ℓ_1-regularized linear regression: Persistence and oracle inequalities[5]

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

February 4, 2011

Abstract

We study the predictive performance of ℓ_1-regularized linear regression in a model-free setting, including the case where the number of covariates is substantially larger than the sample size. We introduce a new analysis method that avoids the boundedness problems that typically arise in model-free empirical minimization. Our technique provides an answer to a conjecture of Greenshtein and Ritov [17] regarding the "persistence" rate for linear regression and allows us to prove an oracle inequality for the error of the regularized minimizer. It also demonstrates that empirical risk minimization gives optimal rates (up to log factors) of convex aggregation of a set of estimators of a regression function.

1 Introduction

In this article, we study the problem of linear regression with an ℓ_1 regularization. Consider a random pair $(X, Y) \in \mathbb{R}^d \times \mathbb{R}$ of which we have n independent samples $X_1, Y_1, \ldots, X_n, Y_n$. For a fixed $\rho > 0$, consider the

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

[5]The research leading to these results has received funding from the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013) / ERC grant agreement n^o [203134], from the Israel Science Foundation grant 666/06 and from the Australian Research Council grant DP0986563.

We gratefully acknowledge the support of the NSF through grant DMS-0707060.

ℓ_1-regularized estimate $\hat{\beta}$, defined by

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

where ρ_n is a parameter of the problem, $\langle \cdot, \cdot \rangle$ denotes the Euclidean inner product and $\| \cdot \|_{\ell_1^d}$ is the norm $\|x\|_{\ell_1^d} = \sum_{i=1}^{d} |x_i|$ on $\mathbb{R}^d$. This is known as "LASSO" regression and it is often motivated by the fact that it tends to select solutions $\hat{\beta}$ that are sparse [37] (that is, it selects some $\hat{\beta} \in \mathbb{R}^d$ with considerably fewer than d non-zero coordinates), particularly when compared with ordinary least squares or with ℓ_2-regularized (or "Ridge") regression. 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 classical setup, the dimension d is fixed, while the sample size n grows to infinity. A more modern problem – which will be the focus of this article – is the behavior of $\hat{\beta}$ when its dimension grows with the number of samples. There are three problems that are typically studied in this setting, of which we will consider only the first:

1. whether $\hat{\beta}$ performs well on future samples (ie. whether $\mathbb{E}(\langle X, \hat{\beta} \rangle - Y)^2$ is small),
2. whether $\hat{\beta}$ closely approximates some "true" parameter β^* (ie. whether $\|\beta^* - \beta\|$ is small with high probability), or
3. whether $\hat{\beta}$ correctly identifies the relevant coordinates of some "true," sparse parameter β^* (ie. whether $(\beta^* = 0) \iff (\hat{\beta} = 0)$ with high probability).

The first of these properties has been studied, for example, in [12, 39, 17, 4, 7, 21, 16] while the latter two are studied in [5, 41, 42, 26, 25, 23, 11, 7, 24, 21].

Note that the second and third questions above both require some notion of a "true" model, while the first does not necessarily. Nevertheless, most previous results for the prediction question *have* assumed underlying models. For example, Bickel et al [4] prove that, for any $\epsilon > 0$,

$$\mathbb{E}(\langle X, \hat{\beta} \rangle - Y)^2 \leq (1 + \epsilon) \inf_{\beta} \left((\langle X, \hat{\beta} \rangle - Y)^2 + O(n^{-1} \log M \|\beta\|_0) \right),$$

provided that (X, Y) are generated according to a linear model with additive Gaussian noise and the coordinates of X are, in some sense, very far from being linearly dependent. In our general setting, we achieve a slower error

rate of $n^{-1/2}$, but we do it with an *exact* oracle inequality (ie. with a constant of 1 instead of $(1+\epsilon)$). Moreover, we don't require structure on the joint distribution of X, nor do we need (X,Y) to follow a linear model with additive Gaussian noise. Even in the broad generality of our set-up, fast rates seem to be possible if we allow an approximate oracle inequality as in [4]: the method that we develop here is used in [22], which contains a $(1+\epsilon)$-oracle inequality with fast rates in a very general case.

A notable exception to this model-based approach is that of van de Geer [39], which does not assume that the data follow a linear model (and also contains some further generalizations, like allowing for link functions and different losses). However, [39] requires a Lipschitz condition on the loss, which does not hold for linear functions unless the covariates are bounded in the ℓ_2 norm (actually, [39] weakens this somewhat: it is enough for $\mathbb{E}(Y|X) - \langle X, \hat{\beta} \rangle$ to be uniformly bounded). Also, [39] is interested in the sparse case and so its results are in terms of an explicit sparsity assumption.

Indeed, the problem of unbounded classes of functions (or non-Lipshitz losses) has traditionally been difficult to handle in model-free results because some important tools (namely, contraction inequalities and Talagrand's concentration inequality) do not apply. It is possible to work around unboundedness by dividing the set of linear functions into a hierarchy of bounded classes, but the bounds obtained by applying the usual methods to this hierarchy of classes are loose (see [29] for a detailed discussion in a slightly different setting).

Our main contribution is a model-free method for tackling the prediction question which mitigates the problems of unboundedness and requires only mild assumptions. For $\beta \in \mathbb{R}^d$, let $\ell_\beta : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}$ be the function $\ell_\beta(x,y) = (\langle x, \beta \rangle - y)^2$. Then our main result will be presented as an answer to a question posed by Greenshtein and Ritov [17] about the persistence rate of linear regression.

Let $(d_n)_{n=1}^\infty$ be an increasing sequence, consider a sequence of probability 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. Fix some increasing sequence b_n and consider, for every n, the empirical minimizer in $b_n B_1^{d_n} = \{x \in \mathbb{R}^{d_n} : \|x\|_{\ell_1^{d_n}} \leq b_n\}$:

$$\hat{\beta}_n = \operatorname*{argmin}_{\|\beta\|_{\ell_1^{d_n}} \leq b_n} P_n \ell_\beta$$

where P_n is the empirical distribution of $X_1, Y_1, \ldots, X_n, Y_n$, so that $P_n \ell_\beta(X,Y) =$

$n^{-1}\sum_{i=1}^n(\langle X_i,\beta\rangle - Y_i)^2$. The sequence $\hat{\beta}_n$ is called *persistent* if

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

in probability, where P denotes the conditional expectation given $X_1, Y_1, \ldots, X_n, Y_n$, so that $P\ell_{\hat{\beta}_n}(X,Y)$ is a random variable that depends on the data.

Empirical minimization gives a persistent sequence $(\hat{\beta}_n)$ provided that the sequences (b_n) and (d_n) do not increase too rapidly. When (d_n) is at most polynomially large in n, Greenshtein and Ritov asked for the most quickly increasing sequence (b_n) such that empirical minimization is persistent. 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 each coordinate of X be bounded almost surely. We answer this question in the affirmative (up to the power of the logarithm) under even milder assumptions on μ_n. In fact, for $\hat{\beta} = \mathrm{argmin}_{\beta\in bB_1^d} P_n\ell_\beta$ we will establish almost sharp estimates (up to the power of the log factor) on the quantity

$$P\ell_{\hat{\beta}} - \inf_{\beta\in bB_1^d} P\ell_\beta$$

as a function of the radius b, the dimension d and the sample size n. For example, we will show that if each coordinate of X is subexponential then, up to poly-logarithmic factors in b, d and n, the error of the empirical minimizer is bounded by $\frac{b}{\sqrt{n}}(1+\frac{b}{\sqrt{n}})$. If, moreover, μ is log-concave and isotropic then, up to poly-logarithmic factors in b, d and n, the error of the empirical minimizer can also be bounded by $\frac{b^2}{n}+\frac{d}{n}$.

Before stating the result, let us be precise about what we mean by the "subexponential" condition that was mentioned in the previous paragraph.

Assumption 1.1 *For each n, let μ_n be a probability measure on $\mathbb{R}^{d_n}\times\mathbb{R}$ such that, if (X,Y) is distributed according to μ_n then, for every $1\le i\le d_n$ and every $t\ge 1$,*

$$Pr\left(|\langle X, e_i\rangle| \ge t\right) \le 2\exp(-ct)$$

and $\mathbb{E}Y^2\le c$, where c is an absolute constant and $e_1,\ldots,e_d$ is the standard basis in $\mathbb{R}^d$.

Note that the coordinates of X do not need to be either independent or identically distributed.

Theorem 1.1 *Suppose that (d_n) is an increasing sequence and that the sequence (μ_n) satisfies 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 $|\langle X, e_i\rangle| \leq C$ almost surely for every μ_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 we do not believe them to be best possible.

Of course, persistence is not of real practical interest. Given samples $(X_1, Y_1), \ldots, (X_n, Y_n) \in \mathbb{R}^d \times \mathbb{R}$, Theorem 1.1 suggests that we might get a reasonable estimator by picking some $b_n \ll \sqrt{n}$ and doing empirical minimization in $b_n B_1^d$. However, Theorem 1.1 does not tell us whether, for example, $n^{1/3}$ is better than $n^{1/4}$, which is a very important question if we actually want to find an estimator. Fortunately, results like Theorem 1.1 can be used to study the predictive performance of the LASSO estimator. In doing so, we have assumed additional regularity (specifically, uniform boundedness) on the distributions of X and Y. This additional regularity is very convenient because it allows us to use certain strong concentration properties. However, it is probably not necessary; we will comment a little bit more on this matter when it comes time to present the proofs.

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 sequence of probability measures on $\mathbb{R}^{d_n} \times \mathbb{R}$. For $n \geq 1$ and taking (X, Y) distributed according to μ_n, suppose that each coordinate of X is bounded almost surely (in absolute value) by M and that $|Y|$ is also bounded almost surely by M. If we define*

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

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

$$\hat{\beta} = \operatorname*{argmin}_{\beta \in \mathbb{R}^{d_n}} (P_n \ell_\beta + \rho_n \|\beta\|_{\ell_1^{d_n}}),$$

satisfies

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

Remark 1.3 *Observe that Theorem 1.2 is an exact oracle inequality. As the proof of Theorem 1.2 reveals, we can achieve a confidence of* $1-\exp(-u)$ *for any* u *if we include a term of the form* $\frac{u}{n}\|\beta\|^2_{\ell_1^{d_n}}$ *in the regularization and error terms. It is not clear, however, whether such a term is meaningful or whether it merely appears as an artifact of our analysis. We chose to present the theorem in its present form because then the estimator* $\hat{\beta}$ *is exactly the well-studied LASSO estimator.*

Our results also have consequences for the question of optimal aggregation schemes (see, for example, [38]). Suppose that the coordinates of X correspond to a dictionary of d functions, and we set $b_n = 1$. Then the estimator we consider simply minimizes the empirical risk over the convex hull of this dictionary. Our results show that if, for example, the functions in the dictionary are uncorrelated and appropriately scaled, then the distribution of X is isotropic and so the error rate (how much $P\ell_{\hat{\beta}_n}$ exceeds $\inf\{P\ell_\beta : \|\beta\|_{\ell_1^d} \leq 1\}$) is $\min\left(d/n, \sqrt{\frac{\log d}{n}}\right)$, up to log factors in d and n. When $d \gg \sqrt{n}$, we get the same rate without assuming that the dictionary is uncorrelated. Tsybakov showed [38] that this rate cannot be improved, and that a complex estimator that extends ideas of Catoni [9] achieves this rate. Bunea, Tsybakov and Wegkamp showed [6] that this rate is also achieved by an estimator that minimizes a weighted ℓ_1-penalized least squares criterion, with data dependent weights. Our results imply that, up to log factors, this optimal rate is achieved by the simpler estimator that minimizes squared error over the simplex.

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.

Let $|x|$ denote the Euclidean norm of x. A subset B of a vector space is called symmetric if $x \in B$ implies that $-x \in B$. For every $1 \leq p < \infty$ and every integer d, $\|\cdot\|_{\ell_p^d}$ denotes the norm $\|x\|^p_{\ell_p^d} = \sum |x_i|^p$ (with the usual extension for $p = \infty$) and B_p^d is its unit ball.

A significant part of our work will be devoted to the study of the supremum of a stochastic process indexed by a subset of $\mathbb{R}^d$. 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 $t \geq 1$*

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

For example, given any $T \subset \mathbb{R}^d$, the Gaussian and Rademacher processes

$$\left\{\sum_{i=1}^d g_i x_i : x \in T\right\} \text{ and } \left\{\sum_{i=1}^d \varepsilon_i x_i : x \in T\right\}$$

are subgaussian with respect to the Euclidean metric, 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.

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 $\sup_{t\in T} |Z_t|$ in terms of metric invariants of the index set. In particular, we will rely heavily on an entropy integral result. The entropy integral mechanism was introduced by Dudley [13] and then extended by Pisier [33] and Talagrand [35]. The bounds we present could be tightened (by logarithmic factors) by avoiding an entropy integral and bounding Talagrand's γ_2 functional directly. For our purposes, however, the gains are not worth the additional complications involved. We refer the reader to [36] for an extensive survey of chaining methods and their applications.

Definition 2.2 *Let (T, d) be a metric space. Define $N(\epsilon, T, d)$ to be the smallest number of open balls (with respect to the metric d) needed to cover T. Define*

$$\mathcal{D}(T,d) = \int_0^{diam(T,d)} \sqrt{\log N(\epsilon, T, d)}\, d\epsilon.$$

Theorem 2.3 *[36] There exists an absolute constant c such that for every metric space (T, d), every subgaussian process $\{Z_t\}$ indexed by T and every $t_0 \in T$,*

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

The most important processes for us will be 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. The empirical process indexed by F is the collection of $Z_f = P_n f - Pf = n^{-1} \sum_{i=1}^n f(X_i) - \mathbb{E}f$ for each $f \in F$. We denote

$$\|P_n - P\|_F = \sup_{f \in F} |P_n f - Pf| .$$

Unfortunately, an empirical process is not subgaussian under typical assumptions on F and μ. Indeed, by Bernstein's inequality (e.g. [40]) – which is sharp is many cases – the typical tail behavior of $P_n f - Pf$ is a mixture of subgaussian and subexponential tails. One way around this problem is to use a symmetrization argument, due to Giné and Zinn [14], from which we obtain a subgaussian process with respect to a random metric.

Theorem 2.4 *Let F be a class of functions 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 the Euclidean norm $|\cdot|$) indexed by the random coordinate projections

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

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

The particular function classes of interest to us are the sets of linear functionals $\{\langle t, \cdot \rangle : t \in T\}$, where $T \subset \mathbb{R}^d$ is a compact, 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 $T = bB_1^d = \{\beta \in \mathbb{R}^d : \|\beta\|_{\ell_1^d} \leq b\}$ 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 μ be a probability measure on $\mathbb{R}^d$ and consider a 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 (more precisely, by a function f_β where $\beta \in T$) with respect to the squared loss, using empirical data, which is a random sample $(X_i, Y_i)_{i=1}^n$ drawn from 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 compact, convex class of functions. By the strict convexity of the $L_2(\mu)$ norm, $P\ell_f$ 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.

Let $\hat{f}$ be the empirical minimizer

$$\hat{f} = \operatorname*{argmin}_{f \in F} P_n \ell_f.$$

With this notation, our problem is to obtain a high-probability bound on the conditional expectation

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

as a function of the sample size n.

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

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

Lemma 3.1 *Let $F \subset L_2$ be a compact, convex set of functions and let $\mathcal{L}_F$ be the as above. 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 $P\mathcal{L}_f$ is small then f must be close to the true minimizer f^* with respect to the $L_2(\mu)$ norm.

This lemma appeared implicitly in several places (see, for example [27], 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).

It is well known [3] 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 $P\mathcal{L}_f \geq \lambda$,

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

Under such an event, the empirical minimizer must satisfy $P\mathcal{L}_{\hat{f}} \leq \lambda$: if not then we could conclude from $P_n\mathcal{L}_{\hat{f}} \leq 0$ that $P\mathcal{L}_{\hat{f}} \leq 0$.

One way to find such a λ is to bound $\mathbb{E}\|P_n - P\|_{G_\lambda}$, where

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

Indeed, it was first noted in [3] that if $\mathbb{E}\|P_n - P\|_{G_\lambda} \leq \alpha\lambda$ for some $0 < \alpha < 1$ and one has a strong concentration phenomenon for $\|P_n - P\|_{G_\lambda}$ around its expectation then, with high probability, the risk of the empirical minimizer is at most $c(\alpha)\lambda$. In fact, one can obtain the same result without the strong concentration if one is willing to have confidence that is less than exponential. Although the proof is essentially a trivial modification of the proof in [3], it is fairly short and so we include it for completeness.

Theorem 3.2 *Define*

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

If $\mathbb{E}\|P_n - P\|_{G_\lambda} \leq \delta\lambda$ *then with probability at least* $1 - 2\delta$, $P\mathcal{L}_{\hat{f}} \leq \lambda$.

Proof. By rewriting G_λ as

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

it is evident that

$$\sup_{\{\mathcal{L}_f : P\mathcal{L}_f \geq \lambda\}}\left|\frac{P_n\mathcal{L}_f - P\mathcal{L}_f}{P\mathcal{L}_f}\right| = \frac{\|P_n - P\|_{G_\lambda}}{\lambda}.$$

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 $\{\mathcal{L}_f : P\mathcal{L}_f \geq \lambda\}$: by (3.1), with probability at least $1-2\delta$, for all $\mathcal{L}_f$ with $P\mathcal{L}_f \geq \lambda$,

$$\frac{1}{2}P\mathcal{L}_f \leq P_n\mathcal{L}_f \leq \frac{3}{2}P\mathcal{L}_f.$$

Since the loss function of the empirical minimizer, $\mathcal{L}_{\hat{f}}$, does not satisfy this inequality (because $P_n\mathcal{L}_{\hat{f}} \leq 0$), then $P\mathcal{L}_{\hat{f}} \leq \lambda$, 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 [20], if $A \subset \mathbb{R}^n$ then the Rademacher process indexed by A, given by $x \mapsto \sum_{i=1}^n \varepsilon_i x_i = Z_x$ 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 T. This overcomes 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 the metric d 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 $v, u \in \mathbb{R}^d$,

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

Recall that $\beta^* \in T$ is an element for which $\inf_{\beta\in T} P\ell_{f_\beta}$ is attained. By Lemma 3.1,

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

where we use the notation $a + B = \{a + b : b \in B\}$. (Recall that $\|\cdot\|_{L_2}$ is not necessarily the standard Euclidean norm, but rather it is induced by the distribution of X).

Now, 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 $(a+b)^2 \leq 2a^2+2b^2$. The result follows from Höffding's inequality. ∎

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

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

The following is a result from [18].

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,*

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

In particular,

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

Thus, the dominating term in the expectation of the worst deviation of $P_n f^2$ from the mean Pf^2 can be upper bounded in terms of the L_2 norm of the random entropy integral $\mathcal{D}(2T \cap \sqrt{\lambda} D, d_{\infty,n})$.

The next 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 $\lambda > 0$,*

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

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

Proof. By Theorem 2.4, Lemma 3.1 and the definition of the L_2 metric on $\mathbb{R}^d$ endowed by X,

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

where

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

recalling that $D = \{x \in \mathbb{R}^d : \mathbb{E} |\langle x, X \rangle|^2 \leq 1\}$ and that $\|\beta\|_{L_2}^2 = \mathbb{E} \langle \beta, X \rangle^2$.

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. By Theorem 2.3,

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

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

Note that the bound that we have established thus far is for $\mathbb{E}\|P_n - P\|_{\mathcal{L}_{F,\lambda}}$ where $\mathcal{L}_{F,\lambda} = \{\mathcal{L}_f : P\mathcal{L}_f \leq \lambda\}$ 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 [28].

To simplify notation, define

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

where (as before) $K_\lambda = 2T \cap \sqrt{\lambda}D$.

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 $\lambda > 0$

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

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

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

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

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

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

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

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 $P\ell_{\beta^*} \leq \|Y\|_{L_2}^2$, one obtains an error estimate for the empirical minimization problem:

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

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

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 $U_n(bB_1^d)$.

Remark 3.8 *Corollary 3.7 and the second and third parts of Theorem 3.6 follow from the trivial estimate that $K_\lambda \subset 2T$, which is rather loose unless T*

is very small. The fact that the complexity of the indexing set is governed by the intersections $2T \cap \sqrt{\lambda}D$ 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 persistence problem addressed in [17]. 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 4.5.

4 Empirical minimization is persistent

From the results of the previous section, it is evident that one can prove persistence for empirical minimization by bounding $\mathbb{E}\mathcal{D}(bB_1^d \cap \sqrt{\lambda}D, d_{\infty,n})^2$. In Section 4.1, we bound this quantity using the fact that $bB_1^d \cap \sqrt{\lambda}D \subseteq bB_1^d$. In Section 4.2, we bound it using $bB_1^d \cap \sqrt{\lambda}D \subseteq \sqrt{\lambda}D$, and we combine these results.

To obtain these bounds, we need some assumption on the probability distribution μ that gives control of the tails of $\|X_i\|_{\ell_\infty^d}$. We consider two examples which provide the control we need. In the first example, the random variable $\|X\|_{\ell_\infty^d}$ bounded almost surely, while the second consists of log-concave and isotropic measures.

Definition 4.1 *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 $\theta \in \mathbb{R}^d$ with $|\theta| = 1$,

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

where X is distributed according to μ.

The main result of this section is that, for either of these two families ($\|X\|_{\ell_\infty^d}$ bounded in L_∞ or the distribution of X log-concave and isotropic), with high probability, the error $P\mathcal{L}_{\hat{\beta}}$ of the empirical minimizer is bounded by an expression that grows as

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

up to a poly-logarithmic factor in n and d_n.

4.1 Error rates from the entropy integral of bB_1^d

The idea behind the following result appeared in [18] but we will present a detailed proof for the sake of completeness.

Lemma 4.2 *There exists an absolute constant c for which the following holds. For every $b > 0$,*

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

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 $X_1, ..., X_n \in \mathbb{R}^d$ and define the quasi-norm $\|\cdot\|_{H_n}$ by

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

Let $H_n = \{u : \max_{1\leq i\leq n} |\langle u, X_i\rangle| \leq 1\}$ be the unit ball of $\|\cdot\|_{H_n}$.

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 [8], 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, [34] 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,

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

Hence,

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

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 $\mathcal{D}$ on n and d, since our estimates yield a poly-logarithmic dependency in those parameters. Using a much more delicate approach – a generic chaining bound rather than an entropy integral – the power of the logarithms can be reduced (though not completely eliminated). This was done in [19].

We measure the decay of the tails of $|\langle X, e_i \rangle|$ using the Orlicz norm.

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 [10, 40]. A well known fact that follows from Borell's inequality (see, e.g. [30], 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.

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*

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

If μ is log-concave then

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

and if μ is log-concave and isotropic then

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

Proof. A well-known observation due to Pisier (see, e.g. [40]) is that if $Z_1, ..., Z_m$ are random variables then

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

where c_1 is an absolute constant.

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

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

If μ is log-concave,

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

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

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

■

We are now ready to formulate the first error rate estimate for $T = bB_1^d$, which follows directly from Lemmas 4.2 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 $\rho = \max_{1\le j\le d} \|\langle X, e_j\rangle\|_{\psi_1}$.*

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

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

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 $\rho = \max_{1\le j\le d} \|\langle X, e_j\rangle\|_{L_2}$, and if μ is distributed according to a measure that is both log-concave and isotropic then (4.2) holds with $\rho = 1$.

By taking a slowly decreasing sequence δ_n, Theorem 4.5 immediately implies Theorem 1.1.

4.2 Error rates from the entropy integral of $bB_1^d \cap \sqrt{\lambda}D$

Theorem 4.5 yields an estimate on the error rate of the empirical minimizer for each choice of b, d and n, but a careful look at the estimate there shows that it is suboptimal for certain choices. 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 comes from its use of the inclusion $bB_1^d \cap \sqrt{\lambda}D \subset bB_1^d$. However, if b and d are constant with respect to n and the distribution of X is isotropic, then $bB_1^d \cap \sqrt{\lambda}D = bB_1^d \cap \sqrt{\lambda}B_2^d = \sqrt{\lambda}B_2^d$ as long as $\lambda \le b^2/d$. Hence, if there is any hope that the error rate λ_n converges 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. In this section, we do this, and combine the result with the result of the previous section.

Note that the work in this section may not be interesting for the study of ℓ_1-penalized regression, in which the regime $d \gg n$ may be the more interesting one. However, the case $d < n$ has implications for aggregation. As mentioned in the introduction, we can achieve rates of (up to logarithms) d/n when $d < n$, a claim which we will prove in this section.

It turns out that for certain cases (e.g. if X is an isotropic, Gaussian vector) one can prove sharp bounds for the "complexity" of the interpolation body $bB_1^d \cap \sqrt{\lambda}D$ for all values of n, b, d and λ (see [15]). 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.

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 $G_\lambda = \{\lambda \mathcal{L}_f / P\mathcal{L}_f : P\mathcal{L}_f \geq \lambda\}$.

Theorem 4.6 *There exists an absolute constant c for which the following holds. If*

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

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

$$P\mathcal{L}_{\hat{\beta}} \leq \frac{c}{\delta^2} \max\left\{ \frac{U_n^2(T)}{n}, \frac{U_n^2(D)}{n} P\ell_{\beta^*} \right\}$$

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

Proof. Recall that

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

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

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

Setting $A_i = U_n(K_{2^{i+1}\lambda})/\sqrt{n}$, we have

$$\begin{aligned}
2^{-i}\phi_n(2^{i+1}\lambda) &\leq \left(2^{-i} \left(A_i \lambda^{1/2} + A_i^{3/2} \lambda^{1/4} + A_i^2 \right) + 2^{-i} A_i \left(P\ell_{\beta^*} \right)^{1/2} \right) \\
&\leq 2^{-i} \left(\frac{U_n(T)}{\sqrt{n}} \lambda^{1/2} + \left(\frac{U_n(T)}{\sqrt{n}} \right)^{3/2} \lambda^{1/4} + \left(\frac{U_n(T)}{\sqrt{n}} \right)^2 \right) \\
&\quad + 2^{-i} \frac{U_n(D)}{\sqrt{n}} \left(2^{i+1} \lambda P\ell_{\beta^*} \right)^{1/2},
\end{aligned}$$

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

$$c_2\left(\frac{U_n(T)}{\sqrt{n}}\lambda^{1/2}+\frac{U_n(D)}{\sqrt{n}}(\lambda P\ell_{\beta^*})^{1/2}+\left(\frac{U_n(T)}{\sqrt{n}}\right)^{3/2}\lambda^{1/4}+\left(\frac{U_n(T)}{\sqrt{n}}\right)^2\right).$$

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

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

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 $U_n(D)$. 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, whenever $|\theta|=1$, $A\theta$ is on the boundary of D, and thus

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

Lemma 4.7 *There is an absolute constant c for which the following holds. Let $X_1,...,X_n\in\mathbb{R}^d$ and set $Z=\max\|A^*X_i\|_{\ell_2^d}$. Then,*

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

In particular,

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

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, $\|Ax\|_{H_n} = \|x\|_{H_n'}$ and thus the function $A : (\mathbb{R}^d, H_n') \to (\mathbb{R}^d, H_n)$ 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 [31],

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

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 $\varepsilon_0 = Z\sqrt{\log n}/\sqrt{d}$. By a volumetric argument, for $\varepsilon < \varepsilon_0$,

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

Also, $\sup_{v \in B_2^d} \|v\|_{H_n'} \leq \max_{1 \leq i \leq n} \|A^* X_i\|_{\ell_2^d} = Z$. Using an entropy integral argument,

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

■

Combining the two error bounds, the first obtained in the previous section by using $bB_1^d \cap \sqrt{\lambda} D \subseteq bB_1^d$ and the second obtained by using $bB_1^d \cap \sqrt{\lambda} D \subseteq \sqrt{\lambda} D$, we obtain an improved error bound for empirical minimization.

Corollary 4.8 *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*

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

$$\lambda_2 = \frac{c}{\delta^2} \max \left\{ \frac{b^2}{n} \left(\|Q\|_{L_2}^2 h_1^2 \log^3 n \right), \frac{\|Z\|_{L_2}^2}{n} \left(h_2 P \ell_{\beta^*} \right) \right\},$$

where $Z = \max_{1\le i\le n} \|A^* X_i\|_{\ell_2^d}$, $Q = \max_{1\le i\le n} \|X_i\|_{\ell_\infty^d}$, *and* A *is the linear operator satisfying* $D = AB_2^d$. *Then* $P\mathcal{L}_{\hat{\beta}} \le \min\{\lambda_1, \lambda_2\}$ *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 (i.e. $D = B_2^d$).

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

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

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

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

Note that this implies that empirical minimization over the simplex gives the optimal rate for convex aggregation up to log factors, as mentioned in Section 1.

For the second example, assume that μ is an isotropic, log-concave measure on $\mathbb{R}^d$. As we showed above, in this case $\|Q\|_{L_2} \le c\log nd \le ch_1^2$. To bound Z in a sharp way, we will use a deep result of Paouris [32]:

Theorem 4.9 *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* $d \le n \le \exp(c_1\sqrt{d})$ *and* $X_1, ..., X_n$ *are independent copies of* X *then*

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

Thus, one obtains an estimate on λ_1 and λ_2:

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

Again, the error rate is

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

up to a poly-logarithmic factor in n and d.

5 An oracle inequality for error rates

Now that we have good bounds for the complexity of the bodies bB_1^d, we can prove the oracle inequality (Theorem 1.2) that was mentioned in the introduction. It is important to note once again that the oracle inequality we obtain is exact (that is, with constant 1). The price for this exact inequality is the resulting slow rate of $1/\sqrt{n}$. However, as shown in [22], our methods can be used to get an oracle inequality with a leading constant of $1+\epsilon$, and much faster rates.

Our main result in this section, is that if $\|X_i\|_\infty$ is bounded almost surely then the LASSO estimator

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

performs almost as well as the empirical minimizer over bB_1^d for the best choice of b. For convenience, let us denote the approximation error by

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

Clearly, $\mathcal{A}_d(b)$ is a decreasing function of b. In general, we would expect it to be bounded away from zero, 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 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-isomorphic" result we need is very similar to one which first appeared in [3] and has appeared several times since then.

Theorem 5.1 *[29] 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, $G_\lambda = \{\theta \mathcal{L}_f : 0 \leq \theta \leq 1 \text{ and } \theta P\mathcal{L}_f = \lambda\}$). 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 $f \in F$*

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

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, $\max\{\sup_{f\in F}\|f\|_\infty, \|Y\|_\infty\} \le \max\{1,b\}M$ and we obtain the following corollary of Theorem 5.1, Theorem 3.6 and Lemma 4.2:

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

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

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 $\{F_r; r\ge 1\}$ be a collection of subsets of F. We say that $\{F_r; r\ge 1\}$ is an* ordered, parameterized hierarchy of F *if the following conditions hold:*

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

Define, for $f \in F$,

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

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 $r \ge 1$ 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 trivial to check is the third condition. A proof of this fact is given in [29] 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 was 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 $f \in F$,*

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

then a regularized minimizer

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

satisfies

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

where c and c' are absolute constants.

Note that the hypothesis in Theorem 5.4 is one that we are prepared for: 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 every F_r simultaneously. 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 $f \in F_r$,*

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

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

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

where

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

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 $i \ge 1$, $u_i = u + \ln(\pi^2/6) + 2\ln i$. 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 $i \ge 1$,

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

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 $r \ge 1$ 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

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

while in the other direction, we get

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

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

$$\begin{aligned} r_i &\leq 2r_{i-1} \\ P\mathcal{L}_{r_i, f^*_{r_{i-1}}} &\leq \rho_n(r_i, u_i). \end{aligned} \tag{5.3}$$

Note that choosing the largest number is not a problem because both $\rho_n(r,u)$ and $P\mathcal{L}_{r,f^*_{r_{j-1}}}$ 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 $i \geq 1$,

$$i \leq \frac{P\ell_{f^*_{r_1}}}{\rho_n(r_1,u_1)} - \frac{P\ell_{f^*_{r_i}}}{\rho_n(r_i,u_i)} + \log_2(2r_i) \leq \frac{P\ell_{f^*_{r_1}}}{\rho_n(r_1,u_1)} + \log_2(2r_i). \tag{5.4}$$

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 $P\ell_{f^*_{r_{i-1}}} = P\ell_{f^*_{r_i}} + \rho_n(r_i,u_i)$. 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

$$i - 1 \leq \frac{P\ell_{f^*_{r_1}}}{\rho_n(r_1,u_1)} - \frac{P\ell_{f^*_{r_{i-1}}}}{\rho_n(r_{i-1},u_{i-1})} + \log_2(2r_{i-1})$$

then

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

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

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

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 $f \in F_r$,

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

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 an oracle inequality for our problem:

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 $\mathbb{R}^{d_n} \times \mathbb{R}$. For every n, for $(X,Y) \sim \mu_n$, assume that $\|X\|_{\ell_\infty^{d_n}} \leq M$ and $|Y| \leq M$ 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

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

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

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

Proof. With Corollary 5.2 in mind, define

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

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

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

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

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

■

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 $\|\beta\|_1$. In addition, it contains the (unknown) approximation error $\mathcal{A}_{d_n}(r)$. Our next and final proof 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 b, 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,

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

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 a proof of this section's main result:

Proof of Theorem 1.2. Define

$$\begin{aligned}\tilde{\rho}_n(r, u) = c(1 + M^2) \max \Bigg\{ & r \frac{\log^{3/2} n \log^{1/2}(d_n n)}{\sqrt{n}}, \\ & r^2 \frac{\log^3 n \log(d_n n)}{n}, r^2 \frac{u}{n}, \frac{r^2 \log\log r}{n} \Bigg\}\end{aligned}$$

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

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

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

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

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,

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

Indeed, the first case is immediate and 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 $S_n(\beta) \leq T_n(\beta)$, which implies that $\inf_\beta S_n(\beta) \leq \inf_\beta T_n(\beta)$ 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 $\hat{T}_n(\alpha) \geq \hat{S}_n(\alpha)$. Since $\hat{T}_n(\beta) \leq \hat{S}_n(\beta)$ 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] P. L. Bartlett. Fast rates for estimation error and oracle inequalities for model selection. *Econometric Theory*, 24(02):545–552, 2008.

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

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

[4] P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of LASSO and Dantzig selector. *The Annals of Statistics*, 39(4):1705–1732, 2009.

[5] F. Bunea, A. Tsybakov, and M. Wegkamp. Sparsity oracle inequalities for the Lasso. *Electron. J. Stat.*, 1:169–194 (electronic), 2007.

[6] F. Bunea, A. B. Tsybakov, and M. H. Wegkamp. Aggregation and sparsity via ℓ_1 penalized least squares. In *Proceedings of the 19th Annual Conference on Learning Theory (COLT 2006)*, volume 4005 of *Lecture Notes in Artificial Intelligence*, pages 379–391. Springer, 2006.

[7] E. J. Candès and Y. Plan. Near-ideal model selection by l1 minimization. *The Annals of Statistics*, 37(5A):2145–2177, 2009.

[8] B. 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.

[9] O. Catoni. *Statistical Learning Theory and Stochastic Optimization*, volume 1851 of *Ecole d'Eté de Probabilités de Saint-Flour 2001, Lecture Notes in Mathematics*. Springer, 2004.

[10] V. H. de la Peña and E. Giné. *Decoupling: From dependence to independence*. Probability and its Applications (New York). Springer-Verlag, New York, 1999.

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

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

[13] R. M. Dudley. The sizes of compact subsets of Hilbert space and continuity of Gaussian processes. In E. Giné, V. Koltchinskii, and R. Norvaisa, editors, *Selected Works of R.M. Dudley*, Selected Works in Probability and Statistics, pages 125–165. Springer New York, 2010.

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

[15] 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.

[16] E. Greenshtein. Best subset selection, persistence in high-dimensional statistical learning and optimization under ℓ_1 constraint. *The Annals of Statistics*, 34(5):2367–2386, 2006.

[17] E. Greenshtein and Y. Ritov. Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli*, 10(6):971–988, 2004.

[18] O. Guédon, S. Mendelson, A. Pajor, and N. Tomczak-Jaegermann. Subspaces and orthogonal decompositions generated by bounded orthogonal systems. *Positivity*, 11(2):269–283, 2007.

[19] 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.

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

[21] V. Koltchinskii. Sparsity in penalized empirical risk minimization. *Annales de l'Institut Henri Poincaré-Probabilités et Statistiques*, 45(1):7–57, 2009.

[22] G. Lecué and S. Mendelson. General oracle inequalities and applications to high dimensional data analysis, preprint.

[23] C. Leng, Y. Lin, and G. Wahba. A note on the lasso and related procedures in model selection. *Statist. Sinica*, 16(4):1273–1284, 2006.

[24] K. Lounici. Sup-norm convergence rate and sign concentration property of Lasso and Dantzig estimators. *Electronic Journal of Statistics*, 2:90–102, 2008.

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

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

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

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

[29] S. Mendelson and J. Neeman. Regularization in Kernel Learning. *Ann. Stat.*, 30(1):526–565, 2010.

[30] V. D. Milman and G. Schechtman. *Asymptotic theory of finite-dimensional normed spaces*, volume 1200 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 1986.

[31] A. Pajor and N. 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.

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

[33] G. Pisier. Some applications of the metric entropy condition to harmonic analysis. *Banach Spaces, Harmonic Analysis, and Probability Theory*, pages 123–154, 1983.

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

[35] M. Talagrand. Regularity of gaussian processes. *Acta Mathematica*, 159:99–149, 1987. 10.1007/BF02392556.

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

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

[38] A. B. Tsybakov. Optimal rates of aggregation. In *Computational Learning Theory 2003*, volume 2777 of *Lecture Notes in Artificial Intelligence*, pages 303–313. Springer, 2003.

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

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

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

[42] T. Zhang. Some sharp performance bounds for least squares regression with ℓ_1 regularization. *The Annals of Statistics*, 37(5A):2109–2144, 2009.