Learning subgaussian classes : Upper and minimax bounds

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

May 22, 2013

Abstract

We obtain sharp oracle inequalities for the empirical risk minimization procedure in the regression model under the assumption that the target Y and the model $\mathcal{F}$ are subgaussian. The bound we obtain is sharp in the minimax sense if $\mathcal{F}$ is convex. Moreover, under mild assumptions on $\mathcal{F}$, the error rate of ERM remains optimal even if the procedure is allowed to perform with constant probability. A part of our analysis is a new proof of minimax results for the gaussian regression model.

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 stand point, each X_i can be viewed as an input associated with an output Y_i. For 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 $\ell(f(X), Y)$. 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] Partially supported by the Mathematical Sciences Institute – The Australian National University and the European Research Council (under ERC grant agreement n^o [203134]).

for any measurable function $f : \mathcal{X} \mapsto \mathbb{R}$ and any statistic $\hat{f}_N$.

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 the Learning setup, on which we focus here, one is given a function class (also called a model) and the goal is to construct a procedure $\hat{f}_N$ satisfying a *sharp* or *exact oracle inequality*: ensuring 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.

Thus, the procedure $\hat{f}_N$ is a map from the set of N samples to L_2, and it performs with accuracy $\varepsilon_N = \varepsilon_N(\mathcal{F})$ and confidence $1 - \delta_N = 1 - \delta_N(\mathcal{F})$, if for every reasonable class $\mathcal{F}$ and any reasonable target Y, (1.1) is satisfied on an event of measure at least $1 - \delta_N$ and a residue that is at most ε_N.

Clearly, the risk functional is unknown 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 over $\mathcal{F}$. This procedure is called *empirical risk minimization* (*ERM*) and is defined by

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

ERM has been studied extensively over the last 20 years (see, e.g. [26], [13], [10]). The main focus has been to identify the connections between the structure of $\mathcal{F}$ and the residual term for ERM.

In particular, one would like to study the following questions:

1. Given any $0 < \delta_N < 1/2$, what are the error rates ε_N that one may obtain using ERM, and what features of $\mathcal{F}$ govern these rates?
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 an algorithm that can yield 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).

Our aim here is to go beyond the bounded case and proceed without any assumption on the envelope of $\{\ell(f(X),Y) : f \in \mathcal{F}\}$. Instead, we will consider the subgaussian setup.

Recall that if X is distributed according to a probability measure μ, then the ψ_2-norm of a function f is defined by

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

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

Definition 1.1 *A function class $\mathcal{F}$ is L-subgaussian with respect to the probability measure μ if for every $f, h \in \mathcal{F} \cup \{0\}$, $\|f-h\|_{\psi_2(\mu)} \leq L\|f-h\|_{L_2(\mu)}$.*

Our strategy for proving an oracle inequality for ERM is via the isomorphic method, introduced in [2]. Before presenting this method, recall that the excess loss of f is

$$\mathcal{L}_f(x,y) = \ell(f(x),y) - \ell(f^*(x),y) = (f(x)-y)^2 - (f^*(x)-y)^2, \qquad (1.2)$$

where we assume that f^* (an *oracle*) is a fixed element in the set of true minimizers $\operatorname{argmin}_{f\in\mathcal{F}} R(f)$. Only minor changes are needed if the infimum is not attained, an issue that will not be addressed here.

Set

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

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

$$\frac{1}{2}P\mathcal{L}_f \leq P_N\mathcal{L}_f \leq \frac{3}{2}P\mathcal{L}_f. \qquad (1.3)$$

On Ω_0,

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

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

Therefore, to obtain an exact oracle inequality with a confidence parameter δ_N, it suffices to identify λ_N for which Ω_0 has probability at least $1-\delta_N$.

Clearly, this is equivalent to identifying the level λ_N for which the supremum of the ratio process satisfies that

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

with probability at least $1-\delta_N$.

We will assume that the classes in question have reasonable structure in the following sense:

Definition 1.2 *A class $\mathcal{F}$ is B-Bernstein relative to the target Y, if for every $f\in\mathcal{F}$,*

$$\mathbb{E}\big(f(X)-f^*(X)\big)^2\leq BP\mathcal{L}=B\mathbb{E}\big((Y-f(X))^2-(Y-f^*(X))^2\big). \quad (1.4)$$

The class is star-shaped in f_0 if for every $f\in\mathcal{F}$,

$$\{\lambda f_0+(1-\lambda)f:0\leq\lambda\leq1\}\subset\mathcal{F}.$$

One may use the 2-convexity of L_2 and show that if $\mathcal{F}$ is convex then for any target $Y\in L_2$, $\mathcal{F}$ is 1-Bernstein; and that $\mathcal{F}-\mathcal{F}$ is star-shaped in all its elements.

As in many other estimates on the performance of ERM (e.g. [25, 10, 13]), the choice or the residual term is driven by a fixed point argument. Let $d_{\mathcal{F}}(L_2)$ be the diameter of $\mathcal{F}$ in $L_2(\mu)$ and set $\{G_f:f\in\mathcal{F}\}$ to be the canonical gaussian process indexed by $\mathcal{F}$; that is, with a covariance structure endowed by $L_2(\mu)$. Given a set $\mathcal{F}'$, denote by $\mathbb{E}\|G\|_{\mathcal{F}'}$ the expectation of the supremum of $\{G_f:f\in\mathcal{F}'\}$. We will assume throughout that $\mathbb{E}\|G\|_{\mathcal{F}'}$ is finite for all the sets $\mathcal{F}'$ in question.

Definition 1.3 *Let D be the unit ball in $L_2(\mu)$. For every $\eta>0$, let*

$$s_N^*(\eta)=\inf\left\{0<s\leq d_{\mathcal{F}}(L_2):\mathbb{E}\|G\|_{sD\cap(\mathcal{F}-\mathcal{F})}\leq\eta s^2\sqrt{N}\right\}, \quad (1.5)$$

and for every $Q>0$, set

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

In both cases, if the set is empty, set $s_N^*(\eta)=d_{\mathcal{F}}(L_2)$ (resp. $r_N^*(Q)=d_{\mathcal{F}}(L_2)$).

If either $s_N^*(\eta)$ or $r_N^*(Q)$ are equal to $d_{\mathcal{F}}(L_2)$, the resulting upper bound is trivial. Therefore, throughout 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.

One may show that if $\mathcal{F} - \mathcal{F}$ is star-shaped in 0, $s = s_N^*(\eta)$ and $r = r_N^*(Q)$, then $\mathbb{E}\|G\|_{sD\cap(\mathcal{F}-\mathcal{F})} = \eta s^2\sqrt{N}$ and $\mathbb{E}\|G\|_{rD\cap(\mathcal{F}-\mathcal{F})} = \theta r\sqrt{N}$. Indeed, the star-shape property implies that $H(s) = s^{-1}\mathbb{E}\|G\|_{sD\cap(\mathcal{F}-\mathcal{F})}$ is continuous from the left. Since $\mathbb{E}\|G\|_{sD\cap(\mathcal{F}-\mathcal{F})}$ is increasing, the choice of $s_N^*(\eta)$ leads to the equality. A similar argument proves the claim regarding $r_N^*(Q)$.

Moreover, of $s \geq s_N^*(\eta)$ then $\mathbb{E}\|G\|_{sD\cap(\mathcal{F}-\mathcal{F})} \leq \eta s^2\sqrt{N}$, and if $r \geq r_N^*(Q)$, $\mathbb{E}\|G\|_{rD\cap(\mathcal{F}-\mathcal{F})} \leq Qr\sqrt{N}$.

With these definitions in place, one may formulate the upper bound on the performance of ERM.

Theorem A. *For every $L \geq 1$ and $B \geq 1$ there exist constants c_1, c_2, c_3 and c_4 that depend only on B and L for which the following holds. Let $\mathcal{F}$ be an L-subgaussian and B-Bernstein class of functions relative to the target Y. Assume that $\mathcal{F} - \mathcal{F}$ is star-shaped in 0 and that $\|Y - f^*(X)\|_{\psi_2} \leq \sigma$. 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 $1-4\exp(-c_4 NQ^2)$,*
$$R(\hat{f}) \leq \inf_{f\in\mathcal{F}} R(f) + (r_N^*(Q))^2.$$

As mentioned above, if $\mathcal{F}$ is convex, the structural assumptions of Theorem A hold for every $Y \in L_2$.

We will show in what follows that the parameters involved 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 there were no "noise" in the problem – 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 $\tau = (X_i)_{i=1}^N$ 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, that does not depend on the noise level, can serve as an upper bound. And, indeed, $s_N^*(\eta)$

measures the interaction of the "noise" $f^* - Y$ with 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, which is very well established in Statistics (see, e.g., [22, 28, 29, 4, 3] for more details). Standard minimax bounds are based on information-theoretical results such as Fano's Lemma, Assouad's Lemma or Pinsker's inequalities, but unfortunately, these results do not yield lower bounds in the "high probability" realm, as needed to show the optimality of the rate obtained in Theorem A.

We therefore establish a minimax bound that is based on the gaussian shift theorem (and therefore on the gaussian isoperimetric inequality). It allows one to obtain a high probability minimax bound, and, as will be explained below, to recover the known constant probability minimax bound as well.

Consider the gaussian model, in which $(X_i, Y_i)_{i=1}^N$ is an independent sample of

$$Y_f = f(X) + W, \tag{1.6}$$

where $f \in \mathcal{F}$ and $W \sim \mathcal{N}(0, \sigma^2)$ is a gaussian noise, independent of X.

Theorem A′. *There exist absolute constants c_1, c_2 and c_3 for which the following holds. Let $\mathcal{F} \subset L_2$ be a class that is star-shaped in one of its points. If $\tilde{f}_N$ is a statistic constructed from a sample of cardinality N of the model (1.6) 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 $\eta \sim_{L,B} 1/\sigma$, (as is the case in Theorem A when the noise level is 'non-trivial'), then 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 $\delta_N \sim 1/2$, still, when the noise level is non-trivial.

It is not clear whether the isomorphic method, used to prove Theorem A, can yield a better accuracy if one is willing to accept a constant confidence. The next result shows that the answer is negative.

Theorem B. *Under mild assumptions on $\mathcal{F}$, X, Y and η (see Definition 3.1),*

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

This leads one to wonder if a better bound is possible at all, even if a different procedure than ERM is used.

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 for a confidence $\delta_N \sim 1/2$.

To formulate this observation, recall that if A and B are two subsets of L_2, then $N(A,\varepsilon B)$ is the minimal number of translates of εB needed to cover A.

Consider the "Sudakov" analog of the gaussian-based parameter $s_N^*(\eta)$: recall that by Sudakov's inequality (see, for example, [11])

$$\sup_{\varepsilon>0} \varepsilon \log^{1/2} N((\mathcal{F}-\mathcal{F})\cap rD, \varepsilon D) \lesssim \mathbb{E}\|G\|_{(\mathcal{F}-\mathcal{F})\cap rD}. \tag{1.7}$$

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

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

Theorem C. *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 $\tilde{f}_N$ is a procedure for that has a confidence parameter $\delta_N < 1/4$, then its accuracy satisfies $\varepsilon_N \geq c_1(q_N^*(c_2/\sigma))^2$.*

Theorem C is more classical and follows from Theorem 2.5 in [22] or from [28], though the proof presented here is new, and we feel it is more transparent than existing proofs. An added value is that it follows the same path as the proof of Theorem A′, and thus gives a scheme that may be used to prove lower bounds at every confidence level.

With Theorem A in mind, Theorem C shows that if the gaussian parameter $s_N^*(\eta)$ and the Sudakov-based one, $q_N^*(\eta)$, are equivalent for $\eta \sim 1/\sigma$

when $\sigma \gtrsim r_N^*$, the minimax rate in the constant probability realm is attained by ERM.

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

Definition 1.4 *Let $\mathcal{F}$ be a class of functions. For every sample $\tau = (X_1, ..., X_N)$ and $f \in \mathcal{F}$, set*

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

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

Clearly, if $\sigma = 0$ then for every sample τ, ERM selects $\hat{f} \in K(f^*, \tau)$. And since $Y \in \mathcal{F}$, $R(f) = \|f - f^*\|_{L_2}^2$. Thus, $R(\hat{f}) \leq \mathcal{D}(f^*, \tau)$. It is natural to ask whether the reverse direction is true, and also, to try and identify the correct rate when $0 < \sigma < r_N^*$.

The following result shows that the largest "typical" value of $\mathcal{D}(t, \tau)$ is a constant probability minimax bound, regardless of the choice of σ. It combines a "compressed sensing" type of a minimax results (see, e.g., [7, 5]) and statistics ones (e.g. [22, 28, 29, 4, 3]).

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

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

One natural example in which Theorem D may be used is when T is a convex, symmetric subset of $\mathbb{R}^d$ and $\mathcal{F}$ is the class of linear functionals $\{\langle t, \cdot \rangle : t \in T\}$. 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 $\Gamma = \sum_{i=1}^N \langle X_i, \cdot \rangle e_i$, then $\mathcal{D}(0, \tau)$ is the diameter of $\text{Ker}(\Gamma) \cap T$.

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

$$c_N(T) \leq \text{diam}\,(K(t_0, \tau) - t_0) \leq 2D(0, \tau),$$

and $c_N(T)/8$ is a lower estimate on a constant probability minimax bound.

In cases where $r_N^* \sim c_N(T)$, it follows that for every $0 \leq \sigma \lesssim r_N^*$, r_N^* is the minimax rate, and it is achieved by ERM. We will present one such example in Section 5.

One should note that the lower bound of Theorem B and the ones in Theorem A′, C and D are of different nature. Theorem B holds for any L-subgaussian class, input and target that satisfy certain regularity conditions. The others are minimax results and therefore hold only for the "worst" possible distribution according to the model in question.

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). The values of these constants 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 $C(b)$ depending only on b.

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

The first three sections of this article are devoted to the proofs of the four theorems. We then present two examples (regression in B_1^d and low-rank matrix inference) in which the rates established in Theorem A are proved to be sharp. The last section is devoted to concluding remarks on how the results may be extended to cases that are not covered here, and to a comparison with previous results.

2 Learning subgaussian classes

Since subgaussian classes play a central role in this article, we begin this section with a few examples of such classes. One may show that

$$\|f\|_{\psi_2} \sim \sup_{p \geq 2} \frac{\|f\|_{L_p}}{\sqrt{p}}.$$

Thus, if $\mathcal{F}$ is an L-subgaussian class of functions, for every $f, h \in \mathcal{F} \cup \{0\}$, $\sup_{p \geq 2} \|f - h\|_{L_p}/\sqrt{p} \leq L\|f - h\|_{L_2}$.

A measure μ on $\mathbb{R}^d$ is L-subgaussian if for every $t \in \mathbb{R}^d$, the linear functional $\langle t, \cdot \rangle$ is L-subgaussian. Hence, every class of linear functionals on $\mathbb{R}^d$ is L-subgaussian relative to the measure μ.

1. Let x be a mean-zero, variance 1 real-valued random variable which is L-subgaussian and let $x_1, \ldots, x_d$ be independent copies of x. It is straightforward to verify that for every $a \in \mathbb{R}^d$,

$$\Big\| \sum_{i=1}^{d} a_i x_i \Big\|_{\psi_2} \lesssim \|a\|_{\ell_2^d} \|x\|_{\psi_2}.$$

Thus, the random vector $X = (x_1, ..., x_d)$ is cL-subgaussian for a suitable absolute constant c. Moreover, it is isotropic (that is, for every $x \in \mathbb{R}^d$, $\mathbb{E}|\big\langle X, x \big\rangle|^2 = \|x\|_{\ell_2^d}^2$). Thus, for example, the uniform measure on $\{-1,1\}^d$ or on $[-1,1]^d$ are isotropic and L-subgaussian for an absolute constant L.

2. The uniform measure on $d^{1/p} B_p^d$ is also L-subgaussian for an absolute constant L (see [1]), despite the fact that its coordinates are not independent.

3. Let $X = (x_i)_{i=1}^d$ be an *unconditional* random vector, meaning that for every choice of signs $(\varepsilon_i)_{i=1}^d$, $(\varepsilon_i x_i)_{i=1}^d$ has the same distribution as X. If $\mathbb{E}x_i^2 \geq c^2$ for every i and X is supported in RB_∞^d then it is L-subgaussian for $L \lesssim R/c$. Indeed, by Khintchine's inequality [11], for any $p \geq 2$,

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

Also,

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

proving the claim.

4. 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 follows the same path as in example 1.

2.1 Proof of Theorem A

When considering the parameters r_N^* and s_N^* that appear in the upper bound on the error rate, what seems odd at first glance is the different normalization in their definition – the first is linear and the second quadratic. The two originate from the need to compare the behaviour of two processes. Indeed, recall that

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

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

In contrast, s_N^* is designed for the multiplier process, originating from the linear term $(f-f^*)(f^*-Y)$. To compare the resulting "linear" term 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 rather less-natural normalization in the definition of $s_N^*(\eta)$.

It goes without saying that an essential component of the proof of Theorem A must be an accurate analysis of the quadratic and linear terms, and both will be based on results from [20].

The first estimate we require is a bound on the squared empirical process:

Theorem 2.1 *[20] 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 $1-2\exp(-c_2 t^2(\mathbb{E}\|G\|_{\mathcal{H}}/Ld_{\mathcal{H}}(L_2))^2)$,*

$$\sup_{h\in\mathcal{H}} \left|\sum_{i=1}^N h^2(X_i) - \mathbb{E}h^2\right| \leq c_3 L^2\left(t^3 d_{\mathcal{H}}(L_2)\mathbb{E}\|G\|_{\mathcal{H}}\sqrt{N} + t^2(\mathbb{E}\|G\|_{\mathcal{H}})^2\right).$$

Here is a simple application of Theorem 2.1 that explains the role of r_N^*.

Lemma 2.2 *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 in 0 and set $f^* \in \mathcal{F}$. If $0 < Q \leq 1$ and $r \geq r_N^*(Q)$, then with probability at least $1-2\exp\big(-c_1\big(\mathbb{E}\|G\|_{rD\cap(\mathcal{F}-\mathcal{F})}/(Lr)\big)^2\big)$,*

$$\sup_{h\in(\mathcal{F}-f^*)\cap rD} \left|\frac{1}{N}\sum_{i=1}^N h^2(X_i) - \mathbb{E}h^2\right| \leq c_2 L^2 r^2 Q. \tag{2.1}$$

Moreover, if $Q \leq \min\{c_3/L^2, 1\}$, then with the same probability estimate, for every $f \in \mathcal{F}$ satisfying $\|f - f^*\|_{L_2} \geq r$,

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

Proof. The first part of the claim is an immediate corollary of Theorem 2.1 and the fact that if $r \geq r_N^*(Q)$, $\mathbb{E}\|G\|_{(\mathcal{F}-\mathcal{F})\cap rD}/\sqrt{N} \leq Qr$.

Even though the second part is known, we present it for the sake of completeness. Denote by Ω_0 the event on which (2.1) holds. Fix $f \in \mathcal{F}$ for which $\|f - f^*\|_{L_2} \geq r$ and set $h = r(f-f^*)/\left\|f - f^*\right\|_{L_2}$. Since $\mathcal{F}-\mathcal{F}$ is star-shaped in 0, $h \in (\mathcal{F} - \mathcal{F}) \cap rD$. Therefore, on Ω_0, and if $Q \leq \min(c_3/L^2, 1)$,

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

■

The second ingredient we require is a bound on multiplier processes.

Theorem 2.3 *[20] There exist absolute constants c_1, c_2 and c_3 for which the following holds. If $\mathcal{H}$ is an L-subgaussian class of functions and $\xi \in L_{\psi_2}$, then for every $t \geq c_1$, with probability at least $1-2\exp(-c_2 t^2(\mathbb{E}\|G\|_{\mathcal{H}}/L d_{\mathcal{H}}(L_2))^2)$,*

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

Note that in Theorem 2.3 one does not assume that ξ and X are independent, a fact that will be significant in what follows.

Theorem A follows immediately from Theorem 2.4 and the isomorphic method described in the introduction.

Theorem 2.4 *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 in 0 and that $\|Y - f^*(X)\|_{\psi_2} \leq \sigma$. Let $\eta = c_1/\sigma$ and $Q = c_2$.*

1. *If $\sigma \geq c_0 r_N^*(Q)$, then with probability at least $1-4\exp\left(-c_3\eta^2(s_N^*(\eta))^2 N\right)$,*

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

2. If $\sigma \le c_0 r_N^*(Q)$, then with probability at least $1 - 4\exp\left(-c_3 Q^2 N\right)$,

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

Proof. Let $\xi = (f^*(X) - Y)$ and observe that

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

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

$$\left\| \frac{f - f^*}{(P\mathcal{L}_f)^{1/2}} \right\|_{L_2}^2 \le B \text{ and } \left\| \frac{f - f^*}{P\mathcal{L}_f} \right\|_{L_2}^2 \le \frac{B}{P\mathcal{L}_f} \le \frac{B}{\lambda}. \tag{2.2}$$

Therefore,

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

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 put $\mathcal{H} = (\mathcal{F} - \mathcal{F}) \cap \sqrt{\lambda B} D$, where $D = B(L_2)$. Since $\mathcal{F} - \mathcal{F}$ is star-shaped in 0, it follows from (2.2) that

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

and

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

Fix $Q = c_0$ and $\eta = c_1/\sigma$ to be named later. Set $\lambda = (s_N^*(\eta))^2/B$ and $r = r_N^*(Q)$, and observe that

$$\mathbb{E}\|G\|_{(\mathcal{F}-\mathcal{F})\cap rD} = Qr\sqrt{N} = \frac{Q}{r\eta} \cdot \eta r^2\sqrt{N} \geq \eta r^2\sqrt{N},$$

provided that $\sigma \geq c_1 r_N^*(Q)/Q$. Therefore, $r_N^*(Q) \leq s_N^*(\eta) \equiv s$. Let $Q \lesssim 1/L^2B$; by Lemma 2.2 and since $\mathbb{E}\|G\|_{(\mathcal{F}-\mathcal{F})\cap sD} = \eta s^2\sqrt{N}$,

$$\sup_{w \in W_\lambda} \left| \frac{1}{N}\sum_{i=1}^N w^2(X_i) - \mathbb{E}w^2 \right| \leq c_2 L^2 QB \leq \frac{1}{4}$$

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 2.3, 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), set $\lambda = (r_N^*(Q))^2/B$. It follows from Lemma 2.2 and Theorem 2.3, that with probability at least $1 - 4\exp(-c_4Q^2N)$,

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

■

3 Lower bounds on the isomorphic method

The proof of the lower bound on the isomorphic method (formulated in Theorem B) is based on several estimates from [20]. Let $\mathcal{F}$ be a convex, symmetric class of functions (i.e., if $f \in \mathcal{F}$ then $-f \in \mathcal{F}$). Thus $\mathcal{F}-\mathcal{F} = 2\mathcal{F}$, and $s_N^*(\eta) = \inf\{r > 0 : \mathbb{E}\|G\|_{2\mathcal{F}\cap rD} \leq \eta r^2\sqrt{N}\}$. In addition, one may consider scaled versions of r_N^*: for every $1 \leq k \leq N$ and $Q > 0$, set

$$r_k(Q) = \inf\{r > 0 : \mathbb{E}\|G\|_{2\mathcal{F}\cap rD} \leq Qr\sqrt{k}\}.$$

The parameters $r_k(Q)$ measure the radii at which $2\mathcal{F} \cap rD$ has the same "complexity" as a k-dimensional Euclidean ball of radius r. Let $k^*_{\mathcal{F},Q}$ be the first integer larger than $(\mathbb{E}\|G\|_{\mathcal{F}}/Qd_{\mathcal{F}}(L_2))^2$. Thus, it is the first integer k for which $r_k(Q)$ exists. In what follows, we will sometimes write r_k and $k^*_{\mathcal{F}}$ instead of $r_k(Q)$ and $k^*_{\mathcal{F},Q}$.

Definition 3.1 *A class of functions $\mathcal{F}$ is c-skeletal if for every $k \geq k^*_{\mathcal{F}}$ there is a subset $\mathcal{F}_k \subset \mathcal{F} \cap r_k D$ of cardinality at most $\exp(k)$, for which*

$$\mathbb{E}\|G\|_{\mathcal{F}\cap r_k D} \leq c\mathbb{E}\|G\|_{\mathcal{F}_k}.$$

The existence of a skeleton implies that $\mathbb{E}\|G\|_{\mathcal{F}\cap r_k D}$ is exhibited by $\exp(k)$ points. It turns out that under such an assumption, a typical subgaussian projection of $\mathcal{F} \cap r_k D$ of dimension larger than k inherits some of the structure of $\mathcal{F} \cap r_k D$, since all the distances between the points of the skeleton are essentially preserved by the projection (see more details in [20] and in the proof of Theorem 3.3, below).

Among the examples of skeletal sets are convex, symmetric classes with a regular modulus of continuity of the gaussian process $\{G_f : f \in \mathcal{F}\}$:

Lemma 3.2 *[20] If $H(r) = \mathbb{E}\|G\|_{\mathcal{F}\cap rD}$ and there are $\alpha < 1$ and $0 < \beta < 1/2$ satisfying that for every $0 < r \leq d_{\mathcal{F}}(L_2)$,*

$$H(\alpha r) \leq \beta H(r), \tag{3.1}$$

then $\mathcal{F}$ is a c_1-skeletal set for $c_1 = c_1(\alpha,\beta)$.

Another feature of classes that satisfy (3.1) is that at every scale $r > 0$,

$$\log N(\mathcal{F} \cap rD, \alpha rD) \sim_{\alpha,\beta} (\mathbb{E}\|G\|_{\mathcal{F}\cap rD}/r)^2,$$

and the estimate following from Sudakov's inequality is sharp.

Indeed, let $\mathcal{A}$ be an αr-separated subset of $\mathcal{F} \cap rD$ and for every $f \in \mathcal{F}$, let $a_f \in \mathcal{A}$ satisfy that $\|a_f - f\|_{L_2} \leq \alpha r$. Then

$$\begin{aligned} H(r) =&\mathbb{E}\|G\|_{\mathcal{F}\cap rD} \leq \mathbb{E}\sup_{f\in\mathcal{A}} G_f + \mathbb{E}\sup_{f\in\mathcal{F}\cap rD} G_{f-a_f} \leq \mathbb{E}\sup_{f\in\mathcal{A}} G_f + \mathbb{E}\sup_{f\in 2\mathcal{F}\cap\alpha rD} G_f \\ \leq&\mathbb{E}\sup_{f\in\mathcal{A}} G_f + 2\mathbb{E}\sup_{f\in\mathcal{F}\cap\alpha rD} G_f = \mathbb{E}\sup_{f\in\mathcal{A}} G_f + 2H(\alpha r) \\ \leq&\mathbb{E}\sup_{f\in\mathcal{A}} G_f + 2\beta H(r). \end{aligned}$$

Thus, $\mathbb{E}\sup_{f\in\mathcal{A}} G_f \geq (1-2\beta)H(r)$. On the other hand, since $\mathcal{A}\subset rD$, $r\log^{1/2}|\mathcal{A}| \gtrsim \mathbb{E}\sup_{f\in\mathcal{A}} G_f$; hence

$$\alpha r\log^{1/2} N(\mathcal{F}\cap rD, \alpha r) \gtrsim \alpha(1-2\beta)\mathbb{E}\|G\|_{\mathcal{F}\cap rD},$$

as claimed. ■

Another example of a skeletal set is $\mathcal{F}=\{\langle t,\cdot\rangle : t\in B_1^d\}$, assuming that μ is an isotropic measure. One can show that

$$\log N(B_1^d\cap rB_2^d, \alpha r B_2^d) \sim_\alpha (\mathbb{E}\|G\|_{B_1^d\cap rB_2^d}/r)^2,$$

despite the fact that (3.1) does not hold for B_1^d.

The main ingredient in the proof of the lower bound on the ratio estimate is the following theorem.

Theorem 3.3 *For every c, Q, R, $L\geq 1$ and $q>2$, there exist constants c_0, c_1, c_2 and c_3 that depend only on c, Q, R, L and q for which the following holds. Let $\mathcal{F}$ be a c-skeletal and symmetric set, assume that $\xi\in L_q$ is a mean-zero, variance 1 random variable with $\|\xi\|_{L_q}\leq R$. Assume further that ξ satisfies the small-ball property $\mathbb{P}(|\xi|\leq t)\leq c_0 t$.*

Then for every $N\geq c_1 k\geq k^*_{\mathcal{F},Q}$,

$$\mathbb{E}\sup_{f\in\mathcal{F}\cap(r_kD\setminus c_2r_kD)}\left|\sum_{i=1}^N \varepsilon_i\xi_i f(X_i)\right| \geq c_3\sqrt{N}\mathbb{E}\|G\|_{\mathcal{F}\cap r_kD}.$$

The proof is almost identical to the proof of Theorem 6.1 from [20]. We will present full details of the minor differences between the two proofs and outline the rest.

Proof. First, one may show that if $V\subset\mathbb{R}^m$ is a symmetric set, $0<\theta<1$, and for every $1\leq p\leq\theta m$ and every $u,w\in V$,

$$\left\|\sum_{i=1}^m \varepsilon_i(u-w)_i\right\|_{L_p} \geq \rho\left\|\sum_{i=1}^m g_i(u-w)_i\right\|_{L_p},$$

then

$$\mathbb{E}_\varepsilon\sup_{v\in V}\sum_{i=1}^m \varepsilon_i v_i \geq c_0\rho\mathbb{E}_g\sup_{v\in V}\sum_{i=1}^m g_i v_i,$$

where c_0 depends only on θ (see Lemma 6.4 in [20]).

Second, it is well known that for every $v \in V$, $\| \sum_{i=1}^m g_i v_i \|_{L_p} \sim \sqrt{p} \|v\|_{\ell_2^m}$, and [17] showed that

$$\left\| \sum_{i=1}^m \varepsilon_i v_i \right\|_{L_p} \sim \sum_{i=1}^p v_i^* + \sqrt{p} \left(\sum_{i>p} (v_i^2)^* \right)^{1/2}$$

where, given $v \in \mathbb{R}^N$, $(v_i^*)_{i=1}^N$ is a monotone non-increasing rearrangement of $(|v_i|)_{i=1}^N$.

The next observation is that if $\mathcal{F}$ is skeletal, one may assume that the skeleton $\mathcal{F}_k$ is symmetric and is contained in $\mathcal{F} \cap (r_k D \backslash c_1 r_k D)$. Indeed, the symmetry of $\mathcal{F}_k$ follows from the symmetry of $\mathcal{F}$. For the second part, let $\mathcal{F}_k'$ be a c-skeleton of $\mathcal{F} \cap r_k D$, and let $0 < \alpha < 1$. By standard properties of the gaussian process and the definition of $r_k = r_k(Q)$,

$$\begin{aligned} &\mathbb{E} \sup_{f \in \mathcal{F}_k' \cap \alpha r_k D} G_f \lesssim \alpha r_k \log^{1/2} |\mathcal{F}_k'| \leq \alpha r_k \sqrt{k} \\ &\leq (\alpha/Q) \mathbb{E} \sup_{f \in 2\mathcal{F} \cap r_k D} G_f \leq 2(\alpha/cQ) \mathbb{E} \sup_{f \in \mathcal{F}_k'} G_f. \end{aligned}$$

Thus, for a sufficiently small α, $\mathcal{F}_k = \mathcal{F}_k' \cap (r_k D \backslash \alpha r_k D)$ satisfies that

$$\mathbb{E} \sup_{f \in \mathcal{F}_k} G_f \geq (1/2) \mathbb{E} \sup_{f \in \mathcal{F}_k'} G_f \geq (c/2) \mathbb{E} \|G\|_{\mathcal{F} \cap r_k D}.$$

Next, one may also show that if $N \geq c_1(L) k$, then with probability at least $1 - 2\exp(-c_2(L) k)$, vectors in the set

$$P_\sigma \mathcal{F}_k = \{ (f(X_i))_{i=1}^N : f \in \mathcal{F}_k \}$$

have the following structure: for every $f_1, f_2 \in \mathcal{F}_k$,

$$\frac{1}{2} \|f_1 - f_2\|_{L_2} \leq \left(\frac{1}{N} \sum_{i=1}^N (f_1 - f_2)^2 (X_i) \right)^{1/2} \leq \frac{3}{2} \|f_1 - f_2\|_{L_2}, \tag{3.2}$$

and for every $J \subset \{1, ..., N\}$,

$$\left(\sum_{j \in J} (f_1 - f_2)^2 (X_j) \right)^{1/2} \lesssim_L \|f_1 - f_2\|_{L_2} \left(\sqrt{k} + \sqrt{|J| \log(eN/|J|)} \right).$$

Fix $0 < \beta < 1$ to be named later and let $I \subset \{1, ..., N\}$, $|I| \geq (1-\beta)N$. Set $u = P_\sigma f_1$, $w = P_\sigma f_2$ and observe that for every $1 \leq p \leq N$,

$$\begin{aligned}
&\|(P_I(u-w))^*_{i \geq p}\|_{\ell_2^N} \geq \|u-w\|_{\ell_2^N} - \|(u-w)^*_{i \leq p}\|_{\ell_2^N} - \|(u-w)^*_{i \leq \beta N}\|_{\ell_2^N} \\
\geq &\|f_1 - f_2\|_{L_2} \left(\frac{\sqrt{N}}{\sqrt{2}} - c_3 \left(\sqrt{k} + \max\left\{ \sqrt{p} \log(\frac{eN}{p}), \sqrt{\beta N \log(\frac{e}{\beta})} \right\} \right) \right) \\
\gtrsim &\sqrt{N} \|f_1 - f_2\|_{L_2},
\end{aligned}$$

provided that $p, k \leq c_4(L)N$ and $\beta \leq c_5(L)$. On the other hand,

$$\|P_I(u-v)\|_{\ell_2^N} \leq \|u-v\|_{\ell_2^N} \lesssim \sqrt{N} \|f_1 - f_2\|_{L_2}.$$

Therefore, given $f_1, f_2 \in \mathcal{F}_k$, $p \leq c_4 N$, $I \subset \{1, ..., N\}$ of cardinality $|I| \geq (1-\beta)N$ and $\beta \leq c_5$,

$$\begin{aligned}
&\left\| \sum_{i \in I} \varepsilon_i (f_1 - f_2)(X_i) \right\|_{L_p(\mu_\varepsilon)} \gtrsim \sqrt{p} \left\| (P_I(P_\sigma f_1 - P_\sigma f_2))^*_{i \geq p} \right\|_{\ell_2^N} \gtrsim \sqrt{pN} \left\| f_1 - f_2 \right\|_{L_2} \\
&\gtrsim \sqrt{p} \left\| P_I(P_\sigma f_1 - P_\sigma f_2) \right\|_{\ell_2^N} \gtrsim \left\| \sum_{i \in I} g_i (f_1 - f_2)(X_i) \right\|_{L_p(\mu_g)},
\end{aligned}$$

where μ_ε and μ_g are the measures endowed by the random vectors $(\varepsilon_i)_{i=1}^N$ and $(g_i)_{i=1}^N$ respectively. Thus, recalling that $\mathcal{F}_k$ is symmetric, it follows that on that event and for every such a subset $I \subset \{1, \ldots, N\}$,

$$\mathbb{E}_\varepsilon \sup_{f \in \mathcal{F}_k} \sum_{i \in I} \varepsilon_i f(X_i) \gtrsim_L \mathbb{E}_g \sup_{f \in \mathcal{F}_k} \sum_{i \in I} g_i f(X_i).$$

By Slepian's Lemma (see, e.g. [11]), combined with (3.2), and since $\mathbb{E}\|G\|_{\mathcal{F}_k} \gtrsim_c \mathbb{E}\|G\|_{\mathcal{F} \cap r_k D}$,

$$\mathbb{E}_g \sup_{f \in \mathcal{F}_k} \sum_{i \in I} g_i f(X_i) \gtrsim_c \sqrt{N} \mathbb{E}\|G\|_{\mathcal{F} \cap r_k D}.$$

Next, recall that ξ satisfies the small ball property $\mathbb{P}(\|\xi\| \leq t) \leq c_6 t$. If β is as above, then by a binomial estimate,

$$\begin{aligned}
&\mathbb{P}(|\{i : |\xi_i| \leq t\}| \geq \beta N) \leq \binom{N}{\beta N} \left(\mathbb{P}(|\xi| \leq t) \right)^{\beta N} \\
\leq &\exp\left(\beta N (\log(e/\beta) - \log(1/c_6 t)) \right).
\end{aligned}$$

Hence, if $t = c_7\beta$, then with probability at least $1 - 2\exp(-c_8\beta N)$,

$$|\{i : |\xi_i| \leq c_7\}| \leq \beta N.$$

Let $I = \{i : |\xi_i| \geq c_7\}$ and note that $|I| \geq (1-\beta)N$. By the symmetry of $\mathcal{F}_k$ and the contraction principle for Bernoulli processes (see, for example, chapter 4 in [11]), with probability at least $1 - 2\exp(-c_2 k) - 2\exp(-c_8\beta N)$,

$$\mathbb{E}_\varepsilon \sup_{f \in \mathcal{F}_k} \left| \sum_{i=1}^N \varepsilon_i \xi_i f(X_i) \right| \geq c_7 \mathbb{E}_\varepsilon \sup_{f \in \mathcal{F}_k} \sum_{i \in I} \varepsilon_i f(X_i) \geq c_9 \sqrt{N} \mathbb{E} \|G\|_{\mathcal{F} \cap r_k D},$$

and c_9 depends only on L, q and $\|\xi\|_{L_q}$. ■

Having established Theorem 3.3, one may turn to the proof of the lower bound. Let $\mathcal{F}$ be a convex, symmetric, L-subgaussian and c-skeletal class. Assume that the target Y has mean-zero and variance one, belongs to L_q for some $q > 2$ and satisfies a small-ball property. Assume further that Y is orthogonal to span$(\mathcal{F})$, and thus $f^* = 0$ and $\xi = f^*(X) - Y = -Y$. Therefore,

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

and $P\mathcal{L}_f = \|f\|_{L_2}^2$. Clearly $\mathcal{F} - f^* = \mathcal{F}$, and for every $\lambda > 0$ the resulting ratio process is

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

To upper bound the quadratic term, fix Q to be named later, consider $\lambda = r_k^2(Q)$ for some $1 \leq k \leq N$ and let $\mathcal{H} = \{f/\|f\|_{L_2} : \|f\|_{L_2} \geq r_k(Q)\}$. Since $\mathcal{F}$ is star-shaped in 0, $\mathcal{H} \subset \frac{1}{r_k}(\mathcal{F} \cap r_k D)$, and by Theorem 2.1,

$$\mathbb{E} \sup_{h \in \mathcal{H}} \left| \frac{1}{N} \sum_{i=1}^N h^2(X_i) - \mathbb{E}h^2 \right| \lesssim_L \frac{1}{r_k^2} \cdot r_k^2 \sqrt{\frac{k}{N}} (Q + Q^2) \lesssim_L Q \sqrt{\frac{k}{N}},$$

provided that $Q \leq 1$.

To lower bound the "linear" term, let $\mathcal{F}_k \subset \mathcal{F} \cap (r_k D \backslash c_1 r_k D)$ be the corresponding skeleton of $\mathcal{F}$ at the level r_k, and observe that by a symmetrization and contraction argument and Theorem 3.3,

$$\begin{aligned}\mathbb{E} \sup_{\{f:\mathbb{E}f^2 \geq c_1 r_k^2\}} \left| \sum_{i=1}^N Y_i \frac{f(X_i)}{\mathbb{E}f^2} \right| \gtrsim & \mathbb{E} \sup_{f \in \mathcal{F}_k} \left| \sum_{i=1}^N \varepsilon_i Y_i \frac{f(X_i)}{\mathbb{E}f^2} \right| \\ \gtrsim \frac{1}{r_k^2} \mathbb{E} \sup_{f \in \mathcal{F}_k} \left| \sum_{i=1}^N \varepsilon_i Y_i f(X_i) \right| \geq & \frac{c_2}{r_k^2} \sqrt{k} r_k \sqrt{N} = c_2 \frac{\sqrt{kN}}{r_k}.\end{aligned}$$

Therefore,

$$\mathbb{E} \sup_{\{f:P\mathcal{L}_f \geq c_1 r_k^2\}} \left| \sum_{i=1}^N \frac{\mathcal{L}_f(X_i,Y_i)}{P\mathcal{L}_f} - 1 \right| \geq c_3 \left(\frac{1}{r_k}\sqrt{\frac{k}{N}} - L^2 Q \sqrt{\frac{k}{N}} \right) \geq 1$$

provided that $r_k(Q) \leq c_4 Q \sqrt{k/N}$, and c_4 depends only on L, q and $\|Y\|_{L_q}$.

Corollary 3.4 *Let $\mathcal{F}$ and Y be as above, set*

$$A = \left\{ k \geq k^*_{\mathcal{F},Q} : r_k(Q) \leq c_0 Q \sqrt{\frac{k}{N}} \right\}$$

for a constant c_0 that depends only on c, L, q and $\|Y\|_{L_q}$, and put $k_1 =$ $\min A$. Then for $Q \leq 1$,

$$\mathbb{E} \sup_{\{f:\mathbb{E}\mathcal{L}_f \geq r_{k_1}^2(Q)\}} \left| \frac{\mathcal{L}_f(X_i,Y_i)}{\mathbb{E}\mathcal{L}_f} - 1 \right| \geq 1.$$

To conclude the proof of Theorem B, one has to identify the connections between $s_N^*(\eta)$ and $r_k(Q)$.

Lemma 3.5 *Using the same notation as above, if $\eta \leq 2/c_0$, $r_{k_1}(Q)/2 \leq s_N^*(\eta)$ and if $\eta \geq 1/c_0$, $r_{k_1-1}(Q) \geq s_N^*(\eta)$.*

Proof. Observe that for $k \in A$, $\sqrt{k} \geq r_k(Q)\sqrt{N}/(c_0 Q)$. If $\eta \leq 2/c_0$ then

$$\mathbb{E}\|G\|_{\mathcal{F}\cap(r_k(Q)/2)D} \geq Q\sqrt{k} r_k(Q)/2 \geq (1/2c_0) r_k^2(Q)\sqrt{N} \geq \eta\sqrt{N} r_k^2(Q)/4,$$

implying that $r_k(Q)/2 \leq s_N^*(\eta)$.

In the reverse direction, let $r_k(Q)$ be the largest fixed point satisfying $r_k(Q) < s_N^*(\eta)$. If $k \notin A$ then $\sqrt{k} \leq r_k(Q)\sqrt{N}/c_0 Q$. Therefore,

$$\mathbb{E}\|G\|_{\mathcal{F}\cap r_k(Q)D} \leq Q\sqrt{k} r_k(Q) \leq r_k^2(Q)\sqrt{N}/c_0,$$

and if $1/c_0 \leq \eta$ then $s_N^*(\eta) \leq r_k(Q)$, which is impossible. Hence, $r_{k_1-1}(Q) \geq$ $s_N^*(\eta)$, as claimed. ■

Combining Corollary 3.4 with Lemma 3.5 shows that if $Q < 1$ and $\eta \geq 1/c_0$ then

$$\mathbb{E} \sup_{\{f: P\mathcal{L}_f \geq s_N^*(\eta)\}} \left| \frac{\mathcal{L}_f(X_i,Y_i)}{P\mathcal{L}_f} - 1 \right| \geq 1,$$

proving Theorem B and complementing the upper bound in Theorem 2.4.

4 Minimax lower bounds

Here, we will study the gaussian model, given by $Y = f(X) + W$, where $\mathcal{F}$ is a class of functions on a probability space (Ω, μ) and $f \in \mathcal{F}$. For any $\tau = (x_1, \ldots, x_N) \in \Omega^N$ and $f \in \mathcal{F}$, consider the conditional probability measure $\nu_{f,\tau}$ of $(Y_i | X_i = x_i)_{i=1}^N$ given by

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

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

Let

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

for the squared excess loss functional with $Y = f(X) + W$; namely, $\mathcal{L}_h(X,Y) = (Y - h(X))^2 - (Y - f(X))^2$.

Note that if a procedure $\tilde{f}_N$ has accuracy ε_N with a confidence parameter δ_N then for every $f \in \mathcal{F}$,

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

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

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

Theorem 4.1 *There exists an absolute constant c_1 for which the following holds. If $\mathcal{F}$ is star-shaped in one of its points and $\tilde{f}_N$ is a procedure with a confidence parameter $\delta_N < 1/4$ then its accuracy satisfies*

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

Theorem 4.1 shows that if a procedure has a confidence parameter $\delta_N = \exp(-c_0\gamma N)$, then its accuracy is, at best, $\varepsilon_N \geq c_2\sigma^2\gamma$. Taking $\gamma = \eta(s_N^*(\eta))^2$ for $\eta \sim \sigma^{-1}$ proves the second part of Theorem A$'$, and shows that ERM achieves the minimax rate for the confidence established in Theorem A if the noise level is nontrivial.

The proof of Theorem 4.1 requires several preliminary steps.

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

Lemma 4.2 *For every $f \in \mathcal{F}$,*

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

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

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

Since $\|\rho\|_{L_\infty} \leq 1$ and $\mathbb{E}\rho(\tau) \geq 1 - \delta_N$, by the Paley-Zygmund Theorem, $\mathbb{P}(\rho(\tau) \geq x) \geq 1 - \delta_N/(1-x)$ for every $x > 0$. The result follows by selecting $x = 1 - \sqrt{\delta_N}$. ■

Observe that for every $f \in \mathcal{F}$ and $\tau = (x_1, ..., x_N)$, $\nu_{f,\tau}$ is a gaussian measure on $\mathbb{R}^N$ with mean $P_\tau f = (f(x_i))_{i=1}^N$ and covariance matrix $\sigma^2 I_N$. Denote by $t \mapsto \Phi(t) = \mathbb{P}(g \leq t)$, the cumulative distribution function of a standard gaussian random variable g on $\mathbb{R}$.

Lemma 4.3 *Let $u, v \in \mathbb{R}^N$ and consider two gaussian measures $\nu_u \sim \mathcal{N}(u, \sigma^2 I_N)$ and $\nu_v \sim \mathcal{N}(v, \sigma^2 I_N)$ on $\mathbb{R}^N$. If A is a measurable subset of $\mathbb{R}^N$ 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 4.3 is a version of the gaussian shift theorem.

Theorem 4.4 *[12] 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 4.3. 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 4.4, the corresponding halfspace is

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

and therefore, if $w^\perp$ denotes the space of vectors 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 4.4 and the definition of w. ■

Proof of Theorem 4.1. Let $\tilde{f}_N$ be a procedure with accuracy $\varepsilon_N\leq d_{\mathcal{F}}^2(L_2)/4$ and a confidence parameter δ_N. Shifting $\mathcal{F}$ if needed, and since $\mathcal{F}$ is star-shaped in one of its points, one may assume that $u=0\in\mathcal{F}$ and that $v\in\mathcal{F}$ satisfies that $4\varepsilon_N\leq\|v\|_{L_2}^2\leq 8\varepsilon_N$. By Chebyshev's inequality, $\mathbb{P}\big(\|P_\tau v\|_{\ell_2^N}^2\geq 4N\|v\|_{L_2}^2\big)\leq 1/4$, and thus, for $\tau=(X_i)_{i=1}^N$ is a set of μ^N-probability at least $3/4$, $\|P_\tau v\|_{\ell_2^N}\leq c_1\sqrt{N}\|v\|_{L_2}$.

Consider the sets

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

which, by the choice of v, are disjoint. Since $\tilde{f}_N$ has accuracy ε_N and a confidence parameter δ_N, $\nu_{0,\tau}\otimes\mu^N(\mathcal{A}_0)\geq 1-\delta_N$ and $\nu_{v,\tau}\otimes\mu^N(\mathcal{A}_v)\geq 1-\delta_N$. Applying Lemma 4.2, with μ^N-probability at least $1-2\sqrt{\delta_N}$,

$$\nu_{0,\tau}(\mathcal{A}_0|\tau)\geq 1-\sqrt{\delta_N},\ \text{ and }\ \nu_{v,\tau}(\mathcal{A}_v|\tau)\geq 1-\sqrt{\delta_N}. \qquad (4.1)$$

Let Ω_0 be the set of samples $\tau=(X_i)_{i=1}^N\subset\Omega^N$ for which $\|P_\tau v\|_{\ell_2^N}\leq c_1\sqrt{N}\|v\|_{L_2}$ and (4.1) holds. Hence, $\mathbb{P}(\Omega_0)\geq 3/4-2\sqrt{\delta_N}$, and by Lemma 4.3 applied to the set $\mathcal{A}_0|\tau$,

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

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 $\|P_\tau v\|_{\ell_2^N}\leq\sigma|\Phi^{-1}(\sqrt{\delta_N})|$ then $(*)>1/2$.

Since $\tau\in\Omega_0$, $\|P_\tau v\|_{\ell_2^N}\leq c_1\sqrt{N}\|v\|_{L_2}$; therefore, if

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

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

Thus,

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

and by the choice of v,

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

as claimed. ■

Next, we turn to the proof of Theorem C which is a straightforward application of the next observation:

Theorem 4.5 *There exists an absolute constant c_0 for which the following holds. Let $\mathcal{F}$ and Y be as above, and assume that $\tilde{f}_N$ is a procedure with accuracy $\varepsilon_N = a_N^2$ and a confidence parameter $\delta \leq 1/4$. 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. Let $a = a_N$, set $D(f,r) = \{h \in \mathcal{F} : \|f - h\|_{L_2} \leq r\}$ and put Λ to be a maximal $2a$-separated set of $\mathcal{F} \cap (f + \theta a D)$ with respect to the L_2 norm; thus, $(D(f,a) : f \in \Lambda)$ is a family of disjoint sets in $\mathcal{F} \cap (f + \theta a D)$.

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

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

If $u \neq v$ in Λ and $A \subset \mathbb{R}^N$ then by Lemma 4.3,

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

Fix $v_0 \in \Lambda$, and since $\big(\mathcal{A}_v|\tau, v \in \Lambda\big)$ is a family of disjoint sets,

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

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

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

Taking the expectation with respect to τ,

$$1 \geq \sum_{v\in\Lambda} \mathbb{E}_\tau \int_{z_\tau(v)}^{\infty} \varphi(x)dx, \tag{4.2}$$

and it remains to lower bound each expectation.

Since

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

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

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

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

$$\|P_\tau v_0 - P_\tau v\|_{\ell_2^N} \leq (3/2)\sqrt{N}\|v_0-v\|_{L_2} \leq (3/2)\theta a\sqrt{N},$$

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

$$z_\tau(v) \leq \beta + (3/2)\sqrt{N}\theta a/\sigma$$

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

$$\mathbb{E}_\tau \int_{z_\tau(v)}^{\infty} \varphi(x)dx \geq \frac{1}{3}\int_{\beta+(3/2)\sqrt{N}\theta a_N/\sigma}^{\infty} \varphi(x)dx \gtrsim \exp\Big(-\frac{c_2 N\theta^2 a^2}{\sigma^2}\Big).$$

Thus, by (4.2), $1 \gtrsim |\Lambda|\exp\big(-c_3 N\theta^2 a^2/\sigma^2\big)$, as claimed. ■

We conclude this section with the proof of Theorem D, which is presented for a random design, though a proof for a deterministic design is almost identical. The idea behind the proof is that if $\tau = (X_1, ..., X_N)$ and $P_\tau f_1 = P_\tau f_2$, then the two functions are indistinguishable on a sample $(X_i, Y_i)_{i=1}^N$ of a model $Y^{f_1} = f_1(X) + V$. Therefore, it seems unlikely that one may find a procedure that performs better than the "worse" typical L_2 diameter of sets

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

which is denoted by $\mathcal{D}(f,\tau)$.

Fix $f \in \mathcal{F}$ and let $\tilde{f}_N$ be a given procedure. Define an $\mathcal{F}$-valued random variable h_f, as follows. Let $h_{1,\tau}(f)$ and $h_{2,\tau}(f)$ be almost L_2-diametric points in $K(f,\tau)$. Let δ be a $\{0,1\}$-valued random variable with mean $1/2$, which is independent of X and V, and set

$$h_f = (1-\delta)h_{1,\tau}(f) + \delta h_{2,\tau}(f). \tag{4.3}$$

Note that for every realization of δ, $\mathcal{D}(h_f,\tau) = \mathcal{D}(f,\tau)$. Let $I(A)$ be the indicator of the set A and observe that for every realization of the random variable δ,

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

Put

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

and

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

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

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

Note that for any sample τ, $h_{1,\tau}(X_i)+V_i = h_{2,\tau}(X_i)+V_i$; therefore,

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

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

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

To conclude the proof, observe that the squared excess risk of $\tilde{f}_N$ for the model $Y^f = f(X)+V$ is the square of the L_2 distance between $\tilde{f}_N$ and f. ■

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

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

$$K(v_0, \tau) = \{v \in T : \Gamma v = \Gamma v_0\} = v_0 + (T \cap \ker\Gamma),$$

and since T is convex, the largest diameter is attained for $v_0 = 0$.

Let $d_N = d_N(\rho)$ satisfy that with probability at least $1-\rho$, $\mathcal{D}(0, \tau) \geq d_N$. Then, by Theorem D, any procedure with a confidence parameter $\delta_N \leq 1/2 + \rho$ has its accuracy parameter ε_N larger than $d_N(\rho)/4$.

On the other hand, a straightforward application of Lemma 2.2 shows that with probability at least $1 - 2\exp(-c_1 N Q^2)$, $\mathcal{D}(0, \tau) \leq r_N^*(Q)$. In certain cases, $c_N(T) \sim r_N^*(Q)$ for a suitable absolute constant Q. Thus, with the same probability estimate,

$$r_N^*(Q) \lesssim c_N(T) \leq \mathcal{D}(0, \tau) \leq r_N^*(Q),$$

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

5 Examples

Here, we will present two examples of problems in which our results may be used. 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.

5.1 Learning over the B_1^d ball

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 $\|Y - f^*\|_{\psi_2} \leq \sigma$.

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 gaussian process is given by $t \to \sum_{i=1}^d g_i t_i$, and D is the unit ball in ℓ_2^d.

One may show (see, for example, [9]) that for every $1/\sqrt{d} \leq s \leq 2$,

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

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

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

Therefore, 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,

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

where c_1 and c_2 are constants that depend only on Q. 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 and $\eta = c_0/\sigma$, and let $N \leq c_1 d$. If $\sigma \geq r_N^*$ then also $\sigma^2 d^2 \gtrsim N$. Hence,

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

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

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

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

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

where c_3, c_4, c_5 depend on L, B and the choice of Q.

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

Turning to the lower estimate and as noted in Theorem A′, if $\tilde{f}_N$ is a procedure with accuracy ε_N, that has to achieve 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.

Moreover, since $T = B_1^d$ is skeletal, then by Theorem B, the isomorphic method cannot be used to improve the rate of $(s_N^*(c_0/\sigma))^2$ for σ sufficiently small.

For a lower bound with constant probability, recall that to apply Theorem C, one has to bound the covering numbers

$$\log N(B_1^d \cap r B_2^d, \theta r B_2^d)$$

from below for some $\theta < 1$.

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

$$\bigcup_{|I|=k} r S^I \subset B_1^d \cap r B_2^d.$$

Recall the well known fact (see, e.g., [14]) that there is a collection $\mathcal{B}$ of subsets of $\{1, ..., d\}$ of cardinality k, that is $c_1 k$ separated in the Hamming distance, and $\log |\mathcal{B}| \geq c_2 k \log(ed/k)$. The set $\Lambda = \{r \sum_{i \in I} e_i : I \in \mathcal{B}\}$ is a $c_8 r$-separated subset of $B_1^d \cap r B_2^d$ relative to the ℓ_2^d distance. Hence,

$$\log N(B_1^d \cap r B_2^d, c_8 r B_2^d) \geq c_9 \frac{\log(edr^2)}{r^2}.$$

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

$$\frac{\log(edr^2)}{r^2} \lesssim \log N(B_1^d \cap r B_2^d, c_8 r B_2^d) \lesssim \frac{Nr^2}{\sigma^2}.$$

Therefore,

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

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

If $r \leq 1/\sqrt{d}$, $\log N(B_1^d \cap rB_2^d, c_{10}rB_2^d) \gtrsim 2^d$, and

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

if $d^2\sigma^2 \leq N$. Thus, $\tilde{f}_N$ cannot outperform ERM in the noisy case, even if allowed to succeed only with 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 D 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 [8],

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

Thus, for $0 \leq \sigma \lesssim r_N^*(Q)$, $\tilde{f}_N$ does not outperform ERM.

5.2 Low-rank matrix inference via the max-norm

In this type of problem, the goal is to explain an output 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,

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

A similar estimator has been studied in [21] for $Y = \langle A^*, X\rangle + W$, a random vector X that is selected uniformly from the canonical bases of $\mathbb{R}^{p\times q}$, a noise vector W that is either gaussian or sub-exponential noise with independent coordinates, and matrices in $\mathcal{B}_{max}$ with bounded entries.

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

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

It is straightforward to verify that if the X is not an isotropic vector, but rather, only equivalent to an isotropic one, similar estimates to the ones presented below hold, and the modifications required in the proofs are minor.

Let A^* be the true minimizer of the squared loss in $\mathcal{B}_{max}$ and set the "noise parameter" $\left\|Y - \left\langle X, A^* \right\rangle\right\|_{\psi_2} \leq \sigma$. Since $\mathcal{F}$ is convex, the true minimizer is unique and the Bernstein and star-shape 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 $\eta \sim_L \sigma^{-1}$.

Set B_F to be the unit ball relative to the Frobenius norm and observe that since X is isotropic, the relative L_2 unit ball is

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

and the corresponding gaussian process has a covariance structure given by

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

A simple application of Grothendieck's inequality (see, e.g., [18]) shows that

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

where K_G is the Grothendieck constant and $\mathcal{X}_\pm = \{uv^\top : u \in \{\pm 1\}^p, v \in \{\pm 1\}^q\}$.

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

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

By standard properties of gaussian processes,

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

In the reverse direction, by Lemma 3.1 in [21], if

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

then

$$s \log^{1/2} N(\mathcal{B}_{max} \cap s\sqrt{pq} B_F, s\sqrt{pq/2} B_F) \gtrsim \sqrt{p+q}. \tag{5.1}$$

Hence, in that range of s,

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

and

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

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

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

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

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

To see that this estimate is sharp in the minimax sense when $\sigma \gtrsim \sqrt{(p+q)/N}$, consider the gaussian regression model $Y=\langle A^*,X\rangle+W$ and 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 C and (5.1), any procedure with confidence parameter $\delta_N \le 1/4$ has accuracy $\varepsilon_N \gtrsim \sigma\sqrt{\frac{p+q}{N}}$, matching the upper bounds in the noisy regime.

6 Concluding remarks and comparisons with existing results

Subgaussian classes of functions play a central role in our presentation. The reason for focusing on such classes is that, on one hand, there are many natural examples that fall within the subgaussian framework, and on the other, because the substantial technical machinery needed to establish Theorem A and B is not known in general. Perhaps surprisingly, the difficult part in developing such a theory is not the slow decay of tails of individual class members, but rather, the lack of a framework that captures the “global” complexity of the class – as $\mathbb{E}\|G\|_F$ does in the subgaussian case.

There are cases, though, in which such a theory exists (e.g., unconditional, log-concave vectors in $\mathbb{R}^d$) and one may prove analogous results to the ones presented here. Since the technical cost is rather substantial and would obscure the main message of this article, we decided to leave these generalizations to future work. For more details on these directions, we refer the reader to [16, 15].

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, $\sigma \lesssim r_N^*$, it is known to be sharp only when $\sigma = 0$ (the error rate is a typical value of $\mathcal{D}^2(f^*, \tau)$) or if $\sigma \sim r_N^*$, where 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 situation is even less clear. In the noisy case, when $\sigma \gtrsim r_N^*$, the upper bound of $(s_N^*(c/\sigma))^2$ is sharp only if the gaussian parameter $s_N^*(c/\sigma)$ and the Sudakov-based one, $q_N^*(c/\sigma)$ are equivalent. 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.

Therefore, he have shown that for the gaussian noise, ERM achieves the minimax rate of convergence $\max\big((s_N^*(c/\sigma))^2, (r_N^*(Q))^2\big)$ in the constant probability regime for both ranges of noise, if $\mathcal{F}$ is a convex subgaussian class, satisfying

1. $q_N^* \log^{1/2} N(\mathcal{F} \cap 2q_N^* D, q_N^* D) \sim \mathbb{E}\, \|G\|_{\mathcal{F} \cap q_N^* D}$ – meaning that there is no gap in the Sudakov inequality at scale $q_N^* = q_N^*(c/\sigma)$;
2. $c_N(\mathcal{F}) \sim r_N^*(\mathcal{F})$ – meaning that $\sqrt{N} c_N(\mathcal{F} \cap r_N^* D) \sim \mathbb{E}\, \|G\|_{\mathcal{F} \cap r_N^* D}$, and there is no gap in the Pajor-Tomczak-Jaegermann estimate on the N-Gelfand width. (see [19]).

It seems unlikely that these conditions on the regularity of $\mathcal{F}$ are necessary; the second one if less likely than the first, as an estimate on the "random" width rather than the minimal one suffices for the lower bound. Another issue is that the isomorphic method only leads to an upper bound on the performance of ERM, which is another possible reason for a suboptimal estimate in the constant probability regime. Since $\hat{f}$ minimizes $f \mapsto P_n \mathcal{L}_f$ in $\mathcal{F}$, ERM selects a point in the "sphere" $\{f : P\mathcal{L}_f = r\}$ that minimizes

$$\inf_{\{f : P\mathcal{L}_f = r\}} \Big(\frac{1}{N} \sum_{i=1}^N (f - f^*)^2(X_i) + \frac{1}{N} \sum_{i=1}^N \xi_i (f - f^*)(X_i) \Big). \tag{6.1}$$

If $r \gtrsim r_N^*$, the first term in (6.1) is essentially $\|f - f^*\|_{L_2}^2$, and when the noise level is high, one expects the minimum to be attained by $r \gtrsim r_N^*$.

Thus, the problem of identifying the minimum is restricted to obtaining sharp upper and lower estimates on the multiplier process. On the other hand, for a low noise level, the minimum is likely to be below r_N^*, where there is an additional source of difficulty – that there is no clear way of estimating the quadratic term, making the problem much harder.

The parameter s_N^* is comparable with the ones used in [23, 4, 24, 25], where the fixed points have been associated with a Dudley's entropy integral for localized sets of the class. In [4], it has been shown 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 improves that result, because the complexity measure used here is based on the gaussian mean width, which is always smaller than Dudley's entropy integral. Moreover, no restrictions on the noise level have been imposed.

In this exposition, we tried to underline that the study of the gaussian regression model requires the analysis of two regimes: high and low noise levels (regardless of the desired estimates on the probability). This reveals the two different sources of statistical complexity that are intrinsic to this model. When estimating f in L_2 from the data $(X_i, Y_i)_{i=1}^N$, one source of an error is that f is known only through its coordinate projection $P_\sigma f = (f(X_i))_{i=1}^N$, while the other is that only a noisy version of this projection is observed. The two, projection and noise, lead to different complexity terms and are associated with two different empirical processes: the quadratic, studied in Theorem 2.1, and the multiplier, studied in Theorem 2.3.

One issue that has been neglected in this article is the geometry of the class, which is as important as its metric complexity.

We believe ERM is an optimal procedure if and only if the class is convex, and the importance of convexity has been obscured by the assumption that the class satisfies a Bernstein condition. However, as we show next, a uniform Bernstein condition implies that the class is convex, at least for classes with an error rate that converges to zero.

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 in an immediate outcome of Theorem C 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 6.1 *Let μ be a probability measure and set X to be a random* *variable distributed according to μ. If $\mathcal{F}$ is a locally compact subset of $L_2(\mu)$, the following statements are equivalent:*

i) for any real valued random variable $Y \in L_2$, there exists a unique minimizer in $\mathcal{F}$ of the functional $\mathbb{E}(Y - f(X))^2$. If f^* is that minimizer, then and for every $f \in \mathcal{F}$,

$$\mathbb{E}\big(f(X) - f^*(X)\big)^2 \leq \mathbb{E}\big((Y - f(X))^2 - (Y - f^*(X))^2\big). \tag{6.2}$$

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

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

$$\begin{aligned}&\mathbb{E}\big((Y - f(X))^2 - (Y - f^*(X))^2\big)\\ &\quad = \|f(X) - f^*(X)\|_2^2 + 2\big\langle f^*(X) - Y, f(X) - f^*(X)\big\rangle\\ &\quad \geq \|f(X) - f^*(X)\|_2^2 .\end{aligned}$$

and $\mathcal{F}$ is $1-$Bernstein.

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,

$$\|f_1^*(X) - f_2^*(X)\|_2^2 \leq B\mathbb{E}\big((Y - f_2^*(X))^2 - (Y - f_1^*(X))^2\big) = 0,$$

and $f_1^* = f_2^*$ in $L_2(\mu)$.

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 [27], (see also [6], 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 and Shahar Mendelson. Empirical minimization. *Probab. Theory Related Fields*, 135(3):311–334, 2006.

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

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

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

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

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

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

[9] 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.

[10] 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].

[11] 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.

[12] 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.

[13] 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.

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

[15] Shahar Mendelson and Gregory Paouris. On the singular values of random matrices. Technical report, To appear in Journal of the European Mathematical Society, 2012.

[16] Shahar Mendelson and Grigoris Paouris. On generic chaining and the smallest singular value of random matrices with heavy tails. *J. Funct. Anal.*, 262(9):3775–3811, 2012.

[17] S. J. Montgomery-Smith. The distribution of Rademacher sums. *Proc. Amer. Math. Soc.*, 109(2):517–522, 1990.

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

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

[20] Mendelson Shahar. On the geometry of subgaussian coordinate projections. Technical report, Technion, 2013.

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

[22] 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.

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

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

[25] 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.

[26] 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.

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

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

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