Local Complexities for Empirical Risk Minimization

Peter L. Bartlett[1], Shahar Mendelson[2], and Petra Philips[2]

[1] Division of Computer Science and Department of Statistics
University of California, Berkeley
367 Evans Hall #3860, Berkeley, CA 94720-3860
bartlett@stat.berkeley.edu
[2] RSISE, The Australian National University
Canberra, 0200 Australia
shahar.mendelson@anu.edu.au, petra.philips@anu.edu.au

Abstract. We present sharp bounds on the risk of the empirical minimization algorithm under mild assumptions on the class. We introduce the notion of isomorphic coordinate projections and show that this leads to a sharper error bound than the best previously known. The quantity which governs this bound on the empirical minimizer is the largest fixed point of the function $\xi_n(r) = \mathbb{E}\sup\{|\mathbb{E}f - \mathbb{E}_n f| : f \in F, \mathbb{E}f = r\}$. We prove that this is the best estimate one can obtain using "structural results", and that it is possible to estimate the error rate from data. We then prove that the bound on the empirical minimization algorithm can be improved further by a direct analysis, and that the correct error rate is the maximizer of $\xi_n'(r) - r$, where $\xi_n'(r) = \mathbb{E}\sup\{\mathbb{E}f - \mathbb{E}_n f : f \in F, \mathbb{E}f = r\}$.

Keywords: statistical learning theory, empirical risk minimization, generalization bounds, concentration inequalities, isomorphic coordinate projections, data-dependent complexity.

1 Introduction

Error bounds for learning algorithms measure the probability that a function produced by the algorithm has a small error. Sharp bounds give an insight into the parameters that are important for learning and allow one to assess accurately the performance of learning algorithms. The bounds are usually derived by studying the relationship between the expected and the empirical error. It is now a standard result that, for every function, the deviation of the expected from the empirical error is bounded by a complexity term which measures the size of the function class from which the function was chosen. Complexity terms which measure the size of the entire class are called *global complexity measures*, and two such examples are the VC-dimension and the Rademacher averages of the function class (note that there is a key difference between the two; the

VC-dimension is independent of the underlying measure, and thus captures the worst case scenario, while the Rademacher averages are measure dependent and lead to sharper bounds).

Moreover, estimates which are based on comparing the empirical and the actual structures (for example empirical vs. actual means) uniformly over the class are loose, because this condition is stronger than necessary. Indeed, in the case of the empirical risk minimization algorithm, it is more likely that the algorithm produces functions with a small expectation, and thus one only has to consider a small subclass. Taking that into account, error bounds should depend only on the complexity of the functions with small error or variance. Such bounds in terms of *local complexity measures* were established in [10, 15, 13, 2, 9].

In this article we will show that by imposing very mild structural assumptions on the class, these local complexity bounds can be improved further. We will state the best possible estimates which can be obtained by a comparison of empirical and actual structures. Then, we will pursue the idea of leaving the "structural approach" and analyzing the empirical minimization algorithm directly. The reason for this is that structural results comparing the empirical and actual structures on the class have a limitation. It turns out that if one is too close to the true minimizer the class is too rich at that scale and the structures are not close at a small enough scale to yield a useful bound. On the other hand, with the empirical minimizer one can go beyond the structural limit.

We consider the following setting and notation: let $\mathcal{X} \times \mathcal{Y}$ be a measurable space, and let P be an unknown probability distribution on $\mathcal{X} \times \mathcal{Y}$. Let $((X_1, Y_1), ..., (X_n, Y_n)) \in (\mathcal{X} \times \mathcal{Y})^n$ be a finite training sample, where each pair (X_i, Y_i) is generated independently according to P. The goal of a learning algorithm is to estimate a function $h : \mathcal{X} \longrightarrow \mathcal{Y}$ (based on the sample), which predicts the value of Y given X. The possible choices of functions are all in a function class H, called the hypothesis class. A quantitative measure of how accurate a function $h \in H$ approximates Y is given by a loss function $l : \mathcal{Y}^2 \longrightarrow \mathbb{R}$. Typical examples of loss functions are the 0-1 loss for classification defined by $l(r, s) = 0$ if $r = s$ and $l(r, s) = 1$ if $r \neq s$ or the square-loss for regression tasks $l(r, s) = (r - s)^2$. In what follows we will assume a bounded loss function and therefore, without loss of generality, $l : \mathcal{Y}^2 \longrightarrow [-b, b]$. For every $h \in H$ we define the associated loss function $l_h : (\mathcal{X} \times \mathcal{Y}) \longrightarrow [-b, b]$, $l_h(x, y) = l(h(x), y)$ and denote by $F = \{l_h : (\mathcal{X} \times \mathcal{Y}) \longrightarrow [-b, b] \ : \ h \in H\}$ the loss class associated with the learning problem. The best estimate $h^* \in H$ is the one for which the expected loss (also called risk) is as small as possible, that is, $\mathbb{E}l_{h^*} = \inf_{h \in H} \mathbb{E}l_h$, and we will assume that such an h^* exists and is unique. We call $F' = \{l_h - l_{h^*} : h \in H\}$ the excess loss class. Note that all functions in F' have a non-negative expectation, though they can take negative values, and that $0 \in F'$.

Empirical risk minimization algorithms are based on the philosophy that it is possible to approximate the expectation of the loss functions using their empirical mean, and choose instead of h^* the function $\hat{h} \in H$ for which $\frac{1}{n} \sum_{i=1}^n l_{\hat{h}}(x_i, y_i) \approx \inf_{h \in H} \frac{1}{n} \sum_{i=1}^n l_h(x_i, y_i)$. Such a function is called the empirical minimizer.

In studying the loss class F we will simplify notation and assume that F consists of bounded, real-valued functions defined on a measurable set $\mathcal{X}$, that is, instead of $\mathcal{X} \times \mathcal{Y}$ we only write $\mathcal{X}$. Let $X_1, \ldots, X_n$ be independent random variables distributed according to P. For every $f \in F$, we denote by

$$P_n f = \mathbb{E}_n f = \frac{1}{n} \sum_{i=1}^n f(X_i), \qquad Pf = \mathbb{E} f, \qquad R_n f = \frac{1}{n} \sum_{i=1}^n \sigma_i f(X_i),$$

where $\mathbb{E} f$ is the expectation of the random variable $f(X)$ with respect to P and $\sigma_1, \ldots, \sigma_n$ are independent Rademacher random variables, that is, symmetric, $\{-1, 1\}$-valued random variables. We further denote

$$\|P - P_n\|_F = \sup_{f \in F} |\mathbb{E} f - \mathbb{E}_n f|, \qquad R_n F = \sup_{f \in F} R_n f.$$

The Rademacher averages of the class F are defined as $\mathbb{E} R_n F$, where the expectation is taken with respect to all random variables X_i and σ_i. An empirical version of the Rademacher averages is obtained by conditioning on the sample,

$$\mathbb{E}_\sigma R_n F = \mathbb{E} \left(\sup_{f \in F} \frac{1}{n} \sum_{i=1}^n \sigma_i f(X_i) \middle| X_1, \ldots, X_n \right).$$

Let

$$F_r = \{f \in F : \mathbb{E} f = r\}, \qquad F^n_{r_1, r_2} = \{f \in F : r_1 \leq \mathbb{E}_n f \leq r_2\}.$$

For a given sample, denote by $\hat{f}$ the corresponding empirical risk minimizer, that is, a function that satisfies: $\mathbb{E}_n \hat{f} = \min_{f \in F} \mathbb{E}_n f$. If the minimum does not exist, we denote by $\hat{f} \in F$ any ρ-approximate empirical minimizer, which is a function satisfying

$$\mathbb{E}_n \hat{f} \leq \inf_{f \in F} \mathbb{E}_n f + \rho,$$

where $\rho \geq 0$. Denote the conditional expectation $\mathbb{E}(\hat{f}(X) | X_1, \ldots, X_n)$ by $\mathbb{E} \hat{f}$.

In the following we will show that if the class F is star-shaped and the variance of every function can be bounded by a reasonable function of its expectation, then the quantity which governs both the structural behaviour of the class and the error rate of the empirical minimizer is the function

$$\xi_n(r) = \mathbb{E} \sup_{f \in F_r} |\mathbb{E} f - \mathbb{E}_n f| = \mathbb{E} \|P - P_n\|_{F_r},$$

or minor modifications of $\xi_n(r)$. Observe that this function measures the expectation of the empirical process $\|P - P_n\|$ indexed by the subset F_r. In the classical result, involving a global complexity measure, the resulting bounds are given in terms of $\mathbb{E} \|P - P_n\|$ indexed by the whole set F, and in [10, 15, 13, 2, 9] in terms of the fixed point of $\mathbb{E} \|P - P_n\|$ indexed by the subsets $\{f \in F : \mathbb{E} f \leq r\}$ or $\{f \in F : \mathbb{E} f^2 \leq r\}$, which are all larger sets than F_r. For an empirical minimizer, these structural comparisons lead to the estimate that $\mathbb{E} \hat{f}$ is essentially bounded by $r^* = \inf \left\{ r : \xi_n(r) \leq \frac{r}{4} \right\}$. This result can be improved further: we show that the loss of the empirical minimizer is concentrated around the value $s^* = \operatorname{argmax} \{\xi'_n(r) - r\}$, where $\xi'_n(r) = \mathbb{E} \sup \{\mathbb{E} f - \mathbb{E}_n f : f \in F_r\}$.

2 Preliminaries

In order to obtain the desired results we will require some minor structural assumptions on the class, namely, that F is star-shaped around 0 and satisfies a Bernstein condition.

Definition 1. *We say that F is a (β, B)-Bernstein class with respect to the probability measure P (where $0 < \beta \leq 1$ and $B \geq 1$), if every $f \in F$ satisfies*

$$\mathbb{E}f^2 \leq B(\mathbb{E}f)^\beta.$$

We say that F has Bernstein type β with respect to P if there is some constant B for which F is a (β, B)-Bernstein class.

There are many examples of loss classes for which this assumption can be verified. For example, for nonnegative bounded loss functions, the associated loss function classes satisfy this property with $\beta = 1$. For convex classes of functions bounded by 1, the associated excess squared-loss class satisfies this property as well with $\beta = 1$, a result that was first shown in [12] and improved and extended in [16, 3] e.g. to other power types of excess losses.

Definition 2. *F is called star-shaped around 0 if for every $f \in F$ and $0 \leq \alpha \leq 1$, $\alpha f \in F$.*

We can always make a function star-shaped by replacing F with $\mathrm{star}(F, 0) = \{\alpha f : f \in F, 0 \leq \alpha \leq 1\}$. Although $F \subset \mathrm{star}(F, 0)$, one can show that the complexity measure $\xi_n(r)$ does not increase too much. For star-shaped classes, the function $\xi_n(r)/r$ is non-increasing, a property which will allow us to estimate the largest fixed point of $\xi_n(r)$:

Lemma 1. *If F is star-shaped around 0, then for any $0 < r_1 < r_2$,*

$$\frac{\xi_n(r_1)}{r_1} \geq \frac{\xi_n(r_2)}{r_2}.$$

In particular, if for some α, $\xi_n(r) \geq \alpha r$ then for all $0 < r' \leq r$, $\xi_n(r') \geq \alpha r'$.

Proof: Fix $\tau = (X_1, ..., X_n)$ and without loss of generality, suppose that $\sup_{f \in F_{r_2}} |\mathbb{E}f - \mathbb{E}_n f|$ is attained at f. Then $f' = \frac{r_1}{r_2} f \in F_{r_1}$ satisfies

$$|\mathbb{E}f' - \mathbb{E}_n f'| = \frac{r_1}{r_2} \sup_{f \in F_{r_2}} |\mathbb{E}f - \mathbb{E}_n f|.$$

■

The tools used in the proofs of this article are mostly concentration inequalities. We first state the main concentration inequality used in this article, which is a version of Talagrand's inequality [21, 20, 11].

Theorem 1. *Let F be a class of functions defined on $\mathcal{X}$ and set P to be a probability measure such that for every $f \in F$, $\|f\|_\infty \leq b$ and $\mathbb{E}f = 0$. Let $X_1, ..., X_n$ be independent random variables distributed according to P and set $\sigma^2 = n \sup_{f \in F} \operatorname{var}[f]$. Define*

$$Z = \sup_{f \in F} \sum_{i=1}^{n} f(X_i),$$
$$\bar{Z} = \sup_{f \in F} \left| \sum_{i=1}^{n} f(X_i) \right|.$$

Then there is an absolute constant K such that, for every $x > 0$ and every $\rho > 0$, the following holds:

$$Pr\left(\left\{ Z \geq (1+\rho)\mathbb{E}Z + \sigma\sqrt{Kx} + K(1+\rho^{-1})bx \right\} \right) \leq e^{-x},$$
$$Pr\left(\left\{ Z \leq (1-\rho)\mathbb{E}Z - \sigma\sqrt{Kx} - K(1+\rho^{-1})bx \right\} \right) \leq e^{-x},$$

and the same inequalities hold for $\bar{Z}$.

The inequality for $\bar{Z}$ is due to Massart [14]. The one sided versions were shown by Rio [19] and Klein [7]. For $b = 1$, the best estimates on the constants in all cases are due to Bousquet [6].

Setting $\bar{Z} = \|P - P_n\|_F$ we obtain the following corollary:

Corollary 1. *For any class of functions F, and every $x > 0$, if*

$$\lambda \geq C \max\left\{ \mathbb{E}\|P - P_n\|_F, \sigma_F \sqrt{\frac{x}{n}}, \frac{bx}{n} \right\}, \tag{1}$$

where $\sigma_F^2 = \sup_{f \in F} \operatorname{var}[f]$ and $b = \sup_{f \in F} \|f\|_\infty$, then with probability at least $1 - e^{-x}$, every f in F satisfies

$$|\mathbb{E}f - \mathbb{E}_n f| \leq \lambda.$$

This global estimate is essentially the result obtained in [8, 1, 18]. It is a worst-case result in the sense that it holds uniformly over the entire class, but $\mathbb{E}\|P - P_n\|_F$ is a better measure of complexity than the VC-dimension since it is measure dependent and it is well known that for binary valued classes, $\mathbb{E}\|P - P_n\|_F \leq c\sqrt{VC(F)/n}$. One way of understanding this result is as a method to compare the empirical and actual structure on the class additively up to λ. Condition (1) arises from the two extra terms in Talagrand's concentration inequality. The result is sharp since it can be shown that for large enough n, $\mathbb{E}\|P - P_n\|_F \geq \sigma_F\sqrt{x/n}$, and that with high probability $\|P - P_n\|_F \geq c\mathbb{E}\|P - P_n\|_F$ for a suitable absolute constant c, see e.g. [4]. Therefore, asymptotically, the difference of empirical and actual structures in this sense is controlled by the global quantity $\mathbb{E}\|P - P_n\|_F$, and the error rate obtained using this approach cannot decay faster than $O(1/\sqrt{n})$. In particular, for any ρ-approximate

empirical minimizer, if r satisfies the global condition of the theorem, then with probability at least $1 - e^{-x}$, $\mathbb{E}\hat{f} \leq \mathbb{E}_n\hat{f} + \rho + r$.

The following symmetrization theorem states that the expectation of $\|P - P_n\|_F$ is upper bounded by the Rademacher averages of F, see for example [17].

Theorem 2. *Let F be a class of functions defined on $\mathcal{X}$, set P to be a probability measure on $\mathcal{X}$ and $X_1, ..., X_n$ independent random variables distributed according to P. Then,*

$$\mathbb{E}\|P - P_n\|_F \leq 2\mathbb{E}R_nF\,.$$

The next lemma, following directly from a theorem in [5], shows that the Rademacher averages of a class can be upper bounded by the empirical Rademacher averages of this class. The following formulation can be found in [2].

Theorem 3. *Let F be a class of bounded functions defined on $\mathcal{X}$ taking values in $[a, b]$, P a probability measure on $\mathcal{X}$, and $X_1, ..., X_n$ be independent random variables distributed according to P. Then, for any $0 \leq \alpha \leq 1$ and $x > 0$, with probability at least $1 - e^{-x}$,*

$$\mathbb{E}R_nF \leq \frac{1}{1-\alpha}\mathbb{E}_\sigma R_nF + \frac{(b-a)x}{4n\alpha(1-\alpha)}\,.$$

3 Isomorphic coordinate projections

We now introduce a multiplicative (rather than additive, as in Corollary 1) notion of similarity of the expected and empirical means which characterizes the fact that, for the given sample, for all functions in the class, $|\mathbb{E}f - \mathbb{E}_nf|$ is at most a constant times its expectation.

Definition 3. *For $\tau = (X_1, \ldots, X_n)$, we say that the coordinate projection $\Pi_\tau : f \mapsto (f(X_1), \ldots, f(X_n))$ is an ϵ-isomorphism if for every $f \in F$,*

$$(1-\epsilon)\mathbb{E}f \leq \mathbb{E}_nf \leq (1+\epsilon)\mathbb{E}f.$$

We observe that for star-shaped classes, if, for a given sample τ, a coordinate projection Π_τ is an ϵ-isomorphism on the subset F_r, then the same holds for the larger set $\{f \in F : \mathbb{E}f \geq r\}$.

Lemma 2. *Let F be star-shaped around 0 and let $\tau \in \mathcal{X}^n$. For any $r > 0$ and $0 < \epsilon < 1$, the projection Π_τ is an ϵ-isomorphism of F_r if and only if it is an ϵ-isomorphism of $\{f \in F : \mathbb{E}f \geq r\}$.*

Proof: Let $f \in F$ such that $\mathbb{E}f = t > r$, and since F is star-shaped around 0, $g = rf/t \in F_r$; hence, $(1-\epsilon)\mathbb{E}f \leq \mathbb{E}_n f \leq (1+\epsilon)\mathbb{E}f$ if and only if the same holds for g. ■

Thus, for star-shaped classes, it suffices to analyze this notion of similarity on the subsets F_r. The next result, which establishes this fact, follows from Theorem 1. It states that for every subset F_r, if $\xi_n(r)$ is slightly smaller than r then most projections are ϵ-isomorphisms on F_r (and by Lemma 2 also on $\{f \in F : \mathbb{E}f \geq r\}$). On the other hand, if $\xi_n(r)$ is slightly larger than r, most projections are not ϵ-isomorphisms. Hence, at the value of r for which $\xi_n(r) \sim r$, there occurs a phase transition: above that point the class is small enough and a structural result can be obtained. Below the point, the class F_r, which consists of scaled down versions of all functions $\{f \in F : \mathbb{E}f > r\}$ and "new atoms" with $\mathbb{E}f = r$, is too saturated and statistical control becomes impossible.

Theorem 4. *There is an absolute constant c for which the following holds. Let F be a class of functions, such that for every $f \in F$, $\|f\|_\infty \leq b$. Assume that F is a (β, B)-Bernstein class. Suppose $r \geq 0$, $0 < \epsilon < 1$, and $0 < \alpha < 1$ satisfy*

$$r \geq c \max\left\{ \frac{bx}{n\alpha^2\epsilon}, \left(\frac{Bx}{n\alpha^2\epsilon^2}\right)^{1/(2-\beta)} \right\}.$$

1. *If $\mathbb{E}\left\|P - P_n\right\|_{F_r} \geq (1+\alpha)r\epsilon$, then*

$$\Pr\left\{\Pi_\tau \text{ is not an } \epsilon\text{-isomorphism of } F_r\right\} \geq 1 - e^{-x}.$$

2. *If $\mathbb{E}\left\|P - P_n\right\|_{F_r} \leq (1-\alpha)r\epsilon$, then*

$$\Pr\left\{\Pi_\tau \text{ is an } \epsilon\text{-isomorphism of } F_r\right\} \geq 1 - e^{-x}.$$

Proof: The proof follows in a straightforward way from Theorem 1. Define $Z = n\left\|P - P_n\right\|_{F_r}$, set $\sigma^2 = n\sup_{f\in F_r} \mathrm{var}\left[f\right]$ and note that Π_τ is an ϵ-isomorphism of F_r if and only if $Z \leq \epsilon r n$.

To prove the first part of our claim, recall that by Theorem 1, for every $\rho, x > 0$, with probability larger than $1 - e^{-x}$,

$$Z > (1-\rho)\mathbb{E}Z - \sigma\sqrt{Kx} - K\left(1 + \frac{1}{\rho}\right)bx.$$

To ensure that $Z > \epsilon r n$, select $\rho = \alpha/(2(1+\alpha))$, and observe that by the assumption that F is a Bernstein class, it suffices to show that

$$\frac{1}{2}\alpha n r\epsilon \geq (Bnr^\beta Kx)^{1/2} + K\left(1 + \frac{2(1+\alpha)}{\alpha}\right)xb,$$

which holds by the condition on r.

The second part of the claim also follows from Theorem 1: for every $\rho, x > 0$, with probability larger than $1 - e^{-x}$,

$$Z < (1+\rho)\mathbb{E}Z + \sigma\sqrt{Kx} + K\left(1 + \frac{1}{\rho}\right)bx.$$

Choosing $\rho = \alpha/(2(1-\alpha))$, we see that $Z < nr\epsilon$ if

$$\frac{1}{2}\alpha nr\epsilon \geq (Bnr^{\beta}Kx)^{1/2} + K\left(1 + \frac{2(1-\alpha)}{\alpha}\right)xb,$$

so the condition on r again suffices. ■

Corollary 2. *Let F be a class of functions bounded by b, which is star-shaped around 0 and is a (β, B)-Bernstein class. Then there exists an absolute constant c for which the following holds.If $0 < \epsilon, \alpha < 1$, and $r, x > 0$, satisfy*

$$r \geq \max\left\{\frac{\xi_n(r)}{(1-\alpha)\epsilon}, c\frac{bx}{n\alpha^2\epsilon}, c\left(\frac{Bx}{n\alpha^2\epsilon^2}\right)^{1/(2-\beta)}\right\}, \tag{2}$$

then with probability at least $1 - e^{-x}$, every $f \in F$ satisfies

$$\mathbb{E}f \leq \max\left\{\frac{\mathbb{E}_n f}{1-\epsilon}, r\right\}.$$

Proof: The proof follows directly from Theorem 4. ■

Clearly, Corollary 2 is an improvement on the result in Corollary 1 for most interesting loss classes, for which $0 < \beta \leq 1$. The condition (2) allows one to control the expectation of the empirical minimizer asymptotically up to the scale $O(1/n^{1/2-\beta})$, and for classes with $\beta = 1$ even at the best possible scale $O(1/n)$, as opposed to $O(1/\sqrt{n})$ in Corollary 1. The quantity $\xi_n(r) = \mathbb{E}\left\|P - P_n\right\|_{F_r}$ is also an improvement on $\lambda \sim \mathbb{E}\left\|P - P_n\right\|_F$ from Corollary 1, since the supremum is taken only on the subset F_r which can be much smaller than F.

Corollary 2 also improves the localized results from [2]. In [2] the indexing set is the set of functions with a small variance, $\{f \in F : Pf^2 \leq r\}$, or a sub-root function upper bounding the empirical process indexed by $\{f \in F : Pf \leq r\}$. The advantage of Corollary 2 is that the indexing set F_r is smaller, and that the upper bound in terms of the fixed point can be proved without assuming the sub-root property. The property of $\xi_n(r)$ in Lemma 1, a "sub-linear" property, is sufficient to lead to the following estimate on the empirical minimizer:

Theorem 5. *Let F be a (β, B)-Bernstein class of functions bounded by b which is star-shaped around 0. Then there is an absolute constant c such that if*

$$r' = \max\left\{\inf\left\{r : \xi_n(r) \leq r/4\right\}, \frac{cbx}{n}, c\left(\frac{Bx}{n}\right)^{1/(2-\beta)}\right\},$$

then with probability at least $1 - e^{-x}$, a ρ-approximate empirical minimizer $\hat{f} \in F$ satisfies

$$\mathbb{E}\hat{f} \leq \max\{2\rho, r'\}.$$

Proof: The proof follows from Corollary 2 by taking $\epsilon = \alpha = 1/2$ and $r = r'$. In particular, Lemma 1 shows that if $r' \geq \inf\left\{r : \xi_n(r) \leq \frac{r}{4}\right\}$, then $\xi_n(r') \leq r'/4$. Thus, with large probability, if $f \in F$ satisfies $\mathbb{E}f \geq r'$, then $\mathbb{E}f \leq 2\mathbb{E}_n f$. Since $\hat{f}$ is a ρ-approximate empirical minimizer and F is star-shaped at 0, it follows that $\mathbb{E}_n \hat{f} \leq \rho$, so either $\mathbb{E}f \leq r'$ or $\mathbb{E}f \leq 2\rho$, as claimed. ■

Thus, with high probability, $r^* = \inf\left\{r : \xi_n(r) \leq \frac{r}{4}\right\}$ is an upper bound for $\mathbb{E}\hat{f}$, as long as $r^* \geq c/n$.

This result holds in particular for any empirical minimizer of the excess loss class if the true minimizer f^* exists. In this case, $0 \in F$, and any empirical minimizer over F is also an empirical minimizer over $\mathrm{star}(F, 0)$.

Data-dependent estimation of $\xi_n(r)$ and r^*

The next question we wish to address is how to estimate the function $\xi_n(r)$ and the fixed point

$$r^* = \inf\left\{r : \xi_n(r) \leq \frac{r}{4}\right\}$$

empirically, in cases where the global complexity of the function class, for example the covering numbers or the combinatorial dimension, is not known.

To estimate r^* we will find an empirically computable function $\hat{\xi}_n(r)$ which is, with high probability, an upper bound for the function $\xi_n(r)$. Therefore, it will hold that its fixed point $\hat{r}^* = \inf\{r : \hat{\xi}_n(r) \leq \frac{r}{4}\}$ is with high probability an upper bound for r^*. Since $\hat{\xi}_n(r)/r$ will be a non-increasing function, we will be able to determine $\hat{r}^*$ using a binary search algorithm.

Assume that F is a star-shaped (β, B)-Bernstein class and $\sup_{f \in F} \|f\|_\infty \leq b$. Let $\tau = (X_1, ..., X_n)$ be a sample, where each X_i is drawn independently according to P.

From Theorem 4, for $\alpha = 1/2, \epsilon = 1/2$, if $r \geq c \max\left\{\frac{bx}{n}, \left(\frac{Bx}{n}\right)^{1/(2-\beta)}\right\}$ and $\xi_n(r) \leq \frac{r}{4}$, then with probability larger than $1 - e^{-x}$, every $f \in F_r$ satisfies that

$$\forall f \in F_r : \mathbb{E}_n f \in \left[\frac{r}{2}, \frac{3r}{2}\right].$$

Since F is star-shaped, and by Lemma 1, it holds that $\xi_n(r) \leq \frac{r}{4}$ if and only if $r \geq r^*$. Therefore, if $r \geq \max\left\{r^*, \frac{cbx}{n}, c\left(\frac{Bx}{n}\right)^{1/(2-\beta)}\right\}$, then with probability larger than $1 - e^{-x}$, $F_r \subset F^n_{\frac{r}{2}, \frac{3r}{2}}$, which implies that

$$\mathbb{E}_\sigma R_n\left(F_r\right) \leq \mathbb{E}_\sigma R_n\left(F^n_{\frac{r}{2}, \frac{3r}{2}}\right),$$

where $F^n_{r_1, r_2} = \{f \in F : r_1 \leq \mathbb{E}_n f \leq r_2\}$.

By symmetrization (Theorem 2) and concentration of Rademacher averages around their mean (Theorem 3), it follows that with probability at least $1 - 2e^{-x}$,

$$\xi_n(r) \leq 2\mathbb{E}R_n(F_r) \leq 4\mathbb{E}_\sigma R_n(F_r) + \frac{bx}{n} \leq 4\mathbb{E}_\sigma R_n\left(F^n_{\frac{r}{2}, \frac{3r}{2}}\right) + \frac{r}{c},$$

where we used the fact that $r \geq \frac{cbx}{n}$ (and clearly we can assume that $c > 8$).

Set

$$r' = \max\left\{r^*, \frac{cbx}{n}, c\left(\frac{Bx}{n}\right)^{1/(2-\beta)}\right\}, \text{and}$$

$$R = \left\{\frac{1}{n}, \frac{2}{n}, \ldots, \frac{\lceil bn \rceil}{n}\right\} \cap \left[\frac{\lfloor r'n \rfloor}{n}, \frac{\lceil bn \rceil}{n}\right].$$

Applying the union bound, and since $|R| \leq bn + 1$, with probability at least $1 - 2(bn+1)e^{-x}$, $\xi_n(r) \leq 4\mathbb{E}_\sigma R_n\left(F^n_{\frac{r}{2},\frac{3r}{2}}\right) + \frac{r}{c}$ for every $r \in R$. By Lemma 1, if $r \in [k/n, (k+1)/n]$, then $\xi_n(r) \leq \xi_n\left(\frac{k}{n}\right)\frac{nr}{k}$, and thus, with probability at least $1 - 2(bn+1)e^{-x}$, every $r \in [r', b]$ satisfies

$$\xi_n(r) \leq \xi_n\left(\frac{k}{n}\right)\frac{nr}{k} \leq \left(4\mathbb{E}_\sigma R_n\left(F^n_{\frac{k}{2n},\frac{3k}{2n}}\right) + \frac{k}{cn}\right)\frac{nr}{k} \leq 8\mathbb{E}_\sigma R_n\left(F^n_{c_1 r, c_2 r}\right) + \frac{r}{c},$$

where c_1, c_2 are positive constants. We define therefore

$$\hat{\xi}_n(r) = 8\mathbb{E}_\sigma R_n\left(F^n_{c_1 r, c_2 r}\right) + \frac{r}{c}.$$

Then it follows that with probability at least $1 - 2(bn+1)e^{-x}$

$$\forall r \in [r', b] : \xi_n(r) \leq \hat{\xi}_n(r)\,.$$

Let $\hat{r}^* = \inf\{r : \hat{\xi}_n(r) \leq \frac{r}{4}\}$, then we know that with probability at least $1 - 2(bn+1)e^{-x}$, $\hat{r}^* \geq r^*$. Since $\hat{\xi}_n(r)/r$ is non-increasing, it follows that $r \geq \hat{r}^*$ if and only if $\hat{\xi}_n(r) \leq \frac{r}{4}$.

With this, given a sample of size n, we are ready to state the following algorithm to estimate the upper bound on $\hat{r}^*$ based on the data:

Algorithm RSTAR(F, $X_1, \ldots, X_n$)

Set $r_L = 0$, $r_R = b$.
If $\hat{\xi}_n(r_R) \leq r_R/4$ then
  for $l = 0$ to $\lceil \log_2 bn \rceil$
    set $r = \frac{r_R - r_L}{2}$;
    if $\hat{\xi}_n(r) > r/4$ then set $r_L = r$,
      else set $r_R = r$.
Output $\bar{r} = r_R$.

By the construction, $\bar{r} - \frac{1}{n} \leq \hat{r}^* \leq \bar{r}$. For every n and every sample, with probability larger than $1 - 2(bn+1)e^{-x}$, $r^* \leq \bar{r}$.

Theorem 6. *Let F be a (β, B)-Bernstein class of functions bounded by b which is star-shaped around 0. With probability at least $1 - (2bn+3)e^{-x}$, a ρ-approximate empirical minimizer $\hat{f} \in F$ satisfies*

$$\mathbb{E}\hat{f} \leq \max\{2\rho, r''\},$$

where

$$r'' = \max\left\{\bar{r}, \frac{cbx}{n}, c\left(\frac{Bx}{n}\right)^{1/(2-\beta)}\right\},$$

and $\bar{r} = RSTAR(F, \tau)$.

RSTAR(F, τ) is essentially the fixed point of the $\mathbb{E}_\sigma R_n\left(F^n_{c_1 r, c_2 r}\right)$. This function measures the complexity of the function class $F^n_{c_1 r, c_2 r}$ which is the subset of functions having the empirical mean in an interval whose length is proportional to r. The main difference from the data-dependent estimates in [2] is that instead of taking the whole empirical ball, here we only measure the complexity of an empirical "belt" around r, since $c_1 r > 0$.

We can tighten this bound further by narrowing the size of the belt by replacing the empirical set $F^n_{r/2,3r/2}$ with $F^n_{r-r/\log n, r+r/\log n}$. The price we pay is an extra $\log n$ factor.

With the same reasoning as before, by Theorem 4 for $\alpha = 1/2, \epsilon = 1/\log n$, and since F is star-shaped, then, if $r \geq \max\left\{r^*, \frac{cbx\log n}{n}, c\left(\frac{Bx\log^2 n}{n}\right)^{1/(2-\beta)}\right\}$, with probability larger than $1 - e^{-x}$, $F_r \subset F^n_{r-r/\log n, r+r/\log n}$. We define

$$\hat{\xi}_n(r) = \left(4\mathbb{E}_\sigma R_n\left(F^n_{k/n - k/(n\log n), k/n + k/(n\log n)}\right) + \frac{k}{cn\log n}\right)\frac{n}{k}r,$$

if $r \in [k/n, (k+1)/n]$. Again, with probability at least $1 - 2(bn+1)e^{-x}$, it holds that for all $r \in [r', b] : \xi(r) \leq \hat{\xi}_n(r)$, where

$$r' = \max\left\{r^*, \frac{cbx\log n}{n}, c\left(\frac{Bx\log^2 n}{n}\right)^{1/(2-\beta)}\right\}.$$

Since $\hat{\xi}_n(r)/r$ is non-increasing, we can compute

$$\hat{r}^* = \inf\left\{r : \hat{\xi}_n(r) \leq \frac{r}{2\log n}\right\}$$

with a slight modification of RSTAR (we replace the test in the if-clause, $\hat{\xi}_n(r) > r/4$, with $\hat{\xi}_n(r) > r/2\log n$). For every n and every sample of size n, with probability larger than $1 - 2(bn+1)e^{-x}$, $r^* \leq \bar{r}$.

4 Direct concentration result for empirical minimizers

In this section we will now show that a direct analysis of the empirical minimizer leads to sharper estimates than those obtained in the previous section. We will show that $\mathbb{E}\hat{f}$ is concentrated around the value $s^* = \text{argmax}\{\xi'_n(r) - r\}$, where

$$\xi'_n(r) = \mathbb{E}\sup\left\{\mathbb{E}f - \mathbb{E}_n f : f \in F, \mathbb{E}f = r\right\}.$$

To understand why it makes sense to expect that with high probability $\mathbb{E}\hat{f} \sim s^*$, fix one value of r such that $\xi_n'(s^*) - s^* > \xi_n'(r) - r$. Consider a perfect situation in which one could say that with high probability,

$$\xi_n'(r) \sim \sup\{\mathbb{E}f - \mathbb{E}_n f : f \in F, \mathbb{E}f = r\} = r - \inf\{\mathbb{E}_n f : f \in F, \mathbb{E}f = r\}.$$

(Of course, this is not the case, as Talagrand's inequality contains additional terms which blow-up as the multiplicative constant represented by $\sim$ tends to one; this fact is the crux of the proof.) In that case, it would follow that

$$-\inf\{\mathbb{E}_n f : f \in F, \mathbb{E}f = s^*\} > -\inf\{\mathbb{E}_n f : f \in F, \mathbb{E}f = r\}$$

and the empirical minimizer will not be in F_r. In a similar manner, one has to rule out all other values of r, and to that end we will have to consider a belt around s^* rather than s^* itself.

For $\epsilon > 0$, define

$$r_{\epsilon,+} = \sup\left\{0 \le r \le b : \xi_n'(r) - r \ge \sup_s (\xi_n'(s) - s) - \epsilon\right\},$$
$$r_{\epsilon,-} = \inf\left\{0 \le r \le b : \xi_n'(r) - r \ge \sup_s (\xi_n'(s) - s) - \epsilon\right\}.$$

The following theorem is the main result:

Theorem 7. *For any $c_1 > 0$, there is a constant c (depending only on c_1) such that the following holds. Let F be a (β, B)-Bernstein class that is star-shaped at 0. Define $r_{\epsilon,+}$, and $r_{\epsilon,-}$ as above, and set*

$$r' = \max\left\{\inf\{r : \xi_n'(r) \le r/4\}, \frac{cb(x+\log n)}{n}, c\left(\frac{B(x+\log n)}{n}\right)^{1/(2-\beta)}\right\}.$$

For $0 < \rho < r'/2$, let $\hat{f}$ denote a ρ-approximate empirical risk minimizer. If

$$\epsilon \ge c\left(\max\left\{\sup_s (\xi_n'(s) - s), r'^{\beta}\right\}\frac{(B+b)(x+\log n)}{n}\right)^{1/2} + \rho,$$

then

1. *With probability at least $1 - e^{-x}$,*
$$\mathbb{E}\hat{f} \le \max\left\{\frac{1}{n}, r_{\epsilon,+}\right\}.$$
2. *If*
$$\xi_n'(0, c_1/n) < \sup_s (\xi_n'(s) - s) - \epsilon,$$
then with probability at least $1 - e^{-x}$,
$$\mathbb{E}\hat{f} \ge r_{\epsilon,-}.$$

Note that this result is considerably sharper than the bound resulting from Theorem 5, as long as the function $\xi_n'(r) - r$ is not flat. (This corresponds to no "significant atoms" appearing at a scale below some r_0, and thus, for $r < r_0$, F_r is just a scaled down version of F_{r_0}; if $\xi_n'(r) - r$ is flat, the two bounds will be of the same order of magnitude.)

Indeed, by Lemma 1, since $\xi_n'(r)/r$ is non-increasing,

$$\inf\{r : \xi_n'(r) \le r\} \le \inf\left\{r : \xi_n'(r) \le \frac{r}{4}\right\}.$$

Clearly, $\xi_n'(r) \ge 0$, since $\xi_n'(r) \ge \mathbb{E}(\mathbb{E}f - \mathbb{E}_n f) = 0$ for any fixed function, and thus $0 \le s^* \le \inf\{r : \xi_n'(r) \le r\} \le r^*$. The same argument shows that if $\xi_n'(r) - r$ is not "flat" then $s^* \ll r$. Now, for $\beta = 1$, $\epsilon \sim \sqrt{\frac{s^*}{n}} \ll s^*$ and $r_{\epsilon,+}, r_{\epsilon,-}$ will be of the order of s^*.

5 Discussion

Now, we will give an example which shows that, for any given sample size n, we can construct a function class and a probability measure such that the bound on the empirical minimizer differs significantly when using r^* from Section 3 versus s^* from Section 4.

We first prove the existence of two types of function classes, which are both bounded and Bernstein.

Lemma 3. *For every positive integer n and all $m \ge 2(n^2 + n)$, the following holds. If P is the uniform probability measure on $\{1, ..., m\}$, then for every $\frac{1}{n} \le \lambda \le 1/2$ there exists a function class G_λ such that*

1. *For every $g \in G_\lambda$, $-1 \le g(x) \le 1$, $\mathbb{E}g = \lambda$ and $\mathbb{E}g^2 \le 2\mathbb{E}g$.*
2. *For every set $\tau \subset \{1, ..., m\}$ with $|\tau| \le n$, there is some $g \in G_\lambda$ such that for every $i \in \tau$, $g(i) = -1$.*

Also, there exists a function class H_λ such that

1. *For every $h \in H_\lambda$, $0 \le h(x) \le 1$, $\mathbb{E}h = \lambda$.*
2. *For every set $\tau \subset \{1, ..., m\}$ with $|\tau| \le n$, there is some $h \in H_\lambda$ such that for every $i \in \tau$, $h(i) = 0$.*

Proof: The proof is constructive. Let $J \subset \{1, ..., m\}$, $|J| = n$; for every $I \subset J$ define $g = g_{I,J}$ in the following manner. For $i \in I$, set $g(i) = 1$, if $i \in J \backslash I$, set $g(i) = -1$, and for $i \notin J$ put $g(i) = t$, where

$$t = \frac{\lambda m + |J\backslash I| - |I|}{m - n}.$$

Observe that if $m \ge n^2 + 2n$, then $0 < t \le 2\lambda \le 1$ for every I, J. By the definition of t, $\mathbb{E}g_{I,J} = \lambda$, and

$$\begin{aligned}\mathbb{E}g^2 &= \frac{1}{m}\left(|I| - |J\backslash I| + t^2(m-n) + 2|J\backslash I|\right) \le \mathbb{E}g + \frac{2|J\backslash I|}{m} \\ &\le \mathbb{E}g + 2\frac{n}{m} < \mathbb{E}g + \frac{1}{n} \le 2\mathbb{E}g,\end{aligned}$$

where the last inequality holds because $\mathbb{E}g = \lambda \geq 1/n$, and $m \geq 2n^2$.

The second property of G_λ is clear by the construction, and the claims regarding H_λ can be verified using a similar argument. ■

Given a sample size n, we can choose a large enough m and the uniform probability measure P on $\{1, \ldots, m\}$, and define the function class $F = \text{star}(\tilde{F}, 0)$, where $\tilde{F} = H_{1/4} \cup G_{1/n}$ from Lemma 3. F is star-shaped and (1,2) Bernstein.

Theorem 8. *If $0 < \delta < 1$ and $n > N_0(\delta)$, then for any corresponding $F = star(\tilde{F}, 0)$ as above, the following holds:*

1. *For every $X_1, ..., X_n$ there is a function $f \in F$ with $\mathbb{E}f = 1/4$ and $\mathbb{E}_n f = 0$.*
2. *For the class F, the function ξ'_n satisfies*

$$\xi'_n(r) = \begin{cases} (n+1)r & \text{if } 0 < r \leq 1/n, \\ r & \text{if } 1/n < r \leq 1/4, \\ 0 & \text{if } r > 1/4. \end{cases}$$

   Thus, $\inf\{r > 0 : \xi'_n(r) \leq r/4\} = 1/4$.
3. *If $\hat{f}$ is a ρ-approximate empirical minimizer, where $0 < \rho < 1/8$, then with probability larger than $1 - \delta$,*

$$\frac{1}{n}\left(1 - c\sqrt{\frac{\log n}{n}} - \rho\right) \leq \mathbb{E}\hat{f} \leq \frac{1}{n}.$$

The proof can be found in [4].

References

1. P.L. Bartlett, S. Boucheron, G. Lugosi: Model selection and error estimation. Machine Learning 48, 85-113, 2002.
2. P.L. Bartlett, O. Bousquet, S. Mendelson: Local Rademacher Complexities. Submitted, 2002 (available at http://www.stat.berkeley.edu/∼bartlett/publications/recent-pubs.html).
3. P.L. Bartlett, M.I. Jordan, J.D. McAuliffe: Convexity, classification, and risk bounds. Tech. Rep. 638, Dept. of Stat., U.C. Berkeley, 2003.
4. P.L. Bartlett, S. Mendelson: Empirical minimization. Submitted, 2003 (available at http://axiom.anu.edu.au/∼shahar).
5. S. Boucheron, G. Lugosi, P. Massart: Concentration inequalities using the entropy method. Ann. of Prob. 31, 1583-1614, 2003.
6. O. Bousquet: Concentration Inequalities and Empirical Processes Theory Applied to the Analysis of Learning Algorithms. PhD. Thesis, 2002.
7. T. Klein: Une inégalité de concentration gauche pour les processus empiriques. C. R. Math. Acad. Sci. Paris 334(6), 501-504, 2002.
8. V. Koltchinskii, Rademacher penalties and structural risk minimization. IEEE Trans. on Info. Th. 47(5), 1902-1914, 2001.
9. V. Koltchinskii: Local Rademacher Complexities and Oracle Inequalities in Risk Minimization. Tech. Rep., Univ. of New Mexico, August 2003.

10. V. Koltchinskii and D. Panchenko, Rademacher processes and bounding the risk of function learning. In E. Gine and D. Mason and J. Wellner (Eds.), *High Dimensional Probability II*, 443-459, 2000.
11. M. Ledoux: *The concentration of measure phenomenon*. Mathematical Surveys and Monographs, Vol 89, AMS, 2001.
12. W.S. Lee, P.L. Bartlett, R.C. Williamson: The Importance of Convexity in Learning with Squared Loss. IEEE Trans. on Info Th., 44(5), 1974-1980, 1998.
13. G. Lugosi and M. Wegkamp, Complexity regularization via localized random penalties. Ann. of Stat., to appear, 2003.
14. P. Massart: About the constants in Talagrand's concentration inequality for empirical processes. Ann. of Prob., 28(2), 863-884, 2000.
15. P. Massart. Some applications of concentration inequalities to statistics. Ann. de la Faculté des Sciences de Toulouse, IX: 245-303, 2000.
16. S. Mendelson, Improving the sample complexity using global data. IEEE Trans. on Info. Th. 48(7), 1977-1991, 2002.
17. S. Mendelson: A few notes on Statistical Learning Theory. In *Proc. of the Machine Learning Summer School, Canberra 2002, S. Mendelson and A. J. Smola (Eds.)*, LNCS 2600, Springer, 2003.
18. S. Mendelson, Rademacher averages and phase transitions in Glivenko-Cantelli classes. IEEE Transactions on Information Theory 48(1), 251-263, 2002.
19. E. Rio: Inégalités de concentration pour les processus empiriques de classes de parties. Probab. Theory Related Fields 119(2), 163-175, 2001.
20. M. Talagrand: New concentration inequalities in product spaces. Invent. Math., 126, 505-563, 1996.
21. M. Talagrand: Sharper bounds for Gaussian and empirical processes. Ann. of Prob., 22(1), 28-76, 1994.