Empirical Minimization

Peter L. Bartlett

Division of Computer Science and Department of Statistics

University of California at Berkeley

367 Evans Hall

Berkeley, CA 94720-3860

bartlett@cs.berkeley.edu

Shahar Mendelson

Centre for Mathematics and its Applications

The Australian National University

Canberra, ACT 0200, Australia

shahar.mendelson@anu.edu.au

October 12, 2005

Abstract

We investigate the behavior of the empirical minimization algorithm using various methods. We first analyze it by comparing the empirical, random, structure and the original one on the class, either in an additive sense, via the uniform law of large numbers, or in a multiplicative sense, using isomorphic coordinate projections. We then show that a direct analysis of the empirical minimization algorithm yields a significantly better bound, and that the estimates we obtain are essentially sharp. The method of proof we use is based on Talagrand's concentration inequality for empirical processes.

Keywords: empirical processes, error bounds, isomorphic coordinate projections, empirical minimization.

1 Introduction

Let F be a class of real-valued functions defined on a set $\mathcal{X}$, and suppose that $X_1, \ldots, X_n, X \in \mathcal{X}$ are independent and identically distributed. An

empirical minimizer $\hat{f} \in F$ is a function that minimizes

$$\mathbb{E}_n f = \frac{1}{n} \sum_{i=1}^{n} f(X_i).$$

In case no such minimum exists, we consider ρ-approximate empirical minimizers, which are functions $\hat{f} \in F$ satisfying

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

where $\rho \geq 0$.

In this article, we study the expectation of the empirical minimizer, defined as

$$\mathbb{E}\left[\hat{f}(X)|X_1, \ldots, X_n\right],$$

and for brevity, we write this conditional expectation as $\mathbb{E}\hat{f}$. For reasons that will be made clear immediately, it is natural to assume that for every $f \in F$, $\mathbb{E}f \geq 0$, although functions in F can take negative values.

The study of bounds on $\mathbb{E}\hat{f}$ that hold with high probability arises in many applied areas, including the analysis of randomized optimization methods involving Monte Carlo estimates of integrals, and prediction problems that arise in machine learning and nonparametric statistics. We focus on the latter motivation here: Suppose that a learning algorithm 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. For $f(x, y) = \ell(g(x), y)$, 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 $F = \{(x, y) \mapsto \ell(g(x), y) : g \in G\}$. More frequently, we are concerned with *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

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

Notice that $\mathbb{E}f \geq 0$ for all $f \in F$, but functions in F can take negative values. Indeed, the case of functions that can be negative is prevalent; if ℓ is a metric and for each x, y there is a $g \in G$ with $g(x) = y$, the assumption that every $f \in F$ is nonnegative corresponds to assuming that $Y = g^*(X)$ almost surely. The existence of such a g^* is typically an unreasonable assumption about the probability distribution P, even more so that this function is in the class G.

For most of the remainder of the paper, we ignore the underlying $\mathcal{Y}$-valued class G, and consider classes F of uniformly bounded real functions. The following lemma shows that, under mild conditions, such a class corresponds to an excess loss class. The proof is presented in the appendix.

Lemma 1.1 *Suppose that $(\mathcal{X}, \mathcal{F})$ is a measurable space, $F \subseteq [-1, 1]^{\mathcal{X}}$ is a set of measurable functions, $0 \in F$, and $x \mapsto \inf\{f(x) : f \in F\}$ is measurable. Suppose also that $\ell : \mathcal{Y}^2 \to \mathbb{R}^+$ is such that for some $y_0 \in \mathcal{Y}$, $\{\ell(y, y_0) : y \in \mathcal{Y}\}$ contains a closed interval of length $\sup\{f_1(x) - f_2(x) : x \in \mathcal{X}, f_1, f_2 \in F\}$. Then there is a class $G \subseteq \mathcal{Y}^{\mathcal{X}}$ and a function $g^* \in G$ for which*
$$F = \{x \mapsto \ell(g(x), y_0) - \ell(g^*(x), y_0) : g \in G\}$$

and $x \mapsto \ell(g(x), y_0)$ is measurable for each $g \in G$. Thus, if the distribution of X is such that $\mathbb{E}f \geq 0$ for all $f \in F$, then $g^* \in G$ minimizes $\mathbb{E}\ell(g(X), y_0)$ and F is the excess loss class for G.

One case in which Lemma 1.1 clearly applies is when $\mathcal{Y} = [-1, 1]$ and $\ell(y, y_0) = (y - y_0)^2$ when one takes $y_0 = 0$. Thus, subject to mild measurability assumptions, every class of functions bounded by 1 with a nonnegative expectation is a squared excess loss of some class G.

We consider several approaches to estimating the expectation of the empirical minimizer, all of which depend on expectations of the following centered empirical processes, indexed by certain subsets of F.

$$\xi_n(r_1, r_2) = \mathbb{E}\sup\{\mathbb{E}f - \mathbb{E}_n f : f \in F, \, r_1 \leq \mathbb{E}f < r_2\},$$
$$\xi_n(r) = \mathbb{E}\sup\{\mathbb{E}f - \mathbb{E}_n f : f \in F, \, \mathbb{E}f = r\}.$$

The first two approaches are based on the ability to relate the empirical (random) structure endowed on F by the measure $P_n = n^{-1}\sum_{i=1}^{n} \delta_{X_i}$ with

the real one, endowed by μ. In the classical approach, which involves a uniform law of large numbers argument, one estimates the "worst deviation" of $\mathbb{E}f$ from $\mathbb{E}_n f$ over the entire class. It is possible to show that typically the dominant term in the upper bound for $\mathbb{E}\hat{f}$ is in this case

$$\sup\{r > 0 : \xi_n(0,1) - r \geq 0\}. \tag{1.1}$$

Essentially equivalent results were presented in [11, 1, 20].

If the class F satisfies an additional regularity condition (namely that it is star-shaped around zero—see Section 2.2 for the definition) and if variances of functions in F are bounded by their expectations, then this result can be improved. Indeed, one can show that for $0 < \epsilon < 1$, with high probability (which depends on ϵ), for a "large portion" of F, which contains the functions with "large expectations",

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

Note that this notion of similarity is multiplicative, and means that, for a large subset of F, the empirical and actual structures are equivalent, in the sense that a random coordinate projection of that portion of F preserves the L_1 structure. In this case, the dominant term in the estimate on $\mathbb{E}\hat{f}$ becomes

$$\sup\{r > 0 : \xi_n(r) - \theta r \geq 0\}$$

for some $0 < \theta < 1$.

This is an improvement on the estimate (1.1) (that is, on $\xi_n(0,1)$). Indeed, it is possible to show under the star-shape assumption on F the the sets $A = \{r \geq 0 : \xi_n(r) \geq \theta r\}$ and $B = \{r \geq 0 : \xi_n(0,r) \geq \theta r\}$ are intervals containing 0 (see Lemma 2.10). Denote by $a^* = \sup A$ and $b^* = \sup B$ and observe that, $a^* \leq b^*$, without loss of generality, $b^* > 0$ and (again, by lemma 2.10), $\xi_n(0,r)$ is continuous and monotone. Thus, $\xi_n(0,b^*) = \theta b^*$, implying that $\theta b^* \leq \theta \xi_n(0,b^*) \leq \xi_n(0,1)$, and in particular, that $\theta a^* \leq \xi_n(0,1)$.

This estimate also improves earlier related error bounds from [18, 19, 12, 2, 16]. In particular, the function $\xi_n(r) = \mathbb{E}\sup\{\mathbb{E}f - \mathbb{E}_n f : f \in F, \mathbb{E}f = r\}$ that appears in our bound is replaced in these earlier results by various upper bounds on the larger function $\mathbb{E}\sup\{\mathbb{E}f - \mathbb{E}_n f : f \in F, \mathbb{E}f^2 \leq r\}$. It is important to emphasize that the bound obtained in Section 2.2 using the indexing set $\{f : \mathbb{E}f = r\}$ is significantly sharper than any bound that could be established using the localization $\{f : \mathbb{E}f^2 \leq r\}$, and can lead to improved convergence rates in some examples. The reason is that the latter set can be much larger than the former for small values of r. Moreover,

4

in some applications (see, for example, [22]), the fact that the indexing set is determined by the expectation rather than the second moment plays an integral part in obtaining the error bound for a squared loss class.

The same comments apply in comparison with convergence rate results for M-estimators in terms of a fixed point of the modulus of continuity of the relevant empirical process (see [26, 29, 28, 27]).

The proof of the error bound in Section 2.2 is surprisingly simple (particularly in light of the considerable effort required for the proofs of the results it improves), and is based on a ratio limit theorem type of argument. Although the ratio limit theorem we prove is new and could have implications in other areas of mathematics (e.g. various embedding problems in the local theory of normed spaces), it is close in nature to other ratio limit theorems [7, 13]. The main novelty is the way in which the ratio limit theorem, combined with the mild structural assumptions on F, yield the required bound on $\mathbb{E}\hat{f}$, which is the optimal estimate one can obtain using this strategy.

It turns out that the latter estimate can be improved even further, using a direct analysis of the empirical minimization algorithm, rather than by means of a structural result which holds for every function in the class. We show that the dominant term in the upper bound on $\mathbb{E}\hat{f}$ is, roughly,

$$\arg \max_{r>0} \left(\xi_n(r) - r \right),$$

and that this bound is essentially sharp. Moreover, it significantly improves the structural estimates. To that end, we present an example where the upper bound decreases from $1/4$ using the structural approach to $1/n$ using the direct analysis.

1.1 Concentration inequalities

In this section we present the concentration inequalities we require. The first is Bernstein's inequality (see, for example, [29]).

Theorem 1.2 *Let P be a probability measure on $\mathcal{X}$ and set $X_1, ..., X_n$ to be independent random variables distributed according to P. Given a function $f : \mathcal{X} \to \mathbb{R}$, set $Z = \sum_{i=1}^{n} f(X_i)$, let $b = \|f\|_\infty$ and put $\sigma^2 = n\mathbb{E}f^2$. Then*

$$Pr\left\{|Z - \mathbb{E}Z| \geq x\right\} \leq 2\exp\left(-\frac{x^2}{2(\sigma^2 + bx/3)}\right).$$

The second concentration result is a functional version of Bernstein's inequality, due to Talagrand [25, 14]. The random variable Z defined by a

single function in Bernstein's inequality becomes the supremum of a centered empirical process.

Theorem 1.3 *Let F be a class of functions defined on $\mathcal{X}$ and set P to be a probability measure such that for every $f \in F$, $\|f\|_\infty \leq b$ and $\mathbb{E}f = 0$. Let $X_1, ..., X_n$ be independent random variables distributed according to P and set $\sigma^2 = n \sup_{f \in F} \mathrm{var}\,[f]$. Define*

$$Z = \sup_{f \in F} \sum_{i=1}^{n} f(X_i),$$

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

Then, for every $x > 0$,

$$Pr\left(\{|Z - \mathbb{E}Z| \geq x\}\right) \leq C \exp\left(-\frac{x}{Kb} \log\left(1 + \frac{bx}{\sigma^2 + b\mathbb{E}\bar{Z}}\right)\right), \qquad (1.2)$$

where C and K are absolute constants. The same inequality is also true when $\bar{Z}$ replaces Z in (1.2).

In most of the applications we explore, it is easier to use the following version of Talagrand's inequality.

Theorem 1.4 *There is an absolute constant K for which the following holds. Let F, Z and $\bar{Z}$ be as in Theorem 1.3. Then, for every $x > 0$ and every $\alpha > 0$,*

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

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

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

The inequality for $\bar{Z}$ is due to Massart [17]. The one sided versions were shown by Rio [24] and Klein [10]. The best estimates on the constants in all cases are due to Bousquet [5].

Finally, a notational convention. All absolute constants are positive numbers, which will be denoted by C, c, K, k. Their values may change from line to line.

2 Comparing the empirical and actual structures

In this section we investigate various notions of similarity and conditions which ensure that with high probability, the empirical and the actual structures on a class (that is, the expectations and empirical means) are sufficiently close. This is important from our point of view because, when the two structures are comparable, an empirical minimizer must have a small expectation.

2.1 The uniform law of large numbers

The first notion of similarity we explore is based on the uniform law of large numbers. Recall that a class of functions F satisfies the uniform law of large numbers with respect to a probability measure P if, for every $\epsilon > 0$,

$$\lim_{n \to \infty} Pr\left(\{\|P - P_n\|_F \geq \epsilon\}\right) = 0,$$

where

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

$$\mathbb{E}_n f = \frac{1}{n} \sum_{i=1}^{n} f(X_i),$$

$X_1, ..., X_n$ are independent random variables distributed according to P, and $P_n = n^{-1} \sum_{i=1}^{n} \delta_{X_i}$, which is the empirical measure supported on $X_1, ..., X_n$.

This leads to the first notion of similarity between the empirical and actual structures.

Definition 2.1 *Given an integer n and a probability measure P, we say that the empirical and actual structures on F are (λ, δ)-close if*

$$Pr\left(\{\|P - P_n\|_F \geq \lambda\}\right) \leq \delta.$$

In this case, the measure of similarity is additive, uniformly on the entire class. Observe that if $0 \in F$, the empirical and actual structures are (λ, δ)-close, and f satisfies $\mathbb{E}_n f \leq \inf_{f \in F} \mathbb{E}_n f + \rho$, then with probability at least $1 - \delta$, $\mathbb{E}f \leq \lambda + \rho$. In particular, if $\hat{f}$ is an empirical minimizer then with probability larger than $1 - \delta$, $\mathbb{E}\hat{f} \leq \lambda$.

The following results reveal the benefits and limitations of this notion of similarity. Although they are not new, we present them for the sake of completeness.

Theorem 2.2 *There exists an absolute constant C for which the following holds. For any class of functions F, and every $0 < \delta < 1$, the empirical and actual structures are (λ_n, δ) close, provided that*

$$\lambda_n \geq C \max \left\{ \mathbb{E} \left\| P - P_n \right\|_F, \sigma_F \sqrt{\frac{\log(1/\delta)}{n}}, \frac{b \log(1/\delta)}{n} \right\},$$

where $\sigma_F^2 = \sup_{f \in F} \mathrm{var}\,[f]$ and $b = \sup_{f \in F} \|f\|_\infty$.

The proof of this claim follows immediately from Theorem 1.4, and is omitted.

The following theorem shows that the estimates in Theorem 2.2 cannot be improved by more than a constant factor, unless n is small.

Theorem 2.3 *There are absolute constants c, c' and C for which the following holds. Let F be a class of functions satisfying $\sup_{f \in F} \|f\|_\infty \leq 1$ and set $\sigma_F^2 = \sup_{f \in F} \mathrm{var}\,[f]$. Then,*

$$\mathbb{E} \left\| P - P_n \right\|_F \geq c \frac{\sigma_F}{\sqrt{n}}.$$

Furthermore, for every integer $n \geq 1/\sigma_F^2$, with probability at least c',

$$\left\| P - P_n \right\|_F \geq C \mathbb{E} \left\| P - P_n \right\|_F.$$

Theorem 2.3 is most likely not new, but we could not locate an appropriate reference. We include the proof in the appendix.

These upper and lower bounds clearly reveal the limitation of this notion of similarity. Even for "very small" classes, one cannot hope to have λ_n decay to 0 faster than $O(1/\sqrt{n})$, while for larger classes, the dominating term becomes the "global" average $\mathbb{E} \left\| P - P_n \right\|_F$. In particular, it would be impossible to use this notion of similarity to obtain an asymptotic result stronger than $\mathbb{E}\hat{f} \leq 1/\sqrt{n}$ with high probability.

As an example (which is also well known), consider a class of binary-valued functions which has a finite Vapnik-Chervonenkis dimension (see [30]).

Lemma 2.4 *There exist absolute constants C and c for which the following holds. Let F be a class of $\{0, 1\}$-valued functions, such that $\mathrm{vc}(F) \leq d$. Then for any probability measure and every integer n,*

$$\mathbb{E} \left\| P - P_n \right\|_F \leq C \sqrt{\frac{d}{n}},$$

In particular, for every probability measure P, the empirical and actual structures are (λ_n, δ)-close, provided that

$$\lambda_n \geq C \max \left\{ \sqrt{\frac{d}{n}}, \sqrt{\frac{\log(1/\delta)}{n}} \right\}.$$

On the other hand, for any two integers n and d, there exists a probability measure P for which

$$\mathbb{E} \, \|P - P_n\|_F \geq c \min \left\{ \sqrt{\frac{d}{n}}, 1 \right\}.$$

The proof of Lemma 2.4 is standard and thus omitted.

Finally, notice that $\mathbb{E}\|P - P_n\|_F = \xi_n(0,1)$ for $\sup_{f \in F} \|f\|_\infty \leq 1$. To conclude, this notion of similarity involves bounding $\mathbb{E} \, \|P - P_n\|_F$. No significant structural assumptions (other than an L_∞ bound on the elements of F) are required, and the empirical and actual structures are "close" on the entire class. Unfortunately, λ_n cannot decrease faster than $1/\sqrt{n}$, which limits the usefulness of this approach to estimate $\mathbb{E}\hat{f}$.

2.2 Isomorphic coordinate projections

Here, we focus on a slightly different notion of similarity of the empirical and actual structures. The question we investigate is when "most" random coordinate projections are isomorphisms.

Definition 2.5 *For $\tau = (X_1, \ldots, X_n)$, we say that the coordinate projection $\Pi_\tau : f \mapsto (f(X_1), \ldots, f(X_n))$ is an ϵ-isomorphism if for every $f \in F$,*

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

The reason for the name ϵ-isomorphism is that if $F = \{|g_1 - g_2| : g \in G\}$, then Π_τ is an ϵ-isomorphism if and only if

$$(1 - \epsilon)\|f\|_{L_1(P)} \leq \|\Pi_\tau f\|_{L_1^n} \leq (1 + \epsilon)\|f\|_{L_1(P)},$$

and thus the random projection $\Pi_\tau : (G, L_1(P)) \to (G, L_1^n)$ is a bi-Lipschitz function.

Observe that Π_τ is an ϵ-isomorphism on F in the sense of Definition 2.5 if and only if the same holds for the linear span of F.

In order to ensure that a coordinate projection is a good isomorphism for a single function, the "mass" of the function must be more-or-less evenly

spread on the space (Ω, P). To that end, it suffices to have a lower bound on the expectation of the function (total mass) and an upper bound on, say, the L_∞ norm of the function. Thus, the mass can not be concentrated on "few" atoms in Ω and a random choice of coordinates is a good representation of the function. The use of a random coordinate projection approach for individual functions is common in asymptotic geometry, most notably, in the context of embedding finite dimensional subspaces of L_p in ℓ_p^n (see [9] and references therein). Here, we establish a similar bound that holds uniformly over a set of functions and not just for an individual function. More significantly, we use this approach to obtain an improved error bound with a simple proof.

It turns out that while most projections are not ϵ-isomorphisms for the entire class F in the sense of Definition 2.5, most projections are ϵ-isomorphisms for a large portion of F, which suffices for our investigation.

We make three mild structural assumptions about the class. The first, as in the previous section, is the assumption that functions in F are bounded by b. The second is that the class F is star-shaped around 0, that is, for every $0 \le a \le 1$ and any $f \in F$, $af \in F$. For the third assumption we require the following definition.

Definition 2.6 *We say that F is a (β, B)-Bernstein class with respect to the probability measure P (where $0 < \beta \le 1$ and $B \ge 1$), if every f in F satisfies*

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

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.

Observe that if f belongs to a Bernstein class then $\mathbb{E}f \ge 0$. The name "Bernstein class" arises because this property allows better concentration, and thus a faster rate of convergence via "function class" generalizations of Bernstein's inequality. Obviously, if F consists of nonnegative functions bounded by b then F is a $(1, b)$-Bernstein class with respect to any probability measure. One such example is when F is a loss class, that is, $F = \{(x, y) \mapsto \ell(g(x), y) : g \in G\}$ for some function $\ell : \mathbb{R}^2 \to [0, \infty)$ satisfying $\sup_{f \in F} \|f\|_\infty = b < \infty$. In fact, many loss classes that do not consist of nonnegative functions have similar properties. For example, let G be a convex class of functions bounded by 1. Let $\ell : \mathbb{R} \times \mathcal{Y} \to [0, 1]$ be a loss function for which, for some constants L, c, r, and for all $y \in \mathcal{Y}$, the function $\hat{y} \mapsto \ell(\hat{y}, y)$ is L-Lipschitz and has modulus of convexity satisfying $\delta(\epsilon) \ge c\epsilon^r$. Here, the modulus of convexity of a function $f : \mathbb{R} \to \mathbb{R}$ is defined as

$$\delta(\epsilon) = \inf \left\{ (f(x) + f(y))/2 - f((x + y)/2) : |x - y| \ge \epsilon \right\}.$$

10

Recall that for a probability distribution P on $\mathcal{X} \times \mathcal{Y}$, $g^* \in G$ is the minimizer of $\mathbb{E}\ell(g(X), Y)$, and the excess loss class is defined by

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

Then F is a β-Bernstein class, with $\beta = \min\{1, 2/r\}$. This is true, in particular, for $\ell(\hat{y}, y) = (y - \hat{y})^p$, where $\beta = 1$ for $1 \leq p \leq 2$ and $\beta = 2/p$ for $2 < p < \infty$. See [15, 19, 23, 3].

Theorem 2.7 *There is an absolute constant c for which the following holds. Let F be a class of functions, such that for every $f \in F$, $\mathbb{E}f = \lambda$ and $\|f\|_\infty \leq b$. Assume that F is a (β, B)-Bernstein class, and suppose that $0 < \epsilon < 1$ and $0 < \theta < 1$ satisfy*

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

1. If $\mathbb{E}\|P - P_n\|_F \geq (1 + \theta)\lambda\epsilon$, then

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

2. If $\mathbb{E}\|P - P_n\|_F \leq (1 - \theta)\lambda\epsilon$, then

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

For example, if F consists of nonnegative functions bounded by 1, then $\beta = 1$ and $b = B = 1$, and the condition on λ becomes

$$\lambda \geq \frac{x}{n\theta^2\epsilon^2}.$$

Proof: The proof follows in a straightforward way from Theorem 1.4. Define $Z = n\|P - P_n\|_F$, set $\sigma^2 = n \sup_{f \in F} \text{var}[f]$ and note that Π_τ is an ϵ-isomorphism of F if and only if $Z \leq \epsilon\lambda n$.

To prove the first part of our claim, recall that by Theorem 1.4, for every $\alpha, x > 0$, with probability larger than $1 - e^{-x}$,

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

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

$$\frac{1}{2}\theta n\lambda\epsilon \geq (Bn\lambda^\beta Kx)^{1/2} + K\left(1 + \frac{2(1 + \theta)}{\theta}\right)xb,$$

11

which holds by the condition on λ.

The second part of the claim also follows from Theorem 1.4: for every $\alpha, x > 0$, with probability larger than $1 - e^{-x}$,

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

Choosing $\alpha = \theta/(2(1 - \theta))$, we see that $Z < n\lambda\epsilon$ if

$$\frac{1}{2}\theta n\lambda\epsilon \geq (Bn\lambda^\beta Kx)^{1/2} + K\left(1 + \frac{2(1 - \theta)}{\theta}\right)xb,$$

so the condition on λ again suffices. ∎

Next, let us turn to a similar result, without the assumption that all class members have the same expectation. From here on, denote $F_\lambda = \{f \in F : \mathbb{E}f = \lambda\}$. The assumption that F is star-shaped around 0 ensures that the sets F_λ become "richer" as λ approaches 0. As our results show, there is a critical value of λ below which the sets F_λ are too rich to allow a comparison between the empirical and actual structures. In the next lemma we show that if one can control the structures on the set F_λ, it automatically guarantees the same for $\{f \in F : \mathbb{E}f \geq \lambda\}$.

Lemma 2.8 *Let F be star-shaped around 0 and let $\tau \in \mathcal{X}^n$. For any $\lambda > 0$ and $0 < \epsilon < 1$, the projection Π_τ is an ϵ-isomorphism of F_λ if and only if it is an ϵ-isomorphism of $\{f \in F : \mathbb{E}f \geq \lambda\}$.*

Proof: It suffices to show that if Π_τ is an ϵ-isomorphism of F_λ, then the same holds on $\{f \in F : \mathbb{E}f \geq \lambda\}$. To that end, observe that if $\mathbb{E}f = t > \lambda$, and since F is star-shaped around 0, $g = \lambda f/t \in F_\lambda$; hence, $(1 - \epsilon)\mathbb{E}f \leq \mathbb{E}_n f \leq (1 + \epsilon)\mathbb{E}f$ if and only if the same holds for g. ∎

From this result one easily obtains the following error bound:

Corollary 2.9 *Let F be a class of functions bounded by b, which is star-shaped around 0 and is a (β, B)-Bernstein class. For $0 < \epsilon, \lambda, \theta < 1$ and $x > 0$, if*

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

and $\mathbb{E}\|P - P_n\|_{F_\lambda} \leq (1 - \theta)\lambda\epsilon$, then with probability at least $1 - e^{-x}$, every $f \in F$ satisfies

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

Proof: By our assumption on $\mathbb{E}\left\|P - P_n\right\|_{F_\lambda}$ and λ, Theorem 2.7 implies that, with "large" probability, $(1 - \epsilon)\mathbb{E}f \leq \mathbb{E}_n f \leq (1 + \epsilon)\mathbb{E}f$, for every $f \in F_\lambda$. By Lemma 2.8, the same is true for every $f \in F$ that satisfies $\mathbb{E}f \geq \lambda$. ∎

Let us present a similar "one sided" result, which will be used later. Define

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

Lemma 2.10 *If F is star-shaped at 0, then for $0 < \alpha, \lambda < 1$,*

$$\xi_n(\alpha\lambda) \geq \alpha\xi_n(\lambda).$$

In particular, if $\alpha\lambda \leq \xi_n(\lambda)$ then for all $0 < \lambda' \leq \lambda$, $\alpha\lambda' \leq \xi_n(\lambda')$.
The same result is true for the function $\xi_n(0, r)$.

Observe that Lemma 2.10 implies that for every $0 < \alpha < 1$, the set $\{0 \leq r \leq 1 : \xi_n(r) \geq \alpha r\}$ is an interval containing 0.
Proof: Fix $\tau = (X_1, ..., X_n)$ and without loss of generality, suppose that $\sup_{f \in F_\lambda} \mathbb{E}f - \mathbb{E}_n f$ is attained at f. Then for any $0 < \alpha < 1$, $f' = \alpha f \in F_{\alpha\lambda}$ satisfies

$$\mathbb{E}f' - \mathbb{E}_n f' = \alpha \sup_{f \in F_\lambda} \mathbb{E}f - \mathbb{E}_n f,$$

and the first part follows.
For the second part, note that if $\lambda' \leq \lambda$,

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

∎

Here is a one-sided version of Corollary 2.9.

Theorem 2.11 *There exists an absolute constant c for which the following holds. Let F be a (β, B)-Bernstein class of functions bounded by b which is star-shaped around 0. Then for any $0 < \theta, \epsilon, \lambda < 1$ satisfying*

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

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

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

13

In particular, there is an absolute constant c such that if

$$r' = \max\left\{\inf\left\{r > 0 : \xi_n(r) \le r/4\right\}, \frac{cbx}{n}, c\left(\frac{Bx}{n}\right)^{1/(2-\beta)}\right\},$$

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

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

The way one should interpret Theorem 2.11 is as follows. The second and third terms in the definition of r' are the natural restrictions one has to impose to ensure that a random coordinate projection is a good isomorphism for a single function. Indeed, if $\mathbb{E}f$ is larger than those two terms, (combined with the assumption on the L_∞ norm and the second moment of f), then the mass of f is equally spread on Ω and a random coordinate projection would preserve its expectation. The first, and more significant term, is a complexity measure for the entire class, and is needed to ensure that a simultaneous coordinate selection is possible.

Proof: The first part of the claim follows a similar path to that of the previous result (with the one-sided concentration result in Theorem 1.4), and is omitted. The second part is evident by taking $\epsilon = \theta = 1/2$ and applying the first part for $\lambda = r'$. In particular, Lemma 2.10 shows that if

$$r' \ge \inf\left\{r > 0 : \xi_n(r) \le \frac{r}{4}\right\}$$

then $\xi_n(r') \le r'/4$. Thus, with large probability, if $f \in F$ satisfies $\mathbb{E}f \ge r'$, then $\mathbb{E}f \le 2\mathbb{E}_n f$. Since $\hat{f}$ is a ρ-approximate empirical minimizer and F is star-shaped at 0, it follows that $\mathbb{E}_n\hat{f} \le \rho$, so either $\mathbb{E}\hat{f} \le r'$ or $\mathbb{E}\hat{f} \le 2\rho$, as claimed. ∎

As mentioned in the introduction, this error bound improves previous error bounds, in which the dominating terms were various upper bounds on the fixed point of $\phi(r) = \mathbb{E}\sup|\mathbb{E}f - E_n f|$, where the supremum is taken with respect to $\{f \in F : \mathbb{E}f^2 \le r\}$ (see, for example [2, 12, 27, 28, 29]).

We end this section with the example of a binary-valued class with a finite VC dimension. (The first result of this form is due to Vapnik and Chervonenkis; see, for example, [4]).

Theorem 2.12 *There are absolute constants c and c' for which the following holds. Let G be a class of binary-valued functions which contains 0,*

such that $\mathrm{vc}(G) \leq d$, *and set* $F = \mathrm{star}(G,0)$. *Then, for every* $0 < \epsilon < 1$,
$\mathbb{E}\|P - P_n\|_{F_\lambda} \leq \epsilon\lambda/2$ *provided that*

$$\lambda \geq \frac{c}{\epsilon^2} \cdot \frac{d}{n} \log\left(\frac{n}{ed}\right).$$

In particular, for every $n \geq d$, *with probability larger than* $1 - \left(\frac{ed}{n}\right)^{c'd}$, *if*
$\mathbb{E}_n \hat{g} \leq \inf_{g \in G} \mathbb{E}_n g + \rho$, *then*

$$\mathbb{E}\hat{g} \leq c \max\left\{\frac{d}{n}\log\left(\frac{n}{ed}\right), \rho\right\}.$$

Before presenting the proof, recall that the covering number of a metric space (X, d) at scale ϵ is the minimal number of open balls (with respect to the metric d) of radius ϵ needed to cover X, and is denoted by $N(\epsilon, X, d)$.

Proof: Since G is a VC class, then by Haussler's inequality [8],

$$N\left(\epsilon, G, L_2\left(P\right)\right) \leq Cd(4e)^d \epsilon^{-2d},$$

where $d = \mathrm{vc}(G)$. A standard argument shows (see, e.g. [19]) the covering numbers of the star-shaped hull of G with 0 satisfy

$$N\left(\epsilon, F, L_2\left(P\right)\right) \leq N\left(\frac{\epsilon}{2}, G, L_2\left(P\right)\right) \cdot \left(\left\lceil\frac{2}{\epsilon}\right\rceil + 1\right),$$

which implies that

$$\log N\left(\epsilon, F, L_2\left(P\right)\right) \leq Cd \log\left(\frac{2}{\epsilon}\right).$$

Applying a symmetrization argument, and since a Rademacher process is subgaussian (see [21] for more details),

$$\mathbb{E}\|P - P_n\|_{F_\lambda} \leq C \max\left\{\frac{d}{n}\log\left(\frac{1}{\lambda}\right), \sqrt{\frac{d\lambda}{n}\log\left(\frac{1}{\lambda}\right)}\right\},$$

and the result follows from an easy computation. ∎

3   Empirical minimization

In this section we investigate the properties of the empirical minimizer, and compare the estimates we obtain to the ones obtained via the structural

15

results in the previous section. In particular, we show that there are cases where a direct analysis of the empirical minimization yields much sharper estimates than the structural approach. The approach we use bears some similarity to the technique of peeling. Recall that

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

The main result of this section is that the expectation of the empirical minimizer is essentially the maximizer of the function $\xi_n(r) - r$. For the sake of simplicity, we shall assume that the supremum is achieved at some s, that is,

$$\xi_n(s) - s = \sup\{\xi_n(r) - r : r > 0\},$$

and if this is not the case, a standard limiting argument can be applied.

It is easy to verify that this estimate does not violate the upper bounds of the previous section. By fixing any function in F_r, it is evident that $\xi_n(r) \geq 0$. Hence, considering r near zero shows that the maximal value of $\xi_n(r) - r$ must be at least 0. Thus, the maximizer s cannot be larger than

$$\inf \left\{ r > 0 : \xi_n(r) \leq r \right\} \leq \inf \left\{ r > 0 : \xi_n(r) \leq \frac{r}{4} \right\},$$

and this is no larger than r' introduced in the previous section.

To obtain upper and lower bounds on the expectation of the empirical minimizer, we consider values of r that do not quite maximize $\xi_n(r) - r$. Specifically, for $\epsilon > 0$, define

$$r_{\epsilon,+} = \sup \left\{ 0 \leq r \leq b : \xi_n(r) - r \geq \sup_s \left(\xi_n(s) - s \right) - \epsilon \right\},$$

$$r_{\epsilon,-} = \inf \left\{ 0 \leq r \leq b : \xi_n(r) - r \geq \sup_s \left(\xi_n(s) - s \right) - \epsilon \right\}.$$

Clearly, if s denotes the maximum and either $r > r_{\epsilon,+}$ or $r < r_{\epsilon,-}$, then

$$\xi_n(s) - s > \xi_n(r) - r + \epsilon.$$

The following theorem shows that, with a suitable choice of ϵ, the expectation of the empirical minimizer is approximately between $r_{\epsilon,-}$ and $r_{\epsilon,+}$. For the lower bound, we need an additional condition on the complexity of the subset of functions in F with "small" expectations. To that end, recall that for $r_1 < r_2$,

$$\xi_n(r_1, r_2) = \mathbb{E} \sup \left\{ \mathbb{E}f - \mathbb{E}_n f : f \in F, \ r_1 \leq \mathbb{E}f < r_2 \right\}.$$

Theorem 3.1 *For any $c_1 > 0$, there is a constant c (depending only on c_1) such that the following holds. Let F be a (β, B)-Bernstein class that is star-shaped at 0. Define s, $r_{\epsilon,+}$, and $r_{\epsilon,-}$ as above, and set*

$$r' = \max \left\{ \inf \left\{ r > 0 : \xi_n(r) \leq r/4 \right\}, \frac{cb(x + \log n)}{n}, c \left(\frac{B(x + \log n)}{n} \right)^{1/(2-\beta)} \right\},$$

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

$$\epsilon \geq c \left(\max \left\{ \sup_{s>0} \left(\xi_n(s) - s \right), r'^\beta \right\} \frac{(B + b)(x + \log n)}{n} \right)^{1/2} + \rho,$$

then

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

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

2. *If*

$$\xi_n(0, c_1/n) < \sup_{s>0} \left(\xi_n(s) - s \right) - \epsilon,$$

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

$$\mathbb{E}\hat{f} \geq r_{\epsilon,-}.$$

It is easy to verify that if F consists of nonnegative functions, then Theorem 3.1 recovers the error bound established in the previous section, but does not improve it. Also, the lower bound is vacuous in this case, as $f = 0$ is always an empirical minimizer, and thus it is impossible to obtain any nontrivial lower bound for such a class. Of course, as stated in the introduction, classes of nonnegative functions are not of interest here.

The proof of Theorem 3.1 involves splitting F into the subsets $F_r = \{f \in F : \mathbb{E}f = r\}$, and using concentration to show that there is likely to be a function with $\mathbb{E}f = s$ for which the empirical mean is smaller than any function with $\mathbb{E}f > r_{\epsilon,+}$ or $\mathbb{E}f < r_{\epsilon,-}$. (Here, s is the value which maximizes the difference $\xi_n(r) - r$.) We do this by progressively eliminating subsets F_r. For the upper bound, we first use the results of the previous section to show that it is unlikely that $\hat{f} \in F_r$ for $r \geq r'$, and then we split the interval $(r_{\epsilon,+}, r')$ into intervals of width Δ, and separately eliminate each of these. For the lower bound, we first use the condition on $\xi_n(0, c_1/n)$ to eliminate any $r < c_1/n$, and then separately eliminate intervals of width Δ from the interval $[c_1/n, r_{\epsilon,-})$. The following lemma is the main tool in eliminating these intervals.

Lemma 3.2 *Let $\hat{f}$ be a ρ-approximate empirical risk minimizer from F and set $r, s, \Delta \geq 0$ and $0 < \alpha < 1$. If*

$$\xi_n(s) - s \geq \xi_n(r, r + \Delta) - r + \alpha\xi_n(s) + \alpha\xi_n(r, r + \Delta)$$

$$+ \sqrt{\frac{BKx}{n}}\left(s^{\beta/2} + (r + \Delta)^{\beta/2}\right) + 2K\left(1 + \frac{1}{\alpha}\right)\frac{bx}{n} + \rho$$

then with probability at least $1 - 2e^{-x}$,

$$\mathbb{E}\hat{f} \notin [r, r + \Delta).$$

Note that the proof of the lemma uses the full strength of Talagrand's concentration inequality, particularly the fact that the constant α in Theorem 1.4 can be made arbitrarily close to 0.

Proof: By Theorem 1.4 (together with some easy manipulations), it follows that with probability at least $1 - 2e^{-x}$, both

$$\sup\left\{\mathbb{E}f - \mathbb{E}_n f : \mathbb{E}f = s\right\} - s$$

$$> (1 - \alpha)\xi_n(s) - s - \sqrt{\frac{BKxs^{\beta}}{n}} - K\left(1 + \frac{1}{\alpha}\right)\frac{bx}{n}$$

and

$$\sup\left\{\mathbb{E}f - \mathbb{E}_n f : r \leq \mathbb{E}f < r + \Delta\right\} - r$$

$$< (1 + \alpha)\xi_n(r, r + \Delta) - r + \sqrt{\frac{BKx(r + \Delta)^{\beta}}{n}} + K\left(1 + \frac{1}{\alpha}\right)\frac{bx}{n}.$$

In that case, if

$$(1 - \alpha)\xi_n(s) - s$$

$$\geq (1 + \alpha)\xi_n(r, r + \Delta) - r + \sqrt{\frac{BKxs^{\beta}}{n}} + \sqrt{\frac{BKx(r + \Delta)^{\beta}}{n}}$$

$$+ 2K\left(1 + \frac{1}{\alpha}\right)\frac{bx}{n} + \rho$$

then

$$\sup\left\{\mathbb{E}f - \mathbb{E}_n f : \mathbb{E}f = s\right\} - s > \sup\left\{\mathbb{E}f - \mathbb{E}_n f : r \leq \mathbb{E}f < r + \Delta\right\} - r + \rho$$

and thus

$$\inf\left\{\mathbb{E}_n f : \mathbb{E}f = s\right\} < \inf\left\{\mathbb{E}_n f : r \leq \mathbb{E}f < r + \Delta\right\} - \rho,$$

18

as claimed. ∎

Since the lemma compares $\xi_n(s) - s$ with $\xi_n(r, r+\Delta) - r$, we need to relate $\xi_n(r, r+\Delta)$ to $\xi_n(r)$. The following result will suffice, provided $r > 0$ and Δ is sufficiently small. (For the proof of the lower bound, it does not give a useful bound on $\xi_n(0, \Delta)$, and we need to deal with that case separately.)

Lemma 3.3 *If F is star-shaped around 0, then for every $r, \Delta > 0$,*

$$\xi_n(r) \leq \xi_n(r, r+\Delta) \leq \xi_n(r)\left(1 + \frac{\Delta}{r}\right).$$

Proof: The first inequality is immediate from the definitions. For the second, we can assume that $\xi_n(r, r+\Delta) > 0$. Fix $(X_1, \ldots, X_n)$, some $f \in F$ and $\delta > 0$ for which $r \leq \mathbb{E}f < r + \Delta$ and

$$\mathbb{E}f - \mathbb{E}_n f \geq \sup\{\mathbb{E}g - \mathbb{E}_n g : r \leq \mathbb{E}g < r + \Delta\} - \delta > 0.$$

Set $\tilde{f} = rf/\mathbb{E}f$ and note that $\mathbb{E}\tilde{f} = r$ and that

$$\sup\{\mathbb{E}g - \mathbb{E}_n g : \mathbb{E}g = r\} \geq \mathbb{E}\tilde{f} - \mathbb{E}_n \tilde{f} = \frac{r}{\mathbb{E}f}\left(\mathbb{E}f - \mathbb{E}_n f\right)$$

$$> \frac{r}{r + \Delta}\left(\mathbb{E}f - \mathbb{E}_n f\right)$$

$$> \frac{1}{1 + \Delta/r}\left(\sup\{\mathbb{E}g - \mathbb{E}_n g : r \leq \mathbb{E}g < r + \Delta\} - \delta\right).$$

The assertion follows by taking the expectation with respect to $X_1, \ldots, X_n$, and letting $\delta \to 0$. ∎

We are now ready to prove Theorem 3.1.

Proof: *(of Theorem 3.1) (1)* Fix $x > 0$, which might be different from the x of the theorem statement. First, Theorem 2.11 and the fact that $\rho \leq r'/2$ imply that, with probability at least $1 - e^{-x}$,

$$\mathbb{E}\hat{f} \leq r'.$$

Next, for any $\epsilon > 0$, if $r > \max\{1/n, r_{\epsilon,+}\}$, then $\xi_n(s) - s > \xi_n(r) - r + \epsilon$. Therefore, by Lemma 3.3,

$$\xi_n(s) - s > \xi_n(r, r+\Delta) - r + \epsilon - \frac{\Delta}{r}\xi_n(r).$$

Let

$$\epsilon_0 = \alpha \xi_n(s) + \left(\alpha \left(1 + \frac{\Delta}{r} \right) + \frac{\Delta}{r} \right) \xi_n(r)$$
$$+ \sqrt{\frac{BKx}{n}} \left(s^{\beta/2} + (r + \Delta)^{\beta/2} \right) + 2K \left(1 + \frac{1}{\alpha} \right) \frac{bx}{n} + \rho,$$

and fix $\Delta = \min\{\alpha/n, r' - r\}$, where $\alpha \le 1$ will be specified later.

For any $\epsilon \ge \epsilon_0$, Lemma 3.2 and Lemma 3.3 show that with probability at least $1 - 2e^{-x}$, we have $\mathbb{E}f \notin [r, r + \Delta)$. Since $r \ge 1/n$, then

$$\epsilon_0 \le \alpha(\xi_n(s) - s) + c\alpha(\xi_n(r) - r) + \alpha(s + cr)$$
$$+ \sqrt{\frac{BKx}{n}} \left(s^{\beta/2} + (r + \Delta)^{\beta/2} \right) + 2K \left(1 + \frac{1}{\alpha} \right) \frac{bx}{n} + \rho$$
$$\le c \left(\alpha(\xi_n(s) - s) + \alpha r' + \sqrt{\frac{Bxr'^\beta}{n}} + \frac{bx}{n\alpha} \right) + \rho.$$

Observe that by the definition of r', if we select

$$\alpha = \sqrt{\frac{bx}{n \max\{\xi_n(s) - s, r'\}}},$$

then $\alpha \le 1$, hence

$$\epsilon_0 \le c \max \left\{ \sqrt{\frac{bx(\xi_n(s) - s)}{n}}, \sqrt{\frac{bxr'}{n}}, \sqrt{\frac{Bxr'^\beta}{n}} \right\} + \rho.$$

Thus, we have shown that if ϵ satisfies the condition of the theorem, with probability at least $1 - 2e^{-x}$, we have $\mathbb{E}\hat{f} \notin [r, r + \Delta)$. To complete the proof, we repeatedly apply this result to a grid V of values of r, ranging from $\max\{1/n, r_{\epsilon,+}\}$ to r'. Clearly,

$$\log |V| \le \log \left\lceil \frac{r'}{\Delta} \right\rceil = \log \left\lceil \frac{r'n}{\alpha} \right\rceil \le c \log n,$$

and the result is evident by the union bound.

(2) We start by showing that $\mathbb{E}\hat{f}$ is probably outside the interval $[0, c_1/n)$. Indeed, since

$$\xi_n(0, c_1/n) < \sup_{s > 0} (\xi_n(s) - s) - \epsilon$$

20

and by Lemma 3.2, if

$$\epsilon \geq c \left(\alpha \left(\xi_n(s) - s \right) + \alpha r' + \sqrt{\frac{Bx}{n}} r'^\beta + \frac{bx}{\alpha n} \right) + \rho,$$

then with probability at least $1 - 2e^{-x}$, $\mathbb{E}\hat{f} \notin [0, c_1/n)$. The same argument as in part (1) shows that it suffices to choose

$$\epsilon \geq c \max \left\{ \sqrt{\frac{bx(\xi_n(s) - s)}{n}}, \sqrt{\frac{bxr'}{n}}, \sqrt{\frac{Bxr'^\beta}{n}} \right\} + \rho.$$

Next, we split the interval $[c_1/n, r_{\epsilon,-})$ into smaller intervals, $[r, r + \Delta)$, and show that $\mathbb{E}\hat{f}$ is unlikely to be in one of these intervals. For any $\delta > 0$, if $r \leq r_{\epsilon,-} - \delta$ then

$$\xi_n(s) - s > \xi_n(r) - r + \epsilon.$$

Since $c_1/n \leq r < r_{\epsilon,-} < r'$, we can use Lemmas 3.2 and 3.3 in the same way as in part (1). Letting δ approach zero completes the proof. ∎

4 Direct approach vs. structural results

Finally, we show that a direct analysis of the empirical minimization algorithm can yield much better estimates than the structural results presented in Section 2.2 under the assumptions we used throughout this article, namely, that F is a star-shaped class of uniformly bounded functions, which satisfies a Bernstein condition (and in particular, $\mathbb{E}f \geq 0$ for every $f \in F$).

Indeed, for every fixed integer n, we can construct a class and a probability measure for which every coordinate projection will not be an isomorphism (for any $0 < \epsilon < 1$) on the set $\{f : \mathbb{E}f \geq 1/4\}$, in the sense that for every sample $X_1, ..., X_n$, there will be a function f, with $\mathbb{E}_n f = 0$, but $\mathbb{E}f = 1/4$. (There is no magic in the number $1/4$; any sufficiently small positive constant will do.) Thus, any kind of a structural approach will only yield a trivial upper bound on $\mathbb{E}\hat{f}$. On the other hand, we will show that with probability larger than $1 - \delta$, we have $\mathbb{E}\hat{f} \leq 1/n$.

Although the theorem in this section is formulated as an existence result, it is clear that the structural results of Section 2 will be loose whenever the set of functions with expectation near zero is sufficiently rich. We use the following lemma to construct an example of such a function class.

Lemma 4.1 *For every positive integer n and all $m \geq 2(n^2 + n)$, the following holds. If P is the uniform probability measure on $\{1, ..., m\}$, then for every $\frac{1}{n} \leq \lambda \leq 1/2$ there exists a function class G_λ such that*

1. *For every $g \in G_\lambda$, $-1 \leq g(x) \leq 1$, $\mathbb{E}g = \lambda$ and $\mathbb{E}g^2 \leq 2\mathbb{E}g$.*

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

Also, there exist a function class H_λ such that

1. *For every $h \in H_\lambda$, $0 \leq h(x) \leq 1$, $\mathbb{E}h = \lambda$.*

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

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

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

Observe that if $m \geq n^2 + 2n$, then $0 < t \leq 2\lambda \leq 1$ for every I, J. Also, by the definition of t, $\mathbb{E}g_{I,J} = \lambda$. Next, note that

$$\mathbb{E}g^2 = \frac{1}{m}\left(|I| - |J\backslash I| + t^2(m - n) + 2|J\backslash I|\right) \leq \mathbb{E}g + \frac{2|J\backslash I|}{m}$$

$$\leq \mathbb{E}g + 2\frac{n}{m} < \mathbb{E}g + \frac{1}{n} \leq 2\mathbb{E}g,$$

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

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

$\blacksquare$

Theorem 4.2 *There is an absolute constant c for which the following holds. If $0 < \delta < 1$ and $n > N_0(\delta)$ there is a probability measure P and a star-shaped class F, which consists of functions bounded by 1 and has Bernstein type 1 with constant 2, such that*

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

2. For the class F, the function ξ_n satisfies

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

Thus, $\inf \{r > 0 : \xi_n(r) \le r/4\} = 1/4$.

3. If $\hat{f}$ is a ρ-approximate empirical minimizer, where $0 < \rho < 1/8$, then with probability larger than $1 - \delta$,

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

Proof: For any integer n, let m and P be as in Lemma 4.1, put $\tilde{F} = H_{1/4} \cup G_{1/n}$, and set $F = \text{star}(\tilde{F}, 0)$. Observe that $H_{1/4}$ consists of nonnegative functions and that $G_{1/n}$ is a Bernstein class of type 1 with constant 2. Thus, as a star-shaped hull of a Bernstein class, F has type 1 with a constant 2.

Next, we estimate the function $\xi_n(r)$ associated with F. Clearly, ξ_n vanishes for $r > 1/4$. For $r = 1/4$, and since $|\{X_1, ..., X_n\}| \le n$, there is a function in $H_{1/4}$ which is nonnegative, vanishes on $(X_1, ..., X_n)$, but its expectation is $1/4$. Thus, $\sup_{f \in F_{1/4}} \mathbb{E}f - \mathbb{E}_n f = 1/4$, and $\xi_n(1/4) = 1/4$. It is easy to see that for $1/n < r < 1/4$,

$$F_r = \{4rf : f \in H_{1/4}\},$$

and thus, on $(1/n, 1/4)$, $\xi_n(r) = r$. As for $r = 1/n$, recall that if $\tau \subset \{1, ..., m\}$, $|\tau| \le n$, then there is some $f \in F_{1/n}$ which is -1 on τ, implying that

$$\xi_n(1/n) = \mathbb{E} \sup \{\mathbb{E}f - \mathbb{E}_n f : f \in F_{1/n}\} \ge \frac{1}{n} + 1.$$

Clearly, this is also an upper bound on $\xi_n(1/n)$, and ξ_n decays linearly to 0 for $r < 1/n$.

We next consider the conditions of Theorem 3.1. It is easy to verify that $r' = 1/4$ if n is sufficiently large, and that

$$\sup_{s > 0} (\xi_n(s) - s) = 1.$$

Fix $0 < c_1 < 1/2$, let c be as in Theorem 3.1 and choose

$$\epsilon = c \left(\frac{\log(1/\delta) + \log n}{n}\right)^{1/2} + \rho.$$

23

Observe that $r_{\epsilon,+} = 1/n$ and that

$$r_{\epsilon,-} = \frac{1-\epsilon}{n} = \frac{1}{n}\left(1 - c\left(\frac{\log(n/\delta)}{n}\right)^{1/2} - \rho\right).$$

Thus, by Theorem 3.1, $\mathbb{E}\hat{f} \leq 1/n$, with probability at least $1 - \delta$. For the lower bound, note that $\xi_n(0, c_1/n) = c_1(1 + 1/n)$; hence for suitably large n $\xi_n(0, c_1/n) < \sup_{s>0}(\xi_n(s) - s) - \epsilon$, and by Theorem 3.1,

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

with probability at least $1 - \delta$. ■

A   Appendix: Proofs

Proof: *(of Lemma 1.1)* Suppose that, $a < b \in [0,1]$ satisfies $b - a = \sup\{f_1(x) - f_2(x) : x \in \mathcal{X}, f_1, f_2 \in F\}$ and $[a,b] \subseteq \{\ell(y, y_0) : y \in \mathcal{Y}\}$. Choose a mapping $u : [a,b] \to \mathcal{Y}$ such that $\ell(u(\alpha), y_0) = \alpha$ for $\alpha \in [a,b]$ (for example, u can be taken as a selection of the pre-image of $\ell(\cdot, y_0)$). For $f \in F$, define $g_f(x) = u(f(x) - \inf\{f'(x) : f' \in F\} + a)$, and note that $x \mapsto \ell(g_f(x), y_0)$ is measurable by assumption. Let $G = \{g_f : f \in F\}$ and set $g^* = g_0$. Therefore,

$$\{x \mapsto \ell(g(x), y_0) - \ell(g^*(x), y_0) : g \in G\}$$
$$= \{x \mapsto \ell(g_f(x), y_0) - \ell(g_0(x), y_0) : f \in F\}$$
$$= F.$$

If $\mathbb{E}f \geq 0$ for every $f \in F$ then clearly the choice $g^* = g_0$ minimizes $\mathbb{E}\ell(g(X), y_0)$, as claimed. ■

The proof of Theorem 2.3 uses the following lemma.

Lemma A.1 *For independent random variables $X_1, \ldots, X_n$, define*

$$Y = \sum_{i=1}^{n}(X_i - \mathbb{E}X_i)$$

and $\sigma^2 = \mathbb{E}Y^2$. If $|X_i| \leq 1$ and $\sigma^2 \geq 1$, then

$$\Pr\left(|Y| \geq \frac{\sigma}{2}\right) \geq c$$

for some universal constant c.

Proof: First we show that there is an absolute constant K such that

$$\mathbb{E}Y^2\chi_{\{|Y|\geq K\sigma\}} \leq \frac{\sigma^2}{4}.$$

Indeed, for every integer k,

$$\begin{aligned}
\mathbb{E}Y^2\chi_{\{|Y|\geq k\sigma\}} &= \sum_{m=k}^{\infty} \mathbb{E}Y^2\chi_{\{m\sigma\leq|Y|\leq(m+1)\sigma\}} \\
&\leq \sigma^2 \sum_{m=k}^{\infty}(m+1)^2 \Pr\{(|Y|\geq m\sigma\}) \\
&\leq 2\sigma^2 \sum_{m=k}^{\infty}(m+1)^2 e^{-3m/8},
\end{aligned}$$

where the last inequality follows from Bernstein's inequality and the fact that $\sigma^2 \geq 1$. Thus, the assertion follows by taking k sufficiently large.

Since $\mathbb{E}Y^2\chi_{\{|Y|\leq\sigma/2\}} \leq \sigma^2/4$, then

$$\begin{aligned}
\sigma^2 = \mathbb{E}Y^2 \\
&\leq \frac{\sigma^2}{4} + \mathbb{E}Y^2\chi_{\{\sigma/2\leq|Y|\leq K\sigma\}} + \frac{\sigma^2}{4} \\
&\leq \frac{\sigma^2}{2} + K^2\sigma^2 Pr\left(\left\{\frac{\sigma}{2}\leq|Y|\leq K\sigma\right\}\right),
\end{aligned}$$

and the result follows. ∎

Proof: *(of Theorem 2.3)* Without loss of generality, assume that $\sigma_F^2 = $ var$[g]$ for some $g \in F$. Let $Y = \sum_{i=1}^{n}(g(X_i) - \mathbb{E}g)$ and set $v = \mathbb{E}Y^2 = n\sigma_F^2$. By the assumption, $v \geq 1$, and thus, Lemma A.1 implies that

$$Pr\left(\left\{\|P-P_n\|_F \geq \frac{\sigma_F}{2\sqrt{n}}\right\}\right) \geq Pr\left(\left\{\frac{1}{n}|Y|\geq\frac{\sqrt{v}}{2n}\right\}\right) \geq c$$

for some absolute constant c. Integrating, $\mathbb{E}\|P-P_n\|_F \geq c\sigma_F/(2\sqrt{n})$. Since $n\sigma_F^2 \geq 1$,

$$\sigma_F\sqrt{\frac{x}{n}} + \frac{x}{n} \leq 2\sigma_F\sqrt{\frac{x}{n}} \leq \frac{1}{4\max\left\{3K,\sqrt{K}\right\}}\mathbb{E}\|P-P_n\|_F,$$

where K is the constant in Theorem 1.4, and the last inequality holds for an appropriate choice of x, which will be an absolute constant. The claim

now follows from Talagrand's inequality; by Theorem 1.4, with probability at least $1 - e^{-x}$ and selecting $\rho = 1/2$,

$$\|P_n - P\|_F \geq \frac{1}{2}\mathbb{E}\|P_n - P\|_F - \sqrt{K}\sigma_F\sqrt{\frac{x}{n}} - 3\frac{Kx}{n}$$

$$\geq \frac{1}{2}\mathbb{E}\|P_n - P\|_F - \max\left\{3K, \sqrt{K}\right\}\left(\sigma_F\sqrt{\frac{x}{n}} + \frac{x}{n}\right)$$

$$\geq \frac{1}{4}\mathbb{E}\|P_n - P\|_F.$$

■

References

[1] P. L. Bartlett, S. Boucheron, and G. Lugosi. Model selection and error estimation. *Machine Learning*, 48:85–113, 2002.

[2] P. L. Bartlett, O. Bousquet, and S. Mendelson. Localized Rademacher complexity. In *Proceedings of the Conference on Computational Learning Theory*, pages 44–58, 2002.

[3] Peter L. Bartlett, Michael I. Jordan, and Jon D. McAuliffe. Large margin classifiers: convex loss, low noise, and convergence rates. In *Advances in Neural Information Processing Systems, 16*, 2003. (To appear. http://www.stat.berkeley.edu/~bartlett/papers/bjm-lmcclln-03.ps.gz).

[4] Anselm Blumer, A. Ehrenfeucht, David Haussler, and Manfred K. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *Journal of the ACM*, 36(4):929–965, 1989.

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

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

[7] Evarist Giné, Vladimir Koltchinskii, and Jon A. Wellner. Ratio limit theorems for empirical processes. In Evarist Giné, Christian Houdré, and David Nualart, editors, *Statistical Inequalities and Applications*, pages 249–278. 2003.

[8] D. Haussler. Sphere packing numbers for subsets of the Boolean n-cube with bounded Vapnik-Chervonenkis dimension. *Journal of Combinatorial Theory, Series A*, 69(2):217–232, 1995.

[9] W. B. Johnson and G. Schechtman. Finite dimensional subspaces of ℓ_p. In W.B. Johnson and J. Lindenstrauss, editors, *Handbook of the Geometry of Banach Spaces, Vol 1*. North Holland, 2001.

[10] T. Klein. Une inégalité de concentration gauche pour les processus empiriques. [A left concentration inequality for empirical processes]. *C. R. Math. Acad. Sci. Paris*, 334(6):501–504, 2002.

[11] V. Koltchinskii. Rademacher penalties and structural risk minimization. *IEEE Transactions on Information Theory*, 47(5):1902–1914, July 2001.

[12] V. I. Koltchinskii and D. Panchenko. Rademacher processes and bounding the risk of function learning. In Evarist Giné, David M. Mason, and Jon A. Wellner, editors, *High Dimensional Probability II*, volume 47, pages 443–459. Birkhäuser, 2000.

[13] Vladimir Koltchinskii. Local Rademacher complexities and oracle inequalities in risk minimization. Technical report, University of New Mexico, 2003.

[14] M. Ledoux. *The Concentration of Measure Phenomenon*, volume 89. AMS, 2001.

[15] W. S. Lee, P. L. Bartlett, and R. C. Williamson. Efficient agnostic learning of neural networks with bounded fan-in. *IEEE Transactions on Information Theory*, 42(6):2118–2132, 1996.

[16] Gábor Lugosi and Marten Wegkamp. Complexity regularization via localized random penalties. *Annals of Statistics*, 32(4):to appear, 2004.

[17] P. Massart. About the constants in Talagrand's concentration inequality. *Annals of Probability*, 28:863–885, 2000.

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

[19] Shahar Mendelson. Improving the sample complexity using global data. *IEEE Transactions on Information Theory*, 48(7):1977–1991, 2002.

[20] Shahar Mendelson. Rademacher averages and phase transitions in Glivenko-Cantelli classes. *IEEE Transactions on Information Theory*, 48(1):251–263, 2002.

[21] Shahar Mendelson. A few notes on statistical learning theory. In S. Mendelson and A. J. Smola, editors, *Advanced Lectures in Machine Learning*, volume 2600 of *Lecture Notes in Computer Science*, pages 1–40. Springer, 2003.

[22] Shahar Mendelson. On the performance of kernel classes. *Journal of Machine Learning Research*, 4:759–771, 2003.

[23] Shahar Mendelson. Geometric parameters in learning theory. *Geometric aspects of Functional Analysis (GAFA Seminar Notes)*. Lecture Notes in Mathematics 1850, 193-236, 2004.

[24] E. Rio. Inégalités de concentration pour les processus empiriques de classes de parties [Concentration inequalities for set-indexed empirical processes]. *Probability Theory and Related Fields*, 119(2):163–175, 2001.

[25] M. Talagrand. Sharper bounds for Gaussian and empirical processes. *Annals of Probability*, 22:28–76, 1994.

[26] S. van de Geer. A new approach to least-squares estimation, with applications. *Annals of Statistics*, 15:587–602, 1987.

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

[28] A. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998.

[29] Aad W. van der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes*. Springer, 1996.

[30] Vladimir 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.