ESAIM: PS October 2010, Vol. 14, p. 315–337
DOI: 10.1051/ps:2008036

ESAIM: Probability and Statistics
www.esaim-ps.org

ON THE OPTIMALITY OF SAMPLE-BASED ESTIMATES OF THE EXPECTATION OF THE EMPIRICAL MINIMIZER [*], [**]

Peter L. Bartlett[1], Shahar Mendelson[2, 3] and Petra Philips[4]

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 due to Bartlett and Mendelson, which relates expectations to sample averages. Second, we show that these structural upper bounds can be loose, compared to previous bounds. 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.

Mathematics Subject Classification. 62G08, 68Q32.

Received July 12, 2007. Revised September 26, 2008.

1. Introduction

The empirical minimization algorithm is a statistical procedure that chooses a function that minimizes an empirical loss functional on a given class of functions. Known as an M-estimator in statistical literature, it has been studied extensively [11,29,31]. 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 $X_1, ..., X_n$.

Keywords and phrases. Error bounds; empirical minimization; data-dependent complexity.

[*] *This work was supported in part by National Science Foundation Grant 0434383.*

[**] *This work was supported in part by the Australian Research Council Discovery Grant DP0559465.*

[1] Computer Science Division and Department of Statistics, 367 Evans Hall #3860, University of California, Berkeley, CA 94720-3860, USA; bartlett@cs.berkeley.edu

[2] Centre for Mathematics and its Applications (CMA), The Australian National University Canberra, ACT 0200, Australia.

[3] Department of Mathematics, Technion I.I.T., Haifa 32000, Israel.

[4] Friedrich Miescher Laboratory of the Max Planck Society, Tübingen 72076, Germany.

Article published by EDP Sciences

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 article, 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 [8] 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

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

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, [31]). 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, for example, 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 [3,12,16,18,28].

Recent results [1] 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 [1] 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 measure the complexity of the sets F_r:

$$\xi_{n,F,\mu}(r) = \mathbb{E} \sup \{|\mathbb{E}f - \mathbb{E}_n f| : f \in F_r \} , \tag{1.2}$$

$$\xi'_{n,F,\mu}(r) = \mathbb{E} \sup \{\mathbb{E}f - \mathbb{E}_n f : f \in F_r \} . \tag{1.3}$$

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 $\sup_{f \in F_r} |\mathbb{E}f - \mathbb{E}_n f|$ and $\sup_{f \in F_r} (\mathbb{E}f - \mathbb{E}_n f)$ around their expectations. 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 [1] that provide bounds using the ξ'_n and ξ_n is Talagrand's concentration inequality for empirical processes (see Thm. 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 $f \in F_r$,

$(1-\lambda)\mathbb{E}f \leq \mathbb{E}_n f \leq (1+\lambda)\mathbb{E}f$. If the equivalence holds for a sample $(X_1, ..., X_n)$ for such an r_0, then every $f \in F$ satisfies that

$$\mathbb{E}f \leq \max\left\{\frac{\mathbb{E}_n f}{1-\lambda}, r_0\right\}, \tag{1.4}$$

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

$$r_n^* = \inf\left\{r : \xi_{n,G}(r) \leq r/4\right\}, \tag{1.5}$$

where $G = \{\theta f : 0 \leq \theta \leq 1,\ f \in F\}$. In fact, since in (1.4) only a "one-sided" condition is required, one can actually use

$$r_n'^* = \inf\left\{r : \xi'_{n,G}(r) \leq r/4\right\}. \tag{1.6}$$

For the rest of this section we will assume that F is star-shaped around 0 (that is, $G = F$), and we will explain the significance of this property later.

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

$$s_n^* = \sup\left\{r : \xi'_n(r) - r = \max_s \left\{\xi'_n(s) - s\right\}\right\}. \tag{1.7}$$

To see why s_n^* is a likely candidate, note that for any empirical minimizer, the function of r defined as $\sup_{f\in F_r}(\mathbb{E}f - \mathbb{E}_n f) - r = -\inf_{f\in F_r}\mathbb{E}_n f$ 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 $r > 0$,

$$\sup_{f\in F_r}(\mathbb{E}f - \mathbb{E}_n f) \ \approx\ \mathbb{E}\sup_{f\in F_r}(\mathbb{E}f - \mathbb{E}_n f) \ =\ \xi'_n(r).$$

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 $\varepsilon > 0$,

$$r_{n,\varepsilon,+} = \sup\left\{r : \xi'_{n,F,\mu}(r) - r \geq \sup_s\left(\xi'_{n,F,\mu}(s) - s\right) - \varepsilon\right\}, \tag{1.8}$$

$$r_{n,\varepsilon,-} = \inf\left\{r : \xi'_{n,F,\mu}(r) - r \geq \sup_s\left(\xi'_{n,F,\mu}(s) - s\right) - \varepsilon\right\}. \tag{1.9}$$

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 Thm. 2.6 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 and show that 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 $1/n$. Let us mention that there is a construction related to this one in [1]: for every n there is a function class F_n for which this phenomenon occurs. The construction we present here is stronger, since it shows that, for some function class and probability distribution, the true convergence rate for a fixed class is far from the structural bound. The idea behind the construction is based on the one presented in [1], 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 restricted to 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 select a function $g : \mathcal{X} \to \mathcal{Y}$ that makes an accurate prediction of the outcome for each observation. We assume that $(X, Y), (X_1, Y_1), ..., (X_n, Y_n)$ are chosen independently from a probability distribution P on $\mathcal{X} \times \mathcal{Y}$, but P is unknown. The quality of the prediction is measured using a bounded *loss function*, $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, b]$, where $\ell(\hat{y}, y)$ represents the cost incurred for the prediction $\hat{y}$ 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 $(X_1, Y_1), ..., (X_n, Y_n)$ 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 that minimizes the sample average of $\ell(g(x), y)$, which corresponds to choosing $f \in F$ that minimizes $\mathbb{E}_n f$, 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 $g \in G$ that minimizes the risk (respectively, empirical risk) again corresponds to choosing $f \in F$ that minimizes $\mathbb{E}f$ (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. Assumptions on F

Throughout this article, we assume that F is a class of functions defined on a probability space (Ω, μ) satisfying the following conditions:

(1) Each functions in F maps to the bounded interval $[-b, b]$.
(2) Each function in F has nonnegative expectation.
(3) F contains 0.
(4) F has Bernstein type $\beta > 0$.

We shall see shortly why these conditions are natural for many practical nonparametric and machine learning methods. The Bernstein condition, defined precisely below, is that the second moment of every function is bounded by a power of its expectation, uniformly over the class.

Definition 2.1. We say that F is a (β, B)-Bernstein class with respect to the probability measure P (where $0 < \beta \leq 2$ 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.

These conditions are satisfied by a large variety of loss classes arising in statistical settings. One simple example is the loss class,

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

in the case where some function $g^* \in G$ has zero loss, that is, $\mathbb{E}\ell(g^*(X), Y) = 0$. Clearly, if F contains 0, functions in F are bounded and have nonnegative expectations, and trivially F has Bernstein type 1:

$$\mathbb{E}f^2 \leq b\mathbb{E}f.$$

However, in practical problems, the assumption that there is some function $g^* \in G$ that has zero loss is often unreasonable.

More realistic examples are excess loss classes,

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

where g^* in G achieves the minimal risk over G. Clearly, functions in F are bounded and have nonnegative expectation, and F contains zero. As the following examples show, the boundedness and Bernstein conditions also frequently arise naturally.

Low noise classification: In two-class pattern classification, we have $\mathcal{Y} = \{\pm 1\}$, and $\ell(\hat{y}, y)$ is the 0-1 loss, that is, the indicator of $\hat{y} \neq y$. Clearly, the boundedness condition holds. A key factor in the difficulty of a pattern classification problem is the behavior of the conditional probability $\eta(x) = \Pr(Y = 1 | X = x)$, and in particular how likely it is to be near the critical value of 1/2. Starting with Tsybakov [27], many authors have considered [2,4,5,11,19,26] pattern classification when there is a constant ϵ such that the conditional probability satisfies

$$\Pr\left(\left|\eta(X) - \frac{1}{2}\right| < \epsilon\right) = 0. \tag{2.1}$$

Suppose that we assume, as in [27], that the class $\mathcal{G}$ contains the minimizer g^* of the expected loss (the *Bayes classifier*), which is the indicator of $\eta(x) > 1/2$. Then it is easy to show that this implies the excess loss class is of Bernstein type 1. Indeed, one can verify that (2.1) is equivalent to the assertion that all measurable functions $g : \mathcal{X} \to \{\pm 1\}$ satisfy

$$\Pr\left(g(X) \neq g^*(X)\right) \leq \frac{1}{2\epsilon}\mathbb{E}\left(\ell(g(X), Y) - \ell(g^*(X), Y)\right)$$

(see, for example, Lem. 5 in [4]). Therefore,

$$\mathbb{E}\left(\ell(g(X), Y) - \ell(g^*(X), Y)\right)^2 = \Pr\left(g(X) \neq g^*(X)\right) \leq \frac{1}{2\epsilon}\mathbb{E}\left(\ell(g(X), Y) - \ell(g^*(X), Y)\right).$$

Similarly, if there is a constant $\kappa \geq 0$ such that

$$\Pr\left(\left|\eta(X) - \frac{1}{2}\right| \leq \epsilon\right) \leq c\epsilon^{\kappa} \tag{2.2}$$

for some c and all $\epsilon > 0$ (see [27]), and the class $\mathcal{G}$ contains the Bayes classifier, then this implies the excess loss class is of Bernstein type $\kappa/(1+\kappa)$ (see, for example, Lem. 5 in [4]).

Boosting with an ℓ_1 constraint: Large margin classification methods, such as AdaBoost and support vector machines, minimize the sample average of a convex criterion over a class of real-valued functions. For example, Lugosi and Vayatis [15] consider empirical minimization with an exponential loss over a class of ℓ_1-constrained linear combinations of binary functions: define, for a given class $\mathcal{H}$ of $\{\pm 1\}$-valued functions, the class

$$G_\lambda = \left\{ \sum_i \alpha_i h_i : h_i \in \mathcal{H} \text{ and } \sum_i |\alpha_i| \leq \lambda \right\}.$$

Let $\ell(\hat{y}, y) = \exp(-y\hat{y})$, and consider the excess loss class

$$F_\lambda = \{(x,y) \mapsto \ell(g(x),y) - \ell(g^*(x),y) : g \in G_\lambda\},$$

where g^* is the minimizer in G_λ of the risk. Then for all probability distributions, functions in F_λ are bounded by $b = \exp(\lambda)$ and have Bernstein type 1 (see Lem. 7 and Tab. 1 in [4]).

Support vector machines with low noise: The support vector machine is a method for pattern classification that chooses a function $f : \mathcal{X} \to \mathbb{R}$ from a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ with kernel $k : \mathcal{X}^2 \to \mathbb{R}$ so as to minimize the regularized empirical risk criterion

$$\frac{1}{n}\sum_{i=1}^{n} \ell(f(x_i), y_i) + \lambda \|f\|^2,$$

where $y_i \in \{\pm 1\}$, the loss is the hinge loss,

$$\ell(\hat{y}, y) = \max\{0, 1 - \hat{y}y\}, \tag{2.3}$$

and $\|f\|$ denotes the norm in the RKHS. This is equivalent, for some r, to solving the constrained optimization problem

$$\begin{array}{rl} \min_f & \frac{1}{n}\sum_{i=1}^{n} \ell(f(x_i), y_i) \\ \text{s.t.} & f \in \mathcal{H} \\ & \|f\|^2 \leq r^2. \end{array}$$

Define $H_r = \{g \in \mathcal{H} : \|g\|^2 \leq r^2\}$ and the excess loss class

$$F_r = \{(x,y) \mapsto \ell(g(x),y) - \ell(g^*(x),y) : g \in \mathcal{H}_r\}, \tag{2.4}$$

where $g^* \in \mathcal{H}_r$ is the minimizer of the risk. Then if the kernel of the RKHS satisfies

$$k(X,X) \leq B \qquad \text{almost surely,} \tag{2.5}$$

all functions in F_r are bounded by $2Br$. Furthermore, if the probability distribution satisfies the low noise condition (2.2) and F_r contains the Bayes classifier, then Lemma 7 of [2] shows that F_r has Bernstein type $\kappa/(1+\kappa)$.

Thus, our assumptions are satisfied in this case, and the results in this article give estimates of the excess risk, that is, the difference between the expected loss and the infimum over all measurable functions of the expected loss. In fact, this also leads to an estimate of the excess risk as measured

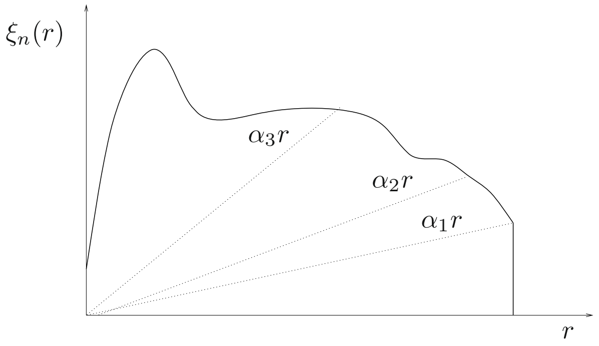

FIGURE 1. The graph of a function ξ_n that is "sub-linear" (*cf.* Lem. 2.3).

by the 0-1 loss: for all large margin classification methods, which minimize the sample average of a surrogate loss function, there is a general, optimal inequality relating the excess risk as measured by the surrogate loss to the excess risk as measured by the 0-1 loss [4].

Kernel ridge regression for classification: If, in the support vector machine, we replace the hinge loss (2.3) with the quadratic loss,

$$\ell(\hat{y}, y) = (\hat{y} - y)^2,$$

we obtain the kernel ridge regression method for pattern classification. Defining the class F_r as in (2.4), if the kernel satisfies the bound (2.5), then every function in F_r is bounded by $2Br$. Furthermore, without any constraints on the probability distribution, the uniform convexity of the loss function implies that F_r has Bernstein type 1 [14].

Kernel regression with convex loss: Similar examples can be obtained when the quadratic loss is replaced by any power loss (see [20,21]). In kernel regression also, if the response variable satisfies $|Y| \leq B$ almost surely, then the boundedness of the kernel implies boundedness of functions in the excess loss class, and uniform convexity of the loss implies that the excess loss class is Bernstein.

2.3. Star-shaped classes

We begin with the following definition:

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

We will show below that if F is an excess loss class, then any empirical minimizer in F is also an empirical minimizer in the set

$$\text{star}(F, 0) = \{\alpha f : f \in F, \, 0 \leq \alpha \leq 1\}.$$

Hence, one can replace F with star$(F, 0)$ in the analysis of the empirical minimization problem. Moreover, since $\mathbb{E}f$ and $\mathbb{E}_n f$ are linear functionals in f, the "localized complexity" of 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, and thus the analysis of the empirical minimization problem becomes simpler. 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.

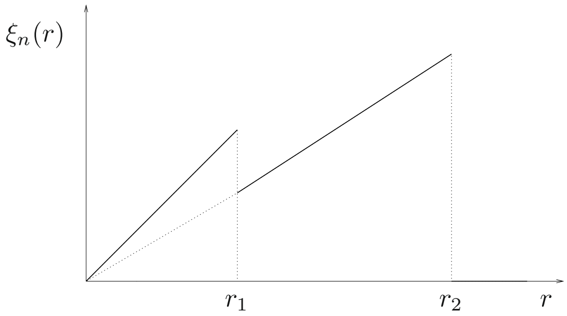

FIGURE 2. An example of a graph of a function ξ_n for the class star$(F,0)$, where F contains only functions with expectations r_1 and r_2.

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 $(0,0)$. For the sake of completeness we include the proof of Lemma 2.3, which was originally stated in [1].

Proof (of Lem. 2.3). Fix a sample $X_1, ..., X_n$ and, without loss of generality, suppose that $\sup_{f \in F_{r_2}} |\mathbb{E}f - \mathbb{E}_n f|$ is attained at f. Since F is star-shaped, then $f' = \frac{r_1}{r_2} f \in F_{r_1}$ satisfies

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

and the first part follows.

The second part follows directly from the first part by noting that

$$\xi_n(r') \geq \frac{r'}{r} \xi_n(r) \geq \frac{r'}{r} \alpha r = \alpha r'.$$

The proof for ξ_n' is analogous. □

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 that are either equal to r_1 or to r_2.

The following lemma allows one to use star$(F,0)$ in the analysis of the empirical minimization problem and obtain results regarding the empirical minimization problem over F.

Lemma 2.4. *Let F be a class of functions that contains* 0.

(1) *If F is a (β, B)-Bernstein class then* star$(F,0)$ *is also a (β, B)-Bernstein class.*

(2) *For every $x_1, ..., x_n$, set*

$$I_1 = \inf \left\{ \sum_{i=1}^n f(x_i) : f \in F \right\},$$
$$I_2 = \inf \left\{ \sum_{i=1}^n f(x_i) : f \in \text{star}(F,0) \right\}.$$

Then $I_1 = I_2$.

Moreover, for every $\varepsilon > 0$ the set $\{f \in \operatorname{star}(F,0) : \sum_{i=1}^n f(x_i) \le I_1 + \varepsilon\}$ has a nonempty intersection with F.

Note that by Lemma 2.4, if the set of ε-approximate empirical minimizers relative to $\operatorname{star}(F,0)$ is contained in some set A, then the set of ε-approximate empirical minimizers relative to F is also contained in A. In particular, consider the set $A = \{f : \gamma \le \mathbb{E}f \le \beta\}$. Thus, upper and lower estimates of the expectation of the empirical minimizer in $\operatorname{star}(F,0)$ would imply the same fact for all empirical minimizers in F.

Proof of Lemma 2.4. Every $g \in \operatorname{star}(F,0)$ is of the form $g = \alpha f$ for some $f \in F$ and $0 \le \alpha \le 1$. Since $\beta \le 2$ and F is a (β, B)-Bernstein class,

$$\mathbb{E}g^2 = \alpha^2 \mathbb{E}f^2 \le B\alpha^2(\mathbb{E}f)^\beta \le B(\mathbb{E}\alpha f)^\beta = B(\mathbb{E}g)^\beta.$$

To prove the second part, notice that $I_2 \le I_1$. Since $0 \in F$, we have $I_1 \le 0$ and thus, if $I_2 = 0$ then the claim is obvious. Therefore, assume that $I_2 < 0$ and for the sake of simplicity, assume that the infimum is attained in $g = \alpha f$ for some $f \in F$ and $0 < \alpha \le 1$. If $\alpha < 1$ then

$$I_1 \le \sum_{i=1}^n f(x_i) = \alpha^{-1}\sum_{i=1}^n g(x_i) = \alpha^{-1} I_2 < I_2,$$

which is impossible. Thus $\alpha = 1$ and $I_1 = I_2$.

The final claim of the lemma follows using a similar argument. □

Motivated by these observations, we redefine the set F_r as

$$F_r = \{f \in \operatorname{star}(F,0) : \mathbb{E}f = r\}.$$

For the remainder of the article, we use this in the definitions of the complexity parameters $\xi_{n,F,\mu}(r)$, $\xi'_{n,F,\mu}(r)$ in (1.2–1.3), and hence in the definitions of r_n^*, s_n^*, $r_{n,\varepsilon,+}$, and $r_{n,\varepsilon,-}$ in (1.5–1.9) as well.

2.4. **Preliminary results**

If F is star-shaped around 0 one can derive the following estimates for the empirical minimizer. (Recall the definition $r_n^* = \inf\{r : \xi_n(r) \le r/4\}$ and $r_n'^* = \inf\{r : \xi_n'(r) \le r/4\}$, where ξ_n and ξ_n' were defined above in (1.2) and (1.3).)

Theorem 2.5 ([1]). *Let F be a (β, B)-Bernstein class of functions bounded by b that contains 0. Then there is an absolute constant c such that with probability at least $1 - \mathrm{e}^{-x}$, any empirical minimizer $\hat{f} \in F$ satisfies*

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

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

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

Thus, with high probability, r_n^* is an upper bound for $\mathbb{E}\hat{f}$, as long as $r_n^* \ge c/n^{1/(2-\beta)}$, 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 [14] and for low noise classification from [19,27], 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 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 $c\sqrt{\log n/n}$. In addition, if the class is not too "rich" around 0, then with high probability, $\mathbb{E}\hat{f} \ge r_{n,\varepsilon,-}$. (To recall the definitions of s_n^*, $r_{n,\varepsilon,+}$, and $r_{n,\varepsilon,-}$ see (1.7)–(1.9).) The result follows immediately from the main result of [1], together with the observations above about star-shaped classes.

Theorem 2.6. *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 contains 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

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

then

(1) *With probability at least $1 - \mathrm{e}^{-x}$,*

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

(2) *If*

$$\mathbb{E}\sup\left\{ \mathbb{E}f - \mathbb{E}_n f : f \in \mathrm{star}(F,0), \mathbb{E}f \le c_1/n \right\} < \sup_s \left(\xi'_{n,F,\mu}(s) - s\right) - \varepsilon,$$

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

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

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 Lem. 2.3). It follows that if $\xi'_n(r) - r$ is not flat around s_n^*, then the bound resulting from Theorem 2.6 improves the structural bound of Theorem 2.5. Figure 3 illustrates graphically such a case.

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 2.5 and the expectation of the empirical minimizer, as estimated in Theorem 2.6. 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 [1]) of two Bernstein classes of functions satisfying the following for any fixed n. The functions are defined on a finite set $\{1, ..., m\}$

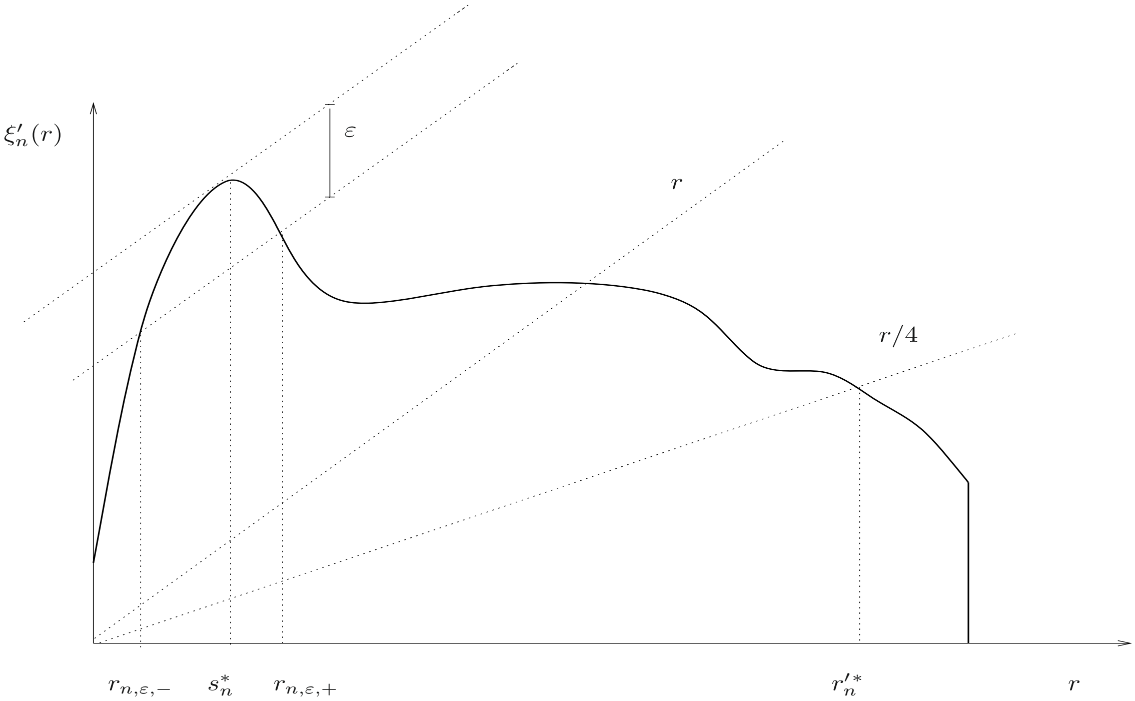

FIGURE 3. The graph of a function ξ'_n, and the corresponding values for r'^*_n, s^*_n, $r_{n,\varepsilon,+}$, and $r_{n,\varepsilon,-}$. If $s^*_n \ll r'^*_n$ and $\xi'_n(r) - r$ is peaked around s^*_n, then $r_{n,\varepsilon,+}$ is smaller than r'^*_n.

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 2.6 is considerably better than the one resulting from Theorem 2.5 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.

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

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

Also, there exists a function class H^n_λ such that

(1) *For every $h \in H^n_\lambda$, $0 \le h(x) \le 1$, $\mathbb{E}h = \lambda$, and $\mathbb{E}h^2 \le \mathbb{E}h$.*
(2) *For every set $\tau \subset (0,1]$ with $|\tau| \le n$, there is some $h \in H^n_\lambda$ 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 $((i-1)/m, i/m]$, $1 \le i \le m$, and set G^n_λ to be the function class containing all functions taking the value -1 on exactly n such intervals; that is, each function in G^n_λ 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

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

Since $0 \le \lambda \le 1/2$, $0 \le t_\lambda \le 1$ 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 $g \in G^n_\lambda$,

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

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 contains 0 and is a (1,2)-Bernstein class, it satisfies the assumptions of Theorems 2.5 and 2.6. Moreover, it is star-shaped around 0 and for any $n \ge 5$ 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 $n \ge 5$,*

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

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 $n \ge N(\varepsilon)$ and every $k \le n/c$,*

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

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

Note that by the properties of F mentioned above, for every sample of cardinality $n \ge 5$, the graph of ξ'_n for the class $F_n \cup H^n_{1/4}$ (which is the same as for the class $\mathrm{star}(F_n \cup H^n_{1/4}, 0)$) is as in Figure 4, with $r'^*_n = 1/4$ and $s^*_n = 1/n$. For the star-shaped hull of the union of all these sets, the function ξ'_n can still be described in closed-form for values of $r > 1/5$ and $r \le 1/n$, because $\sup_{f \in F_r}(\mathbb{E}f - \mathbb{E}_n f)$ is independent of the sample and is reached at a scaled-down function from H and respectively G; this is proved in part 1 of the theorem.

On the other hand, for $1/n < r < 1/5$ this supremum is no longer independent of the sample and thus we cannot provide a simple closed-form for ξ'_n. Despite that, $\xi'_n(r) - r$ still achieves its maximum at $1/n$ and decays rapidly for $r > 1/n$, ensuring that $r_{n,\varepsilon,+} \ll r'^*_n$, which is the second part of the theorem. Figure 5 illustrates the qualitative behavior of ξ'_n.

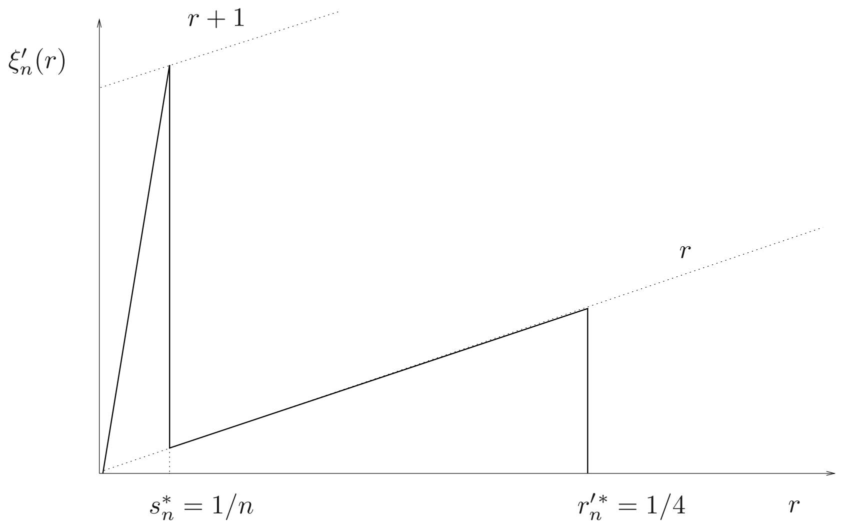

FIGURE 4. $\xi'_{n,F_n \cup H^n_{1/4},\mu}$ (as in the proof of Thm. 3.2).

Proof of Theorem 3.2. For the first part of the proof, observe that the subsets F_r consisting of functions with expectation $\mathbb{E}f = r$ are

$$F_r = \begin{cases} H_r \cup G_r & \text{if } r < 1/5 \\ H_r & \text{if } r \in (1/5, 1/4] \\ \emptyset & \text{if } r > 1/4, \end{cases}$$

where H_r and G_r are the scaled-down versions of H and G, and $G_r = \cup_{k=5}^{\lfloor 1/r \rfloor} \{krg \,:\, g \in G^k_{1/k}\}$. The first part of the theorem follows from the definition of the function ξ'_n and the fact that for any fixed sample of size n, the infimum $\inf_{f \in F_r} \mathbb{E}_n f$ is equal to 0 and reached at a scaled-down function from $H^n_{1/4}$ for $r \in (1/5, 1/4]$, and it is equal to -1 and reached at a scaled-down function from $G^k_{1/k}$ whenever $r \in (1/(k+1), 1/k]$ and $k \geq n$.

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 $\sup_{f \in G^k_{1/k}} (\mathbb{E}f - \mathbb{E}_n f)$ for a fixed $X_1, ..., X_n$. Let $m = 2(k^2 + k)$ and note that by the construction of $G^k_{1/k}$, each $g \in G^k_{1/k}$ 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 $\mathrm{vc}(\Phi) \leq k$, since no set of $k+1$ distinct points in $(0, 1]$ can be shattered by Φ (actually, $\mathrm{vc}(\Phi) = k$ since the set $\{1/k, 1/(k-1), ..., 1\}$ is shattered by Φ). Recall that if Φ is a class of binary-valued functions and if the VC-dimension $\mathrm{vc}(\Phi) \leq k$, then as a special case of Theorem A.5, the Rademacher averages (see p. 330, Eq. (4.1) for the definition) can be bounded by

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

for some absolute constant c_2.

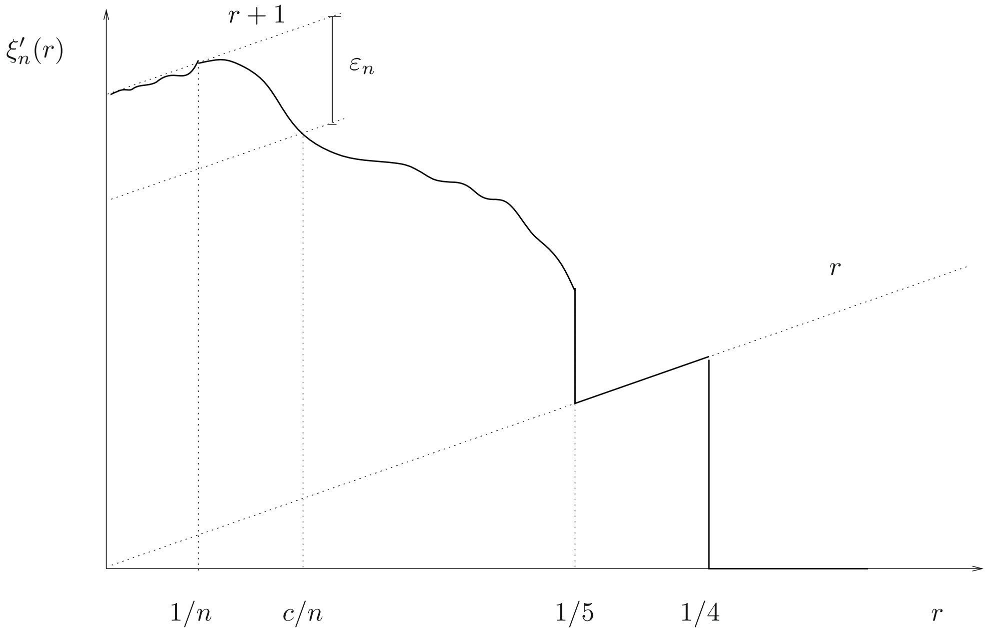

FIGURE 5. Qualitative behavior of $\xi'_{n,F,\mu}$.

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,

$$\sup_{f\in G^k_{1/k}} (\mathbb{E}f - \mathbb{E}_n f) = \frac{1}{k} + \frac{2(k+1)^2 \sup_J \ell_J - 3kn - 2n}{kn(2k+1)}\cdot$$

From Talagrand's concentration inequality (Thm. 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 $1 - \mathrm{e}^{-c_1 n t^2}$,

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

where the last inequality holds by (3.3).

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

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

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

for every $k \le n/c_1$.

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 $G^k_{1/k}$ or are scaled down versions of $G^{k'}_{1/k'}$ for $k' \le k$. Hence,

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

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

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

provided that $k \le n/c$. □

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 $n \ge N(x)$,*

(1) *With probability at least $1 - \mathrm{e}^{-x}$, $\mathbb{E}\hat{f} \le c/n \sim s_n^*$.*
(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 $F_r = \{f \in \mathrm{star}(F, 0) : \mathbb{E}f = r\}$ 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. [1] *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* $\xi_n(r) \geq (1+\alpha)r\lambda$, *then with probability at least* $1 - \mathrm{e}^{-x}$,

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

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

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

(3) *If* $\xi'_n(r) \geq (1+\alpha)r\lambda$, *then with probability at least* $1 - \mathrm{e}^{-x}$,

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

(4) *If* $\xi'_n(r) \leq (1-\alpha)r\lambda$, *then with probability at least* $1 - \mathrm{e}^{-x}$,

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

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*

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

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

If we define the "empirical shell,"

$$F^n_{\frac{r}{2},\frac{3r}{2}} := \{f \in \mathrm{star}(F, 0) \;:\; r/2 \leq \mathbb{E}_n f \leq 3r/2\},$$

the corollary shows that, for suitably 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. For this, define the random variables

$$R_n f = \frac{1}{n}\sum_{i=1}^{n} \sigma_i f(X_i) \quad \text{and} \quad R_n(F) = \sup_{f \in F} R_n f, \tag{4.1}$$

where $\sigma_1, ..., \sigma_n$ 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 $X_1, ..., X_n$,

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

Theorem 4.3. *There are absolute constants* c, c_1, c_2, *and* c_3 *for which the following holds. Let* F *be a* (β, B)*-Bernstein class that contains* 0 *and for which* $\sup_{f \in F} \|f\|_\infty \leq b$. *If*

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

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

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

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

Proof. By Lemma 2.3, $\xi_n(r) \leq \frac{r}{4}$ if and only if $r \geq r^*_n$. Thus, by Corollary 4.2 (for appropriately chosen c), if $r \geq \tilde{r}'_n$, then with probability larger than $1-\mathrm{e}^{-x}$, $F_r \subseteq F^n_{\frac{r}{2},\frac{3r}{2}}$, 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 (Thm. A.2) and concentration of Rademacher averages around their mean (Thm. A.3), and since $r \geq \frac{cbx}{n}$, it follows that with probability at least $1-2\mathrm{e}^{-x}$,

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

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 that are 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 $1-2(bn+1)\mathrm{e}^{-x}$,

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

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 $1-2(bn+1)\mathrm{e}^{-x}$, every $r \in [\tilde{r}'_n, b]$ satisfies

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

where k satisfies that $r \in [k/n, (k+1)/n]$ 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 $\hat{r}^*_n = \inf\{r : \hat{\xi}_n(r) \leq \frac{r}{4}\}$. By Theorem 4.3, with probability at least $1-2(bn+1)\mathrm{e}^{-x}$, $\hat{r}^*_n \geq r^*_n$. Moreover, since $\hat{\xi}_n(r)/r$ is non-increasing, $r \geq \hat{r}^*_n$ if and only if $\hat{\xi}_n(r) \leq \frac{r}{4}$.

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 $\hat{\xi}_n(r_R) \leq r_R/4$ 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, $\bar{r} - \frac{1}{n} \leq \hat{r}_n^* \leq \bar{r}$. Hence, for every n, with probability larger than $1 - 2(bn+1)\mathrm{e}^{-x}$, $r_n^* \leq \bar{r}$.

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 contains 0. For every integer n, any $x > 0$, and any sample X of size n, with probability at least $1 - (2bn+3)\mathrm{e}^{-x}$, $\mathbb{E}\hat{f} \leq \mathrm{RSTAR}(F, X)$.*

Note that $\mathrm{RSTAR}(F, X)$ 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 [3] is that instead of taking the whole empirical ball as in [3], 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 (β, B)-Bernstein class that contains 0 such that $\sup_{f \in F} \|f\|_\infty \leq b$. 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 $1 - 2(bn+1)\mathrm{e}^{-x}$

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

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 $r \geq \tilde{r}_n$ then with probability larger than $1 - \mathrm{e}^{-x}$, $F_r \subset F^n_{(1-\varepsilon_n)r, (1+\varepsilon_n)r}$. 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 $1 - 2(bn+1)\mathrm{e}^{-x}$, for every $r \in [\tilde{r}_n, b]$, $\xi(r) \leq \hat{\xi}_n(r)$. □

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

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

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 $1 - 2bne^{-x}$, $r_n^* \leq \bar{r}$, 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 [1] 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 $\tau = \{x_1, ..., x_n\}$, 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 functions in F_1^k have expectations of the order of $1/k$ while 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, $P_\tau F_2^\ell = P_\tau F_1^\ell = \{-1, 0, 1\}^n$.

Therefore, F_1 and F_2 are star-shaped, Bernstein classes that have identical coordinate projections, making it impossible to distinguish 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* $1-\mathrm{e}^{-x}$, $\mathbb{E}\hat{f} \le c/n \sim s_n^*(F_1)$.
(2) *For* F_2, *with probability* 1, $\mathbb{E}\hat{f} \ge 1/4 \sim r_n^*(F_2)$.

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. 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 $f \in F_1^k$,

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

Clearly, for a class of functions bounded by 1, $\xi'_{n,F,\mu}(r) - r \le 1$, 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 $r \ge c/n$, $\xi'_{n,F_1,\mu}(r) - r \le 1/2$. This is the case because the sets F_1^k are not "rich" enough when projected onto samples of size n as long as $k \le n/c$.

Indeed, the function class F_1^n has low complexity in terms of the combinatorial dimension $\mathrm{vc}(F_1^n, \varepsilon)$ (see Def. A.4). In particular, the definitions imply that $\mathrm{vc}(F_1^k, \varepsilon) \le 2k$ for all $0 < \varepsilon \le 2$ 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 $\mathbb{E}R_n(F_1^k) \le c_2\sqrt{k/n}$. 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 $1 - \mathrm{e}^{-c_1 n t^2} = 1 - \mathrm{e}^{-c_1' n}$,

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

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

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

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, with probability at least $1 - n\mathrm{e}^{-c_1' n}$, for any $k \le n/c$,

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

Taking the expectation,

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

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

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

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

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

$$F_2 = \bigcup \{\alpha f : f \in F_2^k,\, k \in \mathbb{N},\, \alpha \in [0,1]\},$$

and empirical minimization is a linear operation, it is clear that the empirical minimum will be attained at $\alpha = 1$ (using a similar argument to the one used in Lem. 2.4). Since all the functions in $\cup_{k \in \mathbb{N}} F_2^k$ have expectation greater than $1/4$, then 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

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

where if $\alpha_k = 1$ then $F_{\alpha_k}^k = F_1^k$ and if $\alpha_k = 2$ then $F_{\alpha_k} = F_2^k$. 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_1^j or F_2^j. 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 observation worth noting is that the class F_1 is not a Glivenko-Cantelli class. The classes F_1^k become richer as k grows – *i.e.*, in the part of F_1 in which the expectation of functions is smaller. The reason one can still obtain a generalization bound even for classes that are not Glivenko-Cantelli is because the method of [1] uses the expectation of the empirical process indexed by $\{f \in \text{star}(F,0) : \mathbb{E}f = r\}$, 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 [3], it would be impossible.

A. Additional material

The main technical tool we require is Talagrand's celebrated concentration theorem for empirical processes [13,24,25]. The version we use is due to Bousquet [7], building on Massart's argument (see also [10,17,22]).

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 $\rho > 0$,

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

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.

Recall the definition of the Rademacher averages of a class from page 330, equation (4.1). 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 [30].

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 [6], 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 [3].

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 $1-\mathrm{e}^{-x}$,*

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

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

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

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 $\mathrm{vc}(F)$ 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

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

The following result is a recent extension, due to Rudelson and Vershynin [23] to well-known estimates (see [9]) 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,*

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

References

[1] P.L. Bartlett and S. Mendelson, Empirical minimization. *Probab. Theory Relat. Fields* **135** (2006) 311–334.

[2] P.L. Bartlett and M.H. Wegkamp, Classification with a reject option using a hinge loss. *J. Machine Learn. Res.* **9** (2008) 1823–1840.

[3] P.L. Bartlett, O. Bousquet and S. Mendelson, Local Rademacher Complexities. *Ann. Statist.* **33** (2005) 1497–1537.

[4] P.L. Bartlett, M.I. Jordan and J.D. McAuliffe, Convexity, classification, and risk bounds. *J. Am. Statist. Assoc.* **101** (2006) 138–156.

[5] G. Blanchard, G. Lugosi and N. Vayatis, On the rate of convergence of regularized boosting classifiers. *J. Mach. Learn. Res.* **4** (2003) 861–894.

[6] S. Boucheron, G. Lugosi and P. Massart, Concentration inequalities using the entropy method. *Ann. Probab.* **31** (2003) 1583–1614.

[7] O. Bousquet, *Concentration Inequalities and Empirical Processes Theory Applied to the Analysis of Learning Algorithms*. Ph.D. thesis, École Polytechnique, 2002.

[8] R.M. Dudley, *Uniform Central Limit Theorems*, Cambridge University Press (1999).

[9] D. Haussler, Sphere Packing Numbers for Subsets of the Boolean n-cube with Bounded Vapnik-Chervonenkis Dimension. *J. Combin. Theory Ser. A* **69** (1995) 217–232.

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

[11] V. Koltchinskii, Local Rademacher Complexities and Oracle Inequalities in Risk Minimization. *Ann. Statist.* **34** (2006).

[12] V. Koltchinskii and D. Panchenko, Rademacher processes and bounding the risk of function learning. High Dimensional Probability, Vol. II (2000) 443–459.

[13] M. Ledoux, The Concentration of Measure Phenomenon, volume 89 of *Mathematical Surveys and Monographs*. American Mathematical Society (2001).

[14] W.S. Lee, P.L. Bartlett and R.C. Williamson, The Importance of Convexity in Learning with Squared Loss. *IEEE Trans. Informa. Theory* **44** (1998) 1974–1980.

[15] G. Lugosi and N. Vayatis, On the Bayes-risk consistency of regularized boosting methods (with discussion), *Ann. Statist.* **32** (2004) 30–55.

[16] G. Lugosi and M. Wegkamp, Complexity regularization *via* localized random penalties. *Ann. Statist.* **32** (2004) 1679–1697.

[17] P. Massart, The constants in Talagrand's concentration inequality for empirical processes. *Ann. Probab.* **28** (2000) 863–884.

[18] P. Massart, Some applications of concentration inequalities to statistics. *Ann. Fac. Sci. Toulouse Math.* **IX** (2000) 245–303.

[19] P. Massart and E. Nédélec, Risk bounds for statistical learning. *Ann. Statist.* **34** (2006) 2326–2366.

[20] S. Mendelson, Improving the sample complexity using global data. *IEEE Trans. Inform. Theory* **48** (2002) 1977–1991.

[21] 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).

[22] E. Rio, Inégalités de concentration pour les processus empiriques de classes de parties. *Probab. Theory Relat. Fields* **119** (2001) 163–175.

[23] M. Rudelson and R. Vershynin, Combinatorics of random processes and sections of convex bodies. *Ann. Math.* **164** (2006) 603–648.

[24] M. Talagrand, Sharper Bounds for Gaussian and Empirical Processes. *Ann. Probab.* **22** (1994) 20–76.

[25] M. Talagrand, New concentration inequalities in product spaces. *Inventiones Mathematicae* **126** (1996) 505–563.

[26] B. Tarigan and S.A. Van de Geer, Adaptivity of support vector machines with ℓ_1 penalty. Technical Report MI 2004-14, University of Leiden (2004).

[27] A. Tsybakov, Optimal aggregation of classifiers in statistical learning. *Ann. Statist.* **32** (2004) 135–166.

[28] S.A. Van de Geer, A new approach to least squares estimation, with applications. *Ann. Statist.* **15** (1987) 587–602.

[29] S.A. Van de Geer, *Empirical Processes in M-Estimation*, Cambridge University Press (2000).

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

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