Regularization and the small-ball method I: sparse recovery

Guillaume Lecué[1,3]   Shahar Mendelson[2,4,5]

January 20, 2016

Abstract

We obtain bounds on estimation error rates for regularization procedures of the form

$$\hat{f} \in \underset{f \in F}{\operatorname{argmin}} \left(\frac{1}{N} \sum_{i=1}^{N} (Y_i - f(X_i))^2 + \lambda \Psi(f) \right)$$

when Ψ is a norm and F is convex.

Our approach gives a common framework that may be used in the analysis of learning problems and regularization problems alike. In particular, it sheds some light on the role various notions of sparsity have in regularization and on their connection with the size of subdifferentials of Ψ in a neighbourhood of the true minimizer.

As 'proof of concept' we extend the known estimates for the LASSO, SLOPE and trace norm regularization.

1 Introduction

The focus of this article is on *regularization*, which is one of the most significant methods in modern statistics. To give some intuition on the method and on the reasons behind its introduction, consider the following standard problem.

Let (Ω, μ) be a probability space and set X to be distributed according to μ. F is a class of real-valued functions defined on Ω and Y is the unknown random variable that one would like to approximate using functions in F. Specifically, one would like to identify the best approximation to Y in F, say in the L_2 sense, and find the function f^* that minimizes in F the *squared loss functional* $f \to \mathbb{E}(f(X) - Y)^2$; that is,

$$f^* = \operatorname{argmin}_{f \in F} \mathbb{E}(f(X) - Y)^2,$$

with the underlying assumption that f^* exists and is unique.

Unlike problems in approximation theory, neither the target Y nor the underlying measure μ are known. Therefore, computing the L_2 distance between functions in F and Y is

[1] CNRS, CREST, ENSAE, 3, avenue Pierre Larousse, 92245 MALAKOFF. France.

[2] Department of Mathematics, Technion, I.I.T., Haifa, Israel and Mathematical Sciences Institute, The Australian National University, Canberra, Australia

[3] Email: guillaume.lecue@ensae.fr

[4] Email: shahar@tx.technion.ac.il

[5] Supported by the Israel Science Foundation.

impossible. Instead, one is given partial information: a random sample $(X_i, Y_i)_{i=1}^N$, selected independently according to the joint distribution of X and Y.

Because of the random nature of the sample and the limited information it provides, there is no real hope of identifying f^*, but rather, only of approximating it. In an *estimation problem* one uses the sample to produce a random function $\hat{f} \in F$, and the success of the choice is measured by the distance between $\hat{f}$ and f^* in the $L_2(\mu)$ sense. Thus, one would like to ensure that with high probability with respect to the samples $(X_i, Y_i)_{i=1}^N$, the *error rate*

$$\left\| \hat{f} - f^* \right\|_{L_2(\mu)}^2 = \mathbb{E}\Big(\big(\hat{f}(X) - f^*(X)\big)^2 | (X_i, Y_i)_{i=1}^N \Big)$$

is small. More accurately, the question is to identify the way in which the error rate depends on the structure of the class F and scales with the sample size N and the required degree of confidence (probability estimate).

It is not surprising (and rather straightforward to verify) that the problem becomes harder the larger F is. In contrast, if F is small, chances are that $f^*(X)$ is very far from Y, and identifying it, let alone approximating it, is pointless.

In situations we shall refer to as *learning problems*, the underlying assumption is that F is indeed small, and the issue of the approximation error – the distance between Y and f^* is ignored.

While the analysis of learning problems is an important and well-studied topic, the assumption that F is reasonably small seems somewhat restrictive; it certainly does not eliminate the need for methods that allow one to deal with very large classes.

Regularization was introduced as an alternative to the assumption on the 'size' of F. One may consider large classes, but combine it with the belief that f^* belongs to a relatively small substructure in F. The idea is to penalize a choice of a function that is far from that substructure, which forces the learner to choose a function in the 'right part' of F.

Formally, let E be a vector space, assume that $F \subset E$ is a closed and convex set and let $\Psi : E \to \mathbb{R}^+$ be the penalty. Here, we will only consider the case in which Ψ is a norm on E.

Let $\lambda > 0$ and for a sample $(X_i, Y_i)_{i=1}^N$, set

$$\hat{f} \in \operatorname*{argmin}_{f \in F} \left(\frac{1}{N} \sum_{i=1}^N (Y_i - f(X_i))^2 + \lambda \Psi(f) \right);$$

$\hat{f}$ is called a regularization procedure, Ψ is the regularization function and λ is the regularization parameter.

In the classical approach to regularization, the substructure of f^* is quantified directly by Ψ. The underlying belief is that $\Psi(f^*)$ is not 'too big' and one expects the procedure to produce $\hat{f}$ for which $\Psi(\hat{f})$ is of the order of $\Psi(f^*)$. Moreover, the anticipated error rate $\|\hat{f} - f^*\|_{L_2(\mu)}$ depends on $\Psi(f^*)$. In fact, an optimistic viewpoint is that regularization could perform as well as the best learning procedure in the class $\{f : \Psi(f) \leq \Psi(f^*)\}$, but without knowing $\Psi(f^*)$ beforehand.

Among the regularization schemes that are based on the classical approach are reproducing kernel Hilbert spaces (RKHS), in which the RKHS norm serves at the penalty. Since RKHS norms capture various notions of smoothness, in RKHS regularization one is driven towards a choice of a smooth $\hat{f}$ – as smooth as f^* is.

In more modern regularization problems the situation is very different. Even when penalizing with a norm Ψ, one no longer cares whether or not $\Psi(f^*)$ is small; rather, one knows (or at least believes) that f^* is *sparse* in some sense, and the hope is that this sparsity will be reflected in the error rate.

In other words, although one uses certain norms as regularization functions – norms that seemingly have nothing to do with 'sparsity' – the hope is that the sparse nature of f^* will be exposed by the regularization procedure, while $\Psi(f^*)$ will be of little importance.

The most significant example in the context of sparsity-driven regularization is the celebrated LASSO estimator [34]. Let $F = \{\langle t, \cdot \rangle : t \in \mathbb{R}^d\}$ and set t^* to be a minimizer in $\mathbb{R}^d$ of the functional $t \to \mathbb{E}(\langle t, X\rangle - Y)^2$. The LASSO is defined by

$$\hat{t} \in \underset{t \in \mathbb{R}^d}{\operatorname{argmin}} \Big(\frac{1}{N} \sum_{i=1}^{N} \big(\langle t, X_i \rangle - Y_i\big)^2 + \lambda \Psi(t) \Big)$$

for the choice $\Psi(t) = \|t\|_1 = \sum_{i=1}^d |t_i|$.

The remarkable property of the LASSO (see [8] and [3]) is that for a well-chosen regularization parameter λ, if t^* is supported on at most s coordinates (and under various assumptions on X and Y to which we will return later), then with high probability,

$$\|\hat{t} - t^*\|_2^2 \lesssim \frac{s \log(ed)}{N}.$$

Thus, the error rate of the LASSO does not depend on $\Psi(t^*) = \|t^*\|_1$, but rather on the degree of sparsity of t^*, measured here by the cardinality of its support $\|t^*\|_0 = |\{i : t_i^* \neq 0\}|$.

This fact seems almost magical, because to the naked eye, the regularization function $\|t\|_1$ has nothing to with sparsity; yet ℓ_1 regularization leads to a sparsity-driven error rate.

A standard (yet somewhat unconvincing) explanation of this phenomenon is that the penalty $\|t\|_1$ is a convexified version of $\|t\|_0 = |\{i : t_i \neq 0\}|$, though this loose connection hardly explains why $\|t^*\|_0$ has any effect on the error rate of the LASSO.

A similar phenomenon occurs for other choices of Ψ, such as the SLOPE and trace-norm regularization, which will be explored in detail in what follows. In all these cases and others like them, the regularization function is a norm that does not appear to be connected to sparsity, nor to other natural notions of low-dimensional structures for that matter. Yet, and quite mysteriously, the respective regularization procedure emphasize those very properties of t^*.

The aim of this note is to offer a framework that can be used to tackle standard learning problems (small F) and regularized problems alike. Moreover, using the framework, one may explain how certain norms lead to the emergence of sparsity-based bounds.

In what follows we will show that two parameters determine the error rate of regularization problems. The first one captures the 'complexity' of each set in the natural hierarchy in F

$$F_\rho = \{f \in F : \Psi(f - f^*) \le \rho\}.$$

Applying results from [22, 24, 21], the 'complexity' of each F_ρ turns out to be the optimal (in the minimax sense) error rate of the learning problem in that set.

To be more precise, the main ingredient in obtaining a sharp error rate in a learning problem in a class H is an accurate analysis of the empirical excess squared loss functional

$$f \to P_N \mathcal{L}_f = \frac{1}{N}\sum_{i=1}^{N}(f(X_i) - Y_i)^2 - \frac{1}{N}\sum_{i=1}^{N}(f^*(X_i) - Y_i)^2. \tag{1.1}$$

Since the minimizer $\hat{f}$ of the functional (1.1) satisfies $P_N \mathcal{L}_{\hat{f}} \le 0$, one may obtain an estimate on the error rate by showing that with high probability, if $\|f - f^*\|_{L_2(\mu)} \ge r$ then $P_N \mathcal{L}_f > 0$. This excludes functions in the set $\{f \in H : \|f - f^*\|_{L_2(\mu)} \ge r\}$ as potential empirical minimizers.

One may show that this 'critical level' is the correct (minimax) error rate of a learning problem in H, and that the same parameter is of central importance in regularization problems: namely, the 'critical level' $r(\rho)$ for each one of the sets $\{f \in F : \Psi(f - f^*) \le \rho\}$ (see Section 2.1 for an accurate definition of $r(\rho)$ and its role in the analysis of learning problems and regularization problems).

The second parameter, which is the main ingredient in our analysis of regularization problems, measures the 'size' of the subdifferential of Ψ in points that are close to f^* – recall that the subdifferential of Ψ in f is

$$(\partial\Psi)_f = \{z^* \in E^* : \Psi(f + h) \ge \Psi(f) + z^*(h) \text{ for every } h \in E\}$$

where E^* is the dual space of (E, Ψ).

Indeed, fix $\rho > 0$, and let $\Gamma_{f^*}(\rho)$ be the collection of functionals that belong to the subdifferential $(\partial\Psi)_f$ for some $f \in F$ that satisfies $\Psi(f - f^*) \le \rho/20$. Set

$$H_\rho = \{f \in F : \Psi(f - f^*) = \rho \text{ and } \|f - f^*\|_{L_2(\mu)} \le r(\rho)\}$$

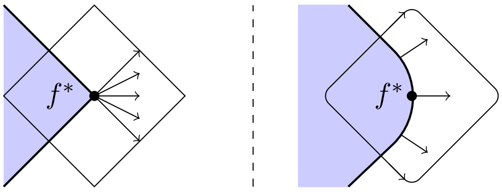
and let

$$\Delta(\rho) = \inf_{h \in H_\rho} \sup_{z^* \in \Gamma_{f^*}(\rho)} z^*(h - f^*).$$

It is well known that the subdifferential of a norm Ψ in f consists of all the norm one linear functionals z^* for which $z^*(f) = \Psi(f)$. Hence, $\Gamma_{f^*}(\rho)$ is a subset of the unit sphere of E^*. And, since H_ρ consists of functions whose Ψ norm is ρ, $\Delta(\rho) \le \rho$. Therefore, the fact that $\Delta(\rho) \ge \alpha\rho$ implies that $\Gamma_{f^*}(\rho)$ is rather large: for every $h \in H_\rho$ there is some $z^* \in \Gamma_{f^*}(\rho)$ for which $z^*(h)$ is 'almost extremal': at least $\alpha\rho$.

Our main result (Theorem 3.2 below) is that if $\Gamma_{f^*}(\rho)$ is large enough to ensure that $\Delta(\rho) \ge 4\rho/5$, and the regularization parameter λ is set to be of the order of $\frac{r^2(\rho)}{\rho}$, then with high probability, the regularized minimizer in F, $\hat{f}$, satisfies that $\|\hat{f} - f^*\|_{L_2(\mu)} \le r(\rho)$ and $\Psi(\hat{f} - f^*) \le \rho$.

Theorem 3.2 implies that one may address a regularization problem by selecting ρ wisely, keeping in mind that points in a Ψ-ball of radius $\sim \rho$ around f^* must generate a sufficiently large class of subdifferentials.



It is essential that the functionals in $\Gamma_{f^*}(\rho)$ need to be 'almost extremal' only for points in H_ρ rather than for the entire sphere; otherwise, it would have forced $\Gamma_{f^*}(\rho)$ to be unreasonably large – close to the entire dual sphere.

One may show that sparsity, combined with the right choice of Ψ, contributes in two places: firstly, if f^* is sparse in some sense, and Ψ is not smooth on sparse elements, then $\Gamma_{f^*}(\rho)$, which contains the subdifferential $(\partial\Psi)_{f^*}$ is large; secondly, for the right choice of ρ the 'localization' H_ρ consists of elements that are well placed: if $\Psi(f - f^*) = \rho$ and $\|f - f^*\|_{L_2(\mu)} \leq r(\rho)$, there is some $z^* \in \Gamma_{f^*}(\rho)$ for which $z^*(f - f^*)$ is large enough – and thus $\Delta(\rho)$ is large enough. The fact that H_ρ is well placed is an outcome of some compatibility between Ψ and the $L_2(\mu)$ norm.

Of course, to find the right choice of ρ one must first identify $r(\rho)$, which is, in itself, a well-studied yet nontrivial problem.

Before we dive into technical details, let us formulate some outcomes of our main result. We will show how it can be used to obtain sparsity-driven error rates in three regularization procedures: the LASSO, SLOPE and trace norm regularization. In all three cases our results actually extend the known estimates in various directions.

The LASSO.

The LASSO is obtained when $F = \{\langle t, \cdot \rangle : t \in \mathbb{R}^d\}$. Identifying the linear functional $\langle t, \cdot \rangle$ with t, the regularization function is

$$\Psi(t) = \|t\|_1 = \sum_{i=1}^{d} |t_i|,$$

and the resulting regularization problem is

$$\hat{t} \in \underset{t \in \mathbb{R}^d}{\operatorname{argmin}} \left(\frac{1}{N} \sum_{i=1}^{N} \left(\langle t, X_i \rangle - Y_i \right)^2 + \lambda \|t\|_1 \right).$$

The LASSO has been studied extensively in the last two decades. It has been THE benchmark estimator of high-dimensional statistics because the high dimensional parameter space does not significantly affect its performance as long as t^* is sparse. This is indeed the case for estimation and sparse oracle inequalities in [3, 15, 36, 37, 19, 28, 35], support recovery results in [17, 39, 2] or testing and confidence intervals results in [16, 18]. We refer the reader to the books [5, 8] for more results and references on the LASSO.

SLOPE.

In some sense, SLOPE, introduced in [4, 32], is actually an extension of the LASSO, even though it has been introduced as an extension of multiple-test procedures. Again, the underlying class is $F = \{\langle t, \cdot \rangle : t \in \mathbb{R}^d\}$, and to define the regularization function let $\beta_1 \geq \beta_2 \geq ... \geq \beta_d > 0$ and set

$$\Psi(t) = \sum_{i=1}^{d} \beta_i t_i^*,$$

where $(t_i^*)_{i=1}^d$ denotes the non-increasing re-arrangement of $(|t_i|)_{i=1}^d$. Thus, the SLOPE norm is a sorted, weighted ℓ_1-norm, and for $(\beta_1, ..., \beta_d) = (1, ..., 1)$, SLOPE regularization coincides with the LASSO.

Trace-norm regularization.

Consider the trace inner-product on $\mathbb{R}^{m \times T}$. Let $F = \{\langle A, \cdot \rangle : A \in \mathbb{R}^{m \times T}\}$ and given a target Y put A^* to be the matrix that minimizes $A \to \mathbb{E}(\langle A, X \rangle - Y)^2$.

The regularization function is the *trace norm.*

Definition 1.1 *Let A be a matrix and set $(\sigma_i(A))$ to be its singular values, arranged in a non-increasing order. For $p \geq 1$, $\|A\|_p = (\sum \sigma_i^p(A))^{1/p}$ is the p-Schatten norm. The trace-norm is the 1-Schatten norm, the Hilbert-Schmidt norm is the 2-Schatten norm and the operator norm is the ∞-Schatten norm.*

The trace norm regularization procedure is

$$\hat{A} \in \underset{A \in \mathbb{R}^{m \times T}}{\operatorname{argmin}} \left(\frac{1}{N} \sum_{i=1}^{N} (Y_i - \langle X_i, A \rangle)^2 + \lambda \|A\|_1 \right)$$

and it was introduced for the reconstruction of low-rank, high-dimensional matrices [29, 9, 30, 6, 7, 27].

As will be explained in what follows, our main result holds in rather general situations and it may be implemented in examples once the ‘critical levels’ $r(\rho)$ are identified. Since the examples we present serve mainly as “proof of concept”, we will focus only on one scenario in which $r(\rho)$ may be completely characterized for an arbitrary class of functions.

Definition 1.2 *Let ℓ_2^M be an M-dimensional inner product space and let μ be a measure on ℓ_2^M. The measure μ is isotropic if for every $x \in \ell_2^M$,*

$$\int \langle x, t \rangle^2 d\mu(t) = \|x\|_{\ell_2^M}^2;$$

it is L-subgaussian if for every $p \geq 2$ and every $x \in \ell_2^M$,

$$\|\langle x, \cdot \rangle\|_{L_p(\mu)} \leq L\sqrt{p} \|\langle x, \cdot \rangle\|_{L_2(\mu)}.$$

Hence, the covariance structure of an isotropic measure coincides with the inner product in ℓ_2^M, and if μ is an L-subgaussian measure the $L_p(\mu)$ norm of a linear form does not grow faster than the L_p norm of the corresponding gaussian variable.

Assumption 1.1 *Assume that the underlying measure μ is isotropic and L-subgaussian, and that for $f^* = \langle t^*, \cdot \rangle$ (or $f^* = \langle A^*, \cdot \rangle$), the noise*[1] *$\xi = f^*(X) - Y$ belongs to L_q for some $q > 2$.*

When dealing with the LASSO and SLOPE, the natural Euclidean structure is the standard one in $\mathbb{R}^d$, and for trace norm regularization, the Euclidean structure is endowed by the trace inner product in $\mathbb{R}^{m \times T}$.

The second part of Assumption 1.1, that $\xi \in L_q$ for some $q > 2$, is rather minimal. Indeed, for the functional $f \to \mathbb{E}(f(X) - Y)^2$ to be well defined, one must assume that $f(X) - Y \in L_2$; the assumption here is only slightly stronger.

Applying our main result we will show the following:

Theorem 1.3 *Consider the LASSO under Assumption 1.1. Let $0 < \delta < 1$. Assume that there is some $v \in \mathbb{R}^d$ supported on at most s coordinates for which*

$$\|t^* - v\|_1 \leq c_1(\delta) \|\xi\|_{L_q} s \sqrt{\frac{\log(ed)}{N}}.$$

If $\lambda = c_2(L, \delta)\|\xi\|_{L_q}\sqrt{\log(ed)/N}$, then with probability at least $1 - \delta$ the LASSO estimator with regularization parameter λ satisfies that for every $1 \leq p \leq 2$

$$\left\|\hat{t} - t^*\right\|_p \leq c_3(L, \delta) \|\xi\|_{L_q} s^{1/p} \sqrt{\frac{\log(ed)}{N}}.$$

The error rate in Theorem 1.3 coincides with the standard estimate on the LASSO (cf. [3]), but in a broader context: t^* need not be sparse but only approximated by a sparse vector; the target Y is arbitrary and the noise ξ may be heavy tailed and need not be independent of X.

Turning to SLOPE, let us recall the estimates from [32], where the setup is somewhat restricted:

Let X be a gaussian vector on $\mathbb{R}^d$, set W to be a gaussian random variable with variance σ^2 that is independent of X and put $Y = \langle t^*, X \rangle + W$. Consider some $q \in (0, 1)$, let $\Phi^{-1}(\alpha)$ be the α-th quantile of the standard normal distribution and put $\beta_i = \Phi^{-1}(1 - iq/(2d))$.

[1] In what follows we will refer to ξ as 'the noise' even though it depends in general on Y and X. The reason for using that term comes from the situation in which $Y = f^*(X) - W$ for a symmetric random variable W that is independent of X (independent noise); thus $\xi = W$. We have opted to call ξ 'the noise' because its role in the general case and its impact on the error rate is rather similar to what happens for independent noise.

Theorem 1.4 *[32] Let $1 \leq s \leq d$ satisfy that $s/d = o(1)$ and $(s \log d)/N = o(1)$ when $N \to \infty$. If $0 < \varepsilon < 1$, $N \to \infty$ and $\lambda = 2\sigma/\sqrt{N}$, the SLOPE estimator with weights $(\beta_i)_{i=1}^d$ and regularization parameter λ satisfies*

$$\sup_{\|t^*\|_0 \leq s} Pr\Big(\frac{N \left\| \hat{t} - t^* \right\|_2^2}{2\sigma^2 s \log(d/s)} > 1 + 3\varepsilon \Big) \to 0.$$

Note that Theorem 1.4 is asymptotic in nature and not 'high-dimensional'. Moreover, it only holds for a gaussian X, independent gaussian noise W, a specific choice of weights $(\beta_i)_{i=1}^d$ and t^* that is s-sparse.

We consider a more general situation. Let $\beta_i \leq C\sqrt{\log(ed/i)}$ and set $\Psi(t) = \sum_{i=1}^d t_i^* \beta_i$.

Theorem 1.5 *There exists constants c_1, c_2 and c_3 that depend only on L, δ and C for which the following holds.*

Under Assumption 1.1, if there is $v \in \mathbb{R}^d$ that satisfies $|\text{supp}(v)| \leq s$ and

$$\Psi(t^* - v) \leq c_1 \|\xi\|_{L_q} \frac{s}{\sqrt{N}} \log\left(\frac{ed}{s}\right),$$

then for $N \geq c_2 s \log(ed/s)$ and with the choice of $\lambda = c_2 \|\xi\|_{L_q}/\sqrt{N}$, one has

$$\Psi(\hat{t} - t^*) \leq c_3 \|\xi\|_{L_q} \frac{s}{\sqrt{N}} \log\left(\frac{ed}{s}\right) \quad \text{and} \quad \left\| \hat{t} - t^* \right\|_2^2 \leq c_3 \|\xi\|_{L_q}^2 \frac{s}{N} \log\left(\frac{ed}{s}\right)$$

with probability at least $1 - \delta$.

Finally, let us consider trace norm regularization.

Theorem 1.6 *Under Assumption 1.1 and if there is $V \in \mathbb{R}^{m \times T}$ that satisfies that $\text{rank}(V) \leq s$ and*

$$\|A^* - V\|_1 \leq c_1 \|\xi\|_{L_q} s \sqrt{\frac{\max\{m, T\}}{N}},$$

one has the following. Let $N \geq c_2 s \max\{m, T\}$ and $\lambda = c_3 \|\xi\|_{L_q} \sqrt{\frac{\max\{m,T\}}{N}}$. Then with probability at least $1 - \delta$, for any $1 \leq p \leq 2$

$$\left\| \hat{A} - A^* \right\|_p \leq c_4 \|\xi\|_{L_q} s^{1/p} \sqrt{\frac{\max\{m, T\}}{N}}.$$

The constants c_1, c_2, c_3 and c_4 depends only on L and δ.

A result of a similar flavour to Theorem 1.6 is Theorem 9.2 from [8].

Theorem 1.7 *Let X be an isotropic and L-subgaussian vector, and W that is mean-zero, independent of X and belongs to the Orlicz space L_{ψ_α} for some $\alpha \geq 1$. If $Y = \langle A^*, X \rangle + W$ and*

$$\lambda \geq c_1(L) \max\left\{ \|\xi\|_2 \sqrt{\frac{m(t + \log m)}{N}}, \|\xi\|_{\psi_\alpha} \log^{1/\alpha}\left(\frac{\|\xi\|_{\psi_\alpha}}{\|\xi\|_{L_2}}\right) \frac{\sqrt{m}(t + \log N)(t + \log m)}{N} \right\},$$

then with probability at least $1 - 3\exp(-t) - \exp(-c_2(L)N)$

$$\left\| \hat{A} - A^* \right\|_2^2 \leq c_3 \min\left\{ \lambda \left\| A^* \right\|_1, \lambda^2 \mathrm{rank}(A^*) \right\}. \tag{1.2}$$

Clearly, the assumptions of Theorem 1.7 are more restrictive than those of Theorem 1.6, as the latter holds for a heavy tailed ξ that need not be independent of X, and for A^* that can be approximated by a low-rank matrix. Moreover, if $\|A^*\|_1$ is relatively large and the error rate in Theorem 1.7 is the sparsity-dominated $\lambda^2\mathrm{rank}(A^*)$, then the error rate in Theorem 1.6 is better by a logarithmic factor.

The proofs of the error rates in all the three examples will be presented in Section 5.

1.1 Notation

We end the introduction with some standard notation.

Throughout, absolute constants are denoted by $c, c_1 ...$, etc. Their value may change from line to line. When a constant depends on a parameter α it will be denoted by $c(\alpha)$. $A \lesssim B$ means that $A \leq cB$ for an absolute constant c, and the analogous two-sided inequality is denoted by $A \sim B$. In a similar fashion, $A \lesssim_\alpha B$ implies that $A \leq c(\alpha) B$, etc.

Let $E \subset L_2(\mu)$ be a vector space and set Ψ to be a norm on E. For a set $A \subset E$, $t \in E$ and $r > 0$, let $rA + t = \{ra + t : a \in A\}$.

Denote by $B_\Psi = \{w \in E : \Psi(w) \leq 1\}$ the unit ball of (E, Ψ) and set $S_\Psi = \{f \in E : \Psi(f) = 1\}$ to be the corresponding unit sphere. $B_\Psi(\rho, f)$ is the ball of radius ρ centred in f and $S_\Psi(\rho, f)$ is the corresponding sphere. Also, set D to be the unit ball in $L_2(\mu)$, S is the unit sphere there, and $D(\rho, f)$ and $S(\rho, f)$ are the ball and sphere centred in f and of radius ρ, respectively.

A class of spaces we will be interested in are ℓ_p^d, that is, $\mathbb{R}^d$ endowed with the ℓ_p norm; B_p^d denotes the unit ball in ℓ_p^d and $S(\ell_p^d)$ is the unit sphere.

For every $x = (x_i)_{i=1}^d$, $(x_i^*)_{i=1}^d$ denotes the non-increasing rearrangement of $(|x_i|)_{i=1}^d$.

Finally, if $(X_i, Y_i)_{i=1}^N$ is a sample, $P_N h = \frac{1}{N}\sum_{i=1}^N h(X_i, Y_i)$ is the empirical mean of h.

2 Preliminaries: The regularized functional

Let $F \subset E$ be a closed and convex class of functions. Recall that for target Y, f^* is the minimizer in F of the functional $f \to \mathbb{E}(f(X) - Y)^2$. Since F is closed and convex, the minimum exists and is unique.

Let $\mathcal{L}_f(X, Y) = (f(X) - Y)^2 - (f^*(X) - Y)^2$ be the excess squared loss functional and for $\lambda > 0$ let

$$\mathcal{L}_f^\lambda(X, Y) = \mathcal{L}_f + \lambda(\Psi(f) - \Psi(f^*))$$

be its regularized counterpart. Thus, for a random sample $(X_i, Y_i)_{i=1}^N$, the empirical (regularized) excess loss functional is

$$P_N \mathcal{L}_f^\lambda = \frac{1}{N} \sum_{i=1}^N \mathcal{L}_f(X_i, Y_i) + \lambda(\Psi(f) - \Psi(f^*)),$$

Note that if $\ell_f(x,y)=(y-f(x))^2$ and $\hat{f}$ minimizes $P_N\ell_f+\lambda\Psi(f)$ then $\hat{f}$ also minimizes $P_N\mathcal{L}_f^\lambda$. Moreover, since $\mathcal{L}_{f^*}^\lambda=0$, it is evident that $P_N\mathcal{L}_{\hat{f}}^\lambda\leq 0$.

This simple observation shows that the random set $\{f\in F : P_N\mathcal{L}_f^\lambda>0\}$ may be excluded from our considerations, as it does not contain potential minimizers. Therefore, if one can show that with high probability,

$$\{f\in F : P_N\mathcal{L}_f^\lambda\leq 0\}\subset\{f\in F : \|f-f^*\|_{L_2(\mu)}\leq r\},$$

then on that event, $\|\hat{f}-f^*\|_{L_2(\mu)}\leq r$.

We will study when $P_N\mathcal{L}_f^\lambda>0$ by considering the two parts of the empirical functional: the empirical excess loss $P_N\mathcal{L}_f$ and the regularized part $\lambda(\Psi(f)-\Psi(f^*))$.

Because of its crucial role in obtaining error estimates in learning problems, the functional $f\to P_N\mathcal{L}_f$ has been studied extensively using the *small-ball method*, (see, e.g., [22, 24, 21]). Thus, the first component in the machinery we require for explaining both learning problems and regularization problems is well understood and ready-to-use; its details are outlined below.

2.1 The natural decomposition of $P_N\mathcal{L}_f$

Set $\xi=\xi(X,Y)=f^*(X)-Y$ and observe that

$$\begin{aligned}\mathcal{L}_f(X,Y)=&(f-f^*)^2(X)+2(f-f^*)(X)\cdot(f^*(X)-Y)\\ =&(f-f^*)^2(X)+2\xi(f-f^*)(X).\end{aligned}$$

Since F is convex, the characterization of the nearest point map in a Hilbert space shows that

$$\mathbb{E}(f-f^*)(X)\cdot(f^*(X)-Y)\geq 0$$

for every $f\in F$. Hence, setting $\xi_i=f^*(X_i)-Y_i$, one has

$$\begin{aligned}P_N\mathcal{L}_f^\lambda\geq&\frac{1}{N}\sum_{i=1}^N(f-f^*)^2(X_i)+2\Big(\frac{1}{N}\sum_{i=1}^N\xi_i(f-f^*)(X_i)-\mathbb{E}\xi(f-f^*)(X)\Big)\\ &+\lambda(\Psi(f)-\Psi(f^*)).\end{aligned}$$

To simplify notation, for $w\in L_2(\mu)$ set $\mathcal{Q}_w=w^2$ and $\mathcal{M}_w=\xi w-\mathbb{E}\xi w$. Thus, for every $f\in F$,

$$P_N\mathcal{L}_f^\lambda\geq P_N\mathcal{Q}_{f-f^*}+2P_N\mathcal{M}_{f-f^*}+\lambda(\Psi(f)-\Psi(f^*)). \tag{2.1}$$

The decomposition of the empirical excess loss to the quadratic component ($\mathcal{Q}_{f-f^*}$) and the multiplier one ($\mathcal{M}_{f-f^*}$) is the first step in applying the small-ball method to learning problems. One may show that on a large event, if $\|f-f^*\|_{L_2(\mu)}$ is larger than some critical level then $P_N\mathcal{Q}_{f-f^*}\geq\theta\|f-f^*\|_{L_2}^2$ and dominates $P_N\mathcal{M}_{f-f^*}$; hence $P_N\mathcal{L}_f>0$.

To identify this critical level, let us define the following parameters:

Definition 2.1 *Let $H \subset F$ be a convex class that contains f^*. Let $(\varepsilon_i)_{i=1}^N$ be independent, symmetric, $\{-1,1\}$-valued random variables that are independent of $(X_i, Y_i)_{i=1}^N$.*

For $\gamma_Q, \gamma_M > 0$ set

$$r_Q(H,\gamma_Q) = \inf\left\{ r>0 : \mathbb{E} \sup_{h \in H \cap D(r,f^*)} \left| \frac{1}{N}\sum_{i=1}^N \varepsilon_i (h-f^*)(X_i) \right| \leq \gamma_Q r \right\},$$

let

$$\phi_N(H,s) = \sup_{h \in H \cap D(s,f^*)} \left| \frac{1}{\sqrt{N}} \sum_{i=1}^N \varepsilon_i \xi_i (h-f^*)(X_i) \right|,$$

and put

$$r_M(H,\gamma_M,\delta) = \inf\left\{ s>0 : Pr\left(\phi_N(H,s) \leq \gamma_M s^2 \sqrt{N}\right) \geq 1-\delta \right\}.$$

The main outcome of the small-ball method is that for the right choices of γ_M and γ_Q, $r = \max\{r_M, r_Q\}$ is the above-mentioned 'critical level' in H, once H satisfies a weak small-ball condition.

Assumption 2.1 (The small ball condition) *Assume that there are constants $\kappa > 0$ and $0 < \varepsilon \leq 1$, for which, for every $f, h \in F \cup \{0\}$,*

$$Pr\left(|f-h| \geq \kappa \|f-h\|_{L_2(\mu)}\right) \geq \varepsilon.$$

There are numerous examples in which the small-ball condition may be verified for κ and ε that are absolute constants. We refer the reader to [12, 21, 10, 23, 24, 31] for some of them.

Theorem 2.2 ([24]) *Let H be a closed, convex class of functions that contains f^* and satisfies Assumption 2.1 with constants κ and ε. If $\theta = \kappa^2\varepsilon/16$ then for every $0 < \delta < 1$, with probability at least $1-\delta-2\exp(-N\varepsilon^2/2)$ one has:*

- *for every $f \in H$,*

$$|P_N \mathcal{M}_{f-f^*}| \leq \frac{\theta}{8} \max\left\{ \|f-f^*\|_{L_2(\mu)}^2, r_M^2\left(H, \theta/10, \delta/4\right) \right\},$$

- *If $f \in H$ and $\|f-f^*\|_{L_2(\mu)} \geq r_Q\left(H, \kappa\varepsilon/32\right)$ then*

$$P_N \mathcal{Q}_{f-f^*} \geq \theta \|f-f^*\|_{L_2(\mu)}^2.$$

In particular, with probability at least $1-\delta-2\exp(-N\varepsilon^2/2)$,

$$P_N \mathcal{L}_f \geq \frac{\theta}{2} \|f-f^*\|_{L_2(\mu)}^2$$

for every $f \in H$ that satisfies

$$\|f-f^*\|_{L_2(\mu)} \geq \max\left\{ r_M\left(H, \theta/10, \delta/4\right), r_Q\left(H, \kappa\varepsilon/32\right) \right\}.$$

From now on, we will assume that F satisfies the small-ball condition with constants κ and ε, and that $\theta = \kappa^2\varepsilon/16$.

Definition 2.3 *Let $\rho > 0$ and set*

$$r_M(\rho) = r_M\big(F \cap B_\Psi(\rho, f^*), \frac{\theta}{10}, \frac{\delta}{4}\big) \quad \textit{and} \quad r_Q(\rho) = r_Q\big(F \cap B_\Psi(\rho, f^*), \frac{\kappa\varepsilon}{32}\big).$$

In what follows we will abuse notation and omit the dependence of r_M and r_Q on f^*, κ, ε and δ.

Let $r(\cdot)$ be a function that satisfies

$$r(\rho) \geq \sup_{f^* \in F} \max\{r_Q(\rho), r_M(\rho)\}.$$

Finally, put

$$\mathcal{O}(\rho) = \sup_{f \in F \cap B_\Psi(\rho, f^*) \cap D(r(\rho), f^*)} \big|P_N \mathcal{M}_{f-f^*}\big|.$$

Theorem 2.2 implies the following:

Corollary 2.4 ([24]) *Using the notation introduced above, on an event of probability at least $1 - \delta - 2\exp(-N\varepsilon^2/2)$, if $f \in F \cap B_\Psi(\rho, f^*)$ and $\|f - f^*\|_{L_2(\mu)} \geq r(\rho)$ then*

$$P_N \mathcal{L}_f \geq \frac{\theta}{2}\|f - f^*\|_{L_2(\mu)}^2.$$

Moreover, on the same event,

$$\mathcal{O}(\rho) \leq \frac{\theta}{8} r^2(\rho).$$

Remark 2.5 *Let us stress once again that $r(\rho)$ plays a central role in the analysis of empirical risk minimization in the set $F \cap B_\Psi(\rho, f^*)$. Theorem 2.2 implies that with high probability, the empirical risk minimizer $\tilde{h}$ in $F \cap B_\Psi(\rho, f^*)$ satisfies*

$$\|\tilde{h} - h^*\|_{L_2(\mu)} \leq r(\rho).$$

Moreover, if follows from [11] and [25] that under mild structural assumptions on F, $r(\rho)$ is the best possible error rate of any learning procedure in $F \cap B_\Psi(\rho, f^*)$ – i.e., the minimax rate in that class.

Let $\mathcal{A}$ be the event from Corollary 2.4 and set

$$\gamma_\mathcal{O}(\rho) = \sup_{w \in \mathcal{A}} \mathcal{O}(\rho).$$

$\gamma_\mathcal{O}$ will be of little importance in what follows, because it may be upper bounded by $(\theta/8)r^2(\rho)$. However, it will be of the utmost importance in [13], where complexity-based regularization is studied (see Section 6 for more details).

3 The main result

Let us turn to the second part of the regularized functional – namely, $\lambda(\Psi(f)-\Psi(f^*))$. Let E^* be the dual space to (E,Ψ) and set Ψ^* to be the dual norm. B_{Ψ^*} and S_{Ψ^*} denote the dual unit ball and unit sphere, respectively; i.e., B_{Ψ^*} consists of all the linear functionals z^* on E for which $\sup_{\Psi(x)=1}|z^*(x)|\leq 1$.

Definition 3.1 *The functional $z^*\in S_{\Psi^*}$ is a norming functional for $z\in E$ if $z^*(z)=\Psi(z)$.*

In the language of convex functions, a functional is norming for x if and only if it belongs to $(\partial\Psi)_x$, the subdifferential of Ψ in x.

Let $\Gamma_{f^*}(\rho)$ be the collection of functionals that are norming for some $f\in B_\Psi(\rho/20,f^*)$. In particular, $\Gamma_{f^*}(\rho)$ contains all the norming functionals of f^*.

Set

$$\Delta(\rho)=\inf_h \sup_{z^*\in\Gamma_{f^*}(\rho)} z^*(h-f^*),$$

where the infimum is taken in the set

$$F\cap S_\Psi(\rho,f^*)\cap D(r(\rho),f^*)=\{h\in F:\Psi(h-f^*)=\rho \text{ and } \|h-f^*\|_{L_2(\mu)}\leq r(\rho)\}.$$

Note that if $z^*\in\Gamma_{f^*}(\rho)$ and $h\in S_\Psi(\rho,f^*)$ then $|z^*(h-f^*)|\leq\Psi(h-f^*)=\rho$. Thus, a lower bound of the form $\Delta(\rho)\geq(1-\delta)\rho$ implies that $\Gamma_{f^*}(\rho)$ is a relatively large subset of the dual unit sphere: each point in $F\cap S_\Psi(\rho,f^*)\cap D(r(\rho),f^*)$ has an 'almost norming' functional in $\Gamma_{f^*}(\rho)$.

Our main result is that if $\Gamma_{f^*}(\rho)$ is indeed large enough to ensure that $\Delta(\rho)\geq 4/5\rho$ then with high probability $\|\hat{f}-f^*\|_{L_2(\mu)}\leq r(\rho)$ and $\Psi(\hat{f}-f^*)\leq\rho$.

Theorem 3.2 *Let $\rho>0$ and set $\mathcal{A}$ to be an event on which Corollary 2.4 holds. If $\Delta(\rho)\geq 4\rho/5$ and*

$$3\frac{\gamma_{\mathcal{O}}(\rho)}{\rho}\leq\lambda\leq\frac{\theta}{2}\cdot\frac{r^2(\rho)}{\rho},$$

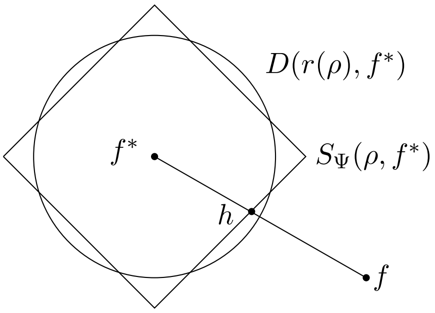
then on the event $\mathcal{A}$, a regularized empirical minimizer $\hat{f}\in\operatorname{argmin}_{f\in F}P_N\mathcal{L}_f^\lambda$ satisfies

$$\Psi(\hat{f}-f^*)\leq\rho \text{ and } \|\hat{f}-f^*\|_{L_2(\mu)}\leq r(\rho).$$

Moreover, since $r_{\mathcal{O}}(\rho)\leq(\theta/8)r^2(\rho)$, the same assertion holds if

$$\frac{3\theta}{8}\cdot\frac{r^2(\rho)}{\rho}\leq\lambda\leq\frac{\theta}{2}\cdot\frac{r^2(\rho)}{\rho},$$

The proof of the theorem follows in three steps: first, one has to show that $P_N\mathcal{L}_f^\lambda$ is positive on the set $F\cap S_\Psi(\rho,f^*)$. Second, thanks to certain homogeneity properties of the functional, it is positive in $F\backslash B_\Psi(\rho,f^*)$, because it is positive on the 'sphere' $F\cap S_\Psi(\rho,f^*)$. Finally, one has to study the functional in $F\cap B_\Psi(\rho,f^*)$ and verify that it is positive in that set, provided that $\|f-f^*\|_{L_2(\mu)}\geq r(\rho)$.



Proof. Fix $h \in F \cap S_\Psi(\rho, f^*)$ and we shall treat two different cases: when $\|h - f^*\|_{L_2(\mu)} \geq r(\rho)$ and when $\|h - f^*\|_{L_2(\mu)} \leq r(\rho)$.

If $\|h - f^*\|_{L_2} \geq r(\rho)$, then by the triangle inequality for Ψ,

$$\Psi(h) - \Psi(f^*) = \Psi(h - f^* + f^*) - \Psi(f^*) \geq -\Psi(h - f^*).$$

Hence, for $(X_i, Y_i)_{i=1}^N \in \mathcal{A}$ and by the upper estimate in the choice of λ,

$$P_N \mathcal{L}_h^\lambda \geq \frac{\theta}{2}\|h - f^*\|_{L_2(\mu)}^2 - \lambda\Psi(h - f^*) \geq \frac{\theta}{2} r^2(\rho) - \lambda\rho > 0. \tag{3.1}$$

Next, if $\|h - f^*\|_{L_2(\mu)} \leq r(\rho)$ then

$$P_N \mathcal{L}_h^\lambda \geq -2\mathcal{O}(\rho) + \lambda(\Psi(h) - \Psi(f^*)).$$

Consider $u, v \in E$ that satisfy $f^* = u + v$ and $\Psi(u) \leq \rho/20$. Let z^* be any norming functional of v; thus, $z^* \in S_{\Psi^*}$ and $z^*(v) = \Psi(v)$. Since $\Psi(h) = \sup_{x^* \in B_{\Psi^*}} x^*(h)$ it follows that

$$\Psi(h) - \Psi(f^*) \geq \Psi(h) - \Psi(v) - \Psi(u) \geq z^*(h - v) - \Psi(u) \geq z^*(h - f^*) - 2\Psi(u).$$

This holds for any $v \in B_\Psi(\rho/20, f^*)$, and by the definition of $\Delta(\rho)$ and for an optimal choice of z^*,

$$P_N \mathcal{L}_h^\lambda \geq -2\mathcal{O}(\rho) + \lambda(z^*(h - f^*) - 2\Psi(u)) \geq -2\mathcal{O}(\rho) + \lambda(\Delta(\rho) - \rho/10) > 0, \tag{3.2}$$

where the last inequality holds because $\Delta(\rho) \geq 4\rho/5$ and $\lambda \geq 3\gamma_\mathcal{O}(\rho)/\rho$. Also, since $\gamma_\mathcal{O}(\rho) \leq (\theta/8)r^2(\rho)$, it suffices that $\lambda \geq (3\theta/8)r^2(\rho)/\rho$ to ensure that $P_N \mathcal{L}_h^\lambda > 0$ in (3.2). This completes the proof of the first step – that $P_N \mathcal{L}_h^\lambda > 0$ on $F \cap S_\Psi(\rho, f^*)$.

Turning to the second step, one has to establish a similar inequality for functions outside $B_\Psi(\rho, f^*)$. To that end, let $f \in F \backslash B_\Psi(\rho, f^*)$. Since F is convex and Ψ is homogeneous, $f = f^* + \alpha(h - f^*)$ for some $h \in F \cap S_\Psi(\rho, f^*)$ and $\alpha > 1$. Therefore,

$$P_N \mathcal{Q}_{f-f^*} = \alpha^2 P_N \mathcal{Q}_{h-f^*} \text{ and } P_N \mathcal{M}_{f-f^*} = \alpha P_N \mathcal{M}_{h-f^*};$$

moreover, $\Psi(f - f^*) = \alpha\Psi(h - f^*)$ and for every functional z^*, $z^*(f - f^*) = \alpha z^*(h - f^*)$.

Thus, by (3.1), when $\|h - f^*\|_{L_2(\mu)} \geq r(\rho)$, $P_N \mathcal{L}_f^\lambda > 0$, and when $\|h - f^*\|_{L_2(\mu)} \leq r(\rho)$,

$$\begin{aligned} P_N \mathcal{L}_f^\lambda &\geq \alpha^2 P_N \mathcal{Q}_{h-f^*} + 2\alpha P_N \mathcal{M}_{h-f^*} + \lambda(\alpha z^*(h - f^*) - 2\Psi(u)) \\ &\geq \alpha\big(P_N \mathcal{Q}_{h-f^*} + 2P_N \mathcal{M}_{h-f^*} + \lambda(z^*(h - f^*) - 2\Psi(u))\big) > 0. \end{aligned}$$

Finally, when $h \in F \cap B_\Psi(\rho, f^*)$ and $\|h - f^*\|_{L_2(\mu)} \geq r(\rho)$, (3.1) shows that $P_N \mathcal{L}_f^\lambda > 0$. ■

Remark 3.3 *Note that if $\rho \geq \Psi(f^*)$ there is no upper limitation on the choice of λ. Indeed, if $\|h - f^*\|_{L_2(\mu)} \geq r(\rho)$ and $\Psi(h) = \rho \geq \Psi(f^*)$ then $\lambda(\Psi(h) - \Psi(f^*)) \geq 0$, and $P_N \mathcal{L}_h^\lambda > 0$ just as in (3.1). The rest of the proof remains unchanged.*

It follows from the proof that the quadratic component $P_N \mathcal{Q}_{f-f^*}$ and the regularization one $\lambda(\Psi(f) - \Psi(f^*))$ dominate the multiplier component $2P_N \mathcal{M}_{f-f^*}$ in different parts of F. The behaviour of $P_N \mathcal{Q}_{f-f^*}$ allows one to exclude the set $(F \cap B_\psi(\rho, f^*)) \backslash D(r(\rho), f^*)$, as well as any point in F for which the interval $[f, f^*]$ intersects $(F \cap S_\psi(\rho, f^*)) \backslash D(r(\rho), f^*)$. This exclusion is rather free-of-charge, as it holds with no assumptions on the norm Ψ.

The situation is more subtle when trying to exclude points for which the interval $[f, f^*]$ intersects $F \cap S_\psi(\rho, f^*) \cap D(r(\rho), f^*)$. That is precisely the region in which the specific choice of Ψ is important and the regularization component is the reason why $P_N \mathcal{L}_f^\lambda > 0$.

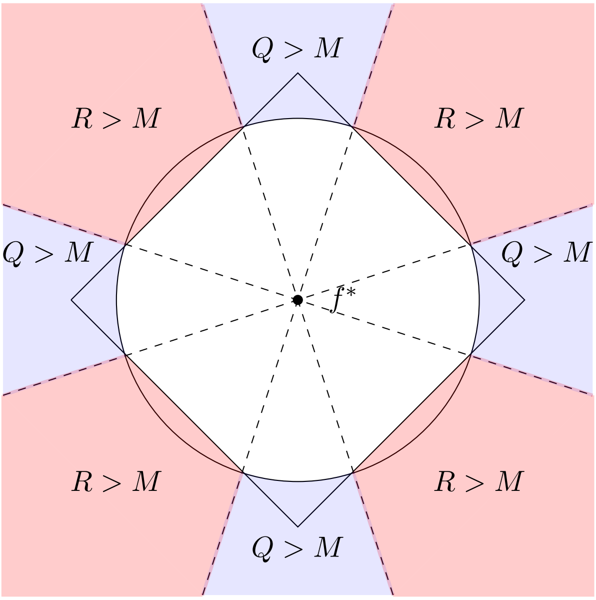


Figure 1: $P_N \mathcal{L}_f^\lambda > 0$ for two different reasons: either $Q > M$ – the quadratic component dominates the multiplier component, or $R > M$ – the regularization component dominates the multiplier component.

4 The role of $\Delta(\rho)$

It is clear that $\Delta(\rho)$ plays a crucial role in the proof of Theorem 3.2, and that the larger $\Gamma_{f^*}(\rho)$ is, the better the lower bound on $\Delta(\rho)$.

Having many norming functionals of points in $B_\Psi(\rho/20, f^*)$ can be achieved somewhat artificially, by taking $\rho \sim \Psi(f^*)$. If ρ is large enough, $B_\Psi(\rho/20, f^*)$ contains a Ψ-ball centred in 0. Therefore, $\Gamma_{f^*}(\rho)$ is the entire dual sphere and $\Delta(\rho) = \rho$. This is the situation when one attempts to derive complexity-based bounds (see Section 6 and [13]), i.e., when one wishes to find $\hat{f}$ that inherits some of f^*'s 'good qualities' that are captured by $\Psi(f^*)$.

Here, we are interested in cases in which ρ may be significantly smaller than $\Psi(f^*)$ and enough norming functionals have to be generated by other means.

If Ψ is smooth, each $f \neq 0$ has a unique norming functional, and for a small ρ, the norming functionals of points in $B_\Psi(\rho/20, f^*)$ are close to the (unique) norming functional of f^*; hence there is little hope that $\Gamma_{f^*}(\rho)$ will be large enough to ensure that $\Delta(\rho) \sim \rho$. It is therefore reasonable to choose Ψ that is not smooth in f^* or in a neighbourhood of f^*.

Another important fact is that $\Gamma_{f^*}(\rho)$ need not be as large as the entire dual sphere to ensure that $\Delta(\rho) \sim \rho$. Indeed, it suffices if $\Gamma_{f^*}(\rho)$ contains 'almost norming' functionals only to points that satisfy $\|w\|_{L_2(\mu)} \leq r(\rho)/\rho$ and $\Psi(w) = 1$, rather than to every point in the sphere S_Ψ.

4.1 $\Delta(\rho)$ and sparsity

It turns out that the combination of the right notion of sparsity with a wise choice of a norm Ψ ensures that $\Gamma_{f^*}(\rho)$ contains enough 'almost norming' functionals precisely for the subset of the sphere one is interested in.

To give an indication of how this happens, let us show the following:

Lemma 4.1 *Let $Z \subset S_{\Psi^*}$, $W \subset S_\Psi$ and $0 < \eta_1, \eta_2 < 1$. If every $w \in W$ can be written as $w = w_1 + w_2$, where $\Psi(w_1) \leq \eta_1 \Psi(w)$ and $\sup_{z^* \in Z} z^*(w_2) \geq (1-\eta_2)\Psi(w_2)$, then*

$$\inf_{w \in W} \sup_{z^* \in Z} z^*(w) \geq (1-\eta_1)(1-\eta_2) - \eta_1$$

In particular, if $\eta_1, \eta_2 \leq 1/20$ then $\inf_{w \in W} \sup_{z^* \in Z} z^*(w) \geq 4/5$.

Proof. Let $w = w_1 + w_2$ and observe that $\Psi(w_2) \geq \Psi(w) - \Psi(w_1) \geq (1-\eta_1)\Psi(w)$. Thus, for the optimal choice of $z^* \in Z$,

$$\begin{aligned} z^*(w_1 + w_2) \geq & (1-\eta_2)\Psi(w_2) + z^*(w_1) \geq (1-\eta_2)\Psi(w_2) - \eta_1 \Psi(w). \\ \geq & \big((1-\eta_1)(1-\eta_2) - \eta_1\big)\Psi(w), \end{aligned}$$

and the claim follows because $w \in S_\Psi$. ■

Let $E = \mathbb{R}^d$ viewed as a class of linear functionals on $\mathbb{R}^d$. Set μ to be an isotropic measure on $\mathbb{R}^d$; thus $\{t \in \mathbb{R}^d : \mathbb{E}\langle t, X\rangle^2 \leq 1\} = B_2^d$.

Assume that for $t \in \mathbb{R}^d$ that is supported on $I \subset \{1, ..., d\}$, the set of its norming functionals consists of functionals of the form $z_0^* + (1-\eta_2)u^*$ for some fixed z_0^* that is supported on I and *any* $u \in B_{\Psi^*}$ that is supported on I^c (such is the case, for example, when $E = \ell_1^d$).

For every such t, consider $w \in \rho S_\Psi$ and set $w_1 = P_I w$ and $w_2 = P_{I^c} w$, the coordinate projections of w onto $\mathrm{span}(e_i)_{i \in I}$ and $\mathrm{span}(e_i)_{i \in I^c}$, respectively. Hence, there is a functional $z^* = z_0^* + (1-\eta_2)u^*$ that is norming for t and also satisfies

$$z^*(w_2) = (1-\eta_2)u^*(w_2) = (1-\eta_2)\Psi(w_2).$$

Therefore, Lemma 4.1 may be applied once $\Psi(P_I w) \leq \eta_1 \Psi(w)$.

Naturally, such a shrinking phenomenon need not be true for *every* $w \in S_\Psi$; fortunately, it is only required for $w \in S_\Psi \cap (r(\rho)/\rho)D$ – and we will show that it is indeed the case in the three examples we present. In all three, the combination of sparsity and the right choice of the norm helps in establishing a lower bound on $\Delta(\rho)$ in two ways: firstly, the set $\Gamma_{t^*}(\rho)$ consists of functionals that are 'almost norming' for any x whose support is disjoint from the support of t^*; and secondly, a coordinate projection 'shrinks' the Ψ norm of points in $\rho S_\Psi \cap r(\rho)D$.

4.2 $\Delta(\rho)$ in the three examples

Let us show that in the three examples, the LASSO, SLOPE and trace norm regularization, $\Delta(\rho) \geq (4/5)\rho$ for the right choice of ρ, and that choice depends on the degree of sparsity in each case.

In all three examples, we will assume that the underlying measure is isotropic; thus the $L_2(\mu)$ norm coincides with the natural Euclidean structure: the ℓ_2^d norm for the LASSO and SLOPE, and the Hilbert-Schmidt norm for trace-norm regularization.

The LASSO.

Observe that if $\langle t^*, \cdot \rangle$ is the true minimizer of the functional $\langle t, \cdot \rangle \to \mathbb{E}(\langle t, X \rangle - Y)^2$, then any function $h_t = \langle t, \cdot \rangle$ for which $\|h_t - f^*\|_{L_2} \leq r(\rho)$ and $\Psi(h_t - f^*) = \rho$ is of the form $h_t = \langle t, \cdot \rangle = \langle w + t^*, \cdot \rangle$, where

$$w \in \rho S(\ell_1^d) \cap r(\rho) B_2^d.$$

Recall that the dual norm to $\|\cdot\|_1$ is $\|\cdot\|_\infty$, and thus

$$\Delta(\rho) = \inf_{w \in \rho S(\ell_1^d) \cap r(\rho) B_2^d} \sup_{z \in \Gamma_{t^*}(\rho)} \langle z, w \rangle,$$

where $\Gamma_{t^*}(\rho)$ is the set of all vectors $z^* \in \mathbb{R}^d$ that satisfy

$$\|z^*\|_\infty = 1 \ \text{ and } \ z^*(v) = \|v\|_1 \text{ for some } v, \ \|v - t^*\|_1 \leq \rho/20.$$

Lemma 4.2 *There exists an absolute constant c for which the following hold. If $t^* = v + u$ for $u \in (\rho/20)B_1^d$ and $|\mathrm{supp}(v)| \leq c(\rho/r(\rho))^2$ then $\Delta(\rho) \geq 4\rho/5$.*

In other words, if t^* is well approximated with respect to the ℓ_1^d norm by some $v \in \mathbb{R}^d$ that is s-sparse, and s is small enough relative to the ratio $(\rho/r(\rho))^2$, then $\Delta(\rho) \geq (4/5)\rho$.

Just as noted earlier, we shall use two key properties of the ℓ_1 norm and sparse vectors: firstly, that if x and y have disjoint supports, there is a functional that is simultaneously norming for x and y, i.e., $z^* \in B_\infty^d$ for which

$$z^*(x) = \|x\|_1 \ \text{ and } \ z^*(y) = \|y\|_1; \tag{4.1}$$

secondly, that if $\|x\|_1 = \rho$ and $\|x\|_2$ is significantly smaller than ρ, a coordinate projection 'shrinks' the ℓ_1^d norm: $\|P_I x\|_1$ is much smaller than $\|x\|_1$.

Proof. Let $w \in \rho S(\ell_1^d) \cap r(\rho) B_2^d$. Since $\|t^* - v\|_1 \leq \rho/20$ there exists $z^* \in \Gamma_{t^*}(\rho)$ that is norming for v. Moreover, if $I = \text{supp}(v)$, then according to (4.1) one can choose z^* that is also norming for $P_{I^c} w$. Thus, $\|P_{I^c} w\|_1 = z^*(P_{I^c} w)$ and

$$z^*(w) = z^*(P_I w) + z^*(P_{I^c} w) \geq \|P_{I^c} w\|_1 - \|P_I w\|_1 \geq \|w\|_1 - 2\|P_I w\|_1 .$$

Since $\|w\|_2 \leq r(\rho)$, one has $\|P_I w\|_1 \leq \sqrt{s}\,\|P_I w\|_2 \leq \sqrt{s} r(\rho)$. Therefore,

$$\langle z, w\rangle \geq \rho - 2\sqrt{s} r(\rho) \geq 4\rho/5$$

when $100s \leq (\rho/r(\rho))^2$. ■

SLOPE.

Let $\beta_1 \geq \beta_2 \geq ... \geq \beta_d > 0$, and recall that $\Psi(t) = \sum_{i=1}^d \beta_i t_i^*$.

Note that $\Psi(t) = \sup_{z \in Z} \langle z, t\rangle$, for

$$Z = \left\{ \sum_{i=1}^d \varepsilon_i \beta_{\pi_i} e_i : \ (\varepsilon_i)_{i=1}^d \in \{-1, 1\}^d, \ \pi \text{ is a permulation of } \{1, ..., d\} \right\}.$$

Therefore, the extreme points of the dual unit ball are of the form $\sum_{i=1}^d \varepsilon_i \beta_{\pi_i} e_i$.

Following the argument outlined above, let us show that if x is supported on a reasonably small $I \subset \{1, ..., d\}$, the set of norming functionals of x consists of 'almost norming' functionals for any y that is supported on I^c. Moreover, and just like the ℓ_1^d norm, if $\Psi(x) = \rho$ and $\|x\|_2$ is significantly smaller than ρ, a coordinate projection of x 'shrinks' its Ψ norm.

Lemma 4.3 *There exists an absolute constant c for which the following holds. Let $1 \leq s \leq d$ and set $\mathcal{B}_s = \sum_{i \leq s} \beta_i/\sqrt{i}$. If t^* is $\rho/20$ approximated (relative to Ψ) by an s-sparse vector and if $\mathcal{B}_s \leq c\rho/r(\rho)$ then $\Delta(\rho) \geq 4\rho/5$.*

Proof. Let $t^* = u + v$, for v that is supported on at most s coordinates and $u \in (\rho/20) B_\Psi$. Set $I \subset \{1, ..., d\}$ to be the support of v and let $z = (z_i)_{i=1}^d$ be a norming functional for v to be specified later; thus, $z \in \Gamma_{t^*}(\rho)$.

Given t for which $\Psi(t-t^*)=\rho$ and $\|t-t^*\|_2 \leq r(\rho)$, one has

$$\begin{aligned}z(t-t^*) &= z(t-v)-z(u) = z(P_{I^c}(t-v)) + z(P_I(t-v)) - z(u) \\ &\geq \sum_{i \in I^c} z_i(t-v)_i + \sum_{i \in I} z_i(t-v)_i - \Psi(u) \\ &\geq \sum_{i \in I^c} z_i(t-v)_i - \sum_{i \leq s} \beta_i (t-v-u)_i^* - 2\Psi(u) \\ &= \sum_{i \in I^c} z_i(t-v)_i - \sum_{i \leq s} \beta_i (t-t^*)_i^* - 2\Psi(u) = (*).\end{aligned}$$

Since v is supported in I, one may optimize the choice of z by selecting the right permutation of the coordinates in I^c, and

$$\begin{aligned}&\sum_{i \in I^c} z_i(t-v)_i \geq \sum_{i>s} \beta_i (t-v)_i^* \geq \sum_{i>s} \beta_i (t-v-u)_i^* - \Psi(u) \\ =&\sum_{i=1}^d \beta_i (t-t^*)_i^* - \sum_{i \leq s} \beta_i (t-t^*)_i^* - \Psi(u).\end{aligned}$$

Therefore,

$$(*) \geq \sum_{i=1}^d \beta_i (t-t^*)_i^* - 2\sum_{i \leq s} \beta_i (t-t^*)_i^* - 3\Psi(u) \geq \frac{17}{20}\rho - 2\sum_{i \leq s} \beta_i (t-t^*)_i^*.$$

Since $\|t-t^*\|_2 \leq r(\rho)$, it is evident that $(t-t^*)_i^* \leq r(\rho)/\sqrt{i}$, and

$$\sum_{i=1}^s \beta_i (t-t^*)_i^* \leq r(\rho) \sum_{i=1}^s \frac{\beta_i}{\sqrt{i}} = r(\rho)\mathcal{B}_s.$$

Hence, if $\rho \geq 40 r(\rho)\mathcal{B}_s$ then $\Delta(\rho) \geq 4\rho/5$. ■

Trace-norm regularization.

The trace norm has similar properties to the ℓ_1 norm. Firstly, one may show that the dual norm to $\|\cdot\|_1$ is $\|\cdot\|_\infty$, which is simply the standard operator norm. Moreover, one may find a functional that is simultaneously norming for any two elements with 'disjoint support' (and of course, the meaning of 'disjoint support' has to be interpreted correctly here). Finally, it satisfies a 'shrinking' phenomenon for matrices whose Hilbert-Schmidt norm is significantly smaller than their trace norm.

Lemma 4.4 *There exists an absolute constant c for which the following hold. If $A^* = V+U$, where $\|U\|_1 \leq \rho/20$ and $\mathrm{rank}(V) \leq c(\rho/r(\rho))^2$, then $\Delta(\rho) \geq 4\rho/5$.*

The fact that a low-rank matrix has many norming functionals is well known and follows, for example, from [38].

Lemma 4.5 *Let $V \in R^{m \times T}$ and assume that $V = P_I V P_J$ for appropriate orthogonal projections onto subspaces $I \subset \mathbb{R}^m$ and $J \subset \mathbb{R}^T$. Then, for every $W \in \mathbb{R}^{m \times T}$ there is a matrix Z that satisfies $\|Z\|_\infty = 1$, and*

$$\big\langle Z, V \big\rangle = \|V\|_1, \quad \big\langle Z, P_{I^\perp} W P_{J^\perp} \big\rangle = \|P_{I^\perp} W P_{J^\perp}\|_1,$$

$$\big\langle Z, P_I W P_{J^\perp} \big\rangle = 0 \ \text{ and } \ \big\langle Z, P_{I^\perp} W P_J \big\rangle = 0.$$

Lemma 4.5 describes a similar phenomenon to the situation in ℓ_1^d, but with a different notion of 'disjoint support': if V is low-rank and the projections P_I and P_J are non-trivial, one may find a functional that is norming both for V and for the part of W that is 'disjoint' of V. Moreover, the functional vanishes on the 'mixed' parts $P_I W P_{J^\perp}$ and $P_{I^\perp} W P_J$.

Proof of Lemma 4.4. Recall that S_1 is the unit sphere of the trace norm and that B_2 is the unit ball of the Hilbert-Schmidt norm. Hence,

$$\Delta(\rho) = \inf_{W \in \rho S_1 \cap r(\rho) B_2} \ \sup_{Z \in \Gamma_{A^*}(\rho)} \big\langle Z, W \big\rangle$$

where $\Gamma_{A^*}(\rho)$ is the set of all matrices $Z \in \mathbb{R}^{m \times T}$ that satisfy $\|Z\|_\infty = 1$ and $\big\langle Z, V \big\rangle = 1$ for some V for which $\|A^* - V\|_1 \leq \rho/20$.

Fix a rank-s matrix $V = P_I V P_J$, for orthogonal projections P_I and P_J that are onto subspaces of dimension s. Consider $W \in \mathbb{R}^{m \times T}$ for which $\|W\|_1 = \rho$ and $\|W\|_2 \leq r(\rho)$ and put Z to be a norming functional of V as in Lemma 4.5. Thus, $Z \in \Gamma_{A^*}(\rho)$ and

$$\begin{aligned}\big\langle Z, W \big\rangle =& \big\langle Z, P_{I^\perp} W P_{J^\perp} \big\rangle + \big\langle Z, P_I W P_J \big\rangle = \|P_{I^\perp} W P_{J^\perp}\|_1 - \|P_I W P_J\|_1 \\ \geq& \|W\|_1 - \|P_I W P_{J^\perp}\|_1 - \|P_{I^\perp} W P_J\|_1 - 2\|P_I W P_J\|_1.\end{aligned}$$

All that remains is to estimate the trace norms of the three components that are believed to be 'low-dimension' - in the sense that their rank is at most s.

Recall that $(\sigma_i(A))$ are the singular values of A arranged in a non-increasing order. It is straightforward to verify (e.g., using the characterization of the singular values via low-dimensional approximation), that

$$\sigma_i(P_I W P_{J^\perp}), \sigma_i(P_{I^\perp} W P_J), \sigma_i(P_I W P_J) \leq \sigma_i(W).$$

Moreover, $\|W\|_2 \leq r(\rho)$, therefore, being rank-s operators, one has

$$\|P_I W P_{J^\perp}\|_1, \ \|P_{I^\perp} W P_J\|_1, \ \|P_I W P_J\|_1 \leq \sum_{i=1}^s \sigma_i(W) \leq \sqrt{s}\Big(\sum_{i=1}^s \sigma_i^2(W)\Big)^{1/2} \leq \sqrt{s} r(\rho),$$

implying that

$$\big\langle Z, W \big\rangle \geq \rho - 4r(\rho)\sqrt{s}.$$

Therefore, if $400s \leq (r/r(\rho))^2$, then $\Delta(\rho) \geq 4\rho/5$. ■

5 The three examples revisited

The estimates on $\Delta(\rho)$ presented above show that in all three examples, when f^* is well approximated by a function whose 'degree of sparsity' is $\lesssim (\rho/r(\rho))^2$, then $\Delta(\rho) \geq 4\rho/5$ and Theorem 3.2 may be used. Clearly, the resulting error rates depend on the right choice of ρ, and thus on $r(\rho)$.

Because $r(\rho)$ happens to be the minimax rate of the learning problem in the class $F \cap B_\Psi(\rho, f^*)$, its properties have been studied extensively. Obtaining an estimate of $r(\rho)$ involves some assumptions on X and ξ, and the one setup in which it can be characterized for an arbitrary class F is when the class is L-subgaussian and $\xi \in L_q$ for some $q > 2$ (though ξ need not be independent of X). It is straightforward to verify that an L-subgaussian class satisfies the small-ball condition of Assumption 2.1 for $\kappa = 1/2$ and $\varepsilon = c/L^4$ where c is an absolute constant. Moreover, if the class is L-subgaussian, the natural complexity parameter associated with it is the expectation of the supremum of the canonical gaussian process indexed by the class.

Definition 5.1 *Let $F \subset L_2(\mu)$ and set $\{G_f : f \in F\}$ to be the canonical gaussian process indexed by F; that is, each G_f is a centred gaussian variable and the covariance structure of the process is endowed by the inner product in $L_2(\mu)$. The expectation of the supremum of the process is defined by*

$$\ell_*(F) = \sup\{\mathbb{E} \sup_{f \in F'} G_f : F' \subset F \text{ is finite}\}.$$

It follows from a standard chaining argument that if F is L-subgaussian then

$$\mathbb{E} \sup_{f \in F} \left| \frac{1}{N} \sum_{i=1}^{N} \varepsilon_i (h - f^*)(X_i) \right| \lesssim L \frac{\ell_*(F)}{\sqrt{N}}.$$

Therefore, if

$$F_{\rho,r} = F \cap B_\Psi(\rho, f^*) \cap D(r, f^*)$$

then for every $\rho > 0$ and $f^* \in F$

$$r_Q(\rho) \leq \inf\left\{ r > 0 : \ell_*(F_{\rho,r}) \leq C(L) r \sqrt{N} \right\}.$$

Turning to r_M, we shall require the following fact from [20].

Theorem 5.2 ([20]) *Let $q > 2$ and $L \geq 1$. For every $0 < \delta < 1$ there is a constant $c = c(\delta, L, q)$ for which the following holds. If H is an L-subgaussian class and $\xi \in L_q$, then with probability at least $1 - \delta$,*

$$\sup_{h \in H} \left| \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \varepsilon_i \xi_i h(X_i) \right| \leq c \|\xi\|_{L_q} \ell_*(H).$$

The complete version of Theorem 5.2 includes a sharp estimate on the constant c. However, obtaining accurate probability estimates is not the main feature of this note and deriving such estimates leads to a cumbersome presentation. To keep our message to the point, we have chosen not to present the best possible probability estimates in what follows.

A straightforward application of Theorem 5.2 shows that

$$r_M(\rho) \leq \inf\left\{r > 0 : \|\xi\|_{L_q} \ell_*(F_{\rho,r}) \leq cr^2\sqrt{N}\right\}$$

for a constant c that depends on L, q and δ.

Recall that we have assumed that X is isotropic, which means that the $L_2(\mu)$ norm coincides with the natural Euclidean structure on the space: the standard ℓ_2^d norm for the LASSO and SLOPE and the Hilbert-Schmidt norm for trace norm regularization. Since the covariance structure of the indexing gaussian process is endowed by the inner product, it follows that

$$\ell_*(\rho B_\Psi \cap rD) = \mathbb{E} \sup_{w \in \rho B_\Psi \cap rB_2} \left\langle G, w \right\rangle$$

for the standard gaussian vector $G = (g_1, ..., g_d)$ in the case of the LASSO and SLOPE and the gaussian matrix $G = (g_{ij})$ in the case of trace norm minimization. Hence, one may obtain a bound on $r(\rho)$ by estimating this expectation in each case.

The LASSO and SLOPE. Let $(\beta_i)_{i=1}^d$ be a non-increasing positive sequence and set $\Psi(t) = \sum_{i=1}^d t_i^* \beta_i$.

Since the LASSO corresponds to the choice of $(\beta_i)_{i=1}^d = (1, ..., 1)$, it suffices to identify $\ell_*(\rho B_\Psi \cap rB_2^d)$ for the SLOPE norm and a general choice of weights.

Lemma 5.3 *There exists an absolute constant C for which the following holds. If β and Ψ are as above, then*

$$\mathbb{E} \sup_{w \in \rho B_\Psi \cap rB_2^d} \left\langle G, w \right\rangle \leq C \min_k \left\{ r\sqrt{(k-1)\log\left(\frac{ed}{k-1}\right)} + \rho \max_{i \geq k} \frac{\sqrt{\log(ed/i)}}{\beta_i} \right\}$$

(and if $k = 1$, the first term is set to be 0).

Proof. Fix $1 \leq k \leq d$. Let J be the set of indices of the k largest coordinates of $(|g_i|)_{i=1}^d$, and for every w let I_w be the sets of indices of the k largest coordinates of $(|w_i|)_{i=1}^d$. Put $J_w = J \cup I_w$ and note that $|J_w| \leq 2k$. Hence,

$$\begin{aligned} \sup_{w \in \rho B_\Psi \cap rB_2^d} \sum_{i=1}^d w_i g_i &\leq \sup_{w \in rB_2^d} \sum_{i \in J_w} w_i g_i + \sup_{w \in \rho B_\Psi} \sum_{i \in J_w^c} w_i g_i \\ &\lesssim r\left(\sum_{i<k} (g_i^*)^2\right)^{1/2} + \sup_{w \in \rho B_\Psi} \sum_{i \geq k} w_i^* \beta_i \frac{g_i^*}{\beta_i} \lesssim r\left(\sum_{i<k} (g_i^*)^2\right)^{1/2} + \rho \max_{i \geq k} \frac{g_i^*}{\beta_i}. \end{aligned}$$

As a starting point, note that a standard binomial estimate shows that

$$\begin{aligned}Pr\left(g_i^* \geq t\sqrt{\log(ed/i)}\right) \leq & \binom{d}{i} Pr^i\left(|g| \geq t\sqrt{\log(ed/i)}\right) \\ \leq & 2\exp(i\log(ed/i) - i\log(ed/i)\cdot t^2/2).\end{aligned}$$

Applying the union bound one has that for $t \geq 4$, with probability at least $1-2\exp(-(t^2/2)k\log(ed/k))$,

$$g_i^* \leq c_3 t\sqrt{\log(ed/i)} \text{ for every } i \geq k. \tag{5.1}$$

The same argument shows that $\mathbb{E}(g_i^*)^2 \lesssim \log(ed/i)$.

Let U_k be the set of vectors on the Euclidean sphere that are supported on at most k coordinates. Set

$$\|x\|_{[k]} = \left(\sum_{i\leq k}(x_i^*)^2\right)^{1/2} = \sup_{u\in U_k}\langle x,u\rangle$$

and recall that by the gaussian concentration of measure theorem (see, e.g., Theorem 7.1 in [14]),

$$\left(\mathbb{E}\|G\|_{[k]}^q\right)^{1/q} \leq \mathbb{E}\|G\|_{[k]} + c\sqrt{q}\sup_{u\in U_k}\|\langle G,u\rangle\|_{L_2} \leq \mathbb{E}\|G\|_{[k]} + c_1\sqrt{q}.$$

Moreover, since $\mathbb{E}(g_i^*)^2 \lesssim \log(ed/i)$, one has

$$\mathbb{E}\|G\|_{[k]} \leq \left(\mathbb{E}\sum_{i\leq k}(g_i^*)^2\right)^{1/2} \lesssim \sqrt{k\log(ed/k)}.$$

Therefore, by Chebyshev's inequality for $q \sim k\log(ed/k)$, for $t \geq 1$, with probability at least $1-2t^{-c_1 k\log(ed/k)}$,

$$\left(\sum_{i\leq k}(g_i^*)^2\right)^{1/2} \leq c_2 t\sqrt{k\log(ed/k)}.$$

Turning to the 'small coordinates', by (5.1),

$$\max_{i\geq k}\frac{g_i^*}{\beta_i} \lesssim t\max_{i\geq k}\frac{\sqrt{\log(ed/i)}}{\beta_i}.$$

It follows that for every choice of $1 \leq k \leq d$,

$$\begin{aligned}\mathbb{E}\sup_{w\in\rho B_\Psi\cap rB_2^d}\langle G,w\rangle \lesssim & r\mathbb{E}\left(\sum_{i<k}(g_i^*)^2\right)^{1/2} + \rho\mathbb{E}\max_{i\geq k}\frac{g_i^*}{\beta_i} \\ \lesssim & r\sqrt{(k-1)\log(ed/(k-1))} + \rho\max_{i\geq k}\frac{\sqrt{\log(ed/i)}}{\beta_i},\end{aligned}$$

and, if $k=1$, the first term is set to be 0. ■

If $\beta = (1, ..., 1)$ (which corresponds to the LASSO), then $B_\Psi = B_1^d$, and one may select $\sqrt{k} \sim \rho/r$, provided that $r \leq \rho \leq r\sqrt{d}$. In that case,

$$\mathbb{E} \sup_{w \in \rho B_1^d \cap r B_2^d} \langle G, w \rangle \lesssim \rho \sqrt{\log(edr^2/\rho^2)}.$$

The estimates when $r \geq \rho$ or $r\sqrt{d} \leq \rho$ are straightforward. Indeed, if $r \geq \rho$ then $\rho B_1^d \subset r B_2^d$ and

$$\ell_*(\rho B_1^d \cap r B_2^d) = \ell_*(\rho B_1^d) \sim \rho \sqrt{\log(ed)},$$

while if $r\sqrt{d} \leq \rho$ then $r B_2^d \subset \rho B_1^d$, and

$$\ell_*(\rho B_1^d \cap r B_2^d) = \ell_*(r B_2^d) \sim r\sqrt{d}.$$

The LASSO.

A straightforward computation shows that

$$r_M^2(\rho) \lesssim_{L,q,\delta} \begin{cases} \frac{\|\xi\|_{L_q}^2 d}{N} & \text{if } \rho^2 N \gtrsim_{L,q,\delta} \|\xi\|_{L_q}^2 d^2 \\ \rho \|\xi\|_{L_q} \sqrt{\frac{1}{N} \log\left(\frac{e\|\xi\|_{L_q} d}{\rho\sqrt{N}}\right)} & \text{otherwise,} \end{cases}$$

and

$$r_Q^2(\rho) \lesssim_L \begin{cases} 0 & \text{if } N \gtrsim_L d \\ \frac{\rho^2}{N} \log\left(\frac{c(L)d}{N}\right) & \text{otherwise.} \end{cases}$$

Proof of Theorem 1.3. We will actually prove a slightly stronger result, which gives an improved estimation error if one has prior information on the degree of sparsity.

Using the estimates on r_M and r_Q, it is straightforward to verify that the sparsity condition of Lemma 4.2 holds when $N \gtrsim_{L,q,\delta} s \log(ed/s)$ and for any

$$\rho \gtrsim_{L,q,\delta} \|\xi\|_{L_q} s \sqrt{\frac{1}{N} \log\left(\frac{ed}{s}\right)}.$$

It follows from Lemma 4.2 that if there is an s-sparse vector that belongs to $t^* + (\rho/20) B_1^d$, then $\Delta(\rho) \geq 4\rho/5$. Finally, Theorem 3.2 yields the stated bounds on $\|\hat{t} - t^*\|_1$ and $\|\hat{t} - t^*\|_2$ once we set

$$\lambda \sim \frac{r^2(\rho)}{\rho} \sim_{L,q,\delta} \|\xi\|_{L_q} \sqrt{\frac{1}{N} \log\left(\frac{ed}{s}\right)}.$$

The estimates on $\|\hat{t} - t^*\|_p$ for $1 \leq p \leq 2$ can be easily verified because

$$\|x\|_p \leq \|x\|_1^{-1+2/p} \|x\|_2^{2-2/p}.$$

In case one has no prior information on s, one may take

$$\rho \sim_{L,q,\delta} \|\xi\|_{L_q} s \sqrt{\frac{1}{N} \log(ed)}$$

and

$$\lambda \sim_{L,q,\delta} \|\xi\|_{L_q} \sqrt{\frac{\log(ed)}{N}}.$$

The rest of the argument remains unchanged. ■

SLOPE

Assume that $\beta_i \leq C\sqrt{\log(ed/i)}$, which is the standard assumption for SLOPE [4, 32]. By considering the cases $k=1$ and $k=d$,

$$\mathbb{E} \sup_{w \in \rho B_\Psi \cap r B_2^d} \langle G, w \rangle \lesssim \min\{C\rho, \sqrt{d}r\}.$$

Thus, one may show that

$$r_Q^2(\rho) \lesssim_L \begin{cases} 0 & \text{if } N \gtrsim_L d \\ \frac{\rho^2}{N} & \text{otherwise,} \end{cases} \quad \text{and} \quad r_M^2(\rho) \lesssim_{L,q,\delta} \begin{cases} \|\xi\|_{L_q}^2 \frac{d}{N} & \text{if } \rho^2 N \gtrsim_{L,q,\delta} \|\xi\|_{L_q}^2 d^2 \\ \|\xi\|_{L_q} \frac{\rho}{\sqrt{N}} & \text{otherwise.} \end{cases}$$

Proof of Theorem 1.5. Recall that $\mathcal{B}_s = \sum_{i \leq s} \beta_i/\sqrt{i}$, and when $\beta_i \leq C\sqrt{\log(ed/i)}$, one may verify that

$$\mathcal{B}_s \lesssim C\sqrt{s\log(ed/s)}.$$

Hence, the condition $\mathcal{B}_s \lesssim \rho/r(\rho)$ holds when $N \gtrsim_{L,q,\delta} s\log(ed/s)$ and

$$\rho \gtrsim_{L,q,\delta} \|\xi\|_{L_q} \frac{s}{\sqrt{N}} \log\left(\frac{ed}{s}\right).$$

It follows from Lemma 4.3 that $\Delta(\rho) \geq 4\rho/5$ when there is an s-sparse vector in $t^* + (\rho/20)B_\Psi$; therefore, one may apply Theorem 3.2 for the choice of

$$\lambda \sim \frac{r^2(\rho)}{\rho} \sim_{L,q,\delta} \frac{\|\xi\|_{L_q}}{\sqrt{N}}.$$

■

The trace-norm.

Recall that B_1 is the unit ball of the trace norm, that B_2 is the unit ball of the Hilbert-Schmidt norm, and that the canonical gaussian vector here is the gaussian matrix $G = (g_{ij})$. Since the operator norm is the dual to the trace norm,

$$\ell_*(B_1) = \mathbb{E}\sigma_1(G) \lesssim \sqrt{\max\{m,T\}},$$

and clearly,

$$\ell_*(B_2) = \mathbb{E}\|G\|_2 \lesssim \sqrt{mT}.$$

Thus,

$$\ell_*(\rho B_\Psi \cap rB_2) = \ell_*(\rho B_1 \cap rB_2) \leq \min\left\{\rho\ell_*(B_1), r\ell_*(B_2)\right\} \lesssim \min\{\rho\sqrt{\max\{m,T\}}, r\sqrt{mT}\}.$$

Therefore,

$$r_Q^2(\rho) \lesssim_L \begin{cases} 0 & \text{if } N \gtrsim_L mT \\ \rho^2 \frac{\max\{m,T\}}{N} & \text{otherwise,} \end{cases}$$

and

$$r_M^2(\rho) \lesssim_{L,q,\delta} \begin{cases} \|\xi\|_{L_q}^2 \frac{mT}{N} & \text{if } \rho^2 N \gtrsim_{L,q,\delta} \|\xi\|_{L_q}^2 (mT)(\min\{m,T\})^2 \\ \rho\|\xi\|_{L_q}\sqrt{\frac{\max\{m,T\}}{N}} & \text{otherwise.} \end{cases}$$

Proof of Theorem 1.6. It is straightforward to verify that if $N \gtrsim_{L,q,\delta} s\max\{m,T\}$ then $s \lesssim (\rho/r(\rho))^2$ when

$$\rho \gtrsim_{L,q,\delta} \|\xi\|_{L_q} s\sqrt{\frac{\max\{m,T\}}{N}}$$

as required in Lemma 4.4. Moreover, if there is some $V \in \mathbb{R}^{m\times T}$ for which $\|V - A^*\|_1 \lesssim \rho$ and $\operatorname{rank}(V) \le s$, it follows that $\Delta(\rho) \ge 4\rho/5$. Setting

$$\lambda \sim \frac{r^2(\rho^*)}{\rho^*} \sim_{L,q,\delta} \|\xi\|_{L_q}\sqrt{\frac{\max\{m,T\}}{N}},$$

Theorem 3.2 yields the bounds on $\|\hat{A} - A^*\|_1$ and $\|\hat{A} - A^*\|_2$. The bounds on the Schatten norms $\|\hat{A} - A^*\|_p$ for $1 \le p \le 2$ hold because $\|A\|_p \le \|A\|_1^{-1+2/p} \|A\|_2^{2-2/p}$. ■

6 Concluding Remarks

As noted earlier, the method we present may be implemented in classical regularization problems as well, leading to an error rate that depends on $\Psi(f^*)$ – by applying the trivial bound on $\Delta(\rho)$ when $\rho \sim \Psi(f^*)$.

The key issue in classical regularization schemes is the price that one has to pay for not knowing $\Psi(f^*)$ in advance. Indeed, given information on $\Psi(f^*)$, one may use a learning procedure taking values in $\{f \in F : \Psi(f) \le \Psi(f^*)\}$ such as Empirical Risk Minimization. This approach would result in an error rate of $r(c\Psi(f^*))$, and the hope is that the error rate of the regularized procedure is close that – without having prior knowledge on $\Psi(f^*)$. Surprisingly, as we will show in [13], that is indeed the case.

The problem with applying Theorem 3.2 to the classical setup is the choice of λ. One has no information on $\Psi(f^*)$, and thus setting $\lambda \sim r^2(\rho)/\rho$ for $\rho \sim \Psi(f^*)$ is clearly impossible.

A first attempt of bypassing this obstacle is Remark 3.3: if $\rho \gtrsim \Psi(f^*)$, there is no upper constraint on the choice of λ. Thus, one may consider $\lambda \sim \sup_{\rho>0} \frac{r^2(\rho)}{\rho}$, which suites any $\rho > 0$. Unfortunately, that choice will not do, because in many important examples the supremum happens to be infinite. Instead, one may opt for the lower constraint on λ and select

$$\lambda \sim \sup_{\rho>0} \frac{\gamma_{\mathcal{O}}(\rho)}{\rho}, \tag{6.1}$$

which is also a legitimate choice for any ρ, and is always finite.

We will show in [13] that the choice in (6.1) leads to optimal bounds in many interesting examples – thanks to the first part of Theorem 3.2.

An essential component in the analysis of regularization problems is bounding $r(\rho)$, and we only considered the subgaussian case and completely ignored the question of the probability estimate. In that sense, the method we presented falls short of being completely satisfactory.

Addressing both these issues requires sharp upper estimates on empirical and multiplier processes, preferably in terms of some natural geometric feature of the underlying class. Unfortunately, this is a notoriously difficult problem. Indeed, the final component in the chaining-based analysis used to study empirical and multiplier processes is to translate a metric complexity parameter (e.g., Talagrand's γ-functionals) to a geometric one (for example, the mean-width of the set). Such estimates are known almost exclusively in the subgaussian case – which is, in a nutshell, Talagrand's *Majorizing Measures theory* [33].

The chaining process in [20] is based on a more sensitive metric parameter than the standard subgaussian one. This leads to satisfactory results for other choices of random vectors that are not necessarily subgaussian, for example, unconditional log-concave. Still, it is far from a complete theory – as a general version of the Majorizing Measures Theorem is not known.

Another relevant fact is from [26]. It turns out that if V is a class of linear functionals on $\mathbb{R}^d$ that satisfies a relatively minor symmetry property, and X is an isotropic and subgaussian vector for which

$$\sup_{t \in S^{d-1}} \|\langle X, t\rangle\|_{L_p} \leq L\sqrt{p} \quad \text{for} \quad 2 \leq p \lesssim \log d, \tag{6.2}$$

then the empirical and multiplier processes indexed by V behave as if X were a subgaussian vector. In other words, it suffices to have a subgaussian moment growth up to $p \sim \log d$ to ensure a subgaussian behaviour.

This fact is useful because all the indexing sets considered here (and in many other sparsity-based regularization procedures as well) satisfy the required symmetry property.

Finally, a word about the probability estimate in Theorem 5.2. The actual result from [20] leads to a probability estimate governed by two factors: the L_q space to which ξ belongs and the 'effective dimension' of the class. For a class of linear functionals on $\mathbb{R}^d$ and an isotropic vector X, this effective dimension is

$$D(V) = \left(\frac{\ell^*(V)}{d_2(V)}\right)^2,$$

where $\ell_*(V) = \mathbb{E}\sup_{v \in V} |\langle G, v\rangle|$ and $d_2(V) = \sup_{v \in V} \|v\|_{\ell_2^d}$.

One may show that with probability at least

$$1 - c_1 w^{-q} N^{-((q/2)-1)} \log^q N - 2\exp(-c_2 u^2 D(V)),$$

$$\sup_{v \in V} \left| \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \left(\xi_i \langle V, X_i\rangle - \mathbb{E}\xi\langle X, v\rangle\right) \right| \lesssim Lwu\|\xi\|_{L_q} \ell_*(V). \tag{6.3}$$

If ξ has better tail behaviour, the probability estimate improves; for example, if ξ is subgaussian then (6.3) holds with probability at least $1 - 2\exp(-cw^2 N) - 2\exp(-cu^2 D(V))$.

The obvious complication is that one has to obtain a *lower bound* on the effective dimension $D(V)$. And while it is clear that $D(v) \gtrsim 1$, in many cases (including our three examples) a much better bound is true.

Let us mention that the effective dimension is perhaps the most important parameter in Asymptotic Geometric Analysis. Milman's version of Dvoretzky's Theorem (see, e.g., [1]) shows that $D(V)$ captures the largest dimension of a Euclidean structure hiding in V. In fact, this geometric observation exhibits why that part of the probability estimate in (6.3) cannot be improved.

References

[1] Shiri Artstein-Avidan, Apostolos Giannopoulos, and Vitali D. Milman. *Asymptotic geometric analysis. Part I*, volume 202 of *Mathematical Surveys and Monographs.* American Mathematical Society, Providence, RI, 2015.

[2] Francis R. Bach. Structured sparsity-inducing norms through submodular functions. In *Advances in Neural Information Processing Systems 23: 24th Annual Conference on Neural Information Processing Systems 2010. Proceedings of a meeting held 6-9 December 2010, Vancouver, British Columbia, Canada.*, pages 118–126, 2010.

[3] Peter J. Bickel, Ya'acov Ritov, and Alexandre B. Tsybakov. Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.*, 37(4):1705–1732, 2009.

[4] Małgorzata Bogdan, Ewout van den Berg, Chiara Sabatti, Weijie Su, and Emmanuel J. Candès. SLOPE—Adaptive variable selection via convex optimization. *Ann. Appl. Stat.*, 9(3):1103–1140, 2015.

[5] Peter Bühlmann and Sara van de Geer. *Statistics for high-dimensional data.* Springer Series in Statistics. Springer, Heidelberg, 2011. Methods, theory and applications.

[6] Emmanuel J. Candès and Yaniv Plan. Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *IEEE Trans. Inform. Theory*, 57(4):2342–2359, 2011.

[7] David Gross. Recovering low-rank matrices from few coefficients in any basis. *IEEE Trans. Inform. Theory*, 57(3):1548–1566, 2011.

[8] Vladimir Koltchinskii. *Oracle inequalities in empirical risk minimization and sparse recovery problems*, volume 2033 of *Lecture Notes in Mathematics.* Springer, Heidelberg, 2011. Lectures from the 38th Probability Summer School held in Saint-Flour, 2008, École d'Été de Probabilités de Saint-Flour. [Saint-Flour Probability Summer School].

[9] Vladimir Koltchinskii, Karim Lounici, and Alexandre B. Tsybakov. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *Ann. Statist.*, 39(5):2302–2329, 2011.

[10] Vladimir Koltchinskii and Shahar Mendelson. Bounding the Smallest Singular Value of a Random Matrix Without Concentration. *Int. Math. Res. Not. IMRN*, (23):12991–13008, 2015.

[11] Guillaume Lecué and Shahar Mendelson. Learning subgaussian classes: Upper and minimax bounds. Technical report, CNRS, Ecole polytechnique and Technion, 2013.

[12] Guillaume Lecué and Shahar Mendelson. Sparse recovery under weak moment assumptions. Technical report, CNRS, Ecole Polytechnique and Technion, 2014. To appear in Journal of the European Mathematical Society.

[13] Guillaume Lecué and Shahar Mendelson. Regularization and the small-ball method ii: complexity-based bounds. Technical report, CNRS, ENSAE and Technion, I.I.T., 2015.

[14] Michel Ledoux. *The concentration of measure phenomenon*, volume 89 of *Mathematical Surveys and Monographs.* American Mathematical Society, Providence, RI, 2001.

[15] Karim Lounici. Sup-norm convergence rate and sign concentration property of Lasso and Dantzig estimators. *Electron. J. Stat.*, 2:90–102, 2008.

[16] Nicolai Meinshausen. Hierarchical testing of variable importance. *Biometrika*, 95(2):265–278, 2008.

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

[18] Nicolai Meinshausen, Lukas Meier, and Peter Bühlmann. p-values for high-dimensional regression. *J. Amer. Statist. Assoc.*, 104(488):1671–1681, 2009.

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

[20] Shahar Mendelson. Upper bounds on product and multiplier empirical processes. Technical report. To appear in Stochastic Processes and their Applications.

[21] Shahar Mendelson. Learning without concentration for general loss function. Technical report, Technion, I.I.T., 2013. arXiv:1410.3192.

[22] Shahar Mendelson. Learning without concentration. In *Proceedings of the 27th annual conference on Learning Theory COLT14*, pages pp 25–39. 2014.

[23] Shahar Mendelson. A remark on the diameter of random sections of convex bodies. In *Geometric aspects of functional analysis*, volume 2116 of *Lecture Notes in Math.*, pages 395–404. Springer, Cham, 2014.

[24] Shahar Mendelson. Learning without concentration. *J. ACM*, 62(3):Art. 21, 25, 2015.

[25] Shahar Mendelson. 'local vs. global parameters', breaking the gaussian compexity barrier. Technical report, Technion, I.I.T., 2015.

[26] Shahar Mendelson. On multiplier processes under weak moment assumptions. Technical report, Technion, I.I.T., 2015.

[27] Sahand Negahban and Martin J. Wainwright. Restricted strong convexity and weighted matrix completion: optimal bounds with noise. *J. Mach. Learn. Res.*, 13:1665–1697, 2012.

[28] Sahand N. Negahban, Pradeep Ravikumar, Martin J. Wainwright, and Bin Yu. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statist. Sci.*, 27(4):538–557, 2012.

[29] Benjamin Recht, Maryam Fazel, and Pablo A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Rev.*, 52(3):471–501, 2010.

[30] Angelika Rohde and Alexandre B. Tsybakov. Estimation of high-dimensional low-rank matrices. *Ann. Statist.*, 39(2):887–930, 2011.

[31] Mark Rudelson and Roman Vershynin. Small ball probabilities for linear images of high dimensional distributions. Technical report, University of Michigan, 2014. International Mathematics Research Notices, to appear. [arXiv:1402.4492].

[32] W. Su and E. J. Candès. Slope is adaptive to unknown sparsity and asymptotically minimax. Technical report, Stanford University, 2015. To appear in The Annals of Statistics.

[33] Michel Talagrand. *Upper and lower bounds for stochastic processes*, volume 60 of *Ergebnisse der Mathematik und ihrer Grenzgebiete. 3. Folge. A Series of Modern Surveys in Mathematics [Results in Mathematics and Related Areas. 3rd Series. A Series of Modern Surveys in Mathematics]*. Springer, Heidelberg, 2014. Modern methods and classical problems.

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

[35] Sara van de Geer. Weakly decomposable regularization penalties and structured sparsity. *Scand. J. Stat.*, 41(1):72–86, 2014.

[36] Sara A. van de Geer. The deterministic lasso. Technical report, ETH Zürich, 2007. http://www.stat.math.ethz.ch/ geer/lasso.pdf.

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

[38] G. A. Watson. Characterization of the subdifferential of some matrix norms. *Linear Algebra Appl.*, 170:33–45, 1992.

[39] Peng Zhao and Bin Yu. On model selection consistency of Lasso. *J. Mach. Learn. Res.*, 7:2541–2563, 2006.