ℓ_1-regularized linear regression: Persistence and oracle inequalities

Peter Bartlett[1,2], Shahar Mendelson[3,4] and Joseph Neeman*[1]

March 31, 2009

Abstract

We study the predictive performance of ℓ_1-regularized linear regression, including the case where the number of covariates is substantially larger than the sample size. We introduce a new analysis method that does not require uniformly bounded covariates, an assumption that was often necessary with previous techniques. This technique provides an answer to a conjecture of Greenshtein and Ritov [12] regarding the "persistence" rate for linear regression and allows us to prove an oracle inequality for the error of the regularized minimizer.

1 Introduction

In this article we study the problem of linear regression with an ℓ_1 constraint or with an ℓ_1 regularization. In the ℓ_1 constraint case, one considers a random variable $(X, Y) \in \mathbb{R}^d \times \mathbb{R}$ of which one has n independent samples $X_1, Y_1, \ldots, X_n, Y_n$. For a fixed $b > 0$, the ℓ_1 constraint regression produces $\hat{\beta}$ defined by

$$\hat{\beta} = \operatorname{argmin}_{\{\beta \in \mathbb{R}^d : \|\beta\|_{\ell_1^d} \leq b\}} \sum_{i=1}^{n} \left(\langle X_i, \beta \rangle - Y_i \right)^2 . \tag{1.1}$$

This regression is known as "lasso" regression and it is often motivated by the fact that it tends to select solutions $\hat{\beta}$ that are sparse [29] (that is, it selects some $\hat{\beta} \in \mathbb{R}^d$ with considerably fewer than d non-zero coordinates),

[1]Department of Statistics, University of California, Berkeley, CA 94720, USA.
[2]Computer Science Division, University of California, Berkeley, CA 94720, USA.
[3]Centre for Mathematics and its Applications, The Australian National University, Canberra, ACT 0200, Australia.
[4]Department of Mathematics, Technion, I.I.T, Haifa 32000, Israel.

particularly when compared with least squares or with ℓ_2-regularized (or "ridge") regression. Naturally, from a practical point of view, sparsity is desirable because it allows for fast computation of $\langle X, \hat{\beta} \rangle$ on future samples.

In the standard setup, the dimension d and the ℓ_1^d radius of the set of constraints b are fixed, while the sample size n grows to infinity. In this case, the quality of the prediction of the empirical minimizer $\hat{\beta}$ is determined by an appropriate notion of complexity of the set $\{\beta \in \mathbb{R}^d : \|\beta\|_{\ell_1^d} \leq b\}$.

A more interesting problem is what happens when the dimension (the number of explanatory variables) and the radius increase with the sample size. Motivated by many practical prediction problems, including those that arise in microarray data analysis and natural language processing, this problem has been extensively studied in recent years. The results can be divided into two categories: those that study the predictive power of $\hat{\beta}$ [9, 30, 12] and those that study its sparsity pattern and reconstruction properties [4, 32, 18, 19, 17, 8]; this article falls into the first of these categories.

Unfortunately, thus far there have been no satisfactory bounds on the way the error of the empirical minimizer $\hat{\beta}$ depends on the radius b and the dimension d; the existing estimates in the case where b and d are allowed to grow to infinity with the sample size n were rather loose. The main aim of this article is to identify the predictive power of $\hat{\beta}$ as a function of all three parameters b, d, and n.

A notable difficulty in studying this problem arises from the fact that linear functions are unbounded. And although the problem of empirical minimization in a fixed function class has been extensively studied already (see, for example, [3, 6] and the references therein), the most satisfactory results apply only to function classes that are bounded almost surely. The problem is made more difficult by our use of the quadratic loss, which takes away some potential sources of uniform regularity (for example, we cannot rely on a bounded Lipschitz constant). Thus, the techniques used in the bounded case break down completely in the unbounded, quadratic setting.

Our main contribution is a method, based on Talagrand's "generic chaining," [28] that allows us to avoid the problems arising in the unbounded quadratic case, under some mild assumptions.

Traditionally, in the study of empirical minimization, one separates the risk $P\ell_{\hat{\beta}}(X, Y)$ into two quantities:

$$P\ell_{\hat{\beta}}(X,Y) = \left(P\ell_{\hat{\beta}}(X,Y) - \inf_{\beta \in bB_1^d} P\ell_\beta(X,Y) \right) + \inf_{\beta \in bB_1^d} P\ell_\beta(X,Y) \quad (1.2)$$

where we use the abbreviation $\ell_\beta(X, Y) = (\langle X, \beta \rangle - Y)^2$. The first of these

quantities is called the *sample error* and it measures the success of empirical minimization relative to the best function in our hypothesis class, bB_1^d; the second quantity is called the *approximation error* and it describes the performance of bB_1^d without regard to the sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ that powered our empirical minimization. There is a well-known conflict between these two errors: we can reduce the approximation error by enlarging our function class (that is, by increasing b or d), but doing so will increase the sample error. For a fixed b and d, the sample error shrinks as the sample size increases but the approximation error remains constant. If one wishes to minimize the total error, therefore, one should not consider a fixed radius b and dimension d but rather increasing sequences (b_n) and (d_n), each element of which is chosen to minimize the error (1.2) for that particular sample size.

Here, we will establish almost sharp estimates (up to the exact power of a log factor) on the error of the empirical minimizer in $\{x \in \mathbb{R}^d : \|x\|_{\ell_1^d} \leq b\}$ as a function on the radius b, the dimension d and the sample size n, under mild assumptions on (X, Y). For example, we will show that if μ is an isotropic (that is, its covariance structure coincides with the Euclidean one), log-concave measure on $\mathbb{R}^d$ and $Y \in L_2$, then up to poly-logarithmic factors in b, d and n, the error of the empirical minimizer is bounded by

$$\min\left\{\frac{b^2}{n} + \frac{d}{n}, \frac{b}{\sqrt{n}}\left(1 + \frac{b}{\sqrt{n}}\right)\right\}.$$

An outcome of these estimates is a solution to the question of *persistence*, posed by Greenshtein and Ritov [12], which is defined as follows. Let $(d_n)_{n=1}^{\infty}$ be an increasing sequence, consider a sequence of measures $(\mu_n)_{n=1}^{\infty}$ on $\mathbb{R}^{d_n} \times \mathbb{R}$ and suppose that for every n, one is given n independent samples $(X_1, Y_1), \ldots, (X_n, Y_n)$ drawn according to μ_n. For $\beta \in \mathbb{R}^d$, denote the squared loss of β by $\ell_\beta(x, y) = (\langle x, \beta \rangle - y)^2$. Fix some increasing sequence b_n and consider, for every n, the empirical minimizer in $b_n B_1^{d_n}$:

$$\hat{\beta}_n = \operatorname{argmin}_{\|\beta\|_{\ell_1^{d_n}} \leq b_n} \sum_{i=1}^{n} \ell_\beta(X_i, Y_i).$$

The sequence $\hat{\beta}_n$ is called *persistent* if

$$P_{(X,Y)\sim\mu_n} \ell_{\hat{\beta}_n}(X, Y) - \inf_{\beta_n \in \mathbb{R}^{d_n}} P_{(X,Y)\sim\mu_n} \ell_{\beta_n}(X, Y) \to 0,$$

in probability.

Empirical minimization gives a persistent sequence $(\hat{\beta}_n)$ provided that the sequences (b_n) and (d_n) do not increase too rapidly. Greenshtein and Ritov asked for the most quickly increasing sequence (b_n) such that empirical minimization is persistent. Under the assumption that d_n is at most polynomial in n (and under some conditions on μ_n), they showed that one can take $b_n = o((n/\log(n))^{1/4})$. They also, however, proved persistence for $b_n = o((n/\log(n))^{1/2})$ in the case of Gaussian measures μ_n and showed that this was the best possible rate in the Gaussian case. They asked whether it was possible to improve the persistence result in the non-Gaussian case under the condition (on the sequence μ_n) that $\|X\|_{\ell_\infty^{d_n}}$ be bounded almost surely. We answer this question in the affirmative (up to the exact power of the logarithm) under even milder assumptions on μ_n. We should point out that not only do we improve the persistence rates, we also remove the restriction that d_n be polynomial in n; in fact, our result is non-trivial for sequences almost as large as $d_n \sim \exp(\sqrt{n})$.

To formulate the result we will need the following assumption.

Assumption 1.1 *For every μ_n on $\mathbb{R}^{d_n}$ set $X^{(n)} = (X_1^{(n)}, ..., X_{d_n}^{(n)})$ to be a vector distributed according to μ_n. Assume that there is an absolute constant c such that for every integer n, every $1 \leq i \leq d_n$ and every $t \geq 1$,*

$$Pr\left(|X_i^{(n)}| \geq t\right) \leq 2\exp(-ct).$$

In other words, Assumption 1.1 states that all the coordinates of each $X^{(n)}$ (explanatory variable) are subexponential with uniform constant c.

Theorem 1.1 *Suppose that (d_n) is an increasing sequence and that (μ_n) satisfy Assumption 1.1. Then empirical minimization is persistent provided that*

$$b_n = o\left(\frac{\sqrt{n}}{\log^{3/2} n \cdot \log^{3/2}(nd_n)}\right).$$

Alternatively, suppose $|X_i^{(n)}| \leq C$ almost surely for every $n \in \mathbb{N}$ and every $1 \leq i \leq d_n$. Then empirical minimization is persistent provided that

$$b_n = o\left(\frac{\sqrt{n}}{\log^{3/2} n \cdot \log^{1/2}(nd_n)}\right).$$

We have made no particular effort to optimize the powers of the logarithms in Theorem 1.1 and that we do not believe them to be best possible. Also, observe that this estimate implies an almost optimal bound

for a slightly different question: the case where the constraint set consists of sparse vectors in $\mathbb{R}^{d_n}$. In this case, instead of performing the empirical minimization in the set of vectors of ℓ_1^d norm at most b, one performs minimization in the set $T_{k,d}$, which is the convex hull of vectors in $\mathbb{R}^d$ of Euclidean norm at most 1 that are supported on at most k coordinates. If we denote the Euclidean unit ball by B_2^d and the unit ball in ℓ_1^d by B_1^d, then clearly $T_{k,d} \subset \sqrt{k}B_1^d \cap B_2^d$. Thus, according to Theorem 1.1, one may allow for the number k_n of non-zero coordinates to grow as quickly as $k_n = b_n^2$ and have the analog of the persistence property for the empirical minimizer in $T_{k_n,d}$.

Theorem 1.1 provides an answer to the question posed by Greenshtein and Ritov, but it does not directly address our original question. Recalling the tradeoff between sample and approximation error, our mission was to choose a sequence (b_n) that exploits this tradeoff to minimize the error. Persistence does not give the best sequence (b_n); it gives a rate of increase which is not too fast, but it does not provide information on which of the slower sequences are the best. In other words, the fact that one can select $b_n \ll \sqrt{n}$ yields no information on whether $b_n = n^{1/4}$ is better than $b_n = n^{1/3}$. What's more, choosing the right sequence (b_n) requires some knowledge of the behavior of the approximation error as the radius our the ℓ_1^d ball increases.

One way of addressing this problem is to consider a modified minimization problem. Redefine $\hat{\beta}$ using the ℓ_1 regularization

$$\hat{\beta} = \operatorname{argmin}_{\beta \in \mathbb{R}^d} \left(\frac{1}{n} \sum_{i=1}^{n} \ell_\beta(X_i, Y_i) + \lambda_n \|\beta\|_{\ell_1^d} \right). \tag{1.3}$$

This sort of regularization is well-known [29] and has been extensively studied in the past. A fact we shall require is that if the *regularization parameter* λ_n is chosen carefully then the solution to (1.3) is almost as good as the solution to (1.1) for the best choice of b [1]. Furthermore, it is possible to choose such a λ_n without knowing how the approximation error behaves. Intuitively, this is not so surprising. The heart of the matter is that $\inf_{\{\|\beta\|_1 \le b\}} \frac{1}{n} \sum_i \ell_\beta(X_i, Y_i)$ is likely to be close to the approximation error in bB_1^d. Thus, if the approximation error decreases quickly as b increases the regularized ℓ_1 problem is likely to choose $\hat{\beta}$ with a relatively large ℓ_1^d norm. On the other hand, if the approximation error decreases slowly, (1.3) will select $\hat{\beta}$ of a small ℓ_1^d norm. Thus (1.3) is somehow equivalent to choosing the radius in (1.1) to depend on the approximation error.

The method of analysis that we use for this problem requires, unfor-

tunately, a uniform concentration phenomenon (though it is likely that it could be avoided using a deviation argument by using uniform tail estimates rather than a concentration argument). Therefore, one has to make stronger assumptions on the measures μ_n, namely that both $|Y|$ and $\|X\|_{\ell_\infty^d}$ are bounded almost surely by a constant independent of n.

Theorem 1.2 *There exist absolute constants c and c' for which the following holds. Let $(d_n)_{n\geq 1}$ be any increasing sequence with $\log d_n = o(n)$ and let $(\mu_n)_{n\geq 1}$ be a family of measures on $\mathbb{R}^{d_n}$. For $n \geq 1$, suppose that X is distributed according to μ_n, that $\|X\|_{\ell_\infty^{d_n}} \leq M$ almost surely and that Y is a real-valued random variable with $|Y| \leq M$ almost surely. If we define*

$$\lambda_n = cM^2 \frac{\log^{3/2} n \cdot \log^{1/2}(d_n n)}{\sqrt{n}}$$

then for all sufficiently large n (depending on d_n and M), with probability at least $1 - \exp(-\log^3 n \cdot \log(d_n n))$, for any

$$\hat{\beta} \in \operatorname{argmin}_{\beta\in\mathbb{R}^{d_n}} (P\ell_\beta + \lambda_n \|\beta\|_{\ell_1^d}),$$

we have

$$P\ell_{\hat{\beta}} \leq \inf_{\beta\in\mathbb{R}^{d_n}} (P\ell_\beta + c'\lambda_n(1 + \|\beta\|_{\ell_1^d})).$$

2 Preliminaries

In this section we will present the basic definitions and results that we require. Throughout, all absolute constants (that is, positive numbers that are independent of the other parameters of the problem) are denoted by $C, C_1, ..., c, c_1$ etc. Their value may change from line to line.

We will consider a Euclidean structure on $\mathbb{R}^d$; $|x|$ is the Euclidean norm of x. We shall abuse notation and denote the absolute value and the cardinality of a set by $|\cdot|$ as well.

A subset of a vector space is called symmetric if the fact that x is in the set implies that $-x$ is also in the set. It is a well known fact that if $K \in \mathbb{R}^d$ is convex and symmetric and has a nonempty interior then it defines a norm on $\mathbb{R}^d$ by $\|x\|_K = \inf\{t : t^{-1}x \in K\}$. For every $1 \leq p \leq \infty$ and integer d, B_p^d is the unit ball in ℓ_p^d, that is, $B_p^d = \{x : \sum_{i=1}^d |x_i|^p \leq 1\}$.

A significant part of our work will be devoted to the study of the supremum of a collection of random variables, where each one of them is naturally associated with a point in $\mathbb{R}^d$. This is an example of a rather general idea: to study the supremum of a family of random variables indexed by a metric space using the metric structure of the set.

Definition 2.1 *A process $\{Z_t : t \in T\}$ indexed by a metric space (T, d) is called subgaussian with respect to the metric d if for every $x, y \in T$ and every $t \geq 1$*

$$Pr\left(|Z_x - Z_y| \geq td(x,y)\right) \leq 2\exp(-t^2/2).$$

Two examples of subgaussian process are the following. Let $T \subset \mathbb{R}^d$ and for every $x \in T$ consider the two random variables

$$G_x = \sum_{i=1}^{d} g_i x_i, \quad \text{and} \quad Z_x = \sum_{i=1}^{d} \varepsilon_i x_i,$$

where $(g_i)_{i=1}^d$ are independent standard Gaussian random variables and $(\varepsilon_i)_{i=1}^d$ are independent, symmetric-$\{-1, 1\}$ valued random variables. It is standard to verify that both $\{G_x : x \in T\}$ and $\{Z_x : x \in T\}$ are subgaussian processes with respect to the Euclidean metric on $\mathbb{R}^d$. For the Gaussian process this is evident because G_x is distributed as $g_1|x|$. For the Rademacher process $\{Z_x : x \in T\}$, it is simply a reformulation of Höffding's inequality, that for every $x \in \mathbb{R}^d$ and every $t > 0$,

$$Pr\left(\left|\sum_{i=1}^{d} \varepsilon_i x_i\right| \geq t|x|\right) \leq 2\exp(-t^2/2).$$

When a random process $\{Z_t : t \in T\}$ is subgaussian with respect to a metric d, one can use the generic chaining mechanism to control the random variable $\sup_{t\in T} |Z_t|$ using the so-called γ-functionals.

Definition 2.2 *[28] For a metric space (T, d), an admissible sequence of T is a collection of subsets of T, $\{T_s : s \geq 0\}$, such that for every $s \geq 1$, $|T_s| = 2^{2^s}$ and $|T_0| = 1$. Define the γ_2 functional by*

$$\gamma_2(T, d) = \inf \sup_{t\in T} \sum_{s=0}^{\infty} 2^{s/2} d(t, T_s),$$

where the infimum is taken with respect to all admissible sequences of T.

The following theorem is a corollary of the chaining mechanism. For results of an almost identical flavor and a comprehensive survey on generic chaining and its applications we refer the reader to [28].

Theorem 2.3 *There exist absolute constants c_1 and c_2 for which the following holds. Let $\{Z_t : t \in T\}$ be a subgaussian process with respect to the metric d. Then, for every $u \geq 1$ and any $t_0 \in T$,*

$$Pr\left(\sup_{t\in Z} |Z_t - Z_{t_0}| \geq c_1 u \gamma_2(T,d)\right) \leq 2\exp(-u^2/2).$$

In particular,

$$\mathbb{E}\sup_{t\in T} |Z_t - Z_{t_0}| \leq c_2 \gamma_2(T,d).$$

A straightforward way (though it often leads to suboptimal results) to construct an admissible sequence is using covers of the metric space (T,d). Let $N(\varepsilon,T,d)$ be the smallest number of open balls (with respect to the metric d) needed to cover T. The corresponding set of centers is called an ε-cover of T. For every integer s, let $\varepsilon_s = \inf\{\varepsilon : N(\varepsilon,T,d) \leq 2^{2^s}\}$, and let T_s be a minimal ε_s cover of T. Then using this admissible sequence one can show (see, for example, [28]), that there is an absolute constant c such that

$$\gamma_2(T,d) \leq c\int_0^{\mathrm{diam}(T,d)} \sqrt{\log N(\varepsilon,T,d)}d\varepsilon,$$

that is, the γ_2 functional may be bounded using an appropriate entropy integral.

In our analysis we will be interested in empirical processes. Let F be a class of functions on a probability space (Ω,μ) and let $X_1,...,X_n$ be distributed according to μ. Consider the process indexed by F, given by $Z_f = n^{-1}\sum_{i=1}^n f(X_i) - \mathbb{E}f$ and denote

$$\|P_n - P\|_F = \sup_{f\in F}\left|\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}f\right|$$

and

$$\mathbb{E}\|P_n - P\|_F = \mathbb{E}\sup_{f\in F}\left|\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}f\right|.$$

Unfortunately, under typical assumptions on the class F and the measure μ an empirical process is not subgaussian. Indeed, by Bernstein's inequality (e.g. [31]) which is sharp is many cases, a typical tail behavior of the random variable $n^{-1}\sum_{i=1}^n f(X_i) - \mathbb{E}f$ is a mixture of subgaussian and subexponential tails. One of the possible ways around this problem is to use symmetrization arguments, due to Giné and Zinn [10]; the resulting metric is a random one.

Theorem 2.4 *Let F be a class of functionals on (Ω, μ). Then,*

$$\mathbb{E}\|P_n - P\|_F \leq \frac{2}{n}\mathbb{E}_X\mathbb{E}_\varepsilon \sup_{f\in F}\left|\sum_{i=1}^n \varepsilon_i f(X_i)\right|,$$

where $(\varepsilon_i)_{i=1}^n$ are independent, symmetric $\{-1,1\}$-valued random variables.

Theorem 2.4 implies that estimating the expectation of the supremum of the empirical process indexed by F is reduced to bounding the expectation of the supremum of the Rademacher process (which is subgaussian with respect to $|\cdot|$) of a typical coordinate projection of F,

$$P_\sigma F = \{(f(X_i))_{i=1}^n : f \in F\}.$$

The best understood situations in Learning Theory are when the given function class is bounded in L_∞ (see, for example [3]). Such problems are much simpler than the general, unbounded one for two reasons. The first is that the random variable $\|P_n - P\|_F$ is concentrated around its mean $\mathbb{E}\|P_n - P\|_F$ in the uniformly bounded case. This concentration result was established by Talagrand [27].

The second reason why the unbounded case is much harder is because it often rules out the use of contraction inequalities, which are standard tools in empirical process theory. Contraction inequalities are used in the context of learning as a way of relating the complexity of the loss class (in our case, $\{\ell_\beta = \langle\beta,\cdot\rangle^2 : \beta \in T\}$) to that of the base class $\{\langle\beta,\cdot\rangle : \beta \in T\}$. Since contraction inequalities are not valid without a Lipschitz bound, one has to find other ways of controlling the complexity of an unbounded loss class, which will be the main topic of the next section.

3 Error rates for linear functionals on $\mathbb{R}^d$

The situation we study here is when the class of functions consists of linear functionals $\{\langle t,\cdot\rangle : t \in T\}$, where $T \subset \mathbb{R}^d$ is a convex symmetric set. In this section, we will develop an estimate on the error of the empirical minimizer in T, via an "isomorphic" bound, as will be explained below. This bound, applied to the set $T = bB_1^d = \{\beta \in \mathbb{R}^d : \|x\|_{\ell_1^d} \leq b\}$ will yield a sharp estimate (up to logarithmic factors in b, d and n) on the performance of the empirical minimization algorithm in bB_1^d.

Let us introduce the following notation. Let μ be a probability measure on $\mathbb{R}^d$ and consider an unknown (real-valued) random variable Y. Let $T \subset$

$\mathbb{R}^d$ be a compact, convex, symmetric set and to each $\beta \in T$ associate the function $f_\beta = \langle \beta, \cdot \rangle : \mathbb{R}^d \to \mathbb{R}$. Recall that our goal is to estimate the random variable Y by an element in T (i.e. by a function f_β where $\beta \in T$) with respect to the squared loss, using empirical data, which is a random sample $(X_i, Y_i)_{i=1}^n$ according to the joint distribution of μ and Y.

Set $F = \{\langle \beta, \cdot \rangle : \beta \in T\}$, let $\ell(x, y) = (x - y)^2$ and for every $f \in F$, define $\ell_f = \ell(f(X), Y)$ to be the squared loss associated with f and Y.

Note that if $\mathbb{E}\|X\|_{\ell_2^d} < \infty$ then $F \subset L_2$ is compact and since T is convex, F is a convex class of functions. One can show that in such a case, $\mathbb{E}\ell(f(X), Y)$ has a unique minimizer in F, and we will denote it by $f^* = f_{\beta^*}$, where $\beta^* \in T$ (note that β^* is not unique if the measure μ is supported on a subspace of $\mathbb{R}^d$; our analysis, however, only requires the uniqueness of f_{β^*}). Thus, we can define the excess loss function associated with f by $\mathcal{L}_f = \ell_f - \ell_{f^*}$ and the excess loss class

$$\mathcal{L}_F = \{\ell_f - \ell_{f^*} : f \in F\}.$$

For the sake of simplicity, we shall sometimes abuse notation and write $\mathcal{L}_\beta$ and ℓ_β for $\mathcal{L}_{f_\beta}$ and ℓ_{f_β}, respectively.

It is clear that our problem is how to obtain an estimate on the conditional expectation

$$\hat{R} = \mathbb{E}\left(\mathcal{L}_{\hat{f}} | (X_i, Y_i)_{i=1}^n\right),$$

as a function of the sample size n that holds with high probability.

The function class $\mathcal{L}_F$ has certain properties that will be used in our analysis. First of all, for every $f \in F$, $\mathbb{E}\mathcal{L}_f \geq 0$ and equality holds only for f^*. The second property we require is more delicate. To formulate it, define for any $\lambda \geq 0$,

$$\mathcal{L}_{F,\lambda} = \{\mathcal{L}_f : \mathbb{E}\mathcal{L}_f \leq \lambda\}.$$

Lemma 3.1 *Let $F \subset L_2$ be a compact, convex set of functions and Let $\mathcal{L}_F$ be the squared loss class. Then, for any $\lambda > 0$*

$$\mathcal{L}_{F,\lambda} \subset \{\mathcal{L}_f : \mathbb{E}|f - f^*|^2 \leq \lambda\}.$$

Lemma 3.1 ensures that if $\mathbb{E}\mathcal{L}_f$ is "small" then f must be relatively close to the true minimizer f^* with respect to the $L_2(\mu)$ norm.

This lemma appeared implicitly in several places (see, for example [21], Cor. 3.4 and [2], Lemma 14) and in more general situations (for example, loss functions that are uniformly convex rather than the squared loss). We present the proof of Lemma 3.1 for the sake of completeness.

Proof. Let $\mathbb{E}\mathcal{L}_f \leq \lambda$. Then,

$$\begin{aligned}\lambda &\geq \mathbb{E}\left(f(X) - f^*(X)\right) \cdot \left(f(X) + f^*(X) - 2Y\right) \\ &= \left\langle f(X) - f^*(X), f(X) + f^*(X) - 2Y \right\rangle,\end{aligned}$$

where the inner product here is in L_2 with respect to the joint probability distribution of X and Y, which is denoted by ν. Since f^* is the best approximation to Y in F with respect to the $L_2(\nu)$ and since F is convex, then by the characterization of the nearest point in a compact, convex subset in an inner product space, $\left\langle f^* - Y, f - f^* \right\rangle \geq 0$. Hence,

$$\begin{aligned}\left\langle f(X) - f^*(X), f(X) + f^*(X) - 2Y \right\rangle &= \|f - f^*\|^2_{L_2(\nu)} + \left\langle f^* - Y, f - f^* \right\rangle \\ &\geq \|f - f^*\|^2_{L_2(\nu)} = \mathbb{E}_X |f - f^*|^2.\end{aligned}$$

■

It is well known [16] that one way of obtaining an estimate on the error of the empirical minimizer is by finding a small λ (that depends on n) such that with high probability, for every $f \in F$ with $\mathbb{E}\mathcal{L}_f \geq \lambda$,

$$(1 - \varepsilon)\mathbb{E}\mathcal{L}_f \leq \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}_f(X_i, Y_i) \leq (1 + \varepsilon)\mathbb{E}\mathcal{L}_f.$$

Hence, for functions with large error, the empirical error and the actual one are ε-equivalent, and this fact implies that the random and deterministic structures are "almost isometric" (or for a fixed ε, isomorphic). In particular, the empirical minimizer cannot be a function of large error: if it were then this "almost isometric" structure would apply and we could conclude from $\sum_{i=1}^{n} \mathcal{L}_{\hat{f}}(X_i, Y_i) \leq 0$ that $\mathbb{E}\mathcal{L}_{\hat{f}} \leq 0$.

If one wishes to find a small λ for which one can find an isomorphic condition with, for example, $\varepsilon = 1/2$, this can be achieved by controlling $\mathbb{E}\|P_n - P\|_{G_\lambda}$ where

$$G_\lambda = \{\theta\mathcal{L}_f : f \in F,\ 0 \leq \theta \leq 1,\ P(\theta\mathcal{L}_f) = \lambda\}.$$

Observe that G_λ is the localization at level λ of the star-shaped hull of $\mathcal{L}_F$.

The fact that if $\mathbb{E}\|P_n - P\|_{G_\lambda} \leq \alpha\lambda$ for some $0 < \alpha < 1$ then with high probability, the risk of the empirical minimization algorithm is at most $c(\alpha)\lambda$ was first noted in [3] for cases in which one has a strong concentration phenomenon for $\|P_n - P\|_{G_\lambda}$ around its expectation. In fact, one can obtain the same result without the strong concentration if one is willing to have confidence that is less than exponential.

Theorem 3.2 *Let $\{\mathcal{L}_f : f \in F\}$ be an excess loss class with respect to some loss function ℓ and set*

$$G_\lambda = \{\theta\mathcal{L}_f : 0 \leq \theta \leq 1,\ P(\theta\mathcal{L}_f) = \lambda\}\,.$$

If $\mathbb{E}\|P_n - P\|_{G_\lambda} \leq \delta\lambda$ then with probability at least $1 - 2\delta$, the conditional expectation $\mathbb{E}(\mathcal{L}_{\hat{f}}|X_1, Y_1, \ldots, X_n, Y_n) \leq \lambda$.

Proof. By rewriting G_λ as

$$G_\lambda = \{\theta\mathcal{L}_f : 0 \leq \theta \leq 1,\ P(\theta\mathcal{L}_f) = \lambda\} = \left\{\frac{\lambda\mathcal{L}_f}{\mathbb{E}\mathcal{L}_f} : \mathbb{E}\mathcal{L}_f \geq \lambda\right\}, \tag{3.1}$$

it is evident that

$$\sup_{\{\mathcal{L}_f : \mathbb{E}\mathcal{L}_f \geq \lambda\}} \left|\frac{n^{-1}\sum_{i=1}^n \mathcal{L}_f(X_i, Y_i) - \mathbb{E}\mathcal{L}_f}{\mathbb{E}\mathcal{L}_f}\right| = \frac{\|P_n - P\|_{G_\lambda}}{\lambda}.$$

By Markov's inequality, with probability at least $1 - 2\delta$,

$$\frac{\|P_n - P\|_{G_\lambda}}{\lambda} \leq \frac{1}{2\delta\lambda}\mathbb{E}\|P_n - P\|_{G_\lambda} \leq \frac{1}{2}.$$

This gives an isomorphic condition on $\{\mathcal{L}_f : \mathbb{E}\mathcal{L}_f \geq \lambda\}$: by (3.1), with probability at least $1 - 2\delta$, for all $\mathcal{L}_f$ with $\mathbb{E}\mathcal{L}_f \geq \lambda$,

$$\frac{1}{2}\mathbb{E}\mathcal{L}_f \leq \frac{1}{n}\sum_{i=1}^n \mathcal{L}_f(X_i, Y_i) \leq \frac{3}{2}\mathbb{E}\mathcal{L}_f.$$

Since the loss function of the empirical minimizer, $\mathcal{L}_{\hat{f}}$, does not satisfy this inequality (because $\sum_{i=1}^n \mathcal{L}_{\hat{f}}(X_i, Y_i) \leq 0$), then $\mathbb{E}\mathcal{L}_{\hat{f}} \leq \lambda$, as claimed. ∎

Given a class of functions F and a sample $\sigma = (X_i, Y_i)_{i=1}^n$, recall that $P_\sigma F$ is the coordinate projection of F onto σ, that is,

$$P_\sigma F = \{(f(X_i, Y_i))_{i=1}^n : f \in F\} \subset \mathbb{R}^n.$$

A key part of our analysis is to bound the Rademacher process indexed by coordinate projections of $\mathcal{L}_{F,\lambda}$ which, by symmetrization, leads to the desired bound on $\mathbb{E}\|P_n - P\|_{G_\lambda}$. Recall that by Höffding's inequality [15], if $A \subset \mathbb{R}^n$ then the Rademacher process indexed by A, given by $x \to \sum_{i=1}^n \varepsilon_i x_i = Z_x$ is subgaussian with respect to the Euclidean metric.

Consider the L_2 metric endowed on the parameter space $\mathbb{R}^d$ by the covariance structure $\|\beta\|_{L_2}^2 = \mathbb{E}|\langle X, \beta\rangle|^2$ and denote its unit ball by D. Thus, $D = \{x \in \mathbb{R}^d : \mathbb{E}|\langle X, x\rangle|^2 \leq 1\}$.

The following lemma allows one to control the Rademacher process indexed by $P_\sigma \mathcal{L}_{F,\lambda}$ using the distances between the indexing parameters in $\mathbb{R}^d$. This enables one to overcome the difficulty arising from the fact that $\mathcal{L}_{f_\beta}$ is a shift of $\langle \beta, \cdot \rangle^2$, which leads to a process that is very different and considerably more difficult to handle than the one indexed by the linear functionals $\langle \beta, \cdot \rangle$.

Lemma 3.3 *For every $\sigma = (X_i, Y_i)_{i=1}^n$ the Rademacher process indexed by $P_\sigma \mathcal{L}_{F,\lambda}$ is subgaussian with respect to a metric on T, defined by*

$$d(\beta_1, \beta_2) = 4\|\beta_1 - \beta_2\|_{\infty,n} \left(\sup_{v \in \sqrt{\lambda} D \cap 2T} \sum_{i=1}^n \langle X_i, v\rangle^2 + \sum_{i=1}^n \ell_{f^*}(X_i, Y_i) \right)^{1/2} \tag{3.2}$$

where $\|\beta_1 - \beta_2\|_{\infty,n} = \max_{1 \leq i \leq n} |\langle X_i, \beta_1 - \beta_2\rangle|$.

In other words, $d(\beta_1, \beta_2)$ is the random ℓ_∞ distance, multiplied by what is essentially the empirical ℓ_2 diameter of the localized set $\sqrt{\lambda} D \cap 2T$.

Proof. Denote $\|g\|_{\ell_2^n}^2 = \sum_{i=1}^n g^2(X_i, Y_i)$ and observe that for every $v, u \in \mathbb{R}^d$,

$$\|\mathcal{L}_{f_u} - \mathcal{L}_{f_v}\|_{\ell_2^n}^2 = \|\ell_{f_u} - \ell_{f_v}\|_{\ell_2^n}^2 = \sum_{i=1}^n \langle X_i, u - v\rangle^2 \left(\langle X_i, u + v\rangle - 2Y_i \right)^2.$$

Recall that $\beta^* \in T$ is the element for which $\inf_{\beta \in T} \mathbb{E}\ell_{f_\beta}$ is attained. Then by Lemma 3.1,

$$\begin{aligned} \{v \in T : \mathcal{L}_{f_v} \in \mathcal{L}_{F,\lambda}\} &\subset \{v \in T : \|v - \beta^*\|_{L_2} \leq \sqrt{\lambda}\} = T \cap (\beta^* + \sqrt{\lambda} D) \\ &\subset \beta^* + (2T \cap \sqrt{\lambda} D), \end{aligned}$$

where the last inequality follows from the convexity and symmetry of T and using the notation $a + B = \{a + b : b \in B\}$.

In particular, if $u, v \in T$ and $\|v - \beta^*\|_{L_2}, \|u - \beta^*\|_{L_2} \leq \sqrt{\lambda}$ then

$$(u + v)/2 - \beta^* \in 2T \cap \sqrt{\lambda} D.$$

Thus, for every $\mathcal{L}_u, \mathcal{L}_v \in \mathcal{L}_{F,\lambda}$,

$$\|\mathcal{L}_u - \mathcal{L}_v\|_{\ell_2^n}^2 = \sum_{i=1}^n \big\langle X_i, u-v\big\rangle^2 \left(\big\langle X_i, u+v\big\rangle - 2Y_i\right)^2 \tag{3.3}$$

$$\leq \max_{1\leq i\leq n} |\big\langle X_i, u-v\big\rangle|^2 \cdot 4\sum_{i=1}^n \left(\big\langle X_i, \frac{u+v}{2} - \beta^*\big\rangle + \left(\big\langle X_i, \beta^*\big\rangle - Y_i\right)\right)^2$$

$$\leq 8\|u-v\|_{\infty,n}^2 \left(\sup_{t\in 2T\cap\sqrt{\lambda}D} \sum_{i=1}^n \big\langle X_i, t\big\rangle^2 + \sum_{i=1}^n \ell_{\beta^*}(X_i, Y_i)\right),$$

where the last inequality follows from $\|a+b\|^2 \leq \|a+b\|^2 + \|a-b\|^2$. Hoeffding's inequality implies the result. ∎

The next step is to bound the random diameter

$$\left(\sup_{t\in 2T\cap\sqrt{\lambda}D} \sum_{i=1}^n \big\langle X_i, t\big\rangle^2\right)^{1/2}$$

from above using the random ℓ_∞ metric. To simplify notation, set for a given sample $(X_1, ..., X_n)$ the random metric

$$d_{\infty,n}(f,g) = \max_{1\leq i\leq n} |f(X_i) - g(X_i)|,$$

and for a class of functions F let

$$U_n(F) = (\mathbb{E}\gamma_2^2(F, d_{\infty,n}))^{1/2} \text{ and } \sigma_F = (\sup_{f\in F} \mathbb{E}f^2(X))^{1/2}.$$

The following is a result from [14].

Theorem 3.4 *There exists an absolute constant c for which the following holds. Let F be a class of functions on (Ω, μ), let X be distributed according to μ and set $X_1, \ldots, X_n$ to be independent copies of X. Then,*

$$\mathbb{E}\sup_{f\in F}\left|\sum_{i=1}^n (f^2(X_i) - \mathbb{E}f^2(X))\right| \leq c\max\left(\sqrt{n}\sigma_F U_n(F),\, U_n^2(F)\right). \tag{3.4}$$

In particular,

$$\mathbb{E}\sup_{t\in 2T\cap\sqrt{\lambda}D} \sum_{i=1}^n \big\langle t, X_i\big\rangle^2 \leq n\lambda + c\max\left(\sqrt{n\lambda}U_n(F),\, U_n^2(F)\right). \tag{3.5}$$

Thus, the dominating term in the expectation of the worst deviation of $n^{-1}\sum_{i=1}^n f^2(X_i)$ from the mean $\mathbb{E}f^2$ can be upper bounded in terms of the L_2 norm of $\gamma_2(F, d_{\infty,n})$.

The following theorem is the key technical result. In using the notation $U_n(K)$ for a set $K \subseteq \mathbb{R}^d$, we identify K with the class of functions $\{\langle x, \cdot \rangle : x \in K\}$.

Theorem 3.5 *There exists an absolute constant c for which the following holds. For every convex and symmetric $T \subset \mathbb{R}^d$ and every $\lambda > 0$,*

$$\mathbb{E}\|P_n - P\|_{\mathcal{L}_{F,\lambda}} \le c\frac{U_n(K)}{\sqrt{n}} \cdot \left(\lambda + P\ell_{\beta^*} + \sqrt{\lambda}\frac{U_n(K)}{\sqrt{n}} + \frac{U_n^2(K)}{n}\right)^{1/2},$$

where $K = 2T \cap \sqrt{\lambda}D$.

Proof. By the Giné-Zinn symmetrization theorem, Lemma 3.1 and the definition of the L_2 metric on $\mathbb{R}^d$ endowed by X,

$$\mathbb{E}\|P_n - P\|_{\mathcal{L}_{F,\lambda}} \le \mathbb{E}\mathbb{E}_\varepsilon \sup_{\beta \in W} \frac{2}{n}\left|\sum_{i=1}^n \varepsilon_i \mathcal{L}_{f_\beta}(X_i, Y_i)\right| = (*),$$

where

$$W = \{\beta \in T : \|\beta - \beta^*\|_{L_2} \le \sqrt{\lambda}\} \subset \beta^* + (2T \cap \sqrt{\lambda}D)$$

and $D = \{x \in \mathbb{R}^d : \mathbb{E}|\langle x, X\rangle|^2 \le 1\}$.

By Lemma 3.3, for every fixed sample $(X_i, Y_i)_{i=1}^n$, this Rademacher process is subgaussian with respect to the metric d defined in that lemma. Thus, by the generic chaining mechanism, setting $K = 2T \cap \sqrt{\lambda}D$,

$$\begin{aligned}
(*) &\le \frac{c_1}{n}\mathbb{E}\left(\gamma_2(\beta^* + K, d_{\infty,n})\left(\sup_{t\in K}\sum_{i=1}^n \langle t, X_i\rangle^2 + \sum_{i=1}^n \ell_{\beta^*}(X_i, Y_i)\right)^{1/2}\right) \\
&= \frac{c_1}{n}\mathbb{E}\left(\gamma_2(K, d_{\infty,n})\left(\sup_{t\in K}\sum_{i=1}^n \langle t, X_i\rangle^2 + \sum_{i=1}^n \ell_{\beta^*}(X_i, Y_i)\right)^{1/2}\right), \\
&\le \frac{c_1}{\sqrt{n}}\left(\mathbb{E}\gamma_2^2(K, d_{\infty,n})\right)^{1/2} \cdot \left(\mathbb{E}\sup_{t\in K}\frac{1}{n}\sum_{i=1}^n \langle X_i, t\rangle^2 + \mathbb{E}\ell_{\beta^*}\right)^{1/2},
\end{aligned}$$

where the first equality is evident because the metric $d_{\infty,n}$ is translation invariant, and thus $\gamma_2(\beta^* + K, d_{\infty,n}) = \gamma_2(K, d_{\infty,n})$, and the last inequality is the Cauchy-Schwarz inequality. The claim now follows from Equation (3.5). ∎

Note that the bound that we have established thus far is for $P\|P_n - P\|_{\mathcal{L}_{F,\lambda}}$ where $\mathcal{L}_{F,\lambda} = \{\mathcal{L}_f : \mathbb{E}\mathcal{L}_f \leq \lambda\}$ for any $\lambda > 0$. To control $\mathbb{E}\|P_n - P\|_{G_\lambda}$ we require an additional "peeling" argument, following the same path as in [22].

To simplify notation, define

$$\phi_n(x) = \frac{U_n(K_x)}{\sqrt{n}} \cdot \left(x + P\ell_{\beta^*} + \sqrt{x}\frac{U_n(K_x)}{\sqrt{n}} + \frac{U_n^2(K_x)}{n} \right)^{1/2},$$

where $K_x = 2T \cap \sqrt{x}D$.

Theorem 3.6 *There exist absolute constants c_1, c_2 and c_3 for which the following holds. For every $\lambda > 0$,*

$$\mathbb{E}\|P_n - P\|_{G_\lambda} \leq c_1 \sum_{i=0}^{\infty} 2^{-i} \phi_n(2^{i+1}\lambda).$$

In particular, for every $\lambda > 0$

$$\mathbb{E}\|P_n - P\|_{G_\lambda} \leq c_2 \frac{U_n(T)}{\sqrt{n}} \cdot \left(\lambda + \mathbb{E}\ell_{\beta^*} + \sqrt{\lambda}\frac{U_n(T)}{\sqrt{n}} + \frac{U_n^2(T)}{n} \right)^{1/2},$$

and thus $\mathbb{E}\|P_n - P\|_{G_\lambda} \leq \delta\lambda$ provided that

$$\lambda \geq \frac{c_3}{\delta^2} \max\left\{ \frac{U_n(T)}{\sqrt{n}} \sqrt{\mathbb{E}\ell_{\beta^*}}, \frac{U_n^2(T)}{n} \right\}.$$

Proof. Observe that for every $\lambda > 0$,

$$G_\lambda = \left\{ \frac{\lambda\mathcal{L}_f}{\mathbb{E}\mathcal{L}_f} : P\mathcal{L}_f \geq \lambda \right\} = \bigcup_{i=0}^{\infty} \left\{ \frac{\lambda\mathcal{L}_f}{\mathbb{E}\mathcal{L}_f} : 2^i\lambda \leq \mathbb{E}\mathcal{L}_f \leq 2^{i+1}\lambda \right\}.$$

Hence, setting $H_i = \left\{ \frac{\lambda\mathcal{L}_f}{\mathbb{E}\mathcal{L}_f} : 2^i\lambda \leq \mathbb{E}\mathcal{L}_f \leq 2^{i+1}\lambda \right\}$, then

$$\begin{aligned}
\mathbb{E}\|P_n - P\|_{G_\lambda} &\leq \sum_{i=0}^{\infty} \mathbb{E}\|P_n - P\|_{H_i} \\
&\leq \sum_{i=0}^{\infty} 2^{-i} \mathbb{E} \sup_{\{\mathcal{L}_f : 2^i\lambda \leq \mathbb{E}\mathcal{L}_f \leq 2^{i+1}\lambda\}} \left| \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}_f(X_i, Y_i) - \mathbb{E}\mathcal{L}_f \right| \\
&\leq \sum_{i=0}^{\infty} 2^{-i} \mathbb{E}\|P_n - P\|_{\mathcal{L}_{F,2^{i+1}\lambda}} \\
&\leq c_1 \sum_{i=0}^{\infty} 2^{-i} \phi_n(2^{i+1}\lambda),
\end{aligned}$$

where the last inequality is evident from Theorem 3.5.

The second and third claims follow using the fact that

$$2T \cap \sqrt{2^{i+1}\lambda} D \subset 2T$$

and a straightforward computation. ∎

Combining Theorem 3.2 and Theorem 3.6 with the trivial bound $\mathbb{E}\ell_{\beta^*} \leq \|Y\|_{L_2}^2$, one obtains the following.

Corollary 3.7 *There exists an absolute constant c for which the following holds. Let $T \subset \mathbb{R}^d$ be as above and set $\hat{\beta} \in T$ the parameter selected by the empirical minimization algorithm. Then, for all $0 < \delta \leq 1/2$, with probability at least $1 - 2\delta$,*

$$P\left(\mathcal{L}_{\hat{\beta}} | (X_i, Y_i)_{i=1}^n\right) \leq \frac{c}{\delta^2} \max\left\{ \frac{U_n(T)}{\sqrt{n}} \|Y\|_{L_2}, \frac{U_n^2(T)}{n} \right\}.$$

Thus, to obtain an estimate on the risk of the empirical minimization algorithm, all that one has to do is to bound $U_n(T)$, which, in the case we are interested in, is $U_n(bB_1^d)$. Observe that an estimate on $U_n(bB_1^d)$ would suffice to handle the two cases considered in [12]; for the first, the indexing set is $T = b_n B_1^{d_n}$, and for the second, T is the convex hull of vectors of Euclidean norm at most a that are supported on at most k coordinates. Indeed, the second case follows from the first one: if $|t| \leq a$ is supported on at most k coordinates then by the Cauchy-Schwarz inequality, $\|t\|_{\ell_1^d} \leq \sqrt{k}|t| \leq a\sqrt{k}$. Hence, if T is the convex hull of the set of such vectors then $T \subset a\sqrt{k}B_1^d$.

Remark 3.8 *Corollary 3.7 and the second and third parts of Theorem 3.6 follow from the trivial estimate that $K_x \subset 2T$, which is rather loose unless T is very small. The fact that the complexity of the indexing set is governed by the intersections $2T \cap \sqrt{x}D$ is one of the benefits gained by the localization argument and becomes more significant the larger T is. For the case that interests us, when $T = bB_1^d$, it turns out that for a wide range of choices of $d = d(n)$ and $b = b(n)$ one may safely replace $bB_1^d \cap \sqrt{\lambda}D$ with bB_1^d, and bounding $U_n(bB_1^d)$ is enough to obtain a sharp estimate (up to logarithmic factors) in the problem addressed in [12]. However, when $d \ll n$, $bB_1^d \cap \sqrt{\lambda}D$ is better approximated by $\sqrt{\lambda}D$, as will be explained in Section 4.1.*

4 Empirical minimization is persistent

Based on the results of the previous section, it is evident that if one wishes to prove that empirical minimization is persistent, it suffices to control

$\gamma_2(bB_1^d, d_{\infty,n})$ for every $X_1, ..., X_n$. To that end, we shall use a covering estimate and upper bound γ_2 using an entropy integral.

The idea behind the following result appeared in [14] but we will present a detailed proof for the sake of completeness.

Lemma 4.1 *There exists an absolute constant c for which the following holds. For every $b > 0$,*

$$\gamma_2(bB_1^d, d_{\infty,n}) \leq cbQh(n,d),$$

where $Q = \max_{1\leq i\leq n} \|X_i\|_{\ell_\infty^d}$ and $h(d,n) = \log^{3/2} n \max\{\log^{1/2} d, \log^{1/2} n\}$.

Proof. First, assume that $d \geq n$. Fix $X_1, ..., X_n \in \mathbb{R}^d$, define

$$H_n = \{u : \max_{1\leq i\leq n} |\langle u, X_i\rangle| \leq 1\}$$

and let $\| \ \|_{H_n}$ be the quasi-norm on $\mathbb{R}^d$ whose unit ball is H_n.

Consider the operator $A : \ell_1^n \to \mathbb{R}^d$ defined by $Ae_i = X_i$ and observe that the number of translates of εH_n needed to cover B_1^d, denoted by $N(B_1^d, \varepsilon H_n)$, satisfies

$$N(B_1^d, \varepsilon H_n) = N(A^* B_1^d, \varepsilon B_\infty^n).$$

Indeed, this is the case because $u \in H_n$ if and only if $A^*u \in B_\infty^n$.

Recall that for an operator $A : X \to Y$ between the normed spaces X and Y, the ℓ-entropy number of A is given by

$$e_\ell(A) = \inf\{\varepsilon > 0 : N(AB_X, \varepsilon B_Y) \leq 2^\ell\},$$

where B_X and B_Y are the unit balls in X and Y respectively. By a well known result of Carl [5], if $A : \ell_1^n \to \ell_\infty^d$ then for $\ell \leq n \leq d$,

$$e_\ell(A^*) \leq c_1 \|A^*\|_{\ell_1^d \to \ell_\infty^n} \left(\frac{\log(1+n/\ell)\cdot\log(1+d/\ell)}{\ell}\right)^{1/2},$$

and clearly, $\|A^*\| = \|A\| = \max_{1\leq i\leq n} \|X_i\|_{\ell_\infty^d} \equiv Q$.

Therefore, since $n \leq d$, then for every

$$\varepsilon > c_2 Qb\sqrt{\frac{\log d}{n}} \equiv \varepsilon_0,$$

$$\log N(bB_1^d, \varepsilon H_n) \leq c_3 \frac{b^2 Q^2 \log d \cdot \log n}{\varepsilon^2}.$$

Using a standard volumetric estimate (see, for example, [26] Chapter 5), for every $\varepsilon \leq \varepsilon_0$

$$\begin{aligned}\log N(bB_1^d, \varepsilon H_n) &\leq \log N(bB_1^d, \varepsilon_0 H_n) + \log N(\varepsilon_0 H_n, \varepsilon H_n) \\ &\leq c_3 \frac{b^2 Q^2}{\varepsilon_0^2} \log d \cdot \log n + n \log\left(1 + \frac{\varepsilon_0}{\varepsilon}\right) \\ &\leq c_4 n \left(\log\left(1 + \frac{\varepsilon_0}{\varepsilon}\right) + \log n\right).\end{aligned}$$

Also,

$$\sup_{v \in bB_1^d} \|v\|_{H_n} = b \max_{1 \leq j \leq d} \max_{1 \leq i \leq n} |\langle e_j, X_i \rangle| = b \max_{1 \leq i \leq n} \|X_i\|_{\ell_\infty^d} = bQ.$$

By an entropy integral argument, it is evident that

$$\begin{aligned}\gamma_2(bB_1^d, d_{\infty,n}) &\leq c_5 \int_0^\infty \sqrt{\log N(bB_1^d, \varepsilon H_n)} d\varepsilon = c_5 \int_0^{bQ} \sqrt{\log N(bB_1^d, \varepsilon H_n)} d\varepsilon \\ &\leq c_6 \left(\int_0^{\varepsilon_0} \sqrt{n \log\left(1 + \frac{\varepsilon_0}{\varepsilon}\right)} + \int_{\varepsilon_0}^{bQ} \frac{bQ\sqrt{\log d \cdot \log n}}{\varepsilon} d\varepsilon \right) \\ &\leq c_7 \left(\sqrt{n \log n}\, \varepsilon_0 + bQ \sqrt{\log d \cdot \log n} \log\left(\frac{bQ}{\varepsilon_0}\right) \right) \\ &\leq c_8 bQ \sqrt{\log d} \cdot (\log n)^{3/2}.\end{aligned}$$

as claimed.

If $n \geq d$ then $B_1^d \subset B_1^n$, and one can extend each $X_i \in \mathbb{R}^d$ to $X_i \oplus 0 \in \mathbb{R}^n$. Now the bound is as before, but with d replaced by n. ∎

Let us mention that we have made no effort to optimize the dependency of γ_2 on n and d, since our estimates yield a poly-logarithmic dependency in those parameters. Using a much more delicate approach—a construction of an appropriate admissible sequence of T rather than by an entropy integral argument, as was done in [13]—the power of the logarithms can be reduced (though not completely eliminated).

We will consider two families of measures on $\mathbb{R}^d$. The first is when the random variable $\|X\|_{\ell_\infty^d}$ is bounded in L_∞, and the second is when X is selected according to a measure satisfying that for every $1 \leq i \leq d$, the tail of $|\langle X, e_i \rangle|$ decays quickly. A natural example of such vectors are those distributed according to log-concave measures.

Definition 4.2 *A measure μ on $\mathbb{R}^d$ is called log-concave if for all nonempty and measurable sets $A, B \subset \mathbb{R}^d$ and every $0 \leq \lambda \leq 1$,*

$$\mu(\lambda A + (1-\lambda)B) \geq \mu^{\lambda}(A)\mu^{1-\lambda}(B).$$

A measure μ on $\mathbb{R}^d$ is called isotropic if for every $\theta \in S^{d-1}$,

$$\mathbb{E}\langle X, \theta\rangle^2 = 1,$$

where X is distributed according to μ.

There are many natural examples of log-concave measures, and their study is a central part of modern asymptotic geometric analysis. Among the examples are measures that have a log-concave density, the volume measure on a convex body, and many others.

A central part of our analysis will be based on decay properties of random variables. A useful way of quantifying these properties is using Orlicz norms.

Definition 4.3 *Let Y be a random variable. For $\alpha \geq 1$ define the α-Orlicz norm of Y by*

$$\|Y\|_{\psi_\alpha} = \inf\left\{C > 0 : \mathbb{E}\exp\left(\frac{|Y|^\alpha}{C^\alpha}\right) \leq 2\right\}.$$

For basic facts regarding Orlicz norms we refer the reader to [7, 31]. It is standard to verify that if Y has a bounded ψ_α norm then for every $t \geq 1$,

$$Pr\left(|Y| \geq t\right) \leq 2\exp(-t^\alpha/\|Y\|_{\psi_\alpha}^\alpha).$$

The reverse direction is also true: if Y has a tail bounded by $\exp(-t^\alpha/K^\alpha)$ then $\|Y\|_{\psi_\alpha} \leq c_1 K$. Also, if Y has a bounded ψ_α norm (denoted $Y \in L_{\psi_\alpha}$) then for every $p > 1$, $\|Y\|_{L_p} \leq cp^{1/\alpha}\|Y\|_{\psi_\alpha}$.

A well known fact that follows from Borell's inequality (see, e.g. [23], Appendix III) is that if μ is log-concave and if X is distributed according to μ, then for every $x \in \mathbb{R}^d$,

$$\|\langle X, x\rangle\|_{\psi_1} \leq c\|\langle X, x\rangle\|_{L_1}, \tag{4.1}$$

where c is an absolute constant. In particular, the moments of linear functionals satisfy

$$\|\langle X, x\rangle\|_{L_p} \leq cp\|\langle X, x\rangle\|_{L_1}.$$

Lemma 4.4 *There exists an absolute constant c for which the following holds. Let μ be a measure on $\mathbb{R}^d$ and suppose that $X_1, ..., X_n$ are independent and distributed according to μ. Then*

$$\left(\mathbb{E} \max_{1 \le i \le n} \|X_i\|_{\ell_\infty^d}^2 \right)^{1/2} \le c \log(nd) \cdot \max_{1 \le j \le d} \|\langle X, e_j \rangle\|_{\psi_1}.$$

Therefore, if μ is log-concave then

$$\left(\mathbb{E} \max_{1 \le i \le n} \|X_i\|_{\ell_\infty^d}^2 \right)^{1/2} \le c \log(nd) \cdot \max_{1 \le j \le d} \|\langle X, e_j \rangle\|_{L_2},$$

and if μ is log-concave and isotropic then

$$\left(\mathbb{E} \max_{1 \le i \le n} \|X_i\|_{\ell_\infty^d}^2 \right)^{1/2} \le c \log(nd).$$

Proof. Recall the well-known observation due to Pisier (see, e.g. [31]) that if $Z_1, ..., Z_m$ are random variables then

$$\| \max_{1 \le i \le m} Z_i \|_{\psi_1} \le c_1 \max_{1 \le i \le m} \|Z_i\|_{\psi_1} \log m,$$

where c_1 is an absolute constant.

Since $Q = \max_{1 \le i \le n} \|X_i\|_{\ell_\infty^d} = \max_{i,j} |\langle X_i, e_j \rangle|$ then

$$\|Q\|_{L_2} \le c_2 \|Q\|_{\psi_1} \le c_3 \log(nd) \max_{1 \le j \le d} \|\langle X, e_j \rangle\|_{\psi_1}.$$

If μ is log-concave,

$$\max_{1 \le j \le d} \|\langle X, e_j \rangle\|_{\psi_1} \le c_4 \max_{1 \le j \le d} \|\langle X, e_j \rangle\|_{L_1} \le c_4 \max_{1 \le j \le d} \|\langle X, e_j \rangle\|_{L_2},$$

by (4.1) and Jensen's inequality. If, in addition, μ is isotropic, then

$$\max_{1 \le j \le d} \|\langle X, e_j \rangle\|_{L_2} = \max_{1 \le j \le d} \|e_j\| = 1.$$

■

We are now ready to formulate the first error rate estimate for $T = bB_1^d$, which follows directly from Lemmas 4.1 and 4.4.

Theorem 4.5 *There exists an absolute constant c for which the following holds. Set $h(n,d) = \log^{3/2} n \cdot \log^{3/2}(nd)$ and $\rho = \max_{1\leq j\leq d} \|\langle X, e_j\rangle\|_{\psi_1}$.*

If $T = bB_1^d$ then with probability at least $1-2\delta$, any empirical minimizer $\hat{\beta}$ satisfies

$$\mathbb{E}\mathcal{L}_{\hat{\beta}} \leq \frac{c}{\delta^2} \max\left\{ \frac{bh\rho}{\sqrt{n}} \cdot \sqrt{\mathbb{E}\ell_{\beta^*}}, \frac{b^2h^2\rho^2}{n} \right\}. \tag{4.2}$$

If $\|X\|_{\ell_\infty^d}$ is bounded almost surely by U then (4.2) holds with $\rho = cU$ and $h(n,d) = \log^{3/2} n \cdot \log^{1/2}(nd)$. If X is distributed according to a log-concave measure then (4.2) holds with $\rho = \max_{1\leq j\leq d} \|\langle X, e_j\rangle\|_{L_2}$, and if μ is distributed according to a measure that is both log-concave and isotropic then (4.2) holds with $\rho = 1$.

As an example, let μ_n be a family of isotropic, log-concave measures on $\mathbb{R}^{d_n}$ and assume that $d_n \sim n^\alpha$ for some $\alpha > 1$. Observe that $\mathbb{E}\ell_{\beta^*} \leq \mathbb{E}Y^2$, which is dimension independent. Then, as long as

$$\frac{b_n^2 \log^6 n}{n} \to 0,$$

the empirical minimizer is persistent.

More generally, let us formulate an estimate on the optimal choice of the parameters b_n and k_n as promised in Theorem 1.1. Recall that $T_{k,d}$ is the convex hull of vectors in $\mathbb{R}^d$ of Euclidean norm at most 1 that are supported on at most k coordinates, and therefore $T_{k,d} \subset \sqrt{k}B_1^d$.

Corollary 4.6 *Let T_n be either $b_nB_1^{d_n}$ or $T_{b_n^2,d_n}$.*

1. *Set b_n and d_n to satisfy*

$$\lim_{n\to\infty} \frac{b_n}{\sqrt{n}} \log^{3/2} n \cdot \log^{3/2}(nd_n) = 0$$

and let (μ_n) be a sequence of measures on $\mathbb{R}^{d_n}$. If μ_n is isotropic and log-concave for every n, then for every random variable $Y \in L_2$ the empirical minimization algorithm on T_n is persistent. More generally, if each coordinate of $X^{(n)} \sim \mu_n$ is sub-exponential with a constant not depending on n, then for every random variable $Y \in L_2$ the empirical minimization algorithm on T_n is persistent.

2. *Alternatively, if $\|X\|_{\ell_\infty^d}$ is bounded almost surely and $Y \in L_2$, then the empirical minimization algorithm on T_n is persistent for b_n and d_n satisfying*

$$\lim_{n\to\infty} \frac{b_n}{\sqrt{n}} \log^{3/2} n \cdot \log^{1/2}(nd_n) = 0.$$

4.1 Error rates for empirical minimization

Theorem 1.1 gives the optimal bound (up to logarithmic factors) on the rate at which one may "expand" the dimension d and the radius of B_1^d and still obtain a persistent algorithm. However, if one expands d and b at that rate, the resulting error rate is arbitrarily slow. Of course, Theorem 4.5, is stronger in the sense that it yields an estimate on the error rate for each choice of triplet (b, d, n), but a careful look at the estimate established there shows that it is suboptimal for certain triplets. For example, for fixed values of b and d that do not grow with n, one would expect an error rate that is roughly of the order of $1/n$ rather than $1/\sqrt{n}$. The reason for that looseness in Theorem 4.5 is that it was implicitly assumed in the proof that $bB_1^d \cap \sqrt{\lambda}D$ is essentially equivalent to $2bB_1^d$, enabling us to replace one with the other. However, if b and d are constant with respect to n, then in the isotropic case ($D = B_2^d$), $bB_1^d \cap \sqrt{\lambda}B_2^d = \sqrt{\lambda}B_2^d$ as long as $\lambda \leq b^2/d$. Hence, if there is any hope that the error rate $\lambda_n \to 0$ then one should approximate $bB_1^d \cap \sqrt{\lambda}B_2^d$ by $\sqrt{\lambda}B_2^d$ rather than by bB_1^d.

Let us mention that in certain cases (e.g. if X is an isotropic, Gaussian vector) one can prove sharp bounds for the "complexity" of the interpolation body given by $(\mathbb{E}\gamma_2^2(bB_1^d \cap \sqrt{\lambda}D, d_{\infty,n}))^{1/2}$ for all values of n, b, d and λ (see [11]). This analysis shows that the gap between the exact estimates and the bound given by considering the two "extreme" cases of bB_1^d and $\sqrt{\lambda}D$ is logarithmic in the parameters b, d and n. Since the analysis of the complexity of the interpolation body even in the Gaussian case is technically involved and its gains are rather minimal we shall not present it here. Instead, we will now consider the other extreme case, in which one replaces $bB_1^d \cap \sqrt{\lambda}D$ by $\sqrt{\lambda}D$. Our starting point is a modified version of Theorem 3.6. To formulate it, recall that if $T \subset \mathbb{R}^d$ and $\beta \in T$ then $\mathcal{L}_{f_\beta}$ is the excess loss associated with the parameter β, and $G_\lambda = \{\lambda\mathcal{L}_f / \mathbb{E}\mathcal{L}_f : \mathbb{E}\mathcal{L}_f \geq \lambda\}$.

Theorem 4.7 *There exists an absolute constant c for which the following holds. If*

$$\lambda \geq \frac{c}{\delta^2} \max\left\{ \frac{U_n^2(T)}{n}, \frac{U_n^2(D)}{n} \mathbb{E}\ell_{\beta^*} \right\}$$

then $\mathbb{E}\|P_n - P\|_{G_\lambda} \leq \delta\lambda$. In particular,

$$P\left(\mathcal{L}_{\hat{\beta}} | (X_i, Y_i)_{i=1}^n \right) \leq \frac{c}{\delta^2} \max\left\{ \frac{U_n^2(T)}{n}, \frac{U_n^2(D)}{n} \mathbb{E}\ell_{\beta^*} \right\}$$

with probability at least $1 - 2\delta$.

Proof. Recall that

$$\phi_n(x) = \frac{U_n(K_x)}{\sqrt{n}} \cdot \left(x + \mathbb{E}\ell_{\beta^*} + \sqrt{x}\frac{U_n(K_x)}{\sqrt{n}} + \frac{U_n^2(K_x)}{n} \right)^{1/2},$$

where $K_x = 2T \cap \sqrt{x}D$, and that by Theorem 3.6,

$$\mathbb{E}\|P_n - P\|_{G_\lambda} \leq c_1 \sum_{i=0}^{\infty} 2^{-i} \phi_n(2^{i+1}\lambda).$$

Set $A_i = U_n(K_{2^{i+1}\lambda})/\sqrt{n}$ and thus,

$$2^{-i}\phi_n(2^{i+1}\lambda) \leq c_2 \left(2^{-i/2} \left(A_i \lambda^{1/2} + A_i^{3/2}\lambda^{1/4} + A_i^2 \right) + 2^{-i} A_i \left(\mathbb{E}\ell_{\beta^*} \right)^{1/2} \right)$$
$$\leq c_2 \left(2^{-i/2} \left(\frac{U_n(T)}{\sqrt{n}}\lambda^{1/2} + \left(\frac{U_n(T)}{\sqrt{n}} \right)^{3/2} \lambda^{1/4} + \left(\frac{U_n(T)}{\sqrt{n}} \right)^2 \right) + 2^{-i} \frac{U_n(D)}{\sqrt{n}} \left(2^{i+1}\lambda \mathbb{E}\ell_{\beta^*} \right)^{1/2} \right),$$

where we used $K_{2^{i+1}\lambda} \subset 2^{(i+1)/2}\sqrt{\lambda}D$ for the last term and $K_{2^{i+1}\lambda} \subset T$ for all the others.

Summing over i, it is evident that $\sum_{i=0}^{\infty} 2^{-i}\phi_n(2^{i+1}\lambda)$ is at most

$$c_3 \left(\frac{U_n(T)}{\sqrt{n}}\lambda^{1/2} + \frac{U_n(D)}{\sqrt{n}}(\lambda \mathbb{E}\ell_{\beta^*})^{1/2} + \left(\frac{U_n(T)}{\sqrt{n}} \right)^{3/2} \lambda^{1/4} + \left(\frac{U_n(T)}{\sqrt{n}} \right)^2 \right).$$

Therefore, by a straightforward computation, $\mathbb{E}\|P_n - P\|_{G_\lambda}$ is smaller than $\delta\lambda$ provided that

$$\lambda \geq \frac{c_3}{\delta^2} \max \left\{ \frac{U_n^2(T)}{n}, \frac{U_n^2(D)}{n} \mathbb{E}\ell_{\beta^*} \right\}.$$

The second part of the claim is a direct application of Theorem 3.2. ∎

Since we have already bounded $U_n(T)$ for the sets T we are interested in, it remains to bound $U_n(D)$.

4.1.1 The complexity of D

Note that $D = \{x \in \mathbb{R}^d : \mathbb{E}\langle X, x\rangle^2 \leq 1\}$ is an ellipsoid in $\mathbb{R}^d$, as the unit ball of an inner product on $\mathbb{R}^d$ defined by $[x, y] = \mathbb{E}\langle X, x\rangle\langle X, y\rangle$. Thus, $D = AB_2^d$ for a certain linear operator A. Moreover, if X is a random vector distributed according to μ then A^*X is an isotropic random vector on $\mathbb{R}^d$. Indeed, for every $\theta \in S^{d-1}$, $A\theta$ is on the surface of inD, and thus

$$\mathbb{E}\langle \theta, A^*X\rangle^2 = \mathbb{E}\langle A\theta, X\rangle^2 = 1.$$

Lemma 4.8 *There is an absolute constant c for which the following holds. Let $X_1, ..., X_n \in \mathbb{R}^d$ and set $M = \max \|A^* X_i\|_{\ell_2^d}$. Then,*

$$\gamma_2(D, d_{\infty,n}) \leq cM\sqrt{\log n}\log d.$$

In particular,

$$U_n(D) \leq c(\mathbb{E}M^2)^{1/2}\sqrt{\log n}\log d.$$

Proof. Define

$$\begin{aligned} H_n &= \{x \in \mathbb{R}^d : \max_{1\leq i\leq n} |\langle x, X_i\rangle| \leq 1\}, \\ H_n' &= \{x \in \mathbb{R}^d : \max_{1\leq i\leq n} |\langle x, A^* X_i\rangle| \leq 1\}, \\ \|x\|_{H_n} &= \max_{1\leq i\leq n} |\langle x, X_i\rangle|, \\ \|x\|_{H_n'} &= \max_{1\leq i\leq n} |\langle x, A^* X_i\rangle|. \end{aligned}$$

Again, and at the price of a logarithmic looseness, the proof will be based on a covering numbers argument. Observe that for every $\varepsilon > 0$, $N(D, \varepsilon H_n) = N(B_2^d, \varepsilon H_n')$. Indeed, if $x, y \in D = AB_2^d$, then $x = Au$, $y = Av$, for some $u, v \in B_2^d$ and

$$\begin{aligned} \|x - y\|_{H_n} &= \max_{1\leq i\leq n} |\langle x - y, X_i\rangle| = \max_{1\leq i\leq n} |\langle Au - Av, X_i\rangle| \\ &= \max_{1\leq i\leq n} |\langle u - v, A^* X_i\rangle| = \|u - v\|_{H_n'}, \end{aligned}$$

and thus $A : (R^d, H_n') \to (\mathbb{R}^d, H_n)$ is an isometry, implying that $N(D, \varepsilon H_n) = N(B_2^d, \varepsilon H_n')$.

Let $G = (g_1, ..., g_d) \in \mathbb{R}^d$ be a Gaussian vector on $\mathbb{R}^d$. By the dual Sudakov inequality [24],

$$\log^{1/2} N(B_2^d, \varepsilon H_n') \leq c_1 \frac{\mathbb{E}\|G\|_{H_n'}}{\varepsilon},$$

and observe that

$$\mathbb{E}\|G\|_{H_n'} = \mathbb{E} \max_{1\leq i\leq n} |\langle G, A^* X_i\rangle| \leq c_2\sqrt{\log n} \max \|A^* X_i\|_{\ell_2^d}.$$

Fix $\varepsilon_0 > 0$ to be named later. Applying a volumetric argument, for $\varepsilon < \varepsilon_0$

$$\begin{aligned} \log N(B_2^d, \varepsilon H_n') &\leq \log N(B_2^d, \varepsilon_0 H_n') + \log N(\varepsilon_0 H_n', \varepsilon H_n') \\ &\leq c_3 \left(\frac{\sqrt{\log n}\max_{1\leq i\leq n} \|A^* X_i\|_{\ell_2^d}}{\varepsilon_0} \right)^2 + d\log\left(1 + \frac{\varepsilon_0}{\varepsilon}\right) \\ &\leq (c_3 + 1)d\log\left(1 + \frac{\varepsilon_0}{\varepsilon}\right) \end{aligned}$$

for the choice of $\varepsilon_0 = \sqrt{\log n}\max\|A^*X_i\|_{\ell_2^d}/\sqrt{d}$. Also $\sup_{v\in B_2^d}\|v\|_{H_n'} \leq \max_{1\leq i\leq n}\|A^*X_i\|_{\ell_2^d} \equiv M$. Using an entropy integral argument it is evident that

$$\begin{aligned}\gamma_2(D,d_{\infty,n}) &\leq c_4\left(\sqrt{d}\int_0^{\varepsilon_0}\sqrt{\log\left(1+\frac{\varepsilon_0}{\varepsilon}\right)}d\varepsilon + M\sqrt{\log n}\int_{\varepsilon_0}^{M}\frac{d\varepsilon}{\varepsilon}\right)\\ &\leq c_5\left(\sqrt{d}\varepsilon_0 + M\sqrt{\log n}\log\left(\frac{M}{\varepsilon_0}\right)\right)\\ &\leq c_6 M\sqrt{\log n}\log d.\end{aligned}$$

■

Combining the two error bounds, the first obtained by using $bB_1^d \cap \sqrt{\lambda}D \subseteq bB_1^d$ and the second obtained by using $bB_1^d \cap \sqrt{\lambda}D \subseteq \sqrt{\lambda}D$, the following error bound is evident.

Corollary 4.9 *There is an absolute constant c for which the following holds. Let* $h_1(n,d) = \max\{\sqrt{\log n}, \sqrt{\log d}\}$ *and* $h_2(n,d) = \log n \log^2 d$. *Set*

$$\lambda_1 = \frac{c}{\delta^2}\max\left\{\frac{b}{\sqrt{n}}\left(\|Q\|_{L_2}h_1(\log^{3/2}n)\sqrt{\mathbb{E}\ell_{\beta^*}}\right), \frac{b^2}{n}\left(h_1^2\|Q\|_{L_2}^2\log^3 n\right)\right\},$$
$$\lambda_2 = \frac{c}{\delta^2}\max\left\{\frac{b^2}{n}\left(\|Q\|_{L_2}^2 h_1^2\log^3 n\right), \frac{\|M\|_{L_2}^2}{n}\left(h_2\mathbb{E}\ell_{\beta^*}\right)\right\},$$

where $M = \max_{1\leq i\leq n}\|A^*X_i\|_{\ell_2^d}$, $Q = \max_{1\leq i\leq n}\|X_i\|_{\ell_\infty^d}$, *and A is the linear operator satisfying* $D = AB_2^d$. *Then*

$$\mathbb{E}\left(\mathcal{L}_{\hat{\beta}}|(X_i,Y_i)_{i=1}^n\right) \leq \min\{\lambda_1,\lambda_2\}.$$

with probability at least $1-2\delta$.

Let us return to the two families of measures we considered above and for the sake of simplicity assume in both cases that μ is isotropic, that is, $D = B_2^d$.

First, if $\|X\|_{\ell_\infty^d}$ is bounded in L_∞ by U then $Q \leq U$ and $M \leq U\sqrt{d}$. Hence,

$$\lambda_1 = c\max\left\{\left(U\cdot h_1(\log^{3/2}n)P\ell_{\beta^*}\right)\frac{b}{\sqrt{n}}, \left(U^2\cdot h_1^2\log^3 n\right)\cdot\frac{b^2}{n}\right\},$$
$$\lambda_2 = c\max\left\{\left(U^2\cdot h_1^2\log^3 n\right)\frac{b^2}{n}, \left(h_2\mathbb{E}\ell_{\beta^*}\right)\cdot\frac{d}{n}\right\},$$

Therefore, up to a poly-logarithmic factor in n and d, the error rate is

$$\min\left\{\frac{b^2}{n}+\frac{d}{n},\frac{b}{\sqrt{n}}\left(1+\frac{b}{\sqrt{n}}\right)\right\},$$

as was promised in the introduction.

For the second example, assume that μ is an isotropic, log-concave measure on $\mathbb{R}^d$. As we showed above, in this case $\|Q\|_{L_2} \leq c\log nd \leq ch_1^2$. To bound M, we need the following deep result of Paouris [25]:

Theorem 4.10 *There are absolute constants c_1 and c_2 for which the following holds. Let X be distributed according to an isotropic log-concave measure on $\mathbb{R}^d$. If $d \leq n \leq \exp(c_1\sqrt{d})$ and $X_1,...,X_n$ are independent copies of X then*

$$\left(\mathbb{E}\max_{1\leq i\leq n}\|X_i\|_{\ell_2^d}^2\right)^{1/2} \leq c_2\sqrt{d}.$$

Thus, one obtains the following estimate on λ_1 and λ_2.

$$\lambda_1 = c\max\left\{\frac{b}{\sqrt{n}}\left(h_1^3(\log^{3/2} n)\sqrt{\mathbb{E}\ell_{\beta^*}}\right),\frac{b^2}{n}\left(h_1^6\cdot\log^3 n\right)\right\},$$
$$\lambda_2 = c\max\left\{\frac{b^2}{n}\left(h_1^6\cdot\log^3 n\right),\frac{d}{n}\left(h_2\mathbb{E}\ell_{\beta^*}\right)\right\},$$

Again, up to a poly-logarithmic factor in n and d, the error rate is

$$\min\left\{\frac{b^2}{n}+\frac{d}{n},\frac{b}{\sqrt{n}}\left(1+\frac{b}{\sqrt{n}}\right)\right\}.$$

5 An oracle inequality for error rates

We remarked in the previous section that persistence is not the end of the story. As we increase the radius b_n towards $\sqrt{n}$, the rate of decay of the error of the empirical minimizer $P(\mathcal{L}_{\hat{f}}|(X_i,Y_i))$ becomes arbitrarily slow. On the other hand, if we slow the increase of b_n then the approximation error, $\inf_{\beta\in b_nB_1^d}\mathbb{E}\ell_\beta$, does not decay as a function of the radius b. Without knowing this approximation error in advance, the previous results do not allow us to optimize our choice of b_n. In this section, we show that if μ happens to be isotropic and $\|X_i\|_\infty$ is bounded almost surely, then we will show that the "lasso" estimator

$$\hat{\beta} = \mathrm{argmin}_{\beta\in\mathbb{R}^d}\left(\sum_{i=1}^n(\langle\beta,X_i\rangle - Y_i)^2 + \rho_n\|\beta\|_1\right)$$

performs almost as well as the empirical minimizer for the best value of b_n. For convenience, let us denote the approximation error by

$$\mathcal{A}_d(b) = \inf_{\beta \in bB_1^d} \mathbb{E}\ell_\beta.$$

Clearly, this is a decreasing function of b. In general, we would expect it to be bounded below, but in very nice cases (for example, if there is some "true" noiseless parameter) it might tend to zero as $b \to \infty$.

Our analysis of this problem will rely on two additional ingredients: a model-selection inequality and an "almost-isomorphic" result that holds with exponential confidence. The second component will be based on the estimates we have already established for $\mathbb{E}\|P_n - P\|_{G_\lambda}$.

The "almost-isometric" result we need is very similar to one which first appeared in [3] and has appeared several times since then.

Theorem 5.1 *[20] There exists an absolute constant c for which the following holds. Let $\mathcal{L}_F$ be a squared loss class associated with a convex class F and a random variable Y. Set G_λ to be the localization at level λ of the star-shaped hull of F (that is, $G_\lambda = \{\theta \mathcal{L}_f : 0 \leq \theta \leq 1 \text{ and } P\mathcal{L}_f \leq \lambda\}$). If $R = \max\{\sup_{f\in F} \|f\|_\infty, \|Y\|_\infty\}$ and $\mathbb{E}\|P_n - P\|_{G_\lambda} \leq \lambda/8$, then with probability at least $1 - \exp(-u)$, for every $f \in F$*

$$\frac{1}{2}P_n\mathcal{L}_f - \frac{\lambda}{2} - c(1+R^2)\frac{u}{n} \leq \mathbb{E}\mathcal{L}_f \leq 2P_n\mathcal{L}_f + \frac{\lambda}{2} + c(1+R^2)\frac{u}{n}.$$

To apply this theorem in our case, suppose that $\|X\|_{\ell_\infty^d} \leq M$ and $|Y| \leq M$ almost surely. If $F = \{f_\beta : \beta \in bB_1^d\}$ then $\sup_{f\in F}\|f\|_\infty \leq bM$. In particular, $\max\{\sup_{f\in F}\|f\|_\infty, \|Y\|_\infty\} \leq \max\{1,b\}M$ and we obtain the following corollary of Theorem 5.1, Theorem 3.6 and Lemma 4.1:

Corollary 5.2 *Suppose that X is distributed such that $\max\{\|X\|_{\ell_\infty^d}, |Y|\} \leq M$ almost surely. Then with probability at least $1-\exp(u)$, for every $\beta \in bB_1^d$,*

$$\frac{1}{2}P_n\mathcal{L}_f - \frac{\lambda}{2} - c(1+b^2)\frac{M^2u}{n} \leq \mathbb{E}\mathcal{L}_f \leq 2P_n\mathcal{L}_f + \frac{\lambda}{2} + c(1+b^2)\frac{M^2u}{n}$$

where

$$\lambda = c'M\max\left\{b\frac{\log^{3/2} n \log^{1/2}(dn)\sqrt{\mathcal{A}_{d_n}(b)}}{\sqrt{n}}, b^2M\frac{\log^3 n \log(dn)}{n}\right\},$$

and c, c' are absolute constants.

For the model selection result that we require, we will first need a few definitions:

Definition 5.3 *Let F be a class of functions and let $\{F_r; r \geq 1\}$ be a collection of subsets of F. We say that $\{F_r; r \geq 1\}$ is an* ordered, parameterized hierarchy of F *if the following conditions hold:*

1. *$\{F_r : r \geq 1\}$ is monotone (that is, whenever $r \leq s$, $F_r \subseteq F_s$);*
2. *for every $r \geq 1$, there exists a unique element $f_r^* \in F_r$ such that $\mathbb{E}\ell_{f_r^*} = \inf_{f \in F_r} P\ell_f$;*
3. *the map $r \mapsto \mathbb{E}\ell_{f_r^*}$ is continuous;*
4. *for every $r_0 \geq 1$, $\bigcap_{r > r_0} F_r = F_{r_0}$; and*
5. *$\bigcup_{r \geq 1} F_r = F$.*

Define, for $f \in F$,

$$r(f) = \inf\{r \geq 1; f \in F_r\}.$$

Note that from the semi-continuity property of an ordered, parameterized hierarchy (property 4), it follows that $f \in F_{r(f)}$ for all $f \in F$. Also, the second property of an ordered, parameterized hierarchy allows us to define, for $r \geq 1$ and $f \in F_r$, the excess loss function $\mathcal{L}_{r,f} = (f - Y)^2 - (f_r^* - Y)^2$. That is, $\mathcal{L}_{r,f}$ is the excess loss function with respect to the class F_r.

One can easily check that $F_r = \{f_\beta : \|\beta\|_1 \leq r - 1\}$ defines an ordered parameterized hierarchy of $F = \{f_\beta : \beta \in \mathbb{R}^d\}$ with $r(f) = \|\beta\|_1 + 1$; the only condition that is not completely trivial to check is the third condition. A proof of this fact is given in [20] when F_r is the unit ball of a reproducing kernel Hilbert space, but the same argument works in our case and so we omit it.

The model selection result we require has been established in [1]:

Theorem 5.4 *Let $\{F_r : r \geq 1\}$ be an ordered, parameterized hierarchy and define, for convenience, $\mathcal{L}_f = \mathcal{L}_{r(f),f}$. Suppose that $\rho_n(r)$ is a positive, increasing, continuous function. If for every $f \in F$,*

$$\frac{1}{2} P_n \mathcal{L}_f - \rho_n(r(f)) \leq \mathbb{E}\mathcal{L}_f \leq 2P_n \mathcal{L}_f + \rho_n(r(f))$$

then a regularized minimizer

$$\hat{f} \in \operatorname{argmin}_{f \in F}(P_n \ell_f + c\rho_n(r(f)))$$

satisfies

$$\mathbb{E}\ell_{\hat{f}} \leq \inf_{f \in F} \left(\mathbb{E}\ell_f + c' \rho_n(r(f)) \right),$$

where c and c' are absolute constants.

Note that the hypothesis in Theorem 5.4 is one that we are prepared to handle: it is an "almost-isomorphic" condition of the sort that we obtain from Theorem 5.1. However, Theorem 5.1 only gives us an almost-isomorphic condition for each F_r with high probability, while Theorem 5.4 requires an isomorphic condition for each F_r. Fortunately, the exponential confidence in Theorem 5.1 allows us to apply a union bound to Theorem 5.4, bringing us to the following result:

Theorem 5.5 *Let $\{F_r : r \geq 1\}$ be an ordered, parameterized hierarchy and suppose that $\rho_n(r, x)$ is a positive, continuous function that is increasing in both r and x. Suppose that for every $r \geq 1$, with probability at least $1 - \exp(-x)$, for every $f \in F_r$,*

$$\frac{1}{2} P_n \mathcal{L}_f - \rho_n(r, x) \leq \mathbb{E}\mathcal{L}_f \leq 2 P_n \mathcal{L}_f + \rho_n(r, x).$$

Then for every $x > 0$, with probability at least $1 - \exp(-x)$, every regularized minimizer

$$\hat{f} \in \operatorname{argmin}_{f \in F} (P_n \ell_f + c_1 \rho_n(2r(f), \theta(r(f), x)))$$

satisfies

$$\mathbb{E}\ell_{\hat{f}} \leq \inf_{f \in F} \left(\mathbb{E}\ell_f + c_2 \rho_n(2r(f), \theta(r(f), x)) \right)$$

where

$$\theta(r, x) = x + c_3 + c_4 \log \left(1 + \frac{\mathbb{E}\ell_{f_1^*}}{\rho_n(1, x + c_3)} + \log r \right)$$

and c_1 through c_4 are absolute constants.

Proof. Let $(r_i)_{i=1}^{\infty}$ be an increasing sequence (to be determined later) such that $r_1 = 1$ and $r_i \to \infty$ as $i \to \infty$. Fix $u > 0$ and define, for each $i \geq 1$, $u_i = u + \ln(\pi^2/6) + 2 \ln i$. Then

$$\sum_{i=0}^{\infty} e^{-u_i} = e^{-u}$$

and so, by the union bound, with probability at least $1 - e^{-u}$, for every $i \geq 1$,

$$\frac{1}{2} P_n \mathcal{L}_{r_i,f} - \rho_n(r_i, u_i) \leq \mathbb{E}\mathcal{L}_{r_i,f} \leq 2P_n \mathcal{L}_{r_i,f} + \rho_n(r_i, u_i).$$

If we only cared about a sequence of r_i, this would be enough for our result. However, we need an almost-isomorphic condition for all $r \geq 1$ and so the next step must be to find an almost-isomorphic condition for F_r when $r \in [r_{j-1}, r_j]$. In one direction, we have

$$\begin{aligned}
\mathbb{E}\mathcal{L}_{r,f} &= \mathbb{E}\mathcal{L}_{r_j,f} - \mathbb{E}\mathcal{L}_{r_j,f_r^*} \\
&\leq 2P_n \mathcal{L}_{r_j,f} + \rho_n(r_j, u_j) - \mathbb{E}\mathcal{L}_{r_j,f_r^*} \\
&= 2P_n \mathcal{L}_{r,f} + 2P_n \mathcal{L}_{r_j,f_r^*} + \rho_n(r_j, u_j) - \mathbb{E}\mathcal{L}_{r_j,f_r^*} \\
&\leq 2P_n \mathcal{L}_{r,f} + 5\rho_n(r_j, u_j) + 3\mathbb{E}\mathcal{L}_{r_j,f_r^*} \\
&\leq 2P_n \mathcal{L}_{r,f} + 5\rho_n(r_j, u_j) + 3\mathbb{E}\mathcal{L}_{r_j,f_{r_{j-1}}^*} \qquad (5.1)
\end{aligned}$$

while in the other direction, we get

$$\begin{aligned}
2\mathbb{E}\mathcal{L}_{r,f} &= 2\mathbb{E}\mathcal{L}_{r_j,f} - 2\mathbb{E}\mathcal{L}_{r_j,f_r^*} \\
&\geq P_n \mathcal{L}_{r_j,f} - 2\rho_n(r_j, u_j) - 2\mathbb{E}\mathcal{L}_{r_j,f_r^*} \\
&= P_n \mathcal{L}_{r,f} + P_n \mathcal{L}_{r_j,f_r^*} - 2\rho_n(r_j, u_j) - 2\mathbb{E}\mathcal{L}_{r_j,f_r^*} \\
&\geq P_n \mathcal{L}_{r,f} - \frac{5}{2}\rho_n(r_j, u_j) - \frac{3}{2}\mathbb{E}\mathcal{L}_{r_j,f_r^*} \\
&\geq P_n \mathcal{L}_{r,f} - \frac{5}{2}\rho_n(r_j, u_j) - \frac{3}{2}\mathbb{E}\mathcal{L}_{r_j,f_{r_{j-1}}^*} \qquad (5.2)
\end{aligned}$$

Now we can choose our sequence r_i: recall that $r_1 = 1$ and set r_i, for all $i \geq 2$, to be the largest number satisfying both

$$\begin{aligned}
r_i &\leq 2r_{i-1} \\
\mathbb{E}\mathcal{L}_{r_i,f_{r_{i-1}}^*} &\leq \rho_n(r_i, u_i). \qquad (5.3)
\end{aligned}$$

Note that choosing the largest number is not a problem because both $\rho_n(r, u)$ and $\mathbb{E}\mathcal{L}_{r,f_{r_{j-1}}^*}$ are continuous functions of r; that is, the supremum of the set of r satisfying (5.3) is attained.

Our choice of r_i ensures that, for all $i \geq 1$,

$$i \leq \frac{\mathbb{E}\ell_{f_{r_1}^*}}{\rho_n(r_1, u_1)} - \frac{\mathbb{E}\ell_{f_{r_i}^*}}{\rho_n(r_i, u_i)} + \log_2(2r_i) \leq \frac{\mathbb{E}\ell_{f_{r_1}^*}}{\rho_n(r_1, u_1)} + \log_2(2r_i). \qquad (5.4)$$

Indeed, for $i = 1$ this is trivial. For larger i we can proceed by induction: our definition of r_i ensures that either $r_i = 2r_{i-1}$ or $\mathbb{E}\ell_{f_{r_{i-1}}^*} = \mathbb{E}\ell_{f_{r_i}^*} + \rho_n(r_i, u_i)$.

In the first case, $\log_2 r_i = \log_2 r_{i-1} + 1$ and the inductive step follows. In the second case, assuming that

$$i - 1 \leq \frac{\mathbb{E}\ell_{f^*_{r_1}}}{\rho_n(r_1, u_1)} - \frac{\mathbb{E}\ell_{f^*_{r_{i-1}}}}{\rho_n(r_{i-1}, u_{i-1})} + \log_2 r_{i-1}$$

then

$$\begin{aligned} i &\leq \frac{\mathbb{E}\ell_{f^*_{r_1}}}{\rho_n(r_1, u_1)} - \frac{\mathbb{E}\ell_{f^*_{r_{i-1}}}}{\rho_n(r_{i-1}, u_{i-1})} + 1 + \log_2(2r_i) \\ &\leq \frac{\mathbb{E}\ell_{f^*_{r_1}}}{\rho_n(r_1, u_1)} - \frac{\mathbb{E}\ell_{f^*_{r_{i-1}}}}{\rho_n(r_i, u_i)} + 1 + \log_2(2r_i) \\ &= \frac{\mathbb{E}\ell_{f^*_{r_1}}}{\rho_n(r_1, u_1)} - \frac{\mathbb{E}\ell_{f^*_{r_i}}}{\rho_n(r_i, u_i)} + \log_2(2r_i) \end{aligned}$$

which proves (5.4) by induction. In particular, for any $i \geq 1$ and any $r \geq r_i$, $u_i \leq \theta(r, u)$. Therefore

$$\rho_n(r_i, u_i) \leq \rho_n(2r, \theta(r, u))$$

for any $r \in [r_{i-1}, r_i]$.

Note that (5.4) implies that the sequence r_i tends to infinity with i. Then by (5.1), (5.2) and (5.3), with probability at least $1 - e^{-u}$, for all $r \geq 1$ and all $f \in F_r$,

$$\frac{1}{2} P_n \mathcal{L}_{r,f} - 2\rho_n(2r, \theta(r, u)) \leq \mathbb{E}\mathcal{L}_{r,f} \leq 2P_n \mathcal{L}_{r,f} + 8\rho_n(2r, \theta(r, u)).$$

We conclude the proof by applying Theorem 5.4. ∎

Combining this model selection result with our previous estimates on the complexity of B_1^d, we obtain the following oracle inequality:

Corollary 5.6 *There are absolute constants c and c' for which the following holds. Let (d_n) be any increasing sequence and let (μ_n) be a sequence of measures on $\mathbb{R}^{d_n}$. Assume further that for every $n \geq 1$, X is a random vector in $\mathbb{R}^{d_n}$ distributed according to μ_n and that $\|X\|_{\ell_\infty^{d_n}} \leq M$ almost surely. If Y is a real-valued random variable with $|Y| \leq M$ almost surely, then for all $u > 0$, with probability at least $1 - \exp(-u)$, for any integer n and any*

$$\hat{\beta} \in \operatorname{argmin}_{\beta \in \mathbb{R}^{d_n}} \left(P_n \ell_\beta + \rho_n(1 + \|\beta\|_{\ell_1^d}, u) \right),$$

we have

$$\mathbb{E}\ell_{\hat{\beta}} \leq \inf_{\beta\in\mathbb{R}^{d_n}} \left(\mathbb{E}\ell_\beta + \rho_n(1+\|\beta\|_{\ell_1^d}, u)\right)$$

where $\rho_n(r,u) \geq \tau_n(r,u)$ *and*

$$\tau_n(r,u) = c(1+M)$$
$$\max\left\{ r\frac{\log^{3/2} n \log^{1/2}(d_n n)\sqrt{\mathcal{A}_{d_n}(r)}}{\sqrt{n}}, r^2 M \frac{\log^3 n \log(d_n n)}{n}, r^2 M \frac{u}{n}, \frac{Mr^2 \log\log r}{n} \right\}.$$

Proof. With Corollary 5.2 in mind, define

$$\rho_n(r,u) = c(1+M)\max\left\{ r\frac{\log^{3/2} n \log^{1/2}(d_n n)\sqrt{\mathcal{A}_{d_n}(r)}}{\sqrt{n}}, Mr^2 \frac{\log^3 n \log(dn)}{n}, \frac{Mr^2 u}{n} \right\}.$$

By Corollary 5.2, it is evident that ρ_n satisfies the hypothesis of Theorem 5.5. To complete the proof, we only need to expand the $\theta(r,u)$ function from Theorem 5.5 and simplify. Indeed, $\rho_n(1,u) \geq \rho_n(1,0) \geq cM^2 n^{-1}$ and so

$$\frac{\mathbb{E}\ell_{f_1^*}}{\rho(1,u+c_3)} \leq \frac{M^2}{\rho(1,0)} \leq cn.$$

Then $\theta(r,u) \leq u + c(1+\log n + \log\log r)$ and thus,

$$\rho_n(r,\theta(r,u)) \leq c(1+M)\max\left\{ r\frac{\log^{3/2} n \log^{1/2}(d_n n)\sqrt{\mathcal{A}_{d_n}(r)}}{\sqrt{n}}, \right.$$
$$\left. r^2 M \frac{\log^3 n \log(d_n n)}{n}, \frac{r^2 M u}{n}, \frac{Mr^2 \log\log r}{n} \right\} =: \tau_n(r,u).$$

■

Note that this is not the "lasso"-type regularization that we promised. Indeed, the regularization parameter contains quadratic terms like $\|\beta\|_1^2$ instead of only linear terms like $\|\beta\|_1$. Our next and final result will use the trivial bound $\mathcal{A}_{d_n}(b) \leq \mathcal{A}_{d_n}(0) \leq \|Y\|_{L_2}^2$ to simplify Corollary 5.6 and provide the promised regularization parameter. First, though, let us briefly discuss the case in which $\mathcal{A}_{d_n}(b)$ is, for sufficiently large n and r, zero, which is the case when there is a true, noiseless parameter for all sufficiently large n. Then there exists $s \in \mathbb{R}$ such that for a sufficiently large n,

$$\inf_{\beta\in\mathbb{R}^{d_n}} \left(\mathbb{E}\ell_\beta + \tau_n(1+\|\beta\|_{\ell_1^d}, u)\right) \leq \mathcal{A}_{d_n}(s) + \tau_n(s,u)$$
$$= cs^2(1+M^2)\max\left\{ \frac{\log^3 n \log(d_n n)}{n}, \frac{u}{n}, \frac{\log\log s}{n} \right\}.$$

If, for example, d_n is at most polynomial in n, then one obtains error rates that are $\sim 1/n$ up to logarithmic factors in n.

We conclude with the promised, lasso-type result:

Corollary 5.7 *There exist absolute constants c and c' for which the following holds. Let (d_n), X, Y and M be as in Corollary 5.6. If $\log d_n = o(n)$ then for all sufficiently large n (depending on d_n and M), with probability at least $1 - \exp(-\log^3 n \log(d_n n))$, for any*

$$\hat{\beta} \in \operatorname{argmin}_{\beta \in \mathbb{R}^{d_n}} \left(P_n \ell_\beta + cM^2 \|\beta\|_{\ell_1^d} \frac{\log^{3/2} n \log^{1/2}(d_n n)}{\sqrt{n}} \right),$$

we have

$$\mathbb{E}\ell_{\hat{\beta}} \leq \inf_{\beta \in \mathbb{R}^{d_n}} \left(\mathbb{E}\ell_\beta + c' M^2 (1 + \|\beta\|_{\ell_1^d}) \frac{\log^{3/2} n \log^{1/2}(d_n n)}{\sqrt{n}} \right).$$

Proof. Define

$$\tilde{\rho}_n(r,u) = c(1+M^2) \max \left\{ r \frac{\log^{3/2} n \log^{1/2}(d_n n)}{\sqrt{n}}, r^2 \frac{\log^3 n \log(d_n n)}{n}, r^2 \frac{u}{n}, \frac{r^2 \log\log r}{n} \right\}$$

and note that (for an appropriate choice of the absolute constant c) $\tilde{\rho}_n \geq \tau_n$. Therefore Corollary 5.6 holds with $\rho_n = \tilde{\rho}_n$. To complete the proof, one has to remove the r^2 terms from $\tilde{\rho}_n$. To this end, fix $u = \log^3 n \log(d_n n)$, and define

$$\sigma_n(r) = c(1+M^2) r \frac{\log^{3/2} n \log^{1/2}(d_n n)}{\sqrt{n}},$$

and

$$\begin{aligned}
S_n(\beta) &= \mathbb{E}\ell_\beta + c\tilde{\rho}_n(1 + \|\beta\|_{\ell_1^d}, u) \\
\hat{S}_n(\beta) &= P_n \ell_\beta + c'\tilde{\rho}_n(1 + \|\beta\|_{\ell_1^d}, u) \\
T_n(\beta) &= \mathbb{E}\ell_\beta + c\sigma_n(1 + \|\beta\|_{\ell_1^d}) \\
\hat{T}_n(\beta) &= P_n \ell_\beta + c'\sigma_n(1 + \|\beta\|_{\ell_1^d}).
\end{aligned}$$

We claim that

$$\operatorname{argmin}_{\beta \in \mathbb{R}^{d_n}} \hat{S}_n(\beta) \supset \operatorname{argmin}_{\beta \in \mathbb{R}^{d_n}} \hat{T}_n(\beta) \tag{5.5}$$

and that

$$\inf_{\beta \in \mathbb{R}^{d_n}} S_n(\beta) \leq \inf_{\beta \in \mathbb{R}^{d_n}} T_n(\beta). \tag{5.6}$$

Observe that if (5.5) and (5.6) hold, then they, together with Corollary 5.6, imply the desired result, because $\operatorname{argmin}(P_n\ell_\beta+\sigma_n(1+\|\beta\|_1)) = \operatorname{argmin}(P_n\ell_\beta+\sigma_n(\|\beta\|_1))$, as $\sigma_n(r)$ is a linear function of r.

Suppose there is some α such that $S_n(\alpha) > T_n(\alpha)$. Then $\tilde{\rho}_n(1+\|\alpha\|_1, u) > \sigma_n(1+\|\alpha\|_1)$, which implies (setting $r = 1+\|\alpha\|_1$ for ease of notation) that

$$r\frac{\log^{3/2} n \log^{1/2}(d_n n)}{\sqrt{n}} < \max\left\{ r^2\frac{\log^3 n \log(d_n n)}{n}, r^2\frac{u}{n}, r^2 \log\log r \frac{1}{n} \right\}.$$

With our choice of u, the first two terms on the right hand side are the same, and we infer that either

$$r\frac{\log^{3/2} n \log^{1/2}(d_n n)}{\sqrt{n}} < r^2\frac{\log^3 n \log(d_n n)}{n}$$

or

$$r\frac{\log^{3/2} n \log^{1/2}(d_n n)}{\sqrt{n}} < \frac{r^2 \log\log r}{n}.$$

In either case, for sufficiently large n,

$$r\frac{\log^{3/2} n \log^{1/2}(d_n n)}{\sqrt{n}} > 1$$

(the first case is immediate; note that the second case implies that $\sqrt{n}\log\sqrt{n} \le r\log r$ and so $r \ge \sqrt{n}$). In particular, $T_n(\alpha) \ge c\sigma_n(1+\|\alpha\|_1) \ge c(1+M^2)$. On the other hand,

$$\inf_\beta T_n(\beta) \le T_n(0) \le M + c\sigma_n(1) \le M + \tilde{c}\frac{(1+M^2)\log^{3/2} n \log^{1/2}(d_n n)}{\sqrt{n}}.$$

Therefore, if $\log d_n = o(n)$, then $\inf_\beta T_n(\beta) \le 2M$ for sufficiently large n, and thus, $T_n(\alpha) > \inf_\beta T_n(\beta)$, provided that the c in the definition of T_n satisfies $c > 1$. In other words, the only way to come close to the infimum of $T_n(\beta)$ is if $S_n(\beta) \le T_n(\beta)$, which implies that $\inf_\beta S_n(\beta) \le \inf_\beta T_n(\beta)$ and so (5.6) is confirmed.

Suppose we can choose α such that $\hat{S}_n(\alpha) > \hat{T}_n(\alpha)$. Then $\tilde{\rho}_n(1+\|\alpha\|_1, u) > \sigma_n(1+\|\alpha\|_1)$, and repeating the previous argument, it follows that for sufficiently large n (depending only on M and d_n), α is not a minimizer of $\hat{T}_n$. That is, $\alpha \in \operatorname{argmin}\hat{T}_n$ only if $\hat{T}_n(\alpha) \ge \hat{S}_n(\alpha)$. Since $\hat{T}_n(\beta) \le \hat{S}_n(\beta)$ for every β, then $\hat{T}_n(\alpha) = \hat{S}_n(\alpha)$. Hence, α is a minimizer of $\hat{S}_n$, proving (5.5). ■

References

[1] Peter L. Bartlett. Fast rates for estimation error and oracle inequalities for model selection. *Econometric Theory*, 24(02):545–552, 2008.

[2] Peter L. Bartlett, Michael I. Jordan, and Jon D. McAuliffe. Convexity, classification, and risk bounds. *J. Amer. Statist. Assoc.*, 101(473):138–156, 2006.

[3] Peter L. Bartlett and Shahar Mendelson. Empirical minimization. *Probab. Theory Related Fields*, 135(3):311–334, 2006.

[4] Florentina Bunea, Alexandre Tsybakov, and Marten Wegkamp. Sparsity oracle inequalities for the Lasso. *Electron. J. Stat.*, 1:169–194 (electronic), 2007.

[5] Bernd Carl. Inequalities of Bernstein-Jackson-type and the degree of compactness of operators in Banach spaces. *Ann. Inst. Fourier (Grenoble)*, 35(3):79–118, 1985.

[6] Felipe Cucker and Steve Smale. On the mathematical foundations of learning. *Bull. Amer. Math. Soc. (N.S.)*, 39(1):1–49 (electronic), 2002.

[7] Víctor H. de la Peña and Evarist Giné. *Decoupling*. Probability and its Applications (New York). Springer-Verlag, New York, 1999. From dependence to independence, Randomly stopped processes. U-statistics and processes. Martingales and beyond.

[8] David L. Donoho, Michael Elad, and Vladimir N. Temlyakov. Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Trans. Inform. Theory*, 52(1):6–18, 2006.

[9] David L. Donoho and Iain M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455, 1994.

[10] Evarist Giné and Joel Zinn. Some limit theorems for empirical processes. *Ann. Probab.*, 12(4):929–998, 1984. With discussion.

[11] Y. Gordon, A. E. Litvak, S. Mendelson, and A. Pajor. Gaussian averages of interpolated bodies and applications to approximate reconstruction. *J. Approx. Theory*, 149(1):59–73, 2007.

[12] Eitan Greenshtein and Ya'acov Ritov. Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli*, 10(6):971–988, 2004.

[13] O. Guédon, S. Mendelson, A. Pajor, and N. Tomczak-Jaegermann. Majorizing measures and proportional subsets of bounded orthonormal systems. *Rev. Mat. Iberoamericana*, 24(3):1075–1095, 2008.

[14] Olivier Guédon, Shahar Mendelson, Alain Pajor, and Nicole Tomczak-Jaegermann. Subspaces and orthogonal decompositions generated by bounded orthogonal systems. *Positivity*, 11(2):269–283, 2007.

[15] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.*, 58:13–30, 1963.

[16] Wee Sun Lee, Peter L. Bartlett, and Robert C. Williamson. Efficient agnostic learning of neural networks with bounded fan-in. *IEEE Trans. Inform. Theory*, 42(6, part 2):2118–2132, 1996.

[17] Chenlei Leng, Yi Lin, and Grace Wahba. A note on the lasso and related procedures in model selection. *Statist. Sinica*, 16(4):1273–1284, 2006.

[18] N. Meinshausen and B. Yu. Lasso-type recovery of sparse representations for high-dimensional data. *Ann. Statist.*, 37(1):246–270, 2009.

[19] Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the lasso. *Ann. Statist.*, 34(3):1436–1462, 2006.

[20] S. Mendelson and J. Neeman. Regularization in Kernel Learning. *preprint*, 2008.

[21] Shahar Mendelson. Improving the sample complexity using global data. *IEEE Trans. Inform. Theory*, 48(7):1977–1991, 2002.

[22] Shahar Mendelson. On the performance of kernel classes. *J. Mach. Learn. Res.*, 4(5):759–771, 2004.

[23] Vitali D. Milman and Gideon Schechtman. *Asymptotic theory of finite-dimensional normed spaces*, volume 1200 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 1986. With an appendix by M. Gromov.

[24] Alain Pajor and Nicole Tomczak-Jaegermann. Remarques sur les nombres d'entropie d'un opérateur et de son transposé. *C. R. Acad. Sci. Paris Sér. I Math.*, 301(15):743–746, 1985.

[25] G. Paouris. Concentration of mass on convex bodies. *Geom. Funct. Anal.*, 16(5):1021–1049, 2006.

[26] Gilles Pisier. *The volume of convex bodies and Banach space geometry*, volume 94 of *Cambridge Tracts in Mathematics*. Cambridge University Press, Cambridge, 1989.

[27] M. Talagrand. Sharper bounds for Gaussian and empirical processes. *Ann. Probab.*, 22(1):28–76, 1994.

[28] Michel Talagrand. *The generic chaining*. Springer Monographs in Mathematics. Springer-Verlag, Berlin, 2005. Upper and lower bounds of stochastic processes.

[29] Robert Tibshirani. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58(1):267–288, 1996.

[30] Sara A. van de Geer. High-dimensional generalized linear models and the lasso. *Ann. Statist.*, 36(2):614–645, 2008.

[31] Aad W. van der Vaart and Jon A. Wellner. *Weak convergence and empirical processes*. Springer Series in Statistics. Springer-Verlag, New York, 1996. With applications to statistics.

[32] Cun-Hui Zhang and Jian Huang. The sparsity and bias of the LASSO selection in high-dimensional linear regression. *Ann. Statist.*, 36(4):1567–1594, 2008.