Learning subgaussian classes : Upper and minimax bounds

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

July 17, 2015

Abstract

We obtain oracle inequalities for empirical risk minimization when the class $\mathcal{F}$ and the target Y are subgaussian. The bound we obtain is sharp in the minimax sense if $\mathcal{F}$ is convex. Moreover, under mild additional assumptions on $\mathcal{F}$, the error rate remains optimal even if the procedure is allowed to perform with constant probability. As a part of our analysis we present a new proof of minimax results for the gaussian regression model and study linear regression in the unit ℓ_1^d-ball and matrix regression in the max-norm ball.

1 Introduction and main results

Let $\mathcal{D} := \{(X_i, Y_i) : i = 1, \cdots, N\}$ be a set of N i.i.d random variables with values in $\mathcal{X} \times \mathbb{R}$. From a statistical standpoint, each X_i can be viewed as an input associated with an output Y_i. Given a new input X, one would like to guess its associated output Y, assuming that (X, Y) is distributed according to the same probability distribution that generated the data $\mathcal{D}$. To that end, one may use $\mathcal{D}$ to construct a function $\hat{f}_N(\mathcal{D}, \cdot) = \hat{f}_N(\cdot)$, and the hope is that $\hat{f}_N(X)$ is close to Y in some sense.

Here, we will consider the *squared loss function* $\ell : \mathbb{R} \times \mathbb{R} \mapsto \mathbb{R}$, defined by $\ell(u, v) = (u - v)^2$, as a way of measuring the pointwise error $\ell(f(X), Y)$, and the resulting *squared risk* is

$$R(f) = \mathbb{E}\big(f(X) - Y\big)^2 \text{ and } R(\hat{f}_N) = \mathbb{E}\big(\big(\hat{f}_N(X) - Y\big)^2 | \mathcal{D}\big).$$

[1]CNRS, CMAP, Ecole Polytechnique, 91120 Palaiseau, France.
[2]Department of Mathematics, Technion, I.I.T, Haifa 32000, Israel.
[3]Email: guillaume.lecue@cmap.polytechnique.fr
[4]Email: shahar@tx.technion.ac.il
[5]Supported by the Mathematical Sciences Institute – The Australian National University and by the Israel Science Foundation.

In the classical statistics setup, one usually assumes that the regression function of Y given X belongs to some particular function space (called a *model*). In contrast, in the Learning setup on which we focus here, one is given a function class $\mathcal{F}$ (also called a model), and the goal is to construct a procedure $\hat{f}_N$ that satisfies a *sharp* or *exact oracle inequality* (following [31]; such bounds are called excess risk bounds in [24] and [20]). An exact oracle inequality ensures that with high probability,

$$R(\hat{f}_N) \leq \inf_{f \in \mathcal{F}} R(f) + \text{residue}, \tag{1.1}$$

and one would like to make the residue as small as possible.

For the sake of simplicity, assume that there is some $f^* \in \mathcal{F}$ minimizing the risk in $\mathcal{F}$, though the claims presented here remain true without that assumption.

Observe that if X is distributed according to a measure μ, the procedure $\hat{f}_N$ is a map from the set of N-samples into $\mathcal{F}$. It performs with accuracy $\varepsilon_N = \varepsilon_N(\mathcal{F})$ and confidence $1 - \delta_N = 1 - \delta_N(\mathcal{F})$ if for any reasonable target Y, (1.1) is satisfied on an event of measure at least $1 - \delta_N$ with respect to the product probability measure endowed by (X, Y) and with a residue that is at most ε_N.

Clearly, the risk functional is not known but its empirical version

$$R_N(f) = \frac{1}{N} \sum_{i=1}^{N} \big(f(X_i) - Y_i\big)^2$$

is. Thus, a natural procedure that comes to mind is minimizing the empirical risk in $\mathcal{F}$ – the so-called *empirical risk minimization* (*ERM*), which is defined by

$$\hat{f} \in \operatorname*{argmin}_{f \in \mathcal{F}} R_N(f).$$

ERM has been studied extensively over the last 30 years (see, e.g. [36], [24], [20] and references therein), and the main focus has always been to identify the connections between the structure of $\mathcal{F}$ and the accuracy and confidence that ERM yields.

Among the natural questions regarding the performance of ERM are:

1. Given any $0 < \delta_N < 1/2$, what is the error rate ε_N that one may obtain using ERM, and what features of $\mathcal{F}$ govern that rate?

2. Given any $0 < \delta_N < 1/2$, does ERM achieve the minimax rates for the confidence level δ_N? In other words, is there a procedure that can

perform with a better accuracy than ERM, given the same confidence level?

The majority of results on the performance of ERM have been obtained in the bounded case: when $\sup_{f\in\mathcal{F}}|\ell(Y,f(X))|\leq b$ almost surely, or, alternatively, when the envelope function $\sup_{f\in\mathcal{F}}|\ell(Y,f(X))|$ is well behaved in some weaker sense (e.g., has a sub-exponential tail). A result in this direction is from [2] (see Corollary 5.3 there) which we formulate using the notation of Theorem 5.1 in [20].

For any $\gamma>0$, let

$$k_N(r)=\mathbb{E}\sup\Big(\Big|\frac{1}{\sqrt{N}}\sum_{i=1}^N\varepsilon_i(f-f^*)(X_i)\Big|:f\in\mathcal{F},\|f-f^*\|_{L_2(\mu)}\leq 2r\Big),$$

and set

$$k_N^*(\gamma)=\inf\Big\{r>0:8k_N(r)\leq\gamma r^2\sqrt{N}\Big\}.$$

Theorem 1.1 *There exist absolute constants c_0,c_1 and $q>2$ for which the following holds. If $\mathcal{F}$ is a convex class of functions that are bounded by 1, then for every $t>0$, with probability at least $1-c_0\exp(-t)$,*

$$R(\hat{f})\leq\inf_{f\in\mathcal{F}}R(f)+c_1\max\Big(\big(k_N^*(1/q)\big)^2,\frac{t}{N}\Big)\Big). \tag{1.2}$$

A result of a similar flavour was obtained in [6]: let $N(A,B)$ be the number of translates of B needed to cover A. Set D to be the unit ball in $L_2(\mu)$ and let

$$\sigma^*=\inf\Big\{r>0:\int_{c_0r^2}^{c_1r}\log^{1/2}N\big(\mathcal{F}\cap(2rD),\varepsilon D\big)d\varepsilon\leq c_2r^2\sqrt{N}\Big\}, \tag{1.3}$$

for absolute constants c_0,c_1,c_2.

The result in [6] is that under various assumptions on the class $\mathcal{F}$ (assumptions that allow one to upper bound the function $k_N(r)$ using the entropy integral in (1.3)), $(\sigma^*)^2$ may serve as a residual term.

Both results relay heavily on the assumption that $\mathcal{F}$ is bounded in L_∞ and their proofs do not extend beyond the bounded case.

Our aim here is to go beyond the bounded case and proceed without any assumption on the envelope of $\{\ell(f(X),Y):f\in\mathcal{F}\}$. The subgaussian framework is natural for such results, as it captures many typical applications in which the functions involved are unbounded: for example, regression with a gaussian noise; compressed sensing; matrix completion; phase recovery, etc. [7, 10, 9, 8, 18, 19].

Definition 1.2 *Let X be distributed according to a probability measure μ. Set*

$$\|f\|_{\psi_2(\mu)} = \inf\left\{c > 0 : \mathbb{E}\exp(f^2(X)/c^2) \leq 2\right\},$$

and denote by $L_{\psi_2} = L_{\psi_2(\mu)}$ the space of functions with a finite ψ_2-norm.

Definition 1.3 *A function class $\mathcal{F} \subset L_2(\mu)$ is L-subgaussian with respect to the probability measure μ if for every $f, h \in \mathcal{F} \cup \{0\}$, $\|f-h\|_{\psi_2(\mu)} \leq L\|f-h\|_{L_2(\mu)}$, and the canonical gaussian process indexed by $\mathcal{F}$, $\{G_f : f \in \mathcal{F}\}$ is bounded (see the book [15] for an extensive survey on gaussian processes).*

What is probably the most interesting set of examples that belong to the subgaussian framework is classes of linear functionals on $\mathbb{R}^d$.

Definition 1.4 *A probability measure μ on $\mathbb{R}^d$ is L-subgaussian for some $L > 0$, if for every $t \in \mathbb{R}^d$, $\left\|\langle t, \cdot\rangle\right\|_{\psi_2(\mu)} \leq L\left\|\langle t, \cdot\rangle\right\|_{L_2(\mu)}$. The measure μ is isotropic if $\left\|\langle t, \cdot\rangle\right\|_{L_2(\mu)} = \|t\|_{\ell_2^d}^2$ for every $t \in \mathbb{R}^d$, where $\|\cdot\|_{\ell_2^d}$ denotes the Euclidean norm in $\mathbb{R}^d$.*

There are many natural examples of subgaussian measures on $\mathbb{R}^d$:

• Let x be a real-valued random variable that has mean-zero and variance 1. If $\|x\|_{\psi_2(\mu)} \leq L\|x\|_{L_2(\mu)}$ and $x_1, \ldots, x_d$ are independent copies of x, then for every $a \in \mathbb{R}^d$,

$$\Big\|\sum_{i=1}^d a_i x_i\Big\|_{\psi_2(\mu)} \lesssim \|a\|_{\ell_2^d}\|x\|_{\psi_2(\mu)} \leq L\Big\|\sum_{i=1}^d a_i x_i\Big\|_{L_2(\mu)}$$

where we write $u \lesssim v$ if $u \leq c_0 v$ for an absolute constant c_0. Thus, the random vector $X = (x_1, ..., x_d)$ is L-subgaussian. Moreover, it is clearly isotropic.

Among the examples of such product measures are the uniform measure on the combinatorial cube $\{-1, 1\}^d$, the uniform measure on the cube $[-1, 1]^d$ or the canonical gaussian measure on $\mathbb{R}^d$.

• Let $2 \leq p < \infty$ and denote by B_p^d the unit ball of $(\mathbb{R}^d, \|\ \|_{\ell_p})$. The uniform measure on $d^{1/p}B_p^d$ is L-subgaussian for an absolute constant L (see [1]), despite the fact that its coordinates are not independent.

• Let $X = (x_i)_{i=1}^d$ be an *unconditional* random vector (that is, $(\varepsilon_i x_i)_{i=1}^d$ has the same distribution as X for every choice of signs $(\varepsilon_i)_{i=1}^d$). If $\mathbb{E}x_i^2 \geq c^2$ for

every $1 \le i \le d$ and X is supported in RB_∞^d, then it is L-subgaussian for $L \lesssim R/c$. Indeed, it is straightforward to verify that for every $f \in L_{\psi_2(\mu)}$,

$$c_1\|f\|_{\psi_2(\mu)} \le \sup_{p\ge 2} \frac{\|f\|_{L_p(\mu)}}{\sqrt{p}} \le c_2\|f\|_{\psi_2(\mu)}$$

for suitable absolute constants c_1 and c_2. Thus, it suffices to show that for every $t \in \mathbb{R}^d$ and every $p \ge 2$,

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

By Khintchine's inequality [22],

$$\begin{aligned}\|\langle X, t\rangle\|_{L_p}^p =& \mathbb{E}\Big|\sum_{j=1}^d x_j t_j\Big|^p = \mathbb{E}_X\mathbb{E}_\varepsilon\Big|\sum_{j=1}^d \varepsilon_j x_j t_j\Big|^p \\ \lesssim& p^{p/2}\mathbb{E}_X\Big(\sum_{j=1}^d x_j^2 t_j^2\Big)^{p/2} \lesssim p^{p/2}R^p\|t\|_{\ell_2^d}^p.\end{aligned}$$

Also,

$$\|\langle X, t\rangle\|_{L_2}^2 = \mathbb{E}_X\mathbb{E}_\varepsilon\Big(\sum_{i=1}^d \varepsilon_i x_i t_i\Big)^2 = \mathbb{E}_X\sum_{i=1}^d x_i^2 t_i^2 \ge c^2\|t\|_{\ell_2^d}^2,$$

proving the claim.

• If x is a mean-zero, variance one, L-subgaussian random variable, and $X = (x_{i,j})$ is a matrix whose coordinates are independent copies of x, then X defines a cL subgaussian, isotropic measure on the space of matrices of the right dimensions, relative to the natural trace inner product. The same holds if X has independent rows, distributed according to an isotropic, L-subgaussian random vector. The proof of both facts is straightforward and thus omitted.

The above shows that even the seemingly restricted setup of classes of linear functionals on $\mathbb{R}^d$ endowed with an L-subgaussian measure is encountered in many natural (and well studied) examples.

The strategy we use here for the study of ERM is the isomorphic method, introduced in [3]. Before presenting it, recall that the excess loss of f is

$$\mathcal{L}_f(x, y) = \ell(f(x), y) - \ell(f^*(x), y) = (f(x) - y)^2 - (f^*(x) - y)^2 \qquad (1.4)$$

and set

$$P\mathcal{L}_f = \mathbb{E}\mathcal{L}_f(X,Y) \ \text{ and } \ P_N\mathcal{L}_f = \frac{1}{N}\sum_{i=1}^N \mathcal{L}_f(X_i,Y_i).$$

A rather obvious but very useful observation is that for every $f \in \mathcal{F}\backslash\{f^*\}$, $P\mathcal{L}_f > 0$, while the empirical minimizer $\hat{f}$ satisfies that $P_N\mathcal{L}_{\hat{f}} \leq 0$.

The isomorphic method is based on the following idea. Consider an event Ω_0, on which for every function in the set $\{f \in \mathcal{F} : P\mathcal{L}_f \geq \lambda_N\}$,

$$\frac{1}{2}P\mathcal{L}_f \leq P_N\mathcal{L}_f \leq \frac{3}{2}P\mathcal{L}_f. \tag{1.5}$$

It follows that on Ω_0, ERM produces $\hat{f}$ that satisfies

$$R(\hat{f}) \leq \inf_{f\in\mathcal{F}} R(f) + \lambda_N,$$

because $P_N\mathcal{L}_{\hat{f}} \leq 0$ and thus $\hat{f} \notin \{f \in \mathcal{F} : P\mathcal{L}_f \geq \lambda_N\}$.

Consequently, an exact oracle inequality with a confidence parameter δ_N may be derived by identifying λ_N for which Ω_0 has probability at least $1-\delta_N$; that is, the level λ_N for which

$$\sup_{\{f\in\mathcal{F}:P\mathcal{L}_f\geq\lambda_N\}} \Big|\frac{1}{N}\sum_{i=1}^N \frac{\mathcal{L}_f(X_i,Y_i)}{P\mathcal{L}_f} - 1\Big| \leq \frac{1}{2}$$

with probability at least $1-\delta_N$ (see Theorem 4.4 in [20] for results of a similar flavour).

Remark 1.5 *Note that only the lower estimate in* (1.5) *is needed for the argument outlined above to work. This observation is the key in the application of the small-ball method to learning problems in [?, ?], which allows one to deal with heavy-tailed scenarios.*

Just like k_N^* and σ^* in (1.2) and (1.3) respectively, and other well known estimates on the performance of ERM (e.g. [35, 20, 24]), the residual term we use is defined in terms of fixed points.

Let $\{G_f : f \in \mathcal{F}\}$ be the canonical gaussian process indexed by $\mathcal{F}$. Given a set $\mathcal{F}' \subset \mathcal{F}$, denote by $d_{L_2}(\mathcal{F}')$ its diameter in $L_2(\mu)$ and put

$$\mathbb{E}\|G\|_{\mathcal{F}'} = \sup\{\mathbb{E}\sup_{h\in\mathcal{H}} G_h \ : \ \mathcal{H} \subset \mathcal{F}' \text{ is finite}\}.$$

Recall that D is the unit ball in $L_2(\mu)$, and set $sD = \{f \in L_2(\mu) : \|f\|_{L_2(\mu)} \leq s\}$ and $\mathcal{F} - \mathcal{F} = \{f - g : f, g \in \mathcal{F}\}$.

For every $\eta > 0$, let

$$s_N^*(\eta) = \inf\left\{0 < s \leq d_{\mathcal{F}}(L_2) : \mathbb{E}\|G\|_{sD\cap(\mathcal{F}-\mathcal{F})} \leq \eta s^2\sqrt{N}\right\}, \tag{1.6}$$

and for every $Q > 0$, set

$$r_N^*(Q) = \inf\left\{0 < r \leq d_{\mathcal{F}}(L_2) : \mathbb{E}\|G\|_{rD\cap(\mathcal{F}-\mathcal{F})} \leq Qr\sqrt{N}\right\}.$$

Remark 1.6 *We will assume without explicitly stating it that $s_N^*(\eta), r_N^*(Q) < d_{\mathcal{F}}(L_2)$, which is always the case if N is large enough.*

With these definitions in place, one may formulate a restricted version of the upper bound on the performance of ERM – for a convex, L-subgaussian class of functions.

Theorem A. *For every $L \geq 1$ there exist constants c_1, c_2, c_3 and c_4 that depend only on L for which the following holds. Let $\mathcal{F} \subset L_2(\mu)$ be a convex, L-subgaussian class of functions, assume that $\|Y - f^*(X)\|_{\psi_2} \leq \sigma$ and set $\eta = c_1/\sigma$ and $Q = c_2$.*

1. *If $\sigma \geq c_3 r_N^*(Q)$ then with probability at least $1 - 4\exp(-c_4 N\eta^2(s_N^*(\eta))^2)$,*

$$R(\hat{f}) \leq \inf_{f\in\mathcal{F}} R(f) + (s_N^*(\eta))^2.$$

2. *If $\sigma \leq c_3 r_N^*(Q)$ then with probability at least $1 - 4\exp(-c_4 NQ^2)$,*

$$R(\hat{f}) \leq \inf_{f\in\mathcal{F}} R(f) + (r_N^*(Q))^2.$$

Hence, with probability at least $1 - 4\exp\left(-c(L)N\min\left\{1, \eta^2(s_N^*(\eta))^2\right\}\right)$,

$$R(\hat{f}) \leq \inf_{f\in\mathcal{F}} R(f) + \max\left\{(s_N^*(\eta))^2, (r_N^*(Q))^2\right\}.$$

We will show in what follows that the parameters involved in the upper bound have very clear roles. r_N^* is an upper estimate (that is often sharp but not always) on the error rate one could have if the problem were noise-free – that is, if $\sigma = 0$. This intrinsic error occurs because it is impossible to

distinguish between $f_1, f_2 \in \mathcal{F}$ on the sample $\mathbb{X} = (X_i)_{i=1}^N$ if $(f_1(X_i))_{i=1}^N = (f_2(X_i))_{i=1}^N$.

Once noise is introduced to the problem and passes a certain threshold, it is no longer realistic to expect that an intrinsic parameter, which does not depend on the noise level, can serve as an upper bound. And, indeed, $s_N^*(\eta)$ measures the interaction between the 'noise' $f^*(X) - Y$ and the class through the choice of $\eta \sim 1/\sigma$. Beyond a trivial threshold on σ, $s_N^*(c/\sigma)$ becomes the dominant term in the upper bound.

Of course, Theorem A is better justified if one can obtain matching lower bounds, showing that ERM is an optimal procedure. To that end, it seems natural to employ minimax theory (see, e.g., [32, 38, 39, 6, 5] for more details).

Standard minimax bounds are based on information-theoretical results such as Fano's Lemma, Assouad's Lemma or Pinsker's inequalities. Unfortunately, these results do not yield lower bounds in the high probability realm, which the range needed here if one is to show the optimality of the rate obtained in Theorem A; rather, these results are restricted to the constant probability regime. To treat the high probability regime, we present a new minimax bound which is based on the gaussian shift theorem (and therefore on the gaussian isoperimetric inequality).

Fix $f \in \mathcal{F}$ and set $W \sim \mathcal{N}(0, \sigma^2)$, i.e., a gaussian noise that is independent of X. Consider the case in which $(X_i, Y_i)_{i=1}^N$ is an independent sample of

$$Y^f = f(X) + W. \tag{1.7}$$

Theorem A′. *There exist absolute constants c_1, c_2 and c_3 for which the following holds. Let $\mathcal{F} \subset L_2(\mu)$ be a class that is star-shaped around one of its points (i.e., for some $f_0 \in \mathcal{F}$ and every $f \in \mathcal{F}$, $[f_0, f] \subset \mathcal{F}$). If $\tilde{f}_N$ is constructed from a sample of cardinality N of (1.7), and has a confidence parameter δ_N, then its accuracy satisfies*

$$\varepsilon_N \geq c_1 \sigma^2 \frac{\log(1/\delta_N)}{N}.$$

In particular, if $\delta_N = \exp(-c_2 \eta^2 (s_N^*(\eta))^2 N)$ for $\eta = c(L)/\sigma$, (as in Theorem A when the noise level is nontrivial), the best accuracy that may be achieved by *any procedure* is

$$\varepsilon_N \geq c_3 \sigma^2 \eta^2 (s_N^*(\eta))^2 \sim (s_N^*(\eta))^2.$$

Thus, ERM achieves the minimax rate for that confidence level.

The second question we wish to address is what happens when the desired confidence is an absolute constant, say $\delta_N \sim 1/2$, but still, when the noise level is nontrivial. We will show that Theorem A (and in particular, the isomorphic method) is optimal in a minimax sense under some regularity assumptions on $\mathcal{F}$, even when $\delta_N \sim 1/2$.

Consider the 'Sudakov analog' of the gaussian-based parameter $s_N^*(\eta)$: recall that by Sudakov's inequality (see, for example, [22]), for any $r > 0$,

$$\sup_{\varepsilon>0} \varepsilon \log^{1/2} N((\mathcal{F}-\mathcal{F}) \cap rD, \varepsilon D) \lesssim \mathbb{E}\|G\|_{rD\cap(\mathcal{F}-\mathcal{F})}. \tag{1.8}$$

Put $C(r) = \sup_{f\in\mathcal{F}} r \log^{1/2} N((\mathcal{F}-f)\cap 2rD, rD)$ and set

$$q_N^*(\eta) = \inf\{s > 0 : C(s) \leq \eta s^2\sqrt{N}\},$$

where again, we assume implicitly that $q_N^*(\eta) \leq d_{\mathcal{F}}(L_2)$.

Theorem B. *There exist absolute constants c_1 and c_2 for which the following holds. Let $\mathcal{F}$ be a class of functions, set $W \sim \mathcal{N}(0,\sigma^2)$ and for every $f \in \mathcal{F}$, put $Y^f = f(X) + W$. If $\tilde{f}_N$ performs with a confidence parameter $\delta_N < 1/4$ for every such target Y^f, then its accuracy can not be better than $c_1(q_N^*(c_2/\sigma))^2$.*

Theorem B is known, and may be derived from Theorem 2.5 in [32] or from [38]. The proof presented here is new, and follows the same path as the proof of Theorem A$'$.

With Theorem A in mind, Theorem B shows that if $s_N^*(\eta)$ and $q_N^*(\eta)$ are equivalent for $\eta \sim 1/\sigma$ and $\sigma \gtrsim r_N^*$, the minimax rate in the constant probability regime is attained by ERM, and therefore ERM is a minimax procedure.

Finally, let us consider the low-noise case, in which $\sigma \lesssim r_N^*$. Although it is not clear if r_N^* is an optimal bound in that range (except when $\sigma \sim r_N^*$), it is not far from optimal.

Definition 1.7 *Let $\mathcal{F}$ be a class of functions. For every sample $\mathbb{X} = (X_1, ..., X_N)$ and $f \in \mathcal{F}$, set*

$$K(f, \mathbb{X}) = \{h \in \mathcal{F} : (f(X_i))_{i=1}^N = (h(X_i))_{i=1}^N\},$$

which is the "level set" in $\mathcal{F}$ given by the values of f on the sample. Let $\mathcal{D}(f,\mathbb{X})$ be the $L_2(\mu)$ diameter of $K(f,\mathbb{X})$.

Clearly, if $\sigma = 0$ then for every sample $\mathbb{X}$, ERM selects $\hat{f} \in K(f^*,\mathbb{X})$ and since $Y = f^*(X)$, $R(f) = \|f - f^*\|^2_{L_2(\mu)}$. Thus, $R(\hat{f}) \leq \mathcal{D}^2(f^*,\mathbb{X})$. It is natural to ask whether the reverse direction is true, and also study what happens when $0 < \sigma < r^*_N$.

The following result shows that the largest typical value of $\mathcal{D}(f,\mathbb{X})$ is a constant-probability minimax bound, regardless of the choice of σ. It is a combination of compressed sensing type of minimax results (see, e.g., [14, 11]) and more classical ones (e.g. [32, 38, 39, 6, 5]).

Theorem C. *For every $f \in \mathcal{F}$ and V that is independent of X, set $Y^f = f(X) + V$. Then, for any procedure $\tilde{f}_N$,*

$$\sup_{f\in\mathcal{F}} \mathbb{P}\left(\|\tilde{f}_N((Y_i^f, X_i)_{i=1}^N) - f\|_{L_2(\mu)} \geq \frac{1}{4}\mathcal{D}(f,\mathbb{X}) \right) \geq 1/2.$$

One natural example in which Theorem C may be used is when T is a convex, centrally-symmetric subset of $\mathbb{R}^d$ (i.e., if $t \in T$ then $-t \in T$), and $\mathcal{F}$ is the class of linear functionals associated with T, $\{\langle t, \cdot\rangle : t \in T\}$. Let $X_1, ..., X_N$ be an independent sample selected according to an isotropic probability measure on $\mathbb{R}^d$. If $\{e_1, \ldots, e_N\}$ is the canonical basis of $\mathbb{R}^N$ and $\Gamma = \sum_{i=1}^N \langle X_i, \cdot\rangle e_i$ is the random matrix whose rows are $(X_i)_{i=1}^N$, then $\mathcal{D}(0,\mathbb{X})$ is the diameter of $\ker(\Gamma) \cap T$.

Recall that the Gelfand N-width of T is the smallest ℓ_2^d-diameter of an N-codimensional section of T, and denote it by $c_N(T)$. Hence, for every $t_0 \in T$,

$$c_N(T) \leq \operatorname{diam}\left(K(t_0,\mathbb{X}) - t_0, \ell_2^d\right) \leq 2\mathcal{D}(0,\mathbb{X}),$$

and by Theorem C, $c_N(T)/8$ is a lower estimate on a constant probability minimax bound. Therefore, in cases where $r^*_N \sim c_N(T)$, it follows that for every $0 \leq \sigma \lesssim r^*_N$, r^*_N is the constant-probability minimax rate, and that rate is achieved by ERM. We will present such an example in Section 4.

Finally, it should be noted that although our presentation focuses on oracle inequalities in the given class, oracle inequalities for model selection and regularized procedures can be derived from the isomorphic method in general, and in particular, directly from Theorem 2.7 below. This strategy is

rather standard and has been used, for example, in [4, 27] and in Chapter 3.6 of [21]. We will not present results in that direction in what follows.

We end this introduction with a word about notation. Throughout, absolute constants or constants that depend on other parameters are denoted by c, C, c_1, c_2, etc., (and, of course, we will specify when a constant is absolute and when it depends on other parameters); their values may change from line to line. The notation $x \sim y$ (resp. $x \lesssim y$) means that there exist absolute constants $0 < c < C$ for which $cy \leq x \leq Cy$ (resp. $x \leq Cy$). If $b > 0$ is a parameter then $x \lesssim_b y$ means that $x \leq C(b)y$ for some constant $C(b)$ that depends only on b.

Let ℓ_p^d be $\mathbb{R}^d$ endowed with the norm $\|x\|_{\ell_p^d} = \big(\sum_{j=1}^d |x_j|^p\big)^{1/p}$. The unit ball in ℓ_p^d is denoted by B_p^d, and the unit Euclidean sphere in $\mathbb{R}^d$ is S^{d-1}.

2 Proof of Theorem A

The proof of Theorem A shows that it is more general than stated. Rather than convexity, the two properties that are actually needed are the following:

Definition 2.1 *A class $\mathcal{H}$ is star-shaped around $h_0 \in \mathcal{H}$ if for every $h \in \mathcal{H}$, the interval $[h, h_0] \subset \mathcal{H}$.*

We will assume that $\mathcal{F} - \mathcal{F} = \{f - h : f, h \in \mathcal{F}\}$ is star-shaped around 0, otherwise, one may consider the star-shaped hull of $\mathcal{F} - \mathcal{F}$ with 0, that is, the set

$$\{\lambda(f - h) : 0 \leq \lambda \leq 1,\ f, h \in \mathcal{F}\}$$

which is not much larger than $\mathcal{F} - \mathcal{F}$.

The second property required is a variant of the Bernstein condition.

Definition 2.2 *A class $\mathcal{F}$ is B-Bernstein relative to the target Y, if for every $f \in \mathcal{F}$,*

$$\mathbb{E}\big(f(X) - f^*(X)\big)^2 \leq BP\mathcal{L}_f = B\mathbb{E}\big((Y - f(X))^2 - (Y - f^*(X))^2\big). \quad (2.1)$$

Definition 2.2 is far less restrictive than it appears at first glance. Indeed, by the 2-convexity of L_2, if $\mathcal{F}$ is convex then for any target $Y \in L_2$, $\mathcal{F}$ is 1-Bernstein relative to Y. Moreover, the results from [26] show that for every class $\mathcal{F}$ and every target Y, the Bernstein constant depends only on the distance of Y from the set of targets Z for which the functional $\mathbb{E}(f - Z)^2$ has multiple minimizers in $\mathcal{F}$.

If one wishes $\mathcal{F}$ to satisfy a Bernstein condition relative to *every* target Y, it forces $\mathcal{F}$ to be convex (see Section 5).

Since all these extensions are well known, we will assume from here on that $\mathcal{F}-\mathcal{F}$ is star-shaped around 0 and that $\mathcal{F}$ satisfies (2.1).

The next lemma shows that the assumption that $\mathcal{F}-\mathcal{F}$ is star-shaped around 0 adds some regularity to the gaussian process $\{G_f : f \in \mathcal{F}\}$.

Lemma 2.3 *Let $\psi(s) = \mathbb{E}\,\|G\|_{sD\cap(\mathcal{F}-\mathcal{F})}$. If $\mathcal{F}-\mathcal{F}$ is star-shaped around 0 then $\psi(s)/s$ is non-increasing in $(0,\infty)$.*

Proof. Fix $s_1 > s_2 > 0$ and $f, h \in \mathcal{F}$. Assume that $s_2 \le \|f-h\|_{L_2(\mu)} \le s_1$ and observe that since $\mathcal{F}-\mathcal{F}$ is star-shaped around 0 and $0 < s_2/\|f-h\|_{L_2(\mu)} < 1$,

$$u = s_2\frac{f-h}{\|f-h\|_{L_2(\mu)}} \in s_2 D \cap (\mathcal{F}-\mathcal{F}).$$

Therefore,

$$G_{f-h} = \frac{\|f-h\|_{L_2(\mu)}}{s_2}G_u \le (s_1/s_2)\sup_{w\in s_2 D\cap(\mathcal{F}-\mathcal{F})} G_w. \tag{2.2}$$

Since (2.2) clearly holds if $\|f-h\|_{L_2(\mu)} \le s_1$, the claim follows by taking the supremum over all possible choices of $f-h \in s_1 D \cap (\mathcal{F}-\mathcal{F})$. ∎

An immediate corollary of Lemma 2.3 is that if $s > s^*_N(\eta)$ then $\psi(s) \le \eta s^2\sqrt{N}$, and if $s < s^*_N(\eta)$, the reverse inequality holds. A similar observation is true for $r^*_N(Q)$.

When considering the parameters r^*_N and s^*_N, what seems odd at first glance is the different normalization in their definition – the first condition is linear, while the second is quadratic. The two originate from the need to compare the way in which two processes scale with $\|f-f^*\|_{L_2(\mu)}$. Indeed, note that

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

The quadratic term is noise-free, and as will be explained below, r^*_N measures the lowest level r at which if $\|f-f^*\|_{L_2(\mu)} \ge r$, $\mathbb{E}(f-f^*)^2 \sim N^{-1}\sum_{i=1}^N (f-f^*)^2(X_i)$.

In contrast, s_N^* is designed for the multiplier process, originating from the linear term $(f^*(X)-Y)\cdot(f-f^*)(X)$. To compare the resulting multiplier component with $\mathbb{E}(f-f^*)^2$, one has to study

$$f \to \frac{1}{N}\sum_{i=1}^N (f^*(X_i)-Y_i)\cdot\frac{(f-f^*)(X_i)}{\mathbb{E}(f-f^*)^2},$$

which is the source of the seemingly less-natural normalization in the definition of $s_N^*(\eta)$.

Let us begin with an estimate on the quadratic component, which is based on a functional Bernstein type inequality:

Theorem 2.4 *[?] There exist absolute constants c_1, c_2 and c_3 for which the following holds. If $\mathcal{H}$ is an L-subgaussian class, then for every $t \geq c_1$, with probability at least $1-2\exp(-c_2t^2(\mathbb{E}\,\|G\|_{\mathcal{H}}/Ld_{L_2}(\mathcal{H}))^2)$,*

$$\sup_{h\in\mathcal{H}}\left|\sum_{i=1}^N h^2(X_i)-\mathbb{E}h^2\right| \leq c_3L^2\left(td_{L_2}(\mathcal{H})\mathbb{E}\,\|G\|_{\mathcal{H}}\sqrt{N}+t^2(\mathbb{E}\,\|G\|_{\mathcal{H}})^2\right).$$

A different proof of Theorem 2.4 may be found in [**?**].

A straightforward application of Theorem 2.4 leads to an isomorphic estimate and illustrates the role of r_N^*:

Lemma 2.5 *There exist absolute constants c_1, c_2 and c_3 for which the following holds. Let $\mathcal{F}$ be an L-subgaussian class, assume that $\mathcal{F}-\mathcal{F}$ is star-shaped around 0 and let $f^* \in \mathcal{F}$. If $0 < Q \leq 1$ and $r \geq 2r_N^*(Q)$, then with probability at least $1-2\exp\big(-c_1(\mathbb{E}\,\|G\|_{rD\cap(\mathcal{F}-\mathcal{F})}/Lr)^2\big)$,*

$$\sup_{h\in rD\cap(\mathcal{F}-f^*)}\left|\frac{1}{N}\sum_{i=1}^N h^2(X_i)-\mathbb{E}h^2\right| \leq c_2L^2r^2Q. \tag{2.3}$$

If $Q \leq \min\{c_3/L^2, 1\}$ then on the same event, for every $f \in \mathcal{F}$ that satisfies $\|f-f^*\|_{L_2(\mu)} \geq r$,

$$\frac{1}{2}\mathbb{E}(f-f^*)^2 \leq \frac{1}{N}\sum_{i=1}^N (f-f^*)^2(X_i) \leq \frac{3}{2}\mathbb{E}(f-f^*)^2.$$

Proof. The first part of the claim is an immediate corollary of Theorem 2.4 and the fact that if $r \geq 2r_N^*(Q)$, $\mathbb{E}\|G\|_{rD\cap(\mathcal{F}-\mathcal{F})}/\sqrt{N} \leq Qr$. The second part is, by now, well understood and is presented for the sake of completeness.

Let Ω_0 be the event on which (2.3) holds. If $f \in \mathcal{F}$ and $\|f - f^*\|_{L_2(\mu)} \geq r$, set $h = r(f - f^*)/\|f - f^*\|_{L_2(\mu)}$. Since $\mathcal{F} - \mathcal{F}$ is star-shaped around 0, $h \in rD \cap (\mathcal{F} - \mathcal{F})$. Therefore, if $Q \leq 1/2c_2L^2$ and $(X_i)_{i=1}^N \in \Omega_0$,

$$\left| \frac{1}{N} \sum_{i=1}^N h^2(X_i) - \mathbb{E}h^2 \right| \leq c_2 Q L^2 r^2 \leq \frac{r^2}{2},$$

and

$$\left| \frac{1}{N} \sum_{i=1}^N (f - f^*)^2(X_i) - \mathbb{E}(f - f^*)^2 \right| \leq \frac{1}{2} \mathbb{E}(f - f^*)^2.$$

■

The second ingredient required for the proof of Theorem A is a bound on multiplier processes.

Theorem 2.6 *[?] There exist absolute constants c_1, c_2 and c_3 for which the following holds. If $\mathcal{H}$ is an L-subgaussian class and $\xi \in L_{\psi_2}$, then for every $t, w \geq c_1$, with probability at least $1 - 2\exp(-c_2 t^2 (\mathbb{E}\|G\|_{\mathcal{H}}/Ld_{L_2}(\mathcal{H}))^2) - 2\exp(-c_2 N \min\{w^2, w^4\})$,*

$$\sup_{h \in \mathcal{H}} \left| \sum_{i=1}^N \xi_i h(X_i) - \mathbb{E}\xi h(X) \right| \leq c_3 L t w \sqrt{N} \|\xi\|_{L_{\psi_2}} \mathbb{E}\|G\|_{\mathcal{H}}.$$

Note that in Theorem 2.6 one does not assume that ξ and X are independent, a fact that will be significant in what follows.

Combining the estimates on the quadratic and multiplier process leads to the following ratio estimate:

Theorem 2.7 *For every $L \geq 1$ and $B \geq 1$ there exist constants c_0, c_1, c_2 and c_3 that depend only on B and L, for which the following holds. Let $\mathcal{F}$ be an L-subgaussian class which is B-Bernstein relative to the target Y. Assume that $\mathcal{F} - \mathcal{F}$ is star-shaped around 0 and that $\|Y - f^*(X)\|_{\psi_2} \leq \sigma$. Let $\eta = c_0/\sigma$ and $Q = c_1$ and assume that $s_N^*(\eta), r_N^*(Q) \leq d_{\mathcal{F}}(L_2)$.*

1. *If $\sigma \geq c_2 r_N^*(Q)$, then with probability at least $1 - 4\exp\left(-c_3 \eta^2 (s_N^*(\eta))^2 N\right)$,*

$$\sup_{\{f \in \mathcal{F} : P\mathcal{L}_f \geq 2(s_N^*(\eta))^2\}} \left| \frac{1}{N} \sum_{i=1}^N \frac{\mathcal{L}_f(X_i, Y_i)}{P\mathcal{L}_f} - 1 \right| \leq \frac{1}{2}.$$

2. If $\sigma \le c_2 r_N^*(Q)$, then with probability at least $1 - 4\exp\left(-c_3 Q^2 N\right)$,

$$\sup_{\{f\in\mathcal{F}: P\mathcal{L}_f \ge 2(r_N^*(Q))^2\}} \left| \frac{1}{N}\sum_{i=1}^N \frac{\mathcal{L}_f(X_i,Y_i)}{P\mathcal{L}_f} - 1 \right| \le \frac{1}{2}.$$

Proof. Set $\xi = (f^*(X) - Y)$ and thus

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

Fix $\lambda > 0$ and let $\mathcal{F}_\lambda = \{f \in \mathcal{F} : P\mathcal{L}_f \ge \lambda\}$. Since $\mathcal{F}$ satisfies the B-Bernstein condition relative to Y, it follows that for every $f \in \mathcal{F}$, $\|f - f^*\|_{L_2(\mu)}^2 \le BP\mathcal{L}_f$. Moreover, if $f \in \mathcal{F}_\lambda$ then

$$\left\| \frac{f-f^*}{(P\mathcal{L}_f)^{1/2}} \right\|_{L_2(\mu)}^2 \le B \quad \text{and} \quad \left\| \frac{f-f^*}{P\mathcal{L}_f} \right\|_{L_2(\mu)}^2 \le \frac{B}{P\mathcal{L}_f} \le \frac{B}{\lambda}. \tag{2.4}$$

Therefore,

$$\begin{aligned}
&\sup_{f\in\mathcal{F}_\lambda} \left| \frac{1}{N}\sum_{i=1}^N \frac{\mathcal{L}_f(X_i,Y_i)}{P\mathcal{L}_f} - 1 \right| = \sup_{f\in\mathcal{F}_\lambda} \left| \frac{1}{N}\sum_{i=1}^N \frac{\mathcal{L}_f(X_i,Y_i) - P\mathcal{L}_f}{P\mathcal{L}_f} \right| \\
&\le \sup_{f\in\mathcal{F}_\lambda} \left| \sum_{i=1}^N \left(\frac{f-f^*}{\left(P\mathcal{L}_f\right)^{1/2}} \right)^2 (X_i) - \mathbb{E}\left(\frac{f-f^*}{(P\mathcal{L}_f)^{1/2}} \right)^2 \right| \\
&+2 \sup_{f\in\mathcal{F}_\lambda} \left| \frac{1}{N}\sum_{i=1}^N \xi_i \left(\frac{f-f^*}{P\mathcal{L}_f} \right)(X_i) - \frac{\mathbb{E}\xi(f-f^*)}{P\mathcal{L}_f} \right|.
\end{aligned}$$

Set

$$W_\lambda = \left\{ \frac{f-f^*}{(P\mathcal{L}_f)^{1/2}} : f \in \mathcal{F}_\lambda \right\}, \quad V_\lambda = \left\{ \frac{f-f^*}{P\mathcal{L}_f} : f \in \mathcal{F}_\lambda \right\},$$

and $\mathcal{H} = (\mathcal{F}-\mathcal{F}) \cap \sqrt{\lambda B} D$. Since $\mathcal{F}-\mathcal{F}$ is star-shaped around 0, it follows from (2.4) that

$$W_\lambda \subset \frac{1}{\sqrt{\lambda}}(\mathcal{F}-\mathcal{F}) \cap \sqrt{B}D \subset \frac{1}{\sqrt{\lambda}} \left((\mathcal{F}-\mathcal{F}) \cap \sqrt{\lambda B} D \right) = \frac{\mathcal{H}}{\sqrt{\lambda}},$$

and

$$V_\lambda \subset \frac{1}{\lambda}(\mathcal{F}-\mathcal{F}) \cap \left(\sqrt{\frac{B}{\lambda}} \right) D \subset \frac{1}{\lambda} \left((\mathcal{F}-\mathcal{F}) \cap \sqrt{\lambda B} D \right) = \frac{\mathcal{H}}{\lambda}.$$

Fix $Q = c_0$ and $\eta = c_1/\sigma$ for a choice of constants to be named later. Set $r = r_N^*(Q)$ and observe that

$$\begin{aligned}\mathbb{E}\|G\|_{rD\cap(\mathcal{F}-\mathcal{F})} \geq & \mathbb{E}\|G\|_{(r/2)D\cap(\mathcal{F}-\mathcal{F})} \geq Q(r/2)\sqrt{N} \\ = & \frac{Q}{2r\eta}\cdot \eta r^2\sqrt{N} > \eta r^2\sqrt{N},\end{aligned}$$

provided that $\sigma \geq 2c_1 r_N^*(Q)/Q$; therefore, $r_N^*(Q) \leq s_N^*(\eta)$. Next, set $\lambda = (s_N^*(\eta))^2/B$, $s = 2s_N^*(\eta)$ and consider $Q \lesssim 1/L^2B$. Note that $\mathbb{E}\|G\|_{sD\cap(\mathcal{F}-\mathcal{F})} \leq \eta s^2\sqrt{N}$ and that by Lemma 2.5,

$$\sup_{w\in W_\lambda}\left|\frac{1}{N}\sum_{i=1}^N w^2(X_i) - \mathbb{E}w^2\right| \leq c_2L^2QB \leq \frac{1}{4}$$

with probability at least $1 - 2\exp(-c_3\eta^2(s_N^*(\eta))^2N)$.

Moreover, if $\eta \lesssim 1/(B\sigma)$ then by Theorem 2.6, with the same probability estimate,

$$\sup_{v\in V_\lambda}\left|\frac{1}{N}\sum_{i=1}^N \xi_i v(X_i) - \mathbb{E}\xi v\right| \lesssim LB\sigma\eta \leq \frac{1}{4}.$$

Thus, for every $Q \lesssim 1/(L^2B)$ and $\eta \lesssim \sigma^{-1}\min\{B^{-1}, Q\}$, if $\sigma \gtrsim Q^{-1}r_N^*(Q)$ then with probability at least $1 - 4\exp(-c_3\eta^2(s_N^*(\eta))^2N)$,

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

on the set $\{f \in \mathcal{F} : P\mathcal{L}_f \geq \lambda\}$.

Next, if $\sigma \lesssim Q^{-1}r_N^*(Q)$ (for the same choice of constants as above) and $\lambda = 2(r_N^*(Q))^2/B$, it follows from Lemma 2.5 and Theorem 2.6 that with probability at least $1 - 4\exp(-c_4Q^2N)$,

$$\sup_{w\in W_\lambda}\left|\frac{1}{N}\sum_{i=1}^N w^2(X_i) - \mathbb{E}w^2\right| \leq \frac{1}{4} \text{ and } \sup_{v\in V_\lambda}\left|\frac{1}{N}\sum_{i=1}^N \xi_i v(X_i) - \mathbb{E}\xi v\right| \leq \frac{1}{4}.$$

■

Theorem A is an immediate outcome of Theorem 2.7 and the isomorphic method described in the introduction.

3 Minimax lower bounds

Let $\mathcal{F}$ be a class of functions on a probability space (Ω, μ), fix $f \in \mathcal{F}$, let W be a centred gaussian random variable that is independent of X and consider the target function $Y^f = f(X) + W$. For any $\mathbb{X} = (x_1, \dots, x_N) \in \Omega^N$ let $\nu_{f,\mathbb{X}}$ be the conditional probability measure of $(Y_i^f | X_i = x_i)_{i=1}^N$, which is given by

$$d\nu_{f,\mathbb{X}}(y) = \exp\left(-\frac{\|y - (f(x_i))_{i=1}^N\|_{\ell_2^N}^2}{2\sigma^2} \right) \cdot \frac{dy}{(\sqrt{2\pi}\sigma)^N},$$

and set $\nu_{f,\mathbb{X}} \otimes \mu^N$ to be the probability measure on $(\mathbb{R} \otimes \Omega)^N$ that generates the sample $(Y_i^f, X_i)_{i=1}^N$.

Let

$$\mathcal{B}(f, r) = \{h \in \mathcal{F} : \mathbb{E}\mathcal{L}_h \leq r\} = \{h \in \mathcal{F} : \mathbb{E}(f - h)^2 \leq r\},$$

where $\mathcal{L}_h(X, Y^f) = (Y^f - h(X))^2 - (Y^f - f(X))^2$.

If a procedure $\tilde{f}_N$ performs with accuracy ε_N and has a confidence parameter δ_N, then for every $f \in \mathcal{F}$,

$$(\nu_{f,\mathbb{X}} \otimes \mu^N)\left(\tilde{f}_N^{-1}(\mathcal{B}(f, \varepsilon_N)) \right) \geq 1 - \delta_N.$$

In other words, for every $f \in \mathcal{F}$ the set of data points $(y_i, x_i)_{i=1}^N$ that are mapped by the procedure $\tilde{f}_N$ into the set $\{h \in \mathcal{F} : \mathbb{E}\mathcal{L}_h \leq \varepsilon_N\}$ is of $\nu_{f,\mathbb{X}} \otimes \mu^N$ measure at least $1 - \delta_N$.

The first estimate presented here is the high probability lower bound, formulated in Theorem A$'$.

Theorem 3.1 *There exists an absolute constant c_1 for which the following holds. If $\mathcal{F}$ is star-shaped around one of its points and $\tilde{f}_N$ is a procedure with a confidence parameter $\delta_N < 1/4$, then its accuracy satisfies*

$$\varepsilon_N \geq \min\left\{ c_1\sigma^2 \frac{\log(1/\delta_N)}{N}, \frac{1}{4} d_{\mathcal{F}}(L_2) \right\}.$$

Observe that if $\delta_N = \exp(-c_0\gamma N)$, then $\varepsilon_N \geq c_2\sigma^2\gamma$. Taking $\gamma = \eta(s_N^*(\eta))^2$ for $\eta \sim \sigma^{-1}$ proves the second part of Theorem A$'$: ERM achieves the minimax rate for the confidence established in Theorem A.

The proof of Theorem 3.1 requires several preliminary steps.

Let $\mathbb{X} = (x_i)_{i=1}^N \in \Omega^N$ and consider the conditional probability measure $\nu_{f,\mathbb{X}}$ defined above. Put $\mathcal{A}_f = \tilde{f}_N^{-1}(\mathcal{B}(f,\varepsilon_N))$ and let $\mathcal{A}_f|\mathbb{X} = \{y \in \mathbb{R}^N : (y,\mathbb{X}) \in \mathcal{A}_f\}$ denote the corresponding fiber of $\mathcal{A}_f$.

Lemma 3.2 *For every $f \in \mathcal{F}$,*

$$\mu^N\big(\{\mathbb{X} = (x_i)_{i=1}^N : \nu_{f,\mathbb{X}}(\mathcal{A}_f|\mathbb{X}) \geq 1 - \sqrt{\delta_N}\}\big) \geq 1 - \sqrt{\delta_N}.$$

Proof. Fix $f \in \mathcal{F}$ and let $\rho(\mathbb{X}) = \nu_{f,\mathbb{X}}(\mathcal{A}_f|\mathbb{X})$. Then,

$$1 - \delta_N \leq \nu_{f,\mathbb{X}} \otimes \mu^N(\mathcal{A}_f) = \mathbb{E}\rho(X_1,...,X_N).$$

Since $\|\rho\|_{L_\infty} \leq 1$ and $\mathbb{E}\rho(\mathbb{X}) \geq 1 - \delta_N$, by the Paley-Zygmund Theorem (see Chapter 3.3 in [12]), $\mathbb{P}(\rho(\mathbb{X}) \geq x) \geq (\mathbb{E}\rho(\mathbb{X}) - x)/(1-x) \geq 1 - \delta_N/(1-x)$ for every $0 < x < 1$. The claim follows by selecting $x = 1 - \sqrt{\delta_N}$. ∎

Observe that for every $f \in \mathcal{F}$ and $\mathbb{X} = (x_1,...,x_N)$, $\nu_{f,\mathbb{X}}$ is a gaussian measure on $\mathbb{R}^N$ with mean $P_{\mathbb{X}}f = (f(x_i))_{i=1}^N$ and covariance matrix $\sigma^2 I_N$.

Lemma 3.3 *Let $t \mapsto \Phi(t) = \mathbb{P}(g \leq t)$ be the cumulative distribution function of a standard gaussian random variable on $\mathbb{R}$. Let $u, v \in \mathbb{R}^N$ and consider the two gaussian measures $\nu_u \sim \mathcal{N}(u, \sigma^2 I_N)$ and $\nu_v \sim \mathcal{N}(v, \sigma^2 I_N)$. If $A \subset \mathbb{R}^N$ is measurable, then*

$$\nu_v(A) \geq 1 - \Phi\big(\Phi^{-1}(1 - \nu_u(A)) + \|u - v\|_{\ell_2^N}/\sigma\big).$$

The main component in the proof of Lemma 3.3 is a version of the gaussian shift theorem.

Theorem 3.4 *[23] Let ν be the standard gaussian measure on $\mathbb{R}^N$ and consider $B \subset \mathbb{R}^N$ and $w \in \mathbb{R}^N$. If $H_+ = \{x \in \mathbb{R}^N : \langle x, w\rangle \geq b\}$ is a halfspace satisfying that $\nu(H_+) = \nu(B)$, then $\nu(w + B) \geq \nu(w + H_+)$.*

Proof of Lemma 3.3. Let ν be the standard gaussian measure on $\mathbb{R}^N$. A straightforward change of variables shows that

$$\nu_u(A) = \nu\big((A-u)/\sigma\big) \text{ and } \nu_v(A) = \nu\big((A-v)/\sigma\big).$$

Let $B = (A-u)/\sigma$, $w = (u-v)/\sigma$ and set $\nu(B) = \alpha$. Using the notation of Theorem 3.4, the corresponding halfspace is

$$H_+ = \{x : \langle x, w/\|w\|_{\ell_2^N}\rangle \geq \Phi^{-1}(1-\alpha)\},$$

and therefore, if $w^\perp \subset \mathbb{R}^N$ is the subspace orthogonal to w,

$$w + H_+ = \{(\lambda+1)w + w^\perp : \ \lambda \geq \Phi^{-1}(1-\alpha)/\|w\|_{\ell_2^N}\}.$$

Clearly,

$$\nu(w + H_+) = \mathbb{P}\big(g \geq \Phi^{-1}(1-\alpha) + \|w\|_{\ell_2^N}\big),$$

and the claim follows from Theorem 3.4 and the definition of w. ∎

Proof of Theorem 3.1. Let $\tilde{f}_N$ be a procedure that performs with accuracy $\varepsilon_N \le d^2_{\mathcal{F}}(L_2)/4$ and a confidence parameter δ_N. Shifting $\mathcal{F}$ if needed, and since $\mathcal{F}$ is star-shaped around one of its points, one may assume that $u = 0 \in \mathcal{F}$ and consider $v \in \mathcal{F}$ for which $4\varepsilon_N \le \|v\|^2_{L_2(\mu)} \le 8\varepsilon_N$. By Chebyshev's inequality, $\mathbb{P}\big(\|P_{\mathbb{X}} v\|^2_{\ell_2^N} \ge 4N\|v\|^2_{L_2(\mu)}\big) \le 1/4$, and thus, for $\mathbb{X} = (X_i)_{i=1}^N$ in a set of μ^N-probability at least $3/4$, $\|P_{\mathbb{X}} v\|_{\ell_2^N} \le c_1\sqrt{N}\|v\|_{L_2(\mu)}$.

Let

$$\mathcal{A}_0 = \tilde{f}_N^{-1}(\mathcal{B}(0,\varepsilon_N)) \ \text{ and } \ \mathcal{A}_v = \tilde{f}_N^{-1}(\mathcal{B}(v,\varepsilon_N)),$$

which, by the choice of v, are disjoint. Since $\tilde{f}_N$ performs with accuracy ε_N and has a confidence parameter δ_N, $\nu_{0,\mathbb{X}} \otimes \mu^N(\mathcal{A}_0) \ge 1 - \delta_N$ and $\nu_{v,\mathbb{X}} \otimes \mu^N(\mathcal{A}_v) \ge 1 - \delta_N$. Applying Lemma 3.2, with μ^N-probability at least $1 - 2\sqrt{\delta_N}$,

$$\nu_{0,\mathbb{X}}(\mathcal{A}_0|\mathbb{X}) \ge 1 - \sqrt{\delta_N}, \ \text{ and } \ \nu_{v,\mathbb{X}}(\mathcal{A}_v|\mathbb{X}) \ge 1 - \sqrt{\delta_N}. \tag{3.1}$$

Let Ω_0 be the set of samples $\mathbb{X} = (X_i)_{i=1}^N \subset \Omega^N$ for which $\|P_{\mathbb{X}} v\|_{\ell_2^N} \le c_1\sqrt{N}\|v\|_{L_2(\mu)}$ and (3.1) holds. Hence, $\mathbb{P}(\Omega_0) \ge 3/4 - 2\sqrt{\delta_N}$, and by Lemma 3.3 applied to the set $\mathcal{A}_0|\mathbb{X}$,

$$\nu_{v,\mathbb{X}}(\mathcal{A}_0|\mathbb{X}) \ge 1 - \Phi\left(\Phi^{-1}(\sqrt{\delta_N}) + \|P_{\mathbb{X}} v\|_{\ell_2^N}/\sigma\right) = (*).$$

Observe that if $\delta_N < 1/4$ then $\Phi^{-1}(\sqrt{\delta_N}) < 0$ and $|\Phi^{-1}(\sqrt{\delta_N})| \sim \sqrt{\log(1/\delta_N)}$. Moreover, if $\|P_{\mathbb{X}} v\|_{\ell_2^N} \le \sigma|\Phi^{-1}(\sqrt{\delta_N})|$ then $(*) > 1/2$.

Since $\mathbb{X} \in \Omega_0$, $\|P_{\mathbb{X}} v\|_{\ell_2^N} \le c_1\sqrt{N}\|v\|_{L_2(\mu)}$; therefore, if

$$\|v\|_{L_2(\mu)} \lesssim \sigma\sqrt{\frac{\log(1/\delta_N)}{N}},$$

it follows that $\nu_{v,\mathbb{X}}(\mathcal{A}_0|\mathbb{X}) > 1/2$. On the other hand, $\mathcal{A}_0|\mathbb{X}$ and $\mathcal{A}_v|\mathbb{X}$ are disjoint and $\nu_{v,\mathbb{X}}(\mathcal{A}_v|\mathbb{X}) \ge 1 - \sqrt{\delta_N}$, which is impossible if $\delta_N < 1/4$.

Thus,

$$\|v\|_{L_2(\mu)} \gtrsim \sigma\sqrt{\frac{\log(1/\delta_N)}{N}},$$

and by the choice of v,

$$8\varepsilon_N \ge \|v\|^2_{L_2(\mu)} \gtrsim \sigma^2\frac{\log(1/\delta_N)}{N},$$

as claimed. ∎

Next, let us turn to the proof of Theorem B, which is a straightforward application of the next observation:

Theorem 3.5 *There exists an absolute constant c_0 for which the following holds. Let $\mathcal{F}$ and Y^f as above, and assume that $\tilde{f}_N$ is a procedure that performs with accuracy $\varepsilon_N = a_N^2$ and has a confidence parameter $\delta_N \leq 1/4$. Then, for any $\theta \geq 4$ and $f \in \mathcal{F}$, if Λ is a $2a_N$-separated subset of $\mathcal{F} \cap (f + \theta a_N D)$,*

$$\log |\Lambda| \leq c_0 N \left(\frac{\theta a_N}{\sigma} \right)^2 .$$

Proof. Observe that if $a_N \geq (1/2) d_{\mathcal{F}}(L_2)$ then $|\Lambda| = 1$ and Theorem 3.5 is trivially true. Hence, one may assume that $a_N < (1/2) d_{\mathcal{F}}(L_2)$.

Let $a = a_N$, set $D(f, r) = \{h \in \mathcal{F} : \|f - h\|_{L_2(\mu)} \leq r\}$ and put Λ to be a maximal $2a$-separated subset of $\mathcal{F} \cap (f + \theta a D)$ with respect to the $L_2(\mu)$ norm. Thus, $\{D(f, a) : f \in \Lambda\}$ is a family of disjoint subsets of $\mathcal{F} \cap (f + \theta a D)$.

Recall that for any $\mathbb{X} = (x_1, \ldots, x_N) \in \Omega^N$, $\mathcal{A}_f | \mathbb{X}$ is the fiber of $\mathcal{A}_f = \tilde{f}_N^{-1}(D(f, a))$. Since $\tilde{f}_N$ performs with accuracy a^2 and has a confidence parameter $\delta_N = 1 - \alpha$, it follows that for any $f \in \Lambda$,

$$\mathbb{E}_{\mathbb{X}} \nu_{f,\mathbb{X}}(\mathcal{A}_f | \mathbb{X}) = \nu_{f,\mathbb{X}} \otimes \mu^N(\mathcal{A}_f) \geq \alpha.$$

If $u \neq v$ in Λ and $A \subset \mathbb{R}^N$ then by Lemma 3.3,

$$\nu_{u,\mathbb{X}}(A) \geq 1 - \Phi\big(\Phi^{-1}(1 - \nu_{v,\mathbb{X}}(A)) + \|P_{\mathbb{X}} v - P_{\mathbb{X}} u\|_{\ell_2^N} / \sigma\big).$$

Fix $v_0 \in \Lambda$. Since $\{\mathcal{A}_v | \mathbb{X}, v \in \Lambda\}$ is a family of disjoint sets,

$$\begin{aligned} 1 &\geq \sum_{v \in \Lambda} \nu_{v_0,\mathbb{X}}(\mathcal{A}_v | \mathbb{X}) \\ &\geq \sum_{v \in \Lambda} \Big(1 - \Phi\big(\Phi^{-1}(1 - \nu_{v,\mathbb{X}}(\mathcal{A}_v | \mathbb{X})) + \|P_{\mathbb{X}} v_0 - P_{\mathbb{X}} v\|_{\ell_2^N} / \sigma\big) \Big) \\ &= \sum_{v \in \Lambda} \int_{z_{\mathbb{X}}(v)}^{\infty} \varphi(x) dx, \end{aligned}$$

where φ is a density function of a the standard gaussian $\mathcal{N}(0, 1)$ and

$$z_{\mathbb{X}}(v) = \Phi^{-1}(1 - \nu_{v,\mathbb{X}}(\mathcal{A}_v | \mathbb{X})) + \|P_{\mathbb{X}} v_0 - P_{\mathbb{X}} v\|_{\ell_2^N} / \sigma.$$

Taking the expectation with respect to $\mathbb{X}$,

$$1 \geq \sum_{v \in \Lambda} \mathbb{E}_{\mathbb{X}} \int_{z_{\mathbb{X}}(v)}^{\infty} \varphi(x) dx, \tag{3.2}$$

and it remains to lower bound each expectation.

Since

$$\mathbb{E}_{\mathbb{X}}\nu_{v,\mathbb{X}}\big((\mathcal{A}_v|\mathbb{X})^c\big) \le 1-\alpha \le 1/4,$$

it follows from Chebyshev's inequality that $\mathbb{P}_{\mathbb{X}}\big(\nu_{v,\mathbb{X}}(\mathcal{A}_v|\mathbb{X}) \ge 3/4\big) \le 1/3$. Therefore, with μ^N-probability at least $2/3$,

$$\Phi^{-1}\big(1-\nu_{v,\mathbb{X}}(\mathcal{A}_v|\mathbb{X})\big) = \Phi^{-1}\big(\nu_{v,\mathbb{X}}\big((\mathcal{A}_v|\mathbb{X})^c\big)\big) \le \Phi^{-1}(3/4) := \beta.$$

Another application of Chebyshev's inequality shows that with μ^N-probability at least $2/3$,

$$\|P_{\mathbb{X}}v_0 - P_{\mathbb{X}}v\|_{\ell_2^N} \le (3/2)\sqrt{N}\|v_0-v\|_{L_2(\mu)} \le (3/2)\theta a\sqrt{N},$$

because $v \in D(v_0,\theta a)$. Therefore, with μ^N-probability at least $1/3$,

$$z_{\mathbb{X}}(v) \le \beta + (3/2)\sqrt{N}\theta a/\sigma$$

and since $\beta + (3/2)\sqrt{N}\theta a/\sigma > 0$,

$$\mathbb{E}_{\mathbb{X}}\int_{z_{\mathbb{X}}(v)}^{\infty}\varphi(x)dx \ge \frac{1}{3}\int_{\beta+(3/2)\sqrt{N}\theta a_N/\sigma}^{\infty}\varphi(x)dx \gtrsim \exp\Big(-\frac{c_2N\theta^2a^2}{\sigma^2}\Big).$$

Thus, by (3.2), $1 \gtrsim |\Lambda|\exp\big(-c_3N\theta^2a^2/\sigma^2\big)$, as claimed. ∎

We end this section with the proof of Theorem C, which is presented for a random design, though the proof for a deterministic design is almost identical. The idea is that if $\mathbb{X} = (X_1,...,X_N)$ and $P_{\mathbb{X}}f_1 = P_{\mathbb{X}}f_2$, the two functions are indistinguishable on a sample $(X_i,Y_i)_{i=1}^N$ of $Y^{f_1} = f_1(X)+V$. Therefore, no procedure can perform with a better accuracy than the largest typical $L_2(\mu)$ diameter of the sets

$$K(f,\mathbb{X}) = \{h\in\mathcal{F} : P_{\mathbb{X}}h = P_{\mathbb{X}}f\}.$$

Fix $f\in\mathcal{F}$ and for every sample $\mathbb{X}$ let $\mathcal{D}(f,\mathbb{X})$ be the $L_2(\mu)$-diameter of $K(f,\mathbb{X})$. Define an $\mathcal{F}$-valued random variable h^f as follows. Let $h^f_{1,\mathbb{X}}$ and $h^f_{2,\mathbb{X}}$ be almost $L_2(\mu)$-diametric points in $K(f,\mathbb{X})$, set δ to be a $\{0,1\}$-valued random variable with mean $1/2$, which is independent of X and V, and put

$$h^f = (1-\delta)h^f_{1,\mathbb{X}} + \delta h^f_{2,\mathbb{X}}. \tag{3.3}$$

Note that for every realization of δ, $h^f \in K(f,\mathbb{X})$ and $\mathcal{D}(h^f,\mathbb{X}) = \mathcal{D}(f,\mathbb{X})$. Denote by $\mathbb{P}_{X,V}$ (resp. $\mathbb{E}_{X,V}$) the probability distribution of (resp.

expectation w.r.t.) $(X_i, V_i)_{i=1}^N$. Let $I(A)$ be the indicator of the set A and observe that for every realization of the random variable δ,

$$\begin{aligned}
&\sup_{f\in\mathcal{F}} \mathbb{P}_{X,V}\left(\|\tilde{f}_N\left((X_i, f(X_i)+V_i)_{i=1}^N\right) - f\|_{L_2(\mu)} \geq \mathcal{D}(f,\mathbb{X})/4\right)\\
\geq &\sup_{f\in\mathcal{F}} \mathbb{P}_{X,V}\left(\|\tilde{f}_N\left((X_i, h^f(X_i)+V_i)_{i=1}^N\right) - h^f\|_{L_2(\mu)} \geq \mathcal{D}(h^f,\mathbb{X})/4\right)\\
= &\sup_{f\in\mathcal{F}} \mathbb{P}_{X,V}\left(\|\tilde{f}_N\left((X_i, h^f(X_i)+V_i)_{i=1}^N\right) - h^f\|_{L_2(\mu)} \geq \mathcal{D}(f,\mathbb{X})/4\right) = (*)
\end{aligned}$$

because $h^f \in \mathcal{F}$ and $\mathcal{D}(f,\mathbb{X}) = \mathcal{D}(h^f,\mathcal{X})$.

For every $f \in \mathcal{F}$ put

$$A_1^f = \left\{\|\tilde{f}_N\left((X_i, h^f_{1,\mathbb{X}}(X_i)+V_i)_{i=1}^N\right) - h^f_{1,\mathbb{X}}\|_{L_2(\mu)} \geq \mathcal{D}(f,\mathbb{X})/4\right\},$$

and

$$A_2^f = \left\{\|\tilde{f}_N\left((X_i, h_{2,\mathbb{X}}(X_i)+V_i)_{i=1}^N\right) - h_{2,\mathbb{X}}\|_{L_2(\mu)} \geq \mathcal{D}(f,\mathbb{X})/4\right\}.$$

Taking the expectation in $(*)$ with respect to δ,

$$\begin{aligned}
\mathbb{E}_\delta(*) \geq &\sup_{f\in\mathcal{F}} \mathbb{E}_{X,V}\mathbb{E}_\delta I\left(\tilde{f}_N\left((X_i, h^f(X_i)+V_i)_{i=1}^N\right) - h^f\|_{L_2(\mu)} \geq \mathcal{D}(f,\mathbb{X})/4\right)\\
= &\sup_{f\in\mathcal{F}} \mathbb{E}_{X,V}\frac{1}{2}(I(A_1^f)+I(A_2^f)).
\end{aligned}$$

Note that for any sample $\mathbb{X}$, $h^f_{1,\mathbb{X}}(X_i)+V_i = h^f_{2,\mathbb{X}}(X_i)+V_i$; therefore,

$$\tilde{f}_N\left((X_i, h^f_{1,\mathbb{X}}(X_i)+V_i)_{i=1}^N\right) = \tilde{f}_N\left((X_i, h^f_{2,\mathbb{X}}(X_i)+V_i)_{i=1}^N\right) \equiv f_0.$$

Since $h^f_{1,\mathbb{X}}$ and $h^f_{2,\mathbb{X}}$ are almost diametric in $K(f,\mathbb{X})$, either $\|h^f_{1,\mathbb{X}} - f_0\|_{L_2(\mu)} \geq \mathcal{D}(f,\mathbb{X})/4$ or $\|h^f_{2,\mathbb{X}} - f_0\|_{L_2(\mu)} \geq \mathcal{D}(f,\mathbb{X})/4$. Thus, $I(A_1^f)+I(A_2^f) \geq 1$ almost surely, and

$$\sup_{f\in\mathcal{F}} \mathbb{P}_{X,V}\left(\|\tilde{f}_N\left((X_i, f(X_i)+V_i,)_{i=1}^N\right) - f\|_{L_2(\mu)} \geq \mathcal{D}(f,\mathbb{X})/4\right) \geq 1/2.$$

■

Remark. It is straightforward to verify that if $\sigma = 0$, then ERM satisfies $\hat{f} \in K(f^*,\mathbb{X})$ for every sample $\mathbb{X}$. Therefore, a typical value of $\mathcal{D}(f^*,\mathbb{X})$ is a lower bound on the minimax rate in the noise-free case.

As an example, let $T \subset \mathbb{R}^d$ be a convex, centrally-symmetric set, put μ to be an isotropic, L-subgaussian measure on $\mathbb{R}^d$ and set $\mathcal{F}$ to be the class of linear functionals indexed by T. Given a sample $\mathbb{X} = (X_1, ..., X_N)$, set $\Gamma_{\mathbb{X}} = \sum_{i=1}^N \langle X_i, \cdot \rangle e_i$ and put $P_{\mathbb{X}} t = \Gamma_{\mathbb{X}} t$. Therefore,

$$K(v_0, \mathbb{X}) = \{v \in T : \Gamma_{\mathbb{X}} v = \Gamma_{\mathbb{X}} v_0\} \subset 2T \cap \ker(\Gamma_{\mathbb{X}}).$$

Let $d_N = d_N(\rho)$ satisfy that with probability at least $1-\rho$, $\mathcal{D}(0, \mathbb{X}) \geq d_N$. Then, by Theorem C, any procedure with a confidence parameter $\delta_N \leq 1/2 + \rho$ cannot perform with a better accuracy parameter than $d_N(\rho)/4$.

On the other hand, a straightforward application of Lemma 2.5 shows that with probability at least $1 - 2\exp(-c_1 N Q^2)$, $\mathcal{D}(0, \mathbb{X}) \lesssim r_N^*(Q)$. Therefore, if $d_N(T) \sim r_N^*(Q)$ for a suitable absolute constant Q, then with probability at least $1 - 2\exp(-c_1 Q^2 N)$,

$$r_N^*(Q) \lesssim d_N(T) \leq \mathcal{D}(0, \mathbb{X}) \leq r_N^*(Q),$$

and if $\sigma \lesssim r_N^*(Q)$, the error rate obtained in Theorem A is the minimax rate in the constant probability range.

4 Examples

Here, we will present two examples in which our results lead to sharp upper and lower bounds. Although there are many other examples that follow the same path, and for which the estimates of Theorem A are sharp, we will not present them here for the sake of brevity.

4.1 Learning in B_1^d

Let $\mathcal{F}$ be a class of linear functionals, indexed by $T = B_1^d$, the unit ball in ℓ_1^d. Assume that μ is an isotropic, L-subgaussian measure on $\mathbb{R}^d$, that $Y \in L_{\psi_2}$ and that $\|Y - f^*\|_{\psi_2} \leq \sigma$. For instance, if W is a centred random variable that is independent of X and $Y = f(X) + W$ for $f \in \mathcal{F}$, then $W = Y = f^*(X)$.

The upper bound of Theorem A is based on estimates on $\mathbb{E}\|G\|_{2\mathcal{F} \cap sD}$. Because the measure μ is isotropic, the canonical gaussian process is given by $t \to \sum_{i=1}^d g_i t_i$, and $2\mathcal{F} \cap sD = 2B_1^d \cap sB_2^d$.

One may show (see, for example, [17]) that for every $1/\sqrt{d} \leq s \leq 2$,

$$\mathbb{E} \sup_{t \in 2B_1^d \cap sB_2^d} \left| \sum_{i=1}^d g_i t_i \right| \sim \sqrt{\log(eds^2)},$$

while if $s \leq 1/\sqrt{d}$, $sB_2^d \subset 2B_1^d \cap sB_2^d \subset 2sB_2^d$ and

$$\mathbb{E} \sup_{t \in 2B_1^d \cap sB_2^d} \Big| \sum_{i=1}^d g_i t_i \Big| \sim s\sqrt{d}.$$

Setting $\eta = c_0/\sigma$, it is straightforward to verify that

$$(s_N^*(\eta))^2 \sim \begin{cases} \sigma \sqrt{\frac{\log(ed^2\sigma^2/N)}{N}} & \text{if } N \leq \sigma^2 d^2, \\ \sigma^2 d/N & \text{otherwise.} \end{cases}$$

Also,

$$(r_N^*(Q))^2 \quad \begin{cases} \sim_Q \frac{1}{N} \log\left(\frac{ed}{N}\right) & \text{if } N \leq c_1 d, \\ \lesssim_Q \frac{1}{d} & \text{if } c_1 d \leq N \leq c_2 d \\ = 0 & \text{if } N > c_2 d, \end{cases}$$

where c_1 and c_2 are constants that depend only on Q, and one must keep in mind that N has to be large enough to ensure that $s_N^*(\eta), r_N^*(Q) \leq d_{\mathcal{F}}(L_2) = 1$.

When $N \sim d$, r_N^* decays rapidly from $N^{-1/2}\log^{1/2}(ed/N)$ to 0. Thus, when $c_1 d \leq N \leq c_2 d$ one only has an upper estimate on r_N^*, and we will only consider the cases $N \leq c_1 d$ and $N \geq c_2 d$.

Fix Q to be a constant depending on L, set $\eta = c_0/\sigma$ and let $N \leq c_1 d$. If $\sigma \geq r_N^*$ then $\sigma^2 d^2 \gtrsim N$, and

$$(s_N^*(c_0/\sigma))^2 \sim \sigma \sqrt{\frac{\log(ed^2\sigma^2/N)}{N}}.$$

Applying Theorem A, if $\sigma \geq c_3\sqrt{\log(ed/N)/N}$, then with probability at least $1 - 2\exp(-c_4\sigma^{-1}\log(ed^2\sigma^2/N))$,

$$R(\hat{f}) \leq \inf_{f \in \mathcal{F}} R(f) + c_5 \sigma \sqrt{\frac{\log(ed^2\sigma^2/N)}{N}}.$$

And, if $\sigma \leq c_3\sqrt{\log(ed/N)/N}$, then with probability at least $1 - 2\exp(-c_4 N)$,

$$R(\hat{f}) \leq \inf_{f \in \mathcal{F}} R(f) + \frac{c_5}{N} \log\left(\frac{ed}{N}\right),$$

for constants c_3, c_4, c_5 that depend on L and the choice of Q.

In a similar fashion, if $N \geq c_2 d$ then $r_N^* = 0$, and thus, if $\sigma \neq 0$, $\sigma \geq r_N^*$. Therefore, the error rate of ERM is determined solely by s_N^*.

Turning to the lower estimate, and as noted in Theorem A′, if $\tilde{f}_N$ performs with accuracy ε_N and achieves the same confidence obtained in Theorem A, then in the noisy case ($\sigma \gtrsim r_N^*$)

$$\varepsilon_N \gtrsim \sigma^2 \frac{\log(1/\delta_N)}{N} = (s_N^*(c/\sigma))^2.$$

Thus, ERM achieves the minimax rate in that regime.

For a lower bound that holds with constant probability we shall apply Theorem B. To that end, let us bound the covering numbers $\log N(B_1^d \cap rB_2^d, \theta r B_2^d)$ from below for some $\theta < 1$.

Fix $1/\sqrt{d} \le r \le 1$, and without loss of generality assume that $k = 1/r^2$ is an integer. For $I \subset \{1, ..., d\}$, let S^I be the Euclidean sphere supported on the coordinates I, and note that

$$\bigcup_{|I|=k} rS^I \subset B_1^d \cap rB_2^d.$$

It is a well known fact (see, e.g., [25]) that there is a collection of subsets of $\{1, ..., d\}$ of cardinality k, which will be denoted by $\mathcal{B}$, that is $c_0 k$ separated in the Hamming distance and for which $\log |\mathcal{B}| \ge c_1 k \log(ed/k)$. Thus, the set $\Lambda = \{r \sum_{i \in I} e_i : I \in \mathcal{B}\}$ is a $c_2 r$-separated subset of $B_1^d \cap rB_2^d$ with respect to the ℓ_2^d norm, and

$$\log N(B_1^d \cap rB_2^d, c_3 r B_2^d) \ge c_4 \frac{\log(edr^2)}{r^2}.$$

By Theorem B, given a procedure with a confidence parameter $\delta_N \le 1/4$, its accuracy $\varepsilon_N = r^2 \ge 1/d$ satisfies

$$\frac{\log(edr^2)}{r^2} \lesssim \log N(B_1^d \cap rB_2^d, c_3 r B_2^d) \lesssim \frac{Nr^2}{\sigma^2}.$$

Therefore,

$$\varepsilon_N \gtrsim \sigma \sqrt{\frac{\log(ed^2\sigma^2/N)}{N}},$$

provided that $d^2\sigma^2 \ge N$ (otherwise, $r \le 1/\sqrt{d}$).

If $r \le 1/\sqrt{d}$, $\log N(B_1^d \cap rB_2^d, c_5 r B_2^d) \gtrsim 2^d$, and thus

$$\varepsilon_N = r^2 \gtrsim \sigma^2 \frac{d}{N}$$

when $d^2\sigma^2 \leq N$. Therefore, $\tilde{f}_N$ cannot outperform ERM in the noisy case, even if it is allowed to succeed with only constant probability.

Finally, turning to the trivial noise level, one has to show that the estimate of r_N^* is sharp. Recall that by Theorem C it suffices to show that the Gelfand N-width of B_1^d satisfies $c_N(B_1^d) \sim r_N^*$. By a result due to Garanaev and Gluskin [16],

$$c_N(B_1^d) \sim \min\left\{1, \sqrt{\frac{\log(ed/N)}{N}}\right\} \sim r_N^*.$$

Thus, for $0 \leq \sigma \lesssim r_N^*(Q)$, $\tilde{f}_N$ does not outperform ERM, proving ERM's optimality in both regimes.

4.2 Low-rank matrix inference via the max-norm

In this type of problem, the goal is to estimate Y by a linear function of a low-rank (or approximately low rank) matrix. Since the rank is not a convex constraint, one may consider the convex relaxation given by the factorization-based norm

$$\|A\|_{max} = \min_{A=UV^\top} \|U\|_{2\to\infty} \|V\|_{2\to\infty}.$$

Let $\mathcal{B}_{max}$ be the unit ball relative to that norm and set $\mathcal{F} = \{f_A = \langle \cdot, A\rangle : A \in \mathcal{B}_{max}\}$. Thus,

$$\hat{A}_N \in \operatorname*{argmin}_{\|A\|_{max}\leq 1} \frac{1}{N}\sum_{i=1}^N \big(Y_i - \langle X_i, A\rangle\big)^2.$$

A similar estimator has been studied in [30] for $Y = \langle A^*, X\rangle + W$, a random vector X that is selected uniformly from the canonical basis of $\mathbb{R}^{p\times q}$, a noise vector W that is either gaussian or sub-exponential with independent coordinates, and matrices in $\mathcal{B}_{max}$ with uniformly bounded entries.

Assume that X is isotopic and L-subgaussian relative to the normalized Frobenius norm

$$\big\|\langle X, A\rangle\big\|_{L_2} = (pq)^{-1/2}\|A\|_F, \quad \big\|\langle X, A\rangle\big\|_{\psi_2} \leq L(pq)^{-1/2}\|A\|_F,$$

let A^* be the minimizer of the loss in $\mathcal{B}_{max}$ and set $\sigma = \big\|Y - \langle X, A^*\rangle\big\|_{\psi_2}$. Since $\mathcal{F}$ is convex, the minimizer is unique and the conditions of Theorem A are satisfied.

To apply Theorem A, one has to estimate the fixed points $r_N^*(Q)$ and $s_N^*(\eta)$ for Q that depends only on L and $\eta \sim_L \sigma^{-1}$.

Let B_F be the unit ball relative to the Frobenius norm. Since X is isotropic, the relative L_2 unit ball is

$$D = \{f_A : \mathbb{E}|\langle X, A\rangle|^2 \le 1\} = \{\langle \cdot, A\rangle : A \in \sqrt{pq}B_F\},$$

and the corresponding gaussian process has a covariance structure given by

$$\mathbb{E}G_{f_A}G_{f_B} = (pq)^{-1}\langle A, B\rangle = (pq)^{-1}\mathrm{Tr}(A^\top B).$$

A simple application of Grothendieck's inequality (see, e.g., [28]) shows that

$$\mathrm{conv}(\mathcal{X}_\pm) \subset \mathcal{B}_{max} \subset K_G\mathrm{conv}(\mathcal{X}_\pm)$$

where K_G is the Grothendieck constant and $\mathcal{X}_\pm = \{uv^\top : u \in \{\pm 1\}^p, v \in \{\pm 1\}^q\}$; in particular, $\mathrm{diam}(\mathcal{B}_{max}, L_2) \sim 1$.

Let $\mathfrak{G} = (g_{ij})_{1\le i\le p:1\le j\le q}$ be a matrix with independent, centered gaussian entries with variance $(pq)^{-1}$. Thus, for every $s > 0$,

$$\begin{aligned}\mathbb{E}\,\|G\|_{(\mathcal{F}-\mathcal{F})\cap sD} &= \mathbb{E}\sup_{A\in 2\mathcal{B}_{max}\cap s\sqrt{pq}B_F}|\langle \mathfrak{G}, A\rangle| \le 2\mathbb{E}\sup_{A\in\mathcal{B}_{max}}|\langle \mathfrak{G}, A\rangle| \\ &\le 2K_G\mathbb{E}\sup_{A\in\mathrm{conv}(\mathcal{X}_\pm)}|\langle \mathfrak{G}, A\rangle|.\end{aligned}$$

By standard properties of gaussian processes,

$$\mathbb{E}\sup_{A\in\mathrm{conv}(\mathcal{X}_\pm)}|\langle \mathfrak{G}, A\rangle| \lesssim \max_{A\in\mathcal{X}_\pm}\frac{\|A\|_F}{\sqrt{pq}}\sqrt{\log|\mathcal{X}_\pm|} \lesssim \sqrt{p+q}.$$

In the reverse direction, by Lemma 3.1 in [30], if

$$\frac{1}{\min(p,q)} \lesssim s^2 \lesssim 1,$$

then

$$s\log^{1/2} N(\mathcal{B}_{max}\cap s\sqrt{pq}B_F, s\sqrt{pq/2}B_F) \gtrsim \sqrt{p+q}. \tag{4.1}$$

Hence, in that range of s,

$$\mathbb{E}\,\|G\|_{sD\cap(\mathcal{F}-\mathcal{F})} \sim \sqrt{p+q},$$

and

$$(s_N^*(c/\sigma))^2 \sim \sigma\sqrt{\frac{p+q}{N}}, \quad (r_N^*(Q))^2 \sim \frac{p+q}{N}$$

as long as both are smaller than 1.

Applying Theorem A, if $\sigma \gtrsim_{Q,L} \sqrt{(p+q)/N}$ then with probability at least $1-2\exp(-c_1\sqrt{N(p+q)}/\sigma)$, ERM satisfies that

$$\mathbb{E}(Y-\langle \hat{A}, X\rangle)^2 \leq \inf_{A\in\mathcal{B}_{max}} \mathbb{E}(Y-\langle A, X\rangle)^2 + c_2(Q,L)\sigma\sqrt{\frac{p+q}{N}},$$

and if $\sigma \lesssim_{Q,L} \sqrt{(p+q)/N}$, then with probability at least $1-2\exp(-c_1 N)$,

$$\mathbb{E}(Y-\langle \hat{A}, X\rangle)^2 \leq \inf_{A\in\mathcal{B}_{max}} \mathbb{E}(Y-\langle A, X\rangle)^2 + c_2(Q,L)\frac{p+q}{N}.$$

To see that this estimate is sharp in the minimax sense when $\sigma \gtrsim \sqrt{(p+q)/N}$ (and as long as $s_N^*, r_N^* \lesssim 1$, i.e., $\sigma \lesssim \sqrt{N/(p+q)}$), observe that Theorem A$'$ implies that ERM achieves the minimax rate for the confidence parameter $\delta_N \lesssim \exp(-c_1\sqrt{N(p+q)}/\sigma)$. Moreover, by Theorem B and (4.1), any procedure with confidence parameter $\delta_N \leq 1/4$ has accuracy $\varepsilon_N \gtrsim \sigma\sqrt{\frac{p+q}{N}}$, matching the upper bound in the noisy regime.

5 Concluding remarks and comparisons with existing results

The reason this note has featured subgaussian classes is that conceptually, such classes are the natural extension of bounded classes: the behaviour of both may be analyzed using two-sided concentration arguments. Unfortunately, this is as far as concentration goes: the substantial technical machinery needed for the proof Theorem A is not true beyond the subgaussian framework, and the analysis of more 'heavy-tailed' problems requires a totally different machinery.

The results presented in this article are sharp in many cases but not in every case. First, in the 'high probability' range, Theorem A and Theorem A$'$ show that when $\sigma \gtrsim r_N^*$ the result is sharp in the minimax sense. However, if $\sigma \lesssim r_N^*$, the estimate is known to be sharp only for $\sigma = 0$, when the error rate is a typical value of $\mathcal{D}^2(f^*, \mathbb{X})$, or for $\sigma \sim r_N^*$, when the error rate is $\sim (r_N^*)^2$. A sharp estimate for $\sigma \in (0, r_N^*)$ is not known, although there are many examples in which r_N^* is equivalent to the 'width' of the class, and then ERM is optimal in the minimax sense in that range as well.

In the constant probability regime, the picture is even less complete. In the noisy case, when $\sigma \gtrsim r_N^*$, the upper bound of $(s_N^*(c/\sigma))^2$ is sharp only

if it happens to be equivalent to $q_N^*(c/\sigma)$. Unfortunately, this is not even true for $\mathcal{F} = \{\langle t, \cdot \rangle : t \in B_p^d\}$ for $1 + 1/\log d < p < 2$.

In the low-noise case (i.e. $\sigma \lesssim r_N^*$), the situation is as described above, with an unknown range – when $\sigma \lesssim r_N^*$.

As an example, our results show that for gaussian noise, ERM achieves the minimax rate $\max\{(s_N^*(c/\sigma))^2, (r_N^*(Q))^2\}$ in the constant probability regime for both ranges of noise, when considering a class of linear functionals, indexed by a convex, centrally-symmetric set T, for an isotropic, L-subgaussian measure μ on $\mathbb{R}^n$, and when

1. $q_N^* \log^{1/2} N(T \cap 2q_N^* B_2^d, q_N^* B_2^d) \sim \mathbb{E}\,\|G\|_{T \cap q_N^* B_2^d}$ – meaning that there is no gap in Sudakov's inequality at scale $q_N^* = q_N^*(c/\sigma)$;
2. $c_N(T) \sim r_N^*(T)$ – meaning that $\sqrt{N} c_N(T \cap r_N^* B_2^d) \sim \mathbb{E}\,\|G\|_{T \cap r_N^* B_2^d}$, and there is no gap in the Pajor-Tomczak-Jaegermann estimate on the Gelfand N-width of T (see [29]).

The parameter s_N^* may be compared with the fixed points used in [33, 6, 34, 35, 2]. In all those cases, the fixed points are associated with Dudley's entropy integral for the localized class, rather than with the smaller localized gaussian process. For example, the results in [6] show that if the noise level is large enough and there is no gap in both Sudakov's AND Dudley's inequalities at the correct level (given by the fixed point), ERM is a minimax procedure in expectation. Theorem A clearly improves that result.

Finally, although the importance of convexity may have been obscured by the Bernstein condition, a uniform Bernstein condition implies that the class is convex, at least if a nontrivial error rate is to be expected.

Indeed, observe that if $\mathcal{F} \subset L_2(\mu)$ is closed but not locally compact in $L_2(\mu)$ then the minimax rate of $Y = f(X) + W$ does not tend to 0 as the sample size tends to infinity. This is an immediate outcome of Theorem B and the fact that there is some $r > 0$ and $f \in \mathcal{F}$ for which $f + rD$ contains an infinite set that is $r/4$ separated in $L_2(\mu)$. Thus, one may restrict oneself to classes that are locally compact, and, in which case, one has the following:

Theorem 5.1 *Let μ be a probability measure and let X be distributed according to μ. If $\mathcal{F}$ is a locally compact subset of $L_2(\mu)$, the following are equivalent:*

i) for any real valued random variable $Y \in L_2$, the minimum of the functional $f \to \mathbb{E}(Y - f(X))^2$ in $\mathcal{F}$ is attained. And, if f^* is such a

minimizer, then for every $f \in \mathcal{F}$,

$$\mathbb{E}\big(f(X) - f^*(X)\big)^2 \le \mathbb{E}\big((Y - f(X))^2 - (Y - f^*(X))^2\big). \qquad (5.1)$$

ii) $\mathcal{F}$ is nonempty and convex.

Proof. If $\mathcal{F}$ is a nonempty, closed and convex subset of a Hilbert space, the metric projection $Y \to f^*$ exists and is unique. By its characterization, $\big\langle f(X) - f^*(X), Y - f^*(X)\big\rangle \le 0$ for every $f \in \mathcal{F}$, and

$$\begin{aligned}
&\mathbb{E}\big((Y - f(X))^2 - (Y - f^*(X))^2\big) \\
&\quad = \|f(X) - f^*(X)\|_{L_2}^2 + 2\big\langle f^*(X) - Y, f(X) - f^*(X)\big\rangle \\
&\quad \ge \|f(X) - f^*(X)\|_{L_2}^2 .
\end{aligned}$$

In the reverse direction, if $\mathcal{F}$ is locally compact, the set-value metric projection onto $\mathcal{F}$ exists, and since it is 1-Bernstein for any Y, the metric projection is unique. Indeed, if $f_1^*, f_2^* \in \mathcal{F}$ are minimizers then by the Bernstein condition,

$$\|f_1^*(X) - f_2^*(X)\|_{L_2}^2 \le B\mathbb{E}\big((Y - f_2^*(X))^2 - (Y - f_1^*(X))^2\big) = 0.$$

Thus, any $Y \in L_2$ has a unique best approximation in $\mathcal{F}$, making $\mathcal{F}$ a locally compact Chebyshev set in a Hilbert space. By a result due to Vlasov [37], (see also [13], Chapter 12), $\mathcal{F}$ is convex. ∎

References

[1] Franck Barthe, Olivier Guédon, Shahar Mendelson, and Assaf Naor. A probabilistic approach to the geometry of the l_p^n-ball. *Ann. Probab.*, 33(2):480–513, 2005.

[2] Peter L. Bartlett, Olivier Bousquet, and Shahar Mendelson. Local Rademacher complexities. *Ann. Statist.*, 33(4):1497–1537, 2005.

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

[4] Peter L. Bartlett, Shahar Mendelson, and Joseph Neeman. ℓ_1-regularized linear regression: persistence and oracle inequalities. *Probab. Theory Related Fields*, 154(1-2):193–224, 2012.

[5] Lucien Birgé. Nonasymptotic minimax risk for Hellinger balls. *Probab. Math. Statist.*, 5(1):21–29, 1985.

[6] Lucien Birgé and Pascal Massart. Rates of convergence for minimum contrast estimators. *Probab. Theory Related Fields*, 97(1-2):113–150, 1993.

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

[8] Emmanuel J. Candes and Terence Tao. The Dantzig selector: statistical estimation when p is much larger than n. *Ann. Statist.*, 35(6):2313–2351, 2007.

[9] Emmanuel J. Candes and Terence Tao. Reflections on compressed sensing. *IEEE Information Theory Society Newsletter*, 58(4):14–17, 2008.

[10] Emmanuel J. Candes and Terence Tao. The power of convex relaxation: near-optimal matrix completion. *IEEE Trans. Inform. Theory*, 56(5):2053–2080, 2010.

[11] Albert Cohen, Wolfgang Dahmen, and Ronald DeVore. Compressed sensing and best k-term approximation. *J. Amer. Math. Soc.*, 22(1):211–231, 2009.

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

[13] Frank Deutsch. *Best approximation in inner product spaces*, volume 7 of *CMS Books in Mathematics*. Springer-Verlag, 2001.

[14] David L. Donoho. Compressed sensing. *IEEE Trans. Inform. Theory*, 52(4):1289–1306, 2006.

[15] Richard M. Dudley. *Uniform central limit theorems*, volume 63 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 1999.

[16] A. Yu. Garnaev and E. D. Gluskin. The widths of a Euclidean ball. *Dokl. Akad. Nauk SSSR*, 277(5):1048–1052, 1984.

[17] 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.

[18] D. Gross. Recovering low-rank matrices from few coefficients in any basis. *Information Theory, IEEE Transactions on*, 57(3):1548–1566, 2011.

[19] Norman E. Hurt. *Phase retrieval and zero crossings*, volume 52 of *Mathematics and its Applications*. Kluwer Academic Publishers Group, Dordrecht, 1989. Mathematical methods in image reconstruction.

[20] 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].

[21] Guillaume Lecué. Interplay between concentration, complexity and geometry in learning theory with applications to high dimensional data analysis. Habilitation à diriger des recherches. 2011.

[22] Michel Ledoux and Michel Talagrand. *Probability in Banach spaces*, volume 23 of *Ergebnisse der Mathematik und ihrer Grenzgebiete (3) [Results in Mathematics and Related Areas (3)]*. Springer-Verlag, Berlin, 1991. Isoperimetry and processes.

[23] Wenbo V. Li and James Kuelbs. Some shift inequalities for Gaussian measures. In *High dimensional probability (Oberwolfach, 1996)*, volume 43 of *Progr. Probab.*, pages 233–243. Birkhäuser, Basel, 1998.

[24] Pascal Massart. *Concentration inequalities and model selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, Berlin, 2007. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard.

[25] S. Mendelson, A. Pajor, and M. Rudelson. The geometry of random $\{-1, 1\}$-polytopes. *Discrete Comput. Geom.*, 34(3):365–379, 2005.

[26] Shahar Mendelson. Obtaining fast error rates in nonconvex situations. *J. Complexity*, 24(3):380–397, 2008.

[27] Shahar Mendelson and Joseph Neeman. Regularization in kernel learning. *Ann. Statist.*, 38(1):526–565, 2010.

[28] Srebro Nathan and Shraibman Adi. Rank, trace-norm and max-norm. *18th Annual Conference on Learning Theory (COLT)*, 2005.

[29] Alain Pajor and Nicole Tomczak-Jaegermann. Subspaces of small codimension of finite-dimensional Banach spaces. *Proc. Amer. Math. Soc.*, 97(4):637–642, 1986.

[30] Cai Toni and Zhou Wenxin. Matrix completion via max-norm constrained optimization. Technical report, Wharton University, 2013.

[31] Alexandre Tsybakov. Optimal rate of aggregation. In *Computational Learning Theory and Kernel Machines (COLT-2003)*, volume 2777 of *Lecture Notes in Artificial Intelligence*, pages 303–313. Springer, Heidelberg, 2003.

[32] Alexandre B. Tsybakov. *Introduction to nonparametric estimation*. Springer Series in Statistics. Springer, New York, 2009. Revised and extended from the 2004 French original, Translated by Vladimir Zaiats.

[33] Sara van de Geer. Estimating a regression function. *Ann. Statist.*, 18(2):907–924, 1990.

[34] Sara van de Geer. Hellinger-consistency of certain nonparametric maximum likelihood estimators. *Ann. Statist.*, 21(1):14–44, 1993.

[35] Sara A. van de Geer. *Applications of empirical process theory*, volume 6 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 2000.

[36] Vladimir N. Vapnik. *Statistical learning theory*. Adaptive and Learning Systems for Signal Processing, Communications, and Control. John Wiley & Sons Inc., New York, 1998. , A Wiley-Interscience Publication.

[37] P.L. Vlasov. Čebyšev sets in banach spaces. *Sov. Math. Dokl.*, 2:1373–1374, 1961.

[38] Yuhong Yang and Andrew Barron. Information-theoretic determination of minimax rates of convergence. *Ann. Statist.*, 27(5):1564–1599, 1999.

[39] Bin Yu. Assouad, Fano, and Le Cam. In *Festschrift for Lucien Le Cam*, pages 423–435. Springer, New York, 1997.