives the result. □

We now derive a similar result for general classes of functions.

Theorem 4. *Let $\mathcal{F}$ be a class of functions with ranges in $[-1,1]$ and assume that there is some B such that for every $f \in \mathcal{F}$, $Pf^2 \leq BPf$. Let ϕ be a function such that $\phi(r)/\sqrt{r}$ is non-increasing for $r > 0$ and*

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

For any $K > 1$ and every $x > 0$, there is a set of probability larger than $1 - e^{-x}$, on which every $f \in \mathcal{F}$ satisfies

$$Pf \leq \frac{K}{K-1} P_n f + 300 K r_0 + \frac{x(10 + 4BK)}{n} \,,$$

where r_0 is the largest solution of the equation $r_0 = \phi(r_0)$.

Notice that by Lemma 1, when $\mathcal{F}$ is star-shaped around 0, one can choose

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

Proof. Let $\mathcal{G}_r$ be defined as in section 5.2. As before we have with probability $1 - e^{-x}$,

$$V_r \leq 2(1+\alpha)\mathbb{E}\left[\sup_{g \in \mathcal{G}_r} |R_n g| \right] + \sqrt{\frac{2rx}{n}} + 2\left(\frac{1}{3} + \frac{1}{\alpha} \right) \frac{x}{n} \,.$$

Also we can prove as in the previous proof,

$$\mathbb{E}\left[\sup_{g \in \mathcal{G}_r} |R_n g| \right] \leq \phi(r) + \sum_{j=0}^{k} \lambda^{-j} \phi(r\lambda^{j+1})) \,.$$

By our assumption it follows that for $\alpha \geq 1$

$$\phi(\alpha r) \leq \sqrt{\alpha}\phi(r) \,,$$

so that we can get as before

$$V_r \leq 10(1+\alpha)\sqrt{r r_0} + \sqrt{\frac{2rx}{n}} + 2\left(\frac{1}{3} + \frac{1}{\alpha} \right) \frac{x}{n} \,.$$

Choosing r^* such that the right-hand side is equal to r^*/BK and using the same reasoning as in the previous proof gives the result. □

6.2 Applications

As an application of Theorem 4 we present oracle inequalities for the empirical risk minimization procedure, i.e. inequalities that relate the performance of minimizer of the empirical risk to that of the best possible function in the class.

For a class $\mathcal{G}$ of functions, we consider the class $\mathcal{F}$ defined as in (1) and we define

$$\bar{\mathcal{F}} = \{f - f^* : f \in \mathcal{F}\},$$

where $f^* = \arg\min_{f \in \mathcal{F}} Pf$ is the function achieving minimal error in the class $\mathcal{F}$. We want to apply Theorem 4 to $\bar{\mathcal{F}}$, we thus have to check that there exists some $B > 0$ such that

$$\forall f \in \mathcal{F},\ P(f - f^*)^2 \leq B(Pf - Pf^*)\,. \tag{2}$$

Corollary 2. *Let $\mathcal{F}$ be a class of functions with range in $[0,1]$, satisfying (2) for some B. Let f_n be a function in $\mathcal{F}$ such that*

$$P_n f_n = \inf_{f \in \mathcal{F}} P_n f\,.$$

Let ϕ and r_0 be defined as in Theorem 4 for the class $\bar{\mathcal{F}}$. Then for every $x > 0$, there is a set of probability larger than $1 - e^{-x}$ on which we have

$$Pf_n - Pf^* \leq 300 r_0 + \frac{x(10 + 4B)}{n}\,.$$

As mentioned before, condition (2) is satisfied by the class of quadratic loss functions associated with a convex function class [7]. It is also satisfied for discrete loss in pattern classification under specific circumstances. For example, suppose that some function in the class has conditional error probability uniformly bounded over x, that is, for some $0 \leq \eta < 1/2$, there is a function $g^* \in \mathcal{G}$ satisfying

$$\forall x,\ \Pr(Y \neq g^*(X) | X = x) \leq \eta.$$

Then if f^* is the corresponding loss function ($f^*(x,y) = \ell(g^*(x), y)$), then it is straightforward to show that, for any $f \in \mathcal{F}$, $P(f - f^*)^2 \leq (Pf - Pf^*)/(1 - 2\eta)$.

6.3 Data Dependent Complexity

In this section we present a *computable* iterative procedure that gives error bounds, at least in the proper case, without having *a priori* knowledge regarding the global structure of the given class.

We present a result due to Koltchinskii and Panchenko [6] which gives an upper bound on the generalization error of the minimizer of the empirical risk in terms of local Rademacher averages computed on empirical balls.

Theorem 5 ([6]). *Let $x > 0$, N be a positive integer and define*

$$r_0 = 1 \quad \textit{and} \quad r_{k+1} = 1 \wedge \left(K_1 \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F} : P_n f < 2r_k} |R_n f| \right] + K_2 \sqrt{\frac{2r_k x}{n}} + K_3 \frac{x}{n} \right) ,$$

Then for some appropriate choice of the constants K_i we have that for any integer $k \leq N$ there is a set of probability larger than $1 - Ne^{-x}$ on which

$$P f_n \leq r_k \, .$$

This result only applies to classes of nonnegative functions that contain 0 (proper learning or zero-error case) and the result is valid for empirical risk minimization only, which limits its applications.

As proved in [6], it turns out that if there exists a non-decreasing concave function ψ such that

$$\mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}, P_n f \leq r} R_n(f) \right] \leq \psi(\sqrt{r}) \, ,$$

then the solution $\hat{r}$ of the equation $r = \psi(\sqrt{r})$, is, with high probability, and up to a constant, an upper bound on r_N for a some large enough N (roughly of the order of $\log \log 1/\hat{r}$).

In order to use the result of Theorem 5, one has to compute, from the training sample, the quantity $\mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}, P_n f \leq r} R_n(f) \right]$. In the classification case, we consider a class $\mathcal{G}$ of $\{-1, 1\}$-valued functions and $\ell(x, y) = \frac{1}{2}|x - y|$. As noticed in [1], we have

$$\mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} R_n(f) \right] = \frac{1}{2} - \mathbb{E}_\sigma \left[\inf_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^{n} \ell(g(x_i), \sigma_i y_i) \right] .$$

This shows that computing the global Rademacher average is equivalent to minimizing the empirical error on a sample where the labels have been randomly switched.

For the local Rademacher averages, we have an extra constraint on $P_n f$ which can be handled, for example, by modifying the functional to optimize. Indeed, one can prove that for all $r > 0$, and all fixed σ, there is a λ such that

$$\sup_{f \in \mathcal{F}, P_n f \leq r} R_n(f) = \frac{1 - \lambda}{2} - \inf_{g \in G} \frac{1}{n} \sum_{i=1}^{n} (1 - \sigma_i \lambda) \ell(g(x_i), \sigma_i y_i) \, .$$

This amounts to minimizing a weighted empirical error after switching of the labels. However, the value of λ which satisfies this equality may depend on σ and r. Since we need to compute the local Rademacher average for various values of r (in order to find r_0), this corresponds to solving the above minimization problem for various values of λ. It remains to be investigated whether this can be done efficiently for real-world data.

Acknowledgements

We are grateful to Gábor Lugosi and Pascal Massart for many inspiring discussions.

References

1. P. L. Bartlett, S. Boucheron and G. Lugosi. Model selection and error estimation. *Machine Learning*, 48, 85–113, 2002.
2. P. L. Bartlett and S. Mendelson. Rademacher and Gaussian Complexities: Risk Bounds and Structural Results. Proceedings of the Fourteenth Annual Conference on Computational Learning Theory, pp. 224–240, 2001.
3. S. Boucheron, G. Lugosi and P. Massart. Concentration inequalities using the entropy method. *Preprint*, CNRS-Université Paris-Sud. 2002.
4. O. Bousquet. A Bennett concentration inequality and its application to empirical processes. *C.R. Acad. Sci. Paris*, Ser. I, 334, pp. 495-500, 2002.
5. V. I. Koltchinskii. Rademacher penalties and structural risk minimization. Technical report, University of New Mexico, 2000.
6. V. I. Koltchinskii and D. Panchenko. Rademacher processes and bounding the risk of function learning. In *High Dimensional Probability II*, Eds. E. Gine, D. Mason and J. Wellner, pp. 443 - 459, 2000.
7. W. S. Lee, P. L. Bartlett and R. C. Williamson. Efficient agnostic learning of neural networks with bounded fan-in. IEEE Transactions on Information Theory, **42**(6), 2118–2132, 1996.
8. P. Massart. Some applications of concentration inequalities to statistics. *Annales de la Faculté des Sciences de Toulouse*, IX:245-303, 2000.
9. S. Mendelson. Improving the sample complexity using global data. *IEEE Transactions on Information Theory*, 2002.
10. S. Mendelson and R. C. Williamson Agnostic learning nonconvex classes of function. To appear in Proceedings of the Fifteenth Annual Conference on Computational Learning Theory, 2002.
11. E. Rio Une inegalité de Bennett pour les maxima de processus empiriques. Colloque en l'honneur de J. Bretagnolle, D. Dacunha-Castelle et I. Ibragimov, 2001.
12. A. W. van der Vaart and J. A. Wellner. Weak Convergence and Empirical Processes. Springer, 1996.