Learning subgaussian classes : Upper and minimax bounds

Guillaume Lecué[1,3,5] Shahar Mendelson[2,4,6]

September 28, 2016

Preface

Most the results contained in this note have been presented at the SMF meeting, which took place in May 2011; the rest have been obtained shortly after the time of the meeting.

The question we study has to do with the optimality of Empirical Risk Minimization as a learning procedure in a convex class – when the problem is subgaussian. Subgaussian learning problems are a natural object because they are the simplest unbounded learning scenarios. However, an additional reason for studying such problems was that at the time of the SMF meeting, the technical machinery required for the analysis of more heavy-tailed problems was simply not known. Since 2011, significant progress has been made in the understanding of learning problems in heavy-tailed situations [36, 33, 25, 28], though this progress does not make the results presented here obsolete. We show that ERM performed in a convex class is an optimal learning procedure (in a sense that will be clarified) when the learning problem is subgaussian. This happens to be a rather special feature of subgaussian learning problems, and under weaker tail assumptions ERM fails to deliver the optimal accuracy/confidence trade-off at the high level of accuracy we are interested in here.

The results presented here are complemented in [31], which also focuses on subgaussian learning problems and addresses some of the cases that have not been resolved in this note.

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 a real-valued 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 generates 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

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

In the classical statistics setup, one usually assumes that the regression function of Y given X belongs to some particular function space (called a *statistical model*). In contrast, in the learning setup on which

[1]CNRS, CREST, ENSAE, 3 avenue Pierre Larousse, 92240 Malakoff, France
[2]Department of Mathematics, Technion, I.I.T, Haifa 32000, Israel.
[3]Email:guillaume.lecue@ensae.fr
[4]Email:shahar@tx.technion.ac.il
[5]Supported by French National Research Agency (ANR) under the grants Labex Ecodec (ANR-11-LABEX-0047) and by the "Chaire Economie et Gestion des Nouvelles Données", under the auspices of Institut Louis Bachelier, Havas-Media and Paris-Dauphine.
[6]Supported by the Mathematical Sciences Institute, The Australian National University and by the Israel Science Foundation.

we focus here, one is given a function class $\mathcal{F}$ (sometimes, called a model as well), and the goal is to construct a procedure $\hat{f}_N$ that satisfies a *sharp* or *exact oracle inequality* (following [44]; such bounds are called excess risk bounds in [29] and [21]). 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 in (1.1) as small as possible.

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

$$f^* \in \operatorname*{argmin}_{f\in\mathcal{F}} R(f).$$

Note that in (1.1) the performance of the procedure $\hat{f}_N$ is compared to the best performance possible in $\mathcal{F}$, i.e., to the risk of the best element $f^* \in \mathcal{F}$. This exhibits the point of view of Learning Theory, where one wishes to identify a function that is almost as good as the best possible in $\mathcal{F}$, regardless of whether the best function in $\mathcal{F}$ has a small risk. It is different from typical questions in classical Statistics, where a statistical model is given and the risk of an estimator is compared to the one of the regression function (or Bayes rule). The latter are usually called *excess risk bounds* (cf. [30]) and are actually very different from exact oracle inequalities like (1.1) (see, for example, [24] or Chapter 1.3 in [22] for more details on those differences).

The performance of a procedure is measured relative to a set of admissible targets Y in some class of random variables $\mathcal{Y}$. Naturally, one would like to make $\mathcal{Y}$ as large as possible, for example, all random variables Y bounded by 1, all the random variables Y in L_p for some $p > 2$, or a similar weak condition of that flavor.

Definition 1.1 *Let $\hat{f}_N$ be a learning procedure, that is, a map from the set $(\Omega\times\mathbb{R})^N$ into $\mathcal{F}$. Let $0 < \delta_N < 1$ and $\zeta_N > 0$. We say that $\hat{f}_N$ performs with* ***accuracy*** *ζ_N and* ***confidence*** *$1-\delta_N$ relative to the set of admissible targets $\mathcal{Y}$ if for any $Y \in \mathcal{Y}$, $R(\hat{f}_N) \leq \inf_{f\in\mathcal{F}} R(f) + \zeta_N$ with probability larger than $1-\delta_N$, and the probability is measured with respect to the product measure endowed by the joint distribution of X and Y.*

Clearly, while the true risk of f is not known, simply because X and Y are not known, one still has access to its empirical counterpart:

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

Thus, a natural procedure that comes to mind is finding a function in $\mathcal{F}$ that best fits the data: a minimizer of the empirical risk in $\mathcal{F}$. 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 40 years (see, e.g. [49], [29], [21] and references therein), and the main goal has always been to identify connections between the structure of $\mathcal{F}$ and the accuracy and confidence that ERM yields, while trying to minimize the restrictions on $\mathcal{Y}$. Among the natural questions regarding the performance of ERM are:

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

2. Given any $0 < \delta_N < 1/2$, is ERM an optimal procedure 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 [21].

For any $\gamma > 0$, let

$$k_N(r) = \mathbb{E} \sup \Big(\Big| \frac{1}{\sqrt{N}} \sum_{i=1}^N \varepsilon_i (f - f^*)(X_i) \Big| : f \in \mathcal{F}, \|f - f^*\|_{L_2(\mu)} \leq 2r \Big), \tag{1.2}$$

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

Theorem 1.2 *There exist absolute constants c_0, c_1 and $q > 2$ for which the following holds. If $\mathcal{Y}$ consists of functions that are bounded by 1 and $\mathcal{F}$ is a convex class of functions that are bounded by 1, then for any $Y \in \mathcal{Y}$ and every $t > 0$, with probability at least $1 - c_0 \exp(-t)$,*

$$R(\hat{f}) \leq \inf_{f\in\mathcal{F}} R(f) + c_1 \max \Big\{ \big(k_N^*(1/q)\big)^2, \frac{t}{N} \Big\}. \tag{1.3}$$

A result of a similar flavor was obtained in [7]: 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

$$\sigma^* = \inf \left\{ r > 0 : \int_{c_0 r^2}^{c_1 r} \log^{1/2} N\big(\mathcal{F} \cap (2rD), \varepsilon D\big) d\varepsilon \leq c_2 r^2 \sqrt{N} \right\}, \tag{1.4}$$

for absolute constants c_0, c_1, c_2.

The result in [7] 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 (1.4)), $(\sigma^*)^2$ may serve as a residual term.

These two facts rely heavily on the assumption that $\mathcal{F}$ and $\mathcal{Y}$ are bounded in L_∞ and their proofs do not extend beyond the bounded case.

Our aim here is to study unbounded problems and without any assumption on the envelope of $\{\ell(f(X), Y) : f \in \mathcal{F}\}$. The next natural step is the subgaussian framework, 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. (see [8, 11, 10, 9, 19, 20]), all of which have been studied in the subgaussian framework.

Definition 1.3 *Let μ be a probability measure and let X be distributed according to μ. The $\psi_2(\mu)$-norm of a function f is*

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

The space of functions with a finite ψ_2-norm is denoted by $L_{\psi_2} = L_{\psi_2(\mu)}$.

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

Note that for any $f \in L_{\psi_2}$, $\|f\|_{L_2(\mu)} \leq \|f\|_{\psi_2(\mu)}$. A class is a subgaussian class when the reverse inequality holds, and in particular when the ψ_2 and L_2 norms are equivalent on $\mathcal{F}$.

Note that norm equivalence is very different from being bounded. Having such a norm equivalence implies that $|f| \sim \|f\|_{L_2(\mu)}$ on a relatively large event. In contrast, even though a bounded function has a finite ψ_2 norm (by selecting $c \sim \|f\|_{L_\infty}$ in the definition of the ψ_2 norm), the fact that f is bounded does

not mean that $\|f\|_{\psi_2}$ is equivalent to $\|f\|_{L_2}$, nor that $|f| \sim \|f\|_{L_2(\mu)}$ on a relatively large event. Because of the substantial difference between the two notions, one should not expect that learning procedures exhibit the same performance when one assumes that $\mathcal{F}$ is bounded in L_∞ or when the $\psi_2(\mu)$ and $L_2(\mu)$ norms are equivalent on $\mathcal{F}$.

Let us turn to some examples of subgaussian classes of functions. Probably the most interesting collection 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**, if for every $t \in \mathbb{R}^d$, $\|\langle t, \cdot\rangle\|_{\psi_2(\mu)} \leq L\|\langle t, \cdot\rangle\|_{L_2(\mu)}$. The measure μ is **isotropic** if $\|\langle t, \cdot\rangle\|_{L_2(\mu)} = \|t\|_{\ell_2^d}^2$ 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 it is straightforward to verify that for every $a \in \mathbb{R}^d$,

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

where here, and throughout this note we write $u \lesssim v$ if $u \leq c_0 v$ for an absolute constant c_0. Thus, the measure associated with the random vector $X = (x_1, ..., x_d)$ is cL-subgaussian. Also, the measure is clearly isotropic.

Natural 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 in $\mathbb{R}^d$.

• Let $2 \leq p < \infty$ and denote by B_p^d the unit ball of $(\mathbb{R}^d, \| \ \|_{\ell_p})$. The uniform probability 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 $1 \leq i \leq d$ and X is supported in RB_∞^d, then it is L-subgaussian for $L \lesssim R/c$. Indeed, one may show that for every $f \in L_{\psi_2(\mu)}$,

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

for suitable absolute constants c_1 and c_2 (see, for instance, Corollary 1.1.6 in [12]). Thus, it suffices to verify that for every $t \in \mathbb{R}^d$ and every $p \geq 2$,

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

By Khintchine's inequality (see, for example, [26]),

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

Also,

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

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 are omitted.

These examples show 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] and analyzed there in the bounded setup. Before presenting it, recall that the excess loss of f is

$$\mathcal{L}_f(x,y) = \ell(f(x),y) - \ell(f^*(x),y) = (f(x)-y)^2 - (f^*(x)-y)^2 \tag{1.5}$$

and set

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

A rather obvious but very useful observation is that for every $f \in \mathcal{F}$, $P\mathcal{L}_f \geq 0$, 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 f in the set $\{f \in \mathcal{F} : P\mathcal{L}_f \geq \lambda_N\}$,

$$\frac{1}{2}P\mathcal{L}_f \leq P_N\mathcal{L}_f \leq \frac{3}{2}P\mathcal{L}_f. \tag{1.6}$$

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

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

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

with probability at least $1-\delta_N$ (see Theorem 4.4 in [21] for results of a similar flavor).

Remark 1.5 *Note that only the lower estimate in (1.6) is needed for the argument outlined above to work. This observation is the key in the application of the recent works on the small-ball method in learning theory (cf. [36]), which allows one to deal with heavy-tailed scenarios that are far more general than subgaussian problems.*

Just like k_N^* in (1.3) and σ^* in (1.4) – and many other well known estimates on the performance of ERM (e.g. [48, 21, 29]) – the residual term we use is defined in terms of fixed points. Unlike k_N^* and σ^*, the geometric complexity measure we use here is based on gaussian averages associated with localizations of the class. We refer the reader to Chapter 12 in [16] for more details on gaussian processes (in particular to Theorem 12.1.3 for the existence of such a process and to Theorem 12.1.4 for its linearity).

Denote by $\{G_f : f \in \mathcal{F}\}$ the canonical gaussian process indexed by $\mathcal{F}$, that is $\mathbb{E}G_f = 0$ and the covariance is given by the inner product in $L_2(\mu)$: $\mathbb{E}G_f G_h = \langle f, h\rangle_{L_2(\mu)} = \mathbb{E}f(X)h(X)$. Given a set $\mathcal{F}' \subset \mathcal{F}$ we put

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

This supremum is called the *lattice supremum* (see Chapter 2.2 in [26] for more details).

As an example, if $\mathcal{F}' = \{\langle \cdot, t\rangle : t \in T\}$ is a set of linear functionals indexed by $T \subset \mathbb{R}^d$ and X is a random vector in $\mathbb{R}^d$ with covariance matrix Σ then for $G \sim \mathcal{N}(0, \Sigma)$, we simply have

$$\mathbb{E}\|G\|_{\mathcal{F}'} = \mathbb{E}\sup_{t\in T}\langle G, t\rangle.$$

We are now in a position to introduce the two complexity parameters that will serve as residual terms in the exact oracle inequalities satisfied by ERM.

Definition 1.6 *For any $s \geq 0$, set $sD = \{f \in L_2(\mu) : \|f\|_{L_2(\mu)} \leq s\}$ and $\mathcal{F} - \mathcal{F} = \{f - h : f, h \in \mathcal{F}\}$. For every $\eta > 0$, let*

$$s_N^*(\eta) = \inf\left\{s > 0 : \mathbb{E}\|G\|_{sD\cap(\mathcal{F}-\mathcal{F})} \leq \eta s^2\sqrt{N}\right\}, \tag{1.7}$$

and for every $Q > 0$, set

$$r_N^*(Q) = \inf\left\{r > 0 : \mathbb{E}\|G\|_{rD\cap(\mathcal{F}-\mathcal{F})} \leq Qr\sqrt{N}\right\}. \tag{1.8}$$

In what follows we will always assume without mentioning it explicitly that the sets in (1.7) and (1.8) are nonempty (for example, this forces that $Q \geq c/\sqrt{N}$).

There are many situations in which sharp estimates on $\mathbb{E}\|G\|_{rD\cap(\mathcal{F}-\mathcal{F})}$ are known and one can identify the fixed points $s_N^*(\eta)$ and $r_N^*(Q)$. We will present several examples of that kind in Section 4.

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 $\eta = c_1/(L\sigma)$ and $Q = c_2/L^2$.*

1. *If $\sigma \geq c_3 r_N^*(Q)$ then with probability at least $1 - 6\exp(-c_4 N\eta^2(s_N^*(\eta))^2)$,*

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

2. *If $\sigma \leq c_3 r_N^*(Q)$ then with probability at least $1 - 6\exp(-c_4 NQ^2)$,*

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

Hence, with probability at least $1 - 6\exp\left(-c_4 N\min\{\eta^2(s_N^*(\eta))^2, Q^2\}\right)$,

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

We will show in what follows that the parameters involved in the upper bound have very clear roles. r_N^* is an upper estimate 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 $f_1, f_2 \in \mathcal{F}$ using the sample $\mathbb{X} = (X_i)_{i=1}^N$ when $(f_1(X_i))_{i=1}^N = (f_2(X_i))_{i=1}^N$.

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'[1] $f^*(X) - Y$ and the class through the choice of $\eta \sim 1/\sigma$. Thus, beyond a certain noise-level σ, which depends on the 'complexity' of the class $\mathcal{F}$, $s_N^*(c/\sigma)$ becomes the dominant term in the upper bound.

Note that in the free-noise case, $\sigma = 0$, one has $s_N^*(c/\sigma) = 0$. Therefore, the error rate of ERM depends only on $r_N^*(Q)$. Also, when the number of observations N is large enough, one also has $r_N^*(Q) = 0$, leading to exact reconstruction.

Of course, Theorem A would be better justified if one could obtain matching lower bounds, showing that ERM is an optimal procedure for subgaussian problems. To that end, it seems natural to employ minimax theory (see, e.g., [45, 51, 52, 7, 6] for more details on minimax bounds).

What is a reasonable way of identifying a lower bound on the performance of a learning procedure is to see what accuracy and confidence it can guarantee for a minimal set of admissible targets $\mathcal{Y}$, and a natural choice of a minimal set of targets is

$$\mathcal{Y} = \{Y^f : \ Y^f = f(X) + W\} \tag{1.9}$$

for every $f \in \mathcal{F}$ and W that is a centered gaussian random variable that has variance σ^2 and is independent of X. Thus, this minimal set of targets consists of 'independent perturbations' of realizable learning problems, and thus is arguably the smallest set of 'noisy' targets. The minimax rate is (at least) the best accuracy/confidence trade-off that a learning procedure may attain in $\mathcal{F}$ for the set targets (1.9). Our main focus will be on the accuracy/confidence tradeoff for the accuracy level described in Theorem A.

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 of Theorem A; rather, these results are restricted to constant confidence or hold in expectation. To treat the high probability regime, we present a new minimax bound that is based on the gaussian shift theorem (and therefore on the gaussian isoperimetric inequality).

Theorem A′. *There exists an absolute constant c_5 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 $f \in \mathcal{F}$, $[f_0, f] \subset \mathcal{F}$), and let $\mathcal{Y}$ be the set of admissible targets from (1.9). If $\tilde{f}_N$ attains an accuracy ζ_N with the confidence level δ_N for any target $Y^f \in \mathcal{Y}$, then*

$$\zeta_N \geq \min\left\{ c_5 \sigma^2 \frac{\log(1/\delta_N)}{N}, \frac{1}{4}\mathrm{diam}^2(\mathcal{F}, L_2(\mu)) \right\}.$$

Note that no assumption on the underlying measure μ is required in Theorem A′. Moreover, Theorem A′ makes a natural connection between accuracy and confidence: the higher the confidence $1 - \delta_N$ the larger ζ_N must be.

An important outcome of Theorem A and Theorem A′ is that for the set of admissible targets $\mathcal{Y}$ as in (1.9), and as long as the class $\mathcal{F}$ is convex and L-subgaussian, ERM is optimal in the following sense:

Theorem A″. *There exist absolute constants $c_1, ..., c_4$ for which the following holds. Let $\mathcal{F}$ be a convex, L-subgaussian class of functions and consider the set of admissible targets $\mathcal{Y}$ as in (1.9). Set $\eta = c_1/(L\sigma)$*

[1] We keep the terminology from Statistics: the difference between the output variable Y and the target function $f^*(X)$ is called the noise. This coincides with the classical definition of noise in Statistics when f^* is the regression function.

and $Q = c_2/L^2$. *If* $\sigma \geq c_3 r_N^*(Q)$ *then for any target* $Y^f \in \mathcal{Y}$, *the ERM* $\hat{f}$ *satisfies*

$$R(\hat{f}) \leq \inf_{f\in\mathcal{F}} R(f) + (s_N^*(\eta))^2 \quad \text{with probability} \quad 1 - 6\exp\big(-c_4 N\eta^2 (s_N^*(\eta))^2\big).$$

Also, for any learning procedure $\tilde{f}$ *there is some* $f \in \mathcal{F}$ *for which, if given the data generated by the target* Y^f *and* $R(\tilde{f}) \leq \inf_{f\in\mathcal{F}} R(f) + (s_N^*(\eta))^2$ *with probability at least* $1-\delta$, *then*

$$\delta \geq \exp\big(-c_5 N\eta^2 (s_N^*(\eta))^2\big).$$

Thus, up to the constant in the exponent, the upper bound and the lower bound match and the ERM achieves this bound.

The second question we wish to address is what happens when the desired confidence is an absolute constant – for example, when $1-\delta_N$ is, say, $1/2$, but the noise level is nontrivial in the sense that s_N^* dominates r_N^*. We will show that in such a situation, Theorem A is optimal in a minimax sense under some regularity assumptions on $\mathcal{F}$. This complements Theorem A$''$ which proves the optimality of ERM (under no extra structural assumption) in the high probability case – when $\delta_N \sim \exp(-c\eta^2 (s_N^*(\eta))^2 N)$.

To explore the constant confidence regime, let us consider the 'Sudakov analog' of the gaussian-based parameter $s_N^*(\eta)$: recall that by Sudakov's inequality (see, for example, [26]), for any $r > 0$,

$$\sup_{\varepsilon>0} \varepsilon \log^{1/2} N((\mathcal{F}-\mathcal{F}) \cap rD, \varepsilon D) \lesssim \mathbb{E}\|G\|_{rD\cap(\mathcal{F}-\mathcal{F})}. \tag{1.10}$$

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 B. *There exists an absolute constant* c_1 *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 cannot be better than* $c_1(q_N^*(c_1/\sigma))^2$.

Theorem B is known, and may be derived from Theorem 2.5 in [45] or from [51]. The proof presented here is new, and follows the same path as the proof of Theorem A$'$.

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

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 turns out that 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 diameter of* $K(f,\mathbb{X})$ *with respect to the* $L_2(\mu)$ *norm.*

Clearly, if $\sigma = 0$ then for every sample $\mathbb{X}$, ERM selects $\hat{f} \in K(f^*,\mathbb{X})$ and since $Y = f^*(X)$, $R(f) = \|f - f^*\|_{L_2(\mu)}^2$. Thus, $R(\hat{f}) \leq \mathcal{D}^2(f^*,\mathbb{X})$. It is natural to ask whether the reverse direction is true. The following result shows that the largest typical value of $\mathcal{D}(f,\mathbb{X})$ is a constant-probability minimax bound.

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 $\tilde{f}_N$,*

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

with the probability taken with respect to the product measures endowed by $(Y_i^f, X_i)_{i=1}^N$.

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 indexed by T, i.e., $\mathcal{F} = \{\langle t, \cdot\rangle : t \in T\}$. 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 the intersection of the kernel of Γ and T : $\mathcal{D}(0,\mathbb{X}) = \ker(\Gamma) \cap T$. If $f^* = \langle t^*, \cdot\rangle$, one can relate $\mathcal{D}(t^*,\mathbb{X})$ to the Gelfand widths of T (see, e.g., [41] and [40] for more details).

Definition 1.8 *Let T be a convex, centrally-symmetric subset of $\mathbb{R}^d$. The **Gelfand N-width of** T is the smallest ℓ_2^d-diameter of an N-codimensional section of T. In particular,*

$$c_N(T) = \inf\left\{\mathrm{diam}(\ker(\Gamma) \cap T, \ell_2^d) : \Gamma \in L(\mathbb{R}^d, \mathbb{R}^N)\right\},$$

where $L(\mathbb{R}^d, \mathbb{R}^N)$ is the set of all linear operators from $\mathbb{R}^d$ to $\mathbb{R}^N$ and $\mathrm{diam}(V, \ell_2^d) = \sup_{u,v\in V} \|u - v\|_2$.

Hence, for every $t_0 \in T$,

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

and by Theorem C, $c_N(T)/8$ is a lower bound on the minimax rate in the constant confidence regime. Therefore, when $r_N^* \sim c_N(T)$, 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.

It should be noted that although our presentation focuses on oracle inequalities in a given class, oracle inequalities for model selection and regularized procedures can be derived from the isomorphic method in general, specifically, from Theorem 2.8 below. This strategy is rather standard and has been used, for example, in [4, 37], in Chapter 3.6 of [23] or recently in [25]. We will not present results on regularization methods or model selection methods in what follows since those may be easily obtained from results on ERM.

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 $\|x\|_{\ell_p^d} = \left(\sum_{j=1}^d |x_j|^p\right)^{1/p}$. 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}. We also denote by $d_{L_2}(\mathcal{F}')$ the diameter of $\mathcal{F}'$ in $L_2(\mu)$.

The proofs of our main results are presented in the next two sections. We then present several examples of applications of those results, in which the rates established in Theorem A are shown to be sharp in both the high and constant confidence regimes. The final section contains some concluding remarks.

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 $[h, h_0]$ is contained in $\mathcal{H}$.*

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 (cf. [3]).

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

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

Definition 2.2 is far less restrictive than it appears at first glance. Indeed, by the 2-convexity of the L_2 norm, 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 [35] show that for every class $\mathcal{F}$ and every target Y, the Bernstein constant depends only on the distance between Y and the set of targets Z for which the functional $f \to \mathbb{E}(f - Z)^2$ has multiple minimizers in $\mathcal{F}$. Finally, note that if one wishes $\mathcal{F}$ to satisfy a Bernstein condition relative to *every* target Y, it forces $\mathcal{F}$ to be convex in the locally-compact case (see Section 5 for more details).

In what follows, we shall assume that $\mathcal{F} - \mathcal{F}$ is star-shaped around 0 and that $\mathcal{F}$ satisfies the Bernstein condition (2.1).

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

Lemma 2.3 *Assume that $\mathcal{F} - \mathcal{F}$ is star-shaped around 0 and let $\psi : s \geq 0 \to \mathbb{E}\,\|G\|_{sD \cap (\mathcal{F} - \mathcal{F})}$. Then the following holds:*

1. *$\phi : s \to \psi(s)/s$ is non-increasing.*
2. *For $\eta > 0$ and any $s \geq s_N^*(\eta)$, $\psi(s) \leq \eta s^2 \sqrt{N}$, and for any $0 < s < s_N^*(\eta)$, $\psi(s) \geq \eta s^2 \sqrt{N}$.*
3. *Let $Q > \sqrt{\pi/2N}$. For any $r \geq r_N^*(Q)$, $\psi(r) \leq Qr\sqrt{N}$ and for any $0 < r < r_N^*(Q)$, $\psi(r) > Qr\sqrt{N}$.*

A straightforward outcome of Lemma 2.3 which will be used later is as follows:

Lemma 2.4 *Let $c, \sigma, Q > 0$, set $\eta = c/\sigma$ and consider $s_N^*(\eta)$ and $r_N^*(Q)$ as introduced in Definition 1.6.*

1. *If $\sigma \geq (c/Q) r_N^*(Q)$ then $s_N^*(\eta) \geq r_N^*(Q)$, and if $\sigma \leq (c/Q) r_N^*(Q)$ then $s_N^*(\eta) \leq r_N^*(Q)$.*
2. *If $s_N^*(\eta) \geq r_N^*(Q)$ then $\eta s_N^*(\eta) \leq 4Q$.*

The proof of Lemma 2.4 will also be presented in the Appendix.

When considering the parameters $r_N^*(Q)$ and $s_N^*(\eta)$, what may seem 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, the quadratic component and the multiplier component of the excess loss functional scale with $\|f-f^*\|_{L_2(\mu)}$. Indeed, note that

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

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

In contrast, s_N^* is designed for dealing with the multiplier process, originating from the term $(f^*(X)-Y)\cdot(f-f^*)(X)$. To compare the resulting multiplier component with $\mathbb{E}(f-f^*)^2$ (which is the order of magnitude of $N^{-1}\sum_{i=1}^N (f-f^*)^2(X_i)$ when $\|f-f^*\|_{L_2(\mu)} \geq r_N^*$), one has to study

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

and that 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 (see [15, 5, 32]).

Theorem 2.5 *There exist absolute constants c_1 and c_2 for which the following holds. Let $\mathcal{H}$ be an L-subgaussian class. For every $u>0$, with probability at least $1-2\exp(-c_1\min(u^2,u\sqrt{N}))$,*

$$\sup_{h\in\mathcal{H}}\left|\frac{1}{N}\sum_{i=1}^N h^2(X_i)-\mathbb{E}h^2\right| \leq c_2L^2\left(\frac{d\gamma}{\sqrt{N}}+\frac{\gamma^2}{N}+\frac{ud^2}{\sqrt{N}}\right) \tag{2.2}$$

where $d=d_{L_2}(\mathcal{H})$ is the diameter in $L_2(\mu)$ of $\mathcal{H}$ and $\gamma=\mathbb{E}\|G\|_{\mathcal{H}}$.

The following result is a straightforward application of Theorem 2.5 and illustrates the role of $r_N^*(Q)$.

Lemma 2.6 *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 $r>r_N^*(Q)$, then with probability at least $1-2\exp\big(-c_1Q^2N\big)$,*

$$\sup_{h\in rD\cap(\mathcal{F}-f^*)}\left|\frac{1}{N}\sum_{i=1}^N h^2(X_i)-\mathbb{E}h^2\right| \leq c_2QL^2r^2.$$

Proof. The claim is an immediate corollary of Theorem 2.5. Indeed, one simply has to apply Theorem 2.5 to the set $\mathcal{H}=rD\cap(\mathcal{F}-\mathcal{F})$ and to recall that by Lemma 2.3, if $r>r_N^*(Q)$ then $\mathbb{E}\|G\|_{rD\cap(\mathcal{F}-\mathcal{F})}\leq Qr\sqrt{N}$. Therefore, for any $u>0$, with probability at least $1-2\exp(-c_1\min(u^2,u\sqrt{N}))$,

$$\sup_{f,h\in\mathcal{F}:\|f-h\|_{L_2(\mu)}\leq r}\left|\frac{1}{N}\sum_{i=1}^N (f-h)^2(X_i)-\mathbb{E}(f-h)^2\right| \leq c_2L^2\left(\frac{d\gamma}{\sqrt{N}}+\frac{\gamma^2}{N}+\frac{ud^2}{\sqrt{N}}\right)$$

where $d=\mathrm{diam}(rD\cap(\mathcal{F}-\mathcal{F}),L_2)\leq r$ and $\gamma=\mathbb{E}\|G\|_{rD\cap(\mathcal{F}-\mathcal{F})}\leq Qr\sqrt{N}$. Hence, for $u=c_2Q\sqrt{N}$, with probability larger than $1-2\exp(-c_3Q^2N)$,

$$\sup_{f,h\in\mathcal{F}:\|f-h\|_{L_2(\mu)}\leq r}\left|\frac{1}{N}\sum_{i=1}^N (f-h)^2(X_i)-\mathbb{E}(f-h)^2\right| \leq c_4QL^2r^2. \tag{2.3}$$

■

Remark. Using the notation of Lemma 2.5, consider $Q \leq \min\{1/(2c_2L^2), 1\}$. If (2.3) holds then for every $f \in \mathcal{F}$ that satisfies $\|f - f^*\|_{L_2(\mu)} \geq r > r_N^*(Q)$, one clearly has

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

this is evident because for $h = f - f^*$,

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

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

Theorem 2.7 *[Theorem 4.4 in [32]] There exist absolute constants c_1 and c_2 for which the following holds. If $\mathcal{H}$ is an L-subgaussian class and $\xi \in L_{\psi_2}$, then for every $u, w \geq 8$, and every integer $s_0 \geq 1$, with probability at least*

$$1 - 2\exp\big(-c_1u^2 2^{s_0}\big) - 2\exp\big(-c_1Nw^2\big),$$

$$\sup_{h\in\mathcal{H}}\left|\frac{1}{N}\sum_{i=1}^N \xi_i h(X_i) - \mathbb{E}\xi h(X)\right| \leq c_2 Luw\frac{\|\xi\|_{L_{\psi_2}}}{\sqrt{N}}\left(\mathbb{E}\,\|G\|_{\mathcal{H}} + 2^{s_0/2}d_{L_2}(\mathcal{H})\right).$$

Note that in Theorem 2.7 one does not assume that ξ and X are independent, a fact that will be significant in what follows. Indeed, we will apply Theorem 2.7 to $\xi = Y - f^*(X)$ and the class $\mathcal{H} = rD \cap (\mathcal{F} - \mathcal{F})$ for $r > s_N^*(\eta)$. In that case, $d_{L_2}(\mathcal{H}) \leq r$ and $\mathbb{E}\,\|G\|_{\mathcal{H}} \leq \eta r^2\sqrt{N}$, and for $2^{s_0/2} \sim \eta r\sqrt{N}$, we obtain that with probability larger than $1 - 4\exp\big(-c_1N\min\{\eta^2r^2, 1\}\big)$,

$$\sup_{f,h\in\mathcal{F}:\|f-h\|_{L_2(\mu)}\leq r}\left|\frac{1}{N}\sum_{i=1}^N \xi_i(f-h)(X_i) - \mathbb{E}\xi(f-h)(X)\right| \leq c_2L\eta\|\xi\|_{L_{\psi_2}}r^2. \tag{2.4}$$

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

Theorem 2.8 *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 that is B-Bernstein relative to the target Y. Assume that $\mathcal{F} - \mathcal{F}$ is star-shaped around 0 and that $\|Y - f^*(X)\|_{\psi_2} \leq \sigma$. Set $\eta = c_0/(LB\sigma)$ and $Q = c_1/(L^2B)$.*

1. *If $\sigma \geq c_2r_N^*(Q)$, then with probability at least $1 - 6\exp\big(-c_3N\cdot\eta^2(s_N^*(\eta))^2\big)$,*

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

2. *If $\sigma \leq c_2r_N^*(Q)$, then with probability at least $1 - 6\exp\big(-c_3Q^2N/B\big)$,*

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

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

$$\left\| \frac{f - f^*}{(P\mathcal{L}_f)^{1/2}} \right\|^2_{L_2(\mu)} \leq B \quad \text{and} \quad \left\| \frac{f - f^*}{P\mathcal{L}_f} \right\|^2_{L_2(\mu)} \leq \frac{B}{P\mathcal{L}_f} \leq \frac{B}{\lambda}. \tag{2.5}$$

Therefore,

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

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 $\mathcal{H} = (\mathcal{F} - \mathcal{F}) \cap \sqrt{\lambda B} D$. Recall that $\mathcal{F} - \mathcal{F}$ is star-shaped around 0, and by (2.5) one has 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 $\eta = c_0/(LB\sigma)$ and $Q = c_1/(L^2 B)$ for suitable absolute constants c_0 and c_1. Set $r > r^*_N(Q)$ and note that by Lemma 2.4, if $\sigma \geq c_2 r^*_N(Q)$ then $r^*_N(Q) \leq s^*_N(\eta)$ and $\eta s^*_N(\eta) \geq 4Q$, and if $\sigma \leq c_2 r^*_N(Q)$ then $r^*_N(Q) \geq s^*_N(\eta)$; also $c_2 = c_0/LBQ = c_0 L/c_1$.

First, consider the case $\sigma \geq c_2 r^*_N(Q)$. Applying Lemma 2.6 for $\lambda = (s^*_N(\eta))^2/B$, it follows that with probability at least $1 - 2\exp(-c_3 Q^2 N)$

$$\sup_{w\in W_\lambda} \left| \frac{1}{N}\sum_{i=1}^N w^2(X_i) - \mathbb{E}w^2 \right| \leq c_4 Q L^2 B \leq \frac{1}{4},$$

provided that $Q \leq 1/(4c_4 L^2 B)$. Moreover, by (2.4), and because $\eta s^*_N(\eta) \geq 4Q$, one has that with probability at least

$$1 - 4\exp(-c_5 N \eta^2 (s^*_N(\eta))^2),$$

$$\sup_{v\in V_\lambda} \left| \frac{1}{N}\sum_{i=1}^N \xi_i v(X_i) - \mathbb{E}\xi v \right| \leq c_6 LB\sigma\eta \leq \frac{1}{8}$$

as long as $\eta \leq 1/(c_7 LB\sigma)$.

Thus, for any $Q \lesssim 1/L^2 B$ and $\eta \lesssim 1/\sigma LB$, if $\sigma \geq c_2 L r^*_N(Q)$ then with probability at least $1 - 6\exp(-c_8 N \cdot \eta^2 (s^*_N(\eta))^2)$, the following holds: for every $f \in \mathcal{F}$ that satisfies that $P\mathcal{L}_f \geq \lambda$,

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

Next, let us consider that case $\sigma \le c_2 r_N^*(Q)$ which follows a very similar path to the first case. Recall that $r_N^*(Q) \ge s_N^*(\eta)$. Setting $\lambda = (r_N^*(Q))^2/B$, it follows from Lemma 2.6 and (2.4) that with probability at least

$$1 - 2\exp(-cQ^2N) - 4\exp\left(-cN\min\left\{\frac{\eta^2(r_N^*(Q))^2}{\sigma^2 B}, 1\right\}\right), \tag{2.6}$$

$$\sup_{w\in W_\lambda}\left|\frac{1}{N}\sum_{i=1}^N w^2(X_i) - \mathbb{E}w^2\right| \le \frac{1}{4} \text{ and } \sup_{v\in V_\lambda}\left|\frac{1}{N}\sum_{i=1}^N \xi_i v(X_i) - \mathbb{E}\xi v\right| \le \frac{1}{8}$$

as long as $Q \lesssim 1/(L^2B)$ and $\eta \lesssim 1/(\sigma LB)$. The claim now follows because $\eta \sim 1/\sigma$ and by the choice of σ, namely, that $r_N^*(Q)/\sigma \ge c_2$. ∎

Theorem A is an immediate outcome of Theorem 2.8 for $B = 1$ and the isomorphic method described in the introduction.

3 Minimax lower bounds (proofs of Theorem A′, B and C)

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 $(\mathbb{R}\times\Omega)^N$ that generates the sample $(Y_i^f, X_i)_{i=1}^N$.

Let

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

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 ζ_N and has a confidence parameter δ_N, then for every $f \in \mathcal{F}$,

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

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 $\{h \in \mathcal{F} : \mathbb{E}\mathcal{L}_h \le \zeta_N\}$ 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 that performs with accuracy ζ_N for any target of the form Y^f with a confidence parameter $\delta_N < 1/4$, then*

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

Theorem 3.1 leads to the lower estimate in Theorem A″. Indeed, if a procedure performs with confidence $\delta_N = \exp(-c_0\gamma N)$ for some γ, then $\zeta_N \ge c_2\sigma^2\gamma$. Setting $\gamma = c_3\eta(s_N^*(\eta))^2$ for $\eta \sim \sigma^{-1}$ leads to the desired outcome. Thus, combined with Theorem A, ERM achieves the minimax rate $(s_N^*(\eta))^2$ for the confidence established in Theorem A (up to the constants in the exponent).

The proof of Theorem 3.1 requires several preliminary steps.

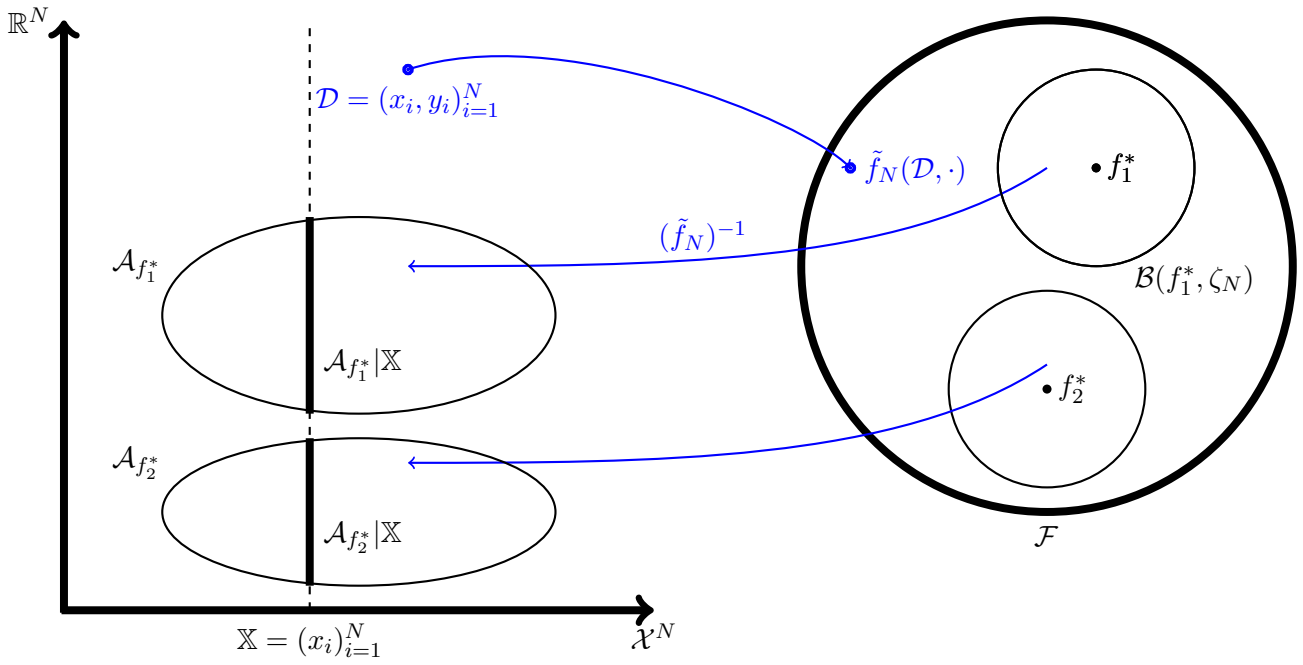

Figure 1: Proof of the minimax lower bounds via the gaussian shift theorem in $\mathbb{R}^N$

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

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

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

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 [13]), $Pr(\rho(\mathbb{X}) \geq x) \geq (\mathbb{E}\rho(\mathbb{X}) - x)/(1-x) \geq 1 - \delta_N/(1-x)$ 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 *[27] 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,

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

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 $\zeta_N \leq d^2_{\mathcal{F}}(L_2)/4$ 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 $4\zeta_N \leq \|v\|^2_{L_2(\mu)} \leq 8\zeta_N$. By Chebyshev's inequality, $Pr\big(\|P_{\mathbb{X}}v\|^2_{\ell_2^N} \geq 4N\|v\|^2_{L_2(\mu)}\big) \leq 1/4$, and thus, for $\mathbb{X} = (X_i)_{i=1}^N$ in a set of μ^N-probability at least $3/4$, $\|P_{\mathbb{X}}v\|_{\ell_2^N} \leq c_1\sqrt{N}\|v\|_{L_2(\mu)}$.

Let

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

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

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

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

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

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

Since $\mathbb{X} \in \Omega_0$, $\|P_{\mathbb{X}}v\|_{\ell_2^N} \leq c_1\sqrt{N}\|v\|_{L_2(\mu)}$; 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 $\nu_{v,\mathbb{X}}(\mathcal{A}_v|\mathbb{X}) \geq 1-\sqrt{\delta_N}$, 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,

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

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 $\zeta_N = a_N^2$ and has a confidence parameter $\delta_N \leq 1/4$. For any $\theta \geq 4$ and $f \in \mathcal{F}$, if Λ is a $2a_N$-separated subset of $\mathcal{F} \cap (f + \theta a_N D)$ then*

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

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,

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

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.

Recall that

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

and by Chebyshev's inequality, $Pr\big(\nu_{v,\mathbb{X}}(\mathcal{A}_v | \mathbb{X}) \geq 3/4\big) \leq 1/3$. Therefore, with μ^N-probability at least $2/3$,

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

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

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

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

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

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

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

Thus, by (3.2), $1 \gtrsim |\Lambda| \exp\big(- c_3 N \theta^2 a^2 / \sigma^2 \big)$, as claimed. ∎

We end this section with the proof of Theorem C, which is presented for a random choice of $(X_1, ..., X_N)$, though the proof for a fixed $(x_1, ..., x_N)$ – the so-called 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 $Pr_{X,V}$ (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,

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

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

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,

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

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

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

■

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 when considering only noise-free targets.

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 than $d_N(\rho)/4$.

On the other hand, a straightforward application of Lemma 2.6 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

In this section, we present two examples in which our results lead to sharp upper and lower minimax bounds, thus showing the optimality (in some minimax sense) of ERM.

4.1 Learning in ρB_1^d

Let $\mathcal{F}$ be the class of linear functionals $\langle \cdot, t \rangle$, indexed by $T = \rho B_1^d$, the unit ball in ℓ_1^d of radius ρ. Assume that μ is an isotropic, L-subgaussian measure on $\mathbb{R}^d$, that $Y \in L_{\psi_2}$ and that $\|Y - f^*(X)\|_{\psi_2} \leq \sigma$.

Since ρB_1^d is centrally symmetric, so is $\mathcal{F}$, and $\mathcal{F} - \mathcal{F} = 2\mathcal{F}$. Thus, the estimates in Theorem A are based only on the behavior of the function $s \to \mathbb{E}\|G\|_{2\mathcal{F} \cap sD}$. And, because the measure μ is isotropic, the canonical gaussian process is given by $t \to G_t = \sum_{i=1}^d g_i t_i$, where $g_1, \ldots, g_d$ are d independent, standard gaussian variables. Moreover, for every $s > 0$, the indexing set $2\mathcal{F} \cap sD$ corresponds to $2\rho B_1^d \cap sB_2^d$. One may show (see, for example, [18]) that for every $2\rho/\sqrt{d} \leq s$,

$$\mathbb{E}\,\|G\|_{2\rho B_1^d \cap sB_2^d} = \mathbb{E} \sup_{t \in 2\rho B_1^d \cap sB_2^d} \left| \sum_{i=1}^d g_i t_i \right| \sim \rho \sqrt{\log(ed \min\{s^2/\rho^2, 1\})},$$

and if $s \leq 2\rho/\sqrt{d}$ then $2\rho B_1^d \cap sB_2^d = sB_2^d$ and

$$\mathbb{E}\,\|G\|_{2\rho B_1^d \cap sB_2^d} = \mathbb{E} \sup_{t \in 2\rho B_1^d \cap sB_2^d} \left| \sum_{i=1}^d g_i t_i \right| \sim s\sqrt{d}.$$

Setting $\eta = c_0/(L\sigma)$ and $Q = c_1/L^2$, it is straightforward to verify that

$$(s_N^*(\eta))^2 \sim_L \begin{cases} \rho\sigma \sqrt{\frac{\log d}{N}} & \text{if } \rho^2 N \leq \sigma^2 \log d, \\ \rho\sigma \sqrt{\frac{1}{N} \log\left(\frac{ed^2\sigma^2}{\rho^2 N}\right)} & \text{if } \sigma^2 \log d \leq \rho^2 N \leq \sigma^2 d^2, \\ \frac{\sigma^2 d}{N} & \text{if } \rho^2 N \geq \sigma^2 d. \end{cases} \qquad (c)$$

Also,

$$(r_N^*(Q))^2 \quad \begin{cases} \sim_L \frac{\rho^2}{N}\log\left(\frac{ed}{N}\right) & \text{if } N \le c_1 d, \\ \lesssim_L \frac{\rho^2}{d} & \text{if } c_1 d \le N \le c_2 d \\ = 0 & \text{if } N > c_2 d, \end{cases}$$

where c_1 and c_2 are constants that depend only on L.

When $N \sim d$, $(r_N^*(Q))^2$ decays rapidly from $(\rho^2/N)\log(ed/N)$ to 0. Thus, when $c_1 d \le N \le c_2 d$ one only has an upper estimate on $(r_N^*(Q))^2$, and we will therefore only consider the cases $N \le c_1 d$ and $N \ge c_2 d$.

Let us present the exact oracle inequalities satisfied by the ERM in ρB_1^d that follow from Theorem A. First, assume that $N \le c_1 d$. If $\sigma \gtrsim r_N^*(Q)$ then $\sigma^2 d^2 \gtrsim N\rho^2$, and

$$(s_N^*(\eta))^2 \sim_L \begin{cases} \rho\sigma\sqrt{\frac{\log d}{N}} & \text{if } \rho^2 N \le \sigma^2 \log d, \\ \rho\sigma\sqrt{\frac{1}{N}\log\left(\frac{ed^2\sigma^2}{\rho^2 N}\right)} & \text{if } \sigma^2 \log d \le \rho^2 N. \end{cases}$$

Setting

$$\delta_N = \begin{cases} 6\exp\left(-\frac{c_4\rho}{\sigma}\sqrt{N\log d}\right) & \text{if } \rho^2 N \le \sigma^2 \log d, \\ 6\exp\left(-\frac{c_4\rho}{\sigma}\sqrt{N\log\left(\frac{ed^2\sigma^2}{\rho^2 N}\right)}\right) & \text{if } \sigma^2 \log d \le \rho^2 N, \end{cases} \tag{4.1}$$

and applying Theorem A, it follows that if $\sigma \ge c_3\rho\sqrt{\log(ed/N)/N}$, then with probability at least $1-\delta_N$,

$$R(\hat{f}) \le \inf_{f\in\mathcal{F}} R(f) + \frac{c_5\rho\sigma}{\sqrt{N}} \begin{cases} \sqrt{\log d} & \text{if } \rho^2 N \le \sigma^2 \log d, \\ \sqrt{\log\left(\frac{ed^2\sigma^2}{\rho^2 N}\right)} & \text{if } \sigma^2 \log d \le \rho^2 N, \end{cases}$$

and if $\sigma \le c_3\rho\sqrt{\log(ed/N)/N}$, then with probability at least $1 - 6\exp(-c_4 N)$,

$$R(\hat{f}) \le \inf_{f\in\mathcal{F}} R(f) + \frac{c_5\rho^2}{N}\log\left(\frac{ed}{N}\right),$$

for constants c_3, c_4, c_5 that depend on L.

In a similar fashion, if $N \ge c_2 d$ then $r_N^* = 0$, and thus, if $\sigma \ne 0$, $\sigma \ge r_N^*$. Therefore, the error rate of ERM is given by s_N^*. When $\sigma = 0$ (the noise-free case) then $s_N^*(\eta) = r_N^*(Q) = 0$ and with probability larger than $1 - 6\exp(-c_4 N)$, $\hat{f} = f^*$, implying exact reconstruction.

Turning to the lower estimate, assume that the set of admissible targets contains every $Y^t = \langle t, x\rangle + W$, for $t \in \rho B_1^d$ and W that is a centered gaussian random variable with variance σ^2 that is independent of X. It follows from Theorem A$''$ that if $\sigma \gtrsim r_N^*(Q)$, ERM is an optimal procedure in the following sense: it achieves the accuracy

$$(s_N^*(c/\sigma))^2 \sim \rho\sigma\sqrt{(1/N)\log(ed^2\sigma^2/(\rho^2 N))}$$

if $\rho^2 N \ge \sigma^2 \log d$, and the accuracy

$$(s_N^*(c/\sigma))^2 \sim \rho\sigma\sqrt{(1/N)\log d}$$

if $(\sigma^2/\log d) \lesssim \rho^2 N \le \sigma^2 \log d$ (note that when $(\sigma^2/\log d) \gtrsim \rho^2 N$ then δ_N in (4.1) is larger than 1 and the probability estimate $1-\delta_N$ is negative).

For a minimax lower bound that holds with constant probability we shall apply Theorem B. To that end, let us bound the covering numbers $\log N(\rho B_1^d \cap 2rB_2^d, rB_2^d)$ from below. First note that

$$N(\rho B_1^d \cap 2rB_2^d, rB_2^d) = N(B_1^d \cap (2r/\rho)B_2^d, (r/\rho)B_2^d)$$

and it suffices to study the covering numbers $N(B_1^d \cap 2rB_2^d, rB_2^d)$ for various choices of r.

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

$$\bigcup_{|I|=k} 2rS^I \subset B_1^d \cap 2rB_2^d.$$

It is a well known fact (see, e.g., [34]) that there is a collection of subsets of $\{1, ..., d\}$ of cardinality k, which will be denoted by $\mathcal{B}$, that is $k/8$ separated in the Hamming distance and for which $\log |\mathcal{B}| \geq c_1 k \log(ed/k)$. Thus, the set $\Lambda = \{(2r)^2 \sum_{i \in I} e_i : I \in \mathcal{B}\}$ is an r-separated subset of $B_1^d \cap 2rB_2^d$ with respect to the ℓ_2^d norm, and for any $1/\sqrt{d} \leq 2r \leq 1$,

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

Moreover, one can prove (via Maurey's empirical method) that this estimate is sharp (see, e.g., [42]). Thus it follows that for any $\rho/\sqrt{d} \leq 2r \leq \rho$,

$$\log N(\rho B_1^d \cap 2rB_2^d, rB_2^d) \sim \frac{\rho^2}{r^2} \log\Big(\frac{edr^2}{\rho^2}\Big).$$

If $2r \leq \rho/\sqrt{d}$ than $\rho B_1^d \cap 2rB_2^d = 2rB_2^d$ and by a volumetric estimate, $\log N(\rho B_1^d \cap 2rB_2^d, rB_2^d) \sim d$. If, on the other hand, $2\rho > 2r \geq \rho$ then $\rho B_1^d \cap 2rB_2^d = \rho B_1^d$ and since $\log N(\rho B_1^d, rB_2^d) \sim \log(edr^2/\rho^2) \sim \log d$ (which is evident from the argument used above), then $\log N(\rho B_1^d \cap 2rB_2^d, rB_2^d) \sim \log d$. Finally, when $2r \geq 2\rho$, $\log N(\rho B_1^d \cap 2rB_2^d, rB_2^d) = 0$.

Therefore,

$$(q_N^*(c_0/\sigma))^2 \sim_L \begin{cases} \rho^2 & \text{if } \rho^2 N \leq \sigma^2 \log d, \\ \rho\sigma\sqrt{\frac{1}{N}\log\left(\frac{ed^2\sigma^2}{\rho^2 N}\right)} & \text{if } \sigma^2 \log d \leq \rho^2 N \leq \sigma^2 d^2, \\ \frac{\sigma^2 d}{N} & \text{if } \rho^2 N \geq \sigma^2 d. \end{cases} \qquad (c')$$

We conclude that when $\sigma \gtrsim r_N^*(Q)$ (and in particular, when $\sigma^2 d^2 \gtrsim N\rho^2$), and if $\rho^2 N \geq \sigma^2 \log d$, then $q_N^*(c_0/\sigma) \sim s_N^*(\eta)$. This estimate also exhibits that Sudakov's inequality for the set $\rho B_1^d \cap 2rB_2^d$ is sharp at the scale $\varepsilon = r$ in the following sense: for every $0 < r < \rho$,

$$r\sqrt{\log N(\rho B_1^d \cap 2rB_2^d, rB_2^d)} \sim \mathbb{E}\,\|G\|_{2\rho B_1^d \cap rB_2^d}.$$

Therefore, by Theorem B, one has that if $\sigma \gtrsim r_N^*(Q)$ and if the set of admissible targets contains every $Y^t = \langle X, t\rangle + W$ as above, then the minimax rate in the constant confidence regime is $(s_N^*(\eta))^2$ and that ERM is optimal procedure when $\rho^2 N \geq \sigma^2 \log d$.

Note that when $\rho^2 N \leq \sigma^2 \log d$ the estimates on $(q_N^*(c_0/\sigma))^2$ and on $(s_N^*(\eta))^2$ do not coincide. And, it turns out that if one extends the set of admissible targets, ERM cannot perform with a better accuracy than $\sim (s_N^*(\eta))^2$ in this range. Indeed, consider the one dimensional case $d = 1$ and a target Y defined as follows: the marginal law of Y given X is

$$Y = \begin{cases} \sigma X & \text{with probability } 1/2 + \delta \\ -\sigma X & \text{with probability } 1/2 - \delta \end{cases} \qquad (4.2)$$

for δ that will be specified later, and X that is distributed uniformly in $\{-1, 1\}$. The corresponding class of one-dimensional linear functionals is $\mathcal{F} = \{f_t = tx \ : \ -\rho \leq t \leq \rho\}$.

It is straightforward to verify that for every $t \in [-\rho, \rho]$,

$$R(t) := R(f_t) = (\sigma^2 + t^2) - 4t\delta\sigma,$$

and if $2\sigma\delta \geq \rho$ then the minimizer of $R(t)$ in $[-\rho, \rho]$ is $t = \rho$.

Next, let us identify the minimizer of the empirical risk $R_N(t) = N^{-1}\sum_{i=1}^N (Y_i - tX_i)^2$. Given the sample $(X_i, Y_i)_{i=1}^N$, let $J = \{i : Y_i = \sigma X_i\}$. Observe that for every $t \in [-\rho, \rho]$,

$$\sum_{i=1}^N (Y_i - tX_i)^2 = \sum_{i\in J}(\sigma - t)^2 X_i^2 + \sum_{i\in J^c}(-\sigma - t)^2 X_i^2 = (\sigma - t)^2|J| + (\sigma + t)^2|J^c|.$$

Hence, if $|J^c| \geq |J|$ then the empirical minimizer satisfies $\hat{t} \leq 0$. By the choice of Y, the random variable $Z = \mathbb{1}_{\{Y=-\sigma X\}}|X$ has mean $1/2 - \delta$ and variance $\tau^2 = 1/4 - \delta^2$. Given $X_1, ..., X_N$, let $Z_i = \mathbb{1}_{\{Y_i=-\sigma X_i\}}|X_i$ and note that $|J^c| = \sum_{i=1}^N Z_i$. It follows from the Berry-Esseen Theorem that if $\delta\sqrt{N}/\tau \leq c_1$ then with probability at least $1/4$, $|J^c| \geq N/2$. And to ensure that $\delta\sqrt{N}/\tau \leq c_1$ it suffices to select $\delta = c_2/\sqrt{N}$. All that remains is to estimate the excess risk of $\hat{t}$, which clearly satisfies

$$R(\hat{t}) - R(t^*) \geq c_3\sigma\rho\delta = \frac{c_4\rho\sigma}{\sqrt{N}}. \tag{4.3}$$

Thus, when $\sigma\delta \geq \rho$ (i.e., when $\rho^2 N \lesssim \sigma^2$), the best accuracy that ERM can achieve with constant probability is $\sim (s_N^*(\eta))^2$.

Finally, turning to the low noise regime ($\sigma \lesssim r_N^*(Q)$), one can show that the rate $(r_N^*(Q))^2$ is actually sharp. Recall that by Theorem C it suffices to show that the Gelfand N-width of ρB_1^d satisfies $c_N(\rho B_1^d) \sim r_N^*$. By a result due to Garanaev and Gluskin [17], when $d \geq N$ one has

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

and $c_N(\rho B_1^d) = 0$ when $d < N$. Therefore, $c_N(\rho B_1^d) \sim r_N^*(Q)$ when either $N \leq c_1 d$ or $N > c_2 d$. In particular, when $0 \leq \sigma \lesssim r_N^*(Q)$, the minimax rate is $(r_N^*(Q))^2$ and it is achieved by the ERM.

4.2 Low-rank matrix inference via the max-norm

In this section, the goal is to estimate the real-valued 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 "a 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 [43] 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, and in particular,

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

Let $A^* \in \operatorname{argmin}_{A\in\mathcal{B}_{max}} \mathbb{E}(Y - \langle A, X\rangle)^2$ be a minimizer of the risk in $\mathcal{B}_{max}$ and set $\sigma = \|Y - \langle X, A^*\rangle\|_{\psi_2}$. 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., [38]) 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 $s > 0$,

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

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 [43], if

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

then

$$s\log^{1/2} N(\mathcal{B}_{max}\cap s\sqrt{pq}B_F, s\sqrt{pq/2}B_F) \gtrsim \sqrt{p+q}. \tag{4.4}$$

Hence, it follows from Sudakov's inequality that in that range of s,

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

and

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

as long as both are smaller than 1 and larger than $1/\min\{p,q\}$; that is, when $p+q \lesssim N \lesssim pq$, $p+q \lesssim \sigma^2 N$ and $\sigma^2(p+q)\min(p,q)^2 \gtrsim N$.

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 the 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 $\delta_N = \exp(-c_1\sqrt{N(p+q)}/\sigma)$. Moreover, by Theorem B and (4.4), any procedure with confidence parameter $\delta_N \le 1/4$ has accuracy $\zeta_N \gtrsim \sigma\sqrt{\frac{p+q}{N}}$, matching the upper bound.

5 Concluding remarks

Subgaussian classes are the first family of unbounded classes one is likely to consider, and it turns out that just like bounded classes, the study of subgaussian learning problems may be carried out using a two-sided concentration argument. Unfortunately, this is as far as concentration goes: the substantial technical machinery needed for the proof of Theorem A is not true beyond the subgaussian framework, and the analysis of more 'heavy-tailed' problems requires a totally different machinery (see [36, 33, 25, 28]). Moreover, in more heavy-tailed situations, ERM does not attain the optimal accuracy/confidence tradeoff.

The results presented in this article are sharp in many cases but not in every case. First, in the 'high probability' range, Theorem A$''$ shows that when $\sigma \gtrsim r_N^*$ the performance of ERM is optimal in the minimax sense. However, if $\sigma \lesssim r_N^*$, the estimate we present happens 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 $\sim (r_N^*)^2$. This gap is filled (almost completely) in [31].

In the constant probability regime the picture presented here is even less complete. For example, in 'noisy situations' – 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 even for $\mathcal{F} = \{\langle t, \cdot \rangle : t \in B_p^d\}$, when $1 + 1/\log d < p < 2$. Again, this gap was addressed in [31] – at least when considering a class of admissible targets of the form $Y^f = f(X) + W$.

The case of linear functional in $\mathbb{R}^d$ is a good indication to what our estimates give in general: if X is L-subgaussian then when considering targets of the form $Y^t = \langle X, t \rangle + W$ for a centered gaussian variable W that is independent of X, and $t \in T$, ERM achieves the accuracy

$$\max\{(s_N^*(c/\sigma))^2, (r_N^*(Q))^2\}$$

as long as T is convex and centrally-symmetric. No procedure can outperform this rate, say with confidence at least $3/4$ provided that:

1. $q_N^* \log^{1/2} N(T \cap 2q_N^* B_2^d, q_N^* B_2^d) \sim \mathbb{E} \|G\|_{2T \cap q_N^* B_2^d}$ – meaning that there is no gap in Sudakov's inequality at scale $\varepsilon = q_N^*$.
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 [39]).

Let us mention once again that a complete characterization of the minimax rate in this case was recently established in [31], and the optimal procedure happens to be a minor modification of ERM: it is ERM performed in an appropriate net in T.

The parameter s_N^* may be compared with the fixed points used in [46, 7, 47, 48, 2]. In all those cases, the fixed points are associated with Dudley's entropy integral for the localized class, rather than with the localized gaussian process; as such, the resulting bounds are always weaker than ours. For example, the results in [7] which deal with the same situation as Theorem A$''$ 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 $Y^f = f(X) + W$ 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 $f \in \mathcal{F}$,

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

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, $\big\langle f(X) - f^*(X), Y - f^*(X)\big\rangle \leq 0$ for every $f \in \mathcal{F}$, and

$$\mathbb{E}\big((Y - f(X))^2 - (Y - f^*(X))^2\big) = \|f(X) - f^*(X)\|_{L_2}^2 + 2\big\langle f^*(X) - Y, f(X) - f^*(X)\big\rangle \geq \|f(X) - f^*(X)\|_{L_2}^2 .$$

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,

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

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 [50], (see also [14], Chapter 12), $\mathcal{F}$ is convex. ∎

A Additional proofs

First note that the canonical gaussian process we are interested in is a restriction of the isonormal process on $L_2(\mu)$ to a subset (see Section 12 in [16]). In particular, it inherits the linearity of the isonormal process – a fact we shall use below.

Proof of Lemma 2.3. Fix $s_1 > s_2 > 0$ and $f, h \in \mathcal{F}$. Assume that $s_2 \leq \|f - h\|_{L_2(\mu)} \leq s_1$ and observe that since $\mathcal{F} - \mathcal{F}$ is star-shaped around 0 and $0 < s_2/\|f - h\|_{L_2(\mu)} < 1$, it follows that

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

Therefore,

$$G_{f-h} = \frac{\|f - h\|_{L_2(\mu)}}{s_2} G_u \leq (s_1/s_2) \sup_{w \in s_2 D \cap (\mathcal{F} - \mathcal{F})} G_w. \tag{A.1}$$

Since (A.1) clearly holds if $\|f - h\|_{L_2(\mu)} \leq s_2$, by taking the supremum over all possible choices of $f - h \in s_1 D \cap (\mathcal{F} - \mathcal{F})$,

$$\sup_{w \in s_1 D \cap (\mathcal{F} - \mathcal{F})} G_w \leq (s_1/s_2) \sup_{w \in s_2 D \cap (\mathcal{F} - \mathcal{F})} G_w,$$

which is equivalent to $\psi(s_1)/s_1 \leq \psi(s_2)/s_2$; therefore, ϕ is non-increasing on $(0, +\infty)$.

The two other parts of the claim can be established using a similar argument and their proofs are omitted. ∎

Proof of Lemma 2.4. First, assume that $\sigma \geq (c/Q)r_N^*(Q)$. Let $r < r_N^*(Q)$ and note that by Lemma 2.3,

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

Hence, if $(Q\sigma)/rc \geq 1$ then $\mathbb{E}\|G\|_{rD\cap(\mathcal{F}-\mathcal{F})} \geq \frac{c}{\sigma}r^2\sqrt{N}$, implying that $r \leq s_N^*(c/\sigma)$. But $(Q\sigma)/rc \geq 1$ is equivalent to $\sigma \geq (c/Q)r$, which holds for any $r < r_N^*(Q)$.

For the reverse direction, let $\sigma \leq (c/Q)r_N^*(Q)$ and set $r > r_N^*(Q)$. Thus, by Lemma 2.3,

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

Hence, if $Q/r \leq c/\sigma$ then $r \geq s_N^*(c/\sigma)$. But $Q/r \leq c/\sigma$ if $\sigma \leq (c/Q)r$, which clearly holds. ∎

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] Withold Bednorz. Concentration via chaining method and its applications. Technical report, University of Warsaw, 2013. ArXiv:1405.0676.

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

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

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

[9] 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.

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

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

[12] Djalil Chafaï, Olivier Guédon, Guillaume Lecué, and Alain Pajor. *Interactions between compressed sensing random matrices and high dimensional geometry*, volume 37 of *Panoramas et Synthèses [Panoramas and Syntheses]*. Société Mathématique de France, Paris, 2012.

[13] 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.

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

[15] S. Dirksen. Tail bounds via generic chaining. *Electron. J. Probab.*, 20:no. 53, 29, 2015.

[16] R. M. Dudley. *Real analysis and probability*, volume 74 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 2002. Revised reprint of the 1989 original.

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

[18] 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.

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

[20] 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.

[21] 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].

[22] Guillaume Lecué. *Interplay between concentration, complexity and geometry in learning theory with applications to high dimensional data analysis*. Habilitation à Diriger des Recherches Université. Paris-Est Marne-la-vallée, December 2011.

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

[24] Guillaume Lecué and Shahar Mendelson. General nonexact oracle inequalities for classes with a subexponential envelope. *Ann. Statist.*, 40(2):832–860, 2012.

[25] Guillaume Lecué and Shahar Mendelson. Regularization and the small-ball method i: sparse recovery. Technical report, CNRS, Ecole Polytechnique and Technion, 2015.

[26] 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.

[27] 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.

[28] G. Lugosi and S. Mendelson. Risk minimization by median-of-means tournaments. Technical report, 2016.

[29] 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.

[30] 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.

[31] S. Mendelson. "local" vs. "global" parameters – breaking the gaussian complexity barrier. Technical report, Technion, 2014. Annals of Statistics, to appear.

[32] S. Mendelson. Upper bounds on product and multiplier empirical processes. Technical report, Technion, I.I.T., 2014. To appear in Stochastic Processes and their Applications.

[33] S. Mendelson. Learning without concentration for general loss functions. Technical report, 2015.

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

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

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

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

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

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

[40] Allan Pinkus. *n-widths in approximation theory*, volume 7 of *Ergebnisse der Mathematik und ihrer Grenzgebiete (3) [Results in Mathematics and Related Areas (3)]*. Springer-Verlag, Berlin, 1985.

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

[42] Carsten Schütt. Entropy numbers of diagonal operators between symmetric Banach spaces. *J. Approx. Theory*, 40(2):121–128, 1984.

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

[44] 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.

[45] 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.

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

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

[48] 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.

[49] 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.

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

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

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