Electronic Journal of Statistics
Vol. 14 (2020) 3563–3605
ISSN: 1935-7524
https://doi.org/10.1214/20-EJS1754

ERM and RERM are optimal estimators for regression problems when malicious outliers corrupt the labels

Geoffrey Chinot

ENSAE, 5 avenue Henri Chatelier, 91120, Palaiseau, France
e-mail: geoffrey.chinot@ensae.fr

Abstract: We study Empirical Risk Minimizers (ERM) and Regularized Empirical Risk Minimizers (RERM) for regression problems with convex and L-Lipschitz loss functions. We consider a setting where $|\mathcal{O}|$ malicious outliers contaminate the labels. In that case, under a local Bernstein condition, we show that the L_2-error rate is bounded by $r_N + AL|\mathcal{O}|/N$, where N is the total number of observations, r_N is the L_2-error rate in the non-contaminated setting and A is a parameter coming from the local Bernstein condition. When r_N is minimax-rate-optimal in a non-contaminated setting, the rate $r_N + AL|\mathcal{O}|/N$ is also minimax-rate-optimal when $|\mathcal{O}|$ outliers contaminate the label. The main results of the paper can be used for many non-regularized and regularized procedures under weak assumptions on the noise. We present results for Huber's M-estimators (without penalization or regularized by the ℓ_1-norm) and for general regularized learning problems in reproducible kernel Hilbert spaces when the noise can be heavy-tailed.

MSC2020 subject classifications: Primary 62G35; secondary 62G08.
Keywords and phrases: regularized empirical risk minimizers, outliers, robustness, minimax-rate-optimality.

Received December 2019.

1. Introduction

Let $(\Omega, \mathcal{A}, P)$ be a probability space where $\Omega = \mathcal{X} \times \mathcal{Y}$. $\mathcal{X}$ denotes the measurable space of the inputs and $\mathcal{Y} \subset \mathbb{R}$ the measurable space of the outputs. Let (X, Y) be a random variable taking values in Ω with joint distribution P and let μ be the marginal distribution of X. Let F denote a class of functions $f : \mathcal{X} \mapsto \mathcal{Y}$. A function f in F is named a *predictor*. The function $\ell : \mathcal{Y} \times \mathcal{Y} \mapsto \mathbb{R}^+$ is a loss function such that $\ell(f(x), y)$ measures the quality of predicting $f(x)$ while the true answer is y. For any function f in F we write $\ell_f(x, y) := \ell(f(x), y)$. For any distribution Q on Ω and any function $f : \mathcal{X} \times \mathcal{Y} \mapsto \mathbb{R}$ we write $Qf = \mathbb{E}_{(X,Y)\sim Q}[f(X, Y)]$. Let $f \in F$, the risk of f is defined as $R(f) := P\ell_f = \mathbb{E}_{(X,Y)\sim P}[\ell(f(X), Y)]$. A prediction function with minimal risk is called an *oracle* and is defined as $f^* \in \operatorname{argmin}_{f\in F} P\ell_f$. For the sake of simplicity, it is assumed that the *oracle* f^* exists and is unique. The joint distribution P of (X, Y) being unknown, computing f^* is impossible. Instead one is given a dataset $\mathcal{D} = (X_i, Y_i)_{i=1}^N$ of N random variables taking values in $\mathcal{X} \times \mathcal{Y}$.

Regularized empirical risk minimization is the most widespread strategy in machine learning to estimate f^*. There exists an extensive literature on its generalization capabilities [56, 31, 30, 35, 18]. However, in the past few years, many papers have highlighted its severe limitations. One main drawback is that a single outlier (X_o, Y_o) (in the sense that nothing is assumed on (X_o, Y_o)) can deteriorate the performances of RERM. Consequently, RERM is in general not robust to outliers. However, the question below naturally follows:

What happens if only the labels $(Y_i)_{i=1}^N$ are contaminated?

For example, in [19], the authors raised the question whether it is possible to attain optimal rates of convergence in outlier-robust sparse regression using regularized empirical risk minimization. They consider the model, $Y_i = \langle X_i, t^* \rangle + \epsilon_i$, where X_i is a Gaussian random vector in $\mathbb{R}^p$ with a covariance matrix satisfying the Restricted Eigenvalue condition [55] and t^* is assumed to be s-sparse. The non-contaminated noise is $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$, while it can be anything when malicious outliers contaminate the labels. The authors prove that the ℓ_1-penalized empirical risk minimizer based on the Huber's loss function has an error rate of the order

$$\sigma\left(\sqrt{s\frac{\log(p/\delta)}{N}} + \frac{|\mathcal{O}|\log(N/\delta)}{N}\right) , \tag{1}$$

with probability larger that $1-\delta$, where $|\mathcal{O}|$ is the number of outliers. Up to a logarithmic factor, the RERM associated with the Huber loss function for the problem or sparse-linear regression is minimax-rate-optimal when $|\mathcal{O}|$ malicious outliers corrupt the labels.

1.1. Setting

In this paper, we consider a setup where $|\mathcal{O}|$ outputs can be contaminated. More precisely, let $\mathcal{I} \cup \mathcal{O}$ denote an unknown partition of $\{1, \cdots, N\}$ where $\mathcal{I}$ is the set of **informative** data and $\mathcal{O}$ is the set of **outliers**.

Assumption 1. *$(X_i, Y_i)_{i\in\mathcal{I}}$ are i.i.d with a common distribution P. The random variables $(X_i)_{i=1}^N$ are i.i.d with law μ.*

Nothing is assumed on the labels $(Y_i)_{i\in\mathcal{O}}$. They can be adversarial outliers making the learning as hard as possible. Without knowing the partition $\mathcal{I} \cup \mathcal{O}$, the goal is to construct an estimator $\hat{f}_N$ that approximates/estimates the *oracle* f^*. A way of measuring the quality of an estimator is via the **error rate** $\|\hat{f}_N - f\|_{L_2(\mu)}$ or **the excess risk** $P\mathcal{L}_{\hat{f}} := P\ell_{\hat{f}_N} - P\ell_{f^*}$. We assume the following:

Assumption 2. *The class F is convex.*

A natural idea to construct robust estimators when the labels might be contaminated is to consider L-Lipschitz loss functions [27, 26] that is a losses satisfying

$$\left|\ell\big(f(x), y\big) - \ell\big(g(x), y\big)\right| \leq L|f(x) - g(x)| \ ,$$

for every $x, y \in \mathcal{X} \times \mathcal{Y}$ and $f, g \in F$. Moreover, for computational purposes we also focus on convex loss functions [53].

Assumption 3. *There exists $L > 0$ such that, for any $y \in \mathcal{Y}$, $\ell(\cdot, y)$ is L-Lipschitz and convex.*

Recall that the Empirical Risk Minimizer (ERM) and the Regularized Empirical Risk Minimizer (RERM) are respectively defined as

$$\hat{f}_N \in \operatorname*{argmin}_{f \in F} \frac{1}{N} \sum_{i=1}^{N} \ell(f(X_i), Y_i) \ ,$$

and

$$\hat{f}_N^{\lambda} \in \operatorname*{argmin}_{f \in F} \frac{1}{N} \sum_{i=1}^{N} \ell(f(X_i), Y_i) + \lambda \|f\| \ ,$$

where $\lambda > 0$ is a tuning parameter and $\|\cdot\|$ is a norm. Under Assumptions 2 and 3 the ERM and RERM are computable using tools from convex optimization [8].

1.2. Our contributions

As exposed in [19], in a setting where $|\mathcal{O}|$ outliers contaminate only the labels, RERM with the Huber loss function is (nearly) minimax-rate-optimal for the sparse-regression problem when the noise and design of non-contaminated data are both Gaussian. It leads to the following questions:

1. *Is the (R)ERM optimal for other loss functions and regression problems than the sparse-regression when malicious outliers corrupt the labels?*
2. *Is the Gaussian assumption on the noise necessary?*

Based on the previous works [18, 16, 17, 1], we study both ERM and RERM for regression problems when the penalization is a norm and the loss function is simultaneously convex and Lipschitz (Assumption 3) and show that:

> In a framework where $|\mathcal{O}|$ outliers may contaminate the labels, under a local Bernstein condition, the error rate for both ERM and RERM can be bounded by
>
> $$r_N + AL\frac{|\mathcal{O}|}{N} \ , \tag{2}$$
>
> where N is the total number of observations, L is the Lipschitz constant from Assumption 3, r_N is the error rate in a non-contaminated setting and A is a constant coming from the local Bernstein condition.

When the proportion of outliers $|\mathcal{O}|/N$ is smaller than the error rate $r_N/(AL)$, both ERM and RERM behave as if there was no contamination. The result holds for any loss function that is simultaneously convex and Lipschitz and not only for the Huber loss function. We obtain theorems that can be used

for many well-known regression problems including structured high-dimensional regression (see Section 3.3), non-parametric regression (see Section 3.4) and matrix trace regression (using the results from [1]). As a proof of concept, for the problem of sparse-linear regression, we improve the rate (1), even when the noise may be heavy-tailed. The next question one may ask is the following:

2. *Is the general bound* (2) *minimax-rate-optimal when* $|\mathcal{O}|$ *malicious outliers may corrupt the labels?*

To answer question 2, we use the results from [13]. The authors established a general minimax theory for the ε-contamination model defined as $P_{(\varepsilon,\theta,Q)} = (1-\varepsilon)P_\theta + \varepsilon Q$ given a general statistical experiment $\{P_\theta, \theta \in \Theta\}$. A deterministic proportion ε of outliers with same the distribution Q contaminates P_θ. When $Y = f_\theta(X) + \epsilon$, $\theta \in \Theta$ and following the idea of [13], we show in Section B that the lower minimax bounds for regression problems in the ε-contamination model are the same when

1. Both the design X and the response variable Y are contaminated.
2. Only the response variable Y is contaminated.

Since in our setting, outliers do not necessarily have the same distribution Q, it is clear that a lower bound on the risk in the ε-contamination model implies a lower bound when $|\mathcal{O}| = \varepsilon N$ arbitrary outliers contaminate the dataset As a consequence, for regression problems, minimax-rate-optimal bounds in the ε-contamination model are also optimal when $N\varepsilon$ malicious outliers corrupt only the labels.

> When the bound (2) is minimax-rate-optimal for regression problems in the ε-contamination model with $\varepsilon = |\mathcal{O}|/N$, then it is also minimax-rate-optimal when $|\mathcal{O}|$ malicious outliers corrupt the labels.

The results are derived under the local Bernstein condition introduced in [18]. This condition enables to obtain fast rates of convergence, even when the noise is heavy-tailed. As a proof of concept, we study Huber's M-estimators in $\mathbb{R}^p$ (non-penalized or regularized by the ℓ_1-norm) when the noise is Cauchy. In these cases, the error rates are respectively

$$L\left(\sqrt{\frac{\mathrm{Tr}(\Sigma)}{N}} + \frac{|\mathcal{O}|}{N}\right) \quad \text{and} \quad L\left(\sqrt{\frac{s\log(p)}{N}} + \frac{|\mathcal{O}|}{N}\right) ,$$

where Σ is the covariance matrix of the design X. We also study learning problems in general Reproducible Kernel Hilbert Space (RKHS). We derive error rates depending on the spectrum of the integral operator as in [46, 43, 10] without assumption on the design and when the noise is heavy-tailed (see section 3.3).

Remark 1. *The general results hold for any Lipschitz and convex loss function. However, for the sake of simplicity, we present applications only for the Huber*

loss function. Using results from [18], it is possible to apply our main results to other Lipschitz loss functions.

1.3. Related literature

Regression problems with possibly heavy-tailed data or outliers cannot be handled by classical least-squares estimators. This lack of robustness of least-squares estimators gave birth to the theory of robust statistics developed by Peter Huber [27, 26, 28], John Tukey [51, 52] and Frank Hampel [23, 24]. The most classical alternatives to least-squares estimators are M-estimators. They consist in replacing the quadratic loss function by other loss functions, less sensitive to outliers [41, 58].

Robust statistics has attracted a lot of attention in the past few years both in the computer science and the statistical communities. For example, although estimating the mean of a random vector in $\mathbb{R}^p$ is one of the oldest and fundamental problems in statistics, it is still a very active research area. Surprisingly, optimal bounds for heavy-tailed data have been obtained only recently [39]. However, their estimator cannot be computed in practice. Using semi-definite programming (SDP), [25] obtained optimal bounds achievable in polynomial time. In a recent works, still using SDP, [32] designed an algorithm computable in nearly linear time, while [37] developed the first tractable optimal algorithm not based on the SDP.

In the meantime, another recent trend in robust statistics has been to focus on finite sample risk bounds that are minimax-rate-optimal when $|\mathcal{O}|$ outliers contaminate the dataset. For example, for the problem of mean estimation, when $|\mathcal{O}|$ malicious outliers contaminate the dataset and the non-contaminated data are assumed to be sub-Gaussian, the optimal rate (measured in Euclidean norm) scales as $\sqrt{p/N}+|\mathcal{O}|/N$. In [13], the authors developed a general analysis for the ε-contamination model. In [12], the same authors proposed an optimal estimator when $|\mathcal{O}|$ outliers with the same distribution contaminate the data. In [22], the authors focused on the problem of high-dimensional linear regression in a robust model where an ε-fraction of the samples can be adversarially corrupted. Robust regression problems have also been studied in [15, 21, 38, 6]. Above-mentioned articles assume corruption both in the design and the label. In such a corruption setting ERM and RERM are known to be poor estimators.

In [19], the authors raised the question whether it is possible to attain optimal rates of convergence in sparse regression using regularized empirical risk minimization when a proportion of malicious outliers contaminate only the labels. They studied ℓ_1 penalized Huber's M-estimators. This work is the closest to our setting and reveals that when only the labels are contaminated, simple procedures, such as penalized Huber's M estimators, still perform well and are minimax-rate-optimal. Their proofs rely on the fact that non-contaminated data are Gaussian. Our approach is different, more general and uses the control of stochastic processes indexed by the class F.

Other alternatives to be robust both for heavy-tailed data and outliers in regression have been proposed in the literature such as Median Of Means (MOM)

based methods [33, 34, 18]. However such estimators are difficult to compute in practice and can lead to sub-optimal rates. For instance, for sparse-linear regressions in $\mathbb{R}^p$ with a sub-Gaussian design, MOM-based estimators have an error rate of the order $L\big(\sqrt{s\log(p)/N}+\sqrt{|\mathcal{O}|/N}\big)$ (see [18]) while the optimal dependence with respect to the number of outliers is $L\big(\sqrt{s\log(p)/N}+|\mathcal{O}|/N\big)$. Finally, there was a recent interest in robust iterative algorithms. It was shown that robustness of stochastic approximation algorithms can be enhanced by using robust stochastic gradients. For example, based on the geometric median [44], [14] designed a robust gradient descent scheme. More recently, [45] showed that a simple truncation of the gradient enhances the robustness of the stochastic mirror descent algorithm.

The paper is organized as follows. In Section 2, we present general results for non-regularized procedures with a focus on the example of the Huber's M-estimator in $\mathbb{R}^p$. Section 3 gives general results for RERM that we apply to ℓ_1-penalized Huber's M-estimators with isotropic design and regularized learning in RKHS. Section A presents simple simulations to illustrate our theoretical findings. In section B, we show that the minimax lower bounds for regression problems in the ε-contamination model are the same when 1) both the design X and the labels are contaminated and 2) when only the labels are contaminated. Section C shows that we can extend the results for ℓ_1-penalized Huber's M-estimator when the covariance matrix of the design X satisfies a Restricted Eigenvalue condition. Finally, the proofs of the main theorems are presented in Section D.

Notations All along the paper, for any f in F, $\|f\|_{L_2}$ will be written instead of $\|f\|_{L_2(\mu)}=\int f^2 d\mu$. We also write L_2 instead of $L_2(\mu)$. Let $\mathbb{B}_2$ and $\mathbb{S}_2$ be respectively the unit ball and the unit sphere with respect to the metric $L_2(\mu)$. The letter c will denote an absolute constant whose value may change from one line to another. For a set T, its cardinality is denoted $|T|$. For two real numbers a,b, $a\vee b$ and $a\wedge b$ denote respectively $\max(a,b)$ and $\min(a,b)$. For any set H for which it makes sense, let $H+f^*=\{h+f^*, h\in H\}$, $H-f^*=\{h-f^*, h\in H\}$.

2. Non-regularized procedures

In this section, we study the Empirical Risk Minimizer (ERM) where we recall the definition below:

$$\hat{f}_N \in \arg\min_{f\in F}\frac{1}{N}\sum_{i=1}^N \ell(f(X_i),Y_i) \ . \tag{3}$$

We establish bounds on the error rate $\|\hat{f}_N-f^*\|_{L_2}$ and the excess risk $P\mathcal{L}_{\hat{f}_N}:=P\ell_{\hat{f}_N}-P\ell_{f^*}$ in two different settings 1) when $F-f^*$ is sub-Gaussian, and 2) when $F-f^*$ is locally bounded. We derive fast rates of convergence under very weak assumptions.

2.1. Complexity measures and parameters

The ERM performs well when the empirical excess risk $f \mapsto P_N\mathcal{L}_f$ uniformly concentrates around its expectation $f \mapsto P\mathcal{L}_f$. From Assumption 3, such uniform deviation results depend on the complexity of the class F. There exist different measures of complexity. In this section, we introduce the two main complexity measures we will use throughout this article. Let $H \subseteq F \subseteq L_2$.

1. Let $(G_h)_{h\in H}$ be the centered Gaussian process indexed by H where the covariance structure of $(G_h)_{h\in H}$ is given by $\mathbb{E}(G_{h_1} - G_{h_2})^2 = \mathbb{E}(h_1(X) - h_2(X))^2$ for all $h_1, h_2 \in H$.
   The **Gaussian mean-width** of H is defined as
   $$w(H) = \mathbb{E}\sup_{h\in H} G_h \ . \tag{4}$$
   For example, for $\mu = \mathcal{N}(0,\Sigma)$, $T \subset \mathbb{R}^p$ and $F = \{\langle t,\cdot\rangle, t \in T\}$, we have $w(F) = \mathbb{E}\sup_{t\in T}\Sigma^{1/2}\langle t, \mathbf{G}\rangle$, where $\mathbf{G} \sim \mathcal{N}(0, I_p)$.
2. Let $S \subseteq \{1,\cdots,N\}$ and $(\sigma_i)_{i\in S}$ be i.i.d Rademacher random variables $(P(\sigma_i = 1) = P(\sigma_i = -1) = 1/2)$ independent to $(X_i)_{i\in S}$.
   The **Rademacher complexity** of H, indexed by S is defined as
   $$\mathrm{Rad}_S(H) = \mathbb{E}\sup_{h\in H}\sum_{i\in S}\sigma_i h(X_i) \ , \tag{5}$$
   where this expectation is taken both with respect to $(X_i)_{i\in S}$ and $(\sigma_i)_{i\in S}$. The Rademacher complexity has been extensively used in the literature as a measure of complexity [2, 3, 7].

Depending on the context, we will use either the Gaussian mean-width or the Rademacher complexity as a measure of complexity. In particular, the Gaussian mean-width naturally appears when dealing with sub-Gaussian classes of functions (see Definition 1) while the Rademacher complexity is convenient when dealing with bounded class of functions.

Definition 1. *A class $H \subseteq F \subset L_2$ is called B-**sub-Gaussian** if for every $\lambda > 0$ and $h \in H$*

$$\mathbb{E}\exp(\lambda h(X)/\|h\|_{L_2}) \le \exp(B^2\lambda^2/2) \ .$$

Now, let us define the two complexity parameters that will drive the rates of convergence of the ERM.

Definition 2. *For any $A > 0$, let*

$$r_{\mathcal{I}}^{SG}(A) = \inf\left\{r > 0 : AL w\big(F \cap (f^* + r\mathbb{B}_2)\big) \le c\sqrt{|\mathcal{I}|}r^2\right\}$$

and

$$r_{\mathcal{I}}^{B}(A) = \inf\left\{r > 0 : AL\mathrm{Rad}_{\mathcal{I}}\big(F \cap (f^* + r\mathbb{B}_2)\big) \le c|\mathcal{I}|r^2\right\}$$

where $c > 0$ denotes an absolute constant and L is the Lipschitz constant from Assumption 3. Finally, for any $A, \delta > 0$ set

$$r^{SG}(A,\delta) \geq c\left(r^{SG}_{\mathcal{I}}(A) \vee AL\sqrt{\frac{\log(1/\delta)}{N}} \vee AL\frac{|\mathcal{O}|}{N}\right) , \tag{6}$$

and

$$r^{B}(A,\delta) \geq c\left(r^{B}_{\mathcal{I}}(A) \vee AL\sqrt{\frac{\log(1/\delta)}{N}} \vee AL\frac{|\mathcal{O}|}{N}\right) , \tag{7}$$

where $c > 0$ is an absolute constant.

In Section 2.2 we use $r^{SG}(A,\delta)$ when the class $F - f^*$ is assumed to be sub-Gaussian while we use $r^{B}(A,\delta)$ when the class $F - f^*$ is (locally) bounded.

2.2. Local Bernstein conditions and main results

To obtain fast rates of convergence, it is necessary to impose assumptions on the distribution P. For instance, the margin assumptions [40, 50, 54] and the Bernstein conditions from [4] have been widely used in statistics and learning theory. A class F is called $(1, A)$ Bernstein [3] if for all f in F, $P(\mathcal{L}_f)^2 \leq AP\mathcal{L}_f$. Under Assumption 3, F is $(1, AL^2)$ Bernstein if for all f in F, $\|f - f^*\|^2_{L_2} \leq AP\mathcal{L}_f$. This condition means that the variance of the problem is not too large. In this paper, we use the second version of the Bernstein condition stated above. Moreover, in the spirit of [18], we introduce the (much) weaker **local Bernstein assumption**. Contrary to the global Bernstein condition, our assumption is required to hold only locally around the *oracle* f^* and not for every f in F. As we will see in applications, it allows to consider heavy-tailed noise without deteriorating the convergence rates.

Assumption 4. *Let $\delta > 0$ and $r(\cdot,\delta) \in \{r^{SG}(\cdot,\delta), r^{B}(\cdot,\delta)\}$, where $r^{SG}(\cdot,\delta)$ and $r^{B}(\cdot,\delta)$ are respectively defined in Equations (6) and (7). Assume that there exists a constant $A > 0$ such that for all $f \in F \cap \left(f^* + r(A,\delta)\mathbb{S}_2\right)$, we have $\|f - f^*\|^2_{L_2} \leq AP\mathcal{L}_f$.*

Assumption 4 holds locally around the *oracle* f^*. The bigger $r(\cdot,\delta)$ the stronger Assumption 4. Assumption 4 has been extensively studied in [18, 17] for different Lipschitz and convex loss functions. For the sake of brevity, in applications we will only focus on the Huber loss function in this paper. We are now in position to state the main theorem for the ERM.

Theorem 1. *Let $\mathcal{I} \cup \mathcal{O}$ be a partition of $\{1, \cdots, N\}$ where $|\mathcal{O}| \leq |\mathcal{I}|$. Grant Assumptions 1, 2 and 3. Let $\delta \in (0,1)$.*

1. *Let us assume that the class $F - f^*$ is 1-sub-Gaussian and that Assumption 4 holds for $r(\cdot,\delta) = r^{SG}(\cdot,\delta)$ and $A > 0$. With probability larger than $1 - \delta$, the estimator $\hat{f}_N$ defined in Equation (3) satisfies*

$$\|\hat{f}_N - f^*\|_{L_2} \leq r^{SG}(A,\delta) \quad \text{and} \quad P\mathcal{L}_{\hat{f}_N} \leq c\frac{(r^{SG}(A,\delta))^2}{A} .$$

2. *Let us assume that Assumption 4 holds for* $r(\cdot,\delta) = r^B(\cdot,\delta)$ *and* $A > 0$ *and that*

$$\forall f \in F \cap \left(f^* + r^B(A,\delta)\mathbb{B}_2\right) \text{ and } x \in \mathcal{X} \quad |f(x) - f^*(x)| \leq 1 \tag{8}$$

Then, with probability larger than $1-\delta$*, the estimator* $\hat{f}_N$ *defined in Equation* (3) *satisfies*

$$\|\hat{f}_N - f^*\|_{L_2} \leq r^B(A,\delta) \quad \textit{and} \quad P\mathcal{L}_{\hat{f}_N} \leq c\frac{(r^B(A,\delta))^2}{A}$$

The constant 1 in the sub-Gaussian assumption or in Equation (8) may be replaced by any other constants.

There are two cases in Theorem 1:

1. When class $F - f^*$ is 1-sub-Gaussian, the complexity-parameter driving the convergence rates depends on the Gaussian mean-width.
2. When the class $F - f^*$ is locally bounded (see Equation 8), the complexity-parameter driving the convergence rates depends on the Rademacher complexity. Equation (8) requires L_∞-boundedness only for functions f in $F \cap (f^* + r^B(A,\delta)\mathbb{B}_2)$. For example, let $F = \{\langle t, \cdot\rangle, t \in \mathbb{R}^p\}$ and X be an isotropic random variable, that is $\mathbb{E}\langle X, t\rangle^2 = \|t\|_2^2$ for all $t \in \mathbb{R}^p$. Let t^* be such that $f^*(\cdot) = \langle t^*, \cdot\rangle$ and f be in $F \cap (f^* + r^B(A,\delta)\mathbb{B}_2)$. Then, $|(f - f^*)(x)| = |\langle t - t^*, x\rangle| \leq \|t - t^*\|_2\|x\|_2 \leq \|x\|_2 r^B(A,\delta)$. Simple computations (see [31]) show that when $r^B(A,\delta) = r^B_{\mathcal{I}}(A)$, the complexity parameter $r^B(A,\delta)$ is of the order $\sqrt{p/|\mathcal{I}|}$ and Equation (8) holds if

$$\|x\|_2 \leq c\sqrt{|\mathcal{I}|/p}$$

The more informative data we have, the larger the euclidean radius of $\mathcal{X}$ can be.

In the case of equality in Equations (6) or (7), Theorem 1 holds if the local Bernstein condition 4 is satisfied for all functions f in F such that:

$$\|f - f^*\|_{L_2} = c\left(r_{\mathcal{I}}(A) \vee AL\frac{|\mathcal{O}|}{N} + AL\sqrt{\frac{\log(1/\delta)}{N}}\right) ,$$

that is on an L_2-sphere around f^* with a radius equal to the rate of convergence. The bound on the error rate can be decomposed as the error rate in the non-contaminated setting and the proportion of outliers $AL|\mathcal{O}|/N$. As long as $AL|\mathcal{O}|/N \leq r_{\mathcal{I}}(A)$, the error rate remains constant and equal the one in a non-contaminating setting. On the other hand, if $AL|\mathcal{O}|/N \geq r_{\mathcal{I}}(A)$, the error rate in the contaminated setting becomes linear with respect to the proportion of outliers $|\mathcal{O}|/N$. Theorem 1 shows that when $r_{\mathcal{I}}(A)$ is minimax-rate-optimal in a non-contaminated setting, the ERM remains optimal when less than $Nr_{\mathcal{I}}(A)/(AL)$ outliers contaminate the labels. We also show in Section 2.3 that this dependence with respect to the number of outliers is minimax-rate-optimal for linear regression in $\mathbb{R}^p$ when $|\mathcal{O}|$ outliers may corrupt the labels.

2.3. A concrete example: the class of linear functionals in $\mathbb{R}^p$ with Huber loss function

To put into perspective the results obtained in Sections 2.2, we apply Theorem 1 in the sub-Gaussian framework for linear regression in $\mathbb{R}^p$. Let $F = \{\langle t, \cdot \rangle, t \in \mathbb{R}^p\}$, which satisfies assumption 2. Let $(X_i, Y_i)_{i=1}^N$ be random variables defined by the following linear model:

$$Y_i = \langle X_i, t^* \rangle + \epsilon_i \ , \tag{9}$$

where $(X_i)_{i=1}^N$ are i.i.d Gaussian random vectors in $\mathbb{R}^p$ with zero mean and covariance matrix Σ. The random variables $(\epsilon_i)_{i\in\mathcal{I}}$ are assumed to be symmetric and independent to $(X_i)_{i=1}^N$. For the moment, nothing more is assumed for $(\epsilon_i)_{i\in\mathcal{I}}$. It is clear that assumption 1 holds. The Empirical Risk Minimizer with the Huber loss function is defined as

$$\hat{t}_N^\gamma = \operatorname*{argmin}_{t\in\mathbb{R}^p} \frac{1}{N}\sum_{i=1}^N \ell^\gamma(\langle X_i, t\rangle, Y_i) \tag{10}$$

where $\ell^\gamma(\cdot,\cdot)$ is the Huber loss function defined for any $\gamma > 0$, $u, y \in \mathcal{Y} = \mathbb{R}$, by

$$\ell^\gamma(u,y) = \begin{cases} \frac{1}{2}(y-u)^2 & \text{if } |u-y| \le \gamma \\ \gamma|y-u| - \frac{\gamma^2}{2} & \text{if } |u-y| > \gamma \end{cases} \ ,$$

which satisfies assumption 3 for $L = \gamma$. Let $t, v \in \mathbb{R}^p$ such that $f(\cdot) = \langle t, \cdot\rangle$ and $g(\cdot) = \langle v, \cdot\rangle$. Since $\mu = \mathcal{N}(0,\Sigma)$, we have $\|f-g\|_{L_2}^2 = \mathbb{E}\langle t-v, X_1\rangle^2 = (t-v)^T\Sigma(t-v)$ and $\lambda(f(X_1)-g(X_1))/\|f-g\|_{L_2} = \big(\lambda/(t-v)^T\Sigma(t-v)\big)(t-v)^T X_1 \sim \mathcal{N}(0,\lambda^2)$. If follows that $F - f^*$ is 1-sub-Gaussian.

Let us turn to the computation of the complexity parameter $r^{SG}(A,\delta)$, for $A, \delta > 0$. Well-known computations (see [48]) give:

$$w\big(F \cap (f^* + r\mathbb{B}_2)\big) \le r\sqrt{\mathrm{Tr}(\Sigma)} \quad \text{and} \quad r_{\mathcal{I}}^{SG}(A) = cA\gamma\sqrt{\frac{\mathrm{Tr}(\Sigma)}{N}} \ ,$$

for $c > 0$ an absolute constant.

To apply Theorem 1, it remains to study the local Bernstein assumption for the Huber loss function. We recall the following result from [18].

Proposition 1 ([18], Theorem 7)**.** *Let $r > 0$ and let $F_{Y|X=x}$ be the conditional cumulative function of Y given $X = x$. Let us assume that the following holds.*

a) There exist $\varepsilon, C' > 0$ such that, for all f in F, $\|f - f^*\|_{L_{2+\varepsilon}} \le C'\|f - f^*\|_{L_2}$.

b) Let ε, C' be the constants defined in a). There exists $\alpha > 0$ such that, for all $x \in \mathbb{R}^p$ and all $z \in \mathbb{R}$ satisfying $|z - f^*(x)| \le (\sqrt{2}(C'))^{(2+\varepsilon)/\varepsilon} r$, $F_{Y|X=x}(z+\gamma) - F_{Y|X=x}(z-\gamma) \geqslant \alpha$.

Then, for all $f \in F \cap (f^* + r\mathbb{B}_2)$, $(4/\alpha)P\mathcal{L}_f^\gamma \ge \|f - f^*\|_{L_2}^2$, where $P\mathcal{L}_f^\gamma$ denotes the excess risk associated with the Huber loss with parameter $\gamma > 0$.

Since $\mu = \mathcal{N}(0, \Sigma)$, the point a) holds with $C' = 3$. Moreover, from the model (9), the point b) can be rewritten as: $\forall x \in \mathbb{R}^p, \forall z \in \mathbb{R} : |z - \langle x, t^* \rangle| \leq 18r$,

$$\mathbb{P}\Big(z - \gamma \leq \langle x, t^* \rangle + \epsilon \leq z + \gamma\Big) = F_\epsilon(z + \gamma - \langle x, t^* \rangle) - F_\epsilon(z - \gamma - \langle x, t^* \rangle) \geq \alpha$$

which is satisfied if

$$F_\epsilon(\gamma - 18r) - F_\epsilon(18r - \gamma) \geq \alpha \tag{11}$$

where F_ϵ denotes the cumulative distribution of ϵ distributed as ϵ_i, for any $i \in \mathcal{I}$. Condition (11) simply implies that the noise puts enough mass around zero.

We are now in position to apply Theorem 1 for Huber's M-estimator in $\mathbb{R}^p$.

Theorem 2. *Let $\mathcal{I} \cup \mathcal{O}$ denote a partition of $\{1, \cdots, N\}$ such that $|\mathcal{I}| \geq |\mathcal{O}|$. Let $(X_i, Y_i)_{i=1}^N$ be random variables valued in $\mathbb{R}^p \times \mathbb{R}$ such that $(X_i)_{i=1}^N$ are i.i.d random variable with $X_1 \sim \mathcal{N}(0, \Sigma)$ and for all $i \in \{1, \cdots, N\}$*

$$Y_i = \langle X_i, t^* \rangle + \epsilon_i \ .$$

Let

$$r(\alpha, \delta) = c\frac{\gamma}{\alpha}\left(\sqrt{\frac{\mathrm{Tr}(\Sigma) \vee \log(1/\delta)}{N}} \vee \frac{|\mathcal{O}|}{N}\right) \ .$$

Let $(\epsilon_i)_{i \in I}$ be i.i.d symmetric random variables independent to $(X_i)_{i \in \mathcal{I}}$ such that there exists $\alpha > 0$ such that

$$F_\epsilon\Big(\gamma - 18r(\alpha, \delta)\Big) - F_\epsilon\Big(18r(\alpha, \delta) - \gamma\Big) \geq \alpha \tag{12}$$

where F_ϵ denotes the cdf of ϵ distributed as ϵ_i for i in $\mathcal{I}$. With probability larger than $1 - \delta$ the estimator $\hat{t}_N^\gamma$ defined in Equation (10) satisfies

$$\|\Sigma^{1/2}(\hat{t}_N^\gamma - t^*)\|_2 \leq r(\alpha, \delta) \quad \textit{and} \quad P\mathcal{L}_{\hat{t}_N^\gamma} \leq c\alpha r^2(\alpha, \delta)$$

Theorem 2 holds under no assumption on $|\mathcal{O}|$ except $|\mathcal{O}| \leq |\mathcal{I}|$. There are two situations

1. The number of outliers $|\mathcal{O}|$ is smaller than $\sqrt{\mathrm{Tr}(\Sigma)N}$. We obtain the rate of convergence $\gamma\sqrt{\mathrm{Tr}(\Sigma)/N}$. When $\mathbb{E}[\epsilon_i^2] = \sigma^2$, $i \in \mathcal{I}$ and $\gamma = \sigma$, it corresponds to the minimax-optimal rate of convergence.
2. The number of outliers $|\mathcal{O}|$ exceeds $\sqrt{\mathrm{Tr}(\Sigma)N}$. In this case, the error rate and the excess risk are deteriorated and the dependence is linear with respect to the proportion of outliers.

Let $\varepsilon = |\mathcal{O}|/N$. From [13], this rate is minimax-optimal in the ε-Huber contamination model and hence also minimax-optimal when $|\mathcal{O}|$ outliers contaminate only the labels (see Theorem 7). In Section A, we run simple simulations to illustrate the linear dependence between the error rate and the proportion of outliers.

Theorem 2 handles many different distributions for the noise as long as Equation (12) is satisfied. We illustrate the fact that the local Bernstein condition is very weak with the following example. Let $\epsilon \sim C(1)$ be a standard Cauchy distribution. For all $t \in \mathbb{R}$, $F_\epsilon(t) = 1/2 + \arctan(t)/\pi$. From easy computations, Equation (11) can be rewritten as

$$\arctan(\gamma - 18r) \geq \pi\alpha/2 \ . \tag{13}$$

For

$$r(\alpha, \delta) = \frac{\gamma}{\alpha}\left(\sqrt{\frac{\mathrm{Tr}(\Sigma) \vee \log(1/\delta)}{N}} \vee \frac{|\mathcal{O}|}{N}\right) \ ,$$

Equation (12) becomes

$$\arctan\left(\gamma\left[1 - \frac{c}{\alpha}\left(\sqrt{\frac{\mathrm{Tr}(\Sigma) \vee \log(1/\delta)}{N}} \vee \frac{|\mathcal{O}|}{N}\right)\right]\right) \geq \pi\alpha/2 \ ,$$

which is satified for $\alpha = 1/4$ and $\gamma = 2\tan(\pi/8)$ if

$$c\left(\sqrt{\frac{\mathrm{Tr}(\Sigma) \vee \log(1/\delta)}{N}} \vee \frac{|\mathcal{O}|}{N}\right) \leq 1 \ .$$

Proposition 2. *In same framework as in Theorem 2, when $\epsilon_i \sim C(1)$, for $i \in \mathcal{I}$, the local Bernstein condition is verified for $\alpha = 1/4$ and $\gamma = 2\tan(\pi/8)$ if*

$$\left(\sqrt{\frac{\mathrm{Tr}(\Sigma) \vee \log(1/\delta)}{N}} \vee \frac{|\mathcal{O}|}{N}\right) \leq 1$$

Remark 2. *The local Bernstein condition also holds many other noise distributions. In the case of Gaussian noise we can take α as a constant and $\gamma = \sigma$. In this case we recover the rate $\sigma\sqrt{\mathrm{Tr}(\Sigma)/N}$.*

3. High dimensional setting

In Section 2, we studied non-regularized procedures. If the class of predictors F is too small there is no hope to approximate Y with $f^*(X)$. It is thus necessary to consider large classes of functions leading to a large error rate unless some extra low-dimensional structure is expected on f^*. Adding a regularization term to the empirical loss is a wide-spread method to induce this low-dimensional structure. More formally, let $F \subset E \subset L_2$ and $\|\cdot\| \mapsto \mathbb{R}^+$ be a norm defined on the linear space E. For any $\lambda > 0$, the regularized empirical risk minimizer (RERM) is defined as

$$\hat{f}_N^\lambda = \operatorname*{argmin}_{f \in F} \frac{1}{N}\sum_{i=1}^N \ell(f(X_i), Y_i) + \lambda\|f\| \tag{14}$$

For example, the use of the ℓ_1-norm promotes sparsity [49] for regression and classification problems in $\mathbb{R}^p$, while the 1-Schatten norm promotes low rank

solutions for matrix reconstruction. The main result of this section has the same flavor as the one in Section 2. The error rate can be bounded by

$$r_N + AL\frac{|\mathcal{O}|}{N} \ .$$

where r_N denotes the (sparse or low-dimensional) error rate in a non contaminated setting, L is the Lipschitz constant from Assumption 3 and A is a parameter coming from the local Bernstein condition. When $|\mathcal{O}| \leq r_N N/(AL)$, the RERM behaves as if there was no contamination.

3.1. Complexity parameters and sparsity equation

To analyze regularized procedures, we first need to redefine the complexity parameter.

Definition 3. *Let $\mathbb{B}$ be the unit ball induced by the regularization norm $\|\cdot\|$. For any $A, \rho > 0$, let $\tilde{r}_{\mathcal{I}}^{SG}(A,\rho)$ and $\tilde{r}_{\mathcal{I}}^{B}(A,\rho)$ be defined as*

$$\tilde{r}_{\mathcal{I}}^{SG}(A,\rho) = \inf\{r > 0 : cALw\big(F \cap (f^* + r\mathbb{B}_2 \cap \rho\mathbb{B})\big) \leq \sqrt{|\mathcal{I}|}r^2\} \ ,$$

and,

$$\tilde{r}_{\mathcal{I}}^{B}(A,\rho) = \inf\{r > 0 : cAL\mathrm{Rad}_{\mathcal{I}}\big(F \cap (f^* + r\mathbb{B}_2 \cap \rho\mathbb{B})\big) \leq |\mathcal{I}|r^2\} \ ,$$

where $c > 0$ denotes an absolute constant and L is the Lipschitz constant from assumption 3. For any $A, \delta, \rho > 0$ let $\tilde{r}^{SG}(A,\rho,\delta)$ and $\tilde{r}^{B}(A,\rho,\delta)$ be such that

$$\tilde{r}^{SG}(A,\rho,\delta) \geq \tilde{r}_{\mathcal{I}}^{SG}(A,\rho) \vee AL\sqrt{\frac{\log(1/\delta)}{N}} \vee AL\frac{|\mathcal{O}|}{N} \ , \tag{15}$$

and,

$$\tilde{r}^{B}(A,\rho,\delta) \geq \tilde{r}_{\mathcal{I}}^{B}(A,\rho) \vee AL\sqrt{\frac{\log(1/\delta)}{N}} \vee AL\frac{|\mathcal{O}|}{N} \ . \tag{16}$$

The main difference between the complexity parameters from Definition 3 and the ones from Definition 2 is the localization $\rho\mathbb{B}$. Parameters in Definition 3 measure the local complexity of F around f^*, where the localization is defined with respect to the metric induced by the regularization norm.

To deal with the regularization part, we use the tools from [35]. The idea is the following: the ℓ_1 norm induces sparsity properties because it has large subdifferentials at sparse vectors. Therefore, to obtain "sparsity dependent bounds", i.e bounds depending on the unknown sparsity of the *oracle* f^*, it seems natural to look at the size of the subdifferential of $\|\cdot\|$ in f^*. We recall that the subdifferential of $\|\cdot\|$ in f is defined as

$$(\partial\|.\|)_f = \{z^* \in E^* \ : \ \|f+h\| - \|f\| \geq z^*(h) \text{ for every } h \in E\} \ ,$$

where E^* is the dual space of the normed space $(E, \|\cdot\|)$. It can be also written as

$$(\partial \|\cdot\|)_f = \begin{cases} \{z^* \in \mathbb{S}^* : z^*(f) = \|f\|\} & \text{if } f \neq 0 \\ \mathbb{B}^* & \text{if } f = 0 \end{cases} \tag{17}$$

where $\mathbb{B}^*$ and $\mathbb{S}^*$ denote respectively the unit ball and the unit sphere with respect to the dual norm $\|\cdot\|^*$ defined as $z^* \in E^* \to \|z^*\|^* = \sup_{\|f\|\leq 1} z^*(f)$. When $f \neq 0$, the subdifferential of $\|\cdot\|$ in f is the set of all vectors z^* in the unit dual sphere $\mathbb{S}^*$ which are norming for f. For any $\rho > 0$, let

$$\Gamma_{f^*}(\rho) = \bigcup_{f \in F \cap (f^* + (\rho/20)\mathbb{B})} (\partial \|\cdot\|)_f \ .$$

Instead of looking at the subdifferential of $\|\cdot\|$ exactly in f^* we consider the collection of subdifferentials for functions $f \in F$ "close enough" to the *oracle* f^*. It enables to handle *oracles* f^* that are not exactly sparse but approximatively sparse. The main technical tool to analyze regularization procedures is the following sparsity equation [35].

Definition 4. *For any $A, \rho, \delta > 0$, let $\tilde{r}(A, \rho, \delta) \in \{\tilde{r}^{SG}(A, \rho, \delta), \tilde{r}^{B}(A, \rho, \delta)\}$. Define*

$$H_{A,\rho,\delta} = F \cap \left(f^* + \rho\mathbb{B} \cap \tilde{r}(A, \rho, \delta)\mathbb{B}_2\right) \ ,$$

and

$$\Delta(A, \rho, \delta) = \inf_{h \in H_{A,\rho,\delta}} \sup_{z^* \in \Gamma_{f^*}(\rho)} z^*(h - f^*) \ . \tag{18}$$

*A real number $\rho > 0$ satisfies the A, δ-**sparsity equation** if $\Delta(A, \rho, \delta) \geq 4\rho/5$.*

The constant $4/5$ in Definition 4 could be replaced by any constant in $(0, 1)$. The sparsity equation is a very general and powerful tool allowing to derive "sparsity dependent bounds" when taking ρ^* function of the unknown sparsity (see Section 3.3 for a more explicit example or [17, 35] for many other illustrations).

Remark 3. *It is also possible to obtain "norm dependent bounds", i.e bounds depending on the norm of the oracle $\|f^*\|$. By taking $\rho^* = 20\|f^*\|$, we get that $0 \in F \cap (f^* + (\rho^*/20)\mathbb{B})$ and from Equation (17) it follows that $\Gamma_{f^*}(20\|f^*\|) = \mathbb{B}^*$ and for any $A, \delta > 0$, $\Delta(A, \rho^*, \delta) = \rho^*$. In other words, the sparsity equation is always satisfied for $\rho^* = 20\|f^*\|$ (see Section 3.4 for an example)*

3.2. Local Bernstein conditions and main results

In this section, we adapt the local Bernstein assumption to regularized framework.

Assumption 5. *Let $\delta > 0$ and $\tilde{r}(\cdot, \cdot, \delta) \in \{\tilde{r}^{SG}(\cdot, \cdot, \delta), \tilde{r}^{B}(\cdot, \cdot, \delta)\}$. Suppose there exist $A > 0$ and ρ^* satisfying the A, δ-sparsity equation from Definition 4 such that for all $f \in F \cap \left(f^* + \rho\mathbb{B} \cap \tilde{r}(A, \rho^*, \delta)\mathbb{S}_2\right)$ we have $\|f - f^*\|_{L_2}^2 \leq AP\mathcal{L}_f$.*

We are now in position to state the main theorem of this section.

Theorem 3. *Let $\mathcal{I} \cup \mathcal{O}$ denote a partition of $\{1, \cdots, N\}$ such that $|\mathcal{O}| \leq |\mathcal{I}|$. Grant Assumptions 1, 2, 3. Let $\delta > 0$.*

1. *Let $\tilde{r}(\cdot, \cdot, \delta) = \tilde{r}^{SG}(\cdot, \cdot, \delta)$.*
   Assume that the class $F - f^*$ is 1-sub-Gaussian and that assumption 5 holds with $A, \rho^* > 0$. Set:

$$\lambda = c \frac{(\tilde{r}^{SG}(A, \rho^*, \delta))^2}{A\rho^*} \quad .$$

   With probability larger than $1 - \delta$, the estimator $\hat{f}_N^\lambda$ defined in Equation (14) satisfies

$$\|\hat{f}_N^\lambda - f^*\|_{L_2} \leq \tilde{r}^{SG}(A, \rho^*, \delta) \quad , \quad \|\hat{f}_N^\lambda - f^*\| \leq \rho^*$$
$$\text{and} \quad P\mathcal{L}_{\hat{f}_N^\lambda} \leq c \frac{(\tilde{r}^{SG}(A, \rho^*, \delta))^2}{A} \quad .$$

2. *Let $\tilde{r}(\cdot, \cdot, \delta) = \tilde{r}^{B}(\cdot, \cdot, \delta)$.*
   Let us assume that Assumption 5 holds with $A, \rho^* > 0$ and that

$$\forall f \in F \cap \left(f^* + \tilde{r}^B(A, \rho^*, \delta)\mathbb{S}_2 \cap \rho^* \mathbb{B}\right) \text{ and } x \in \mathcal{X} \quad |f(x) - f^*(x)| \leq 1 \quad (19)$$

   Set:

$$\lambda = c \frac{(\tilde{r}^{B}(A, \rho^*, \delta))^2}{A\rho^*} \quad .$$

   With probability larger than $1 - \delta$, the estimator $\hat{f}_N^\lambda$ defined in Equation (14) satisfies

$$\|\hat{f}_N^\lambda - f^*\|_{L_2} \leq \tilde{r}^{B}(A, \rho^*, \delta) \quad , \quad \|\hat{f}_N^\lambda - f^*\| \leq \rho^*$$
$$\text{and} \quad P\mathcal{L}_{\hat{f}_N^\lambda} \leq c \frac{(\tilde{r}^{B}(A, \rho^*, \delta))^2}{A} \quad .$$

When the equality holds in Equations (15) or (16) we have

$$\|\hat{f}_N^\lambda - f^*\|_{L_2} \leq \tilde{r}_{\mathcal{I}}(A, \rho) \vee AL\sqrt{\frac{\log(1/\delta)}{N}} \vee AL\frac{|\mathcal{O}|}{N} \quad ,$$

with probability larger than $1-\delta$. The error rate can be decomposed as the error rate in the non-contaminated setting and the proportion of outliers $AL|\mathcal{O}|/N$.

Equation (19) means that the class $F - f^*$ is locally bounded. As we will see in Section 3.4, requiring the local boundedness instead of the global one has important consequences.

Theorem 3 is a "meta" theorem in the sense that it can used for many practical problems. We use Theorem 3 for ℓ_1-penalized Huber's M-estimator in Section 3.3. It is also possible to use Theorem 3 for many other convex and Lipschitz

loss functions and regularization norms as it is done in [17]. It can also be used for matrix reconstruction problems by penalizing with the 1-Schatten norm [35].

Theorem 3 may seem complicated at a first glance because the parameter A appears in the definition of the complexity parameters, in the sparsity equation and in Assumption 5. However, there is a simple general routine that we may use to apply Theorem 3.

General routine to apply Theorem 3 when the class $F - f^*$ is sub-Gaussian

1. Verify that the class $F - f^*$ is sub-Gaussian and take $\tilde{r}(\cdot,\cdot,\cdot) = \tilde{r}^{SG}(\cdot,\cdot,\cdot)$.
2. Verify assumptions 1, 2 and 3.
3. Compute the localized Gaussian mean width $w\big(F \cap (f^* + r\mathbb{B}_2 \cap \rho\mathbb{B})\big)$ for any $r, \rho > 0$. Deduce the value of $\tilde{r}^{SG}_{\mathcal{I}}(A,\rho)$ for any $A, \rho > 0$.
4. From the computation of $\tilde{r}^{SG}_{\mathcal{I}}(A,\rho)$ deduce the closed form of $\tilde{r}^{SG}(A,\rho,\delta)$.
5. For fixed constants $A, \delta > 0$, find $\rho^* > 0$ satisfying the A,δ- sparsity equation.
6. From the value of ρ^*, compute $\tilde{r}^{SG}(A,\rho^*,\delta)$ for any $A, \delta > 0$.
7. Find a constant $A > 0$ verifying Assumption 5.

General routine to apply Theorem 3 when the class $F - f^*$ is locally bounded

1. Take $\tilde{r}(\cdot,\cdot,\cdot) = \tilde{r}^{B}(\cdot,\cdot,\cdot)$.
2. Verify assumptions 1, 2 and 3.
3. Compute the localized Rademacher complexity $\mathrm{Rad}_{\mathcal{I}}\big(F \cap (f^* + r\mathbb{B}_2 \cap \rho\mathbb{B})\big)$ for any $r, \rho > 0$. Deduce the value of $\tilde{r}^{B}_{\mathcal{I}}(A,\rho)$ for any $A, \rho > 0$.
4. From the computation of $\tilde{r}^{B}_{\mathcal{I}}(A,\rho)$ deduce the closed form of $\tilde{r}^{G}(A,\rho,\delta)$.
5. For fixed constants $A, \delta > 0$, find $\rho^* > 0$ satisfying the A,δ- sparsity equation.
6. From the value of ρ^*, compute $\tilde{r}^{B}(A,\rho^*,\delta)$ for any $A, \delta > 0$.
7. Find a constant $A > 0$ verifying Assumption 5.
8. Verify that the class $F - f^*$ is locally bounded with $\tilde{r}^{B}(A,\rho^*,\delta)$ computed previously.

We will apply these two general routines for practical examples in Section 3.3 and 3.4.

3.3. Application to ℓ_1-penalized Huber's M-estimator with Gaussian design

Let $F = \{\langle t,\cdot\rangle, t \in \mathbb{R}^p\}$ denote the class of linear functionals in $\mathbb{R}^p$. Let $(X_i, Y_i)_{i=1}^N$ be random variables defined by, $Y_i = \langle X_i, t^*\rangle + \epsilon_i$, where $(X_i)_{i=1}^N$ are i.i.d centered standard Gaussian vectors. The random variables $(\epsilon_i)_{i\in\mathcal{I}}$ are symmetric independent to $(X_i)_{i\in\mathcal{I}}$. The *oracle* t^* is assumed to be s-sparse, $\|t^*\|_0 := \sum_{i=1}^p \mathbb{I}\{t_i^* \neq 0\} \leq s$.

The ℓ_1-penalized Huber's M-estimator is defined as

$$\hat{t}_N^{\gamma,\lambda} = \underset{t\in\mathbb{R}^p}{\operatorname{argmin}} \frac{1}{N}\sum_{i=1}^{N} \ell^\gamma(\langle X_i, t\rangle, Y_i) + \lambda\|t\|_1 \tag{20}$$

where $\ell^\gamma(\cdot,\cdot)$ is the Huber loss function. We use the routine of Theorem 3 when $F - f^*$ is sub-Gaussian:

Step 1: As in Section 2.3, the class $F - f^*$ is 1-sub-Gaussian.

Step 2: It is clear that Assumptions 1, 2, 3 with $L = \gamma$ are satisfied.

Step 3 and 4: Let us turn to the computation of the local Gaussian-mean width. Since $X \sim \mathcal{N}(0, I_p)$, for every $t \in \mathbb{R}^p$, we have $w\big(F \cap (f^* + r\mathbb{B}_2 \cap \rho\mathbb{B})\big) = w(r\mathbb{B}_2^p \cap \rho\mathbb{B}_1^p)$ for every $r, \rho > 0$, where $\mathbb{B}_q^p$ denotes the ℓ_q ball in $\mathbb{R}^p$ for $q > 0$. Well-known computations give (see [57] for example)

$$w(\rho B_1^p \cap r B_2^p) \le \rho w(B_1^p) \le c\rho\sqrt{\log(p)} \ ,$$

and consequently,

$$\big(\tilde{r}_{\mathcal{I}}^{SG}(A,\rho)\big)^2 = cA\gamma\rho\sqrt{\frac{\log(p)}{N}} \ ,$$

and let $\tilde{r}^{SG}(A,\rho,\delta)$ be such that

$$\tilde{r}^{SG}(A,\rho,\delta) \ge c\left(\sqrt{A\gamma\rho}\left(\frac{\log(p)}{N}\right)^{1/4} \vee A\gamma\sqrt{\frac{\log(1/\delta)}{N}} \vee A\gamma\frac{|\mathcal{O}|}{N}\right)$$

Step 5 and 6: To verify the A,δ-sparsity equation from Definition 4 for the ℓ_1 norm we use the following result from [35].

Lemma 1 ([35, Lemma 4.2])**.** *Let $\mathbb{B}_1^p$ denote the unit ball induced by $\|\cdot\|_1$. Let us assume that the design X is isotropic. If the oracle t^* is s-sparse and $100s \le \big(\rho/\big(\tilde{r}^{SG}(A,\rho,\delta)\big)^2$ then $\Delta(A,\rho,\delta) \ge (4/5)\rho$.*

Lemma 1 implies that the A,δ-sparsity equation is satisfied with $\rho^* > 0$ if the sparsity s is smaller than $\big(\rho^*/\big(\tilde{r}^{SG}(A,\rho^*,\delta)\big)^2$. From easy computations, it follows

$$\rho^* = A\gamma\left(s\sqrt{\frac{\log(p)}{N}} \vee \sqrt{\frac{s\log(1/\delta)}{N}} \vee \sqrt{s}\frac{|\mathcal{O}|}{N}\right) \ ,$$

and

$$\tilde{r}^{SG}(A,\rho^*,\delta) = A\gamma\left(\sqrt{\frac{s\log(p)\vee\log(1/\delta)}{N}} \vee \frac{|\mathcal{O}|}{N}\right) \ .$$

Step 7: We use Proposition 1 to show that the local Bernstein condition holds for $f \in F \cap \big(f^* + \tilde{r}^{SG}(A,\rho^*,\delta)\mathbb{S}_2 \cap \rho^*\mathbb{B}\big)$. Since $X \sim \mathcal{N}(0, I_p)$, the point a) in Proposition 1 is verified. Moreover, the point b) holds and the local Bernstein condition is verified with $A = 4/\alpha$ if there exists $\alpha > 0$ satisfying

$$F_\epsilon\Big(\gamma - c\tilde{r}^{SG}(4/\alpha, \rho^*, \delta)\Big) - F_\epsilon\Big(c\tilde{r}^{SG}(4/\alpha, \rho^*, \delta) - \gamma\Big) \geq \alpha \ , \tag{21}$$

where F_ϵ denotes the cdf of ϵ distributed as ϵ_i, for $i \in \mathcal{I}$.

We are now in position to state the main result for the ℓ_1-penalized Huber estimator.

Theorem 4. *Let $\mathcal{I} \cup \mathcal{O}$ denote a partition of $\{1, \cdots, N\}$ such that $|\mathcal{I}| \geq |\mathcal{O}|$ and $(X_i, Y_i)_{i=1}^N$ be random variables valued in $\mathbb{R}^p \times \mathbb{R}$ such that $(X_i)_{i=1}^N$ are i.i.d random variable with $X_1 \sim \mathcal{N}(0, I_p)$ and for all $i \in \{1, \cdots, N\}$*

$$Y_i = \langle X_i, t^* \rangle + \epsilon_i \ ,$$

where t^* is s-sparse. For any $\delta, \alpha > 0$, let

$$\tilde{r}^{SG}(\alpha, \delta) = c\frac{\gamma}{\alpha}\left(\sqrt{\frac{s\log(p) \vee \log(1/\delta)}{N}} \vee \frac{|\mathcal{O}|}{N}\right)$$

Let $(\epsilon_i)_{i\in I}$ are i.i.symmetric random variables independent to $(X_i)_{i\in\mathcal{I}}$ such that there exists $\alpha > 0$ such that

$$F_\epsilon\big(\gamma - \tilde{r}^{SG}(\alpha, \delta)\big) - F_\epsilon\big(\tilde{r}^{SG}(\alpha, \delta) - \gamma\big) \geq \alpha \tag{22}$$

where F_ϵ denotes the cdf of ϵ, where ϵ is distributed as ϵ_i, for i in $\mathcal{I}$. Set

$$\lambda = c\gamma\left(\sqrt{\frac{\log(p)}{N}} \vee \sqrt{\frac{\log(1/\delta)}{sN}} \vee \frac{|\mathcal{O}|}{\sqrt{s}N}\right) \ .$$

Then with probability larger than $1 - \delta$, the estimator $\hat{t}_N^{\gamma,\lambda}$ defined in Equation (20) satisfies

$$\|\hat{t}_N^{\gamma,\lambda} - t^*\|_2 \leq \tilde{r}^{SG}(\alpha, \delta), \quad P\mathcal{L}_{\hat{t}_N^{\gamma,\lambda}} \leq c\alpha(\tilde{r}^{SG}(\alpha, \delta))^2$$

$$\text{and} \quad \|\hat{t}_N^{\gamma,\lambda} - t^*\|_1 \leq c\frac{\gamma}{\alpha}\left(s\sqrt{\frac{\log(p)}{N}} \vee \sqrt{\frac{s\log(1/\delta)}{N}} \vee \sqrt{s}\frac{|\mathcal{O}|}{N}\right)$$

There are two situations:

1. When the number of outliers $|\mathcal{O}|$ is smaller than $\sqrt{s\log(p)N}$, the regularization parameter λ does not depend on the unknown sparsity and we obtain the (nearly) minimax-optimal rate in sparse linear regression in $\mathbb{R}^p$ [5, 35, 20]. Using more involved computations and taking a regularization parameter λ depending on the unknown sparsity, we could obtain the exact minimax rate of convergence $s\log(p/s)/N$.
2. When the number of outliers exceeds $\sqrt{s\log(p)N}$ the value of λ depends on the unknown quantities $|\mathcal{O}|$ and s. The error rate is deteriorated and becomes linear with respect to the proportion of outliers $|\mathcal{O}|/N$. Using Theorem 7 and [12], this error rate is minimax optimal (up to a logarithmic term) when $|\mathcal{O}|$ malicious outliers contaminate only the labels.

As in Section 2.3, we can assume that the noise follows a standard Cauchy distribution. In this case we can take $\alpha = 1/4$ and $\gamma = 2\tan(\pi/8)$ and Equation (22) holds if

$$\sqrt{\frac{s\log(p) \vee \log(1/\delta)}{N}} \vee \frac{|\mathcal{O}|}{N} \leq c \ . \tag{23}$$

When $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$, the local Bernstein condition is verified with $\alpha = 1/4$ and $\gamma = c\sigma$ if Equation (23) holds. In Section A, we run simple simulations to illustrate the linear dependence between the error rate and the proportion of outliers.

Remark 4. *In Theorem 4 we assumed that $\mu = \mathcal{N}(0, I_p)$ to apply Lemma 1 and compute the local Gaussian-mean width. It is possible to generalize the result to Gaussian random vectors with covariance matrices Σ verifying $RE(s, 9)$ [55], where s is the sparsity of t^*. Recall that a matrix Σ is said to satisfy the restricted eigenvalue condition $RE(s, c_0)$ with some constant $\kappa > 0$, if $\|\Sigma^{1/2}v\|_2 \geq \kappa\|v_J\|_2$ for any vector v in $\mathbb{R}^p$ and any set $J \subset \{1, \cdots, p\}$ such that $|J| \leq s$ and $\|v_{J^c}\|_1 \leq c_0\|v_J\|_1$. When Σ satisfies the $RE(s, 9)$ condition with $\kappa > 0$ we get the same conclusion as Theorem 4 modulo an extra term $1/\kappa$ in front of $\tilde{r}_{\mathcal{I}}(A, \rho^*, \delta)$ (see Section C for a precise result).*

3.4. Application to RKHS with the huber loss function

We present another example of application of our main results. In particular, we use the routine associated with Theorem 3 in the locally bounded case, for the problem of learning in a Reproducible Kernel Hilbert Space (RKHS) $\mathcal{H}_K$ [47] associated to a bounded positive definite kernel K. We are given N pairs $(X_i, Y_i)_{i=1}^N$ of random variables where the X_i's take their values in some measurable space $\mathcal{X}$ and $Y_i \in \mathbb{R}$. We introduce a kernel $K : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ measuring a similarity between elements of $\mathcal{X}$ i.e $K(x_1, x_2)$ is small if $x_1, x_2 \in \mathcal{X}$ are "similar". The kernel $K(\cdot, \cdot)$ is assumed to be bounded (for all $x \in \mathcal{X}$: $|K(x, x)| \leq 1$). The main idea of kernel methods is to transport the design data X_i's from the set $\mathcal{X}$ to a certain Hilbert space via the application $x \mapsto K(x, \cdot) := K_x(\cdot)$ and construct a statistical procedure in this "transported" and structured space. The kernel K is used to generate a Hilbert space known as Reproducing Kernel Hilbert Space (RKHS). Recall that if K is a positive definite function i.e for all $n \in \mathbb{N}^*$, $x_1, \cdots, x_n \in \mathcal{X}$ and $c_1, \cdots, c_n \in \mathbb{R}$, $\sum_{i=1}^n \sum_{j=1}^n c_i c_j K(x_i, x_j) \geq 0$. By Mercer's theorem there exists an orthonormal basis $(\phi_i)_{i=1}^\infty$ of L_2 such that $\mu \times \mu$ almost surely, $K(x, y) = \sum_{i=1}^\infty \lambda_i \phi_i(x)\phi_i(y)$, where $(\lambda)_{i=1}^\infty$ is the sequence of eigenvalues (arranged in a non-increasing order) of T_K and ϕ_i is the eigenvector corresponding to λ_i where

$$\begin{aligned} T_K &: L_2 \to L_2 \\ (T_K f)(x) &= \int K(x, y) f(y) d\mu(y) \end{aligned} \tag{24}$$

The Reproducing Kernel Hilbert Space $\mathcal{H}_K$ is the set of all functions of the form $\sum_{i=1}^{\infty} a_i K(x_i, \cdot)$ where $x_i \in \mathcal{X}$ and $a_i \in \mathbb{R}$ converging in L_2 endowed with the inner product

$$\Big\langle \sum_{i=1}^{\infty} a_i K(x_i, \cdot), \sum_{i=1}^{\infty} b_i K(y_i, \cdot) \Big\rangle = \sum_{i,j=1}^{\infty} a_i b_j K(x_i, y_i)$$

An alternative way to define a RKHS is via the feature map $\Phi : \mathcal{X} \mapsto \ell_2$ such that $\Phi(x) = \big(\sqrt{\lambda_i}\phi_i(x)\big)_{i=1}^{\infty}$. Since $(\Phi_k)_{k=1}^{\infty}$ is an orthogonal basis of $\mathcal{H}_K$, it is easy to see that the unit ball of $\mathcal{H}_K$ can be expressed as

$$\mathbb{B}_{\mathcal{H}_K} = \{ f_\beta(\cdot) = \big\langle \beta, \Phi(\cdot) \big\rangle_{\ell_2},\ \|\beta\|_2 \leq 1 \} \tag{25}$$

where $\big\langle \cdot, \cdot \big\rangle_{\ell_2}$ is the standard inner product in the Hilbert space ℓ_2. In other words, the feature map Φ can the used to define an isometry between the two Hilbert spaces $\mathcal{H}_K$ and ℓ_2.

The RKHS $\mathcal{H}_K$ is a convex class of functions from $\mathcal{X}$ to $\mathbb{R}$ that can be used as a learning class F. Let us assume that $Y_i = f^*(X_i) + \epsilon_i$ where $(X_i)_{i=1}^N$ are i.i.d random variables taking values in $\mathcal{X}$. The random variables $(\epsilon_i)_{i\in\mathcal{I}}$ are symmetric i.i.d random variables independent to $(X_i)_{i\in\mathcal{I}}$ and f^* is assumed to belong to $\mathcal{H}_K$. It follows that the *oracle* f^* is also defined as

$$f^* \in \operatorname*{argmin}_{f\in\mathcal{H}_K} \mathbb{E}[\ell^\gamma(f(X), Y)]$$

where ℓ^γ is the Huber loss function. For the sake of simplicity, we assume that $\|f^*\|_{\mathcal{H}_K} \leq 1$. Without this assumption, it is possible to obtain error rates depending on $\|f^*\|_{\mathcal{H}_K}$. However, we do not pursue this analysis here to simplify the presentation. Let f be in $\mathcal{H}_K$. By the reproducing property and the Cauchy-Schwarz inequality we have for all x, y in $\mathcal{X}$

$$|f(x) - f(y)| = |\langle f, K_x - K_y \rangle| \leq \|f\|_{\mathcal{H}_K} \|K_x - K_y\|_{\mathcal{H}_K} \tag{26}$$

From Equation (26), it is clear that the norm of a function in the RKHS controls how fast the function varies over $\mathcal{X}$ with respect to the geometry defined by the kernel (Lipschitz with constant $\|f\|_{\mathcal{H}_K}$). As a consequence the norm of regularization $\|\cdot\|_{\mathcal{H}_K}$ is related with its degree of smoothness w.r.t. the metric defined by the kernel on $\mathcal{X}$ and assuming that $\|f^*\|_{\mathcal{H}_K} \leq 1$ is equivalent to assume that the *oracle* f^* is smooth enough. The estimators $\hat{f}_N^{\gamma,\lambda}$ we study in this section is defined as

$$\hat{f}_N^{\gamma,\lambda} = \operatorname*{argmin}_{f\in\mathcal{H}_K} \frac{1}{N} \sum_{i=1}^{N} \ell^\gamma(f(X_i), Y_i) + \lambda \|f\|_{\mathcal{H}_K} \tag{27}$$

We obtain error rates depending on spectrum $(\lambda_i)_{i=1}^{\infty}$ of the integral operator T_K assumed to satisfy the following assumption.

Assumption 6. *The eigenvalues $(\lambda_i)_{i=1}^{\infty}$ of the integral operator T_K satisfy $\lambda_n \leq cn^{-1/p}$ for some $0 < p < 1$ and $c > 0$ an absolute constant.*

In Assumption 6, the value of p is related with the smoothness of the space $\mathcal{H}_K$. Different kinds of spectra could be analyzed. It would only change the computation of the complexity fixed-points. For the sake of simplicity we only focus on this example as it has been also studied in [10, 43] to obtain fast rates of convergence.

Let us use the routine to apply Theorem 3 in the locally bounded setting. Indeed, we will see than the localization with respect to the norm induced by the kernel allows to obtain a locally bounded class of functions.

Step 1: For any $A, \rho, \delta > 0$, let $\tilde{r}(A, \rho, \delta) = \tilde{r}^B(A, \rho, \delta)$.

Step 2: Since every Reproducible Kernel Hilbert space is convex, it is clear that assumptions 1, 2 and 3 with $L = \gamma$ are satisfied.

Step 3: From Theorem 2.1 in [42], for all $\rho, r > 0$

$$\mathrm{Rad}_{\mathcal{I}}\big(\mathcal{H}_K \cap (f^* + r\mathbb{B}_2 \cap \rho\mathbb{B}_{\mathcal{H}_K}\big) \leq c\sqrt{|\mathcal{I}|}\left(\sum_{k=1}^{\infty}\left(\rho^2\lambda_k \wedge r^2\right)\right)^{1/2}.$$

Under assumption 6, straightforward computations give,

$$\left(\sum_{k=1}^{\infty}\left(\rho^2\lambda_k \wedge r^2\right)\right)^{1/2} \leq c\frac{\rho^p}{r^{p-1}},$$

and thus for any $A, \rho > 0$

$$\tilde{r}^B_{\mathcal{I}}(A, \rho) = c(A\gamma)^{1/(p+1)}\frac{\rho^{p/(p+1)}}{N^{1/(2(p+1))}}$$

Step 4: It follows that

$$\tilde{r}^B(A, \rho, \delta) = c\left((A\gamma)^{1/(p+1)}\frac{\rho^{p/(p+1)}}{N^{1/(2(p+1))}} \vee A\gamma\sqrt{\frac{\log(1/\delta)}{N}} \vee A\gamma\frac{|\mathcal{O}|}{N}\right)$$

Step 5: From Remark 3, $\rho^* = 20\|f^*\|_{\mathcal{H}_K} \leq 20$ satisfies the A, δ-sparsity equation for any $A, \delta > 0$.

Step 6: From step 5, we easily get

$$\tilde{r}^b(A, \delta) = c\left(\frac{(A\gamma)^{1/(p+1)}}{N^{1/(2(p+1))}} \vee A\gamma\sqrt{\frac{\log(1/\delta)}{N}} \vee A\gamma\frac{|\mathcal{O}|}{N}\right)$$

Step 7: This step consists in verifying that Assumption 5 holds. To do so, we use a localized version of Theorem 1.

Proposition 3. *Let $r, \rho > 0$ and let $F_{Y|X=x}$ be the conditional cumulative function of Y given $X = x$. Let us assume that:*

a) There exist $\varepsilon, C' > 0$ such that, for all f in $F \cap \big(f^* + \rho\mathbb{B}_{\mathcal{H}_K} \cap r\mathbb{S}_2\big)$, we have $\|f - f^*\|_{L_{2+\varepsilon}} \leq C'\|f - f^*\|_{L_2}$.

b) Let ε, C' be the constants defined in a). There exists $\alpha > 0$ such that, for all $x \in \mathbb{R}^p$ and all $z \in \mathbb{R}$ satisfying $|z - f^*(x)| \leq (\sqrt{2}(C'))^{(2+\varepsilon)/\varepsilon}r$, $F_{Y|X=x}(z + \gamma) - F_{Y|X=x}(z - \gamma) \geqslant \alpha$.

Then, for all $f \in F \cap \big(f^* + \rho\mathbb{B}_{\mathcal{H}_K} \cap r\mathbb{S}_2\big)$, $(4/\alpha)P\mathcal{L}_f^\gamma \geq \|f - f^*\|_{L_2}^2$, *where* $P\mathcal{L}_f^\gamma$ *denotes the excess risk associated with the Huber loss function with parameter* $\gamma > 0$.

The only difference with Proposition 1, is that the point a) and the conclusion hold for functions f in $F \cap \big(f^* + \rho\mathbb{B}_{\mathcal{H}_K} \cap r\mathbb{S}_2\big)$. Proposition 3 is a refinement of Proposition 1 where a localization with respect to the regularization norm is added.

Let f in $\mathcal{H}_K$ such that $\|f - f^*\|_{\mathcal{H}_K} \leq \rho$ and $\|f - f^*\|_{L_2} = r$. Since $|f(x) - g(x)| = |\langle f - g, K_x\rangle|$ for any $f, g \in \mathcal{H}_K$, $x \in \mathcal{X}$ we get

$$\|f - f^*\|_{L_{2+\varepsilon}}^{2+\varepsilon} = \int (f(x) - f^*(x))^{2+\varepsilon} dP_X(x) \leq (\rho)^\varepsilon \|f - f^*\|_{L_2}^2$$

Since $\|f - f^*\|_{L_2} = r$, it follows that

$$\|f - f^*\|_{L_{2+\varepsilon}} \leq \left(\frac{\rho}{r}\right)^{\varepsilon/(2+\varepsilon)} \|f - f^*\|_{L_2}.$$

Therefore, the point a) holds with $C' = (\rho/r)^{\varepsilon/(2+\varepsilon)}$. Let us turn to the point b). From the fact that $C' = (\rho/r)^{\varepsilon/(2+\varepsilon)}$, we have $(\sqrt{2}C')^{(2+\varepsilon)/\varepsilon} r = 2^{(2+\varepsilon)/2\varepsilon}\rho$ and the point b) can be rewritten as, there exists $\alpha > 0$ such that

$$F_\epsilon(\gamma - c\rho) - F_\epsilon(c\rho - \gamma) \geq \alpha \tag{28}$$

where F_ϵ denotes the cdf of ϵ distributed as ϵ_i for $i \in \mathcal{I}$. Equation (28), simply means that the noise ϵ puts enough mass around 0. In this setting we have $\rho = \rho^* = c$ and Equation (28) becomes,

$$F_\epsilon(\gamma - c) - F_\epsilon(c - \gamma) \geq \alpha$$

Step 8: Let us turn to the local boundedness assumption. Since $|f(x) - f^*(x)| = |\langle f - f^*, K_x\rangle|$ for any $f \in \mathcal{H}_K$, $x \in \mathcal{X}$, if $\|f - f^*\|_{\mathcal{H}_K} \leq \rho^*$ we get $|f(x) - f^*(x)| \leq \rho^* = c$ and the local boundedness assumption is well-satisfied.

The fact that the boundedness assumption is only required to hold locally is essential. It is obvious that it does not hold here over the whole class $F = \mathcal{H}_K$. We are now in position to state our main theorem for regularized learning in RKHS with the Huber loss function.

Theorem 5. *Let* $\mathcal{H}_K$ *be a reproducible kernel Hilbert space associated with kernel* K, *where* $|K(x,x)| \leq 1$, *for any* $x \in \mathcal{X}$. *Let* $\mathcal{I} \cup \mathcal{O}$ *denote a partition of* $\{1, \cdots, N\}$ *such that* $|\mathcal{I}| \geq |\mathcal{O}|$ *and* $(X_i, Y_i)_{i=1}^N$ *be random variables valued in* $\mathcal{X} \times \mathbb{R}$ *such that* $(X_i)_{i=1}^N$ *are i.i.d random variable and for all* $i \in \{1, \cdots, N\}$

$$Y_i = f^*(X_i) + \epsilon_i \ ,$$

where f^* *belongs to* $\mathcal{H}_K$ *and* $\|f^*\|_{\mathcal{H}_K} \leq 1$. *Assume that* $(\epsilon_i)_{i\in I}$ *are i.i.d symmetric random variables independent to* $(X_i)_{i\in\mathcal{I}}$ *such that there exists* $\alpha > 0$ *such that*

$$F_\epsilon\big(\gamma - c\big) - F_\epsilon\big(c - \gamma\big) \geq \alpha \tag{29}$$

where F_ϵ denotes the cdf of ϵ where ϵ is distributed as ϵ_i, for i in $\mathcal{I}$. Grant Assumption 6 and let

$$\tilde{r}(\alpha,\gamma) = c\left(\frac{(\gamma/\alpha)^{1/(p+1)}}{N^{1/(2(p+1))}} \vee \frac{\gamma}{\alpha}\sqrt{\frac{\log(1/\delta)}{N}} \vee \frac{\gamma}{\alpha}\frac{|\mathcal{O}|}{N}\right)$$

Set $\lambda = c\alpha\tilde{r}(\alpha,\gamma)$. Then with probability larger than $1-\delta$, the estimator $\hat{f}_N^{\gamma,\lambda}$ defined in Equation (27) satisfies

$$\|\hat{f}_N^{\gamma,\lambda} - f^*\|_2^2 \le \tilde{r}(\alpha,\gamma) \quad \text{and} \quad P\mathcal{L}_{\hat{f}_N^{\gamma,\lambda}} \le c\alpha\tilde{r}(\alpha,\gamma)$$

Theorem 5 holds with no assumption on the design X.

1. When
$$|\mathcal{O}| \le (\alpha/\gamma)^{p/(p+1)} N^{(2p+1)/(2p+2)} \ ,$$
we recover the same rates as [46, 43] even when the target Y is heavy-tailed. In [46, 43] the authors assume that Y is bounded while in [10] the noise is assumed to be light-tailed. We generalize their results to heavy-tailed noise.
2. When
$$|\mathcal{O}| \ge (\alpha/\gamma)^{p/(p+1)} N^{(2p+1)/(2p+2)} \ ,$$
the error rate is deteriorated and becomes linear with respect to the proportion of outliers.

When the noise is Cauchy distributed, we can take γ a large enough absolute constant to verify Equation (29). We obtain an error rate of order $N^{-1/(p+1)}$. Depending on the value of p we have obtained fast rates of convergence for regularized Kernel methods. The faster the spectrum of T_K decreases the faster the rates of convergence.

4. Conclusion and perspectives

We have presented general analyses to study ERM and RERM when $|\mathcal{O}|$ outliers contaminate the labels when 1) the class $F - f^*$ is sub-Gaussian or 2) when the class $F - f^*$ is locally bounded. We use these "meta theorems" to study Huber's M-estimator with no regularization or penalized with the ℓ_1 norm. Under a very weak assumption on the noise (note that it can even not be integrable), we have obtained minimax-optimal rate of convergence for these two examples when $|\mathcal{O}|$ malicious outliers corrupt the labels. We also have obtained fast rates for regularized learning problems in RKHS when the target Y is unbounded and heavy-tailed.

For the sake of simplicity, we have only presented two examples of applications. Many procedures can be analyzed as it has be done in [17] such as Group Lasso, Fused Lasso, SLOPE etc. The results can be easily extented when the sub-Gaussian assumption over $F - f^*$ is relaxed. It would only degrade the confidence in the main theorems (assuming for example that the class is

sub-exponential). The conclusion would be similar. As long as the proportion of outliers is smaller than the rate of convergence, both ERM and RERM behave as if there was to contamination.

Appendix A: Simulations

In this section, we present simple simulations to illustrate our theoretical findings. We consider regression problems in $\mathbb{R}^p$ both non-regularized and penalized with the ℓ_1-norm. For $i = 1, \cdots, N$, let us consider the following model:

$$Y_i = \langle X_i, t^* \rangle + \epsilon_i$$

where $(X_i)_{i=1}^N$ are i.i.d random variables distributed as $\mathcal{N}(0, I_p)$, $(\epsilon_i)_{i\in\mathcal{I}}$ are symmetric independent to X random variables. Nothing is assumed on $(\epsilon_i)_{i\in\mathcal{O}}$. We consider different distributions for the noise $(\epsilon_i)_{i\in\mathcal{I}}$. We consider

- $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ Gaussian distribution.
- $\epsilon_i \sim \mathcal{T}(2)$ Student distribution with 2-degree of freedom.
- $\epsilon_i \sim \mathcal{C}(1)$ Cauchy distribution.

We study Huber's M estimator defined as

$$\hat{t}_N^\gamma \in \operatorname*{argmin}_{t\in\mathbb{R}^p} \frac{1}{N}\sum_{i=1}^N \ell^\gamma(f(X_i), Y_i)$$

where $\ell^\gamma : \mathbb{R}\times\mathbb{R} \mapsto \mathbb{R}^+$ is the Huber loss function defined as, $\gamma > 0$, $u, y \in \mathbb{R}$, by

$$\ell^\gamma(u, y) = \begin{cases} \frac{1}{2}(y-u)^2 & \text{if } |u-y| \le \gamma \\ \gamma|y-u| - \frac{\gamma^2}{2} & \text{if } |u-y| > \gamma \end{cases}$$

Note that other loss functions could be considered as the absolute loss function, or more generally, any quantile loss function. According to Theorem 2, we have

$$\|\hat{t}_N^\gamma - t^*\|_2 \le c\gamma\left(\sqrt{\frac{p}{N}} + \frac{|\mathcal{O}|}{N}\right)$$

where $c > 0$ is an absolute constant. We add malicious outliers following a uniform distribution over $[-10^{-5}, 10^5]$. We expect to obtain an error rate proportional to the proportion of outliers $|\mathcal{O}|/N$. We ran our simulations with $N = 1000$ and $p = 50$. The only hyper-parameter of the problem is γ. For the sake of simplicity we took $\gamma = 1$ for all our simulations. We see on Figure 1 that no matter the noise, the error rate is proportional to the proportion of outliers which matches with our theoretical findings.

In a second experiment, we study ℓ_1 penalized M-Huber's estimator defined as

$$\hat{t}_N^{\lambda,\gamma} \in \operatorname*{argmin}_{t\in\mathbb{R}^p} \frac{1}{N}\sum_{i=1}^N \ell^\gamma(f(X_i), Y_i) + \lambda\|t\|_1$$

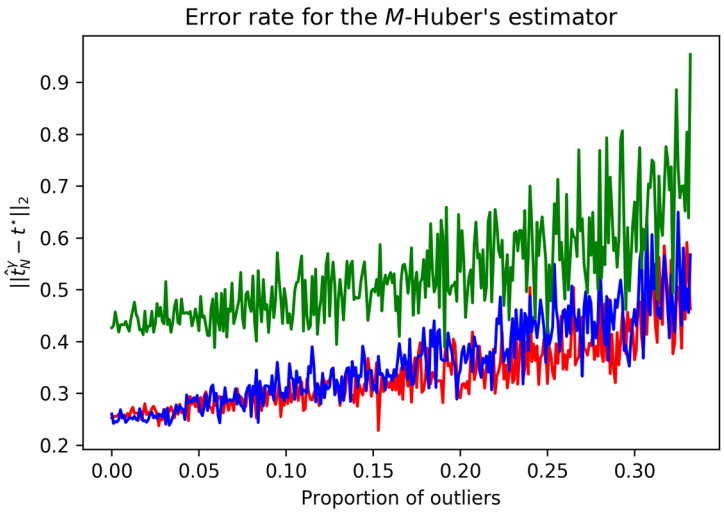

FIG 1. *Error rate for the Huber's M-estimator ($p = 50$ and $N = 1000$)*

where $\ell^\gamma : \mathbb{R} \times \mathbb{R} \mapsto \mathbb{R}^+$ is the Huber loss function and $\lambda > 0$ is a hyper-parameter. According to Theorem 4 we have

$$\|\hat{t}_N^\gamma - t^*\|_2 \leq c\gamma\left(\sqrt{\frac{s\log(p)}{N}} + \frac{|\mathcal{O}|}{N}\right)$$

where $c > 0$ is an absolute constant. We ran our simulations with $N = 1000$ and $p = 1000$ and $s = 50$. The hyper-parameters of the problem are γ and λ. For the sake of simplicity we take $\gamma = 1$ and $\lambda = 10^{-3}$ for all our simulations. We see on Figure 2 that no matter the noise, the error rate is proportional to the proportion of outliers which matches our theoretical findings. The fact that the error rate may be large comes to the fact that we did not optimize the value of λ.

Appendix B: Lower bound minimax risk in regression where only the labels are contaminated

This section is built on the work [13] where the authors establish a general minimax theory for the ε-contamination model defined as $P_{(\varepsilon,\theta,Q)} = (1-\varepsilon)P_\theta + \varepsilon Q$ given a general statistical experiment $\{P_\theta, \theta \in \Theta\}$. A proportion ε of outliers with same the distribution Q contaminate P_θ. Given a loss function $L(\theta_1, \theta_2)$, the minimax rate for the class $\{P_{(\varepsilon,\theta,Q)}, \theta \in \Theta, Q\}$ depends on the modulus of continuity defined as:

$$w(\varepsilon, \Theta) = \sup\left\{L(\theta_1, \theta_2) : TV(P_{\theta_1}, P_{\theta_2}) \leq \frac{\varepsilon}{1-\varepsilon}, \theta_1, \theta_2 \in \Theta\right\} \tag{30}$$

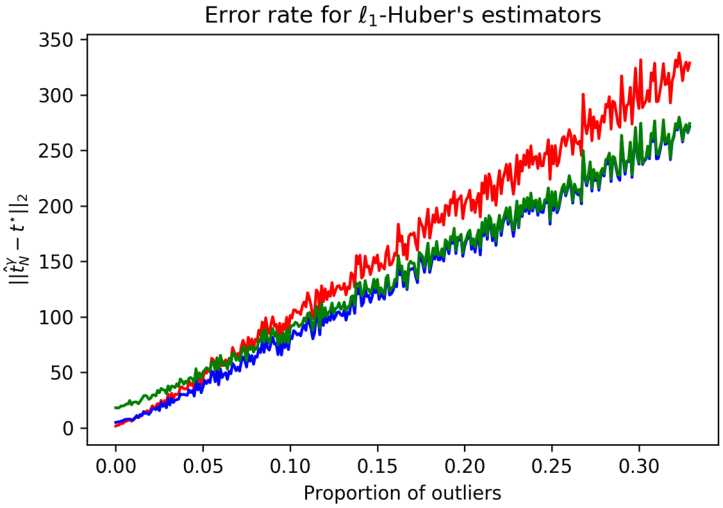

FIG 2. *Error rate for ℓ_1 penalized Huber's M-estimator ($p = 1000$ and $N = 1000$ and $s = 50$)*

where $TV(P_{\theta_1}, P_{\theta_2})$ denotes the total variation distance between P_{θ_1} and P_{θ_2} defined as $TV(P_{\theta_1}, P_{\theta_2}) = \sup_{A\in\mathcal{F}} |P_{\theta_1}(A) - P_{\theta_2}(A)|$, for $\mathcal{F}$ the σ-algebra onto which P_{θ_1} and P_{θ_2} are defined.

Theorem 6 (Theorem 5.1 [13])**.** *Suppose there is some $\mathcal{M}(0)$ such that for $\varepsilon = 0$*

$$\inf_{\hat{\theta}} \sup_{\theta\in\Theta} \sup_{Q} P_{(\varepsilon,\theta,Q)}\left(L(\theta,\hat{\theta}) \geq \mathcal{M}(\varepsilon)\right) \geq c \tag{31}$$

holds. Then, for any $\varepsilon \in [0,1]$ (31) holds for $\mathcal{M}(\varepsilon) = c\big(\mathcal{M}(0) \vee w(\varepsilon,\Theta)\big)$.

$w(\varepsilon,\Theta)$ is the price to pay in the minimax rate when a proportion ε of the samples are contaminated. To illustrate Theorem 6, let us consider the linear regression model:

$$Y_i = \langle X_i, \theta\rangle + \epsilon_i$$

where without contamination $X_i \sim \mathcal{N}(0,\Sigma)$, $\epsilon_i \sim \mathcal{N}(0,\sigma^2)$ are independent. In [12], the authors consider a setting when both the design X and the response variable in the model can be contaminated i.e $(X_1,Y_1),\cdots,(X_N,Y_N) \sim (1-\varepsilon)P_\theta + \varepsilon Q$, with $P_\theta = P(X)P(Y|X)$, $P(X) = \mathcal{N}(0,\Sigma)$ and $P(Y|X) = \mathcal{N}(X^T\theta,\sigma^2)$. They establish that the minimax optimal risk over the class of s-sparse vectors for the metric $L(\theta_1,\theta_2) = \|\theta_1 - \theta\|_2^2$ is given by

$$\sigma^2\left(\frac{s\log(p/s)}{N} \vee \varepsilon^2\right).$$

The question of main interest in our setting is the following: does the minimax risk for regression problem in the ε-contamination model remain the same when only the labels are contaminated?

The following theorem answers to the above question.

Theorem 7. *Let* $\{P_\theta = P^\theta_{(X,Y)}$ *with* $Y = f_\theta(X) + \epsilon, \theta \in \Theta\}$ *be a statistical regression model. For any* $\theta \in \Theta$, $\varepsilon \in [0,1]$ *let*

$$\mathcal{P}_{\theta,\varepsilon} = \Big\{ \big((1-\varepsilon)P_\theta + \varepsilon Q_\theta\big)^{\otimes_{i=1}^N}, P_\theta = P^\theta_{(X,Y)} \text{ with } Y = f_\theta(X) + \epsilon$$
$$Q_\theta = P^\theta_{(X,\tilde{Y})} \text{ with } \tilde{Y} = f_\theta(X) + \tilde{\epsilon} \Big\}$$

Suppose there is some $\mathcal{M}(0)$ *such that for* $\varepsilon = 0$

$$\inf_{\hat{\theta}} \sup_{R_{\theta,\varepsilon} \in \mathcal{P}_{\theta,\varepsilon}, \theta \in \Theta} R_{\theta,\varepsilon}\Big(L(\theta, \hat{\theta}) \geq \mathcal{M}(\varepsilon) \Big) \geq c \tag{32}$$

holds. Then For any $\varepsilon \in [0,1]$ (32) *holds for* $\mathcal{M}(\varepsilon) = c\big(\mathcal{M}(0) \vee w(\varepsilon, \Theta)\big)$

Theorem 7 states that the minimax optimal rates for regression problems in the ε-contamination model are the same when

1. Both the design X and the response variable Y are contaminated.
2. Only the response variable Y is contaminated.

The proof is very similar as the one of Theorem 6. We present it here, for the sake of completeness.

Proof. The case when $\mathcal{M}(\varepsilon) = c\mathcal{M}(0)$ is straightforward. Thus, the goal is to lower bound with a constant the following quantity

$$\inf_{\hat{\theta}} \sup_{R_{\theta,\varepsilon} \in \mathcal{P}_{\theta,\varepsilon}, \theta \in \Theta} R_{\theta,\varepsilon}\Big(L(\theta, \hat{\theta}) \geq w(\varepsilon, \Theta) \Big)$$

We use Le Cam's method with two hypotheses. The first goal is to find θ_1, θ_2 such that $L(\theta_1, \theta_2) \geq w(\varepsilon, \Theta)$. To do so, let θ_1, θ_2 be solution of

$$\max_{\theta_1, \theta_2 \in \Theta} L(\theta_1, \theta_2) \quad \text{s.t} \quad TV(P_{\theta_1}, P_{\theta_2}) = TV(P^{\theta_1}_{(X,Y)}, P^{\theta_2}_{(X,Y)}) \leq \frac{\varepsilon}{1-\varepsilon}$$

Thus there exists $\varepsilon' \leq \varepsilon$ such that $TV(P_{\theta_1}, P_{\theta_2}) = \varepsilon'/(1-\varepsilon')$ and $L(\theta_1, \theta_2) = w(\varepsilon, \Theta)$. To conclude, it is enough to find two distributions $R_{\theta_1,\varepsilon}$ and $R_{\theta_2,\varepsilon}$ in $\mathcal{P}_{\theta_1,\varepsilon}$ and $\mathcal{P}_{\theta_2,\varepsilon}$ such that $R_{\theta_1,\varepsilon} = R_{\theta_2,\varepsilon}$. It would imply that θ_1 and θ_2 are not identifiable from the model and the Le Cam's method would complete the proof.

For $i \in \{1,2\}$ let p_{θ_i} be a density function defined for all $(x,y) \in \mathcal{X} \times \mathcal{Y}$ as

$$p_{\theta_i}(x,y) = \frac{dP^{\theta_i}_{(X,Y)}}{d\big(P^{\theta_1}_{(X,Y)} + P^{\theta_2}_{(X,Y)}\big)}(x,y) \tag{33}$$

By conditioning, it is possible to write $p_{\theta_i}(x,y) = p_X(x) p^{\theta_i}_{Y|X=x}(y)$. Let $R_{\theta_1,\varepsilon}$ and $R_{\theta_2,\varepsilon}$ defined respectively as

$$R_{\theta_1,\varepsilon} = (1-\varepsilon')P^{\theta_1}_{(X,Y)} + \varepsilon' P^{\theta_1}_{(X,\tilde{Y})} \quad \text{and} \quad R_{\theta_2,\varepsilon} = (1-\varepsilon')P^{\theta_2}_{(X,Y)} + \varepsilon' P^{\theta_2}_{(X,\tilde{Y})}$$

where $P^{\theta_1}_{(X,\tilde{Y})}$ and $P^{\theta_2}_{(X,\tilde{Y})}$ are defined by their density functions $\forall (x,y) \in \mathcal{X} \times \mathcal{Y}$,

$$\frac{dP^{\theta_1}_{(X,\tilde{Y})}}{d\big(P^{\theta_1}_{(X,Y)} + P^{\theta_2}_{(X,Y)}\big)}(x,y) = \frac{\big(p_{\theta_2}(x,y) - p_{\theta_1}(x,y)\big)\mathbb{I}\{p_{\theta_2}(x,y) \geq p_{\theta_1}(x,y)\}}{TV\big(P^{\theta_1}_{(X,Y)}, P^{\theta_2}_{(X,Y)}\big)}$$
$$\frac{dP^{\theta_2}_{(X,\tilde{Y})}}{d\big(P^{\theta_2}_{(X,Y)} + P^{\theta_1}_{(X,Y)}\big)}(x,y) = \frac{\big(p_{\theta_1}(x,y) - p_{\theta_2}(x,y)\big)\mathbb{I}\{p_{\theta_1}(x,y) \geq p_{\theta_2}(x,y)\}}{TV\big(P^{\theta_1}_{(X,Y)}, P^{\theta_2}_{(X,Y)}\big)}$$

Using Scheffé's theorem, it is easy to see that $P^{\theta_1}_{(X,\tilde{Y})}$ and $P^{\theta_2}_{(X,\tilde{Y})}$ are probability measures. Moreover, from the facts that $p_{\theta_i}(x,y) = p_X(x)p^{\theta_i}_{Y|X=x}(y)$, $\varepsilon' \leq \varepsilon$ and Lemma 7.2 in [13] we have $R_{\theta_1,\varepsilon} \in \mathcal{P}_{\theta_1,\varepsilon}$ and $R_{\theta_2,\varepsilon} \in \mathcal{P}_{\theta_2,\varepsilon}$.

For any $(x,y) \in \mathcal{X} \times \mathcal{Y}$. Straightforward computations give

$$\begin{aligned}
&\frac{dR_{\theta_1,\varepsilon}}{d\big(P^{\theta_1}_{(X,Y)} + P^{\theta_2}_{(X,Y)}\big)}(x,y) \\
&= (1-\varepsilon')p_{\theta_1}(x,y) + \varepsilon' \frac{\big(p_{\theta_2}(x,y) - p_{\theta_1}(x,y)\big)\mathbb{I}\{p_{\theta_2}(x,y) \geq p_{\theta_1}(x,y)\}}{TV\big(P^{\theta_1}_{(X,Y)}, P^{\theta_2}_{(X,Y)}\big)} \\
&= (1-\varepsilon')p_{\theta_1}(x,y) + \varepsilon' \frac{\big(p_{\theta_2}(x,y) - p_{\theta_1}(x,y)\big)\mathbb{I}\{p_{\theta_2}(x,y) \geq p_{\theta_1}(x,y)\}}{\varepsilon'/(1-\varepsilon')} \\
&= (1-\varepsilon')\big(p_{\theta_1}(x,y) + (p_{\theta_2}(x,y) - p_{\theta_1}(x,y))\mathbb{I}\{p_{\theta_2}(x,y) \geq p_{\theta_1}(x,y)\}\big) \\
&= (1-\varepsilon')\big(p_{\theta_2}(x,y) + (p_{\theta_1}(x,y) - p_{\theta_2}(x,y))\mathbb{I}\{p_{\theta_1}(x,y) \geq p_{\theta_2}(x,y)\}\big) \\
&= \frac{dR_{\theta_2,\varepsilon}}{d\big(P^{\theta_1}_{(X,Y)} + P^{\theta_2}_{(X,Y)}\big)}(x,y)
\end{aligned}$$

□

Appendix C: ℓ_1-penalized Huber's M-estimator with non-isotropic design

In this section, we relax the isotropic assumption on the design X. Recall that a random variable X is isotropic if for every $t \in \mathbb{R}^p$, $\mathbb{E}\langle X, t\rangle^2 = \|t\|_2^2$. Instead, we consider covariance matrices satisfying a Restricted Eigenvalue condition (RE). A matrix Σ is said to satisfy the restricted eigenvalue condition $\mathrm{RE}(s, c_0)$ with some constant $\kappa > 0$, if $\|\Sigma^{1/2}v\|_2 \geq \kappa\|v_J\|_2$ for any vector v in $\mathbb{R}^p$ and any set $J \subset \{1, \cdots, p\}$ such that $|J| \leq s$ and $\|v_{J^c}\|_1 \leq c_0\|v_J\|_1$. We want to derive a result similar to Theorem 4 when $X \sim N(0,\Sigma)$, for Σ satisfying $\mathrm{RE}(s,c)$ for c an absolute constant. With non isotropic design we cannot use Lemma 1 and the computation of the Gaussian mean-width is more involved.

Lemma 2. *Let $\mathbb{B}_1^p$ denote the unit ball induced by $\|\cdot\|_1$. Let us assume that the design X has a covariance matrix satisfying RE$(s,9)$ with constant $\kappa > 0$. If the oracle t^* is s-sparse and $100s \leq \big(\kappa\rho/r\big)^2$ then:*

$$\Delta(A,\rho,\delta) = \inf_{w \in H_{A,\rho,\delta}} \sup_{z^* \in \Gamma_{t^*}(\rho)} \big\langle z^*, w\big\rangle \geq 4\rho/5 \ .$$

The difference with Lemma 1 is the term κ coming from the RE$(s, 9)$ condition.

Proof. The goal is to find ρ such that $\Delta(A, \rho, \delta) \geq (4/5)\rho$. Recall that

$$(\partial \|\cdot\|)_t = \begin{cases} \{z^* \in \mathbb{S}^* : \langle z^*, t\rangle = \|t\|\} & \text{if } t \neq 0 \\ \mathbb{B}^* & \text{if } t = 0 \end{cases}. \tag{34}$$

Since $F = \{\langle t, \cdot\rangle, t \in \mathbb{R}^p\}$, $\|f\|_{L_2} = \|\langle t, X\rangle\|_{L_2} = \|\Sigma^{1/2}t\|_2$. Let w be in $\mathbb{R}^p$ such that $\|w\|_1 = \rho$ and $\|\Sigma^{1/2}w\|_2 \leq r$. Let us denote by I the support of t^* and $P_I w$ the projection of w on $(e_i)_{i\in I}$. By assumption we have $|I| \leq s$. Let z in $(\partial \|\cdot\|)_{t^*}$ such that for every $i \in I$, $z_i = \text{sign}(t_i^*)$, and for every $i \in I^c$, $z_I = \text{sign}(w_i)$. It is clear that z is norming for t^* i.e $\langle z, t^*\rangle = \|t^*\|_1$, $z \in \mathbb{S}_1^* = \mathbb{S}_\infty$ and

$$\langle z, w\rangle = \langle z, P_I w\rangle + \|P_{I^c}\|_1 \geq -\|P_I w\|_1 + \|P_{I^c}\|_1 = \rho - 2\|P_I w\|_1$$

Let us assume that $P_I w$ satisfies $\|P_{I^c} w\|_1 > 9\|P_I w\|_1$ which can be rewritten as $\rho \geq 10\|P_I w\|_1$. It follows that

$$\langle z, w\rangle \geq \rho - 2\|P_I w\|_1 \geq \rho - \frac{1}{5}\rho \geq 4\rho/5,$$

and the sparsity equation is satisfied. Now let us turn to the case when $\|P_{I^c} w\| \leq 9\|P_I w\|_1$. From the RE$(s, 9)$ condition we have $\|P_I w\|_2 \leq \|\Sigma^{1/2}w\|_2/\kappa$ and it follows

$$\rho - 2\|P_I w\|_1 \geq \rho - 2\sqrt{s}\|P_I w\|_2 \geq \rho - \frac{2}{\kappa}\sqrt{s}\|\Sigma^{1/2}w\|_2 \geq \rho - \frac{2}{\kappa}\sqrt{s}r \geq 4\rho/5 \quad \square$$

Now, let us turn to the computation of the Gaussian mean-width when the design X is not isotropic. To do so, we use the following Proposition.

Proposition 4 (Proposition 1 [9])**.** *Let $p \geq 1$ and $M \geq 2$. Let T be the convex hull of M points in $\mathbb{R}^p$ and assume that $T \subset \mathbb{B}_2^p$. Let $\mathbf{G} \sim \mathcal{N}(0, I_p)$. Then for all $s > 0$,*

$$\mathbb{E} \sup_{t \in s\mathbb{B}_2^p \cap T} \langle t, \mathbf{G}\rangle = w(s\mathbb{B}_2^p \cap T) \leq 4\sqrt{\log_+(4eM(s^2 \wedge 1))},$$

where $\log_+(a) = \max(1, \log(a))$.

When $F = \{\langle t, \cdot\rangle, t \in \mathbb{R}^p\}$ and the covariance matrix of X is Σ, for every $r, \rho > 0$ we have

$$w\big(F \cap (f^* + r\mathbb{B}_2 \cap \rho\mathbb{B}_1^p\big) = w\big(r\mathbb{B}_2^p \cap \rho\Sigma^{1/2}\mathbb{B}_1^p\big) = w\big(r\mathbb{B}_2^p \cap \rho T\big)$$

where $T := \Sigma^{1/2}\mathbb{B}_1^p$ is the convex hull of $(\pm\Sigma^{1/2}e_i)_{i=1}^p$. To apply Proposition 4 it is necessary to assume that for every $i = 1, \cdots, p$, $\Sigma^{1/2}e_i \in \mathbb{B}_2^p$ which holds when $\Sigma_{i,i} \leq 1$ and we obtain the following result:

Proposition 5. *Let $F = \{\langle t, \cdot \rangle, t \in \mathbb{R}^p\}$ and assume that Σ, the covariance matrix of X, satisfies $\Sigma_{i,i} \leq 1$ for every $i = 1, \cdots, p$. Then, for every $r, \rho > 0$*

$$w\big(F \cap (f^* + r\mathbb{B}_2 \cap \rho \mathbb{B}_1^p\big) \leq 4\rho\sqrt{\log_+(8ep((r/\rho)^2 \wedge 1))}$$

By taking step by step the computations from Section 3.3 we obtain the following theorem extending Theorem 4 for a non-isotropic design:

Theorem 8. *Let $\mathcal{I} \cup \mathcal{O}$ denote a partition of $\{1, \cdots, N\}$ such that $|\mathcal{I}| \geq |\mathcal{O}|$ and $(X_i, Y_i)_{i=1}^N$ be random variables valued in $\mathbb{R}^p \times \mathbb{R}$ such that $(X_i)_{i=1}^N$ are i.i.d random variable with $X_1 \sim \mathcal{N}(0, \Sigma)$, where Σ satisfies $\Sigma_{i,i} \leq 1$ for $i = 1, \cdots, p$ and verifies $RE(s, 9)$ for some constant $\kappa > 0$. Assume that for all $i \in \{1, \cdots, N\}$*

$$Y_i = \langle X_i, t^* \rangle + \epsilon_i \ ,$$

where t^* is s-sparse. Let

$$\tilde{r}(\alpha, \delta) = c\frac{\gamma}{\alpha}\left(\sqrt{\frac{s\log(p)}{\kappa^2 N}} \vee \sqrt{\frac{\log(1/\delta)}{N}} \vee \frac{|\mathcal{O}|}{N}\right)$$

Assume that $(\epsilon_i)_{i \in I}$ are i.i.d random variables independent to $(X_i)_{i \in \mathcal{I}}$ such that there exists $\alpha > 0$ such that

$$F_\epsilon\big(\gamma - c\tilde{r}(\alpha, \delta)\big) - F_\epsilon\big(c\tilde{r}(\alpha, \delta) - \gamma\big) \geq \alpha \tag{35}$$

where F_ϵ denotes the cdf of ϵ where ϵ is distributed as ϵ_i, for i in $\mathcal{I}$. Set

$$\lambda = c\gamma\left(\sqrt{\frac{\log(p)}{N}} \vee \kappa\sqrt{\frac{\log(1/\delta)}{sN}} \vee \frac{\kappa|\mathcal{O}|}{\sqrt{s}N}\right) \ .$$

Then with probability larger than $1-\delta$ the estimator $\hat{t}_N^{\gamma,\lambda}$ defined in Equation (20) satisfies

$$\|\hat{t}_N^{\gamma,\lambda} - t^*\|_2 \leq \tilde{r}(\alpha, \delta), \quad P\mathcal{L}_{\hat{t}_N^{\gamma,\lambda}} \leq c\alpha(\tilde{r}(\alpha, \delta))^2$$

$$\text{and} \quad \|\hat{t}_N^{\delta,\lambda} - t^*\|_1 \leq c\frac{\gamma}{\kappa\alpha}\left(\frac{s}{\kappa}\sqrt{\frac{\log(p)}{N}} \vee \sqrt{\frac{s\log(1/\delta)}{N}} \vee \sqrt{s}\frac{|\mathcal{O}|}{N}\right)$$

When $\epsilon_i \sim C(1)$, we can use the same argument as in Section 2.3. Equation (35) holds with $\alpha = 1/4$ and $\gamma = 2\tan(\pi/8)$ if

$$\sqrt{\frac{s\log(p)}{\kappa^2 N}} \vee \sqrt{\frac{\log(1/\delta)}{N}} \vee \frac{|\mathcal{O}|}{N} \leq c \ . \tag{36}$$

Now, for the sake of comparizon with [19], let us consider the case where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. For any $t \in \mathbb{R}$, $F_\epsilon(t) = \Phi(t/\sigma)$, where Φ denotes the cdf of a standard gaussian random variable. Thus, Equation (35) can be rewritten as

$$\Phi\left(\frac{\gamma - c\tilde{r}(\alpha, \delta)}{\sigma}\right) \geq \frac{1+\alpha}{2} \ ,$$

which is verified for $\gamma = c\sigma$ and $\alpha = 1/4$ if Equation (36) holds. We improve the main result of [19] in two aspects:

1. We obtain the error rate
$$\sigma\left(\sqrt{\frac{s\log(p)}{\kappa^2 N}} \vee \sqrt{\frac{\log(1/\delta)}{N}} \vee \frac{|\mathcal{O}|}{N}\right) ,$$
while their rate is
$$\sigma\left(\sqrt{\frac{s\log(p/\delta)}{\kappa^2 N}} \vee \frac{|\mathcal{O}|\log(n/\delta)}{N}\right) .$$
In our rate, the term κ is only in factor with the first term and not the probability confidence. The extra term $\log(n/\delta)$ can be problematic in their bound when one wants to obtain exponentially large confidence.
2. Their bound holds for $|\mathcal{O}| \leq cN/\log(N)$ while ours holds for $|\mathcal{O}| \leq cN$.

However, note that when $|\mathcal{O}| \geq \sqrt{s\log(p)N}/\kappa$, our regularization parameter depends on the unknown sparsity and the number of outliers, which is not the case in [19].

Remark 5. *It is possible to replace* $\log(p)$ *by* $\log(p/s)$ *and recover the exact minimax rate of convergence. However, the price to pay is that the regularization parameter* λ *would always depend on the sparsity* s*, even when the number or outliers in small.*

Appendix D: Proofs main theorems

D.1. Proof Theorem 1 in the sub-Gaussian setting

Let $r(\cdot,\delta)$ be such that for all $A > 0$
$$r(A,\delta) \geq c\left(r_{\mathcal{I}}^{SG}(A) \vee AL\sqrt{\frac{\log(1/\delta)}{N}} \vee AL\frac{|\mathcal{O}|}{N}\right) .$$
Moreover, let A satisfying assumption 4 with $r(\cdot,\delta)$.

The proof is split into two parts. First we identify a stochastic argument holding with large probability. Then we show on that event that $\|\hat{f}_N - f^*\|_{L_2} \leq r(A,\delta)$. Finally, at the very end of the proof we show that $P\mathcal{L}_{\hat{f}_N} \leq r^2(A,\delta)/A$.

Stochastic arguments First we identify the stochastic event onto which the proof easily follows. Let $\mathcal{F}_r = F \cap \left(f^* + r(A,\delta)\mathbb{B}_2\right)$ and define
$$\Omega_{\mathcal{I}} = \left\{ \sup_{f\in\mathcal{F}_r} \left|\left(P - P_{\mathcal{I}}\right)\left(\ell_f - \ell_{f^*}\right)\right| \leq c\frac{L}{\sqrt{|\mathcal{I}|}}\left(w(\mathcal{F}_r) + r(A,\delta)\sqrt{\log(1/\delta)}\right)\right\} \tag{37}$$
$$\Omega_{\mathcal{O}} = \left\{ \sup_{f\in\mathcal{F}_r} \left|\left(P - P_{\mathcal{O}}\right)|f - f^*|\right| \leq \frac{c}{\sqrt{|\mathcal{O}|}}\left(w(\mathcal{F}_r) + r(A,\delta)\sqrt{\log(1/\delta)}\right)\right\} \tag{38}$$
where for any $K \subset \{1,\cdots,N\}$, $g : \mathcal{X}\times\mathcal{Y} \mapsto \mathbb{R}$, $P_K g = 1/(|K|)\sum_{i\in K} g(X_i,Y_i)$ and $w(\mathcal{F}_r)$ is the Gaussian mean-width of $\mathcal{F}_r$. Finally let us define $\Omega = \Omega_{\mathcal{I}} \cap \Omega_{\mathcal{O}}$.

Lemma 3. *Grant Assumptions 1, 3, 2 and assume that $F - f^*$ is 1-sub-Gaussian. The event Ω holds with probability larger than $1-\delta$*

The proof of Lemma 3 necessitates several tools from sub-Gaussian random variables that we introduce now.

Let $\psi_2(u) = \exp(u^2) - 1$. The Orlicz space L_{ψ_2} associated to ψ_2 is defined as the set of all random variables Z on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ such that $\|Z\|_{\psi_2} < \infty$ where

$$\|Z\|_{\psi_2} = \inf\{c > 0, \mathbb{E}\psi_2\left(\frac{Z}{c}\right) \leq 1\}$$

Let $H \subset L_2$ and $(X_h)_{h\in H}$ be a stochastic process indexed by the metric space $(H, \|\cdot\|_{L_2})$ satisfying the following Lipschitz condition

$$\text{for all } h, g \in H, \quad \|X_g - X_h\|_{\psi_2} \leq \|g - h\|_{L_2} \ . \tag{39}$$

For such a process it is possible to control the deviation of $\sup_{h\in H} X_h$ in terms of the geometry of $(H, \|\cdot\|_{L_2})$ through the Gaussian mean-width of H.

Theorem 9 ([36], Theorem 11.13)**.** *Let $(X_h)_{h\in H}$ be a random process indexed by $(H, \|\cdot\|_{L_2})$ satisfying Equation (39). Then, there exists an absolute constant $c > 0$ such that for all $u > 0$*

$$\mathbb{P}\left(\sup_{h,g\in H} |X_h - X_g| \geq c(w(H) + uD_{L_2}(H)\right) \leq \exp(-u^2)$$

where $w(H)$ is the Gaussian mean width of H and $D_{L_2}(H)$ is the L_2-diameter.

The following Lemma allows to control the ψ_2-norm of a sum of independent centered random variables.

Lemma 4 ([11], Theorem 1.2.1)**.** *Let $X_1, \cdots, X_N$ be independent real random variables such that for all $i = 1, \cdots, N$, $\mathbb{E}X_i = 0$. Then*

$$\|\sum_{i=1}^N X_i\|_{\psi_2} \leq 16\left(\sum_{i=1}^N \|X_i\|_{\psi_2}^2\right)^{1/2}$$

The following Lemma connects ψ_2-bounded random variable with the control of its Laplace transform.

Lemma 5 ([11], Theorem 1.1.5)**.** *Let Z be a real valued random variable. The following assertions are equivalent*

- *There exists $K > 0$ such that $\|Z\|_{\psi_2} \leq K$*
- *There exist absolute constants $c_1, c_2, c_3 > 0$ such that for every $\lambda \geq c_1/K$*

$$\mathbb{E}\exp(\lambda|Z|) \leq c_3 \exp(c_2\lambda^2 K^2) \tag{40}$$

We are now in position to prove Lemma 3.

Proof. First we prove that $\Omega_{\mathcal{I}}$ holds with probability larger than $1-\delta/2$. Let us assume that for any f, g in $\mathcal{F}_r$, the following condition holds

$$\|(P-P_{\mathcal{I}})(\ell_f-\ell_g)\|_{\psi_2} \leq c(L/\sqrt{|\mathcal{I}|})\|f-g\|_{L_2} \ . \tag{41}$$

Then, from Theorem 9, there exists an absolute constant $c>0$ such that with probability larger than $1-\delta/2$

$$\begin{aligned}\sup_{f\in\mathcal{F}_r}\left|(P-P_{\mathcal{I}})(\ell_f-\ell_{f^*})\right| &\leq \sup_{f,g\in\mathcal{F}_r}\left|(P-P_{\mathcal{I}})(\ell_f-\ell_g)\right| \\ &\leq c\frac{L}{\sqrt{|\mathcal{I}|}}\big(w(\mathcal{F}_r)+\sqrt{\log(1/\delta)}D_{L_2}(\mathcal{F}_r)\big) \\ &\leq c\frac{L}{\sqrt{|\mathcal{I}|}}\big(w(\mathcal{F}_r)+\sqrt{\log(1/\delta)}r(A,\delta)\big) \ ,\end{aligned}$$

concluding the proof for $\Omega_{\mathcal{I}}$. Since $(X_o)_{o\in\mathcal{O}}$ are i.i.d as μ, with the same reasoning if we assume that

$$\big\|(P-P_{\mathcal{O}})|f-g|\big\|_{\psi_2} \leq (c/\sqrt{|\mathcal{O}|})\|f-g\|_{L_2} \ , \tag{42}$$

then, with probability larger than $1-\delta/2$:

$$\sup_{f\in\mathcal{F}_r}\left|(P-P_{\mathcal{O}})|f-f^*|\right| \leq \frac{c}{\sqrt{|\mathcal{O}|}}\big(w(\mathcal{F}_r)+\sqrt{\log(1/\delta)}r(A,\delta)\big) \ .$$

Thus, to finish the proof it remains to show that Equations (41) and (42) hold. From Lemma 4 we get

$$\begin{aligned}\|(P-P_{\mathcal{I}})(\ell_f-\ell_g)\|_{\psi_2} &\leq c\Big(\sum_{i\in\mathcal{I}}\frac{\|(\ell_f-\ell_g)(X_i,Y_i)-\mathbb{E}(\ell_f-\ell_g)(X_i,Y_i)\|_{\psi_2}^2}{|\mathcal{I}|^2}\Big)^{1/2} \\ &= \frac{c}{\sqrt{|\mathcal{I}|}}\|(\ell_f-\ell_g)(X,Y)-\mathbb{E}(\ell_f-\ell_g)(X,Y)\|_{\psi_2}\end{aligned}$$

Thus, it remains to show that $\|(\ell_f-\ell_g)(X,Y)-\mathbb{E}(\ell_f-\ell_g)(X,Y)\|_{\psi_2} \leq cL\|f-g\|_{L_2}$ for $c>0$ an absolute constant. To do so, we use Lemma 5. Let $\lambda \geq cL/(\|f-g\|_{L_2})$. From the symmetrization principle (Lemma 6.3 in [36]) and the contraction principle (Theorem 2.2 in [30]) we get

$$\begin{aligned}\mathbb{E}\exp(\lambda|(\ell_f-\ell_g)(X,Y)-\mathbb{E}(\ell_f-\ell_g)(X,Y)|) &\leq \mathbb{E}\exp(2\lambda\sigma(\ell_f-\ell_g)(X,Y)) \\ &\leq \mathbb{E}\exp(4L\lambda\sigma(f-g)(X)) \\ &\leq \mathbb{E}\exp(4L\lambda|f-g|(X))\end{aligned}$$

where σ is a Rademacher random variable, independent to (X,Y). From the sub-Gaussian Assumption we get

$$\mathbb{E}\exp(\lambda|(\ell_f-\ell_g)(X,Y)-\mathbb{E}(\ell_f-\ell_g)(X,Y)|) \leq \mathbb{E}\exp(16^2\lambda^2L^2\|f-g\|_{L_2}^2)$$

which concludes the proof for $\Omega_{\mathcal{I}}$ with Lemma 5. For $\Omega_{\mathcal{O}}$, we apply the same reasoning without the contraction step. □

Deterministic argument In this paragraph we place ourselves on the event $\Omega = \Omega_{\mathcal{I}} \cap \Omega_{\mathcal{O}}$. The main argument uses the convexity of the class F with the one of the loss function.

From the definition of $\hat{f}_N$, we have $P_N\mathcal{L}_{\hat{f}_N} \leq 0$. To show that $F \cap (f^* + r(A,\delta)\mathbb{B}_2)$ it is sufficient to show that for all functions $f \in F$ such that $F\backslash(f^* + r(A,\delta)\mathbb{B}_2)$ we have $P_N\mathcal{L}_f > 0$. Let f in $F\backslash(f^* + r(A,\delta)\mathbb{B}_2)$. By convexity of F there exists a function $f_1 \in f^* + r(A,\delta)\mathbb{S}_2$ for which

$$f - f^* = \alpha(f_1 - f^*)$$

where $\alpha = \big(\|f - f^*\|_{L_2}/r(A,\delta)\big) \geq 1$. For all $i \in \{1, \cdots, N\}$, let $\psi_i : \mathbb{R} \to \mathbb{R}$ be defined for all $u \in \mathbb{R}$ by

$$\psi_i(u) = \ell(u + f^*(X_i), Y_i) - \ell(f^*(X_i), Y_i).$$

The functions ψ_i are such that $\psi_i(0) = 0$, they are convex under assumption 3. In particular $\alpha\psi_i(u) \leq \psi_i(\alpha u)$ for all $u \in \mathbb{R}$ and $\alpha \geq 1$ and $\psi_i(f(X_i) - f^*(X_i)) = \ell(f(X_i), Y_i) - \ell(f^*(X_i), Y_i)$ so that the following holds:

$$\begin{aligned} P_N\mathcal{L}_f &= \frac{1}{N}\sum_{i=1}^N \psi_i\big(f(X_i) - f^*(X_i)\big) = \frac{1}{N}\sum_{i=1}^N \psi_i(\alpha(f_1(X_i) - f^*(X_i))) \\ &\geq \frac{\alpha}{N}\sum_{i=1}^N \psi_i(f_1(X_i) - f^*(X_i)) = \alpha P_N\mathcal{L}_{f_1}. \end{aligned}$$

From the previous argument it follows that $P_N\mathcal{L}_f \geq \alpha P_N\mathcal{L}_{f_1}$. Therefore it is enough to show that $P_N\mathcal{L}_{f_1} > 0$ for every $f_1 \in F \cap (f^* + r(A,\delta)\mathbb{S}_2)$. We have

$$P_N\mathcal{L}_{f_1} = \frac{|\mathcal{I}|}{N}P_{\mathcal{I}}\mathcal{L}_{f_1} + \frac{|\mathcal{O}|}{N}P_{\mathcal{O}}\mathcal{L}_{f_1}$$

On $\Omega_{\mathcal{I}}$ (see Equation (37)) it follows that

$$P_{\mathcal{I}}\mathcal{L}_{f_1} \geq P\mathcal{L}_{f_1} - \frac{cL}{\sqrt{|\mathcal{I}|}}\left(w(\mathcal{F}_r) + \sqrt{\log(1/\delta)}r(A,\delta)\right)$$

From assumption 4 and the definition $r(A,\delta)$ it follows that

$$\begin{aligned} P_{\mathcal{I}}\mathcal{L}_{f_1} &\geq \frac{r^2(A,\delta)}{A} - \frac{cL}{\sqrt{|\mathcal{I}|}}\left(\frac{\sqrt{|\mathcal{I}|}r^2(A,\delta)}{AL} + \sqrt{\log(1/\delta)}r(A,\delta)\right) \\ &= c\left(\frac{r^2(A,\delta)}{A} - Lr(A,\delta)\sqrt{\frac{\log(1/\delta)}{N}}\right) . \end{aligned}$$

From assumption 3, it follows that

$$P_{\mathcal{O}}\mathcal{L}_{f_1} \geq -P_{\mathcal{O}}|\ell_{f_1} - \ell_{f^*}| \geq -LP_{\mathcal{O}}|f_1 - f^*| .$$

On $\Omega_{\mathcal{O}}$ (see Equation (38)), we get

$$\frac{|\mathcal{O}|}{N}P_{\mathcal{O}}\mathcal{L}_{f_1} \geq -L\frac{|\mathcal{O}|}{N}\|f_1 - f^*\|_{L_1} - \frac{c\sqrt{|\mathcal{O}|}}{N}\left(w(\mathcal{F}_r) + r(A,\delta)\sqrt{\log(1/\delta)}\right)$$
$$\geq -L\frac{|\mathcal{O}|}{N}r(A,\delta) - c\left(\frac{r^2(A,\delta)}{A} + Lr(A,\delta)\sqrt{\frac{\log(1/\delta)}{N}}\right) .$$

Finally, from the definition of $r(A,\delta)$, we obtain

$$P_N\mathcal{L}_{f_1} \geq c\left(\frac{r^2(A,\delta)}{A} - Lr(A,\delta)\sqrt{\frac{\log(1/\delta)}{N}} - Lr(A,\delta)\frac{|\mathcal{O}|}{N}\right) > 0 ,$$

which concludes the proof for the error rate.

We finish the proof by establishing the result for the excess risk. Since $\|\hat{f}_N - f^*\|_{L_2} \leq r(A,\delta)$, on $\Omega_{\mathcal{I}}$ we have

$$P\mathcal{L}_{\hat{f}_N} \leq P_{\mathcal{I}}\mathcal{L}_{\hat{f}_N} + c\left(\frac{r^2(A,\delta)}{A} + Lr(A,\delta)\sqrt{\frac{\log(1/\delta)}{N}}\right)$$
$$= \frac{N}{|\mathcal{I}|}P_N\mathcal{L}_{\hat{f}_N} - \frac{|\mathcal{O}|}{|\mathcal{I}|}P_{\mathcal{O}}\mathcal{L}_{\hat{f}_N} + c\left(\frac{r^2(A,\delta)}{A} + Lr(A,\delta)\sqrt{\frac{\log(1/\delta)}{N}}\right)$$
$$\leq -\frac{|\mathcal{O}|}{|\mathcal{I}|}P_{\mathcal{O}}\mathcal{L}_{\hat{f}_N} + c\left(\frac{r^2(A,\delta)}{A} + Lr(A,\delta)\sqrt{\frac{\log(1/\delta)}{N}}\right) .$$

On $\Omega_{\mathcal{O}}$ we have

$$L\frac{|\mathcal{O}|}{|\mathcal{I}|}P_{\mathcal{O}}|\hat{f}_N - f^*| \leq L\frac{|\mathcal{O}|}{|\mathcal{I}|}r(A,\delta) + c\left(\frac{r^2(A,\delta)}{A} + Lr(A,\delta)\sqrt{\frac{\log(1/\delta)}{N}}\right) .$$

Finally, from the definition of $r(A,\delta)$ we obtain

$$P\mathcal{L}_{\hat{f}_N} \leq c\frac{r^2(A,\delta)}{A}$$

D.2. Proof Theorem 1 in the local bounded framework

The deterministic argument is exactly the same as for the sub-Gaussian case. Recall that $r^b(A,\cdot)$ is defined as

$$r(A,\delta) \geq c\left(r^B_{\mathcal{I}}(A) \vee AL\sqrt{\frac{\log(1/\delta)}{N}} \vee AL\frac{|\mathcal{O}|}{N}\right) ,$$

and satisfies Assumption 4 for $A > 0$. Recall that $\mathcal{F}_r = F \cap (f^* + r(A,\delta)\mathbb{B}_2)$. To study the stochastic argument we use the following result:

Theorem 10 (Theorem 2.6, [29]). *Let $\mathcal{F}$ be a class of functions bounded by M. For all $t>0$, with probability larger than $1-\exp(-t)$*

$$\sup_{f\in\mathcal{F}}|(P_N-P)f| \leq \mathbb{E}\sup_{f\in\mathcal{F}}|(P_N-P)f| + \sqrt{2\frac{t}{N}\left(\sup_{f\in\mathcal{F}}Pf^2+2M\mathbb{E}\sup_{f\in\mathcal{F}}|(P_N-P)f|\right)}+\frac{tM}{N}$$

Let us define:

$$\Omega_{\mathcal{I}} := \left\{ \sup_{f\in\mathcal{F}_r}\left|(P-P_{\mathcal{I}})\mathcal{L}_f\right| \leq c\frac{r^2(A,\delta)}{A}\right\}$$
$$\Omega_{\mathcal{O}} := \left\{ \sup_{f\in\mathcal{F}_r}\left|(P-P_{\mathcal{O}})|f-f^*|\right| \leq c\frac{|\mathcal{I}|}{|\mathcal{O}|}\frac{r^2(A,\delta)}{AL}\right\}$$

Lemma 6. *Grant Assumptions 1, 3 and 2. Assume than for all $f\in F\cap(f^*+r(A,\delta)\mathbb{B}_2)$ and $x\in\mathcal{X}:|f(x)-f^*(x)|\leq 1$. Then, the event $\Omega=\Omega_{\mathcal{I}}\cup\Omega_{\mathcal{O}}$ holds with probability larger than $1-\delta$.*

Proof. Let $(\sigma_i)_{i\in\mathcal{I}}$ be i.i.d Rademacher random variables, assumed to be independent to $(X_i)_{i\in\mathcal{I}}$, from the symmetrization and contraction Lemmas (see [36])

$$\mathbb{E}\sup_{f\in\mathcal{F}_r}|(P_{\mathcal{I}}-P)\mathcal{L}_f| \leq 4L\mathbb{E}\sup_{f\in\mathcal{F}}\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}}\sigma_i(f-f^*)(X_i) \leq c\frac{r^2(A,\delta)}{A}$$

where we used the of $r^B_{\mathcal{I}}(\cdot)$ and the fact that $r^b(A,\delta)\geq r^B_{\mathcal{I}}(A)$ for all $A>0$. Under the local bounded assumption, any function f in $\mathcal{F}_r$, $|\mathcal{L}_f(x,y)|\leq L$ for all $(x,y)\in\mathcal{X}\times\mathcal{Y}$. For any $t>0$, it follows from Theorem 10 that for any function f in $\mathcal{F}_r$

$$\begin{aligned}&|(P_{\mathcal{I}}-P)\mathcal{L}_f|\\ &\leq c\left[\frac{(r^b(A,\delta))^2}{A}+\frac{L\log(1/\delta)}{N}+L\sqrt{\frac{\log(1/\delta)}{|\mathcal{I}|}\left((r^b(A,\delta))^2+\frac{(r^b(A,\delta))^2}{AL}\right)}\right]\\ &\leq c\frac{(r^b(A,\delta))^2}{A}\ .\end{aligned}$$

The proof for $\Omega_{\mathcal{O}}$ uses exactly the same arguments. □

D.3. Proof Theorem 3 in the sub-Gaussian framework

Recall that $\tilde{r}(A,\rho^*,\delta)$ is such that:

$$\tilde{r}(A,\rho^*,\delta)\geq\tilde{r}^{SG}_{\mathcal{I}}(A,\rho^*)\vee AL\sqrt{\frac{\log(1/\delta)}{N}}\vee AL\frac{|\mathcal{O}|}{N}\ ,$$

where ρ^* satisfying the A,δ-sparsity equation with A verifying assumption 5.

The proof is split into two parts and is very similar as the one of Theorem 1. First we identify a stochastic argument holding with large probability. Then, we show on that event that $\hat{f}_N^\lambda \in F \cap \big(f^* + \rho^* \mathbb{B} \cap \tilde{r}(A, \rho^*, \delta)\mathbb{B}_2\big)$. Then, at the very end of the proof we will control the excess risk $P\mathcal{L}\hat{f}_N^\lambda$ where $\hat{f}_N^\lambda$ is defined in equation (14).

Stochastic arguments The stochastic part is the same as the one in the proof of Theorem 1 where a localization with respect to the regularization norm is added. First we identifiate the stochastic event onto which the proof easily follows. Let $\mathcal{F}_{r,\rho} = F \cap \big(f^* + \rho^* \mathbb{B} \cap \tilde{r}(A, \rho^*, \delta)\mathbb{B}_2\big)$ and define

$$\Omega_{\mathcal{I}} = \sup_{f \in \mathcal{F}_{r,\rho}} \Big| \big(P - P_{\mathcal{I}}\big)\big(\ell_f - \ell_{f^*}\big) \Big| \leq c \frac{L}{\sqrt{|\mathcal{I}|}} \left(w(\mathcal{F}_{r,\rho}) + \tilde{r}(A, \rho^*, \delta)\sqrt{\log(1/\delta)} \right)$$
$$\Omega_{\mathcal{O}} = \sup_{f \in \mathcal{F}_{r,\rho}} \Big| \big(P - P_{\mathcal{O}}\big)|f - f^*| \Big| \leq \frac{c}{\sqrt{|\mathcal{O}|}} \left(w(\mathcal{F}_{r,\rho}) + \tilde{r}(A, \rho^*, \delta)\sqrt{\log(1/\delta)} \right)$$

Finally, set $\Omega = \Omega_{\mathcal{I}} \cap \Omega_{\mathcal{O}}$

Lemma 7. *Grant Assumptions 1, 2, 3 and assume that $F - f^*$ is 1-sub-Gaussian. Then the event Ω holds with probability larger than $1 - \delta$*

Proof. The proof is exactly the same as the one in the non-regularized setup where a localization with respect to the regularization norm is added. It is enough to adapt the proof with the definition of $\tilde{r}(A, \rho^*, \delta)$ from Equation (3). □

Deterministic argument In this paragraph we place ourselves on the event Ω. Let us recall that for any function f in F

$$P_N \mathcal{L}_f^\lambda = P_N(\ell_f - \ell_{f^*}) + \lambda(\|f\| - \|f^*\|) \tag{43}$$

From the definition of $\hat{f}_N^\lambda$, we have $P_N \mathcal{L}_{\hat{f}_N^\lambda}^\lambda \leq 0$. To show that $\hat{f}_N^\lambda \in \mathcal{F}_{r,\rho}$ it is sufficient to show that for all functions $f \in F \backslash \mathcal{F}_{r,\rho}$ we have $P_N \mathcal{L}_f^\lambda > 0$. Let f in $f \in F \backslash \mathcal{F}_{r,\rho}$. By convexity of F there exist a function f_1 in F and $\alpha \geq 1$ such that $\alpha(f_1 - f^*) = f - f^*$ and $f_1 \in \partial\mathcal{F}_{r,\rho}$ where $\partial\mathcal{F}_{r,\rho}$ denotes the border of $\mathcal{F}_{r,\rho}$. Using the same convex argument as the one in the proof of Theorem 1 we obtain:

$$P_N \mathcal{L}_f \geq \alpha P_N \mathcal{L}_{f_1} \ ,$$

for $f_1 \in \partial\mathcal{F}_{r,\rho}$. Moreover, by the triangular inequality we obtain

$$\|f\| - \|f^*\| \geq \alpha(\|f_1\| - \|f^*\|),$$

and thus,

$$P_N \mathcal{L}_f^\lambda \geq \alpha P_N \mathcal{L}_{f_1}^\lambda$$

Therefore it is enough to show that $P_N \mathcal{L}_{f_1}^\lambda > 0$ for $f_1 \in \partial\mathcal{F}_{r,\rho}$. By definition of $\partial\mathcal{F}_{r,\rho}$, there are two different cases 1) $f_1 \in F \cap \big(f^* + \rho^* \mathbb{S} \cap \tilde{r}(A, \rho^*, \delta)\mathbb{B}_2\big)$ and

2) $f_1 \in F \cap \big(f^* + \rho^* \mathbb{B} \cap \tilde{r}(A, \rho^*, \delta)\mathbb{S}_2\big)$. In 1) the sparsity equation will help us to show that $P_N \mathcal{L}^{\lambda}_{f_1} > 0$ while in 2), it will be the local Bernstein condition.

Let us begin by the case 1). Let $f_1 \in F \cap \big(f^* + \rho^* \mathbb{S} \cap \tilde{r}(A, \rho^*, \delta)\mathbb{B}_2\big)$.

$$P_N \mathcal{L}_{f_1} = \frac{|\mathcal{I}|}{N} P_{\mathcal{I}} \mathcal{L}_{f_1} + \frac{|\mathcal{O}|}{N} P_{\mathcal{O}} \mathcal{L}_{f_1} \geq \frac{|\mathcal{I}|}{N} P_{\mathcal{I}} \mathcal{L}_{f_1} - L \frac{|\mathcal{O}|}{N} P_{\mathcal{O}} |f_1 - f^*|$$

On $\Omega_{\mathcal{I}}$ if holds that

$$P_{\mathcal{I}} \mathcal{L}_{f_1} \geq \frac{|\mathcal{I}|}{N} \bigg[- c \frac{L}{\sqrt{|\mathcal{I}|}} \Big(w\big(\mathcal{F}_{r,\rho}\big) + \tilde{r}(A, \rho^*, \delta) \sqrt{\log(1/\delta)} \Big) \bigg] \geq - c \frac{\tilde{r}^2(A, \rho^*, \delta)}{A} \ , \tag{44}$$

where we used the definitions of $\tilde{r}^B_{\mathcal{I}}(A, \rho^*)$ and $\tilde{r}(A, \rho^*, \delta)$.

On $\Omega_{\mathcal{O}}$, it holds that

$$\begin{aligned}
&- L \frac{|\mathcal{O}|}{N} P_{\mathcal{O}} |f_1 - f^*| \\
&\geq - L \frac{|\mathcal{O}|}{N} \bigg[\|f_1 - f^*\|_{L_1(\mu)} - \frac{c}{\sqrt{|\mathcal{O}|}} \Big(w\big(\mathcal{F}_{r,\rho}\big) + \tilde{r}(A, \rho^*, \delta) \sqrt{\log(1/\delta)} \Big) \bigg] \\
&\geq - L \frac{|\mathcal{O}|}{N} \tilde{r}(A, \rho, \delta) - c \frac{\tilde{r}^2(A, \rho, \delta)}{A} \\
&\geq - c \frac{\tilde{r}^2(A, \rho, \delta)}{A} \ ,
\end{aligned}$$

where we also used the definitions of $\tilde{r}^B_{\mathcal{I}}(A, \rho^*)$ and $\tilde{r}(A, \rho^*, \delta)$ and the fact that $|\mathcal{O}| \leq |\mathcal{I}|$. Consequently, we get

$$P_N \mathcal{L}_{f_1} \geq - c \frac{\tilde{r}^2(A, \rho, \delta)}{A} \ ,$$

where the constant $c > 0$ is chosen such that $c < 7/17$.

Let us turn to the control of $\lambda(\|f_1\| - \|f^*\|)$. Recall that we are in the case where $\|f_1 - f^*\| = \rho^*$ and $\|f_1 - f^*\|_{L_2} \leq \tilde{r}(A, \rho^*, \delta)$. Let $v \in E$ be such that $\|f^* - v\| \leq \rho^*/20$ and $g \in \partial(\|\cdot\|)_v$. We have

$$\begin{aligned}
&\|f_1\| - \|f^*\| \geq \|f_1\| - \|v\| - \|f^* - v\| \geq \big\langle g, f_1 - v \big\rangle - \|f^* - v\| \\
&\geqslant \big\langle g, f_1 - f^* \big\rangle - 2 \|f^* - v\| \geqslant \big\langle g, f_1 - f^* \big\rangle - \rho^*/10 \ .
\end{aligned}$$

As the latter result holds for all $v \in f^* + (\rho^*/20)\mathbb{B}$ and $g \in \partial \|\cdot\|(v)$, since $f_1 \in \mathcal{F}_{r,\rho}$, we get

$$\|f_1\| - \|f^*\| \geq \Delta(\rho^*) - \rho^*/10 \geq 7\rho^*/10 \ ,$$

where the last inequality holds because ρ^* satisfies the sparsity equation. Finally we obtain

$$P_N \mathcal{L}_{f_1}^{\lambda} \geq -c\frac{\tilde{r}^2(A,\rho,\delta)}{A} - \frac{7}{10}\lambda\rho^* > 0 \ , \tag{45}$$

when $\lambda \geq (10c/7)(\tilde{r}^2(A,\rho,\delta)/(A\rho^*))$.

Let us turn to the second case 2). Let $f_1 \in F \cap \big(f^* + \rho^*\mathbb{B} \cap \tilde{r}(A,\rho^*,\delta)\mathbb{S}_2\big)$. With the same analysis as 1), on Ω, from Assumption 5 it follows that

$$P_N \mathcal{L}_{f_1} \geq \frac{\tilde{r}^2(A,\rho^*,\delta)}{A} - c\frac{\tilde{r}^2(A,\rho^*,\delta)}{A} \ ,$$

where the constant c is the same as the one appearing in Equation (45). As $\|f_1\| - \|f^*\| \geq -\|f_1 - f^*\| \geq -\rho^*$, it follows that

$$P_N \mathcal{L}_{f_1}^{\lambda} \geq (1-c)\frac{\tilde{r}^2(A,\rho^*,\delta)}{A} - \lambda\rho^* > 0 \ ,$$

when $\lambda \leq (1-c)(\tilde{r}^2(A,\rho^*,\delta)/(A\rho^*))$. Note that the condition $c < 7/17$ implies that such a $\lambda > 0$ exists.

We finish the proof by establishing the result for the excess risk. Since $\hat{f}_N^{\lambda} \in F \cap \big(f^* + \rho^*\mathbb{B} \cap \tilde{r}(A,\rho^*,\delta)\mathbb{B}_2\big)$, on Ω

$$P\mathcal{L}_{\hat{f}_N^{\lambda}} \leq P_{\mathcal{I}}\mathcal{L}_{\hat{f}_N^{\lambda}} + c\frac{\tilde{r}^2(A,\rho^*,\delta)}{A}$$

Moreover we have

$$\begin{aligned}
P_{\mathcal{I}}\mathcal{L}_{\hat{f}_N^{\lambda}} &= \frac{N}{|\mathcal{I}|}P_N\mathcal{L}_{\hat{f}_N^{\lambda}} - \frac{|\mathcal{O}|}{|\mathcal{I}|}P_{\mathcal{O}}\mathcal{L}_{\hat{f}_N^{\lambda}} \\
&= \frac{N}{|\mathcal{I}|}P_N\mathcal{L}_{\hat{f}_N^{\lambda}}^{\lambda} + \lambda\frac{N}{|\mathcal{I}|}(\|f^*\| - \|\hat{f}_N^{\lambda}\|) - \frac{|\mathcal{O}|}{|\mathcal{I}|}P_{\mathcal{O}}\mathcal{L}_{\hat{f}_N^{\lambda}} \\
&\leq 2\lambda\rho^* + L\frac{|\mathcal{O}|}{|\mathcal{I}|}P_{\mathcal{O}}|\hat{f}_N^{\lambda} - f^*| \\
&\leq 2\lambda\rho^* + c\frac{\tilde{r}^2(A,\rho^*,\delta)}{A} \\
&= c\frac{\tilde{r}^2(A,\rho^*,\delta)}{A} \ ,
\end{aligned}$$

which concludes the proof for the excess risk.

D.4. Proof Theorem 3 in the local bounded setting

The proof consists in taking the stochastic argument from the proof of Theorem 1 (and adding the localization with respect to the regularization norm) and the deterministic argument from the proof of Theorem 3.

References

[1] Pierre Alquier, Vincent Cottet, Guillaume Lecué, et al. Estimation bounds and sharp oracle inequalities of regularized procedures with Lipschitz loss functions. *The Annals of Statistics*, 47(4):2117–2144, 2019. MR3953446

[2] Peter L. Bartlett, Olivier Bousquet, Shahar Mendelson, et al. Local Rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537, 2005. MR2166554

[3] Peter L. Bartlett and Shahar Mendelson. Empirical minimization. *Probability Theory and Related Fields*, 135(3):311–334, 2006. MR2240689

[4] Peter L. Bartlett and Shahar Mendelson. Empirical minimization. *Probab. Theory Related Fields*, 135(3):311–334, 2006. MR2240689

[5] Pierre C. Bellec, Guillaume Lecué, Alexandre B. Tsybakov, et al. Slope meets lasso: improved oracle bounds and optimality. *The Annals of Statistics*, 46(6B):3603–3642, 2018. MR3852663

[6] Kush Bhatia, Prateek Jain, and Purushottam Kar. Robust regression via hard thresholding. In *Advances in Neural Information Processing Systems*, pages 721–729, 2015.

[7] Olivier Bousquet, Vladimir Koltchinskii, and Dmitriy Panchenko. Some local measures of complexity of convex hulls and generalization bounds. In *International Conference on Computational Learning Theory*, pages 59–73. Springer, 2002. MR2040405

[8] Stephen Boyd, Stephen P. Boyd, and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004. MR2061575

[9] Pierre C. Bellec. Localized gaussian width of m-convex hulls with applications to lasso and convex aggregation. *Bernoulli*, 25(4A):3016–3040, 2019. MR4003572

[10] Andrea Caponnetto and Ernesto De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007. MR2335249

[11] Djalil Chafaï, Olivier Guédon, Guillaume Lecué, and Alain Pajor. *Interactions between compressed sensing random matrices and high dimensional geometry*. Citeseer, 2012. MR3113826

[12] Mengjie Chen, Chao Gao, Zhao Ren, et al. A general decision theory for huber's epsilon-contamination model. *Electronic Journal of Statistics*, 10(2):3752–3774, 2016. MR3579675

[13] Mengjie Chen, Chao Gao, Zhao Ren, et al. Robust covariance and scatter matrix estimation under huber's contamination model. *The Annals of Statistics*, 46(5):1932–1960, 2018. MR3845006

[14] Yudong Chen, Lili Su, and Jiaming Xu. Distributed statistical machine learning in adversarial settings: Byzantine gradient descent. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 1(2):44, 2017.

[15] Yu Cheng, Ilias Diakonikolas, and Rong Ge. High-dimensional robust mean estimation in nearly-linear time. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2755–2771. SIAM,

2019. MR3909640

[16] Geoffrey Chinot. Robust learning and complexity dependent bounds for regularized problems. *arXiv preprint* arXiv:1902.02238, 2019.

[17] Geoffrey Chinot, Guillaume Lecué, and Matthieu Lerasle. Robust high dimensional learning for Lipschitz and convex losses. arXiv:1905.04281, 2019. MR4087486

[18] Geoffrey Chinot, Guillaume Lecué, and Matthieu Lerasle. Robust statistical learning with Lipschitz and convex loss functions. *Probability Theory and Related Fields*, Jul 2019. MR4087486

[19] Arnak Dalalyan and Philip Thompson. Outlier-robust estimation of a sparse linear model using l1-penalized Huber's M-estimator. In *Advances in Neural Information Processing Systems*, pages 13188–13198, 2019.

[20] Arnak S. Dalalyan, Mohamed Hebiri, Johannes Lederer, et al. On the prediction performance of the lasso. *Bernoulli*, 23(1):552–581, 2017. MR3556784

[21] Ilias Diakonikolas, Gautam Kamath, Daniel Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robust estimators in high-dimensions without the computational intractability. *SIAM Journal on Computing*, 48(2):742–864, 2019. MR3945261

[22] Ilias Diakonikolas, Weihao Kong, and Alistair Stewart. Efficient algorithms and lower bounds for robust linear regression. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2745–2754. SIAM, 2019. MR3909639

[23] Frank R. Hampel. A general qualitative definition of robustness. *The Annals of Mathematical Statistics*, pages 1887–1896, 1971. MR0301858

[24] Frank R. Hampel. The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, 69(346):383–393, 1974. MR0362657

[25] Samuel B. Hopkins et al. Mean estimation with sub-gaussian rates in polynomial time. *Annals of Statistics*, 48(2):1193–1213, 2020. MR4102693

[26] P. J. Huber and E. Ronchetti. Robust statistics. In *International Encyclopedia of Statistical Science*, pages 1248–1251. Springer, 2011. MR2488795

[27] Peter J. Huber. Robust estimation of a location parameter. *Breakthroughs in Statistics*, pages 492–518, 1992.

[28] Peter J. Huber et al. The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 221–233. University of California Press, 1967. MR0216620

[29] Vladimir Koltchinskii. Empirical and Rademacher processes. In *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems*, pages 17–32. Springer, 2011. MR2829871

[30] 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]. MR2829871

[31] Vladimir Koltchinskii et al. Local Rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, 34(6):2593–2656, 2006. MR2329442

[32] Guillaume Lecué and Jules Depersin. Robust subgaussian estimation of a mean vector in nearly linear time. *arXiv preprint* arXiv:1906.03058, 2019.

[33] Guillaume Lecué, Matthieu Lerasle, et al. Robust machine learning by median-of-means: theory and practice. *Annals of Statistics*, 48(2):906–931, 2020. MR4102681

[34] Guillaume Lecué, Matthieu Lerasle, and Timlothée Mathieu. Robust classification via mom minimization. *Machine Learning*, pages 1–31, 2020. MR4137195

[35] Guillaume Lecué and Shahar Mendelson. Regularization and the small-ball method I: Sparse recovery. *Ann. Statist.*, 46(2):611–641, 2018. MR3782379

[36] Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*. Springer Science & Business Media, 2013. MR2814399

[37] Zhixian Lei, Kyle Luh, Prayaag Venkat, and Fred Zhang. A fast spectral algorithm for mean estimation with sub-gaussian rates. *arXiv preprint* arXiv:1908.04468, 2019.

[38] Liu Liu, Yanyao Shen, Tianyang Li, and Constantine Caramanis. High dimensional robust sparse regression. *arXiv preprint* arXiv:1805.11643, 2018.

[39] Gábor Lugosi, Shahar Mendelson, et al. Sub-gaussian estimators of the mean of a random vector. *The Annals of Statistics*, 47(2):783–794, 2019. MR3909950

[40] Enno Mammen and Alexandre B. Tsybakov. Smooth discrimination analysis. *Ann. Statist.*, 27(6):1808–1829, 1999. MR1765618

[41] Ricardo Antonio Maronna. Robust m-estimators of multivariate location and scatter. *The Annals of Statistics*, pages 51–67, 1976. MR0388656

[42] Shahar Mendelson. On the performance of kernel classes. *Journal of Machine Learning Research*, 4(Oct):759–771, 2003. MR2075996

[43] Shahar Mendelson, Joseph Neeman, et al. Regularization in kernel learning. *The Annals of Statistics*, 38(1):526–565, 2010. MR2590050

[44] Stanislav Minsker et al. Geometric median and robust estimation in Banach spaces. *Bernoulli*, 21(4):2308–2335, 2015. MR3378468

[45] Alexander V. Nazin, Arkadi S. Nemirovsky, Alexandre B. Tsybakov, and Anatoli B. Juditsky. Algorithms of robust stochastic optimization based on mirror descent method. *Automation and Remote Control*, 80(9):1607–1627, 2019.

[46] Steve Smale and Ding-Xuan Zhou. Learning theory estimates via integral operators and their approximations. *Constructive Approximation*, 26(2):153–172, 2007. MR2327597

[47] Ingo Steinwart and Andreas Christmann. *Support Vector Machines*. Springer Science & Business Media, 2008. MR2450103

[48] Michel Talagrand. *Upper and Lower Bounds for Stochastic Processes: Modern Methods and Classical Problems*, volume 60. Springer Science & Business Media, 2014. MR3184689

[49] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996. MR1379242

[50] Alexandre B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *Ann. Statist.*, 32(1):135–166, 2004. MR2051002

[51] John W. Tukey. A survey of sampling from contaminated distributions. *Contributions to Probability and Statistics*, pages 448–485, 1960. MR0120720

[52] John W. Tukey. The future of data analysis. *The Annals of Mathematical Statistics*, 33(1):1–67, 1962. MR0133937

[53] Sara Van de Geer. Estimation and Testing Under Sparsity. *Lecture Notes in Mathematics*, vol. 2159, 2016. MR3526202

[54] Sara van de Geer. *Estimation and Testing Under Sparsity*, volume 2159 of *Lecture Notes in Mathematics*. Springer, Cham, 2016. Lecture notes from the 45th Probability Summer School held in Saint-Four, 2015, École d'Été de Probabilités de Saint-Flour. [Saint-Flour Probability Summer School]. MR3526202

[55] Sara A. Van De Geer, Peter Bühlmann, et al. On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics*, 3:1360–1392, 2009. MR2576316

[56] Vladimir Naumovich Vapnik. *Statistical Learning Theory*, volume 1. Wiley New York, 1998. MR1641250

[57] Roman Vershynin. High-Dimensional Probability: An Introduction with Applications in Data Science. *Cambridge Series in Statistical and Probabilistic Mathematics*, 2018. MR3837109

[58] Victor J. Yohai and Ricardo A. Maronna. Asymptotic behavior of m-estimators for the linear model. *The Annals of Statistics*, pages 258–268, 1979. MR0520237