Optimal Sample-Based Estimates of the Expectation of the Empirical Minimizer

Peter L. Bartlett
Computer Science Division and Department of Statistics
University of California, Berkeley
367 Evans Hall #3860, Berkeley, CA 94720-3860
bartlett@cs.berkeley.edu

Shahar Mendelson
Centre for Mathematics and its Applications (CMA)
The Australian National University
Canberra, 0200 Australia
shahar.mendelson@anu.edu.au

Petra Philips
Friedrich Miescher Laboratory
of the Max Planck Society
Tübingen, 72076 Germany
petra.philips@tuebingen.mpg.de

October 12, 2005

Abstract

We study sample-based estimates of the expectation of the function produced by the empirical minimization algorithm. We investigate the extent to which one can estimate the rate of convergence of the empirical minimizer in a data dependent manner. We establish three main results. First, we provide an algorithm that upper bounds the expectation of the empirical minimizer in a completely data-dependent manner. This bound is based on a structural result in [3], which relates expectations to sample averages. Second, we show that these structural

AMS 2000 subject classifications. Primary: 62G08; secondary: 68Q32
Key words and phrases: error bounds; empirical minimization; data-dependent complexity
This work was supported in part by National Science Foundation grant 0434383.

upper bounds can be loose, compared to other bounds given in [3]. In particular, we demonstrate a class for which the expectation of the empirical minimizer decreases as $O(1/n)$ for sample size n, although the upper bound based on structural properties is $\Omega(1)$. Third, we show that this looseness of the bound is inevitable: we present an example that shows that a sharp bound cannot be universally recovered from empirical data.

1 Introduction

The empirical minimization algorithm is a statistical procedure that, out of a fixed class of functions, chooses a function that minimizes an empirical loss functional on this class. Known as an M-estimator in the statistical literature, it has been studied extensively [27, 25, 9]. Here, we investigate the limitations of estimates of the expectation of the function produced by the empirical minimization algorithm.

To be more exact, let F be a class of real-valued functions defined on a probability space (Ω, μ) and set $X_1, ..., X_n$ to be independent random variables distributed according to μ. For $f \in F$ define $\mathbb{E}_n f = \frac{1}{n}\sum_{i=1}^n f(X_i)$ and let $\mathbb{E}f$ be the expectation of f with respect to μ. The goal is to find a function that minimizes $\mathbb{E}f$ over F, where the only information available about the unknown distribution μ is through the finite sample $X_1, ..., X_n$. The empirical minimization algorithm produces the function $\hat{f} \in F$ that has the smallest empirical mean, that is, $\hat{f}$ satisfies

$$\mathbb{E}_n \hat{f} = \min \{\mathbb{E}_n f : f \in F\} .$$

Throughout this paper, we assume that such a minimum exists (the modifications required if this is not the case are obvious), that F satisfies some minor measurability conditions, which we omit (see [6] for more details), and that for every $f \in F$, $\mathbb{E}f \geq 0$, which, as we explain later, is a natural assumption in the cases that interest us.

In statistical learning theory, this problem arises when one minimizes the empirical risk, or sample average of a loss incurred on a finite training sample. There, the aim is to ensure that the risk, or expected loss, is small. Thus, $f(X_i)$ represents the loss incurred on X_i. Performance guarantees are typically obtained through high probability bounds on the conditional expectation

$$\mathbb{E}\hat{f} = \mathbb{E}(\hat{f}(X)|X_1, \ldots, X_n). \tag{1.1}$$

In particular, one is interested in obtaining fast and accurate estimates of the rates of convergence of this expectation to 0 as a function of the sample size n.

Classical estimates of this expectation rely on the uniform convergence over F of sample averages to expectations (see, for example, [27]). These estimates are essentially based on the analysis of the supremum of the empirical process $\sup_{f\in F}(\mathbb{E}f - \mathbb{E}_n f)$ indexed by the whole class F. As opposed to these *global* estimates, it is possible to study local subsets of functions of F, which are balls of a given radius with respect to a chosen metric. The supremum of the empirical process indexed by these local subsets as a function of the radius of the balls is called the *modulus of continuity*. Sharper *localized* estimates for the rate of convergence of the expectation can be obtained in terms of the fixed point of the modulus of continuity of the class [24, 15, 10, 13, 1].

Recent results [3] show that one can further significantly improve the high-probability estimates for the convergence rates for empirical minimizers. These results are based on a new localized notion of complexity of subsets of F containing functions with identical expectations and are therefore dependent on the underlying unknown distribution. In this article, we investigate the extent to which one can estimate these high-probability convergence rates in a data-dependent manner, an important aspect if one wants to make these estimates practically useful.

The results in [3] establish upper and lower bounds for the expectation $\mathbb{E}\hat{f}$ using two different arguments. The first is a structural result relating the empirical (random) structure endowed on the class by the selection of the coordinates $(X_1, ..., X_n)$, and the real structure, given by the measure μ. The second is a direct analysis, which yields seemingly sharper bounds. In both cases (and under some mild structural assumptions on the class F), the bounds are given using a function that measures the "localized complexity" of subsets of F consisting of functions with a fixed expectation r, denoted here by $F_r = \{f \in F : \mathbb{E}f = r\}$. For every integer n and probability measure μ on Ω, consider the following two sequences of functions, which are measures for the complexity of the sets F_r:

$$\xi_{n,F,\mu}(r) = \mathbb{E}\sup\{|\mathbb{E}f - \mathbb{E}_n f| : f \in F_r\},$$
$$\xi'_{n,F,\mu}(r) = \mathbb{E}\sup\{\mathbb{E}f - \mathbb{E}_n f : f \in F_r\}.$$

In the following, in cases where the underlying probability measure μ and the class F are clear, we will refer to these functions as ξ_n and ξ'_n. It turns out that these two functions control the generalization ability in F_r

whenever one has a strong degree of concentration for the empirical process suprema $\sup_{f\in F_r}|\mathbb{E}f-\mathbb{E}_nf|$ and $\sup_{f\in F_r}(\mathbb{E}f-\mathbb{E}_nf)$ around their expectation. Thus, ξ_n and ξ_n' can be used to derive bounds on the performance of the empirical minimization algorithm as long as these suprema are sufficiently concentrated. Therefore, the main tool required in the proofs of the results in [3] that provide bounds using the ξ_n' and ξ_n is Talagrand's concentration inequality for empirical processes (see Theorem A.1 in the appendix).

To see how ξ_n' and ξ_n can be used to derive generalization bounds, observe that it suffices to find the "critical point" r_0 for which, with high probability, for a given $0<\lambda<1$, every $r\geq r_0$ and every $f\in F_r$, $(1-\lambda)\mathbb{E}f\leq\mathbb{E}_nf\leq(1+\lambda)\mathbb{E}f$. In particular, for such an r_0, it follows that with high probability, every $f\in F$ satisfies that

$$\mathbb{E}f\leq\max\left\{\frac{\mathbb{E}_nf}{1-\lambda},r_0\right\}, \tag{1.2}$$

and thus, an upper bound on the expectation of the empirical minimizer $\hat{f}$ can be established. It is possible to show that one can take r_0 as r_n^*, where

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

and in fact, since in (1.2) only a "one-sided" condition is required, one can actually use

$$r_n'^*=\inf\left\{r:\,\xi_n'(r)\leq r/4\right\}.$$

A more careful analysis, which uses the strength of Talagrand's concentration inequality for empirical processes, shows that the expectation of the empirical minimizer is governed by approximations of

$$s_n^*=\sup\left\{r:\,\xi_n'(r)-r=\max_s\left\{\xi_n'(s)-s\right\}\right\}.$$

To see why s_n^* is a likely candidate, note that for any empirical minimizer, the function of r defined as $\sup_{f\in F_r}(\mathbb{E}f-\mathbb{E}_nf)-r=-\inf_{f\in F_r}\mathbb{E}_nf$ is maximized for the value $r=\mathbb{E}\hat{f}$. Assume that one has a very strong concentration of empirical processes indexed by F_r around their mean for every $r>0$, that is, with high probability, for every $r>0$,

$$\sup_{f\in F_r}(\mathbb{E}f-\mathbb{E}_nf)\;\approx\;\mathbb{E}\sup_{f\in F_r}(\mathbb{E}f-\mathbb{E}_nf)\;=\;\xi_n'(r).$$

Then, it would make sense to expect that, with high probability, $\mathbb{E}\hat{f}\approx s_n^*$ for $s_n^*=\operatorname{argmax}\{\xi_n'(r)-r\}$.

More precisely, and to overcome the fact that $\mathbb{E}\sup_{f\in F_r}(\mathbb{E}f-\mathbb{E}_n f)$ is only "very close" to $\sup_{f\in F_r}(\mathbb{E}f-\mathbb{E}_n f)$ define for $\varepsilon>0$,

$$r_{n,\varepsilon,+}=\sup\left\{r:\ \xi'_{n,F,\mu}(r)-r\geq\sup_s\left(\xi'_{n,F,\mu}(s)-s\right)-\varepsilon\right\}, \tag{1.3}$$

$$r_{n,\varepsilon,-}=\inf\left\{r:\ \xi'_{n,F,\mu}(r)-r\geq\sup_s\left(\xi'_{n,F,\mu}(s)-s\right)-\varepsilon\right\}. \tag{1.4}$$

Note that $r_{n,\varepsilon,+}$ and $r_{n,\varepsilon,-}$ are respectively upper and lower approximations of s_n^* that become better as $\varepsilon\longrightarrow 0$. They are close to s_n^* if the function $\xi'_n(r)-r$ is peaked around its maximum. Under mild structural assumptions on F, $\mathbb{E}\hat{f}$ can be upper bounded by either $r_n'^*$ or $r_{n,\varepsilon,+}$, and lower bounded by $r_{n,\varepsilon,-}$ for a choice of $\varepsilon=O(\sqrt{\log n/n})$ (see the exact statement in Theorem 2.5 below). Thus, these two parameters—the fixed point of $4\xi'_n$ (denoted by $r_n'^*$) and the points at which the maximum of $\xi'_n(r)-r$ is almost attained—are our main focus.

The first result we present here is that there is a true gap between $r_n'^*$ and s_n^*, which implies that there is a true difference between the bound that could be obtained using the structural approach (i.e. $r_n'^*$) and the true expectation of the empirical minimizer. We construct a class of functions satisfying the required structural assumptions for which for any n, $r_n'^*$ is of the order of a constant (and thus r_n^* is of the order of a constant), but the subsets F_r are very rich when r is close to 0 and s_n^* and $r_{n,\varepsilon,+}$ are of the order of $1/n$. There is a related result in [3], that for every n there is a function class F_n for which this phenomenon occurs. The result here is stronger, since it shows that, for some function class and probability distribution, the true convergence rate is far from the structural bound. The idea behind the construction is based on the one presented in [3], namely that one has complete freedom to choose the expectation of a function, while forcing it to have certain values on a given sample. For the class we construct and any large sample size n, estimates for the convergence rates of the empirical minimizers based on $r_n'^*$ are asymptotically not optimal (as they are $\Theta(1)$ whereas the true convergence rate is $O(1/n)$), and thus the structural bound does not capture the true behavior of the empirical minimizer.

The second question we tackle concerns the estimation of the expectation of the empirical minimizer from data. To that end, in Section 4, we present an efficient algorithm that enables one to estimate r_n^* in a completely data dependent manner. Then, in Section 5, we show that this type of data-dependent estimate is the best one can hope to have if one only has access to the function values on finite samples. We show that in such a case it is

impossible to establish a data dependent upper bound on the expectation of the empirical minimizer that is asymptotically better than r_n^*. The general idea is to construct two classes of functions that look identical when projected on any sample of finite size, but for one class both a typical expectation of the empirical minimizer and r_n^* are of the order of an absolute constant, while for the other a typical expectation is of the order of $1/n$.

2 Definitions and Preliminary Results

2.1 Loss Classes

One of the main applications of our investigations is the analysis of prediction problems, like classification or regression, arising in machine learning. Suppose that one is presented with a sequence of observation-outcome pairs $(x, y) \in \mathcal{X} \times \mathcal{Y}$, and the aim is to choose a function $g : \mathcal{X} \to \mathcal{Y}$ that accurately predicts the outcome given the observation. We assume that $(X, Y), (X_1, Y_1), \ldots, (X_n, Y_n)$ are chosen independently from a probability distribution P on $\mathcal{X} \times \mathcal{Y}$, but P is unknown. The difference between the true outcome and the prediction is measured using a *loss function*, $\ell : \mathcal{Y}^2 \to [0, 1]$, where $\ell(\hat{y}, y)$ represents the cost incurred by predicting $\hat{y}$ when the true outcome is y. The risk of a function $g : \mathcal{X} \to \mathcal{Y}$ is defined as $\mathbb{E}\ell(g(X), Y)$, and the aim is to use the sequence $(X_1, Y_1), \ldots, (X_n, Y_n)$ to choose a function g with minimal risk. Setting $f(x, y) = \ell(g(x), y)$, this task corresponds to minimizing $\mathbb{E}f$. In empirical risk minimization, one chooses g from a set G to minimize the sample average of $\ell(g(x), y)$, which corresponds to choosing $f \in F$ to minimize $\mathbb{E}_n f$, where F is the *loss class*,

$$F = \{(x, y) \mapsto \ell(g(x), y) : g \in G\}.$$

It is sometimes convenient to consider *excess loss functions*,

$$f(x, y) = \ell(g(x), y) - \ell(g^*(x), y),$$

where $g^* \in G$ satisfies $\mathbb{E}\ell(g^*(X), Y) = \inf_{g \in G} \mathbb{E}\ell(g(X), Y)$. Since g^* is fixed, choosing $g \in G$ to minimize risk (respectively, empirical risk) again corresponds to choosing $f \in F$ to minimize $\mathbb{E}f$ (respectively, $\mathbb{E}_n f$), where

$$F = \{(x, y) \mapsto \ell(g(x), y) - \ell(g^*(x), y) : g \in G\}.$$

Thus, for this choice of F, $\mathbb{E}f \geq 0$ for all $f \in F$, but functions in F can have negative values.

2.2 Structural Assumptions

Throughout this article, we assume that F is a class of functions, bounded by b, with nonnegative expectations, that contains 0. These assumptions are justified when one considers the loss or excess loss class F of a function class G containing a minimizer and corresponding to a bounded loss function.

Additionally, we make two mild structural assumptions about the class F, namely, that F is star-shaped around 0 and that F satisfies a Bernstein condition.

Definition 2.1 *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.

Thus, for Bernstein classes of functions, the second moment of every function is bounded by a power of its expectation, uniformly over the class. The significance of this condition is that it implies a better degree of concentration for individual functions, because functions that satisfy a Bernstein condition can not be "too peaky", and thus, random sampling yields a more accurate representation of the expectation.

A Bernstein condition is satisfied by a large variety of loss classes arising in a natural statistical setting. For example, it is satisfied for classes of nonnegative functions that are uniformly bounded, with $\beta = 1$. As was shown in [12, 17, 2], it is also satisfied for excess loss classes associated with learning problems where the hypothesis class is a convex class of functions bounded by 1, and the loss function is a power-type function. In particular, for the squared-error loss function, one can take $\beta = 1$. In addition, it is satisfied for "low noise" classification problems as defined in [23] (see also [16]), in which the conditional expectation of the label given an input x, $p(x) = \mathbb{E}[Y|X = x]$ is, with high probability, not "too close" to $1/2$ (where X denotes the input and Y the label).

Definition 2.2 *F is called star-shaped around 0 if for every $f \in F$ and $0 \leq \alpha \leq 1$, $\alpha f \in F$.*

Observe that if F is an excess loss class, then any empirical minimizer in F is also an empirical minimizer in $\mathrm{star}(F, 0) = \{\alpha f : f \in F, 0 \leq \alpha \leq 1\}$.

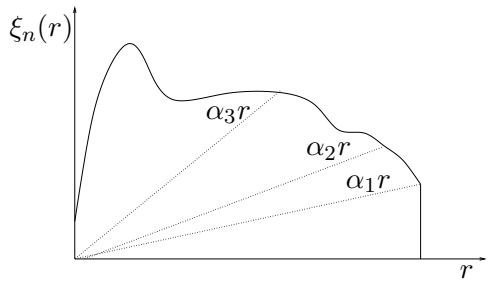

Figure 1: The graph of a function ξ_n that is "sub-linear" (cf. Lemma 2.3).

Hence, one can replace F with $\mathrm{star}(F, 0)$ without changing the empirical minimizer. Moreover, since $\mathbb{E}f$ and $\mathbb{E}_n f$ are linear functionals in f, the "localized complexity" of $\mathrm{star}(F, 0)$ is not considerably larger than that of F (for instance, in the sense of covering numbers). The advantage in considering star-shaped classes is that it adds some regularity to the class. For example, it is easy to see that for star-shaped classes the functions $\xi_n(r)/r$ and $\xi_n'(r)/r$ are non-increasing. Figure 1 illustrates the graph of a typical function with this "sub-linear" property, which is stated formally in the following lemma:

Lemma 2.3 *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 $0 < r' \leq r$, $\xi_n(r') \geq \alpha r'$. Analogous assertions hold for ξ_n'.

In other words, for every r, the graph of ξ_n in the interval $[0, r]$ is above the line connecting $(r, \xi_n(r))$ and $(0, 0)$. The proof of Lemma 2.3 is easy (e.g. a proof for ξ_n can be found in [3]) and is omitted.

As an example, Figure 2 illustrates the graph of a function ξ_n for the star-shaped hull of a class that contains only functions with expectations either equal to r_1 or to r_2.

2.3 Preliminary Results

If F is star-shaped around 0 one can derive the following estimates for the empirical minimizer. (Recall that $r_n^* = \inf\{r : \xi_n(r) \leq r/4\}$ and $r_n'^* = \inf\{r : \xi_n'(r) \leq r/4\}$.)

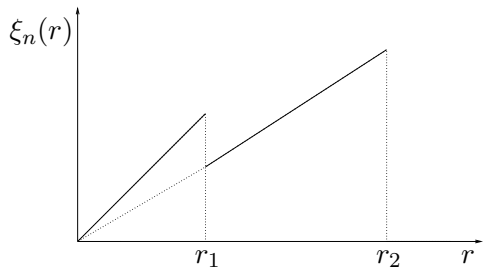

Figure 2: An example of a graph of a function ξ_n for the class $\mathrm{star}(F, 0)$, where F contains only functions with expectations r_1 and r_2.

Theorem 2.4 *[3] Let F be a (β, B)-Bernstein class of functions bounded by b that is star-shaped around 0. Then there is an absolute constant c such that with probability at least $1-e^{-x}$, any empirical minimizer $\hat{f} \in F$ satisfies*

$$\mathbb{E}\hat{f} \leq \max\left\{ r_n^*, \frac{cbx}{n}, c\left(\frac{Bx}{n}\right)^{1/(2-\beta)} \right\}.$$

Also, with probability at least $1 - e^{-x}$, any empirical minimizer $\hat{f} \in F$ satisfies

$$\mathbb{E}\hat{f} \leq \max\left\{ r_n'^*, \frac{cbx}{n}, c\left(\frac{Bx}{n}\right)^{1/(2-\beta)} \right\}.$$

Thus, with high probability, r_n^* is an upper bound for $\mathbb{E}\hat{f}$, as long as $r_n^* \geq c/n^{1/(2-\beta)}$, and the same holds for $r_n'^*$. Note that $r_n'^*$ can be much smaller than r_n^*, and so the convergence rates obtained through $r_n'^*$ are potentially better.

For $\beta = 1$, the estimates based on $r_n'^*$ and r_n^* are at best $1/n$, and in general at best $1/n^{1/(2-\beta)}$. Thus, the degree of control of the variance through the expectation, as measured by the Bernstein condition, influences the best rate of convergence one can obtain in terms of $r_n'^*$ and r_n^* using this method whenever one requires a confidence that is exponentially close to 1. In particular, this approach recovers the better learning rates for convex function classes from [12] and for low noise classification from [23, 16], as both convexity of F for squared-loss and low noise conditions imply that the loss class is Bernstein.

It turns out that this structural bound can be improved using a direct analysis of the empirical minimization process. Indeed, the next theorem, whose proof can be found in [3], shows that one can directly bound $\mathbb{E}\hat{f}$ for

the empirical minimizer without trying to relate the empirical and actual structures of F. It states that $\mathbb{E}\hat{f}$ is concentrated around s_n^* and therefore, with high probability, $\mathbb{E}\hat{f} \le r_{n,\varepsilon,+}$, where ε can be taken smaller than $c\sqrt{\log n/n}$. (The definition of $r_{n,\varepsilon,+}$ was given in the introduction.) In addition, if the class is not too "rich" around 0, then with high probability, $\mathbb{E}\hat{f} \ge r_{n,\varepsilon,-}$.

Theorem 2.5 *[3] 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 of functions bounded by b that is star-shaped around 0. For every n and $\varepsilon > 0$ define $r_{n,\varepsilon,+}$, and $r_{n,\varepsilon,-}$ as above, fix $x > 0$ and set*

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

If

$$\varepsilon \ge c\left(\max\left\{ \sup_s \left(\xi'_{n,F,\mu}(s) - s\right), r'^{\,\beta}_n \right\} \frac{(B+b)(x+\log n)}{n} \right)^{1/2},$$

then

1. *With probability at least $1 - e^{-x}$,*

$$\mathbb{E}\hat{f} \le \max\left\{ \frac{1}{n}, r_{n,\varepsilon,+} \right\}.$$

2. *If*

$$\mathbb{E}\sup\left\{ \mathbb{E}f - \mathbb{E}_n f : f \in F, \mathbb{E}f \le c_1/n \right\} < \sup_s \left(\xi'_{n,F,\mu}(s) - s\right) - \varepsilon,$$

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

$$\mathbb{E}\hat{f} \ge r_{n,\varepsilon,-}\,.$$

To compare this result to the previous one, note that $s_n^* \le r'^*_n$. Indeed, $\xi'_n(r) \ge \mathbb{E}(\mathbb{E}f - \mathbb{E}_n f) = 0$ for any fixed function f, and thus $\xi'_n(0) \ge 0$, $\xi'_n(s_n^*) \ge s_n^*$ and $0 \le s_n^* \le \inf\{r : \xi'_n(r) \le r\} \le r'^*_n$ (where the last inequality holds since $\xi'_n(r)/r$ is non-increasing, by Lemma 2.3). It follows that if $\xi'_n(r) - r$ is not flat around s_n^*, then the bound resulting from Theorem 2.5 improves the structural bound of Theorem 2.4. Figure 3 illustrates graphically such a case.

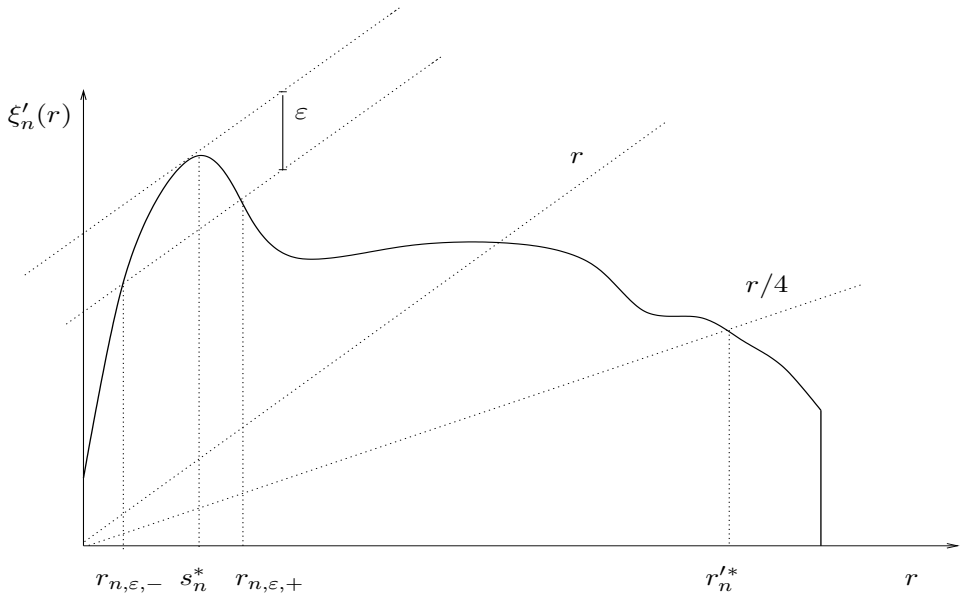

Figure 3: The graph of a function ξ'_n, and the corresponding values for r'^*_n, s^*_n, $r_{n,\varepsilon,+}$, and $r_{n,\varepsilon,-}$. If $s^*_n \ll r'^*_n$ and $\xi'_n(r) - r$ is peaked around s^*_n, then $r_{n,\varepsilon,+}$ is smaller than r'^*_n.

3 A true gap between the expectation of the empirical minimizer and r^*_n

In this section, we construct a class of functions for which there is a clear gap between the structural result of Theorem 2.4 and the expectation of the empirical minimizer, as estimated in Theorem 2.5. The idea behind this construction (as well as in the other construction we present later) is that one has complete freedom to choose the expectation of a function, while forcing it to have certain values on a given sample.

Let us start with an outline of the construction. It is based on the idea (developed in [3]) of two Bernstein classes of functions satisfying the following for any fixed n. The functions are defined on a finite set $\{1, \ldots, m\}$ with respect to the uniform probability measure, where m depends on n. The first class contains all functions that vanish on a set of cardinality n, but have expectations equal to a given constant. The second class consists of functions that each take their minimal values on a set of cardinality n, but have expectations equal to $1/n$. By appropriately choosing the values of the functions, one can show that the star-shaped hull of the union of these two

classes has $r_n'^* \sim c$, whereas $s_n^* \sim r_{n,\varepsilon,+} \sim 1/n$. Thus, the estimate given by Theorem 2.5 is considerably better than the one resulting from Theorem 2.4 for that fixed value of n. To make this example uniform over n, we construct similar sets on $(0,1]$, take the star-shaped hull of the union of all such sets and show that $\xi'_{n,F,\mu}(r) - r$ still achieves its maximum at $1/n$ and decays rapidly for $r > 1/n$, ensuring that $r_{n,\varepsilon,+} \ll r_n'^*$.

The first step in the construction is the following lemma, which states that, for any given n and for $1/n \leq \lambda \leq 1/2$, one can find function classes G_λ^n and H_λ^n defined on $(0,1]$ that are both uniformly bounded and Bernstein with respect to the Lebesgue measure on $(0,1]$, and for which $\xi'_{n,H_\lambda^n,\mu}(\lambda) = \lambda$, $\xi'_{n,G_\lambda^n,\mu}(\lambda) = \lambda + 1$.

Lemma 3.1 *Let μ be the Lebesgue measure on $(0,1]$. Then, for every positive integer n and any $\frac{1}{n} \leq \lambda \leq 1/2$ there exists a function class G_λ^n such that*

1. *For every $g \in G_\lambda^n$, $-1 \leq g(x) \leq 1$, $\mathbb{E}g = \lambda$ and $\mathbb{E}g^2 \leq 2\mathbb{E}g$.*
2. *For every set $\tau \subset (0,1]$ with $|\tau| \leq n$, there is some $g \in G_\lambda^n$ such that for every $s \in \tau$, $g(s) = -1$.*

Also, there exists a function class H_λ^n such that

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

Proof. Let $m = 2(n^2 + n)$. Consider functions that are constant on the intervals $((i-1)/m, i/m]$, $1 \leq i \leq m$, and set G_λ^n to be the function class containing all functions taking the value -1 on exactly n such intervals; that is, each function in G_λ^n is defined as follows: Let $J \subset \{1, ..., m\}$, $|J| = n$ and set

$$g_J(x) = \begin{cases} -1, \text{ if } x \in (\frac{j-1}{m}, \frac{j}{m}] \text{ and } j \in J, \\ t_\lambda, \text{ otherwise,} \end{cases}$$

where

$$t_\lambda = \frac{\lambda m + n}{m - n} = \frac{2\lambda(n^2 + n) + n}{2n^2 + n}. \tag{3.1}$$

Since $0 \leq \lambda \leq 1/2$, $0 \leq t_\lambda \leq 1$ and thus $g_J : (0,1] \to [-1,1]$. It is easy to verify that all the functions in G^n_λ have expectation λ with respect to μ and that G^n_λ is (1,2)-Bernstein, since for any $g \in G^n_\lambda$,

$$\mathbb{E}g^2 = \frac{1}{m}\left(n + t_\lambda^2(m-n)\right) \leq \frac{1}{m}\left(n + t_\lambda(m-n)\right) = \lambda + \frac{1}{n+1} \leq 2\lambda = 2\mathbb{E}g.$$

The construction of H^n_λ is similar, and its functions take the values $\{0, t'_\lambda\}$ for $t'_\lambda = \lambda m/(m-n)$. ■

Using the notation of the lemma, define the following function classes:

$$H = \bigcup_{i=5}^{\infty} H^i_{1/4}, \qquad F_k = G^k_{1/k}, \qquad G = \bigcup_{i=5}^{\infty} F_i,$$

and

$$F = \mathrm{star}(G \cup H, 0). \tag{3.2}$$

Since F is star-shaped around 0 and is a (1,2)-Bernstein class, it satisfies the assumptions of Theorem 2.4 and Theorem 2.5. Also, note that for any $n \geq 5$ and any $X_1, ..., X_n$ there is some $f \in F$ with $\mathbb{E}f = 1/4$ and $\mathbb{E}_n f = 0$, and some $g \in F$ with $\mathbb{E}g = 1/n$ and $\mathbb{E}_n g = -1$. Indeed, f can be taken from $H^n_{1/4}$ and g from $F_n = G^n_{1/n}$.

The following theorem shows that for the class F, for any integer n, $r'^*_n = 1/4$, while the empirical minimizer is likely to be smaller than $r_{n,\varepsilon,+} \sim c/n$.

Theorem 3.2 *For F defined by (3.2), the following holds:*

1. *For every $n \geq 5$,*

$$\xi'_{n,F,\mu}(r) = \begin{cases} r + rk & \text{if } r \in (1/(k+1), 1/k]\,, \text{ where } k \geq n \\ r & \text{if } r \in (1/5, 1/4] \\ 0 & \text{if } r > 1/4, \end{cases}$$

   and in particular, $r'^*_n = 1/4$.

2. *There exists a constant $c > 1$, such that the following holds: for every $\varepsilon < 3/4$, every $n \geq N(\varepsilon)$, and every $k \leq n/c$,*

$$\xi'_{n,F,\mu}(1/k) - 1/k \leq \xi'_{n,F,\mu}(1/n) - 1/n - \varepsilon.$$

   In particular, $r_{n,\varepsilon,+} \leq c/n$.

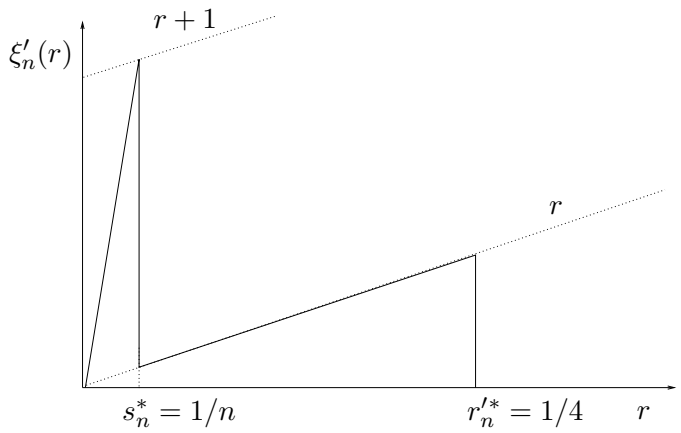

Figure 4: $\xi'_{n,\mathrm{star}(F_n \cup H^n_{1/4}),\mu}$ (as in the proof of Theorem 3.2).

By the properties of F mentioned above, for every sample of cardinality $n \geq 5$, the graph of $\xi'_{n,F,\mu}$ for the class $\mathrm{star}(F_n \cup H^n_{1/4}, 0)$ is as in Figure 4, with $r'^*_n = 1/4$ and $s^*_n = 1/n$. The main part of the proof is to show that for F, which is the star-shaped hull of the union of all these sets, $\xi'_{n,F,\mu}(r) - r$ still achieves its maximum at $1/n$ and decays rapidly for $r > 1/n$, ensuring that $r_{n,\varepsilon,+} \ll r'^*_n$. Figure 5 illustrates the qualitative behavior of ξ'_n.

Proof of Theorem 3.2. The first part of the proof is immediate from the definition of the function ξ'_n and thus omitted. Turning to the second, and more difficult part, note that indeed $r'^*_n = 1/4$ and that the maximal value of $\xi'_{n,F,\mu}(r) - r$ is attained at $r = 1/n$. In order to estimate the value $\xi'_{n,F,\mu}(1/k)$ for $k < n$, consider $\sup_{f \in F_k}(\mathbb{E}f - \mathbb{E}_n f)$ for a fixed $X_1, \ldots, X_n$. Let $m = 2(k^2 + k)$ and note that by the construction of F_k, each $g \in F_k$ is of the form g_J for some set $J \subset \{1, ..., m\}$, $|J| = k$. For each set J let A_J be the union of the intervals $\left(\frac{j-1}{m}, \frac{j}{m}\right]$ where $j \in J$, and let Φ be the following set of indicator functions

$$\Phi = \{\mathbb{1}_{A_J} : J \subset \{1, ..., m\}, |J| = k\}.$$

Clearly, for every $\phi \in \Phi$, $\mathbb{E}\phi = k/m$ and $vc(\Phi) \leq k$, since no set of $k+1$ distinct points in $(0, 1]$ can be shattered by Φ (actually, $vc(\Phi) = k$ since the set $\{1/k, 1/(k-1), \ldots, 1\}$ is shattered by Φ). Recall that if Φ is a class of binary-valued functions and if the VC-dimension $vc(\Phi) \leq k$, then as a special case of Theorem A.5, the Rademacher averages (see page 26 for the

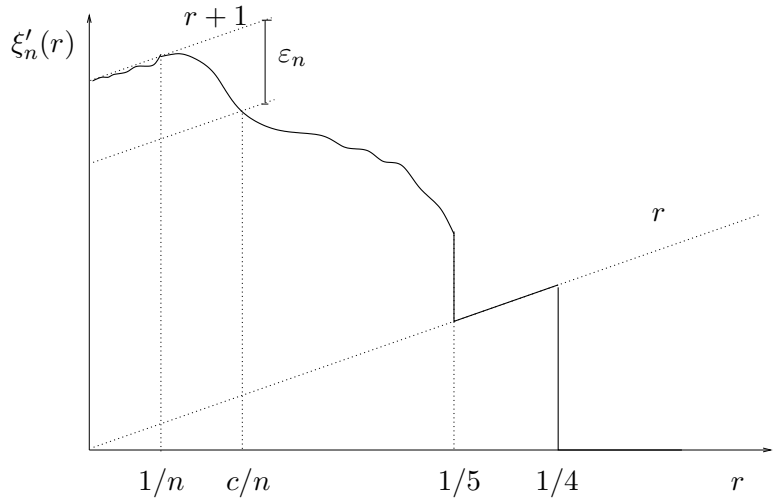

Figure 5: Qualitative behavior of $\xi'_{n,F,\mu}$.

definition) can be bounded by

$$\mathbb{E}R_n(\Phi) \le c_2\sqrt{k/n} \tag{3.3}$$

for some absolute constant c_2.

Define the random variable $\ell_J = \sum_{i=1}^n \mathbb{1}_{A_J}(X_i)$. Thus, ℓ_J is the cardinality of the set $\{i : g_J(X_i) = -1\}$. Note that

$$\mathbb{E}_n g_J = \frac{-2\ell_J(k+1)^2 + 3kn + 2n}{kn(2k+1)},$$

and therefore,

$$\sup_{f\in F_k}(\mathbb{E}f - \mathbb{E}_n f) = \frac{1}{k} + \frac{2(k+1)^2\sup_J \ell_J - 3kn - 2n}{kn(2k+1)}\,.$$

From Talagrand's concentration inequality (Theorem A.1) applied to the set of functions Φ, there exist absolute constants c_1, c_2 such that for any $0 < t \le 1$, with probability larger than $1 - e^{-c_1 nt^2}$,

$$\sup_{f\in\Phi}\sum_{i=1}^n f(X_i) \le \frac{kn}{m} + 2nR_n(\Phi) + 2nt \le \frac{kn}{m} + 2c_2\sqrt{kn} + 2nt\,,$$

where the last inequality holds by (3.3).

Setting t=1/20, and since $kn/m = n/(2(k+1)) < n/10$ for any $k \geq 5$, it is evident that there exists an absolute constant $c > 1$ such that for any $k \leq n/c$, with probability at least $1 - e^{-c_1' n}$, $\sup_J \ell_J \leq n/5 + 2c_2\sqrt{kn} \leq n/4$.

Therefore, applying the union bound for $5 \leq k' \leq k$, it follows that with probability at least $1 - ne^{-c'n}$,

$$\sup_{f \in \cup_{k'=5}^{k} \frac{k'}{k} F_{k'}} (\mathbb{E}f - \mathbb{E}_n f) \leq \frac{(k+1)^2/2 - 3k - 2}{k(2k+1)} \leq \frac{1}{k} + \frac{1}{4}$$

for every $k \leq n/c_1$.

Observe that scaled-down versions of functions from H do not contribute to $\xi'_{n,F,\mu}(1/k)$ and thus, one only has to take care of elements in F with expectation of $1/k$ that come either from F_k or are scaled down versions of $F_{k'}$ for $k' \leq k$. Hence,

$$\begin{aligned}
\xi'_{n,F,\mu}(1/k) &= \mathbb{E} \sup_{f \in \cup_{k'=5}^{k} \frac{k'}{k} F_{k'}} (\mathbb{E}f - \mathbb{E}_n f) \\
&\leq \left(\frac{1}{k} + \frac{1}{4}\right)(1 - ne^{-c'n}) + ne^{-c'n}\left(\frac{1}{k} + 1\right) \\
&= \frac{1}{k} + \frac{1}{4} + \frac{3}{4} ne^{-c'n}.
\end{aligned}$$

Thus, for $\varepsilon < 3/4$, if n is sufficiently large that $3n/4e^{-c'n} \leq 3/4 - \varepsilon$, we have

$$\xi'_{n,F,\mu}(1/k) - 1/k \leq 1 - \varepsilon = \xi'_{n,F,\mu}(1/n) - 1/n - \varepsilon,$$

provided that $k \leq n/c$. ■

To conclude, there exists a true gap between the bound that can be obtained via the structural result (the fixed point $r_n'^*$ of the localized empirical process) and the true expectation of the empirical minimizer as captured by s_n^*.

Corollary 3.3 *For F defined in (3.2), there is an absolute constant $c > 0$ for which the following holds: For any $x > 0$ there is an integer $N(x)$ such that for any $n \geq N(x)$,*

1. *With probability at least $1 - e^{-x}$, $\mathbb{E}\hat{f} \leq c/n \sim s_n^*$.*
2. *$r_n'^* = r_n^* = 1/4$.*

4 Estimating r_n^* from data

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

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

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

A way of estimating r_n^* is to find an empirically computable function $\hat{\xi}_n(r)$ that is, with high probability, an upper bound for the function $\xi_n(r)$ and therefore, its fixed point $\hat{r}_n^* = \inf\{r : \hat{\xi}_n(r) \le \frac{r}{4}\}$ is an upper bound for r_n^*. We shall construct $\hat{\xi}_n$ for which $\hat{\xi}_n(r)/r$ is non-increasing and thus $\hat{r}_n^*$ would be determined using a binary search algorithm. To that end, we require the following result, which states that, for Bernstein classes, there is a phase transition in the behavior of coordinate projections around the point where $\xi_n(r) \sim r$. Above this point, the local subsets $F_r = \{f \in F : \mathbb{E}f = r\}$ are small and the expectation and empirical means are close in a multiplicative sense. Below this point, the sets F_r are too rich to allow this.

Theorem 4.1 *[3] There is an absolute constant c for which the following holds. Let F be a class of functions, such that for every $f \in F$, $\|f\|_\infty \le b$. Assume that F is a (β, B)-Bernstein class. Suppose that $r \ge 0$, $0 < \lambda < 1$, and $0 < \alpha < 1$ satisfy*

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

1. *If $\xi_n(r) \ge (1+\alpha)r\lambda$, then with probability at least $1 - e^{-x}$,*

$$\sup_{f \in F_r} |\mathbb{E}f - \mathbb{E}_n f| \ge \lambda \mathbb{E}f\,.$$

2. *If $\xi_n(r) \le (1-\alpha)r\lambda$, then with probability at least $1 - e^{-x}$,*

$$\sup_{f \in F_r} |\mathbb{E}f - \mathbb{E}_n f| \le \lambda \mathbb{E}f\,.$$

3. *If $\xi_n'(r) \ge (1+\alpha)r\lambda$, then with probability at least $1 - e^{-x}$,*

$$\sup_{f \in F_r} (\mathbb{E}f - \mathbb{E}_n f) \ge \lambda \mathbb{E}f.$$

4. If $\xi_n'(r) \le (1-\alpha)r\lambda$, then with probability at least $1-e^{-x}$,

$$\sup_{f\in F_r}(\mathbb{E}f - \mathbb{E}_n f) \le \lambda \mathbb{E}f.$$

We will make use of the following direct corollary of Theorem 4.1 applied to the case $\alpha = 1/2$, $\lambda = 1/2$.

Corollary 4.2 *There is an absolute constant $c > 0$ for which the following holds. If F is (β, B)-Bernstein, and*

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

and $\xi_n(r) \le \frac{r}{4}$, then with probability larger than $1-e^{-x}$, every $f \in F_r$ satisfies $r/2 \le \mathbb{E}_n f \le 3r/2$.

If we define the "empirical shell,"

$$F^n_{\frac{r}{2},\frac{3r}{2}} := \{f \in F \,:\, r/2 \le \mathbb{E}_n f \le 3r/2\},$$

the corollary shows that, for suitable large r, with high probability,

$$F_r \subseteq F^n_{\frac{r}{2},\frac{3r}{2}}\,.$$

The following theorem shows that the empirical Rademacher average of an empirical shell is with high probability an upper bound for $\xi_n(r)$ for all r larger than the fixed point r_n^*.

Theorem 4.3 *There are absolute constants c, c_1, c_2, and c_3 for which the following holds. Let F be a star-shaped (β, B)-Bernstein class and $\sup_{f\in F}\|f\|_\infty \le b$. For*

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

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

$$\xi_n(r) \le 8\mathbb{E}_\sigma R_n\left(F^n_{c_1 r, c_2 r}\right) + c_3 r$$

for every $r \in [\tilde{r}_n', b]$.

Proof. Since F is star-shaped, then by Lemma 2.3, $\xi_n(r) \leq \frac{r}{4}$ if and only if $r \geq r_n^*$, and thus, by Corollary 4.2 (for appropriately chosen c), if $r \geq \tilde{r}'_n$, then with probability larger than $1 - e^{-x}$, $F_r \subseteq F^n_{\frac{r}{2},\frac{3r}{2}}$, which implies that

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

By symmetrization (Theorem A.2) and concentration of Rademacher averages around their mean (Theorem A.3), and since $r \geq \frac{cbx}{n}$, it follows that with probability at least $1 - 2e^{-x}$,

$$\xi_n(r) \leq 2\mathbb{E}R_n(F_r) \leq 4\mathbb{E}_\sigma R_n(F_r) + \frac{4bx}{n} \leq 4\mathbb{E}_\sigma R_n\left(F^n_{\frac{r}{2},\frac{3r}{2}}\right) + c_3 r\,.$$

To find an upper bound on $\xi_n(r)$ that holds with high probability uniformly for all $r \geq r_n^*$, we divide the interval $[1/n, b]$ into a set of $\lceil bn \rceil$ intervals of length at most $1/n$. (Note that the choice of the starting point $1/n$ restricts the estimates for $\tilde{r}'_n$ to values that are larger than $1/n$. The proof can be easily modified to allow estimates up to the value cbx/n, but since we are only interested in estimates at best of the order of $O(1/n)$ we made this restriction in order to keep the proof simpler.) Let

$$A = \left\{\frac{1}{n}, \frac{2}{n}, \ldots, \frac{\lceil bn \rceil}{n}\right\} \cap \left[\frac{\lfloor c_n n \rfloor}{n}, \frac{\lceil bn \rceil}{n}\right],$$

where

$$c_n = c \max\left\{\frac{bx}{n}, \left(\frac{Bx}{n}\right)^{1/(2-\beta)}\right\}.$$

Since $|A| \leq bn + 1$, the union bound shows that with probability at least $1 - 2(bn+1)e^{-x}$, $\xi_n(r) \leq 4\mathbb{E}_\sigma R_n\left(F^n_{\frac{r}{2},\frac{3r}{2}}\right) + c_3 r$ for every $r \in A$. By Lemma 2.3, for any $1 \leq k \leq n$, if $r \in \left[\frac{k}{n}, \frac{k+1}{n}\right]$, then $\xi_n(r) \leq \xi_n\left(\frac{k}{n}\right)\frac{nr}{k}$. Thus, with probability at least $1 - 2(bn+1)e^{-x}$, every $r \in [\tilde{r}'_n, b]$ satisfies

$$\xi_n(r) \leq \xi_n\left(\frac{k}{n}\right)\frac{nr}{k} \leq \left(4\mathbb{E}_\sigma R_n\left(F^n_{\frac{k}{2n},\frac{3k}{2n}}\right) + \frac{c_3 k}{n}\right)\frac{nr}{k} \leq 8\mathbb{E}_\sigma R_n\left(F^n_{c_1 r, c_2 r}\right) + c_3 r,$$

where k satisfies $r \in [k/n, (k+1)/n]$ and c_1 and c_2 are absolute constants. ∎

Therefore, one can define

$$\hat{\xi}_n(r) = 8\mathbb{E}_\sigma R_n\left(F^n_{c_1 r, c_2 r}\right) + c_3 r.$$

Let $\hat{r}_n^* = \inf\{r : \hat{\xi}_n(r) \le \frac{r}{4}\}$. Then by Theorem 4.3 with probability at least $1-2(bn+1)e^{-x}$, $\hat{r}_n^* \ge r_n^*$. Moreover, since $\hat{\xi}_n(r)/r$ is non-increasing, $r \ge \hat{r}_n^*$ if and only if $\hat{\xi}_n(r) \le \frac{r}{4}$.

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

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

Set $r_L = \max\{1/n, c_n\}$, $r_R = b$.

If $\hat{\xi}_n(r_R) \le r_R/4$ then

for $\ell = 0$ 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, $\bar{r} - \frac{1}{n} \le \hat{r}_n^* \le \bar{r}$. Hence, for every n, with probability larger than $1 - 2(bn+1)e^{-x}$, $r_n^* \le \bar{r}$.

Theorem 4.4 *There exists an absolute constant c for which the following holds. Let F be a (β, B)-Bernstein class of functions bounded by b that is star-shaped around 0. For every integer n, any $x > 0$, and any sample X of size n, with probability at least $1 - (2bn+3)e^{-x}$, $\mathbb{E}\hat{f} \le RSTAR(F, X)$.*

Note that RSTAR(F, X) is essentially the fixed point of the function $r \mapsto \mathbb{E}_\sigma R_n\left(F^n_{c_1 r, c_2 r}\right)$. This function measures the complexity of the function class $F^n_{c_1 r, c_2 r}$, which can be determined empirically by looking at empirical means that fall in an interval whose length is proportional to r. The main difference between that and the data-dependent estimates in [1] is that instead of taking the whole empirical ball as done in [1], here we only measure the complexity of an empirical "shell" around r. However, if the function class is not "regular" around the critical value of r, the complexity of the shell $F(c_1 r, c_2 r)$ might be very different from the complexity of F_r, in which case one would like to make c_1 and c_2 very close to 1.

Indeed, one can tighten this bound further by narrowing the size of the shell and replacing the empirical set $F^n_{\frac{r}{2}, \frac{3r}{2}}$ with $F^n_{(1-\varepsilon_n)r, (1+\varepsilon_n)r}$. This is done by selecting the isomorphism constant in Theorem 4.1 to depend on n and tend to 1 as $n \to \infty$.

Theorem 4.5 *Let F be a star-shaped (β, B)-Bernstein class and $\sup_{f\in F}\|f\|_\infty \le b$. There is an absolute constant c, for which the following holds. If $0 < \varepsilon_n < 1$ and*

$$\tilde{r}_n = \max\left\{ r_n^*, \frac{1}{n}, \frac{cbx}{n\varepsilon_n}, c\left(\frac{Bx}{n\varepsilon_n^2}\right)^{1/(2-\beta)} \right\},$$

then with probability at least $1 - 2(bn+1)e^{-x}$

$$\xi_n(r) \le 4\mathbb{E}_\sigma R_n\left(F^n_{(1-\varepsilon_n)r,(1+\varepsilon_n)r}\right) + \frac{\varepsilon_n r}{c}$$

for every $r \in [\tilde{r}_n, b]$.

Proof. With the same reasoning as before, by Theorem 4.1 for $\alpha = 1/2$ and $\lambda = \varepsilon_n$, if $r \ge \tilde{r}_n$ then with probability larger than $1 - e^{-x}$, $F_r \subset F^n_{(1-\varepsilon_n)r,(1+\varepsilon_n)r}$. We define

$$\hat{\xi}_n(r) = \left(4\mathbb{E}_\sigma R_n\left(F^n_{(1-\varepsilon_n)r,(1+\varepsilon_n)r}\right) + \frac{k\varepsilon_n}{cn}\right)\frac{nr}{k}, \text{ for } r \in \left[\frac{k}{n}, \frac{k+1}{n}\right].$$

Again, with probability at least $1 - 2(bn+1)e^{-x}$, for every $r \in [\tilde{r}_n, b]$, $\xi(r) \le \hat{\xi}_n(r)$. ∎

Since $\hat{\xi}_n(r)/r$ is non-increasing, it is possible to define

$$\hat{r}^* = \inf\left\{ r : \hat{\xi}_n(r) \le \frac{r\varepsilon_n}{2} \right\}$$

with a slight modification of RSTAR (we replace the test in the if-clause, $\hat{\xi}_n(r) > r/4$, with $\hat{\xi}_n(r) > r\varepsilon_n/2$) . It follows that for every n and every sample of size n, with probability larger than $1 - 2bne^{-x}$, $r_n^* \le \bar{r}$, where $\bar{r}$ is generated by the modified algorithm. For example, one can choose $\varepsilon_n = 1/\log n$, which has the advantage that the empirical shells $\hat{F}_{r-\frac{r}{\log n}, r+\frac{r}{\log n}}$ become, with growing sample size, closer to F_r. The price we pay for the advantage is an extra $\log n$ factor in the final estimate, since in this case the estimate of the expectation goes down at the rate of $O(\log n/n)$.

Remark 4.6 Note that a lower bound of a similar nature has to take into account the complexity of the class $F_{0,cr}$. This might happen because one may not have an inclusion $F_r \subseteq F^n_{c_1 r, c_2 r}$ unless $c_1 = 0$. Indeed, if the class F is very rich for r close to 0, it is possible to have functions that have a very small expectation, but for which $\mathbb{E}_n f \sim r$.

5 The limitations of estimating from data

Although the results in [3] show that it is possible to bound the expectation of the empirical minimizer in a far sharper way than by applying a structural result, it was not clear whether such a bound could be estimated from data. In the following we consider a scenario in which one only has access to the function class through the values that class members take on finite samples, that is, the finite dimensional coordinate projections of the class. In this case, we construct an example that shows that, in general, it is impossible to establish a data-dependent estimate of s_n^* that is better than r_n^*. To be precise, we construct two function classes that have identical coordinate projections on every sample. For one class we have $r_n'^* \sim c$, $s_n^* \sim c$ and the expectation of the empirical minimizer is of the order of c with probability 1, while for the other class, $s_n^* \sim 1/n$. If one only has access to the way the classes behave on finite dimensional coordinate projections, that is, samples, the classes are indistinguishable, and it is impossible to predict a better bound than an absolute constant, which could be much worse than the true behavior of the empirical minimizer.

Recall that for a given function class F and a sample $\tau = \{x_1, \ldots, x_n\}$, the coordinate projection of F on τ is

$$P_\tau F = \{(f(x_1), ..., f(x_n)) : f \in F\}.$$

Let μ be the Lebesgue measure on $(0, 1]$. For each $k \in \mathbb{N}$ we construct two function classes F_1^k and F_2^k, both $(1, c)$-Bernstein with respect to μ for a suitable absolute constant c, and take values in $V = \{-1, 0, 1\}$.

In both classes we construct, each function is a constant on the intervals $((j-1)/m_k, j/m_k]$, where $m_k = k^2 + 3k$. The class F_1^k consists of all functions that take the value -1 on k intervals, the value 1 on $2k$ intervals and the value 0 on k^2 intervals. It is easy to verify that for any $f \in F_1^k$, $\mathbb{E}f = k/(k^2 + 3k) \sim 1/k$ and $\mathbb{E}f^2 = 3k/(k^2 + 3k) \sim 1/k$, implying that indeed F_1^k is a $(1, 3)$-Bernstein class.

In contrast, F_2^k consists of all functions that take the value -1 on k intervals, the value 1 on $k^2 + k$ intervals and 0 on k intervals. Therefore, for any function $f \in F_2^k$, $\mathbb{E}f = k^2/(k^2 + 3k) \geq 1/4$ and since $\mathbb{E}f^2 \leq 1$, F_2^k is a $(1, 4)$-Bernstein class. Notice that, whereas functions in F_1^k have expectations of the order of $1/k$, functions in F_2^k have expectations of the order of a constant.

Set

$$F_1 = \text{star}\left(\bigcup_{k\in\mathbb{N}} F_1^k, 0\right), \qquad F_2 = \text{star}\left(\bigcup_{k\in\mathbb{N}} F_2^k, 0\right),$$

and it is easy to verify that for every finite set τ, $P_\tau F_1 = P_\tau F_2$. Indeed, consider a set $\tau = \{x_1, ..., x_n\}$. Without loss of generality, assume that $x_i \neq x_j$ if $i \neq j$. Let ℓ be large enough to ensure that the x_is fall in disjoint intervals $((j-1)/m_\ell, j/m_\ell]$ and that $\ell \geq n$, and thus, $P_\tau F_2^\ell = P_\tau F_1^k = \{-1, 0, 1\}^n$. Note that this actually shows that for every τ with distinct values, $P_\tau F_1 = P_\tau F_2 = [-1, 1]^n$.

Therefore, F_1 and F_2 are star-shaped, Bernstein classes that have identical coordinate projections. Therefore, it is impossible to distinguish between the two, based solely on empirical data. On the other hand, the behavior of the empirical minimizer is very different in the two cases.

Theorem 5.1 *For F_1 and F_2 defined as above, there is an absolute constant $c > 0$ for which the following holds. For any $x > 0$ there is some $N(x)$ such that for any $n \geq N(x)$,*

1. *For F_1, with probability at least $1 - e^{-x}$, $\mathbb{E}\hat{f} \leq c/n \sim s_n^*(F_1)$.*
2. *For F_2, with probability 1, $\mathbb{E}\hat{f} \geq 1/4 \sim r_n^*(F_2)$.*

Theorem 5.1 implies that the estimates for the convergence rate of the empirical minimization algorithm based on s_n^* are significantly better for the class F_1 than for F_2. However, the classes have identical coordinate projections on any sample, and hence are indistinguishable empirically. Thus, one can not get an *empirical* estimate of the convergence rate for F_1 that is significantly better than one based on an empirical estimate of r_n^*.

Proof of Theorem 5.1. We will show that the expectation of the empirical minimizer in F_1 is likely to be smaller than c/n, as opposed to F_2 where it is likely to be of the order of a constant.

For any n, $\inf_{f \in F_1^n} \mathbb{E}_n f = -1$, and therefore $\xi'_{n,F_1,\mu}(s_n) - s_n = 1$, where, for any k and any $f \in F_1^k$,

$$s_k = \mathbb{E}f = \frac{k}{k^2 + 3k} \sim \frac{1}{k}.$$

Clearly, for a class of functions bounded by 1, $\xi'_{n,F,\mu}(r) - r \leq 1$, and thus the maximal value of $\xi'_{n,F_1,\mu}(r) - r$ is attained at $s_n \sim 1/n$. The main part of the proof is to show that there is some absolute constant $c > 1$ such that for large enough values of n and for $r \geq c/n$, $\xi'_{n,F_1,\mu}(r) - r \leq 1/2$. This is the case because the sets F_1^k are not "rich" enough when projected onto samples of size n as long as $k \leq n/c$.

Indeed, the function class F_1^n has low complexity, in terms of the combinatorial dimension $vc(F_1^n, \varepsilon)$ (see Definition A.4). In particular, the definitions imply that $vc(F_1^k, \varepsilon) \leq 2k$ for all $0 < \varepsilon \leq 2$ and all k. Since the class of functions is bounded by 1, Theorem A.5 implies there is an absolute constant c_2 such that $\mathbb{E}R_n(F_1^k) \leq c_2\sqrt{k/n}$. Applying the one sided version of Talagrand's concentration inequality for the empirical process $Z = \sup_{f \in F_1^k}(\mathbb{E}f - \mathbb{E}_n f)$, it follows that for $t = 1/4$, with probability at least $1 - e^{-c_1 n t^2} = 1 - e^{-c_1' n}$,

$$\sup_{f \in F_1^k}(\mathbb{E}f - \mathbb{E}_n f) \leq 2\mathbb{E}R_n(F_1^k) + t \leq 2c_2\sqrt{\frac{k}{n}} + t \leq \frac{1}{2},$$

provided that $k \leq n/c$ for some universal constant c. Let

$$A_k = \bigcup_{k' \leq k} \frac{s_k}{s_{k'}} F_1^{k'},$$

that is, A_k contains the functions in F_1 that have expectations s_k—those either come from F_1^k or are "scaled down" versions of functions from $F_{k'}$ for $k' < k$. Therefore, for any $k \leq n/c$, with probability at least $1 - ne^{-c_1' n}$

$$\sup_{f \in A_k}(\mathbb{E}f - \mathbb{E}_n f) \leq \frac{1}{2}.$$

Taking the expectation,

$$\xi'_{n,F_1,\mu}(s_k) \leq (1 - ne^{-c_1' n})\frac{1}{2} + (1 + s_k)\, ne^{-c_1' n} = \frac{1}{2} + \left(\frac{1}{2} + s_k\right) ne^{-c_1' n},$$

and thus, for all $\varepsilon < 1/2$, $n \geq N(\varepsilon)$ and $k \leq n/c$,

$$\xi'_{n,F,\mu}(s_k) - s_k \leq 1 - \varepsilon - s_k = \xi'_{n,F,\mu}(s_n) - s_n - \varepsilon_n - s_k.$$

This implies that $\xi'_{n,F,\mu}(r) - r \leq \xi'_{n,F,\mu}(s_n) - s_n - \varepsilon_n$ for every $r \geq c'/n$, from which we conclude that $r_{n,\varepsilon,+} \leq c'/n$.

On the other hand, it is easy to verify that for empirical minimization over F_2, $\mathbb{E}\hat{f} \geq 1/4$. Indeed, as we saw for F_1, $\inf_{f \in F_2^n} \mathbb{E}_n f = -1$, which implies $\mathbb{E}_n \hat{f} = -1$. Since we can write $F_2 = \cup\{\alpha f : f \in F_2^k : k \in \mathbb{N}, \alpha \in [0,1]\}$, and empirical minimization is a linear operation, it is clear that the empirical minimum will be attained at $\alpha = 1$. But all functions in $\cup_{k \in \mathbb{N}} F_2^k$ have expectation greater than $1/4$, and so with probability 1, $\mathbb{E}\hat{f} \geq 1/4$ in this case. ∎

Remark 5.2 Note that if one is given the function $\hat{f}$ that the algorithm produced, rather than just the coordinate projections, it becomes possible to distinguish if the class at hand is F_1 or F_2. However, we can define an uncountable collection of function classes

$$\mathbb{F} = \left\{ \text{star}\left(\bigcup_{k\in\mathbb{N}} F^k_{\alpha_k}, 0 \right) \ : \ \alpha_k \in \{1,2\} \text{ for } k \in \mathbb{N} \right\},$$

where if $\alpha_k = 1$ then $F^k_{\alpha_k} = F^k_1$ and if $\alpha_k = 2$ then $F_{\alpha_k} = F^k_2$. Clearly, for every $H, G \in \mathbb{F}$ and every finite $\sigma \subset \Omega$, $P_\sigma(G) = P_\sigma(H)$. If the learner knows that $F \in \mathbb{F}$ and even if $\hat{f}$ is given to him, then the best thing that could be said is that a single "component" of F, say the jth component of F, is F^j_1 or F^j_2. It is impossible to say whether other components of F are of "type 1" or "type 2" and in particular, the convergence rate for the expectation of the empirical minimizer can be as bad as for F_2.

The second remark is that the class F_1 is not a Glivenko-Cantelli class. The classes F^k_1 become richer as k grows - i.e., in the part of F_1 in which the expectation of functions is smaller. The reason we can obtain a generalization bound even for classes that are not Glivenko-Cantelli is because the method of [3] uses the expectation of the empirical process indexed by $\{f \in F : \mathbb{E}f = r\}$, and each one of these sets is a Glivenko-Cantelli class. If one were to try and bound the error of the empirical minimizer using the localization $\{f \in F : \mathbb{E}f \leq r\}$ as in [1], it would be impossible.

A Additional material

The main technical tool we require is Talagrand's celebrated concentration theorem for empirical processes [21, 11]. The version we use is due to Bousquet [5] (see also [14, 19, 8]).

Theorem A.1 *Let F be a class of functions defined on $\mathcal{X}$ and let P 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 $\sigma^2 = n \sup_{f\in F} \text{var} f$. Define*

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

For every $x > 0$ and every $\rho > 0$,

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

and the same inequalities hold for $\bar{Z}$. Here, K is an absolute constant.

The rest of this section is devoted to some results that allow one to estimate $\mathbb{E}\sup_{f\in F}|\mathbb{E}f - \mathbb{E}_n f|$ via the Rademacher process indexed by the class. Define

$$R_n f = \frac{1}{n}\sum_{i=1}^{n}\sigma_i f(X_i) \quad \text{and} \quad R_n(F) = \sup_{f\in F} R_n f,$$

where $\sigma_1, \ldots, \sigma_n$ denote independent Rademacher random variables, that is, symmetric, $\{-1, 1\}$-valued random variables. 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 $X_1, ..., X_n$,

$$\mathbb{E}_\sigma R_n(F) = \mathbb{E}\left(R_n(F)|\, X_1, \ldots, X_n\right).$$

A well known symmetrization argument (due to Giné and Zinn) connects the expectation of $\sup_{f\in F}|\mathbb{E}f - \mathbb{E}_n f|$ to the Rademacher averages of F [26].

Theorem A.2 *Let F be a class of functions defined on (Ω, μ) and let $X_1, ..., X_n$ be independent random variables distributed according to μ. Then,*

$$\mathbb{E}\sup_{f\in F}|\mathbb{E}f - \mathbb{E}_n f| \leq 2\mathbb{E}R_n(F).$$

The next lemma, which follows directly from a self-bounding property of the Rademacher process and the methods developed in [4], shows that $\mathbb{E}_\sigma R_n(F)$ is highly concentrated around its expectation; hence, the Rademacher averages of a class can be upper bounded by their empirical version. The following formulation can be found in [1].

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

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

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

$$\frac{1}{2}\mathbb{E}_\sigma R_n(F) - \frac{cbx}{n} \leq \mathbb{E}R_n(F)$$

where c *is an absolute constant.*

It is possible to bound $\mathbb{E}R_n(F)$ using the combinatorial dimension of a set. Recall that a set $\{x_1, ..., x_n\}$ is shattered by a class of $\{0,1\}$-valued functions F if

$$P_\sigma F = \{(f(x_1), ..., f(x_n)) : f \in F\} = \{0,1\}^n,$$

and that the Vapnik-Chervonenkis dimension d of F denoted by $vc(F)$ is the maximal cardinality of a subset of Ω that is shattered by F. In a similar way, one can define the combinatorial dimension of a class of real-valued functions.

Definition A.4 *For every* $\varepsilon > 0$, *a set* $\sigma = \{x_1, ..., x_n\} \subset \Omega$ *is said to be* ε*-shattered by* F *if there is some function* $s : \sigma \to \mathbb{R}$, *such that for every* $I \subset \{1, ..., n\}$ *there is some* $f_I \in F$ *for which* $f_I(x_i) \geq s(x_i) + \varepsilon$ *if* $i \in I$, *and* $f_I(x_i) \leq s(x_i) - \varepsilon$ *if* $i \notin I$. *Let*

$$\mathrm{vc}(F, \varepsilon) = \sup\{|\sigma| \mid \sigma \subset \Omega,\ \sigma \text{ is } \varepsilon\text{−shattered by } F\}.$$

The following result is a recent extension, due to Rudelson and Vershynin [20] to well-known estimates on $\mathbb{E}R_n(F)$.

Theorem A.5 *There exists an absolute constant* c *for which the following holds. For any class* F *and any probability measure* μ *on* Ω,

$$\mathbb{E}R_n(F) \leq c \int_0^\infty \sqrt{vc(F, \varepsilon)} d\varepsilon.$$

References

[1] P.L. Bartlett, O. Bousquet, S. Mendelson. Local Rademacher Complexities. *The Annals of Statistics*, 33(4), 1497-1537, 2005.

[2] P.L. Bartlett, M.I. Jordan, J.D. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*. To appear.

[3] P.L. Bartlett, S. Mendelson. Empirical minimization. *Probability Theory and Related Fields*. To appear (available at http://www.stat.berkeley.edu/∼bartlett/papers/bm-em-03.pdf).

[4] S. Boucheron, G. Lugosi, P. Massart. Concentration inequalities using the entropy method. *The Annals of Probability* 31, 1583-1614, 2003.

[5] O. Bousquet. Concentration Inequalities and Empirical Processes Theory Applied to the Analysis of Learning Algorithms. PhD thesis, École Polytechnique, 2002.

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

[7] D. Haussler. Sphere Packing Numbers for Subsets of the Boolean n-cube with Bounded Vapnik-Chervonenkis Dimension. Journal of Combinatorial Theory (A), 69(2), 217-232, 1995.

[8] T. Klein. Une inégalité de concentration gauche pour les processus empiriques. C. R. Math. Acad. Sci. Paris 334(6), 501-504, 2002.

[9] V. Koltchinskii. Local Rademacher Complexities and Oracle Inequalities in Risk Minimization. Technical report, University of New Mexico, August 2003.

[10] V. Koltchinskii, D. Panchenko. Rademacher processes and bounding the risk of function learning. High Dimensional Probability, vol. II, 443-459, 2000.

[11] M. Ledoux. The Concentration of Measure Phenomenon. Volume 89 of *Mathematical Surveys and Monographs*. American Mathematical Society, 2001.

[12] W.S. Lee, P.L. Bartlett, R.C. Williamson. The Importance of Convexity in Learning with Squared Loss. *IEEE Transactions on Information Theory*, 44(5), 1974-1980, 1998.

[13] G. Lugosi, M. Wegkamp. Complexity regularization via localized random penalties. *The Annals of Statistics*, 32(4), 1679-1697, 2004.

[14] P. Massart. About the constants in Talagrand's concentration inequality for empirical processes. *The Annals of Probability*, 28(2), 863-884, 2000.

[15] P. Massart. Some applications of concentration inequalities to statistics. *Annales de la Faculté des Sciences de Toulouse*, IX: 245–303, 2000.

[16] P. Massart, E. Nédélec. Risk bounds for statistical learning. Preprint (available at http://www.math.u-psud.fr/∼massart/page5.html).

[17] S. Mendelson. Improving the sample complexity using global data. *IEEE Transactions on Information Theory* 48(7), 1977-1991, 2002.

[18] 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.

[19] E. Rio. Inégalités de concentration pour les processus empiriques de classes de parties. *Probability Theory and Related Fields*, 119(2), 163-175, 2001.

[20] M. Rudelson, R. Vershynin. Combinatorics of random processes and sections of convex bodies. *Annals of Mathematics*, to appear.

[21] M. Talagrand. Sharper Bounds for Gaussian and Empirical Processes. *The Annals of Probability*, 22, 20-76, 1994.

[22] M. Talagrand. New concentration inequalities in product spaces. *Inventiones Mathematicae*, 126, 505-563, 1996.

[23] A. Tsybakov. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32(1), 135-166, 2004.

[24] S.A. van de Geer. A new approach to least squares estimation, with applications. *The Annals of Statistics*, 15(2), 587-602, 1987.

[25] S.A. van de Geer. *Empirical Processes in M-Estimation*, Cambridge University Press, 2000.

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

[27] V.N. Vapnik and A.Y. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2), 264-280, 1971.