Localization methods with applications to robust learning and interpolation

Geoffrey Chinot

INSTITUT POLYTECHNIQUE DE PARIS

ENSAE

IP PARIS

NNT : 2020IPPAG002

Thèse de doctorat

Localization methods with applications to robust learning and interpolation

Thèse de doctorat de l'Institut Polytechnique de Paris
préparée à l'École Nationale de la Statistique et de l'Administration Économique

École doctorale n°574 École Doctorale de Mathématiques Hadamard (EDMH)
Spécialité de doctorat : Mathématiques fondamentales

Thèse présentée et soutenue à Palaiseau, le 22/06/2020, par

GEOFFREY CHINOT

Composition du Jury :

Yannick Baraud Professeur, University of Luxembourg	Examinateur
Alexandra Carpentier Professeur, Universität Magdeburg	Examinateur
Guillaume Lecué Professeur, ENSAE	Directeur de thèse
Gabor Lugosi Professeur, Pompeu Fabra University	Rapporteur
Alexandre Tsybakov Professeur, ENSAE	Président
Sara Van De Geer Professeur, ETH Zurich	Rapporteur

"Vous savez, moi je ne crois pas qu'il y ait de bonne ou de mauvaise situation."

Otis alias Edouard Baer

Remerciements

Mes premières pensées vont naturellement à Matthieu et Guillaume, mes deux directeurs de thèse. Je tiens particulièrement à vous remercier pour tout ce que vous m'avez apporté, tant sur le plan humain que scientifique. Vous avez su me faire confiance et j'espère ne pas vous avoir déçu. J'ai longtemps erré sans savoir quoi faire dans ma vie. Cette thèse m'a apporté une réponse claire à cette question. Tout est limpide maintenant, je suis si épanoui en ce moment que je ne pourrais imaginer faire autre chose. Encore une fois, merci à vous deux. J'espère que ma soutenance sera simplement la fin du premier chapitre dans le livre de nos collaborations.

Gabor Lugosi et Sara van de Geer m'ont fait l'immense honneur d'accepter de rapporter cette thèse. Je ne saurais assez les remercier. C'est pour moi un immense honneur.

Merci également à tout ceux qui m'ont accompagné dans cette folle aventure. Ce ne fut pas tous les jours facile, mais grace à un soutien sans faille, je n'ai jamais baissé la tête. J'aimerais particulièrement te remercier Arnak. Tu es pour moi un modèle de chercheur (et aussi de footballeur). Tu as toujours été là avec tes bons conseils et ta bienveillance. Je voudrais aussi te remercier Sacha, sans toi je n'aurais peut être jamais poussé la porte de la recherche. Cristina, pour ta disponibilité et ta gentillesse. Nicolas pour tous ces cafés remplis de bons conseils. Victor, pour tout ton temps que tu nous offres pour soutenir cette thèse. Pierre, toi qui est parti pour de folles aventures nippones. Cette dernière année fut bien triste sans toi.
Un grand merci à toi Badr, nous avons commencé cette aventure ensemble et la finissons ensemble. Ta bonne humeur est un moteur pour tous les gens qui t'entourent. J'espère que tu apporteras ce rayon de soleil qui te suit en Angleterre. Ils en ont bien besoin !
Merci à tous ceux qui m'ont supporté dans ce bureau 3017, Simo, Gautier, Suzanne et Julien. J'espère ne pas avoir été trop fatiguant. Merci à tous mes compagnons de thèse: Aurélien, Arshak, Nicolas, Jérémy, Léna, Solenne, Lucie, Mamadou, Suzanne, Julien, Mehdi, Alexander, Philip, Avo, Amir, Christophe, Boris, Alexis, Gabriel, François-Pierre, Flore et Meyer.
Merci à mes "co-thésards": Jules, Lucie et Timothée. Nos échanges furent toujours constructifs et très utiles.

Finalement, merci à tous mes proches pour votre soutien. Gautlhier, même si pour toi "je ne travaillerai jamais", tu as su être là pour moi. Merci à vous, Papa et Maman, ça fait bien longtemps que vous ne comprenez plus rien à ma vie, mais tant que vous comprendrez que votre amour me suffit vous aurez tout compris.

Résumé substantiel

Le monde est actuellement en pleine mutation. Beaucoup de transformations majeures des dernières décennies sont directement ou indirectement liées à l'apprentissage statistique. Dans divers secteurs, dont la santé, les sciences, l'éducation et la publicité, les statistiques permettent de résoudre des problèmes qui étaient jusqu'alors inatteignables. Cependant, beaucoup de méthodes statistiques ont été imaginées dans un cadre ancien, où les bases de données étaient de petites tailles. La grande dimension, maintenant omniprésente, apporte de nouveaux challenges. Premièrement, il essentiel de verifier si une méthode est "scalable", c'est-à-dire, applicable à des jeux de données de grandes tailles. Ensuite, les bases de données de grandes dimensions sont susceptibles d'être très corrompues. Dans ce cas, il est essentiel de construire une procédure "robuste", c'est-à-dire fiable lorsque des données aberrantes peuvent contaminer l'information disponible. Plus généralement, la robustesse peut être défini comme la résistance d'une procédure aux hypothèses. Par example, il est souvent commun de supposer que les données sont toutes indépendantes et identiquement distribuées. Que se passe t-il si certaines données sont corrompues ? D'autre part, le cadre théorique de l'apprentissage statistique est basé sur la théorie des probabilités. Les données sont supposées aléatoires. Que ce passe t-il lorsque la variance des données est grande ? Peut-on tout de même en tirer une certaine information ? L'objectif de cette thèse est d'apporter une réponse aux questions précédentes. Nous développons et étudions les propriétés théoriques de différents estimateurs robustes.

Soit $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ un vecteur aléatoire distribué selon P, supposée inconnue. Soit $\mathcal{F}$ une classe de prédicteurs, c'est-à-dire, une classe de fonctions mesurables $f : \mathcal{X} \mapsto \mathcal{Y}$. L'objectif principal de l'apprentissage statistique est de prédire la sortie Y à partir de $f(X)$, pour f dans $\mathcal{F}$. Pour mesurer la qualité de prédiction d'un prédicteur f, nous introduisons une fonction de perte $\ell : \mathcal{Y} \times \mathcal{Y} \mapsto \mathbb{R}_+$ où $\ell(f(X), Y)$ quantifie l'erreur de prédire $f(X)$ alors que le vrai label est Y. Une règle naturelle est de chercher la fonction $f^*_{\mathcal{F}}$ dans $\mathcal{F}$ minimisant le risque intégré, c'est-à-dire l'erreur moyenne de $\ell(f(X), Y)$ par rapport à la distribution P

$$f^*_{\mathcal{F}} \in \operatorname*{argmin}_{f \in \mathcal{F}} R_P(f), \quad \text{où} \quad R_P(f) = \mathbb{E}_{(X,Y)\sim P}\big[\ell\big(f(X), Y\big)\big].$$

P étant inconnue, *l'oracle* $f^*_{\mathcal{F}}$ ne peut être seulement qu'approximé. Pour cela le statisticien dispose d'un jeu de données $\mathcal{D} = (X_i, Y_i)_{i \in [\![1,n]\!]}$ de n observations supposées indépendantes et identiquement distribuées selon P. Une approche très répandue, consiste à remplacer le risque par sa version empirique et la minimiser dans $\mathcal{F}$. Cela se nomme la *minimisation du risque empirique*

$$\hat{f}_n \in \operatorname*{argmin}_{f \in \mathcal{F}} R_n(f), \quad \text{où} \quad R_n(f) = \frac{1}{n} \sum_{i=1}^{n} \ell\big(f(X_i), Y_i\big).$$

La principale contribution de cette thèse est la demonstration mathématique que, lorsque la fonction de perte ℓ est Lipschitz, le minimiseur du risque empirique $\hat{f}_n$ est robuste au "bruit" du problème $Y - f^*_{\mathcal{F}}(X)$ et à un nombre non négligeable de données aberrantes contaminant les variables aléatoires $(Y_i)_{i\in[\![1,n]\!]}$. L'analyse est étendue au minimiseur du risque empirique pénalisé, très répandu chez les praticiens (elastic-net, support-vector-machine, Lasso). Nous développons un nouvel argument d'homogénéité, permettant de localiser l'analyse autour de la solution que l'on cherche à approximer: l'oracle $f^*_{\mathcal{F}}$. Notre approche est générale et permet d'obtenir des résultats optimaux pour de nombreux problèmes bien connues en statistiques.

Cependant le minimiseur du risque empirique n'est pas fiable lorsque la classe de prédicteurs $\mathcal{F}$ n'est pas bornée. Lorsque $\mathcal{F}$ et ℓ ne sont pas bornées, il est nécessaire d'imposer de fortes conditions sur l'enveloppe $\{\ell(f(X),Y), f \in \mathcal{F}\}$ et la distribution P des données. Ces hypothèses sont trop contraignantes et souvent non vérifiées en pratique.
Pour relacher ces hypothèses sur l'enveloppe $\{\ell(f(X),Y), f \in \mathcal{F}\}$, nous étudions les estimateurs minmax-Median Of Means. Soit K un entier tel que K divise n (pour simplifier). Soit $B_1, \cdots, B_K$ une partition de $[\![1,n]\!]$ en K blocks de même taille n/K. Pour tout k dans $[\![1,K]\!]$ et $f \in \mathcal{F}$, soit $P_{B_k}\ell_f = (K/n)\sum_{i\in B_k}\ell(f(X_i),Y_i)$ le risque empririque sur le block B_k. L'estimateur minmax-MOM est défini comme

$$\hat{f}_K^{\mathrm{MOM}} \in \operatorname*{argmin}_{f\in\mathcal{F}} \sup_{g\in\mathcal{F}} \mathrm{Med}\big(P_{B_1}(\ell_f - \ell_g), \cdots, P_{B_K}(\ell_f - \ell_g)\big),$$

où $\mathrm{Med}(\cdot)$ est l'opérateur médiane. Ces estimateurs ne nécessitent aucune hypothèse sur l'enveloppe $\{\ell(f(X),Y), f \in \mathcal{F}\}$. De plus, par construction, ils sont également robustes à $K/2$ outliers contaminant les labels $(Y_i)_{i\in[\![1,n]\!]}$, les entrées $(X_i)_{i\in[\![1,n]\!]}$, ou les deux à la fois.

L'argument d'homogénéité, applicable pour le minimiseur du risque empirique et l'esimateur minmax-MOM, permet d'établir des vitesses rapides, c'est-à-dire de l'ordre $\mathcal{O}(1/n)$, où n est le nombre d'observations. De telles vitesses ne sont pas toujours atteignables. Pour cela, nous introduisons le concept d'hypothèse de Bernstein locale. Moralement, la condition de Bernstein signifie que la variance du problème n'est pas trop grande. Notre analyse permet d'établir des résultats sous une hypothèse de Bernstein, seulement locale. Cette condition relâche l'hypothèse de Bernstein globale et permet d'obtenir des vitesses rapides pour des problèmes où la variance est importante, ce qui n'était pas le cas des analyses précédentes. Si la distribution du bruit est symétrique et "met un peu de masse" autour de 0, alors l'hypothèse de Bernstein locale est vérifiée. Par example, lorsque le bruit est de Cauchy. De plus, notre analyse est simple et permet d'éviter tous les arguments de "peeling", normalement utilisés.

Nous utilisons également des arguments de localisation pour étudier des problèmes d'interpolations. En apprentissage statistique, on dit qu'un estimateur $\hat{f}_n$ interpole, lorsque ce dernier prédit parfaite-

ment sur un jeu d'entrainement, c'est-à-dire $\hat{f}_n(X_i) = Y_i$ pour $n = 1, \cdots, n$. En grande dimension, beaucoup de fonctions peuvent interpoler, et certaines d'entre elles sont bonnes. Dans cette thèse, nous étudions le modèle linéaire Gaussien. Soient $(X_i, Y_i)_{i \in [\![1,n]\!]}$ des vecteurs aléatoires indépendants et vérifiant

$$Y_i = X_i^T \beta^* + \xi_i \ ,$$

où $X_i \sim \mathcal{N}(0, \Sigma)$, $\xi_i \sim \mathcal{N}(0, \sigma^2)$, pour $\sigma > 0$ et $\beta^* \in \mathbb{R}^p$. La dimension p est supposé plus grande que la taille de l'échantillon n. On se place donc dans le cadre de la grande dimension. Nous montrons que l'estimateur interpolant les données de plus petite norme

$$\hat{\beta}_n = \text{argmin} \left\{ \|\beta\|_2 : \beta \in \mathbb{R}^p, \left\langle \beta, X_i \right\rangle = Y_i, i = 1 \cdots, n \right\} \ ,$$

est consistant et atteint même des vitesses rapides sous certaines hypothèses sur le spectre de la matrice de variance-covariance Σ et le bruit σ. Cette méthode souligne qu'une analyse générale, comme pour le minimiseur du risque empirique, peut être imaginée pour les solutions interpolantes. L'idée est simplement de considérer des estimateurs interpolant les données avec une certaines structure, et d'utiliser cette structure pour localiser les estimateurs.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The purpose of this introduction is to describe the main concepts developed in this thesis:

Localization arguments (Section 1.3). We present general techniques to obtain fast rates of convergence. The main idea consists in localizing the analysis around one function of interest, namely, the oracle.

Regularization (Section 1.4). Regularizations are techniques used to reduce the error and reduce the overfitting phenomenon.

Robustness (Section 1.6).*Robustness* in learning can be defined as "the insensitivity to small deviations from the assumptions" (Huber and Ronchetti, 2011). The goal consists in building and analysing estimators under as few assumptions as possible.

Figure 1.1 summarizes the key areas and their interplay in this thesis.

	Chapter 2	Chapter 2	Chapter 3	Chapter 4	Chapter 5
Localization arguments	✓	✓	✓	✓	✓
Regularization		✓	✓	✓	✓
Robustness	✓	✓	✓	✓	

Table 1.1: Schematic interplay of localization arguments, regularization, and robustness in the main chapters of this thesis.

1.1 Statistical learning

Machine learning (ML) is a scientific domain at the interface between applied mathematics, optimization and computer science. It focuses on the study of algorithms and statistical models that computer systems use to perform a specific task. The process of learning begins with observations or data, such as examples, direct experience, or instruction, in order to look for patterns in data and make better decisions in the future based on the examples that we provide. ML has various applications such as email filtering, recommendation systems, natural language processing, bio-informatics, economics, computer vision or even fraud detection. For example, based on a dataset of items and ratings, a recommendation system seeks to predict the "rating" or "preference" a user would give to an new item given its previous preferences. Another example is image classification. The learner received different types of labeled images (dogs or cats for instance). Given this dataset, the goal is to automatically label a new image occurring in the process. This is an example of *supervized Learning*. On the other hand, in community detection, the learner aims to identify communities interacting with each other. This is an example of *unsupervised learning*.
The main focus of this thesis is supervized learning. More precisely, we will be interested in robust supervized learning. Although robust unsupervised learning also exists, it is out the scope of this thesis.

Some definitions In this chapter, we present and use many tools borrowed from empirical processes. Here, are some very useful definitions and notations that we will use all along this chapter.

Definition 1.1: Empirical measure, empirical process

Let $X, X_1, \cdots, X_n$ be independent random variables taking values in a measurable space $(E, \mathcal{E})$ with common distribution P. The *empirical measure* based on the sample $(X_1, \cdots, X_n)$ is defined as

$$P_n := \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i} \ ,$$

where δ_x denotes the Dirac's measure. For a class $\mathcal{F}$ of measurable functions $f : E \mapsto \mathbb{R}$ we write

$$Pf = \mathbb{E}[f(X)] \quad \text{and} \quad P_n f = \frac{1}{n} \sum_{i=1}^{n} f(X_i) \ .$$

The *empirical process* indexed by $\mathcal{F}$ is defined by $\{(P - P_n)(f) : f \in \mathcal{F}\}$, see (Van Der Vaart and Wellner, 1996).

In *supervized machine learning* (Vapnik, 2000; Friedman et al., 2001; Shalev-Shwartz and Ben-David, 2014; Bishop, 2006), the goal is to predict an output Y in $\mathcal{Y}$ based on features X in $\mathcal{X}$, that is to say, to understand the relationship between an ouput Y and inputs X. The set $\mathcal{Y}$ can be either finite (typically $\{0,1\}$) or infinite ($\mathcal{Y} \subset \mathbb{R}$), leading to two important problems in supervized machine learning:

- **Classification**, when $\mathcal{Y}$ is finite. If $|\mathcal{Y}| = 2$, where $|\cdot|$ denotes the cardinality of $\mathcal{Y}$, the problems is binary classification.
- **Regression**, when $\mathcal{Y}$ is a continuous subset of $\mathbb{R}$.

Typically $\mathcal{X} = \mathbb{R}^p$, where p is large and denotes the dimension of the problem. For example:

- For binary classification: $\mathcal{X} = \mathbb{R}^p$ can correspond to a set of images encoded with their p pixels and $\mathcal{Y} = \{0,1\}$, if the label is 1, the image is labelled as a dog, otherwise as a cat.
- For a regression problem, $\mathcal{X}$ can summarize socioeconomic factors and $\mathcal{Y} = [0,100]$ depicts the score of the left-wing during the next presidential election.

The output $Y \in \mathcal{Y}$ is not always a deterministic function of an input $X \in \mathcal{X}$ due to random factors such as measurement errors. Thus, the couple (X,Y) is modeled as a random variable with a certain unknown distribution P. Let P_X denote the marginal distribution of X. The goal becomes to predict the output Y with the input X given that (X,Y) is sampled from P. To do so, we define a *predictor* as a measurable function $f : \mathcal{X} \mapsto \mathcal{Y}$. The random variable $f(X)$ serves to predict Y. The set of possible predictors (i.e measurable functions from $\mathcal{X}$ to $\mathcal{Y}$) is denoted by $\mathcal{F}(\mathcal{X},\mathcal{Y})$. To measure the quality of a predictor f in $\mathcal{F}(\mathcal{X},\mathcal{Y})$, we introduce a loss function

$$\ell : \mathcal{Y} \times \mathcal{Y} \mapsto \mathbb{R}_+ \ ,$$

such that $\ell(f(X),Y)$ measures the error of predicting $f(X)$ while the true answer is Y. It is always assumed that $\ell(y,y) = 0$, for every $y \in \mathcal{Y}$. Two important examples are:

- Example of classification: let $\mathcal{Y} = \{-1,1\}$ and $\ell(y,y') = \mathbf{1}\{y \neq y'\}$. Thus, $\ell(y,y') = 1$ if $y \neq y'$ and $\ell(y,y')$ otherwise. This loss function is often replaced by convex surrogates for computation purposes such as the Hinge loss, $\ell(f(X),Y) = \max(0, 1 - Yf(X))$ or the logistic loss, $\ell(f(X),Y) = \log(1 + \exp(-Yf(X)))$.
- Example of regression: let $\mathcal{Y}$ be a continuous subset of $\mathbb{R}$ and $\ell(y,y') = (y-y')^2/2$. It is also called *least squares regression*.

Given a random couple (X,Y) with distribution P, the quality of a prediction function is measured by its *risk*, or *generalization error*, defined as the averaged loss under the distribution P of the observations:

$$R(f) = \mathbb{E}_{(X,Y)\sim P}[\ell(f(X),Y)] \ .$$

Adopting the notations in Definition 1.1, we also write $R(f) = P\ell_f$. The optimal predictor f_P^{**} is defined, when it exists, as the minimizer of the risk over the set of all measurable functions $\mathcal{F}(\mathcal{X}, \mathcal{Y})$

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

Although the function f_P^{**} may not exist, it does for standard loss functions used in practice. In this case, f_P^{**} is called a *Bayes predictor*. For example, in regression with the square loss, $f_P^{**}(X) = \mathbb{E}_{(X,Y)\sim P}[Y|X]$. However, the distribution P being unknown, the Bayes predictor f_P^{**} is also unknown and must be approximated. To do so, a training dataset of n independent and indentically observations $\mathcal{D} = (X_i, Y_i)_{i\in[\![1,n]\!]}$ in $(\mathcal{X}, \mathcal{Y})^n$ with the same distribution P as (X, Y), is given. One would like to use the dataset $\mathcal{D}$ to predict the output Y associated with the input X. We can formalize the problem through the notion of *learning rule* $\mathcal{Z} : \cup_{n=1}^{\infty}(\mathcal{X} \times \mathcal{Y})^n \mapsto \mathcal{F}(\mathcal{X}, \mathcal{Y})$ defined as a measurable function that maps the set of observations to an estimator $\hat{f}$. Note that the observations $(X_i, Y_i)_{i\in[\![1,n]\!]}$ are random. Consequently, the predictor $\hat{f} = \mathcal{Z}\big((X_i, Y_i)_{i\in[\![1,n]\!]}\big)$ and its risk are also random. One of our main goal, is to find learning rules with a risk close to the one of f_P^{**} with high probability. In a non informative way, we search for learning rules such that with large probability $R(\mathcal{Z}\big((X_i, Y_i)_{i\in[\![1,n]\!]}\big) \approx R(f_P^{**})$. Such results can also be derived in expectation

$$\mathbb{E}_{(X_i,Y_i)_{i\in[\![1,n]\!]}\sim P^{\otimes n}} R(\mathcal{Z}\big((X_i, Y_i)_{i\in[\![1,n]\!]}\big) \approx R(f_P^{**}) \ .$$

In this thesis, we propose results holding with exponentially large probability. In fact, it turns out that such results often imply bounds in expectation (see Section 1.2.2 for an example).

1.2 Empirical Risk Minimization (ERM)

1.2.1 Definition and properties

Since the risk $R(\cdot)$ is unknown, the most common and wide-spread learning rule consists in replacing the expectation with respect to P by the empirical measure and minimize it. This method is known as *Empirical Risk Minimization* (ERM) and is defined as

$$\mathcal{Z}\big((X_i, Y_i)_{i\in[\![1,n]\!]}\big) \in \operatorname*{argmin}_{f \in \mathcal{F}(\mathcal{X},\mathcal{Y})} R_n(f) \quad \text{with} \quad R_n(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(f(X_i), Y_i) \ .$$

Adopting the notations in Definition 1.1 we write $R_n(f) = P_n\ell_f$. This natural idea dates back to works of Gauss in the early 19th-century who introduced the least-squares estimator, which is the ERM for linear predictors with the square loss function. Many important methods of statistical estimation such as maximum likelihood or more general M-estimation (Van de Geer, 2000) are versions of empirical risk minimization. The general theory of empirical risk minimization began with the works of Vapnik and Chervonenkis (Vapnik, 2000) in the late 1970s and the early 1980s. Their main idea was to relate the quality of prediction of the empirical risk minimizer with the

accuracy of approximation of the true distribution P by it empirical counterpart P_n, uniformly over a well-chosen class of functions. Their approach necessitates a uniform control the empirical process $\{(P-P_n)(f), f \in \mathcal{F}\}$, for $\mathcal{F}$ a well-chosen class of functions (see below in this section). The authors introduced a number of natural and important measures of complexity of class of functions, such as entropy and VC-dimension (VC standing for Vapnik-Chervonenkis).

This intuitive learning rules raises a natural question: how does a minimizer of the empirical risk R_n performs, that is, how does its risk behave compared with the one of f_P^{**} ? Although a prediction rule works well on observed points, it does not guarantee that its risk is small. Indeed, the set of all measurable functions is very large, and it is easy to find a prediction function f such that $f(X_i) = Y_i$ for every $i = 1, \cdots, n$. Such a predictor f is a minimizer of the empirical risk. However, fitting perfectly the dataset yields in general to poor generalization properties (Shalev-Shwartz and Ben-David, 2014), a phenomenon known as over-fitting. A standard tool to avoid such pathological situations is to use *regularization methods*. There are two equivalent formulations.

- **Restriction to a small class of functions:** let $\mathcal{F} \subset \mathcal{F}(\mathcal{X}, \mathcal{Y})$ be a class of functions, large enough to reasonably approximate any measurable function, but not too large to avoid the over-fitting phenomenon. The minimization of the empirical risk is restricted to the class of functions $\mathcal{F}$

$$\hat{f}_{\mathcal{F}} \in \operatorname*{argmin}_{f \in \mathcal{F}} P_n \ell_f \ .$$

- **Introduction of a penalization:** Let $\Psi : \mathcal{F}(\mathcal{X}, \mathcal{Y}) \mapsto \mathbb{R}_+$ be a function penalizing the least regular measurable functions and $\lambda > 0$. We define

$$\hat{f}_{\lambda} \in \operatorname*{argmin}_{f \in \mathcal{F}(\mathcal{X}, \mathcal{Y})} \{P_n \ell_f + \lambda \Psi(f)\} \ .$$

Since these two approaches are equivalent, we will focus on the first one, when the minimization of the empirical risk is restricted to a sub-class of measurable functions $\mathcal{F}$. Let $f_{\mathcal{F}}^*$ be the minimizer of the risk over $\mathcal{F}$.

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

With these definitions in mind, we have

$$R(\hat{f}_{\mathcal{F}}) \geq R(f_{\mathcal{F}}^*) \geq R(f_P^*) \ .$$

Let $R(\hat{f}_{\mathcal{F}}) - R(f_P^{**})$ be the *excess risk*. This quantity is always non-negative and the smaller it is, the better $\hat{f}_{\mathcal{F}}$ predicts. The excess

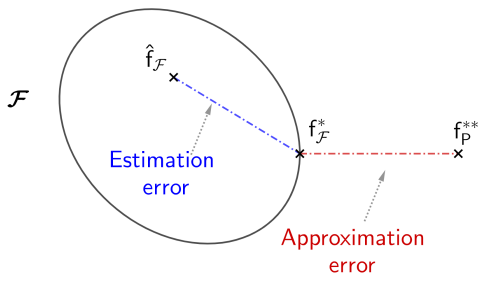


Figure 1.1: Risk decomposition

risk can be decomposed in two terms: the *estimation error* and *approximation error*.

$$R(\hat{f}_{\mathcal{F}}) - R(f_P^{**}) = \underbrace{R(\hat{f}_{\mathcal{F}}) - R(f_{\mathcal{F}}^*)}_{\text{Estimation error}} + \underbrace{R(f_{\mathcal{F}}^*) - R(f_P^{**})}_{\text{Approximation error}} \ .$$

We illustrate this error decomposition in Figure 1.1. The estimation error comes from the fact that $\hat{f}_{\mathcal{F}}$ minimizes the empirical risk instead of the true risk. It increases with the complexity[1] of $\mathcal{F}$. On the other hand, an increasing number of observations makes the empirical risk closer to the true risk and thus reduces the estimation error. The approximation error is due to the fact that $f_{\mathcal{F}}^*$ minimizes the risk only on a subset of all measurable functions. It decreases with the size of $\mathcal{F}$. Consequently, there is a trade-off to find, to optimize the choice of F. It is known as the *bias-variance* trade-off where the variance is due to the estimation error while the bias is due to the approximation error. In this thesis, we will only focus on the estimation error of a given estimator. We want to relate the risk of $\hat{f}_{\mathcal{F}}$ with the one of $f_{\mathcal{F}}^*$, the best risk one can hope using functions in the class $\mathcal{F}$. Taking the point of view of Vapnik and Chervonenkis, we relate the estimation error $R(\hat{f}_{\mathcal{F}}) - R(f_{\mathcal{F}}^*)$ with the empirical process $\{(P_n - P)(\ell_f), f \in \mathcal{F}\}$, indexed by the class $\mathcal{F}$. In particular, we have the following upper bound for the estimation error

$$\begin{aligned} R(\hat{f}_{\mathcal{F}}) - R(f_{\mathcal{F}}^*) &= \underbrace{R(\hat{f}_{\mathcal{F}}) - R_n(\hat{f}_{\mathcal{F}})}_{\leq \sup_{f\in\mathcal{F}} |P\ell_f - P_n\ell_f|} + \underbrace{R_n(\hat{f}_{\mathcal{F}}) - R_n(f_{\mathcal{F}}^*)}_{\leq 0} + \underbrace{R_n(f_{\mathcal{F}}^*) - R(f_{\mathcal{F}}^*)}_{\leq \sup_{f\in\mathcal{F}} |P\ell_f - P_n\ell_f|} \\ &\leq 2 \sup_{f\in\mathcal{F}} |P\ell_f - P_n\ell_f| := \|P_n - P\|_{\ell_{\mathcal{F}}} \end{aligned}$$

To derive upper bounds for the estimation error, it is sufficient to uniformly control the empirical process $\{(P_n\ell_f - P\ell_f), f \in \mathcal{F}\}$ over the class $\mathcal{F}$. Thus, by analysing deviations between the risk and its empirical version it is possible to control the estimation error $R(\hat{f}_{\mathcal{F}}) - R(f_{\mathcal{F}}^*)$. Such deviations are at the heart of the theory of empirical processes (Van Der Vaart and Wellner, 1996; Pollard et al., 1989; Van de Geer, 2000). Concentration results coupled with powerful tools from empirical processes theory allow to prove many non-trivial and deep results in statistical learning. In particular, in the 1990s, Talagrand proved a uniform version of Bernstein's inequality allowing to concentrate $\|P_n - P\|_{\ell_{\mathcal{F}}}$ around its expectation (Talagrand, 1996). This result had a huge impact on the theory of empirical processes and empirical risk minimizer. We will present in Section 1.3 an example of application of this outstanding concentration inequality.

1.2.2 General analysis of the statistical error

In this section, we present standard arguments to bound $\sup_{f\in\mathcal{F}} |P_n\ell_f - P\ell_f| = \|P_n - P\|_{\ell_{\mathcal{F}}}$. Results are derived with high probability and in expectation. The main tools are concentration inequalities and Rademacher complexities. The first step consists in quantifying the deviation of $\|P_n - P\|_{\ell_{\mathcal{F}}}$ around its expectation $\mathbb{E}\|P_n - P\|_{\ell_{\mathcal{F}}}$. To do so, it is necessary to impose strong assumptions over

[1] see Rademacher and Gaussian complexities below.

the class $\ell_{\mathcal{F}}$, where $\ell_{\mathcal{F}} = \{\ell\big(f(\cdot),\cdot\big), f \in \mathcal{F}\}$. For a long time (Devroye et al., 2013; Koltchinskii, 2001; Bartlett, 1998), it has been assumed that the class $\ell_{\mathcal{F}}$ was bounded, that is, there exists a constant $c > 0$ such that $|\ell(f(x), y)| \leq c$, for every $f \in \mathcal{F}$ and $(x, y) \in \mathcal{X} \times \mathcal{Y}$. The main reason is that no concentration results were available to handle unbounded class $\ell_{\mathcal{F}}$. Even if more general results are now available, we focus here on bounded classes $\ell_{\mathcal{F}}$. Two common examples are:

Example 1: if $\ell(y, y') = \mathbf{1}\{y \neq y'\}$ then $c = 1$.

Example 2: if $y \mapsto \ell(\cdot, y)$ is convex and L-Lipschitz ($|\ell(y_1, y) - \ell(y_2, y)| \leq L|y_1 - y_2|$, for every y_1, y_2 in $\mathcal{Y}$) and $\mathcal{F} = \{\big\langle \beta, \Phi(\cdot)\big\rangle, \beta \in \mathbb{R}^p \text{ s.t } \|\beta\|_2 \leq B\}$, where $\Phi(\cdot)$ is a bounded feature map i.e $\|\Phi(x)\|_2 \leq D$ for every $x \in \mathcal{X}$. In this case, $c = LBD$ because for every β in $\mathbb{R}^p$ such that $\|\beta\|_2 \leq B$ and $(x, y) \in \mathcal{X} \times \mathcal{Y}$

$$|\ell\big(\big\langle \beta, \Phi(x)\big\rangle, y\big)| = |\ell\big(\big\langle \beta, \Phi(x)\big\rangle, y\big) - \ell(0, 0)| \leq L|\big\langle \beta, \Phi(x)\big\rangle| \leq L\|\beta\|_2\|\Phi(x)\|_2 \leq LBD \ .$$

In this example, the minimizer of the empirical risk minimizer is unique and defined as

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^p : \|\beta\|_2 \leq B}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1} \ell\big(\big\langle \Phi(X_i), \beta\big\rangle, Y_i\big) \ , \tag{1.1}$$

where the loss is L-Lipschitz.

This boundedness assumption allows to control the deviations of $\|P_n - P\|_{\ell_{\mathcal{F}}}$ around its expectation with high probability. To do so, we use Theorem 1.1, known as Mc Diarmid's inequality that we recall here.

Theorem 1.1: McDiarmid's inequality

Consider independent random variables $X_1, \cdots, X_n \in E$ and a mapping $\psi : E^n \mapsto \mathbb{R}$. If for all $i \in [\![1, n]\!]$ and for all $x_1, \cdots, x_n, x_i'$

$$|\psi(x_1, \cdots, x_i, \cdots, x_n) - \psi(x_1, \cdots, x_i', \cdots, x_n)| \leq c_i \ ,$$

then for every $t > 0$,

$$\mathbb{P}\big(\big|\psi(X_1, \cdots, X_n) - \mathbb{E}[\psi(X_1, \cdots, X_n)]\big| \geq t\big) \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^n c_i^2}\right)$$

For all, $x_1, \cdots, x_n$ in $\mathcal{X}$ and $y_1, \cdots, y_n$ in $\mathcal{Y}$, let $\psi : (\mathcal{X} \times \mathcal{Y})^n \mapsto \mathbb{R}$ be defined as

$$\psi\big((x_1, y_1), \cdots, (x_n, y_n)\big) = \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \ell(f(x_i), y_i) - \mathbb{E}\ell\big(f(X), Y\big) \right| \ .$$

From the boundedness assumption, for all $i \in [\![1, n]\!]$ and $(x_1, y_1), \cdots, (x_n, y_n), (x_i', y_i)'$ in $\mathcal{X} \times \mathcal{Y}$,

$$\begin{aligned} \big|\psi\big((x_1, y_1), \cdots, (x_i, y_i), \cdots, (x_n, y_n)\big) - \psi\big((x_1, y_1), \cdots, (x_i', y_i'), \cdots, (x_n, y_n)\big)\big| \\ \leq \frac{1}{n} \sup_{f \in \mathcal{F}} \big|\ell(f(x_i), y_i) - \ell(f(x_i'), y_i')\big| \leq \frac{2c}{n} \ , \end{aligned}$$

and from Theorem 1.1, with probability larger than $1 - \exp(-t)$

$$R(\hat{f}_{\mathcal{F}}) - R(f^*_{\mathcal{F}}) \leq 2 \sup_{f \in \mathcal{F}} |P\ell_f - P_n\ell_f| \leq \underbrace{2\,\mathbb{E} \sup_{f \in \mathcal{F}} |P\ell_f - P_n\ell_f|}_{(\star)} + c\sqrt{\frac{8t}{n}}$$

Consequently, controlling the estimation error requires to control $(\star)$. A common approach is based on Rademacher complexity that we introduce now.

Definition 1.2: Rademacher complexity

Let $X_1, \cdots, X_n$ be independent random variables taking values in a measurable space $(E, \mathcal{E})$ with common distribution P. Let $\mathcal{F}$ be a class of functions from E to $\mathbb{R}$. The Rademacher complexity of the class $\mathcal{F}$ is defined as

$$\mathcal{R}_n(\mathcal{F}) = \mathbb{E}\left[\sup_{f \in \mathcal{F}} \left(\frac{1}{n} \sum_{i=1}^{n} \sigma_i f(X_i) \right) \right] ,$$

where the variables $\sigma_1, \cdots, \sigma_n$ are i.i.d Rademacher random variables ($\mathbb{P}(\sigma_1 = 1) = \mathbb{P}(\sigma_1 = -1) = 1/2$) independent of $X_1, \cdots, X_n$. The expectation is taken with respect to both the Rademacher random variables and the data $X_1, \cdots, X_n$.

The Rademacher complexity of a class $\mathcal{F}$ quantifies the extent to which some functions in $\mathcal{F}$ can be correlated with a Bernoulli noise sequence. Such a quantity is large if there exists f in $\mathcal{F}$ for which $f(X_i)$ fits well the noise in expectation. Using such a class is very likely to result in over-fitting. This idea could serve as an intuitive explanation why $\mathcal{R}_n(\mathcal{F})$ can be used as a notion of complexity of a class of functions in the analysis of empirical risk minimization. Another reason why Rademacher complexities are very used in practice relies on its appealing properties such as Lemmas 1.1 and 1.2. In particular, the symmetrization Lemma 1.1 gives $(\star) \leq 2\mathcal{R}_n(\ell_{\mathcal{F}})$ and with probability larger than $1 - \exp\big(-t\big)$

$$R(\hat{f}_{\mathcal{F}}) - R(f^*_{\mathcal{F}}) \leq 4\mathcal{R}_n(\ell_{\mathcal{F}}) + c\sqrt{\frac{8t}{n}} \ . \tag{1.2}$$

Although Lemma 1.1 is known for a long time, Rademacher complexities were proposed as a measure of the complexity for the first time only in the early 2000s in (Bartlett et al., 2002a; Koltchinskii, 2001; Mendelson, 2002).

Lemma 1.1: Symmetrization

Let $\mathcal{F}$ be a class of functions from E to $\mathbb{R}$. Then,

$$\mathbb{E}\left[\sup_{f \in \mathcal{F}} \left| \mathbb{E}[f(X)] - \frac{1}{n} \sum_{i=1}^{n} f(X_i) \right| \right] \leq 2\mathcal{R}_n(\mathcal{F}) \ .$$

Equation (1.2) shows that up to a term of order $1/\sqrt{n}$ the estimation error can be bounded by the Rademacher complexity of $\ell_{\mathcal{F}}$, with an exponentially large probability. The computation of $\mathcal{R}_n(\ell_{\mathcal{F}})$ depends on the problem. In the case of Example 2, from Lemma 1.2 we have

$$\mathcal{R}_n(\ell_{\mathcal{F}}) = \mathbb{E}\left[\sup_{t\in\mathbb{R}^p:\|t\|_2\leq B}\left|\frac{1}{n}\sum_{i=1}^n \sigma_i \ell\big(\langle t, \Phi(X_i)\rangle, Y_i\big)\right|\right] \leq 2L\mathbb{E}\left[\sup_{t\in\mathbb{R}^p:\|t\|_2\leq B}\left|\frac{1}{n}\sum_{i=1}^n \sigma_i \langle t, \Phi(X_i)\rangle\right|\right]$$
$$\leq 2LB\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^n \sigma_i \Phi(X_i)\right\|_2 \leq 2LB\left(\left\|\frac{1}{n}\sum_{i=1}^n \sigma_i \Phi(X_i)\right\|_2^2\right)^{1/2} \leq \frac{2LBD}{\sqrt{n}} \ ,$$

and the following result holds

$$\mathbb{P}\left(R(\hat{f}_{\mathcal{F}}) - R(f^*_{\mathcal{F}}) \leq \frac{8LBD}{\sqrt{n}} + c\sqrt{\frac{8t}{n}}\right) \geq 1 - \exp(-t) \ . \tag{1.3}$$

Lemma 1.2: Contraction lemma

Let $\mathcal{F}$ be a class of functions from E to $\mathbb{R}$ and $\phi : \mathbb{R} \mapsto \mathbb{R}$ a L-Lipschitz function

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left(\frac{1}{n}\sum_{i=1}^n \sigma_i \phi\big(f(X_i)\big)\right)\right] \leq 2L\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left(\frac{1}{n}\sum_{i=1}^n \sigma_i f(X_i)\right)\right] \ .$$

Equation (1.3) states that the estimation error of $\hat{f}_{\mathcal{F}}$ defined in Equation (1.1) is controlled with high probability by a term of order $\mathcal{O}(1/\sqrt{n})$ [2]. Since the estimation error of $\hat{f}_{\mathcal{F}}$ is always non-negative, using the integrated tail probability expectation formula we can deduce from (1.3) an upper bound in expectation

$$\mathbb{E}\big(R(\hat{f}_{\mathcal{F}}) - R(f^*_{\mathcal{F}})\big) = \int_0^{+\infty} \mathbb{P}(R(\hat{f}_{\mathcal{F}}) - R(f^*_{\mathcal{F}}) \geq t)\mathrm{d}t$$

where the expectation is taken with respect to the i.i.d sample $(X_i, Y_i)_{i\in[\![1,n]\!]}$ with common distribution P, and straightforward computations give

$$\mathbb{E}\big(R(\hat{f}_{\mathcal{F}}) - R(f^*_{\mathcal{F}})\big) \leq \frac{A(L, B, D, c)}{\sqrt{n}} \ ,$$

where the constant A depends on L, B, D and c. The expected estimation error is also of order $\mathcal{O}(1/\sqrt{n})$. This example reveals that results holding with high probability are more appealing and should be preferred. However, they are usually harder to prove and require the development of concentration inequalities.

[2] The notation $\mathcal{O}$ means that there exists an absolute constant $M > 0$ big enough such that the excess risk is bounded by $M/\sqrt{n}$, for n sufficiently large.

The approach we used is based on a *global* analysis. The estimation error depends on the complexity of the entire class $\mathcal{F}$, measured here by the Rademacher complexity of $\ell_{\mathcal{F}}$. This approach has been used for example in (Bartlett and Mendelson, 2002; Bartlett et al., 2002a; Koltchinskii et al., 2002). Despite its simplicity and generality, this analysis presents two main drawbacks.

1. Global Rademacher complexities are too large. In the example of linear functionals with bounded features it leads to an estimation error of order $\mathcal{O}(1/\sqrt{n})$ while it is possible to obtain rates of order $\mathcal{O}(1/n)$, see Section 1.2.3.

2. To obtain results holding with high probability we assumed that the whole class $\ell_{\mathcal{F}}$ is bounded.

Consequently, a general analysis of the estimation error based on global complexity parameters is not satisfactory. In this thesis, we will present localization arguments allowing to obtain faster rates. We will also develop other approaches to handle unbounded classes $\ell_{\mathcal{F}}$.
We presented a general approach based on the Rademacher complexity $\mathcal{R}_n(\ell_{\mathcal{F}})$. There exist other classical complexities that we will explore in the sequel.

1.2.3 Linear least-squares regression

In Section 1.2.2, we obtained an upper bound on the expected statistical error of order $\mathcal{O}(1/\sqrt{n})$. In this section, we present the example of least-square regression and derive the upper bound $\mathcal{O}(\sigma^2 p/n)$ on the estimation error holding in expectation, where $\sigma^2 > 0$ is the variance of the noise and p is the dimension. We also briefly present the minimax paradigm and claim that the rate $\mathcal{O}(\sigma^2 p/n)$ is minimax-rate-optimal for the problem of least-square regression in a Gaussian setting.

Let (X, Y) be such that $Y|X \sim \mathcal{N}(\langle X, \beta^* \rangle, \sigma^2)$, for $\beta^* \in \mathbb{R}^p$ and $X \sim \mathcal{N}(0, I_p)$. Equivalently, $Y = \langle X, \beta^* \rangle + \xi$, where $\beta^* \in \mathbb{R}^p$, $X \sim \mathcal{N}(0, I_p)$ and $\xi \sim \mathcal{N}(0, \sigma^2)$ is independent of X. Let $\mathcal{F} = \{\langle \beta, \cdot \rangle, \beta \in \mathbb{R}^p\}$ be the class of linear functionals in $\mathbb{R}^p$. For every y, y' in $\mathcal{Y}$ let $\ell(y, y') = (y - y')/2$ be the quadratic loss function. Let $(X_i, Y_i)_{i \in [\![1,n]\!]}$ be i.i.d random variables distributed as (X, Y). The risk associated with β in $\mathbb{R}^p$ is defined as

$$R(\beta) = \frac{1}{2}\mathbb{E}\big[(\langle X, \beta \rangle - Y)^2\big] \ .$$

The parameter β^* minimizes the risk over $\mathbb{R}^p$ and for every $\beta \in \mathbb{R}^p$

$$\begin{aligned} R(\beta) - R(\beta^*) &= \frac{1}{2}\mathbb{E}\big[(\langle X, \beta \rangle - Y)^2\big] - \frac{1}{2}\mathbb{E}\big[(\langle X, \beta^* \rangle - Y)^2\big] \\ &= \mathbb{E}\big[Y\langle X, \beta^* - \beta \rangle\big] + \frac{1}{2}\mathbb{E}\big[\langle X, \beta \rangle^2 - \langle X, \beta^* \rangle^2\big] \\ &= \frac{1}{2}\mathbb{E}\big[\langle X, \beta - \beta^* \rangle^2\big] = \frac{1}{2}\|\beta - \beta^*\|_2^2 \ , \end{aligned}$$

where we used the first order condition to have $\mathbb{E}\big[\big(Y - \langle X, \beta^* \rangle\big)\langle \beta - \beta^*, X \rangle\big] = 0$. Let $\hat{\beta}$ be the minimizer of the empirical risk of $\mathbb{R}^p$:

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \frac{1}{2n} \sum_{i=1}^{n} \big(\langle X_i, \beta \rangle - Y_i\big)^2 = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \frac{1}{2n} \|\boldsymbol{X}\beta - \boldsymbol{Y}\|_2^2 \ ,$$

where $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ denotes the feature matrix whose lines are given by X_i^T, $i = 1, \cdots, n$ and $\boldsymbol{Y} = (Y_1, \cdots, Y_n) \in \mathbb{R}^n$. When $\boldsymbol{X}^T\boldsymbol{X}$ is assumed to be invertible, $\hat{\beta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}/n$, which is the case only if $p \leq n$. This corresponds to a *low dimensional* regime, when the number of observations n is bigger than the number of covariates p. In (Lecué and Mendelson, 2016), the authors establish that

$$\mathbb{E}[R(\hat{\beta}) - R(\beta^*)] \leq c\frac{\sigma^2 p}{n} \ ,$$

where the expectation is taking with respect to the data $(X_i, Y_i)_{i \in [\![1,n]\!]}$ and $c > 0$ is an absolute constant. A natural question is now the following: is a rate of order $\mathcal{O}(\sigma^2 p/n)$ optimal, and in which sense ? The next paragraph gives some elements to answer this question.

Minimax rates of convergence In (Lecué and Mendelson, 2016), the authors provide an upper bound of order $\mathcal{O}(\sigma^2 p/n)$ for the estimation error of the empirical risk minimizer. A large upper bound does not necessarily reflect a bad mathematical analysis. This may be an inevitable consequence of the difficulty of the problem. To study the optimality of a rate, we use the notion of minimax risk, see (Tsybakov, 2008; Massart, 2007) for good references. A *statistical model* $\{P_\theta, \theta \in \Theta\}$, is a set of probability measures indexed by a parameter θ in Θ. The minimax risk associated with the statistical model $\{P_\theta, \theta \in \Theta\}$ is defined as

$$\mathcal{A}_n^* := \inf_{\hat{\theta}_n} \sup_{\theta \in \Theta} \mathbb{E}_\theta[d(\theta, \hat{\theta}_n)] \ ,$$

where the infimum is taken over all estimators $\hat{\theta}_n$, $d(\cdot, \cdot)$ is a distance and $\mathbb{E}_\theta$ denotes the expectation with respect to P_θ. $\mathcal{A}_n^*$ is the best possible rate associated with a statistical model and a distance, one can expect. Given $\mathcal{A}_n^*$ it is now possible to claim that an estimator is optimal. We say that $\hat{\theta}_n$ is *minimax-rate-optimal* for the model $\{P_\theta, \theta \in \Theta\}$ and the distance d is there exists a constant $c > 0$ such that

$$\sup_{\theta \in \Theta} \mathbb{E}_\theta[d(\theta, \hat{\theta}_n)] \leq c\Psi_n \ ,$$

where Ψ_n is the optimal rate of convergence i.e the rate Ψ_n such that there exists two constants $0 < c \leq C$ such that

$$c \leq \liminf_{n \to \infty} \mathcal{A}_n^* \Psi_n \leq \limsup_{n \to \infty} \mathcal{A}_n^* \Psi_n \leq C \ .$$

In (Tsybakov, 2003), the authors provide a lower bound of order $\Omega(\sigma^2 p/n)$[3] for the problem of least-squares regression when the design and the noise are Gaussians. Consequently, the empirical

[3]The notation Ω means that there exists an absolute constant $m > 0$ small enough such that the lower bound is larger than $m\sigma^2 p/n$

risk minimizer is minimax-rate-optimal for the problem of least square regression in a Gaussian model.
The analysis presented in Section 1.2.2 leads to rates of order $\mathcal{O}(1/\sqrt{n})$ while the right order is $\mathcal{O}(1/n)$. Thus, an analysis based on global estimates of the complexity of the class of functions is not satisfactory and more involved arguments have to be used. In section 1.3 we present a new analysis based on local measures of complexity.
In Chapter 6, we will consider the same model when the dimension p may be larger than n and $X \sim \mathcal{N}(0, \Sigma)$.

1.3 Localization methods for ERM

We established upper bounds on the statistical error. In particular, we observed that the general analysis developed in Section 1.2.2, based on global measures of complexity leads to error rates of order $\mathcal{O}(1/\sqrt{n})$. From (Lecué and Mendelson, 2016), the empirical risk minimizer in the problem of least-squares in the Gaussian setting attains an error rate of order $\mathcal{O}(1/n)$. Thus, we would like to derive a general analysis of the empirical risk minimizer leading to error rates of order $1/n$ (when it is possible).

Due to the symmetrization Lemma 1.1, Rademacher complexities have been proposed as an effective notion of complexity measure in (Bartlett et al., 2002a; Koltchinskii, 2001; Bartlett and Mendelson, 2002; Mendelson, 2002). In these papers, the analysis is based on global estimates of the complexity of the class of functions. No further information is used. However, since the risk of the empirical risk minimizer is expected to be small, the complexity of a small neighborhood of the oracle may be sufficient to describe its behavior. It is the main intuition behind localization methods. They appeared first in (Koltchinskii and Panchenko, 2000) for noiseless problems, i.e $R(f_{\mathcal{F}}^*) = 0$. In (Bousquet et al., 2002) the authors performed localization techniques around 0, assumed to belong to $\mathcal{F}$. See also (Lugosi et al., 2004) for localization methods applied to Boolean classes. The first localization around the minimizer of the risk, $f_{\mathcal{F}}^*$ in the class $\mathcal{F}$, was presented by Massart in (Massart, 2000), and then extensively studied in (Bartlett et al., 2005; Bartlett and Mendelson, 2006a,b) for model selection and empirical risk minimization. In their first versions, localization methods were developed for general bounded classes of functions. In this thesis, we will focus on convex classes of functions. We say that the class $\mathcal{F}$ is convex, if for every f, g in $\mathcal{F}$ and $\alpha \in [0, 1]$, the function $\alpha f + (1-\alpha)g$ belongs to $\mathcal{F}$. The role of convexity if twofold. First, from the practitioner's point of view, minimizing the empirical risk is much easier when both the class $\mathcal{F}$ and the loss ℓ are convex. Performing simple gradient descent-based methods allows to converge toward the empirical risk minimizer (Boyd et al., 2004). Secondly, from the theoretical standpoint, it is well known that convexity plays a key role in statistical learning (Lee et al., 1998; Mendelson, 2001).

1.3.1 A general approach of localization in a bounded setting.

The estimation error $R(\hat{f}_{\mathcal{F}}) - R(f^*_{\mathcal{F}})$, that we will call *excess risk* now, is a natural measure of accuracy of the approximation of $f^*_{\mathcal{F}}$ by $\hat{f}_{\mathcal{F}}$. The goal is to find tight upper bounds on the excess risk of $\hat{f}_{\mathcal{F}}$ holding with an exponentially large probability. This bound depends on various measures of complexity that drives the accuracy of approximating the true risk $P\ell_f$ by its empirical counterpart $P_n\ell_f$. Hereafter, we present a simple and general approach to derive upper bounds on the excess risk of the empirical risk minimizer associated with convex loss functions.

For every $f \in \mathcal{F}$ let us recall that $P\ell_f = R(f) = \mathbb{E}_{(X,Y)\sim P}[\ell(f(X), Y)]$ and $P_n\ell_f = R_n(f) = n^{-1}\sum_{i=1}^n \ell(f(X_i), Y_i)$. We will also denote by$P\mathcal{L}_f := P\ell_f - P\ell_{f^*_{\mathcal{F}}}$ and by $P_n\mathcal{L}_f := P_n\ell_f - P_n\ell_{f^*_{\mathcal{F}}}$. Since $\hat{f}_{\mathcal{F}}$ minimizes the empirical risk over $\mathcal{F}$, we have $P_n\ell_{\hat{f}_{\mathcal{F}}} \leq P_n\ell_f$ for every f in $\mathcal{F}$ and in particular $P_n\mathcal{L}_{\hat{f}_{\mathcal{F}}} \leq 0$. Therefore, to control the excess risk of $\hat{f}_{\mathcal{F}}$, it is enough to show that with large probability for every f in $\mathcal{F}$ such that $P\mathcal{L}_f \geq r^*$ we have $P_N\mathcal{L}_f > 0$. With the same probability the excess risk $P\mathcal{L}_{\hat{f}_{\mathcal{F}}}$ will be bounded by r^*. Clearly, the choice of r^* is an important (and complicated) task.
Let $f \in \mathcal{F}$ such that $P\mathcal{L}_f > r^*$. The following "homogeneity lemma" shows that risk bounds for the empirical risk minimizer estimators follow from a concentration of $(P - P_n)\mathcal{L}_f$ over sub-classes of $\mathcal{F}$ around the oracle $f^*_{\mathcal{F}}$.

Lemma 1.3: Homogeneity Lemma (Chinot et al., 2019b)

For every f in $\mathcal{F}$ such that $P\mathcal{L}_f > r^*$ there exists $P\mathcal{L}_f/r^* \geq \alpha > 1$, f_0 in $\mathcal{F}$ such that $\alpha(f_0 - f^*) = f - f^*$ and $P\mathcal{L}_{f_0} = r^*$.

From Lemma 1.3 we obtain

$$P_n\mathcal{L}_f = \frac{1}{n}\sum_{i=1}^n \Big(\ell\big(f(X_i), Y_i\big) - \ell\big(f^*_{\mathcal{F}}(X_i), Y_i\big)\Big) = \frac{1}{n}\sum_{i=1}^n \Big(\ell\big((\alpha f_0 + (1-\alpha)f^*_{\mathcal{F}})(X_i), Y_i\big) - \ell\big(f^*_{\mathcal{F}}(X_i), Y_i\big)\Big) \ ,$$

where $P\mathcal{L}_{f_0} = r^*$ and $\alpha > 1$. Since, the function $y \mapsto \ell(y, y')$ is convex for all y' in $\mathcal{Y}$ we have for every $(x, y) \in \mathcal{X} \times \mathcal{Y}$,

$$\ell\big((\alpha f_0 + (1-\alpha)f^*_{\mathcal{F}})(x), y\big) \geq \alpha\ell\big(f_0(x), y\big) + (1-\alpha)\ell\big(f^*_{\mathcal{F}}(x), y\big) \ ,$$

and it follows that $P_n\mathcal{L}_f \geq \alpha P_n\mathcal{L}_{f_0}$ for f_0 in $\mathcal{F}$ such that $P\mathcal{L}_{f_0} = r^*$. Thus

$$P_n\mathcal{L}_f \geq \alpha P_n\mathcal{L}_{f_0} = \alpha\big(P\mathcal{L}_{f_0} + (P_n - P)\mathcal{L}_{f_0}\big) = \alpha\big(r^* - (P - P_n)\mathcal{L}_{f_0}\big) \geq \alpha\big(r^* - \sup_{f\in\mathcal{F}_{r^*}} |(P_n - P)(\ell_f - \ell_{f^*_{\mathcal{F}}})|\big) \ ,$$

where $\mathcal{F}_{r^*} = \{f \in \mathcal{F} : P\mathcal{L}_f = r^*\}$. Thus, the rest of the analysis consists in finding the smallest $r > 0$ such that with high probability

$$\sup_{f \in \mathcal{F}_r} |(P_n - P)(\ell_f - \ell_{f^*_{\mathcal{F}}})| < r \ .$$

Let $A_n(r)$ be such that, with high probability

$$\sup_{f \in \mathcal{F}_r} |(P_n - P)(\ell_f - \ell_{f^*_{\mathcal{F}}})| < A_n(r) \ .$$

Then with the same probability $P\mathcal{L}_{\hat{f}_{\mathcal{F}}} \leq \inf\{r > 0 : A_n(r) < r\}$.

So far, we have not used any concentration results. There are many different ways to construct upper bounds $A_n(r)$ on the supremum of empirical processes. A very general and common approach is based on Talagrand's inequality. In particular we can use the bounds proved by Bousquet (Bousquet, 2002) and Klein (Klein, 2002; Klein et al., 2005) that we recall here.

Theorem 1.2: Bousquet and Rio-Klein inequalities

Let $\mathcal{F}$ be a class of measurable functions from E into $[0, 1]$ and

$$\sigma_P^2(\mathcal{F}) = \sup_{f \in \mathcal{F}} \left(Pf^2 - (Pf)^2\right)$$

1. **Bousquet bound** (Bousquet, 2002): for all $t > 0$, with probability larger than $1 - \exp(-t)$

$$\sup_{f \in \mathcal{F}} |(P_n - P)(f)| \leq \mathbb{E} \sup_{f \in \mathcal{F}} |(P_n - P)(f)| + \sqrt{2\frac{t}{n}\left(\sigma_P^2(\mathcal{F}) + 2\mathbb{E} \sup_{f \in \mathcal{F}} |(P_n - P)(f)|\right)} + \frac{t}{2n}$$

2. **Rio-Klein bound** (Klein, 2002; Klein et al., 2005): for all $t > 0$, with probability larger than $1 - \exp(-t)$

$$\sup_{f \in \mathcal{F}} |(P_n - P)(f)| \geq \mathbb{E} \sup_{f \in \mathcal{F}} |(P_n - P)(f)| - \sqrt{2\frac{t}{n}\left(\sigma_P^2(\mathcal{F}) + 2\mathbb{E} \sup_{f \in \mathcal{F}} |(P_n - P)(f)|\right)} - \frac{t}{n}$$

The interval $[0, 1]$ can be replaced by any bounded interval by a simple re-scaling argument. Theorem 1.2 states that the supremum of an empirical process over a bounded class of functions concentrates well around its expectation. While Mc Diarmid's inequality 1.1 provides a uniform version of Hoeffding's inequality for bounded classes, Talagrand's inequality depends on the variance of the functions class and can be seen a uniform version of Bernstein's inequality.

If the loss function is 1-Lipschitz and $\ell_{\mathcal{F}} - \ell_{f^*_{\mathcal{F}}} = \{\ell_f - \ell_{f^*_{\mathcal{F}}}, f \in \mathcal{F}\}$ is bounded by 1, from

Theorem 1.2 and Lemma 1.1, with probability larger than $1-\exp(-t)$

$$\sup_{f\in\mathcal{F}_r}|(P_n-P)(\ell_f-\ell_{f^*_{\mathcal{F}}})|\leq 2\mathcal{R}_n(\ell_{\mathcal{F}_r})+\sqrt{\frac{2t}{n}\Big(r+4\mathcal{R}_n(\ell_{\mathcal{F}_r})\Big)}+\frac{t}{2n}\ ,$$

and the following theorem easily follows.

Theorem 1.3: Excess risk for ERM associated with bounded classes $\ell_{\mathcal{F}}$

Let ℓ be a 1-Lipschitz loss function. Let $\mathcal{F}\subset L_2(P_X)$ be a closed convex class of functions such that $\ell_{\mathcal{F}}-\ell_{f^*_{\mathcal{F}}}$ is bounded by 1. There exists constant, $c_1,c_2,c_3>0$ such that the following holds: for every $t>0$, with probability larger than $1-\exp(-c_1t)$ the minimizer of the empirical risk over $\mathcal{F}$ satisfies

$$R(\hat{f}_{\mathcal{F}})-R(f^*_{\mathcal{F}})\leq c_2\max\left((r^*)^2,\frac{t}{n}\right)\ ,$$

where

$$r^*=\mathbb{E}\sup_{f\in\mathcal{F}:P\mathcal{L}_f\leq r}\left|\frac{1}{n}\sum_{i=1}^n\sigma_i(f-f^*_{\mathcal{F}})(X_i)\right|\leq c_3r\ ,$$

for $\sigma_1,\cdots,\sigma_n$ independent symmetric $\{-1,1\}$-valued random variables that are independent of $X_1,\cdots,X_n$. The expectation is taken with respect to both the Rademacher random variables and the inliers $(X_1,\cdots,X_n)$.

The excess risk of the empirical risk minimizer $\hat{f}_{\mathcal{F}}$ is expressed as a fixed-point parameter depending on local Rademacher complexities. Thank to the localization, it is not necessary to assume that whole the class $\ell_{\mathcal{F}}-\ell_{f^*_{\mathcal{F}}}$ is bounded but only that the sub-class $\ell_{\mathcal{F}_{r^*}}-\ell_{f^*_{\mathcal{F}}}=\{\ell_f-\ell_{f^*_{\mathcal{F}}}:P\mathcal{L}_f=r^*\}$ is bounded.

Let us come back to Example 2 presented in Section 1.2.2 and assume that Y is bounded by 1. Under this assumption, the quadratic loss function is 2-Lipschitz and $\ell_{\mathcal{F}}$ is bounded by $2BD$. Let us also assume that $\lambda_{\min}\big(\mathbb{E}\big[\Phi(X)^T\Phi(X)\big]\big)\geq\gamma$, for $\gamma>0$ an absolute constant, where $\lambda_{\min}(\Sigma)$ denotes the smallest eigenvalue of a symmetric matrix Σ. For $r>0$

$$\begin{aligned}\mathbb{E}\sup_{f\in\mathcal{F}:P\mathcal{L}_F\leq r}\left|\frac{1}{n}\sum_{i=1}^n\sigma_i(f-f^*_{\mathcal{F}})(X_i)\right|&\leq\mathbb{E}\sup_{\beta\in\mathbb{R}^p:\mathbb{E}\left\langle\beta-\beta^*,\Phi(X)\right\rangle^2\leq 2r}\left|\frac{1}{n}\sum_{i=1}^n\sigma_i\big\langle\beta-\beta^*,\Phi(X)\big\rangle\right|\\&\leq\mathbb{E}\sup_{\beta\in\mathbb{R}^p:\gamma\|\beta-\beta^*\|_2^2\leq 2r}\left|\frac{1}{n}\sum_{i=1}^n\sigma_i\big\langle\beta-\beta^*,\Phi(X)\big\rangle\right|\\&\leq\sqrt{\frac{2r}{\gamma}}\left\|\frac{1}{n}\sum_{i=1}^n\sigma_i\Phi(X_i)\right\|_2\leq D\sqrt{\frac{2r}{\gamma n}}\ ,\end{aligned}$$

and with probability larger than $1-\exp(-t)$,

$$R(\hat{\beta})-R(\beta^*)=\mathbb{E}\big\langle\hat{\beta}-\beta^*,\Phi(X)\big\rangle^2\leq\max\left(\frac{c(\gamma,D)}{n},\frac{t}{n}\right)\ ,$$

where $c(\gamma, D) > 0$ is a constant depending on D and γ. Using localization arguments we obtain fast rates of convergence i.e rates of order $\mathcal{O}(1/n)$.

1.3.2 Toward a more general analysis

Theorem 1.3 relies heavily on the fact that the class $\ell_{\mathcal{F}}$ is bounded. This setting excludes many natural and common problems in statistics. For example, in Section 1.3, we considered $\mathcal{F} = \{\langle \cdot, t \rangle, t \in \mathbb{R}^p\}$, the class of linear functional indexed by $\mathbb{R}^p$. As soon as X has not a compact support the class $\mathcal{F}$ and thus $\ell_{\mathcal{F}}$ are unbounded. Consequently, Theorem 1.3 does not cover the case where the design X is Gaussian, yet extensively studied in statistics. Moreover, for the quadratic loss function, the noise was also assumed to be bounded. Thus, the *Gaussian model* is excluded from the analysis of Theorem 1.3. The Gaussian model is when $Y = f^*_{\mathcal{F}}(X) + W$, for some $f^*_{\mathcal{F}}$ in $\mathcal{F}$ and W is a centered Gaussian variable with variance σ^2. The target Y consists in noisy measurements of $f^*_{\mathcal{F}}$ corrupted by a Gaussian noise. Despite the fact that boundedness assumptions cannot be used for very standard statistical problems, it has been used very frequently in Learning Theory (Bartlett and Mendelson, 2002; Bartlett et al., 2002a; Massart, 2000) and (Koltchinskii, 2011b) for a good survey. There are two main reasons:

1. Concentration inequalities: Versions of Talagrand's inequality such as Theorem 1.2 were extensively used in Learning Theory. Local Rademacher complexities naturally appear as measures of the complexity.

2. When the class $\mathcal{F}$ and the target are both bounded, the loss function $\ell(y, y') = (y - y')^2/2$ is Lispchitz and one can use contraction argument to show that

$$\mathbb{E} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i \big((\ell(f(X_i), Y_i) - \ell(f^*_{\mathcal{F}}(X_i), Y_i) \big) \right| \leq c \mathbb{E} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i (f - f^*_{\mathcal{F}})(X_i) \right| \ ,$$

   for $c > 0$ an absolute constant.

Lecué and Mendelson (Lecué and Mendelson, 2013) studied learning problems with the quadratic loss function without boundedness assumption on the envelope of $\{\ell(f(X), Y) : f \in \mathcal{F}\}$. They were interested in the most natural setting extending bounded classes: the subgaussian framework (see Definition 1.3). It covers the case of regression with Gaussian noise. Their analysis follows the "isomorphic method" based on the following idea. Let r^* such that with large probability

$$\forall f \in \mathcal{F} : P\mathcal{L}_f \geq r^*, \quad \frac{1}{2} P\mathcal{L}_f \leq P_n \mathcal{L}_f \leq \frac{3}{2} P\mathcal{L}_f \ ,$$

then, with the same probability, the empirical risk minimizer $\hat{f}_{\mathcal{F}}$ satisfies

$$P\mathcal{L}_{\hat{f}_{\mathcal{F}}} \leq r^* \ ,$$

because $P_n\mathcal{L}_{\hat{f}_{\mathcal{F}}} \leq 0$. The idea of the isomorphic method is to identify the right level r^* such that with large probability

$$\sup_{f\in\mathcal{F}:P\mathcal{L}_f\geq r^*} \left|(P_n - P)(\ell_f - \ell_{f^*_{\mathcal{F}}})\right| \leq \frac{1}{2}P\mathcal{L}_f \ .$$

Under the assumption that the noise[4] $Y - f^*_{\mathcal{F}}(X)$ is σ-subgaussian and that the class $\mathcal{F} - f^* = \{f - f^* : f \in \mathcal{F}\}$ is B-subgaussian (see Definition 1.3), the authors derived optimal bounds for the excess risk holding with an exponentially large probability. Their bounds are fixed-points depending on another notion of complexity measure: The Gaussian complexity (see Section 1.5.1 for a precise definition).

Definition 1.3: Subgaussian random variable and subgaussian class

Let P_X be a probability measure on (E, ε) and let X be distributed according to P_X.

1. We say that X is σ-subgaussian if for every $\lambda > 0$
$$\mathbb{E}\exp\left(\lambda X\right) \leq \exp\left(\frac{\lambda^2\sigma^2}{2}\right)$$
2. We say that a class $\mathcal{F} \subset L_2(P_X)$ is B-subgaussian with respect to P_X if for every f, g in $\mathcal{F}$ and $\lambda > 0$
$$\mathbb{E}\exp\left(\lambda\frac{|f(X)|}{\|f\|_{L_2(P_X)}}\right) \leq \exp\left(\frac{\lambda^2B^2}{2}\right)$$

In a path-breaking paper (Mendelson, 2014), Mendelson presented a new general analysis for the quadratic loss function, $\ell(y, y') = (y - y')^2/2$ allowing to handle general classes of functions (not necessarily subgaussian). His starting point is that the isomorphic method consisting in showing that

$$\forall f \in \mathcal{F} \text{ s.t } P\mathcal{L}_f \geq r^*, \quad \frac{1}{2}P\mathcal{L}_f \leq P_n\mathcal{L}_f \leq \frac{3}{2}P\mathcal{L}_f \ ,$$

for a well chosen parameter r^*, is too restrictive. Only the lower estimate $(1/2)P\mathcal{L}_f \leq P_n\mathcal{L}_f$ is actually required. The key assumption leading to the lower bound is the following *small-ball condition* stating that there exist $u, \gamma > 0$ such that

$$\inf_{f,g\in\mathcal{F}} \mathbb{P}\left(|f(X) - g(X)| \geq u\|f - g\|_{L_2(P_X)}\right) \geq \gamma > 0$$

Note that if $\|f\|_{L_4(P_X)} \leq \kappa\|f\|_{L_2(P_X)}$ for every f in $\mathcal{F}$, then by the Paley-Zygmund inequality, for every f, g in $\mathcal{F}$

$$\mathbb{P}\left(|f(X) - g(X)| \geq u\|f - g\|_{L_2(P_X)}\right) \geq \left(\frac{1 - u^2}{\kappa^2}\right)^2 \ .$$

[4] we use the terminology of (Lecué and Mendelson, 2013).

Thus, the small-ball assumption can be understood as an equivalence norm assumption. Note that for subgaussian classes, we have $\|f\|_{L_p(P_X)} \leq \kappa\sqrt{p}\|f\|_{L_2(P_X)}$ for every $p > 0$ and thus the small-ball assumption is automatically satisfied. Under this assumption, Mendelson derives upper bounds on the excess risk depending on fixed-point complexity parameters defined with Rademacher complexities, holding with large probability.

1.4 Localization methods for regularized procedures

1.4.1 Regularized empirical risk minimizer

In Section 1.3, we presented localization arguments to derive fast rates of convergence for the empirical risk minimizer. When the class F is too large, the localization is not sufficient to obtain small excess risk. A regularization term, promoting an expected behavior of the oracle, can be added to enforce a similar structural property of the estimator. This approach is called the regularization method.

Example 1 : Promoting sparsity. Let $\mathcal{F} = \{\langle \beta, \cdot \rangle, \beta \in \mathbb{R}^p\}$ be the class of linear functionals in $\mathbb{R}^p$ and set β^* to be the minimizer of the risk $R(\beta) = \mathbb{E}_{(X,Y)\sim P}[\ell(\langle \beta, X \rangle, Y)]$ in $\mathbb{R}^p$, where ℓ denotes a convex function. The celebrated LASSO estimator (Tibshirani, 1996) is defined as

$$\hat{\beta}_n^\lambda \in \operatorname*{argmin}_{\beta\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^n \ell\big(\langle \beta, X_i \rangle, Y_i\big) + \lambda\|\beta\|_1 \ ,$$

where $\|\beta\|_1 = \sum_{i=1}^p |\beta_i|$, and $\lambda > 0$ is a parameters that has to be chosen carefully. Remarkably, for a well chosen parameter $\lambda > 0$, under some conditions on X and Y, with high probability (Bellec et al., 2018; Van de Geer et al., 2008; Bickel et al., 2009; Bunea et al., 2007), one can show that

$$\|\hat{\beta}_n^\lambda - \beta^*\|_2^2 \leq s\frac{\log(p)}{n} \ ,$$

where $s = \|\beta^*\|_0 = \sum_{i=1}^p \mathbf{1}\{\beta_i^* \neq 0\}$ denotes the sparsity of the oracle β^*. In this example, the penalization $\|\beta\|_1$ promotes sparse solutions.

Example 2 : Promoting low rank matrices. Let $\mathcal{F} = \{\langle M, \cdot \rangle, M \in \mathbb{R}^{m\times T}\}$, where $\langle A, B \rangle = \mathrm{Tr}(A^T B)$ for any matrices A, B in $\mathbb{R}^{m\times T}$. For $A \in \mathbb{R}^{m\times T}$, set $(\sigma_i(A))_{i\in[\![1,\min(m,T)]\!]}$ its singular values arranged in a non-increasing order. The 1-Schatten norm is simply the trace-norm i.e $\|A\|_1 = \mathrm{Tr}\big(\sqrt{A^T A}\big) = \sum_{i=1}^{\min(m,T)} \sigma_i(A)$. The trace norm regularization procedure is defined as following

$$\hat{A}_n^\lambda \in \operatorname*{argmin}_{A\in\mathbb{R}^{m\times T}} \frac{1}{n}\sum_{i=1}^n \ell\big(\langle A, X_i \rangle, Y_i\big) + \lambda\|A\|_1 \ .$$

This procedure was introduced for low-rank reconstruction of high-dimensional matrices (Gross, 2011; Candes and Plan, 2011; Recht et al., 2010; Rohde et al., 2011). The trace norm has

similar properties as the ℓ_1-norm. In this example, the penalization $\|A\|_1$ promotes low rank solutions.

Example 3 : Promoting smooth solutions. Consider a set $\mathcal{X}$ and let $\mathcal{H} \subset \mathbb{R}^{\mathcal{X}}$ be a Hilbert space of real valued functions on $\mathcal{X}$ with the inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$. The function $K : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ is called a *reproducing kernel* of $\mathcal{H}$ if

- For any x in $\mathcal{X}$ the space $\mathcal{H}$ contains the functions $K_x : \mathcal{X} \mapsto \mathbb{R}$ s.t $K_x(y) = K(x, y)$.
- For any $x \in \mathcal{X}$ and $f \in \mathcal{H}$, $\langle K_x, f \rangle = f(x)$, called the *reproducing property.*

If a reproducing kernel exists, the space $\mathcal{H}$ is called reproducing kernel Hilbert space (RKHS). A *positive definite kernel* K is a symmetric function $K : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ such that for any $n \in \mathbb{N}^*$ and $x_1, \cdots, x_n$ in $\mathcal{X}$ the matrix $\big(K(x_i, x_i) : (i, j) \in [\![1, n]\!]^2\big)$ is positive definite. In (Aronszajn, 1950), the author established that for any positive definite kernel K, there exists a unique reproducing kernel Hilbert space reproducing K. Thus, for a positive definite kernel K, let $\mathcal{H}_K$ be the unique RKHS associated with K and let $\mathcal{F} = \mathcal{H}_K$. A very popular approach is the Tikhonov regularization procedure defined as following

$$\hat{f}_n^{\lambda} = \underset{f \in \mathcal{H}_K}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \ell\big(f(X_i), Y_i\big) + \lambda \|f\|_{\mathcal{H}_K}^2 \ ,$$

where $\|\cdot\|_{\mathcal{H}_K}$ denotes the norm derived from the inner product in $\mathcal{H}_K$. From the reproducing property and Cauchy-Schwarz inequality, for every $x, y \in \mathcal{X}$ and $f \in \mathcal{H}_K$

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

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. In this example, the penalization $\|f\|_{\mathcal{H}_K}$ promotes smoothness with respect to the metric induced by the kernel K on $\mathcal{X}$.

The first two examples and the third one are very different in nature. ℓ_1 and S_1 penalizations are used to expose the sparse nature of the oracle $f_{\mathcal{F}}^*$ (sparse of low rank oracle). Although the ℓ_1 norm does not appear to be directly connected to the notion of sparsity, surprisingly, it promotes sparse solution. This "modern" approach of regularization has been extensively studied in the statistical community since the early 2000s.
Example 3 deals with the "classical" point of view in regularization. One may think that the oracle $f_{\mathcal{F}}^*$ has a certain substructure (smooth for example) and that $\Psi(f_{\mathcal{F}}^*)$ is not too big ($\Psi(\cdot)$ being the penalization function). The regularized procedure is expected to produce estimates $\hat{f}_n^{\lambda}$ for which $\Psi(f_n^{\lambda})$ is of the order of $\Psi(f_{\mathcal{F}}^*)$ and its excess risk should depend on $\Psi(f_{\mathcal{F}}^*)$.

More formally, let E be a vector space such that $\mathcal{F} \subset E$ and $\Psi : E \mapsto \mathbb{R}_+$ be a penalty function. It is often assumed that the penalty is a norm (Lecué and Mendelson, 2018) (as in examples 1 and

2). However, in this thesis we adopt a more general point of view and assume that the penalty is only a convex function. For $\mathcal{F}$ a convex class of functions and $\lambda > 0$, the *regularized empirical risk minimizer* (RERM) is defined as

$$\hat{f}_{\mathcal{F}}^{\lambda} = \operatorname*{argmin}_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \ell\big(f(X_i), Y_i\big) + \lambda \Psi(f) \ , \tag{1.4}$$

where ℓ denotes any convex loss function and $(X_i, Y_i)_{i \in [\![1,n]\!]}$ are i.i.d random variables sampled from P. For $\lambda = 0$, we recover the non-penalized empirical risk minimizer. Small values of λ imply that the dominating term in (1.4) is the empirical risk, while large values of λ encourage solutions such that $\Psi(\hat{f}_{\mathcal{F}}^{\lambda})$ is small, even if its empirical risk may be large. The tuning parameter λ has to be chosen carefully. Since in our case the loss function ℓ, the class of functions $\mathcal{F}$ and the penalization Ψ are convex, the estimator $\hat{f}_{\mathcal{F}}^{\lambda}$ is unique and can be easily computed in practice, using tools from convex optimization (Boyd et al., 2004) such as proximal gradient methods (Schmidt et al., 2011).

1.4.2 General approach of localization for regularized procedures

The penalization term $\lambda\Psi(f)$ added to the empirical risk brings a new information that has to be included in the analysis. Since the penalization term $\Psi(\cdot)$ promotes estimators $\hat{f}_{\mathcal{F}}^{\lambda}$ such that $\Psi(\hat{f}_{\mathcal{F}}^{\lambda})$ is small, a simple and natural idea is to add a localization term taking into account the fact that $\hat{f}_{\mathcal{F}}^{\lambda}$ should be close to $f_{\mathcal{F}}^{*}$ with respect to the metric induced by Ψ. We can derive an analysis similar as the one developed in Section 1.3. For the sake of simplicity we will assume that Ψ is norm. The analysis could be extended to more general convex penalization, see Chapter 4

Let $P_n^{\lambda}\ell_f = n^{-1}\sum_{i=1}^{n} \ell\big(f(X_i), Y_i\big) + \lambda\Psi(f)$ and $P_n^{\lambda}\mathcal{L}_f = P_n^{\lambda}(\ell_f - \ell_{f_{\mathcal{F}}^*})$. By definition, $P_n^{\lambda}\ell_{\hat{f}_{\mathcal{F}}^{\lambda}} \geq P_n^{\lambda}\ell_{f_{\mathcal{F}}^*}$ and the proof consists in showing that with high probability, for every $f \in \mathcal{F}$ such that $\Psi(f - f_{\mathcal{F}}^*) > \rho^*$ or $P\mathcal{L}_f > r^*$ we have $P_n^{\lambda}\mathcal{L}_f > 0$. Automatically, with the same probability, we have

$$\Psi(\hat{f}_{\mathcal{F}}^{\lambda} - f_{\mathcal{F}}^{*}) \leq \rho^* \quad \text{and} \quad P\mathcal{L}_{\hat{f}_{\mathcal{F}}^{\lambda}} \leq r^*$$

Let $f \in \mathcal{F}$ such that $\Psi(f - f_{\mathcal{F}}^*) > \rho^*$ or $P\mathcal{L}_f > r^*$. As for the analysis of the empirical risk minimizer in Section 1.3, we want to use the homogeneity Lemma 1.3 to reduce the analysis onto the set $\{f \in \mathcal{F} : P\mathcal{L}_f = r^*\}$. Because of the localization with respect to the regularization norm, the situation is more delicate (see Figure 1.2 for a geometric representation of the problem) and two cases appear:

1. **Cone 1:** $\{f \in \mathcal{F} : P\mathcal{L}_f \leq r^*\} \subset \{f \in \mathcal{F} : \Psi(f - f_{\mathcal{F}}^*) \leq \rho^*\}$
   Use the homogeneity lemma 1.3. There exist $\alpha > 1$ and f_0 in $\mathcal{F}$ s.t $\alpha(f_0 - f_{\mathcal{F}}^*) = f - f_{\mathcal{F}}^*$ with $P\mathcal{L}_{f_0} = r^*$. Automatically $\Psi(f_0 - f_{\mathcal{F}}^*) \leq \rho^*$. By convexity of the penalization term and the loss function it follows that

$$P_n^{\lambda}\mathcal{L}_f = P_n\mathcal{L}_f + \lambda\big(\Psi(f) - \Psi(f_{\mathcal{F}}^*)\big) \geq \alpha P_n^{\lambda}\mathcal{L}_{f_0} \ ,$$

where f_0 satisfies $P\mathcal{L}_{f_0} = r^*$ and $\Psi(f_0 - f_{\mathcal{F}}^*) \leq \rho^*$.

2. **Cone 2:** $\{f \in \mathcal{F} : \Psi(f - f_{\mathcal{F}}^*) \leq \rho^*\} \subset \{f \in \mathcal{F} : P\mathcal{L}_f \leq r^*\}$
Take $\alpha = \Psi(f - f_{\mathcal{F}}^*)/\rho^* > 1$ and f_0 defined by $\alpha(f_0 - f_{\mathcal{F}}^*) = f - f_{\mathcal{F}}^*$. Thus $\Psi(f_0 - f_{\mathcal{F}}^*) = \rho^*$ and automatically $P\mathcal{L}_{f_0} \leq r^*$. As for the first case we have

$$P_n^\lambda \mathcal{L}_f = P_n \mathcal{L}_f + \lambda\big(\Psi(f) - \Psi(f_{\mathcal{F}}^*)\big) \geq \alpha P_n^\lambda \mathcal{L}_{f_0} \ ,$$

where this time, f_0 satisfies $P\mathcal{L}_{f_0} \leq r^*$ and $\Psi(f_0 - f_{\mathcal{F}}^*) = \rho^*$.

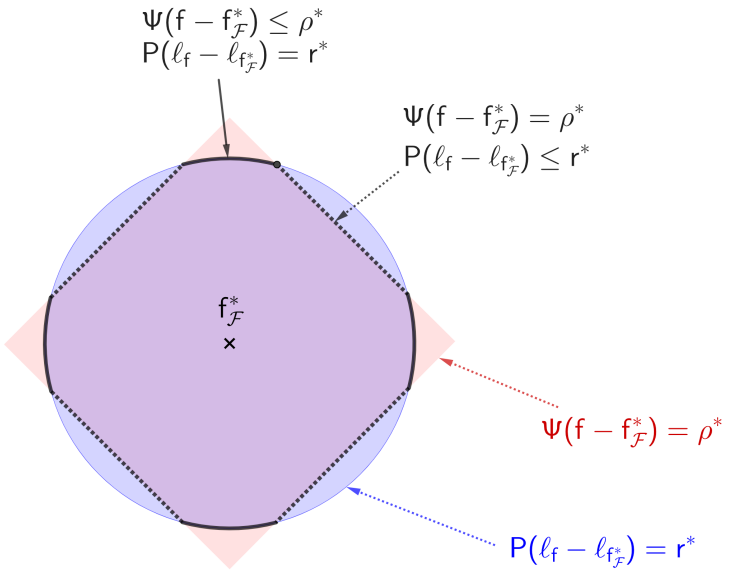


Figure 1.2: Localization cones for regularized empirical risk minimizer

Define $B(\rho^*, r^*) = \{f \in \mathcal{F} : P\mathcal{L}_f \leq r^*\} \cap \{f \in \mathcal{F} : \Psi(f - f_{\mathcal{F}}^*) \leq \rho^*\}$. The rest of the proof consists in proving that $P_n^\lambda \mathcal{L}_f \geq 0$ for every f in $\partial B(\rho^*, r^*)$, where $\partial B(\rho^*, r^*)$ denotes the border of $B(\rho^*, r^*)$.

Analysis in the cone 1: Let f_0 in $\mathcal{F}$ such that $P\mathcal{L}_{f_0} = r^*$ and $\Psi(f_0 - f_{\mathcal{F}}^*) \leq \rho^*$

$$\begin{aligned} P_n^\lambda \mathcal{L}_{f_0} &= P_n \mathcal{L}_{f_0} + \lambda\big(\Psi(f_0) - \Psi(f_{\mathcal{F}}^*)\big) \\ &\geq r^* - \sup_{f \in \mathcal{F}} \left\{ \left| (P_n - P)(\ell_f - \ell_{f_{\mathcal{F}}^*}) \right| : P\mathcal{L}_f = r^*, \Psi(f - f_{\mathcal{F}}^*) \leq \rho^* \right\} - \lambda\rho^* \\ &\geq \frac{r^*}{2} - \sup_{f \in \mathcal{F}} \left\{ \left| (P_n - P)(\ell_f - \ell_{f_{\mathcal{F}}^*}) \right| : P\mathcal{L}_f = r^*, \Psi(f - f_{\mathcal{F}}^*) \leq \rho^* \right\} \ , \end{aligned}$$

for $\lambda \leq r^*/(2\rho^*)$. The rest of the analysis is devoted to find the smallest $r > 0$ such that, with large probability

$$\sup_{f \in \mathcal{F}} \left\{ \left| (P_n - P)(\ell_f - \ell_{f_{\mathcal{F}}^*}) \right| : P\mathcal{L}_f = r, \Psi(f - f_{\mathcal{F}}^*) \leq \rho^* \right\} \leq r/2 \ .$$

Analysis in cone 2: the sparsity equation In (Lecué and Mendelson, 2018), Lecué and Mendelson studied the RERM associated with the quadratic loss function. The authors developed the notion of "sparsity equation" allowing to obtain tight bounds on the excess risk, depending on the structure of the oracle $f_{\mathcal{F}}^*$ such as its sparsity or its rank. Their intuition is based on the fact that the LASSO procedure promotes sparsity because of the large subdifferential of the ℓ_1-norm in sparse vectors. We recall that the subdifferential of Ψ in a point f is defined as

$$\big(\partial\psi\big)_f = \{z^* \in E^* : \Psi(f+h) - \Psi(f) \geq z^*(h), \qquad \forall h \in E\} \ ,$$

where E^* is the dual space of the normed space (E, Ψ). Let $t \in \mathbb{R}^p$,

$$\big(\partial\|\cdot\|_1\big)_t = \{g \in \mathbb{R}^p : \|g\|_\infty \leq 1, \langle t, g\rangle = \|t\|_1\} \ ;$$

and the subdifferential of $\|\cdot\|_1$ is larger for sparse than non-sparse vectors. The penalization would shift the estimates toward subspaces with large subdifferentials. This phenomenon can be extended to other penalization functions. Let

$$S(r,\rho) = \{f \in \mathcal{F} : P\mathcal{L}_f \leq r\} \cap \{f \in \mathcal{F} : \Psi(f - f_{\mathcal{F}}^*) = \rho\} \ ,$$

and

$$\Delta(\rho, r) = \inf_{f \in S(r,\rho)} \sup_{z^* \in (\partial\Psi)_{f_{\mathcal{F}}^*}} z^*(f - f_{\mathcal{F}}^*) \ .$$

It is expected that $\Delta(\rho, r)$ will be large if the subdifferential in the oracle $f_{\mathcal{F}}^*$ is large. Let $f_0 \in S(r^*, \rho^*)$ From the subdifferential definition, for any z^* in $(\partial\Psi)_{f_{\mathcal{F}}^*}$

$$\Psi(f_0) - \Psi(f_{\mathcal{F}}^*) \geq z^*(f_0 - f_{\mathcal{F}}^*) \geq \inf_{f \in S(\rho^*, r^*)} z^*(f - f_{\mathcal{F}}^*) \ ,$$

and since it holds for any z^* in $(\partial\Psi)_{f_{\mathcal{F}}^*}$ it follows that $\Psi(f_0) - \Psi(f_{\mathcal{F}}^*) \geq \Delta(r^*, \rho^*)$ and

$$\begin{aligned} P_n^\lambda \mathcal{L}_{f_0} &= P_n\mathcal{L}_{f_0} + \lambda\big(\Psi(f_0) - \Psi(f_{\mathcal{F}}^*)\big) \\ &\geq -\sup_{f \in \mathcal{F}} \big\{\big|(P_n - P)(\ell_f - \ell_{f_{\mathcal{F}}^*})\big| : P\mathcal{L}_f \leq r^*, \Psi(f - f_{\mathcal{F}}^*) = \rho^*\big\} + \lambda\Delta(r^*, \rho^*) \ . \end{aligned}$$

If $\Delta(r^*, \rho^*) \geq \rho^*/2$, $\lambda = r^*/(2\rho^*)$ and

$$\sup_{f \in \mathcal{F}} \big\{\big|(P_n - P)(\ell_f - \ell_{f_{\mathcal{F}}^*})\big| : P\mathcal{L}_f \leq r^*, \Psi(f - f_{\mathcal{F}}^*) = \rho^*\big\} < r^*/4 \ ,$$

then $P_n^\lambda \mathcal{L}_{f_0} > 0$. It shows why the subdifferential of $f_{\mathcal{F}}^*$ must be large.

Theorem 1.4: Deterministic results for RERM

Let r^* and ρ^* be chosen such that

$$\sup_{f \in \mathcal{F}} \big\{\big|(P_n - P)(\ell_f - \ell_{f_{\mathcal{F}}^*})\big| : P\mathcal{L}_f \leq r^*, \Psi(f - f_{\mathcal{F}}^*) \leq \rho^*\big\} < r^*/4 \ ,$$

and

$$\Delta(r^*, \rho^*) \geq \rho^*/2,$$

Then for $\lambda = r^*/(2\rho^*)$ we have

$$\Psi(\hat{f}_{\mathcal{F}}^{\lambda} - f_{\mathcal{F}}^*) \leq \rho^* \quad \text{and} \quad P\mathcal{L}_{\hat{f}_{\mathcal{F}}^{\lambda}} \leq r^*$$

Theorem 1.4 is completely deterministic. Thus, to obtain upper bound on the excess risk it is sufficient to construct a tight upper bound $A_n(r, \rho)$ such that with high probability.

$$\sup_{f \in \mathcal{F}} \left\{ \left| (P_n - P)(\ell_f - \ell_{f_{\mathcal{F}}^*}) \right| : P\mathcal{L}_f \leq r, \Psi(f - f_{\mathcal{F}}^*) \leq \rho \right\} \leq A_n(r, \rho) \ ,$$

and with the same probability, r^* and ρ^* defined as

$$r^* = \inf \left\{ r > 0 : \Delta(r, \rho^*) \geq \frac{\rho^*}{2} \text{ and } A_n(r, \rho^*) < \frac{r^*}{4} \right\} \ ,$$

satisfy the requirements of Theorem 1.4.

Remark 1.1. *If the norm Ψ is "smooth", in the sense that the subdifferential of Ψ in f is small for any f, then there is little hope to have $\Delta(r, \rho) \geq \rho/2$. In this case we can always take $\rho^* = 3\Psi(f_{\mathcal{F}}^*)$ and for f_0 in the second cone (i.e $P\mathcal{L}_{f_0} \leq r^*, \Psi(f_0 - f_{\mathcal{F}}^*) = \rho^*$) we have*

$$\Psi(f_0) - \Psi(f_{\mathcal{F}}^*) \geq \Psi(f_0 - f_{\mathcal{F}}^*) - 2\Psi(f_{\mathcal{F}}^*) = \Psi(f_{\mathcal{F}}^*) \ ,$$

and by chosing r^* such that

$$\sup_{f \in \mathcal{F}} \left\{ \left| (P_n - P)(\ell_f - \ell_{f_{\mathcal{F}}^*}) \right| : P\mathcal{L}_f \leq r^*, \Psi(f - f_{\mathcal{F}}^*) \leq 3\Psi(f_{\mathcal{F}}^*) \right\} < r^*/4 \ ,$$

with $\lambda = r^*/(6\Psi(f_{\mathcal{F}}^*))$ we have

$$\Psi(\hat{f}_{\mathcal{F}}^{\lambda} - f_{\mathcal{F}}^*) \leq 3\Psi(f_{\mathcal{F}}^*) \quad \text{and} \quad P\mathcal{L}_{\hat{f}_{\mathcal{F}}^{\lambda}} \leq r^* \ .$$

This method can be applied systematically to obtain error bounds depending on $\Psi(f_{\mathcal{F}}^*)$.

1.4.3 Advantages of localization methods for regularized empirical risk minimizers

Using homogeneity arguments, the analysis is reduced to the uniform control of $\left|(P_n - P)(\ell_f - \ell_{f_{\mathcal{F}}^*})\right|$ over "localized" sub-classes of $\mathcal{F}$. There are several advantages:

1. It leads to smaller error rates than the approach based on global complexity parameters. In addition, the regularization adds another localization around the oracle $f_{\mathcal{F}}^*$, often essential to show that the RERM is minimax-rate-optimal (Lecué and Mendelson, 2018).

2. Some proofs are substantially simplified since we no longer use a peeling argument.

3. The localization $\Psi(f-f^*_{\mathcal{F}}) \leq \rho^*$ may imply that the class $\{f \in \mathcal{F} : P\mathcal{L}_f \leq r^* \text{ and } \Psi(f-f^*_{\mathcal{F}}) \leq \rho^*\}$ is bounded. For example, let $\mathcal{F} = \mathcal{H}_K$ be a RKHS associated with a bounded kernel K ($|K(x,y)| \leq 1$ for any $x, y \in \mathcal{X}$). Let us define the penalization by $\Psi(f) = \|f\|_{\mathcal{H}_K}$, where $\|\cdot\|_{\mathcal{H}_K}$ denotes the norm in the RKHS associated with K. Take $\rho^* = 3\|f^*\|_{\mathcal{H}_K}$ as in Remark 1.1. Let f in $\mathcal{F}$ be such that $\Psi(f - f^*_{\mathcal{F}}) \leq \rho^*$ and $x \in \mathcal{X}$. From the reproducing property

$$|f(x)| = |\langle K_x, f\rangle_{\mathcal{H}_K}| \leq \|f\|_{\mathcal{H}_K} \leq \|f - f^*_{\mathcal{F}}\|_{\mathcal{H}_K} + \|f^*\|_{\mathcal{H}_K} \leq 4\|f^*_{\mathcal{F}}\|_{\mathcal{H}_K} \ .$$

In this example, the control of $(P_n - P)(\ell_f - \ell_{f^*_{\mathcal{F}}})$ uniformly over $\{f \in \mathcal{F} : P\mathcal{L}_f \leq r^* \text{ and } \Psi(f - f^*_{\mathcal{F}}) \leq \rho^*\}$ can be done using Talagrand's concentration inequality (see Theorem 1.2), independently from the distribution of X. Note that bounded kernels are very common in machine learning (Shawe-Taylor et al., 2004; Scholkopf and Smola, 2001).

1.5 Complexity parameters in statistical learning

As presented in Sections 1.2 and 1.3, bounding the excess risk of the (R)ERM reduces to the uniform control of $(P_n - P)(\ell_f - \ell_{f^*_{\mathcal{F}}})$ over a sub-class $\tilde{\mathcal{F}} \subset \mathcal{F}$ (the whole class $\mathcal{F}$ for slow-rates or sub-classes when using localization arguments). When the class of functions $\ell_{\tilde{\mathcal{F}}} - f^*_{\mathcal{F}} = \{\ell_f - \ell_{f^*_{\mathcal{F}}} : f \in \tilde{\mathcal{F}}\}$ is bounded, it is possible to show that the supremum of the empirical process indexed by $\ell_{\tilde{\mathcal{F}}} - f^*_{\mathcal{F}}$ concentrates well around its expectation (see Theorem 1.2). Informally, with large probability

$$\sup_{f \in \tilde{\mathcal{F}}} |(P_n - P)(\ell_f - \ell_{f^*_{\mathcal{F}}})| \approx \mathbb{E} \sup_{f \in \tilde{\mathcal{F}}} |(P_n - P)(\ell_f - \ell_{f^*_{\mathcal{F}}})|$$

From the symmetrization principle (see Lemma 1.1), $\mathbb{E}\sup_{f \in \tilde{\mathcal{F}}} |(P_n - P)(\ell_f - \ell_{f^*_{\mathcal{F}}})|$ can be upper bounded by the Rademacher complexity of $\ell_{\tilde{\mathcal{F}}} - \ell_{f^*_{\mathcal{F}}}$ and thus, the performances of the (R)ERM directly depend on the Rademacher complexity $\mathcal{R}_n(\ell_{\tilde{\mathcal{F}}} - \ell_{f^*_{\mathcal{F}}})$. Thus, Rademacher complexities naturally appear when considering bounded classes of functions. However, there exist other ways of measuring the complexity of a class of functions $\mathcal{F}$. This section is devoted to present one of these alternatives: The Gaussian complexity. Since the computation of Gaussian and Rademacher complexities may be involved sometimes, we also present some tools to bound these quantities in practice.

1.5.1 Another examples of complexity: the Gaussian complexity

Definition 1.3 presents the notion of subgaussian classes extending the one of bounded classes. A natural way of measuring the complexity of subgaussian classes is via the Gaussian complexity of $\mathcal{F}$ that we introduce now.

Let $\{G_f : f \in \mathcal{F}\}$ be the canonical Gaussian process indexed by $\mathcal{F} \subset L_2(P_X)$ i.e $\mathbb{E}[G_f] = 0$ for all f in $\mathcal{F}$ and the covariance function is given by the inner product in $L_2(P_X)$ that is $\mathbb{E}[G_f G_h] = \langle f, h \rangle_{L_2(P_X)}$ for every g, h in $\mathcal{F}$. The Gaussian complexity of $\mathcal{F}$ is defined as

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

In (Lecué and Mendelson, 2018), the authors studied ERM associated with subgaussian classes and obtained error rates expressed as fixed point of localized Gaussian complexities defined as

$$\inf\{r > 0 : \mathbb{E}\|G\|_{\mathcal{F}_r} \leq \gamma r \sqrt{n}\} \ ,$$

where $\gamma > 0$ is an absolute constant, n is the number of observations and $\mathcal{F}_r = \{f \in \mathcal{F} : \|f\|_{L_2(P_X)} \leq r\}$. The main reason why Gaussian complexities appear naturally when learning subgaussian classes will be given at the very end of Section 1.5.2. The quantity $\mathbb{E}\|G\|_{\mathcal{F}}$ may look complicated at a first glance but for many applications it has a simple form. For example let $\mathcal{F} = \{\langle t, \cdot \rangle : t \in T \subset \mathbb{R}^p\}$ be the class of linear functionals in $\mathbb{R}^p$ indexed by T and let X be a random vector in $\mathbb{R}^p$ with a covariance matrix Σ, then

$$\mathbb{E}\|G\|_{\mathcal{F}} = \mathbb{E} \sup_{t \in T} \langle G, t \rangle = w^*(\Sigma^{1/2} T) \ ,$$

where $G \sim \mathcal{N}(0, \Sigma)$ and $\Sigma^{1/2} T = \{\Sigma^{1/2} t : t \in T\}$. The quantity $w^*(T) = \mathbb{E} \sup_{t \in T} \langle G, t \rangle$, where $G \sim \mathcal{N}(0, I_p)$, is called the Gaussian mean-width of the set T. It is a well-known quantity, appearing in many phenomena in geometric functional analysis, (Vershynin, 2018; Holmes, 2012). One can think the Gaussian mean-width as one of the basic geometric quantities associated with subsets of $T \subset \mathbb{R}^p$, such as volume, surface area... The Gaussian mean-width of various sets T is known, see for example (Vershynin, 2018; Chatterjee and Goswami, 2019). The Gaussian complexity is also easily computable on several finite dimensional classes of functions such as

1. $\mathcal{F} = \{\langle t, \cdot \rangle : t \in T \subset \mathbb{R}^p\}$ the class of linear functionals in $\mathbb{R}^p$.
2. $\mathcal{F} = \{\langle A, \cdot \rangle : A \in \mathcal{A} \subset \mathbb{R}^{m \times T}\}$ the class of linear functionals in $\mathbb{R}^{m \times T}$ (Lecué and Mendelson, 2018).

Sometimes, the computation of Gaussian complexities is more involved. In Section 1.5.2 we present different tools to bound the expectation of the supremum of a (sub-)Gaussian process and thus $\mathbb{E}\|G\|_{\mathcal{F}}$.

1.5.2 Tools to control Rademacher and Gaussian complexities

Rachemacher and Gaussian complexities measure the richness of class of functions. Empirical risk minimizer and its regularized versions can be analyzed with fixed-point complexity parameters depending on Rademacher or Gaussian complexities. Building upper bounds for these quantities is

thus an important question. In this section we present some basic and more advanced tools to bound the expected suprema of stochastic processes (and obtain upper bounds on the Rademacher and Gaussian complexities).

Definition 1.3 presents the notion of subgaussian class of functions. This definition can be extended to other pseudo-distances. Let (T, d) be a pseudo-metric space and $(X_t)_{t\in T}$ be a stochastic process indexed by T. The process $(X_t)_{t\in T}$ is called subgaussian with respect to the pseudo-metric d if for any t, s in T the increments $X_s - X_t$ are $d(t, s)$-subgaussian i.e for any $\lambda > 0$

$$\mathbb{E}\exp\left(\lambda(X_t - X_s)\right) \leq \exp\left(\frac{\lambda^2 d^2(t,s)}{2}\right) .$$

Note that for $T = \mathcal{F} \subset L_2(P_X)$ and $d(f, g) = \|f - g\|_{L_2(P_X)}$ we recover Definition 1.3. Such stochastic processes have very remarkable properties. Let $N(T, d, \varepsilon)$ be the ε- covering number of (T, d), that is, the minimal number of balls (defined with the metric d) with radius ε needed to completely covers T. The ε-entropy of (T, d) is defined as

$$H(T, d, \varepsilon) = \log N(T, d, \varepsilon) .$$

The supremum of a subgaussian process is bounded from above by Dudley's entropy integral.

Theorem 1.5: Dudley integral

Let $(X_t)_{t\in T}$ be a subgaussian random process with respect to the pseudo-metric d. Then, for every t_0 in T, there exists an absolute constant $c > 0$ such that

$$\mathbb{E}\sup_{t\in T} X_t \leq c\int_0^{D(T)} \sqrt{H(T, d, \varepsilon)}\mathrm{d}\varepsilon \quad \text{and} \quad \mathbb{E}\sup_{t\in T} |X_t - X_{t_0}| \leq c\int_0^{D(T)} \sqrt{H(T, d, \varepsilon)}\mathrm{d}\varepsilon ,$$

where $D(T) = \sup\{d(t, s) : t, s \in T\}$ is the diameter of (T, d).

Theorem 1.5 is derived from chaining techniques (Talagrand, 2006) and is very useful to compute the Rademacher complexity $\mathcal{R}_n(\mathcal{F})$ of a class of functions $\mathcal{F}$. Conditional on $X_1, \cdots, X_n$, the process $\sqrt{n}\mathcal{R}_n(\mathcal{F})$ is subgaussian with respect to the pseudo-distance $L_2(P_n)$ and it follows that

$$R_n(\mathcal{F}) \leq \frac{c}{\sqrt{n}}\mathbb{E}\int_0^{D_n(\mathcal{F})} \sqrt{H(\mathcal{F}, L_2(P_n), \varepsilon)}\mathrm{d}\varepsilon ,$$

where $D_n(\mathcal{F}) = \sup_{f\in\mathcal{F}} P_n f^2$. Note that both the diameter $D_n(\mathcal{F})$ and the entropy with respect to the pseudo metric $L_2(P_n)$ are random. Following the approach developed in (Giné et al., 2006), it is possible to derive upper bounds on the right term under entropy conditions. Roughly, the idea consists in showing that $L_2(P_n)$ and $L_2(P_X)$ are close to each other in a certain sense, and use

bounds on the entropy defined with respect to the metric $L_2(P_X)$.

Theorem 1.5 is very useful to obtain upper bounds on Rademacher and Gaussian complexities. However, a close look at the proof reveals a potential source of looseness. To circumvent this problem, Talagrand developed the idea of *generic chaining* and introduced the so called Talagrand's γ_2 functional introduced in Definition 1.4.

Definition 1.4: Talagrand's γ_2

Let (T, d) be a metric space. A sequence $(T_s)_{s\geq 0}$ of subsets of T is said to be admissible if $|T_0| = 1$ and $1 \leq |T_s| \leq 2^{2^s}$ for every $s \geq 1$. The γ_2-functional of (T, d) is defined as

$$\gamma_2(T, d) = \inf_{(T_s)} \sup_{t\in T} \sum_{s=0}^{\infty} 2^{s/2} d(t, T_s) \ ,$$

where the infimum is with respect to all admissible sequences $(T_s)_{s\geq 0}$ and $d(t, T_s) = \min_{x\in T_s} d(t, x)$.

The γ_2-functional is a refinement of the Dudley's integral. In particular we have

$$\gamma_2(T, d) \leq c \int_0^{D(T)} \sqrt{H(T, d, \varepsilon)} \mathrm{d}\varepsilon \ ,$$

and the following theorem holds:

Theorem 1.6: Generic chaining

Let $(X_t)_{t\in T}$ be a subgaussian process with respect to the pseudo-metric d. Then, for every t_0 in T, there exists an absolute constant $c > 0$ such that

$$\mathbb{E} \sup_{t\in T} X_t \leq c\gamma_2(T, d) \quad \text{and} \quad \mathbb{E} \sup_{t\in T} |X_t - X_{t_0}| \leq c\gamma_2(T, d) \ .$$

Moreover,

$$\mathbb{P}\left(\sup_{t\in T} |X_t - X_{t_0}| \geq c\big(\gamma_2(T, d) + uD(T)\big) \right) \leq \exp(-u^2) \ ,$$

where $D(T)$ denotes the diameter of (T, d)

In practice, the γ_2-functional is often harder to compute that the Dudley's integral. The main advantage of this quantity comes from Talagrand's majorizing measure Theorem, saying that the γ_2-functional gives an optimal bound on Gaussian processes.

Theorem 1.7: Majorizing measure theorem

Let $(X_t)_{t\in T}$ be a Gaussian process with respect to the pseudo-metric d defined as $d(t, s) =$

$\sqrt{\mathbb{E}(X_t - X_s)^2}$. Then, there exist two absolute constants $c, C > 0$ such that

$$c\gamma_2(T,d) \leq \mathbb{E}\sup_{t\in T} X_t \leq C\gamma_2(T,d) \ .$$

Theorems 1.6 and 1.7 explain why the Gaussian complexity appears naturally when learning subgaussian classes. With large probability we can control $\sup_{f\in\tilde{\mathcal{F}}} |(P_n - P)(\ell_f - \ell_{f^*_{\mathcal{F}}})|$ and relate it with its diameter and the Talagrand's γ_2 which is equivalent to the Gaussian complexity.

1.6 Robustness in learning theory

Statistical learning is based on assumptions one makes on the observations. It can be an implicit or an explicit assumption about the randomness and the independence of the data. For example, in Sections 1.2 and 1.3 we assumed that $(X_i, Y_i)_{i\in[\![1,n]\!]}$ were i.i.d and (often) well concentrated. The *robustness* in learning can be defined as "the insensitivity to small deviations from the assumptions" (Huber and Ronchetti, 2011). The goal of *robust learning* is to build estimators under minimal hypotheses. Robustness issues have become popular because collected data are often contaminated, a situation that can be modeled by heavy-tailed distribution or the adjunction of outliers to the dataset. With bigger datasets, this corruption is even more likely. Informally, we say that an estimator is robust if it deviates moderately from its target even when data are not i.i.d and subgaussian. We will give more precise definitions in the sequel.
We begin by presenting the notion of robustness in the problem of mean estimation. Besides giving good insights on the notion of robustness, it will serve as the starting point for the construction of more advanced estimators.

1.6.1 The problem of mean estimation

Univariate case The most simple, yet fundamental problem in statistics, is the mean estimation. Let $X_1, \cdots, X_n$ be i.i.d real random variables with distribution P_X and mean μ_X. The goal is to estimate the mean μ_X. To do so, one can naturally use the empirical mean

$$\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \ .$$

From the law of large number, we know that $\hat{\mu}_n$ converges almost surely to μ_X and the goal is to obtain non-asymptotic results, for a fixed number of observations n. When the distribution P_X is well concentrated around its expectation μ_X, the empirical mean is a "good" estimator of the mean μ_X. For example, let $X_1, \cdots, X_n$ be i.i.d $\mathcal{N}(0, \sigma^2)$ random variables. Straightforward computations show that the empirical mean estimator satisfies defined as

$$\forall \delta \in (0,1), \quad \mathbb{P}\left(|\mu_X - \hat{\mu}_X| \geq c\sigma\sqrt{\frac{\log(1/\delta)}{n}} \right) \leq \delta \ .$$

This result easily extends when $(X_i)_{i\in[\![1,n]\!]}$ are i.i.d σ-subgaussian random variables. An important problem in statistics is to build estimators of the mean achieving the same performance as the empirical mean when the subgaussian assumption is relaxed and some outliers may have corrupted the dataset.

Problem 1: Is it possible to relax the assumption that $(X_i)_{i\in[\![1,n]\!]}$ are subgaussian and assume only the existence of a 2^{nd} moment ?

Problem 2: What if $\mathcal{O}$ outliers corrupt the data ?

The works (Devroye et al., 2016; Lerasle and Oliveira, 2011) answer the problem 1 in the univariate case. The authors define *level-dependent estimators* as estimators depending on the level of confidence δ, and show that there is no estimator independent of the confidence level with the optimal deviations $\sigma\sqrt{\log(1/\delta)/n}$ when the data are only assumed to have a second-order moment. Thus, under the only assumption of a second-order moment, it is possible to obtain the deviation $\sigma\sqrt{\log(1/\delta)/n}$ only if the estimator $\hat{\mu}(\delta)$ depends on the confidence level $\delta \in (0,1)$. Once we accept that the estimator depends on the confidence level δ, different constructions exist (Catoni, 2012; Nemirovsky and Yudin, 1983). The *Median Of Means* (MOM) scheme (Lerasle and Oliveira, 2011; Nemirovsky and Yudin, 1983) is probably the most natural and widespread construction of level-dependent estimator. Let $B_1, \cdots, B_K$ be a partition of $[\![1,n]\!]$ into K blocks of same size (for the sake of simplicity, it is here assumed that K divides n). For each block B_k, $k \in [\![1,K]\!]$, let $P_{B_k}X := (n/K)^{-1}\sum_{i\in B_k} X_i$ be the empirical mean in the block B_k. The MOM-estimator of μ_X is defined as

$$\hat{\mu}_K^{\text{MOM}} = \text{Median}(P_{B_1}X, \cdots, P_{B_K}X) \ .$$

Easy computations (see (Lerasle and Oliveira, 2011; Devroye et al., 2016) for example) show that $\hat{\mu}_K^{\text{MOM}}$ with $K = c\log(1/\delta)$ is a level-dependent *subgaussian estimator* of the mean when only assuming that $X_1, \cdots, X_n$ have a second-order moment: for $\delta \in (0,1)$ the estimator $\hat{\mu}_K^{\text{MOM}}$ with $K = c\log(1/\delta)$ verifies

$$\mathbb{P}\left(|\mu_X - \hat{\mu}_K^{\text{MOM}}(\delta)| \geq c\sigma\sqrt{\frac{\log(1/\delta)}{n}}\right) \leq \delta \ ,$$

where $c > 0$ is an absolute constant.

Problem 2 is related to the notion of breakdown points (Donoho and Huber, 1983), which has been repeatedly investigated in the statistical community. The "Median step" allows to handle a number of outliers $\mathcal{O} \leq cK$. Thus, MOM-estimators are particularly interesting because they can simultaneously solve problems 1 (heavy-tailed distributions) and 2 (corruption by outliers).

Multivariate case The extension of the median to the multivariate case is an interesting problem. In the past few years, this generalization has received a lot of attention. In particular, two different

communities with different notions of robustness have tried to tackle the problem of multivariate robust mean estimation

The statistical community is interested in constructing reliable estimators when the data $(X_i)_{i\in[\![1,n]\!]}$ might be heavy-tailed (Lugosi et al., 2019b; Depersin and Lecué, 2019; Minsker et al., 2015, 2018; Cherapanamjeri et al., 2019; Hopkins, 2018; Chen et al., 2018). Formally, let $(X_i)_{i\in[\![1,n]\!]}$ be i.i.d random vectors in $\mathbb{R}^p$ with mean $\mu_X \in \mathbb{R}^p$ and covariance matrix Σ. We say that $\hat{\mu}(\delta)$ is a *subgaussian-estimator* (Lugosi et al., 2019b) of the mean μ_X at level $\delta \in (0,1)$ if

$$\mathbb{P}\left(\|\mu_X - \hat{\mu}_n(\delta)\|_2 \geq c_1\sqrt{\frac{\mathrm{Tr}(\Sigma)}{n}} + c_2\sqrt{\frac{\|\Sigma\|\log(1/\delta)}{n}} \right) \leq \delta \ ,$$

where $c_1, c_2 > 0$ are two absolute constants and $\|\Sigma\|$ denotes the operator norm of Σ.

The computer science community considers a very different notion of robustness. The goal is to construct robust procedures when $\mathcal{O}$ outliers may contaminate the dataset (Diakonikolas et al., 2019a; Cheng et al., 2019). It covers the Huber ε-contamination model (Huber and Ronchetti, 2011) but also adversarial corrupted data. They want to construct estimators, computable in a polynomial time, with the optimal dependence with respect to the number of outliers $\mathcal{O}$. These results are also different in nature from the previous because the bounds only hold with constant probability $p < 1$.

Recently, combining ideas from (Diakonikolas et al., 2019a) and (Lugosi et al., 2019b), (Depersin and Lecué, 2019) showed that a single algorithm (computable in a nearly-linear time) solves both problems 1 and 2. This estimator is based on MOM ideas.

The construction of robust estimators of the mean can be used to build estimators solving more involved learning tasks.

1.6.2 The notion of robustness in supervized learning

In this section, we come back to the context of supervized learning. Let (X, Y) be distributed as P. Let P_X be the marginal distribution of X and $\mathcal{D} = (X_i, Y_i)_{i\in[\![1,n]\!]}$ be a dataset of n (not necessarily independent) random variables. As explained in Section 1.2, given a convex class of functions $\mathcal{F}$ and a loss $\ell : \mathcal{Y}\times\mathcal{Y} \mapsto \mathbb{R}_+$, the goal is to use the dataset $\mathcal{D}$ to approximate/compute the oracle $f^*_{\mathcal{F}}$ defined as

$$f^*_{\mathcal{F}} = \operatorname*{argmin}_{f\in\mathcal{F}} P\ell_f := \operatorname*{argmin}_{f\in\mathcal{F}} \mathbb{E}_{(X,Y)\sim P}\big[\ell\big(f(X), Y\big)\big] \ .$$

Recall that the empirical risk minimizer is defined as

$$\hat{f}_{\mathcal{F}} = \operatorname*{argmin}_{f\in\mathcal{F}} P_n\ell_f := \operatorname*{argmin}_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \ell\big(f(X_i), Y_i\big) \ .$$

We presented different results on the excess risk for the empirical risk minimizer and its regularized versions. The analysis strongly relies on concentration inequalities (see Section 1.3 and 1.4). In particular, the use of Talagrand inequality requires the boundedness of $\ell_{\mathcal{F}}$ while the framework developed in (Lecué and Mendelson, 2013) focuses on the case when the class of functions $\ell_{\mathcal{F}}$ is subgaussian and thus possesses finite exponential moments. We always assumed that the data were independent and identically distributed as P. We would like to relax these two assumptions. It leads to two different notions of robustness in supervized learning

1. **Robustness with respect to outliers:** Let $\mathcal{I}$ and $\mathcal{O}$ be such that $\mathcal{I}+\mathcal{O}=n$. Let $\mathcal{D}_{\mathcal{I}}$ be a set containing $\mathcal{I}$ informative data $(X_1,Y_1),\cdots(X_{\mathcal{I}},Y_{\mathcal{I}})$. These data are supposed to be i.i.d with distribution P. Let $\mathcal{D}_{\mathcal{O}}$ be a set containing $\mathcal{O}$ outliers $(\tilde{X}_1,\tilde{Y}_1),\cdots(\tilde{X}_{\mathcal{O}},\tilde{Y}_{\mathcal{O}})$. On these data, nothing is assumed. We say that an estimator $\hat{f}_n$ is robust with respect to $\mathcal{O}$ outliers if
$$\mathbb{P}\bigg(R(\hat{f}_n)-R(f^*_{\mathcal{F}})\geq \zeta_n\big|\mathcal{D}_{\mathcal{I}}\cap\mathcal{D}_{\mathcal{O}}\bigg)=\mathbb{P}\bigg(R(\hat{f}_n)-R(f^*_{\mathcal{F}})\geq c\zeta_n\big|\mathcal{D}\bigg)\ ,$$
where $c>0$ is an absolute constant, $\zeta_n>0$ and $\mathcal{D}=(X_i,Y_i)_{i\in[\![1,n]\!]}$ is a dataset containing n i.i.d random variables distributed as P. In other words, an estimators $\hat{f}_n$ is robust to $\mathcal{O}$ outliers if (up to an absolute constant) its risk remains unchanged while introducing at most $\mathcal{O}$ outliers in the dataset.
It is also possible to consider variants of this setting. For example, let $X_1,\cdots,X_n$ be i.i.d random variables distributed as P_X and $Y_1,\cdots,Y_{\mathcal{I}}$ be i.i.d random variables such that $(X_1,Y_1),\cdots,(X_{\mathcal{I}},Y_{\mathcal{I}})$ are i.i.d with common distribution P. Nothing is assumed on the $\mathcal{O}$ outliers $\tilde{Y}_{\mathcal{I}+1},\cdots,\tilde{Y}_n$ and let $\mathcal{D}=(X_i,Y_i)_{i\in[\![1,\mathcal{I}]\!]}\cup(X_i,\tilde{Y}_i)_{i\in[\![\mathcal{I}+1,n]\!]}$. This case covers situations where only the labels are corrupted by outliers.

2. **Robustness with respect to heavy-tailed:** We say that an estimator $\hat{f}_n$ is *robust to heavy-tailed distribution at the order k* if
$$\mathbb{P}\bigg(R(\hat{f}_n)-R(f^*_{\mathcal{F}})\geq \zeta_n|\mathcal{D}_{P_k}\bigg)=\mathbb{P}\bigg(R(\hat{f}_n)-R(f^*_{\mathcal{F}})\geq c\zeta_n|\mathcal{D}_{P^*}\bigg)\ ,$$
where $c>0$ is an absolute constant, $\mathcal{D}_{P_k}=(X_i,Y_i)_{i\in[\![1,n]\!]}$ is a dataset of i.i.d random variables with common distribution P_k and $\mathcal{D}_{P^*}=(X_i,Y_i)_{i\in[\![1,n]\!]}$ is a dataset of i.i.d random variables with common distribution P^*. These distributions are such that, for $(X,Y)\sim P_{k^*}$, the class $\{\ell\big(f(X),Y\big),f\in\mathcal{F}\}$ is subgaussian and for $(X,Y)\sim P_k$ the class $\{\ell\big(f(X),Y\big),f\in\mathcal{F}\}$ has only k-th order moments. In words, an estimator $\hat{f}_n$ is robust with respect to heavy-tailed distribution if one can prove rates of convergence for its excess risk which are as good if the class $\{\ell\big(f(X),Y\big),f\in\mathcal{F}\}$ is subgaussian or if it only satisfies moment conditions.

1.6.3 The limitations of the empirical risk minimizer

The empirical risk minimizer and its regularized counterpart are widespread in machine learning and statistics. They are used for various of real-world applications. Since large datasets are the

most vulnerable to corruption, it is a very important question to know whether ERM and RERM are reliable in such settings.

Let $\ell(y-y') = (y-y')^2/2$ be the quadratic loss function and $\mathcal{F} = \{\langle t, \cdot \rangle, t \in \mathbb{R}^p\}$ be the class of linear functions indexed by $\mathbb{R}^p$. Neither the loss nor the class $\mathcal{F}$ are bounded. Let $\mathcal{D} = (X_i, Y_i)_{i\in[\![1,n]\!]} \in (\mathbb{R}^p \times \mathbb{R})^n$ be a dataset of n random variables. The empirical risk minimizer is defined as

$$\hat{\beta}_n = \operatorname*{argmin}_{\beta\in\mathbb{R}^p} \frac{1}{2n}\sum_{i=1}^{n} \big(\langle X_i, \beta\rangle - Y_i\big)^2 \ .$$

In this case, it is clear that a single outlier (X_o, Y_o) can break down performance of the (R)ERM. This phenomenon occurs if outliers contaminate only the labels $(Y_i)_{i\in[\![1,n]\!]}$, only the inputs $(X_i)_{i\in[\![1,n]\!]}$, or both. Thus, (R)ERM with the quadratic loss function is not robust with respect to outliers at all. In addition, it is quite obvious to see that the estimator $\hat{\beta}_n$ is not robust to heavy-tailed distributions. For heavy-tailed distributions the empirical risk can be very far from the true risk and the empirical risk minimizer is not reliable in such cases. Indeed, let us come back to the problem of univariate mean estimation. Given i.i.d random variables $(X_i)_{i\in[\![1,n]\!]}$ sampled from P_X, the goal is to estimate $\mu_X = \mathbb{E}[X_1]$. Let $\mathcal{F} = \mathbb{R}$ and $\ell(y, y') = (y-y')/2$, for any y, y' in $\mathbb{R}$. We have

$$f^*_{\mathcal{F}} = \operatorname*{argmin}_{f\in\mathcal{F}} \mathbb{E}(f-X)^2/2 = \mu_X \quad \text{and} \quad \hat{f}_n = \operatorname*{argmin}_{f\in\mathcal{F}} \frac{1}{2n}\sum_{i=1}^{n}(f-X_i)^2 = \frac{1}{n}\sum_{i=1}^{n} X_i \ ,$$

and if P_X is heavy-tailed, the estimator $\hat{f}_n = n^{-1}\sum_{i=1}^n X_n$ is very far from the oracle $f^*_{\mathcal{F}} = \mu_X$. Consequently, there is hope to obtain general results of robustness for the (R)ERM associated with the quadratic loss function (Mendelson, 2014; Lecué and Mendelson, 2018).
Now, let $\mathcal{F}$ be a general class of functions and let ℓ denote the quadratic loss. As presented in Section 1.3, the (R)ERM achieves good performances when $P_n(\ell_f - \ell_{f^*})$ concentrated well around its expectation $P(\ell_f - \ell_{f^*})$ uniformly over sub-classes of $\mathcal{F}$. From the following decomposition

$$f(X_i) - Y_i = f(X_i) - f^*_{\mathcal{F}}(X_i) + f^*_{\mathcal{F}}(X_i) - Y_i = (f - f^*_{\mathcal{F}})(X_i) - \xi_i \ ,$$

for $f \in \mathcal{F}$ and where $\xi_i = Y_i - f^*_{\mathcal{F}}(X_i)$, it follows that

$$P_n(\ell_f - \ell_{f^*_{\mathcal{F}}}) = \underbrace{\frac{1}{2n}\sum_{i=1}^{n}(f-f^*_{\mathcal{F}})^2(X_i)}_{\text{quadratic process}} - \underbrace{\frac{1}{n}\sum_{i=1}^{n}\xi_i(f-f^*_{\mathcal{F}})(X_i)}_{\text{multiplier process}} \ .$$

Consequently, the process $\big(P_n(\ell_f - \ell_{f^*_{\mathcal{F}}})\big)_{f\in\mathcal{F}}$ concentrates well around its expectation, if and only if, both the *quadratic and the multiplier processes* concentrate well. It is clear that the quadratic process behaves nicely only when the class $F - f^* = \{f - f^*_{\mathcal{F}}, f \in \mathcal{F}\}$ concentrates well. The concentration of the multiplier process (Mendelson, 2016, 2017) involves both the random variables

$\big(f(X_i) - f^*_{\mathcal{F}}(X_i)\big)_{i\in[\![1,n]\!]}$ and $\big(Y_i - f^*_{\mathcal{F}}(X_i)\big)_{i\in[\![1,n]\!]}$. The random variable $Y_i - f^*_{\mathcal{F}}(X_i)$ can be seen as the noise of the problem [5]. As these processes don't concentrate well under moments conditions, the excess risk of the (R)ERM deteriorates when the noise $Y - f^*_{\mathcal{F}}(X)$ and the class $\mathcal{F} - f^*_{\mathcal{F}}$ are not assumed subgaussian. Thus, even when the class $\mathcal{F} - f^*_{\mathcal{F}}$ is bounded, there is no hope for the (R)ERM with the quadratic loss to be reliable if the noise $Y - f^*_{\mathcal{F}}(X)$ is heavy-tailed.

Consider an unbounded loss function ℓ. Several examples of such functions are given below. The concentration of $P_n(\ell_f - \ell_{f^*_{\mathcal{F}}})$ requires a subgaussian assumption on the class $\mathcal{F} - f^*_{\mathcal{F}}$. However, contrary to the quadratic loss function, the noise does not need to be subgaussian if the loss satisfies the following Lipschitz condition: There exists $L > 0$ such that for every $f, g \in \mathcal{F}$ and $(x, y) \in \mathcal{X}\times\mathcal{Y}$

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

Here are some examples of Lipschitz (and convex) losses

- The **logistic loss** defined, for any $u \in \mathbb{R}$ and $y \in \mathcal{Y} = \{-1, 1\}$, by $\ell(u, y) = \log(1+\exp(-yu))$ is 1-Lipschitz.
- The **hinge loss** defined, for any $u \in \mathbb{R}$ and $y \in \mathcal{Y} = \{-1, 1\}$, by $\ell(u, y) = \max(1 - uy, 0)$ is 1-Lipschitz.
- The **Huber loss** defined, for any $\delta > 0$, $u, y \in \mathbb{R}$, by

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

  is δ-Lipschitz.
- The **quantile loss** defined, for any $\tau \in (0, 1)$, $u, y \in \mathbb{R}$, by $\ell(u, y) = \rho_\tau(u - y)$ where, for any $z \in \mathbb{R}$, $\rho_\tau(z) = z(\tau - I\{z \leq 0\})$ is 1-Lipschitz. For $\tau = 1/2$, the quantile loss is the absolute loss.

From the Lipschitz assumption, a concentration of $\{P_n(\ell_f - \ell_{f^*_{\mathcal{F}}}), f \in \mathcal{F}\}$ follows from a concentration of $\mathcal{F} - f^*_{\mathcal{F}}$ without assumptions on $(Y_i)_{i\in[\![1,n]\!]}$. The (R)ERM is robust with respect to heavy-tailed labels $(Y_i)_{i\in[\![1,n]\!]}$ and to outliers in the labels. This is a standard idea in the theory of robust M-estimators (Huber and Ronchetti, 2011). This intuition is developed in Chapters 2, 3, 4 and 5. We show that heavy-tailed target Y does not affect the performances of the (R)ERM when the loss function is simultaneously convex and Lipschitz. We also demonstrate that the (R)ERM exhibits the minimax-optimal rate when $\mathcal{O}$ outliers corrupt only the labels. Table 1.2 summarizes the robustness results achieved by the empirical risk minimizer and its regularized versions.

[5] Using the terminology of (Mendelson, 2014)

Robustness	Heavy-tailed $\mathcal{F} - f^*$	Heavy-tailed noise $Y - f^*_{\mathcal{F}}(X)$	Outliers $(X_o)_{\mathcal{O}}$	Outliers $(Y_o)_{\mathcal{O}}$
(R)ERM with Lipschitz loss				
(R)ERM with quadratic loss				

Table 1.2: Summary robust properties of the (R)ERM with different loss functions.

1.6.4 More advanced robust estimators

As presented in Section 1.6.3, the (R)ERM with unbounded loss is not robust to heavy-tailed classes of functions $\ell_{\mathcal{F}} - \ell_{f^*_{\mathcal{F}}}$ or to outliers contaminating the inputs $(X_i)_{i\in[\![1,n]\!]}$. Thus, one would like a systematic construction to get reliable estimators when the class $\mathcal{F} - f^*_{\mathcal{F}}$ may be heavy-tailed or when outliers corrupt the inputs $(X_i)_{i\in[\![1,n]\!]}$. Here, we present a simple construction based on the Median Of Means scheme. It relaxes the assumptions on the class $\mathcal{F} - f^*$ and the i.i.d assumption on the data $(X_i, Y_i)_{i\in[\![1,n]\!]}$.

The setting is the following. Let $\mathcal{I}$ and $\mathcal{O}$ be such than $\mathcal{I} + \mathcal{O} = n$ and $\mathcal{D}_{\mathcal{I}} \cup \mathcal{D}_{\mathcal{O}}$ is a partition of $\mathcal{D} = (X_i, Y_i)_{i\in[\![1,n]\!]}$ into two datasets, where $\mathcal{D}_{\mathcal{I}}$ is composed of $\mathcal{I}$ i.i.d informative data distributed as (X, Y) and $\mathcal{D}_{\mathcal{O}}$ is composed of $\mathcal{O}$ outliers for which nothing is assumed.
In Section 1.6.1, we presented the Median Of Means estimator of the mean. This estimator is robust to outliers and it achieves subgaussian deviations from the mean even if data only have two moments. Thus, one would like to apply a similar construction for supervized learning problems. Recall that the oracle $f^*_{\mathcal{F}}$ is defined as

$$f^*_{\mathcal{F}} \in \operatorname*{argmin}_{f\in\mathcal{F}} \mathbb{E}_{(X,Y)\sim P}\big[\ell\big(f(X), Y\big)\big] \ ,$$

where the distribution P is unknown. A first approach is to replace the expectation by the empirical mean and minimize this empirical risk. However, as presented in Section 1.6.3, when the class $\mathcal{F} - f^*_{\mathcal{F}}$ is heavy-tailed, there is no hope to prove deviations of this estimator as good as under subgaussian assumptions. Instead, we estimate the expectation with the (level-dependent) MOM estimator of the mean. Let $K \in [\![1, n]\!]$ assumed to divide n for simplicity. For any $k \in [\![1, K]\!]$ and $f \in \mathcal{F}$, let $P_{B_k}\ell_f = (K/n)\sum_{i\in B_k} \ell\big(f(X_i), Y_i\big)$. The MOM-estimator of the risk $P\ell_f$ is defined as $\mathrm{MOM}_K(f) = \mathrm{Median}\big(P_{B_1}\ell_f, \cdots, P_{B_K}\ell_f\big)$. Although being attractive, minimizing $\mathrm{MOM}_K(f)$ over $\mathcal{F}$ is not sufficient to obtain fast-rates of convergence. A simple explanation comes from the lack of linearity of such a procedure. Recall that the linearity of the empirical process $\{(P_n - P)(f)\}$ is important to use localisation techniques and derive "fast rates" of convergence for ERM (see Section 1.3). Comparing the minimizer of $\mathrm{MOM}_K(f)$ over $\mathcal{F}$ and $f^*_{\mathcal{F}}$ requires to compare differences of median, which does not concentrate as well as the median of the differences. To bypass this issue

of non linearity, the authors in (Lecué and Lerasle, 2019) noticed that the oracle $f_{\mathcal{F}}^*$ also verifies

$$f_{\mathcal{F}}^* \in \operatorname*{argmin}_{f\in\mathcal{F}} \sup_{g\in\mathcal{F}} \mathbb{E}_{(X,Y)\sim P}\big[\ell\big(f(X),Y\big) - \ell\big(g(X),Y\big)\big] \ .$$

Replacing the expectation by the MOM-estimator we finally get the minmax-MOM estimators and their penalized versions defined as

$$\hat{f}_K^{\text{MOM}} \operatorname*{argmin}_{f\in\mathcal{F}} \sup_{g\in\mathcal{F}} \text{Median}\big(P_{B_1}(\ell_f - \ell_g), \cdots, P_{B_K}(\ell_f - \ell_g)\big)$$

$$\hat{f}_{K,\lambda}^{\text{MOM}} \operatorname*{argmin}_{f\in\mathcal{F}} \sup_{g\in\mathcal{F}} \text{Median}\big(P_{B_1}(\ell_f - \ell_g), \cdots, P_{B_K}(\ell_f - \ell_g)\big) + \lambda\big(\Psi(f) - \Psi(g)\big) \ ,$$

where $\Psi : E \mapsto \mathbb{R}_+$ denotes a penalty function and $\lambda > 0$ is a tuning parameter. From these definitions, it is clear that both $\hat{f}_K^{\text{MOM}}$ and $\hat{f}_{K,\lambda}^{\text{MOM}}$ are robust to at least $K/2$ outliers. These outliers, can corrupt both the inputs and the outputs $(X_i, Y_i)_{i\in[\![1,n]\!]}$. Such estimators, have been studied in (Lecué and Lerasle, 2019) for the quadratic loss. The following theorem summarizes their principal results.

Theorem 1.8: MOM with quadratic loss (Lecué and Lerasle, 2019)

Let

$$r_Q = \inf\left\{r > 0 : \forall J \subset \mathcal{I}, |J| \geq \frac{2}{n} \quad \mathbb{E} \sup_{f\in\mathcal{F}:\|f-f_{\mathcal{F}}^*\|_{L_2(P_X)}\leq r} \Big|\sum_{i\in J} \sigma_i (f - f_{\mathcal{F}}^*)(X_i)\Big| \leq c|J|r\right\}$$

$$r_M = \inf\left\{r > 0 : \forall J \subset \mathcal{I}, |J| \geq \frac{2}{n} \quad \mathbb{E} \sup_{f\in\mathcal{F}:\|f-f_{\mathcal{F}}^*\|_{L_2(P_X)}\leq r} \Big|\sum_{i\in J} \sigma_i \xi_i (f - f_{\mathcal{F}}^*)(X_i)\Big| \leq c|J|r^2\right\} \ ,$$

where $(\sigma_i)_{i\in[\![1,n]\!]}$ are i.i.d Rademacher random variables independent of $(X_,Y_i)_{i\in[\![1,n]\!]}$ and $\xi_i = Y_i - f_{\mathcal{F}}^*(X_i)$. Let $K \in [\![1,n]\!]$ be such that $K \geq c\mathcal{O}$. With probability larger than $1 - \exp(cK)$, the estimator $\hat{f}_K^{\text{MOM}}$ satisfies

$$\|\hat{f}_K^{\text{MOM}} - f_{\mathcal{F}}^*\|_{L_2(P_X)}^2 \leq c \max\left(r_Q^2, r_M^2, \frac{K}{n}\right) \quad \text{and} \quad R(\hat{f}_K^{\text{MOM}}) - R(f_{\mathcal{F}}^*) \leq c \max\left(r_Q^2, r_M^2, \frac{K}{n}\right)$$

The rates of convergence of $\hat{f}_K^{\text{MOM}}$ depend on two fixed-point complexity parameters. From Theorems 1.6 and 1.7, when both the class $\mathcal{F} - f_{\mathcal{F}}^*$ and the noise $Y - f_{\mathcal{F}}^*(X)$ are subgaussian, the Rademacher complexities can be replaced by Gaussian complexities. Therefore, this results shows that minimax MOM-estimators achieve the performance of the ERM in the subgaussian case proved in (Lecué and Mendelson, 2013), even when up to cK outliers have corrupted the dataset. When the class $\mathcal{F} - f_{\mathcal{F}}^*$ and the noise $Y - f_{\mathcal{F}}^*(X)$ are heavy-tailed, the computation of Rademacher complexities may be involved (Mendelson, 2016, 2017). In particular, for heavy-tailed noise $Y - f_{\mathcal{F}}^*(X)$ the complexity parameter r_M may be very large which is not entirely satisfactory. To bypass this issue, it is possible to consider robust Lipschitz loss function. From the Lipschitz property, we remove the dependence to the noise. Consequently, the minmax-MOM estimators with Lipschitz loss are

robust to outliers and to heavy-tailed class of functions $\mathcal{F} - f^*_{\mathcal{F}}$. It is also possible to obtain similar results for regularized minmax-MOM estimators $\hat{f}^{\text{MOM}}_{K,\lambda}$ using the sparsity equation developed in Section 1.3.2.

Robustness	Heavy-tailed $\mathcal{F} - f^*$	Heavy-tailed noise $Y - f^*_{\mathcal{F}}(X)$	Outliers $(X_o)_{\mathcal{O}}$	Outliers $(Y_o)_{\mathcal{O}}$
Minmax MOM with Lipschitz loss				
Minmax MOM with quadratic loss				

Table 1.3: Summary robust properties of the minmax-MOM estimators different loss functions.

This general approach allows to construct reliable estimators when the data are corrupted and heavy-tailed. There exist other constructions. For example, in (Loh and Wainwright, 2015; Loh et al., 2017) the authors proposed robust generalized penalized M-estimators. They establish error rates for every stationary points around the oracle $f^*_{\mathcal{F}}$. However, they obtain results only with polynomial probability and for the linear model in $\mathbb{R}^p$. Another line of works investigated robust versions of the gradient descent, based on variants of the multivariate median-of-means technique (Alistarh et al., 2018; Chen et al., 2017; Yin et al., 2018; Prasad et al., 2018). The works (Audibert and Catoni, 2011; Brownlees et al., 2015; Holland and Ikeda, 2017) systematically use the Catoni's approach (Catoni, 2012) to construct robust estimators. Very recently, (Minsker and Mathieu, 2019) presented a "hybrid" approach between the Catoni's and the median-of-means estimators. Their theoretical analysis is accompanied by very encouraging simulations. However, as for the minmax-MOM estimators, there is still no available algorithm able to compute these estimators. It remains an open question.

1.7 Summary of the contributions

Some arguments presented so far are borrowed from Chapters 2 to 6. We provide here a short summary of the results obtained in this thesis chapter by chapter.

Chapter 2 We introduced the homogeneity Lemma 1.3 to study the empirical risk minimizer associated with Lipschitz and convex loss function when the class $\mathcal{F} - f^*_{\mathcal{F}}$ is assumed to be subgaussian. Under a *local Bernstein condition* we derived fast rates with no assumption on the noise $Y - f^*_{\mathcal{F}}(X)$. We also provide precise results to know when this new local Bernstein condition is verified. Very informly, if the distribution of the noise $Y - f^*_{\mathcal{F}}(X)$ puts some mass around 0, then

the local Bernstein condition is granted. This results show that the empirical risk minimizer is robust with respect to heavy-tailed noise $Y - f^*_{\mathcal{F}}(X)$. We obtain minimax-rate-optimal results when applying our main theorem to practical problems.

Theorem 1.9: Robustness of ERM to heavy tailed noise

Let ℓ be a convex and Lipschitz loss function. Let $(X_i, Y_i)_{i\in[\![1,n]\!]}$ be i.i.d random variable distributed as (X, Y). Let us assume that the class $\mathcal{F} - f^*_{\mathcal{F}}$ is a convex 1-subgaussian class, that the distribution of the noise $Y - f^*_{\mathcal{F}}(X)$ is symmetric and puts some mass around 0 (see chapter 2 for a more precise definition). Define

$$r^* = \inf\left\{ r > 0 : \quad \mathbb{E} \sup_{f\in\mathcal{F}:\|f-f^*_{\mathcal{F}}\|_{L_2(P_X)}\leq r} \Big| \sum_{i=1}^n \sigma_i (f - f^*_{\mathcal{F}})(X_i) \Big| \leq cnr^2 \right\}$$

Then with probability larger than $1 - \exp\big(-cn(r^*)^2\big)$, any minimizer of the empirical risk $\hat{f}_n$ verifies

$$\|\hat{f}_n - f^*_{\mathcal{F}}\|_2 \leq r^* \quad \text{and} \quad R(\hat{f}_n) - R(f^*_{\mathcal{F}}) \leq c(r^*)^2$$

Then, we study theoretical properties of minmax-MOM estimators associated with Lipschitz and convex loss function. Such estimators allow to relax the subgaussian assumption on the class $\mathcal{F} - f^*_{\mathcal{F}}$ and the i.i.d assumption.

Theorem 1.10: Minmax-MOM estimators with Lipschitz losses

Let $\mathcal{F}$ be a convex class of functions and ℓ be a Lipschitz loss function. Let $\mathcal{I}$ and $\mathcal{O}$ be such that $\mathcal{I} + \mathcal{O} = n$ and $\mathcal{D}_{\mathcal{I}} \cup \mathcal{D}_{\mathcal{O}}$ be a partition of $\mathcal{D} = (X_i, Y_i)_{i\in[\![1,n]\!]}$ into two datasets, where $\mathcal{D}_{\mathcal{I}}$ is composed of $\mathcal{I}$ i.i.d informative data distributed as (X, Y). Assume that the distribution of $Y - f^*_{\mathcal{F}}(X)$ is symmetric and puts some mass around 0 (as for the ERM) and that $K \geq c\mathcal{O}$. Then, with probability $1 - \exp(-cK)$, the minmax-MOM estimator $\hat{f}^{\text{MOM}}_K$ verifies

$$\|\hat{f}^{\text{MOM}}_K - f^*_{\mathcal{F}}\|^2_{L_2(P_X)} \leq c \max\left((r^*)^2, \frac{K}{n} \right) \quad \text{and} \quad R(\hat{f}^{\text{MOM}}_K) - R(f^*_{\mathcal{F}}) \leq c \max\left((r^*)^2, \frac{K}{n} \right) ,$$

where

$$r^* = \inf\left\{ r > 0 : \forall J \subset \mathcal{I}, |J| \geq \frac{2}{n} \quad \mathbb{E} \sup_{f\in\mathcal{F}:\|f-f^*_{\mathcal{F}}\|_{L_2(P_X)}\leq r} \Big| \sum_{i\in J} \sigma_i (f - f^*_{\mathcal{F}})(X_i) \Big| \leq c|J|r^2 \right\} .$$

Chapter 3 This chapter extends the results obtained in Chapter 2 to regularized problems. We study the estimators

$$\hat{f}^{\lambda}_n = \operatorname*{argmin}_{f\in\mathcal{F}} P_n \ell_f + \lambda\Psi(f) \quad \text{and} \quad \hat{f}^{\text{MOM}}_{K,\lambda} = \text{Median}\big(P_{B_1}(\ell_f - \ell_g), \cdots, P_{B_K}(\ell_f - \ell_g)\big) + \lambda\big(\Psi(f) - \Psi(g)\big) ,$$

when the penalization Ψ is a norm. We develop the analysis presented in Section 1.4 to derive general results for the RERM when the class $\mathcal{F} - f^*_{\mathcal{F}}$ is subgaussian and the loss is simultaneously convex and Lipschitz. We study the sparsity equation for many regularization norms and obtain minimax-rate-optimal results when applying our main theorems. As for the ERM, the RERM associated to Lipschitz loss function is robust with respect to the noise $Y - f^*_{\mathcal{F}}(X)$.
The regularized minmax MOM estimators are studied under the same setting as the RERM except that the i.i.d assumption is relaxed and that the class $\mathcal{F} - f^*_{\mathcal{F}}$ is no longer assumed to be subgaussian. Similar results are also obtained under refinement of the local Bernstein condition introduced in Chapter 2.

Chapter 4 This chapter considers RERM estimators when the regularization is not necessarily a norm. This setting covers important examples such as the elastic net regularization and the Ridge regularization. We derive complexity-dependent bounds i.e depending on $\Psi(f^*_{\mathcal{F}})$, in a setting close to the one studied in Chapter 3.
We also use the homogeneity Lemma 1.3 to show that the subgaussian assumption of $\mathcal{F} - f^*_{\mathcal{F}}$ is not always required. We present the example of Support Vector Machine (SVM) associated with a bounded kernel K. As explained in Section 1.4, the homogeneity Lemma reduces the proof to an upper bound of the empirical process on a bounded subspace of F. In this situation, Talagrand's inequality applies without assuming that the class is subgaussian. In particular, no assumption is necessary on the input $(X_i)_{i\in[\![1,n]\!]}$. We also generalize the results for regularized minmax-MOM when the penalization is not a norm

Chapter 5 This chapter focuses only the ERM and it regularization counterpart. We show that the (R)ERM is robust when $\mathcal{O}$ outliers may contaminate the labels. The main theorem of this chapter is the following.

Theorem 1.11: Robustness of (R)ERM to outliers in the labels

Let $\mathcal{I}$ and $\mathcal{O}$ be such that $\mathcal{I} + \mathcal{O} = n$. Let ℓ be a convex and Lipschitz loss function. Let $(X_i)_{i\in[\![1,n]\!]}$ be i.i.d random variable distributed as X and let $(Y_i)_{i\in[\![1,\mathcal{I}]\!]}$ be i.i.d random variables distributed as Y. Then with probability larger than $1 - \exp\big(-cn(r^*)^2\big)$, the minimizer of the (regularized) empirical risk $\hat{f}_n$ verifies

$$\|\hat{f}_n - f^*_{\mathcal{F}}\|_2 \le c\left(r^* + \frac{\mathcal{O}}{n}\right) \quad \text{and} \quad R(\hat{f}_n) - R(f^*_{\mathcal{F}}) \le c\left((r^*)^2 + \frac{\mathcal{O}}{n}\right)$$

where r^* denotes the error rate in a non-contaminated setting, that is when $(X_i, Y_i)_{i\in[\![1,n]\!]}$ are i.i.d random variables.

From Theorem 1.11, as long as less than nr^* outliers contaminate the labels, the performances of the (R)ERM remain unchanged. When r^* is minimax-rate-optimal, these bounds are also minimax-

rate-optmal in a setting where $\mathcal{O}$ outliers corrupt only the labels. Since in Chapters 2 and 3 we obtain minimax-rate-optimal for many regularized (or not) problems, we show that the (R)ERM is often minimax-rate-optimal when the class $\mathcal{F} - f_{\mathcal{F}}^*$ is subgaussian and $\mathcal{O}$ outliers corrupt the labels.

Chapter 6 Contrary to Chapters 2, 3, 4 and 5, this chapter does not focus on robustness. We study the linear model in the Gaussian setting when the dimension p may be much larger than the number of observations n. Let $(X_i, X_i)_{i\in[\![1,n]\!]}$ be i.i.d random variables distributed as (X, Y) verifying

$$Y = X^T\beta^* + \xi, \quad X \sim \mathcal{N}(0, \Sigma), \quad \xi \sim \mathcal{N}(0, \sigma^2) \ ,$$

for $\beta^* \in \mathbb{R}^p$. We study the interpolating estimator with minimum norm defined as

$$\hat{\beta}_n = \text{argmin}\{\|\beta\|_2 : X_i^T\beta = Y_i, i \in [\![1, n]\!]\} \ .$$

Using localization methods, we investigate the benign overtting phenomenon in the large deviation regime, that is when the bounds on the excess risk hold with probability $1-\exp(-cn)$. Localization with respect to the Euclidean norm allows to obtain fast rates $\mathcal{O}(1/n)$ when the signal-to-noise ratio is large enough. When the signal-to-noise ratio is too low, we also recover the optimal rates at a deviation level $1 - \exp(-cn)$, showing the optimality of our results in this setting.

Chapter 2

Robust Statistical learning with Lipschitz and convex loss functions

In this chapter, we obtain estimation and excess risk bounds for Empirical Risk Minimizers (ERM) and minmax Median-Of-Means (MOM) estimators based on loss functions that are both Lipschitz and convex. Results for the ERM are derived under weak assumptions on the outputs and sub-gaussian assumptions on the design as in (Alquier et al., 2019). The difference with (Alquier et al., 2019) is that the global Bernstein condition of this chapter is relaxed here into a local assumption. We also obtain estimation and excess risk bounds for minmax MOM estimators under similar assumptions on the output and only moment assumptions on the design. Moreover, the dataset may also contains outliers in both inputs and outputs variables without deteriorating the performance of the minmax MOM estimators.

Unlike alternatives based on MOM's principle (Lecué and Lerasle, 2019; Lugosi and Mendelson, 2016), the analysis of minmax MOM estimators is not based on the small ball assumption (SBA) of (Koltchinskii and Mendelson, 2015). In particular, the basic example of non parametric statistics where the learning class is the linear span of localized bases, that does not satisfy SBA (Saumard, 2018) can now be handled. Finally, minmax MOM estimators are analysed in a setting where the local Bernstein condition is also dropped out. It is shown to achieve excess risk bounds with exponentially large probability under minimal assumptions insuring only the existence of all objects.

2.1 Introduction

In this chapter, we study learning problems where the loss function is simultaneously Lipschitz and convex. This situation happens in classical examples such as quantile, Huber and L_1 regression or logistic and hinge classification (van de Geer, 2016). As the Lipschitz property allows to make only weak assumptions on the outputs, these losses have been quite popular in robust statistics (Huber and Ronchetti, 2011). Empirical risk minimizers (ERM) based on Lipschitz losses such as the Huber loss have received recently an important attention (Zhou et al., 2018; Elsener and van de Geer, 2018; Alquier et al., 2019).

Based on a dataset $\{(X_i, Y_i) : i = 1, \dots, N\}$ of points in $\mathcal{X} \times \mathcal{Y}$, a class F of functions and a risk function $R(\cdot)$ defined on F, the statistician want to estimate an oracle $f^* \in \operatorname{argmin}_{f \in F} R(f)$ or to predict an output Y at least as good as $f^*(X)$. The risk function $R(\cdot)$ is often defined as the expectation of a loss function $\ell : (f, x, y) \in F \times \mathcal{X} \times \mathcal{Y} \to \ell_f(x, y) \in \mathbb{R}$ with respect to the unknown distribution P of a random variable $(X, Y) \in \mathcal{X} \times \mathcal{Y}$: $R(f) = \mathbb{E}\ell_f(X, Y)$. Hereafter, the risk is assumed to have this form for a loss function ℓ such that, for any (f, x, y), $\ell_f(x, y) = \bar{\ell}(f(x), y)$, for some function $\bar{\ell} : \bar{\mathcal{Y}} \times \mathcal{Y} \to \mathbb{R}$, where the set $\bar{\mathcal{Y}}$ is a convex set containing all possible values of $f(x)$. The loss function ℓ is said Lipschitz and convex when the following assumption holds.

Assumption 2.1. *There exists $L > 0$ such that, for any $y \in \mathcal{Y}$, $\bar{\ell}(\cdot, y)$ is L-Lipschitz and convex.*

Many classical loss functions satisfy Assumption 2.1 and we recall some of them below.

- The **logistic loss** defined, for any $u \in \bar{\mathcal{Y}} = \mathbb{R}$ and $y \in \mathcal{Y} = \{-1, 1\}$, by $\bar{\ell}(u, y) = \log(1 + \exp(-yu))$ satisfies Assumption 2.1 with $L = 1$.

- The **hinge loss** defined, for any $u \in \bar{\mathcal{Y}} = \mathbb{R}$ and $y \in \mathcal{Y} = \{-1, 1\}$, by $\bar{\ell}(u, y) = \max(1 - uy, 0)$ satisfies Assumption 2.1 with $L = 1$.

- The **Huber loss** defined, for any $\delta > 0$, $u, y \in \mathcal{Y} = \bar{\mathcal{Y}} = \mathbb{R}$, by

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

  satisfies Assumption 2.1 with $L = \delta$.

- The **quantile loss** is defined, for any $\tau \in (0, 1)$, $u, y \in \mathcal{Y} = \bar{\mathcal{Y}} = \mathbb{R}$, by $\bar{\ell}(u, y) = \rho_\tau(u - y)$ where, for any $z \in \mathbb{R}$, $\rho_\tau(z) = z(\tau - I\{z \leq 0\})$. It satisfies Assumption 2.1 with $L = 1$. For $\tau = 1/2$, the quantile loss is the L_1 loss.

All along the paper, the following assumption is also granted.

Assumption 2.2. *The class F is convex.*

When (X, Y) and the data $((X_i, Y_i))_{i=1}^N$ are independent and identically distributed (i.i.d.), for any $f \in F$, the empirical risk $R_N(f) = (1/N)\sum_{i=1}^N \ell_f(X_i, Y_i)$ is a natural estimator of $R(f)$. The empirical risk minimizers (ERM) (Vapnik, 2000) obtained by minimizing $f \in F \to R_N(f)$ are expected to be close to the oracle f^*. This procedure and its regularized versions have been extensively studied in learning theory (Koltchinskii, 2011a). When the loss is both convex and Lipschitz, results have been obtained in practice (Bach et al., 2012; Bubeck, 2015) and theory (van de Geer, 2016). Risk bounds with exponential deviation inequalities for the ERM can be obtained under weak assumptions on the outputs Y, but stronger assumptions on the design X. Moreover, fast rates of convergence (Tsybakov, 2004) can only be obtained under margin type assumptions such as the Bernstein condition (Bartlett and Mendelson, 2006a; van de Geer, 2016).

The Lipschitz assumption and global Bernstein conditions (that hold over the entire F as in (Alquier et al., 2019)) imply boundedness in L_2-norm of the class F, see the discussion preceding Assumption 2.4 for details. This boundedness is not satisfied in linear regression with unbounded design so the results of (Alquier et al., 2019) don't apply to this basic example such as linear regression with a Gaussian design. To bypass this restriction, the global condition is relaxed into a "local" one as in (Elsener and van de Geer, 2018; van de Geer, 2016), see Assumption 2.4 below.

The main constraint in our results on ERM is the assumption on the design. This constraint can be relaxed by considering alternative estimators based on the "median-of-means" (MOM) principle of (Nemirovsky and Yudin, 1983; Birgé, 1984; Jerrum et al., 1986; Alon et al., 1999) and the minmax procedure of (Audibert and Catoni, 2011; Baraud et al., 2017). The resulting minmax MOM estimators have been introduced in (Lecué and Lerasle, 2019) for least-squares regression as an alternative to other MOM based procedures (Lugosi and Mendelson, 2016; Lugosi et al., 2019a; Lecué and Lerasle, 2017). In the case of convex and Lipschitz loss functions, these estimators satisfy the following properties 1) as the ERM, they are efficient under weak assumptions on the noise 2) they achieve optimal rates of convergence under weak stochastic assumptions on the design and 3) the rates are not downgraded by the presence of some outliers in the dataset.

These improvements of MOM estimators upon ERM are not surprising. For univariate mean estimation, rate optimal sub-Gaussian deviation bounds can be shown under minimal L_2 moment assumptions for MOM estimators (Devroye et al., 2016) while the empirical mean needs each data to have sub-Gaussian tails to achieve such bounds (Catoni, 2012). In least-squares regression, MOM-based estimators (Lugosi and Mendelson, 2016; Lugosi et al., 2019a; Lecué and Lerasle, 2017, 2019) inherit these properties, whereas the ERM has downgraded statistical properties under moment assumptions (see Proposition 1.5 in (Lecué and Mendelson, 2016)). Furthermore, MOM procedures are resistant to outliers: results hold in the "$\mathcal{O} \cup \mathcal{I}$" framework of (Lecué and Lerasle, 2017, 2019), where inliers or informative data (indexed by $\mathcal{I}$) only satisfy weak moments assumptions and the dataset may contain outliers (indexed by $\mathcal{O}$) on which no assumption is made, see Section 2.3. This robustness, that almost comes for free from a technical point of view is another important advantage

of MOM estimators compared to ERM in practice. Figure 2.1 [1] illustrates this fact, showing that statistical performance of the standard logistic regression are strongly affected by a single corrupted observation, while the minmax MOM estimator maintains good statistical performance even with 5% of corrupted data.

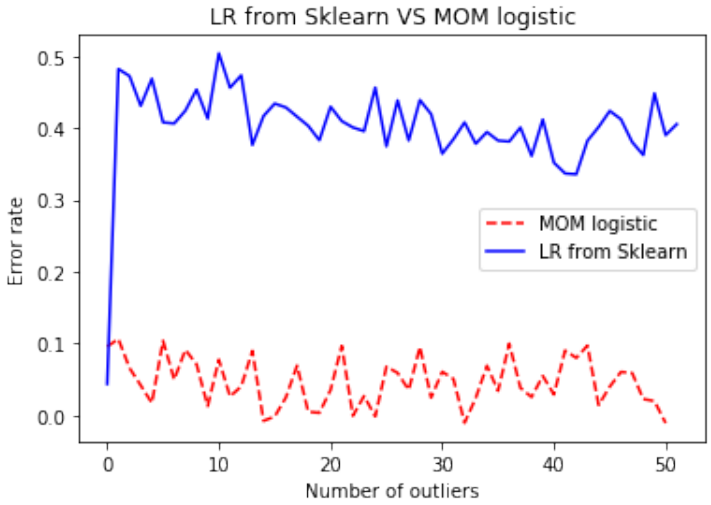


Figure 2.1: MOM Logistic Regression VS Logistic regression from Sklearn ($p = 50$ and $N = 1000$)

Compared to (Lugosi and Mendelson, 2016; Lecué and Lerasle, 2019), considering convex-Lipschitz losses instead of the square loss allows to simplify simultaneously some assumptions and the presentation of the results for MOM estimators: L_2-assumptions on the noise in (Lugosi and Mendelson, 2016; Lecué and Lerasle, 2019) can be removed and complexity parameters driving risk of ERM and MOM estimators only involve a single stochastic linear process, see Eq. (2.3) and (2.6) below. Also, contrary to the analysis in least-squares regression, the small ball assumption (Koltchinskii and Mendelson, 2015; Mendelson, 2014) is not required here. Recall that this assumption states that there are absolute constants κ and β such that, for all $f \in F$, $\mathbb{P}[|f(X) - f^*(X)| \geq \kappa \|f - f^*\|_{L_2}] \geq \beta$. It is interesting as it involves only moments of order 1 and 2 of the functions in F. However, it does not hold with absolute constants in classical frameworks such as histograms, see (Saumard, 2018; Han and Wellner, 2017) and Section 2.5.

Finally, minmax MOM estimators are studied in a framework where the Bernstein condition is dropped out. In this setting, they are shown to achieve an oracle inequality with exponentially large probability (see Section 2.4). The results are slightly weaker in this relaxed setting: the excess risk is bounded but not the L_2 risk and the rates of convergence are "slow" in $1/\sqrt{N}$ in general. Fast rates of convergence in $1/N$ can still be recovered from this general result if a local Bernstein type condition is satisfied though, see Section 2.4 for details. This last result shows that minmax MOM estimators can be safely used with Lipschitz and convex losses, assuming only that inliers data are independent with enough finite moments to give sense to the results. To approximate

[1] All figures can be reproduced from the code available at `https://github.com/lecueguillaume/MOMpower`

minmax MOM estimators, an algorithm inspired from (Lecué and Lerasle, 2019; Lecué et al., 2018) is also proposed. Asymptotic convergence of this algorithm has been proved in (Lecué et al., 2018) under strong assumptions, but, to the best of our knowledge, convergence rates have not been established. Nevertheless, the simulation study presented in Section 2.7 shows that it has good robustness performances.

The paper is organized as follows. Optimal results for the ERM are presented in Section 2.2. Minmax MOM estimators are introduced and analysed in Section 2.3 under a local Bernstein condition and in Section 2.4 without the Bernstein condition. A discussion of the main assumptions is provided in Section 2.5. Section 2.6 presents the theoretical limits of the ERM compared to the minmax MOM estimators. Finally, Section 2.7 provides a simulation study where a natural algorithm associated to the minmax MOM estimator for logistic loss is presented. The proofs of the main theorems are gathered in Sections 2.9, 2.10.1 and 2.10.2.

Notations Let $\mathcal{X}, \mathcal{Y}$ be measurable spaces and let $\bar{\mathcal{Y}}$ denote a convex set $\bar{\mathcal{Y}} \supset \mathcal{Y}$. Let F be a class of measurable functions $f : \mathcal{X} \to \bar{\mathcal{Y}}$ and let $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ be a random variable with distribution P. Let μ denote the marginal distribution of X. For any probability measure Q on $\mathcal{X} \times \mathcal{Y}$, and any function $g \in L_1(Q)$, let $Qg = \int g(x,y)\mathrm{d}Q(x,y)$. Let $\ell : F \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R},\ (f,x,y) \mapsto \ell_f(x,y)$ denote a loss function measuring the error made when predicting y by $f(x)$. It is always assumed that there exists a function $\overline{\ell} : \bar{\mathcal{Y}} \times \mathcal{Y} \to \mathbb{R}$ such that, for any $(f,x,y) \in F \times \mathcal{X} \times \mathcal{Y}$, $\overline{\ell}(f(x),y) = \ell_f(x,y)$. Let $R(f) = P\ell_f = \mathbb{E}\ell_f(X,Y)$ for f in F denote the risk and let $\mathcal{L}_f = \ell_f - \ell_{f^*}$ denote the excess loss. If $F \subset L_1(P) := L_1$ and Assumption 2.1 holds, an equivalent risk can be defined even if $Y \notin L_1$. Actually, for any $f_0 \in F$, $\ell_f - \ell_{f_0} \in L_1$ so one can define $R(f) = P(\ell_f - \ell_{f_0})$. W.l.o.g. the set of risk minimizers is assumed to be reduced to a singleton $\operatorname{argmin}_{f\in F} R(f) = \{f^*\}$. f^* is called the oracle as $f^*(X)$ provides the prediction of Y with minimal risk among functions in F. For any f and $p > 0$, let $\|f\|_{L_p} = (P|f|^p)^{1/p}$, for any $r \geqslant 0$, let $rB_{L_2} = \{f \in F : \|f\|_{L_2} \leq r\}$ and $rS_{L_2} = \{f \in F : \|f\|_{L_2} = r\}$. For any set H for which it makes sense, $H + f^* = \{h + f^* \text{ s.t } h \in H\}$, $H - f^* = \{h - f^* \text{ s.t } h \in H\}$. For any real numbers a, b, we write $a \lesssim b$ when there exists a positive constant c such that $a \leq cb$, when $a \lesssim b$ and $b \lesssim a$, we write $a \asymp b$.

2.2 ERM in the sub-Gaussian framework

This section studies the ERM, improving some results from (Alquier et al., 2019). In particular, the global Bernstein condition in (Alquier et al., 2019) is relaxed into a local hypothesis following (van de Geer, 2016). All along this section, data $(X_i, Y_i)_{i=1}^N$ are independent and identically distributed with common distribution P. The ERM is defined for $f \in F \to P_N\ell_f = (1/N)\sum_{i=1}^N \ell_f(X_i, Y_i)$ by

$$\hat{f}^{ERM} = \operatorname*{argmin}_{f\in F} P_N\ell_f \ . \tag{2.1}$$

The results for the ERM are shown under a sub-Gaussian assumption on the class $F - F$ with

respect to the distribution of X. This result is the benchmark for the following minmax MOM estimators.

Definition 2.1. *Let $B \geq 1$. F is called B-sub-Gaussian (with respect to X) when for all $f \in F$ and all $\lambda > 1$*

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

Assumption 2.3. *The class $F - F$ is B-sub-Gaussian with respect to X, where $F - F = \{f_1 - f_2 : f_1, f_2 \in F\}$.*

Under this sub-Gaussian assumption, statistical complexity can be measured via Gaussian mean-widths.

Definition 2.2. *Let $H \subset L_2$. Let $(G_h)_{h\in H}$ be the canonical centered Gaussian process indexed by H (in particular, the covariance structure of $(G_h)_{h\in H}$ is given by $(\mathbb{E}(G_{h_1} - G_{h_2})^2)^{1/2} = (\mathbb{E}(h_1(X) - h_2(X))^2)^{1/2}$ for all $h_1, h_2 \in H$). The **Gaussian mean-width** of H is $w(H) = \mathbb{E}\sup_{h\in H} G_h$.*

The complexity parameter driving the performance of $\hat{f}^{ERM}$ is presented in the following definition.

Definition 2.3. *The **complexity parameter** is defined as*

$$r_2(\theta) \geq \inf\{r > 0 : 32Lw((F - f^*) \cap rB_{L_2}) \leq \theta r^2\sqrt{N}\}$$

where $L > 0$ is the Lipschitz constant from Assumption 2.1.

Let $A > 0$. In (Bartlett and Mendelson, 2006a), the class F is called $(1, A)$-Bernstein if, for all $f \in F$, $P\mathcal{L}_f^2 \leqslant AP\mathcal{L}_f$. Under Assumption 2.1, F is $(1, AL^2)$-Bernstein if the following stronger assumption is satisfied

$$\|f - f^*\|_{L_2}^2 \leqslant AP\mathcal{L}_f \ . \tag{2.2}$$

This stronger version was used, for example in (Alquier et al., 2019) to study ERM. However, under Assumption 2.1, Eq (2.2) implies that

$$\|f - f^*\|_{L_2}^2 \leqslant AP\mathcal{L}_f \leqslant AL\|f - f^*\|_{L_1} \leqslant AL\|f - f^*\|_{L_2} \ .$$

Therefore, $\|f - f^*\|_{L_2} \leqslant AL$ for any $f \in F$. The class F is bounded in L^2-norm, which is restrictive as this assumption is not verified by the class of linear functions for example. To bypass this issue, the following condition is introduced.

Assumption 2.4. *There exists a constant $A > 0$ such that, for all $f \in F$ satisfying $\|f - f^*\|_{L_2} = r_2(1/(2A))$, we have $\|f - f^*\|_{L_2}^2 \leqslant AP\mathcal{L}_f$.*

In Assumption 2.4, Bernstein condition is granted in a L_2-sphere centered in f^* only. Outside of this sphere, there is no restriction on the excess loss. From the previous remark, it is clear that we necessarily have $r_2(1/(2A)) \leq AL$ (as long as there exists some $f \in F$ such that $\|f - f^*\|_2 \geq r_2(1/2A)$). This relaxed assumption is satisfied for many Lipschitz-convex loss functions under moment assumptions and weak assumptions on the noise as it will be checked in Section 2.5. The following theorem is the main result of this section.

Theorem 2.1. *Grant Assumptions 2.1, 2.2, 2.3 and 2.4, $\hat{f}^{ERM}$ defined in (2.1) satisfies, with probability larger than*

$$1 - 2\exp\left(-C\frac{Nr_2^2(1/(2A))}{(AL)^2}\right), \tag{2.3}$$

$$\|\hat{f}^{ERM} - f^*\|_{L_2}^2 \leq r_2^2(1/(2A)) \text{ and } P\mathcal{L}_{\hat{f}^{ERM}} \leq \frac{r_2^2(1/(2A))}{2A} \ , \tag{2.4}$$

where C is an absolute constant.

Theorem 2.1 is proved in Section 2.9.1. It shows deviation bounds both in L_2 norm and for the excess risk, which are both minimax optimal as proved in (Alquier et al., 2019). As in (Alquier et al., 2019), a similar result can be derived if the sub-Gaussian Assumption 2.3 is replaced by a boundedness in L_∞ assumption. An extension of Theorem 2.1 can be shown, where Assumption 2.4 is replaced by the following hypothesis: there exists κ such that for all $f \in F$ in a L_2-shpere centered in f^*, $\|f - f^*\|_{L_2}^{2\kappa} \leqslant AP\mathcal{L}_f$. The case $\kappa = 1$ is the most classical and its analysis contains all the ingredients for the study of the general case with any parameter $\kappa \geq 1$. More general Bernstein conditions can also be considered as in (van de Geer, 2016, Chapter 7). These extensions are left to the interested reader.

Notice that none of the assumptions 2.1, 2.2, 2.3 and 2.4 involve the output Y directly. All assumptions on Y are done through the oracle f^*. Yet, as will become transparent in the applications in Section 2.5, some assumptions on the distributions of Y are required to check the assumptions of Theorem 2.1. These assumptions are not very restrictive though and Lipschitz losses have been quite popular in robust statistics for this reason.

2.3 Minmax MOM estimators

This section presents and studies minmax MOM estimators, comparing them to ERM. We relax the sub-Gaussian assumption on the class $F - F$ and the i.i.d assumption on the data $(X_i, Y_i)_{i=1}^N$.

2.3.1 The estimators

The framework of this section is a relaxed version of the i.i.d. setup considered in Section 2.2. Following (Lecué and Lerasle, 2017, 2019), there exists a partition $\mathcal{O} \cup \mathcal{I}$ of $\{1, \cdots, N\}$ in two subsets unknown to the statistician. No assumption is granted on the set of "outliers" $(X_i, Y_i)_{i \in \mathcal{O}}$.

"Inliers", $(X_i, Y_i)_{i\in\mathcal{I}}$, are only assumed to satisfy the following assumption. For all $i \in \mathcal{I}$, (X_i, Y_i) has distribution P_i, X_i has distribution μ_i and for any $p > 0$ and any function g for which it makes sense $\|g\|_{L_p(\mu_i)} = (P_i|g|^p)^{1/p}$.

Assumption 2.5. *$(X_i, Y_i)_{i\in\mathcal{I}}$ are independent and, for any $i \in \mathcal{I}$, $\|f - f^*\|_{L_2} = \|f - f^*\|_{L_2(\mu_i)}$ and $P_i\mathcal{L}_f = P\mathcal{L}_f$.*

Assumption 2.5 holds in the i.i.d case but it covers other situations where informative data $(X_i, Y_i)_{i\in\mathcal{I}}$ may have different distributions. Typically, when F is the class of linear functions on $\mathbb{R}^d$, $F = \{\langle t, \cdot\rangle, t \in \mathbb{R}^d\}$ and $(X_i)_{i\in\mathcal{I}}$ are vectors with independent coordinates $(X_{i,j})_{j=1,\dots,d}$, then Assumption 2.5 is met if the coordinates $(X_{i,j})_{i\in\mathcal{I}}$ have the same first and second moments for all $j = 1, \dots, d$.

Recall the definition of MOM estimators of univariate means. Let $(B_k)_{k=1,\dots,K}$ denote a partition of $\{1, \dots, N\}$ into blocks B_k of equal size N/K (if N is not a multiple of K, just remove some data). For any function $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ and $k \in \{1, \dots, K\}$, let $P_{B_k} f = (K/N) \sum_{i\in B_k} f(X_i, Y_i)$. MOM estimator is the median of these empirical means:

$$\mathrm{MOM}_K\big(f\big) = \mathrm{Med}(P_{B_1} f, \cdots, P_{B_K} f) \ .$$

The estimator $\mathrm{MOM}_K\big(f\big)$ achieves rate optimal sub-Gaussian deviation bounds, assuming only that $Pf^2 < \infty$, see for example (Devroye et al., 2016). The number K is a tuning parameter. The larger K, the more outliers are allowed. When $K = 1$, $\mathrm{MOM}_K\big(f\big)$ is the empirical mean, when $K = N$, the empirical median.

Following (Lecué and Lerasle, 2019), remark that the oracle is also solution of the following minmax problem:

$$f^* \in \operatorname*{argmin}_{f\in F} P\ell_f = \operatorname*{argmin}_{f\in F} \sup_{g\in F} P(\ell_f - \ell_g) \ .$$

Minmax MOM estimators are obtained by plugging MOM estimators of the unknown expectations $P(\ell_f - \ell_g)$ in this formula:

$$\hat{f} \in \operatorname*{argmin}_{f\in F} \sup_{g\in F} \mathrm{MOM}_K\big(\ell_f - \ell_g\big) \ . \tag{2.5}$$

The minmax MOM construction can be applied systematically as an alternative to ERM. For instance, it yields a robust version of logistic classifiers. The minmax MOM estimator with $K = 1$ is the ERM.

The linearity of the empirical process P_N is important to use localisation technics and derive "fast rates" of convergence for ERM (Koltchinskii, 2011b), improving "slow rates" derived with the approach of (Vapnik, 1998), see (Tsybakov, 2004) for details on "fast and slow rates". The idea of the minmax reformulation comes from (Audibert and Catoni, 2011), where this strategy allows to overcome the lack of linearity of some alternative robust mean estimators. (Lecué and Lerasle, 2017) introduced minmax MOM estimators to least-squares regression.

2.3.2 Theoretical results

Setting

The assumptions required for the study of estimator (2.5) are essentially those of Section 2.2 except for Assumption 2.3 which is relaxed into Assumption 2.5. Instead of Gaussian mean width, the complexity parameter is expressed as a fixed point of local Rademacher complexities (Bartlett et al., 2005; Boucheron et al., 2005; Bartlett et al., 2005). Let $(\sigma_i)_{i=1,\dots,N}$ denote i.i.d. Rademacher random variables (uniformly distributed on $\{-1,1\}$), independent from $(X_i,Y_i)_{i\in\mathcal{I}}$. Let

$$\tilde{r}_2(\gamma) \geq \inf\left\{ r>0, \forall J\subset\mathcal{I} : |J| \geq \frac{N}{2}, \mathbb{E}\sup_{f\in F:\|f-f^*\|_{L_2}\leq r}\left|\sum_{i\in J}\sigma_i(f-f^*)(X_i)\right| \leq r^2|J|\gamma \right\} . \quad (2.6)$$

The outputs do not appear in the complexity parameter. This is an interesting feature of Lipschitz losses. It is necessary to adapt the Bernstein assumption to this framework.

Assumption 2.6. *There exists a constant $A>0$ such that for all $f\in F$ if $\|f-f^*\|_{L_2}^2 = C_{K,r}$ then $\|f-f^*\|_{L_2}^2 \leqslant AP\mathcal{L}_f$ where*

$$C_{K,r} = \max\left(\tilde{r}_2^2(1/(575AL)), 864A^2L^2\frac{K}{N}\right) . \quad (2.7)$$

Assumptions 2.6 and 2.4 have a similar flavor as both require the Bernstein condition in a L_2-sphere centered in f^* with radius given by the rate of convergence of the associated estimator (see Theorems 2.1 and 2.2). For $K \leq (\tilde{r}_2^2(575/(AL))N)/(846A^2L^2)$ the sphere $\{f\in F : ||f-f^*||_{L_2} = \sqrt{C_{K,r}}\}$ is a L_2-sphere centered in f^* of radius $\tilde{r}_2(575/(AL))$ which can be of order $1/\sqrt{N}$ (see Section 2.3.2). As a consequence, Assumption 2.6 holds in examples where the small ball assumption does not (see discussion after Assumption 2.9).

Main results

We are now in position to state the main result regarding the statistical properties of estimator (2.5) under a local Bernstein condition.

Theorem 2.2. *Grant Assumptions 2.1, 2.2, 2.5 and 2.6 and assume that $|\mathcal{O}| \leq 3N/7$. Let $\gamma = 1/(575AL)$ and $K \in \big[7|\mathcal{O}|/3, N\big]$. The minmax MOM estimator $\hat{f}$ defined in (2.5) satisfies, with probability at least*

$$1-\exp(-K/2016), \quad (2.8)$$

$$\|\hat{f}-f^*\|_{L_2}^2 \leq C_{K,r} \text{ and } P\mathcal{L}_{\hat{f}} \leq \frac{2}{3A}C_{K,r} . \quad (2.9)$$

Suppose that $K=\tilde{r}_2^2(\gamma)N$, which is possible as long as $|\mathcal{O}| \lesssim N\tilde{r}_2^2(\gamma)$. The deviation bound is then of order $\tilde{r}_2^2(\gamma)$ and the probability estimate $1-\exp(-N\tilde{r}_2^2(\gamma)/2016)$. Therefore, minmax MOM estimators achieve the same statistical bound with the same deviation as the ERM as long as $\tilde{r}_2^2(\gamma)$ and $r_2(\theta)$ are of the same order. Using generic chaining (Talagrand, 2014), this comparison is true

under Assumption 2.3. It can also be shown under weaker moment assumption, see (Mendelson, 2017) or the example of Section 2.3.2.

When $\tilde{r}_2^2(\gamma) \asymp r_2(\theta)$, the bounds are rate optimal as shown in (Alquier et al., 2019). This is why these bounds are called rate optimal sub-Gaussian deviation bounds. While these hold for ERM in the i.i.d. setup with sub-Gaussian design in the absence of outliers (see Theorem 2.1), they hold for minmax MOM estimators in a setup where inliers may not be i.i.d., nor have sub-Gaussian design and up to $N\tilde{r}_2^2(\gamma)$ outliers may have contaminated the dataset.

This section is concluded by presenting an estimator achieving (2.5) simultaneously for all K. For all $K \in \{1, \dots, N\}$ and $f \in F$, define $T_K(f) = \sup_{g \in F} \mathrm{MOM}_K\big(\ell_f - \ell_g\big)$ and let

$$\hat{R}_K = \{g \in F : T_K(g) \leq (1/3A)C_{K,r}\} \quad . \tag{2.10}$$

Now, building on the Lepskii's method, define a data-driven number of blocks

$$\hat{K} = \inf\left(K \in \{1, \dots, N\} : \bigcap_{J=K}^{N} \hat{R}_J \neq \emptyset\right) \tag{2.11}$$

and let $\tilde{f}$ be such that

$$\tilde{f} \in \bigcap_{J=\hat{K}}^{N} \hat{R}_J \quad . \tag{2.12}$$

Theorem 2.3. *Grant Assumptions 2.1, 2.2, 2.5 and 2.6 and assume that $|\mathcal{O}| \leq 3N/7$. Let $\gamma = 1/(575AL)$. The estimator $\tilde{f}$ defined in* (2.12) *is such that for all $K \in \big[7|\mathcal{O}|/3, N\big]$, with probability at least*

$$1 - 4\exp(-K/2016),$$

$$\|\tilde{f} - f^*\|_{L_2}^2 \leq C_{K,r} \text{ and } P\mathcal{L}_{\tilde{f}} \leq \frac{2}{3A}C_{K,r} \quad .$$

Theorem 2.3 states that $\tilde{f}$ achieves the results of Theorem 2.2 simultaneously for all $K \geq 7|\mathcal{O}|/3$. This extension is useful as the number $|\mathcal{O}|$ of outliers is typically unknown in practice. However, contrary to $\hat{f}$, the estimator $\tilde{f}$ requires the knowledge of A and $\tilde{r}(\gamma)$. These parameters allow to build confidence regions for f^*, which is necessary to apply Lepski's method. Similar limitations appear in least-squares regression (Lecué and Lerasle, 2019) and even in the basic problem of univariate mean estimation. In this simpler problem, it can be shown that one can build sub-Gaussian estimators depending on the confidence level (through K) under only a second moment assumption. On the other hand, to build estimators achieving the same risk bounds simultaneously for all K, more informations on the distribution of the data are required, see (Devroye et al., 2016, Theorem 3.2). In particular, the knowledge of the variance, which allows to build confidence intervals for the unknown univariate mean, is sufficient. The necessity of extra-information to obtain adaptivity with respect to K is therefore not surprising here.

Some basic examples

The following example illustrates the optimality of the rates provided in Theorem 2.2 even under a simple L_2-moment assumption.

Lemma 2.1 ((Koltchinskii, 2006)). *In the $\mathcal{O} \cup \mathcal{I}$ framework with $F = \{\langle t, \cdot \rangle : t \in \mathbb{R}^d\}$, we have $\tilde{r}_2^2(\gamma) \leq Rank(\Sigma)/(2\gamma^2 N)$, where $\Sigma = \mathbb{E}[XX^T]$ is the $d \times d$ covariance matrix of X.*

The proof of Lemma 2.1 is recalled in Section 2.10.1 for the sake of completeness. Lemma 2.1 grants only the existence of a second moment for X even though the rate obtained $\mathrm{Rank}(\Sigma)/(2\gamma^2 N)$ is the same as the one we would get under a sub-Gaussian assumption given that $r_2(\theta) \sim \mathrm{Rank}(\Sigma)/(2\theta^2 N)$. Moreover, Section 2.5 shows that Assumptions 2.4 and 2.6 are satisfied when $F = \{\langle t, \cdot \rangle : t \in \mathbb{R}^d\}$ and X is a vector with i.i.d. entries having only a few finite moments. Theorem 2.2 applies therefore in this setting and the Minmax MOM estimator (2.5) achieves the optimal fast rate of convergence $\mathrm{Rank}(\Sigma)/N$. This shows that when the model is the entire space $\mathbb{R}^d$, the results for the ERM from Theorem 2.1 obtained under a sub-Gaussian assumption is the same as the one for the minmax MOM from Theorem 2.2 under only weak moment assumption.

However, Lemma 2.1 does not describe a typical situation. Having $\tilde{r}_2^2(\gamma)$ of the same order as $r_2^2(\theta)$ under only a second moment assumption is mainly happening on large models such as the entire space $\mathbb{R}^d$. For smaller size models such as the B_1^d-ball (the unit ball of the ℓ_1^d-norm), the picture is different: $\tilde{r}_2^2(\gamma)$ should be bigger than $r_2^2(\theta)$ unless X has enough moment. To make this statement simple let us consider the case $N = 1$. In that case, we have $r_2^2(\theta) \sim \sqrt{\log d}$. Let us now describe $\tilde{r}_2^2(\gamma)$ under various moment assumptions on X to see when $\tilde{r}_2^2(\gamma)$ compares with $\sqrt{\log d}$.

Let $X = (x_j)_{j=1}^d$ be a random vector. It follows from Equation (3.1) in (Mendelson et al., 2007) that

$$\mathbb{E} \sup_{t \in B_1^d \cap r B_2^d} |\langle t, X \rangle| \lesssim \begin{cases} r\mathbb{E}\left(\sum_{j=1}^d x_j^2\right)^{1/2} & \text{if } r \leq 1/\sqrt{d} \\ r\mathbb{E}\left(\sum_{j=1}^k (x_j^*)^2\right)^{1/2} & \text{if } 1/\sqrt{d} \leq r \leq 1 \\ \mathbb{E}\max_{j=1,\dots,d} |x_j| & \text{if } r \geq 1. \end{cases}$$

where $k = \lceil 1/r^2 \rceil$ and $x_1^* \geq \ldots \geq x_d^*$ is a non-increasing rearrangement of the absolute values of the coordinates $x_j, j = 1, \ldots, d$ of X. Assume that $x_1, \ldots, x_d$ are i.i.d. distributed like x such that $\mathbb{E}x = 0$ and $\mathbb{E}x^2 = 1$. Assume that x has $2p$ moments, for $p \geq 1$ and let c_0 be such that $\mathbb{E}[x^{2p}] \leq c_0$. Then, using Jensen's inequality, we obtain

$$\left(\mathbb{E}\left[\frac{1}{k}\sum_{j=1}^k (x_j^*)^2\right]\right)^p \leq \frac{1}{k}\sum_{j=1}^k \mathbb{E}[(x_j^*)^{2p}] \leq \frac{1}{k}\sum_{j=1}^d \mathbb{E}[x_j^{2p}] \leq \frac{c_0 d}{k}.$$

It follows that

$$\mathbb{E}\left[\left(\frac{1}{k}\sum_{j=1}^k (x_j^*)^2\right)^{1/2}\right] \leq \left(\mathbb{E}\left[\frac{1}{k}\sum_{j=1}^k (x_j^*)^2\right]\right)^{1/2} \leq \left(\frac{c_0 d}{k}\right)^{1/(2p)} \leq \left(\frac{c_0 d}{k}\right)^{1/(2p)} \leq (c_0 d r^2)^{1/(2p)} \ .$$

Hence,

$$\mathbb{E}\sup_{t\in B_1^d\cap rB_2^d}|\langle t,X\rangle| \lesssim \begin{cases} r\sqrt{d} & \text{if } r\le 1/\sqrt{d} \\ (c_0dr^2)^{1/(2p)} & \text{if } 1/\sqrt{d}\le r\le 1 \\ (c_0d)^{1/(2p)} & \text{if } r\ge 1. \end{cases} .$$

Assume that $f^*=\langle t^*,\cdot\rangle$, with $t^*=0$. Then,

$$\tilde r_2(\gamma)=\inf\left\{r>0:\mathbb{E}\sup_{t\in B_1^d\cap rB_2^d}|\langle t,X\rangle|\le\gamma r^2\right\}\lesssim (c_0d)^{1/(4p)}/\sqrt{\gamma}\ .$$

In particular,

$$\begin{cases} \tilde r_2(\gamma)\asymp 1 & \text{when } p\ge\log(c_0d), \\ 1\lesssim r_2(\gamma)\lesssim\sqrt{\log d} & \text{when } \log(c_0d)/\log\log d\le p\le\log(c_0d), \\ \tilde r_2(\gamma)\asymp d^{1/(4p)} & \text{when } p\le\log(c_0d)/\log\log d. \end{cases}$$

Let us now show that these estimates are sharp by considering $x=\epsilon(1+R\eta)$ where ϵ is a Rademacher variable, η is a Bernouilli variable (independent of ϵ) with mean $\delta=1/d$ and $R=d^{1/(4p)}$. We have $\mathbb{E}x=0$ and $\mathbb{E}x^2=1+R\delta\le 2$ because $R\delta\le 1$ when $p\ge 1$. Let $x_j=\epsilon_j(1+R\eta_j), j=1,\dots,d$ be i.i.d. copies of x. We have

$$\begin{aligned}&\mathbb{E}\max_{j=1,\dots,d}|x_j|\ge(1+R)\mathbb{P}\left[\max_{j=1,\dots,d}|x_j|\ge 1+R\right]\\ &=(1+R)\left(1-\mathbb{P}[\eta_j=0,\forall j=1,\dots,d]\right)=(1+R)(1-(1-\delta)^d)\ge(1+R)e^{-1}\gtrsim d^{1/(4p)}.\end{aligned}$$

As a consequence, for all $1\le r\lesssim d^{1/(8p)}$,

$$\mathbb{E}\sup_{t\in B_1^d\cap rB_2^d}|\langle t,X\rangle|=\mathbb{E}\max_{j=1,\dots,d}|x_j|>r^2$$

and so $\tilde r_2^2(\gamma)\gtrsim d^{1/(4p)}$. As a consequence, under only a L_{2p} moment assumption one cannot have $\tilde r_2^2(\gamma)$ better than $d^{1/(4p)}$.

As a consequence, $\tilde r_2(\gamma)$ can be much larger than $r_2(\theta)$ when x has less than $\log(c_0d)/\log(\log d)$ moments, for instance, $\tilde r_2^2(\gamma)$ can be of the order of $d^{1/8}$ when x has only 2 moments. This picture is different from the one given by Lemma 2.1 where we were able to get equivalence between $\tilde r_2(\gamma)$ and $r_2(\theta)$ only under a second moment assumption.

2.4 Relaxing the Bernstein condition

This section shows that minmax MOM estimators satisfy sharp oracle inequalities with exponentially large deviation under minimal stochastic assumptions insuring the existence of all objects. These results are slightly weaker than those of the previous section: the L_2 risk is not controlled and only slow rates of convergence hold in this relaxed setting. However, the bounds are sufficiently

precise to imply fast rates of convergence for the excess risk as in Theorems 2.2 if a slightly stronger Bernstein condition holds.

Given that data may not have the same distribution as (X, Y), the following relaxed version of Assumption 2.5 is introduced.

Assumption 2.7. *$(X_i, Y_i)_{i\in\mathcal{I}}$ are independent and for all $i \in \mathcal{I}$, (X_i, Y_i) has distribution P_i, X_i has distribution μ_i. For any $i \in \mathcal{I}$, $F \subset L_2(\mu_i)$ and $P_i\mathcal{L}_f = P\mathcal{L}_f$ for all $f \in F$.*

When Assumption 2.6 does not necessary hold, the localization argument has to be modified. Instead of the L_2-norm, the excess risk $f \in F \to P\mathcal{L}_f$ is used to define neighborhoods around f^*. The associated complexity is then defined for all $\gamma > 0$ and $K \in \{1, \cdots, N\}$ by

$$\bar{r}_2(\gamma) \geq \inf\left\{ r > 0 : \max\Big(\frac{E(r)}{\gamma}, \sqrt{1536}V_K(r)\Big) \leq r^2 \right\} \tag{2.13}$$

where

$$E(r) = \sup_{J\subset\mathcal{I}:|J|\geq N/2} \mathbb{E} \sup_{f\in F: P\mathcal{L}_f\leq r^2} \left| \frac{1}{|J|}\sum_{i\in J} \sigma_i (f - f^*)(X_i) \right| ,$$

$$\text{and } V_K(r) = \max_{i\in\mathcal{I}} \sup_{f\in F: P\mathcal{L}_f\leq r^2} \sqrt{\mathbb{V}ar_{P_i}(\mathcal{L}_f)}\sqrt{\frac{K}{N}} .$$

There are two important differences between $\bar{r}_2(\gamma)$ on one side and $r_2(\theta)$ in Definition 2.2 or $\tilde{r}_2(\gamma)$ in (2.6) on the other side. The first one is the extra variance term $V_K(r)$. Under the Bernstein condition, this term is negligible in front of the "expectation term" $E(r)$ see (Bartlett and Mendelson, 2006a). In the general setting considered here, the variance term is handled in the complexity parameter. The second important consequence is that $\bar{r}_2$ is a fixed point of the complexity of F localized around f^* *with respect to the excess risk* rather than with respect to the L_2-norm. An important consequence is that this quantity is harder to compute in practical examples. As a consequence, the results of this section are more of theoretical importance.

Theorem 2.4. *Grant Assumptions 2.1, 2.2, 2.7 and assume that $|\mathcal{O}| \leq 3N/7$. Let $\gamma = 1/(768L)$ and $K \in \big[7|\mathcal{O}|/3, N\big]$. The minmax MOM estimator $\hat{f}$ defined in* (2.5) *satisfies, with probability at least $1 - \exp(-K/2016)$, $P\mathcal{L}_{\hat{f}} \leq \bar{r}_2^2(\gamma)$.*

Recall that Assumptions 2.1 and 2.2 are only meaning that the loss function is convex and Lipschitz and that the class F is convex. Assumption 2.7 says that inliers are independent and define the same excess risk as (X, Y) over F. In particular, Theorem 2.4 holds, as Theorem 2.2, without assumptions on the outliers $(X_i, Y_i)_{i\in\mathcal{O}}$ and with weak assumptions on the outputs $(Y_i)_{i\in\mathcal{I}}$ of the inliers (we remrak that excess loss function $f \to P\mathcal{L}_f$ is well-defined under no assumption on Y – even if $Y \notin L_1$ – because $|\mathcal{L}_f| \leq L|f - f^*|$). Moreover, the excess risk bound holds with exponentially large probability without assuming sub-Gaussian design, a small ball hypothesis

or a Bernstein condition. This generality can be achieved by combining MOM estimators with convex-Lipschitz loss functions.

The following result discuss relationships between Theorems 2.2 and 2.4. Introduce the following modification of the Bernstein condition.

Assumption 2.8. *Let $\gamma = 1/(768L)$. There exists a constant $A > 0$ such that for all $f \in F$ if $P\mathcal{L}_f = C'_{K,r}$ then $\|f - f^*\|_{L_2}^2 \leqslant AP\mathcal{L}_f$ where, for $\tilde{r}_2(\gamma)$ defined in* (2.6)*,*

$$C'_{K,r} = \max\left(\frac{\tilde{r}_2^2(\gamma/A)}{A}, \frac{1536AL^2K}{N}\right) .$$

Assumption 2.8 is slightly stronger than Assumption 2.4 since the L_2-metric to define the sphere is replaced by the excess risk metric. If Assumption 2.8 holds then Theorem 2.5 implies the same statistical bounds for (2.5) as Theorem 2.2 up to constants, as shown by the following result.

Theorem 2.5. *Grant Assumptions 2.1, 2.2, 2.7 and assume that $|\mathcal{O}| \leq 3N/7$. Assume that the local Bernstein condition Assumption 2.8 holds. Let $\gamma = 1/(768L)$ and $K \in \big[7|\mathcal{O}|/3, N\big]$. The minmax MOM estimator $\hat{f}$ defined in* (2.5) *satisfies, with probability at least $1 - \exp(-K/2016)$,*

$$\left\|\hat{f} - f^*\right\|_{L_2}^2 \leq \max\left(\tilde{r}_2^2(\gamma/A), \frac{1536L^2A^2K}{N}\right) \text{ and } P\mathcal{L}_{\hat{f}} \leq \max\left(\frac{\tilde{r}_2^2(\gamma/A)}{A}, \frac{1536L^2AK}{N}\right).$$

Proof. First, $V_K(r) \leq LV'_K(r)$ for all $r > 0$ where $V'_K(r) = \sqrt{K/N}\max_{i\in\mathcal{I}}\sup_{f\in F: P\mathcal{L}_f \leq r^2}\|f - f^*\|_{L_2(\mu_i)}$. Moreover, $r \to E(r)/r^2$ and $r \to V'_K(r)/r^2$ are non-increasing, therefore by Assumption 2.8 and the definition of $\tilde{r}_2(\gamma)$, $V'_K(r)$,

$$\frac{1}{\gamma}E\left(\frac{\tilde{r}_2(\gamma/A)}{\sqrt{A}}\right) \leq \frac{\tilde{r}_2^2(\gamma/A)}{A} \quad \text{and} \quad \sqrt{1536}V'_K\left(\sqrt{1536}L\sqrt{\frac{AK}{N}}\right) \leq \frac{1536A^2LK}{N} .$$

Hence, $\bar{r}_2^2(\gamma) \leq \max(\tilde{r}_2^2(\gamma/A)/A, 1536L^2A(K/N))$. ∎

2.5 Bernstein's assumption

This section shows that the local Bernstein condition holds for various loss functions and design X. In Assumption 2.4 and 2.6, the comparizon between $P\mathcal{L}_f$ and $\|f - f^*\|_{L_2}^2$ is only required on a L_2-sphere. In this section, we prove that the local Bernstein assumption can be verified over the entire L_2-ball and not only on the sphere under mild moment conditions. The class $F - \{f^*\}$ is assumed to satisfy a " $L_{2+\varepsilon}/L_2$-norm equivalence assumption", for $\varepsilon > 0$.

Assumption 2.9. *Let $\varepsilon > 0$. There exists $C' > 0$ such that for all $f \in F$, $\|f - f^*\|_{L_{2+\varepsilon}} \leq C'\|f - f^*\|_{L_2}$*

Assumption 2.9 is a "$L_{2+\varepsilon}/L_2$" norm equivalence assumption over $F - \{f^*\}$. A "L_4/L_2" norm equivalence assumption over $F - \{f^*\}$ has been used for the study of MOM estimators (see (Lugosi and Mendelson, 2016)). Examples of distributions satisfying Assumption 2.9 can be found in (Mendelson, 2014, 2015).

There are situations where the constant C' depends on the dimension d of the model. In that case, the results in (Lugosi and Mendelson, 2016; Lecué and Lerasle, 2019) provide sub-optimal statistical upper bounds. For instance, if X is uniformly distributed on $[0,1]$ and $F = \{\sum_{j=1}^d \alpha_j I_{A_j} : (\alpha_j)_{j=1}^p \in \mathbb{R}^d\}$ where I_{A_j} is the indicator of $A_j = [(j-1)/d, j/d]$ then for all $f \in F$, $\|f - f^*\|_{L_{2+\varepsilon}} \leq d^{\varepsilon/(4+2\varepsilon)} \|f - f^*\|_{L_2}$ so $C' = d^{\varepsilon/(4+2\varepsilon)}$. This dependence with respect to the dimension d is inevitable. For instance, in (Lugosi and Mendelson, 2016; Lecué and Lerasle, 2019), a L_4/L_2 norm equivalence is required. In this case, $C' = d^{1/4}$ which ultimately yields sub-optimal rates in this example. On the other hand, as will become clear in this section, the rates given in Theorem 2.2 or Theorem 2.3 are not deteriorated in this example. This improvement is possible since the Bernstein condition is only required in a neighborhood of f^*.

2.5.1 Quantile loss

The proof is based on (Elsener and van de Geer, 2018, Lemma 2.2) and is postponed to Section 2.10.2. Recall that $\ell_f(x,y) = (y - f(x))(\tau - I\{y - f(x) \leq 0\})$.

Assumption 2.10. *Let C' be the constant defined in Assumption 2.9. There exist $\alpha > 0$ and $r > 0$ such that, for all $x \in \mathcal{X}$ and for all z in $\mathbb{R}$ such that $|z - f^*(x)| \leq r(\sqrt{2}C')^{(2+\varepsilon)/\varepsilon}$, we have $f_{Y|X=x}(z) \geq \alpha$, where $f_{Y|X=x}$ is the conditional density function of Y given $X = x$.*

Theorem 2.6. *Grant Assumptions 2.9 (with constant C') and 2.10 (with parameter r and α). Then, for all $f \in F$ satisfying $\|f - f^*\|_{L_2} \leqslant r$, $\|f - f^*\|_{L_2}^2 \leqslant (4/\alpha) P\mathcal{L}_f$.*

Consider the example from Section 2.3.2, assume that $K \lesssim \mathrm{Rank}(\Sigma)$ and let $r^2 = C_{K,r} = \tilde{r}_2^2(\gamma) \leq \mathrm{Rank}(\Sigma)/(2\gamma^2 N)$. If $C' = d^{\varepsilon/(4+2\varepsilon)}$, Assumption 2.10 holds for r and an associated $\alpha \gtrsim 1$ as long as $d^{1/2}\sqrt{\mathrm{Rank}(\Sigma)/N} \lesssim 1$ and, for all $x \in \mathcal{X}$ and for all z in $\mathbb{R}$ such that $|z - f^*(x)| \lesssim 1$, $f_{Y|X=x}(z) \gtrsim 1$. As $\mathrm{Rank}(\Sigma) \leqslant d$, the first condition reduces to $N \gtrsim d^2$. In this situation, the rates given in Theorems 2.2 and 2.3 are still $\mathrm{Rank}(\Sigma)/N$. This gives a partial answer, in our setting, to the issue raised in (Saumard, 2018) regarding results based on the small ball method.

2.5.2 Huber Loss

Consider the Huber loss function defined, for all $f \in F$, $x \in \mathcal{X}$ and $y \in \mathbb{R}$, by $\ell_f(x,y) = \rho_H(y - f(x))$ where $\rho_H(t) = t^2/2$ if $|t| \leq \delta$ and $\rho_H(t) = \delta|t| - \delta^2/2$ otherwise. Introduce the following assumption.

Assumption 2.11. *Let C' be the constant defined in Assumption 2.9. There exist $\alpha > 0$ and $r > 0$ such that for all $x \in \mathcal{X}$ and all z in $\mathbb{R}$ such that $|z - f^*(x)| \leq (\sqrt{2}C')^{(2+\varepsilon)/\varepsilon} r$, $F_{Y|X=x}(z + \delta) - F_{Y|X=x}(z - \delta) \geqslant \alpha$, where $F_{Y|X=x}$ is the conditional cumulative function of Y given $X = x$.*

Under this assumption and a "$L_{2+\epsilon}/L_2$" assumption, the local Bernstein condition is proved to be satisfied in the following result whose proof is postponed to Section 2.10.2.

Theorem 2.7. *Grant Assumptions 2.9 (with constant C') and 2.11 (with parameter r and α). Then, for all $f \in F$ satisfying $\|f - f^*\|_{L_2} \leqslant r$, there exists $\alpha > 0$ (given by Assumption 2.11) such that $\|f - f^*\|_{L_2}^2 \leqslant (4/\alpha) P\mathcal{L}_f$.*

2.5.3 Logistic classification

In this section we consider the logistic loss function.

Assumption 2.12. *There exists $c_0 > 0$ such that*

$$\mathbb{P}\big(|f^*(X)| \leq c_0\big) \geq 1 - \frac{1}{(2C')^{(4+2\varepsilon)/\varepsilon}}.$$

where C' is defined in Assumption 2.9.

The following result is proved in Section 2.10.2.

Theorem 2.8. *Grant Assumptions 2.9 and 2.12. Then, for all $r > 0$ and all $f \in F$ such that $\|f - f^*\|_{L_2} \leqslant r$,*

$$P\mathcal{L}_f \geqslant \frac{e^{-c_0 - r(2C')^{(2+\epsilon)/\epsilon}}}{2\big(1 + e^{c_0 + r(2C')^{(2+\epsilon)/\epsilon}}\big)^2} \|f - f^*\|_{L_2}^2 \ .$$

The proof is postponed to Section 2.10.2. As for the Huber Loss and the Hinge Loss, the rates of convergence are not deteriorated when C' may depend on the dimension as long as $r \times (C')^{(2+\varepsilon)/\varepsilon}$ is smaller than some absolute constant.

2.5.4 Hinge loss

In this section, we show that the local Bernstein condition holds for various design X for the Hinge loss function. We obtain the result under the assumption that the oracle f^* is actually the Bayes rules which is the function minimizing the risk $f \mapsto R(f)$ over all measurable functions from $\mathcal{X}$ to $\mathbb{R}$. Recall that, under this assumption, $f^*(x) = \text{sign}(2\eta(x) - 1)$ where $\eta(X) = \mathbb{P}(Y = 1|X)$. In that case, the Bernstein condition (see (Bartlett and Mendelson, 2006a)) coincides with the margin assumption (see (Tsybakov, 2004; Mammen and Tsybakov, 1999)).

Assumption 2.13. *Let C' be the constant defined in Assumption 2.9. There exist $\alpha > 0$ and $0 < r \leq (\sqrt{2}C')^{-(2+\varepsilon)/\varepsilon}$ such that for all $x \in \mathcal{X}$, for all $z \in \mathbb{R}$, $|z - f^*(x)| \leq (\sqrt{2}C')^{(2+\varepsilon)/\varepsilon} r$*

$$\min\big(\eta(x), 1 - \eta(x), |1 - 2\eta(x)|\big) \geq \alpha \ .$$

Assumption 2.13 is also local and has the same flavor as Assumptions 2.10 and 2.11.

Theorem 2.9. *Grant Assumptions 2.9 (with constant C') and 2.13 (with parameter r and α). Assume that the oracle f^* is the Bayes estimator i.e. $f^*(x) = sign(2\eta(x) - 1)$ for all $x \in \mathcal{X}$. Then, for all $f \in F$ such that $\|f - f^*\|_{L_2} \leqslant r$, $\|f - f^*\|_{L_2}^2 \leqslant \frac{2}{\alpha} P\mathcal{L}_f$.*

The proof is postponed to Section 2.10.2.

2.6 Comparison between ERM and minmax MOM

In this section, we show that robustness properties with respect to heavy-tailed data and to outliers of the minmax MOM estimator in Theorem 2.2 cannot be achieved by the ERM. We prove two lower bounds on the statistical risk of ERM. First, we show that ERM is not robust to contamination in the design X and second that ERM cannot achieve the optimal rate with a sub-Gaussian deviation under only moment assumptions.

We first show the absence of robustness of ERM w.r.t. contamination by even a single input variable. We consider the absolute loss function of linear functionals $\ell_t(x, y) = |y - \langle x, t\rangle|$. Let $X_1, \ldots, X_N$ denote i.i.d. Gaussian vectors, and suppose that there exists t^* such that $Y_i = \langle X_i, t^*\rangle, i = 1, \ldots, N$. Assume that a vector $v \in \mathbb{R}^d$ was added to X_1 (and that this is the only corrupted data). Hence, we are given the dataset $(X_1 + v, Y_1), (X_2, Y_2), \cdots, (X_N, Y_N)$. Consider the ERM constructed on this dataset i.e $\hat{t}^{ERM} \in \operatorname{argmin}_{t\in\mathbb{R}^d} P_N\ell_t$ where $P_N\ell_t = (1/N)|Y_1 - \langle X_1 + v, t\rangle| + (1/N)\sum_{i=2}^N |Y_i - \langle X_i, t\rangle|$. In this context, the following lower bound holds.

Proposition 2.1. *There exist absolute constants c_0 and c_2 such that the following holds. If the contamination vector v satisfies $|\langle v, t^*\rangle| \geq (1/2)\|v\|_2 \|t^*\|_2$, with $\|v\|_2 \geq c_2 N$, then with probability at least $1 - 4\exp(-c_0 N)$, $\left\|\hat{t}^{ERM} - t^*\right\|_2 \geq \|t^*\|_2 /4$.*

When $N \asymp d$, from Theorem 2.2 with $K \asymp d$, minmax MOM estimators yields, with probability at least $1 - 2\exp(-cd)$, $\left\|\hat{t}^{MOM} - t^*\right\|_2 \lesssim d/N$ on the same dataset as the one used by ERM in Proposition 2.1. If $\|t^*\| \gtrsim 1$ then the ERM is suboptimal compared with the minmax MOM estimator.

Proof. To show that $\hat{t}^{ERM}$ is outside $\mathcal{B} = B_2(t^*, (1/4)\|t^*\|_2) = \{t \in \mathbb{R}^d : \|t - t^*\|_2 \leq (1/4)\|t^*\|_2\}$, it is enough to show that $P_N\ell_0$ is smaller than the smallest value of $t \to P_N\ell_t$ over $\mathcal{B}$. It follows from Gaussian concentration that, with probability at least $1 - \exp(-c_0 N)$,

$$P_N\ell_0 = \frac{1}{N}\sum_{i=1}^N |\langle X_i, t^*\rangle| \leq \frac{3}{2}\sqrt{\frac{2}{\pi}}\|t^*\|_2 . \tag{2.14}$$

Let us now bound from bellow the empirical loss function $t \to P_N\ell_t$ uniformly over all t in $\mathcal{B}$. First,

$$|\langle v, t\rangle| \geq |\langle v, t^*\rangle| - |\langle v, t - t^*\rangle| \geq \|v\|_2 (\|t^*\|_2 /2 - \|t - t^*\|_2) \geq \|v\|_2 \|t^*\|_2 /4. \tag{2.15}$$

Then, it follows from Borell-TIS inequality (see Theorem 7.1 in (Ledoux, 2001) or pages 56-57 in (Ledoux and Talagrand, 2013)) that with probability at least $1-2\exp(-c_1N)$, $\|X_1\|_2 \leq \mathbb{E}\|X_1\|_2 + \sqrt{2c_1N} \leq c_3\sqrt{N+d}$. Therefore, $\mathbb{P}(\Omega_1) \geq 1-2\exp(-c_1N)$, where

$$\Omega_1 = \{\forall t \in \mathbb{R}^d, \quad |\langle X_1, t-t^*\rangle| \leq c_3\sqrt{N+d}\,\|t-t^*\|_2\} \ .$$

On Ω_1, we have for all $t \in \mathcal{B}$ that $|\langle X_1, t^*-t\rangle| \leq c_3\sqrt{N+d}\,\|t-t^*\|_2 \leq (c_3/4)\sqrt{N+d}\,\|t^*\|_2$. Therefore, using (2.15) and $\|v\|_2 \geq c_2N$ for a large enough constant c_2,

$$\begin{aligned} P_N\ell_t &= \frac{1}{N}|\langle X_1, t^*-t\rangle - \langle v, t\rangle| + \frac{1}{N}\sum_{i=2}^N |\langle X_i, t^*-t\rangle| \geq \frac{1}{N}|\langle v, t\rangle| - \frac{1}{N}|\langle X_1, t^*-t\rangle| \\ &\geq \frac{1}{4N}\|v\|_2\|t^*\|_2 - \frac{c_3\|t^*\|_2}{4\sqrt{N}} > \frac{3}{2}\sqrt{\frac{2}{\pi}}\|t^*\|_2 \ . \end{aligned} \tag{2.16}$$

■

It follows from Proposition 2.1 that ERM is not consistant when there is even a single outlier among the X_i. By comparison, the minmax MOM has optimal performance even when the dataset has been corrupted by up to d outliers when $N \gtrsim d$. This shows a first advantage of the minmax MOM approach.

Now, we prove a second advantage of the minmax MOM over the ERM by considering heavy-tailed design. We also consider the absolute L_1-loss function as in the previous example and suppose that data are generated from a linear model in dimension $d=1$: $Y = Xt^* + \zeta$ where X and ζ are independent mean zero random variables and $t^* \in \mathbb{R}$ (we choose $d=1$ so that we have access to a canonical definition of median which simplifies the proof). Our aim is to show that if the design X has only a second moment then the ERM $\hat{t}^{ERM}$ cannot achieve the optimal rate $\sqrt{x/N}$ with a sub-Gaussian deviation that is $1-\exp(-c_0x)$ as does the minmax MOM for all $x \in [1, N]$.

Proposition 2.2. *Let $N \geq 8000$ and $10 \leq x \leq N/800$. There exist X and ζ two symmetric and independent random variables such that $\mathbb{E}X^2 \in [1,16]$, $\mathbb{E}\zeta^2 \leq 5x^2$ and, for any $t^* \in \mathbb{R}$ and $Y = Xt^* + \zeta$, we have $\{t^*\} = \operatorname{argmin}_{t\in\mathbb{R}} \mathbb{E}|Y - Xt|$. Let $(X_i, Y_i)_{i=1}^N$ be N i.i.d. copies of (X, Y) such that $Y = Xt^* + \zeta$ for some $t^* \in \mathbb{R}$. Let $\hat{t}^{ERM} \in \operatorname{argmin}_{t\in\mathbb{R}} \sum_{i=1}^N |Y_i - X_it|$. Then, with probability at least $3/(5x)$,*

$$\sqrt{\mathbb{E}[(X(\hat{t}^{ERM} - t^*))^2]} \geq (1/5)\sqrt{x/N} \ .$$

Proof. Let $\delta' = (1/8)\sqrt{x/(2N)}$ and let ζ be uniformly distributed over $[-x-1/2+\delta', -x] \cup [-\delta', \delta'] \cup [x, x+1/2-\delta']$. Let ϵ denote a Rademacher variable, let η be a Bernoulli variable with parameter $\delta = 1/(xN)$ and $R = 4/\sqrt{\delta} = 4\sqrt{xN}$. We assume that ζ, ϵ and η are independent and let $X = \epsilon(1+R\eta)$. Let $t^* \in \mathbb{R}$ and let $\mathcal{D}_N = (X_i, Y_i)_{i=1}^N$ be a dataset of N i.i.d. copies of (X, Y), where $Y = Xt^* + \zeta$.

Since the median of ζ is 0, for all $u \in \mathbb{R}, \mathbb{E}|u-\zeta| \geq \mathbb{E}|\zeta|$ with equality iff $u = 0$. As a consequence, for all $t \in \mathbb{R}, \mathbb{E}|Y - Xt| = \mathbb{E}_X\mathbb{E}_\zeta|\zeta - X(t^*-t)| \geq \mathbb{E}|\zeta|$ and the only minimizer of $t \in \mathbb{R} \to \mathbb{E}|Y - Xt|$

is t^*. In other words, t^* is the oracle. For all $t \in \mathbb{R}$,

$$\mathbb{E}[(X(t-t^*))^2] = \mathbb{E}[X^2](t-t^*)^2 = (1 + 2R\delta + R^2\delta)(t-t^*)^2 \ .$$

Since $R^2\delta = 16$ and $2R\delta \leq \sqrt{8}/100 \leq 1$, we have $(t-t^*)^2 \leq \mathbb{E}(X(t-t^*))^2 \leq 18(t-t^*)^2$, that is, the $L^2(\mu)$-norm is equivalent to the absolute value.

Observe that $\hat{t}^{ERM} - t^*$ is solution of the minimization problem

$$(\hat{t}^{ERM} - t^*) \in \operatorname*{argmin}_{u\in\mathbb{R}} \sum_{i=1}^{N} |X_i| \left| \frac{\zeta_i}{X_i} - u \right| = \operatorname*{argmin}_{u\in\mathbb{R}} \mathbb{E}[|W-u||\mathcal{D}_N] \ .$$

Here, defining $\zeta_i' = \epsilon_i\zeta_i, i \in [N]$, W is a random variable such that

$$\mathbb{P}\left[W = \frac{\zeta_i'}{|X_i|}|\mathcal{D}_N\right] = \frac{|X_i|}{\sum_{i=1}^N |X_i|} \ .$$

Notice that, almost surely, all $\zeta_i'/|X_i|$ are different. In particular, $|\hat{t}^{ERM} - t^*|$ is the absolute value of the empirical median $|\text{Median}(W)|$. Therefore, $|\hat{t}^{ERM} - t^*| \geq c_1\sqrt{x/N}$ when the median of W does not belong to $(-c_1\sqrt{x/N}, c_1\sqrt{x/N})$. This holds when $\mathbb{P}[W \leq -c_1\sqrt{x/N}|\mathcal{D}_N] > 1/2$ or $\mathbb{P}[W \geq c_1\sqrt{x/N}|\mathcal{D}_N] > 1/2$. Introduce the following sets

$$I_{\leq -x} := \{i \in [N] : \zeta_i' \leq -x\}, I_{\delta'} := \{i \in [N] : |\zeta_i'| \leq \delta'\} \text{ and } I_{\geq x} := \{i \in [N] : \zeta_i' \geq x\} \ .$$

Define also the following events

$$\Omega_0 := \left\{ |I_{\delta'}| \leq \sqrt{2xN}, \qquad \big||I_{\leq -x}| - |I_{\geq x}|\big| \leq \sqrt{2xN} \right\} \ ,$$
$$\Omega_1 := \{\forall i \in I_\delta : \eta_i = 0 \text{ and } |\{i \in [N] : \eta_i = 1\}| = 1\} \ .$$

By Hoeffding's inequality (see Chapter 2 in (Boucheron et al., 2013)), as $(\zeta_i')_{i=1}^N$ is a family of i.i.d. random variables distributed like ζ_1, with probability at least $1 - \exp(-x/4)$,

$$|I_{\delta'}| = \sum_{i=1}^{N} I(|\zeta_i'| \leq \delta') \leq N\mathbb{P}[|\zeta_1'| \leq \delta'] + \sqrt{\frac{xN}{2}} = 2\delta' N + \sqrt{\frac{xN}{2}} \leq \sqrt{2xN} \ .$$

Since $\mathbb{P}[\zeta_1' \leq -x] = \mathbb{P}[\zeta_1' \geq x]$, $I(\zeta_i' \leq -x) - I(\zeta_i' \geq x)$ are independent, centered random variables taking values in $[-1, 1]$. By Hoeffding's inequality, with probability at least $1 - 2\exp(-x/2)$,

$$\big||I_{\leq -x}| - |I_{\geq x}|\big| = \left| \sum_{i=1}^{N} I(\zeta_i' \leq -x) - I(\zeta_i' \geq x) \right| \leq \sqrt{2xN}.$$

Using a union bound, we have $\mathbb{P}[\Omega_0] \geq 1 - 2\exp(-x/2) - \exp(-x/4) \geq 1 - 1/(10x)$ when $x \geq 10$. Since the ζ_i''s and the η_i's are independent, on the event Ω_0, we have

$$\mathbb{P}[\forall i \in I_{\delta'} : \eta_i = 0|\mathcal{D}_N] = (1-\delta)^{|I_{\delta'}|} \geq (1-\delta)^{\sqrt{2xN}} \geq 1 - 2\delta\sqrt{2xN} = 1 - \frac{2\sqrt{2}}{\sqrt{xN}} \geq 1 - \frac{1}{10x} \ .$$

The last inequality holds since $x \leq N/800$. Moreover,

$$\mathbb{P}[|\{i \in [N] : \eta_i = 1\}| = 1] = N\delta(1-\delta)^{N-1} \geq \frac{1-2/x}{x} \ .$$

When $x \geq 10$, this implies

$$\mathbb{P}[|\{i \in [N] : \eta_i = 1\}| = 1] \geq \frac{4}{5x} \ .$$

Therefore, on the event Ω_0, $\mathbb{P}[\Omega_1|\mathcal{D}_N] \geq 7/(10x)$ and so

$$\mathbb{P}[\Omega_0 \cap \Omega_1] = \mathbb{E}[\mathbf{1}_{\Omega_0}\mathbb{E}[\mathbf{1}_{\Omega_1}|\mathcal{D}_N] \geq \frac{4}{5x}\left(1 - \frac{1}{10x}\right) \geq 3/(5x) \ .$$

We want to show that $|\text{Median}(W)| \geq c_1\sqrt{x/N}$ on the event $\Omega_0 \cap \Omega_1$, for some well-chosen constant $c_1 > 0$. We have $\mathbb{P}[W \leq -c_1\sqrt{x/N}|\mathcal{D}_N] > 1/2$ if and only if

$$\sum_{i=1}^{N} I\left(\frac{\zeta_i'}{|X_i|} \leq -c_1\sqrt{x/N}\right)|X_i| \geq \frac{1}{2}\sum_{i=1}^{N}|X_i| \ .$$

In particular, if

$$\sum_{i \in I_{\leq -x}} |X_i| \geq \frac{1}{2}\sum_{i=1}^{N}|X_i| \text{ or } \sum_{i \in I_{\geq x}} |X_i| \geq \frac{1}{2}\sum_{i=1}^{N}|X_i| \tag{2.17}$$

then the median of W takes value in $\{\zeta_i'/|X_i| : i \in I_{\leq -x}\}$, resp. in $\{\zeta_i'/|X_i| : i \in I_{\geq x}\}$. Since, for all $i \in I_{\leq -x}, \zeta_i/|X_i| \leq -x/(1+R) < -\delta'$ and for all $i \in I_{\geq x}, \zeta_i/|X_i| \geq x/(1+R) > \delta'$, in these cases, $|\text{Median}(W)| \geq x/(1+R) = x/(1+4\sqrt{xN}) \geq (1/5)\sqrt{x/N}$. Since $|\hat{t}^{ERM} - t^*| = |\text{Median}(W)|$, the proof is finished if (2.17) is proved.

Let us now prove that (2.17) holds on the event $\Omega_0 \cap \Omega_1$. On this event, only one η_i equals to 1. Therefore only one $|X_i|$ equals to $1+R$ and all the others equal 1. Moreover, $\eta_i = 0$ for all $i \in I_{\delta'}$. Therefore, if $i^* \in [N]$ denotes the only index such that $\eta_{i^*} = 1$, then either $i^* \in I_{\leq -x}$ or $i^* \in I_{\geq x}$. If $i^* \in I_{\leq -x}$, on $\Omega_0 \cap \Omega_1$,

$$\sum_{i \in I_{\leq -x}} |X_i| = |I_{\leq x}| - 1 + (1+R) = |I_{\leq -x}| + 4\sqrt{xN} \geq |I_{\geq x}| - \sqrt{\frac{xN}{2}} + |I_{\delta'}| - \sqrt{2xN} + 4\sqrt{xN} \geq |I_{\geq x}| + |I_{\delta'}| \ .$$

Moreover, all the $|X_i|$ equal 1 when $i \in I_{\geq x} \cup I_{\delta'}$. Therefore, $|I_{\geq x}| + |I_{\delta'}| = \sum_{i \in I_{\geq x} \cup I_{\delta'}} |X_i|$. Overall, $\sum_{i \in I_{\leq -x}} |X_i| \geq \sum_{i \in I_{\geq x} \cup I_{\delta'}} |X_i|$ which is equivalent to $\sum_{i \in I_{\leq -x}} |X_i| \geq (1/2)\sum_{i=1}^{N} |X_i|$. Likewise, if $i^* \in I_{\geq x}$ then $\sum_{i \in I_{\geq x}} |X_i| \geq (1/2)\sum_{i=1}^{N} |X_i|$. Therefore, on the event $\Omega_0 \cap \Omega_1$ (2.17) holds. ∎

Proposition 2.2 shows that the distance between the ERM and t^* is larger than $(1/5)\sqrt{x/N}$ with probability at least $3/(5x)$. This probability is larger than $1 - \exp(-x/2016)$ for large values of x, which shows that the ERM does not have sub-Gaussian deviations. Let us now show that, using the same data and a number of blocks $K \asymp x$ (or using the adaptive estimator (2.12)), the minmax MOM estimators achieve the rate $\sqrt{x/N}$ with probability at least $1 - \exp(-x/2016)$. This

will show a second advantage of minmax MOM estimators compared with ERM for heavy-tailed designs.

To apply Theorem 2.2 (or Theorem 2.3 for the adaptive estimator), we show that the local Bernstein condition is satisfied for the example of Proposition 2.2 and compute the complexity parameter $\tilde{r}_2(\gamma)$. We have

$$\mathbb{E}\sup_{t\in\mathbb{R}:\mathbb{E}[(X(t-t^*))^2]\le r^2}\left|\sum_{i=1}^N \sigma_i X_i(t-t^*)\right| = \frac{r}{\sqrt{\mathbb{E}X^2}}\mathbb{E}\left|\sum_{i=1}^N \sigma_i\epsilon_i(1+R\eta_i)\right|$$
$$\le \frac{r\sqrt{2}}{17}\left(\mathbb{E}\left(\sum_{i=1}^N \sigma_i\epsilon_i\right)^2 + R^2\mathbb{E}\left(\sum_{i=1}^N \sigma_i\epsilon_i\eta_i\right)^2\right)^{1/2} \le \frac{r\sqrt{2}}{17}\sqrt{N+R^2N\delta} = r\sqrt{\frac{2N}{17}}.$$

As a consequence, $\tilde{r}_2(\gamma) = (1/\gamma)\sqrt{2/(17N)}$ satisfies (2.6). We now prove Assumption 2.6 in this particular example. Let $K \in [N]$ be such that $K \ge 2(575)^2/(17\times 865)$ so that, for $C_{K,r}$ is defined in (2.7) with $L=1$ and A defined later,

$$C_{K,r} = A^2\max\left(\frac{2(575)^2}{17N}, \frac{865K}{N}\right) = \frac{A^2 865K}{N} \ ,$$

Let $t\in\mathbb{R}$ be such that $\mathbb{E}[(X(t-t^*))^2] = C_{K,r}$. We have to show that $P\mathcal{L}_t \ge A\mathbb{E}[(X(t-t^*))^2]$ for some well chosen A and $P\mathcal{L}_t = \mathbb{E}[|Y-Xt| - |Y-Xt^*|]$. It follows from (2.39) that $P\mathcal{L}_t = \mathbb{E}[g(X,Xt)-g(X,Xt^*)]$ where $g:(x,a)\in\mathbb{R}^2 \mapsto \int \mathbf{1}_{y\ge a}(1-F_{Y|X=x}(y))\mathrm{d}y + (1/2)a$ and $F_{Y|X=x}$ is the cdf of Y given $X=x$. Therefore, if we denote by F the cdf of ζ, we have

$$P\mathcal{L}_t = \mathbb{E}\int_{Xt}^{Xt^*}(1-F_{Y|X=X}(y))dy = \frac{1-\delta}{2}\int_{t-t^*}^0 (1-F(y))dy + \frac{1-\delta}{2}\int_{t^*-t}^0(1-F(y))dy$$
$$+\frac{\delta}{2}\int_{(1+R)(t-t^*)}^0(1-F(y))dy + \frac{\delta}{2}\int_{(1+R)(t^*-t)}^0(1-F(y))dy.$$

Let us choose K such that $\sqrt{C_{K,r}} \le \sqrt{\mathbb{E}X^2}\delta'$ (which holds for instance when $865A^2K \le 17x/128$). In that case, $|t-t^*|\le\delta'$ and so $(1-F(y)) = (1/2-y)$ for all $y\in[-|t-t^*|, |t-t^*|]$. We therefore have

$$\frac{1-\delta}{2}\int_{t-t^*}^0(1-F(y))dy + \frac{1-\delta}{2}\int_{t^*-t}^0(1-F(y))dy = \frac{(1-\delta)(t-t^*)^2}{2}.$$

Moreover, since $(1+R)|t-t^*| = (1+4\sqrt{xN})\sqrt{C_{K,r}/\mathbb{E}X^2} \le 5\sqrt{xN}\delta' = 5x/(8\sqrt{2})$, we have

$$\frac{\delta}{2}\int_{(1+R)(t-t^*)}^0(1-F(y))dy + \frac{\delta}{2}\int_{(1+R)(t^*-t)}^0(1-F(y))dy \ge \frac{-10x\delta}{8\sqrt{2}} = \frac{-5}{4\sqrt{2}N}.$$

Assume that $(1/16)^2 865K \ge 18 * 40/\sqrt{2}$, so $|t-t^*| \ge C_{K,r}/18 \ge 40/(\sqrt{2}N)$. Then, we have

$$P\mathcal{L}_t \ge (1-\delta)(t-t^*)^2/2 - 5/(4\sqrt{2}N) \ge (t-t^*)^2/16 \ .$$

For $A = 1/(16\times 18)$, this yields $P\mathcal{L}_t \ge A\mathbb{E}[X(t-t^*)^2]$, which concludes the proof of the Berntein's assumption.

2.7 Simulation study

This section provides a short simulation study that illustrates our theoretical findings for the minmax MOM estimators. Let us consider the following setup: $X = (\xi_1, \cdots, \xi_d)$, where $(\xi_j)_{j=1}^d$ are independent and identically distributed, with $\xi_1 \sim \mathcal{T}(5)$, and

$$\log\left(\frac{\mathbf{P}(Y=1|X)}{\mathbf{P}(Y=-1|X)}\right) = \langle X, t^* \rangle + \epsilon$$

where $\epsilon \sim \mathcal{LN}(0,1)$. Let $(X_i, Y_i)_{i=1}^N$ be i.i.d with the same distribution as (X,Y). We the study the minmax MOM estimator defined as:

$$\hat{t}_K^{\text{MOM}} \in \arg\min_{t\in\mathbb{R}^p} \sup_{\tilde{t}\in\mathbb{R}^p} \text{MOM}_K(\ell_t - \ell_{\tilde{t}}) \ . \tag{2.18}$$

Following (Lecué and Lerasle, 2019), a gradient ascent-descent step is performed on the empirical incremental risk $(t,\tilde{t}) \to P_{B_k}(\ell_t - \ell_{\tilde{t}})$ constructed on the block B_k of data realizing the median of the empirical incremental risk. Initial points $t_0 \in \mathbb{R}^d$ and $\tilde{t}_0 \in \mathbb{R}^d$ are taken at random. In logistic regression, the step sizes η and $\tilde{\eta}$ are usually chosen equal to $\|\mathbb{X}\mathbb{X}^\top\|_{\text{op}}/4N$, where $\mathbb{X}$ is the $N \times d$ matrix with row vectors equal to $X_1^\top, \cdots, X_N^\top$ and $\|\cdot\|_{\text{op}}$ denotes the operator norm. In a corrupted environment, this choice might lead to disastrous performance. This is why η and $\tilde{\eta}$ are computed at each iteration using only data in the median block: let B_k denote the median block at the current step, then one chooses $\eta = \tilde{\eta} = \|\mathbb{X}_{(k)}\mathbb{X}_{(k)}^\top\|_{\text{op}}/4|B_k|$ where $\mathbb{X}_{(k)}$ is the $|B_k| \times p$ matrix with rows given by X_i^T for $i \in B_k$. In practice, K is chosen by robust cross-validation choice as in (Lecué and Lerasle, 2019).

In a first approach and according to our theoretical results, the blocks are chosen at the beginning of the algorithm. As illustrated in Figure 2.2, this first strategy has some limitations. To understand the problem, for all $k = 1, \ldots, K$, let C_k denote the following set

$$C_k = \left\{ t \in \mathbb{R}^d : P_{B_k}\ell_t = \text{Median}\left\{P_{B_1}\ell_t, \ldots, P_{B_K}\ell_t\right\}\right\} \ .$$

If the minimum of $t \to P_{B_k}\ell_t$ lies in C_k, the algorithm typically converges to this minimum if one iteration enters C_k. As a consequence, when the minmax MOM estimator (2.18) lies in another cell, the algorithm does not converge to this estimator.

To bypass this issue, the partition is changed at every ascent/descent steps of the algorithm, it is chosen uniformly at random among all equipartition of the dataset. This alternative algorithm is described in Algorithm 1. In practice, changing the partition seems to widely accelerate the convergence (see Figure 2.2).

Simulation results are gathered in Figure 2.2. In these simulations, there is no outlier, $N = 1000$ and $d = 100$ with $(X_i, Y_i)_{i=1}^{1000}$ i.i.d with the same distribution as (X,Y). Minmax MOM estimators (2.18) are compared with the Logistic Regression algorithm from the scikit-learn library of (Pedregosa et al., 2011).

Input: The number of block K, initial points t_0 and $\tilde{t}_0$ in $\mathbb{R}^p$ and the stopping criterion $\epsilon > 0$

Output: An estimator of t^*

while $\|t_i - \tilde{t}_i\|_2 \geq \epsilon$ **do**

  Split the data into K disjoint blocks $(B_k)_{k\in\{1,\cdots,K\}}$ of equal sizes chosen at random: $B_1 \cup \cdots \cup B_K = \{1, \cdots, N\}$.

  Find $k \in [K]$ such that $\mathrm{MOM}_K \ell_{t_i} = P_{B_k} \ell_{t_i}$.

  Compute $\eta = \tilde{\eta} = \|\mathbb{X}_{(k)}^T \mathbb{X}_{(k)}\|_{op}/4N$.

  Update $t_{i+1} = t_i - \frac{1}{\eta} \nabla_t (P_{B_k} \ell_t)_{|t=t_i}$ and $\tilde{t}_{i+1} = \tilde{t}_i - \frac{1}{\tilde{\eta}} \nabla_{\tilde{t}} (P_{B_k} \ell_{\tilde{t}})_{|\tilde{t}=\tilde{t}_i}$.

end

Algorithm 1: Descent-ascent gradient method with blocks of data chosen at random at every steps.

The upper pictures compare performance of MOM ascent/descent algorithms with fixed and changing blocks. These pictures give an example where the fixed block algorithm is stuck into local minima and another one where it does not converge. In both cases, the changing blocks version converges to t^*.

Running times of logistic regression (LR) and its MOM version (MOM LR) are compared in the lower picture of Figure 2.2 in a dataset free from outliers. LR and MOM LR are coded with the same algorithm in this example, meaning that MOM gradient descent-ascent and simple gradient descent are performed with the same descent algorithm. As illustrated in Figure 2.2, running each step of the gradient descent on one block only and not on the whole dataset accelerates the running time. The larger the dataset, the bigger the benefit is expected.

The resistance to outliers of logistic regression and its minmax MOM alternative are depicted in Figure 2.1 in the introduction. We added an increasing number of outliers to the dataset. Outliers $\{(X_i, Y_i), i \in \mathcal{O}\}$ in this simulation are such that $X_i \sim \mathcal{LN}(0, 5)$ and $Y_i = -sign(\langle X_i, t \rangle + \epsilon_i)$, with $\epsilon_i \sim \epsilon$ as above. Figure 2.1 shows that logistic classification is mislead by a single outlier while MOM version maintains reasonable performance with up to 50 outliers (i.e 5% of the database is corrupted).

A byproduct of Algorithm 1 is an outlier detection algorithm. Each data receives a score equal to the number of times it is selected in a median block in the random choice of block version of the algorithm. The first iterations may be misleading: before convergence, the empirical loss at the current point may not reveal the centrality of the data because the current point may be far from t^*. Simulations are run with $N = 100$, $d = 10$ and 5000 iterations and therefore only the score obtained by each data in the last 4000 iterations are displayed. 3 outliers $(X_i, Y_i)_{i\in\{1,2,3\}}$ with $X_i = (10)_{j=1}^d$ and $Y_i = -sign(\langle X_i, t \rangle)$ have been introduced at number 42, 62 and 66. Figure 2.3 shows that these are not selected once.

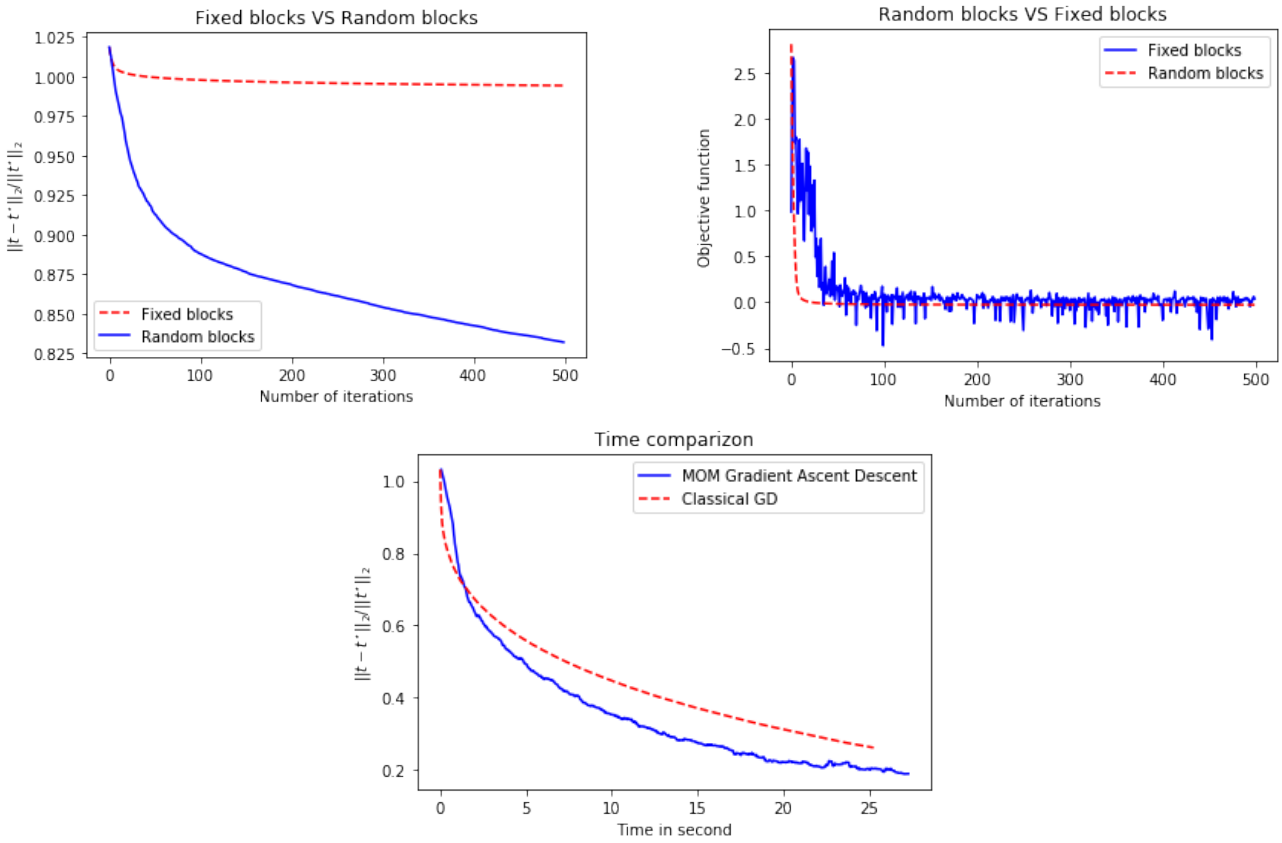


Figure 2.2: Top left and right: Comparizon of the algorithm with fixed and changing blocks. Bottom: Comparizon of running time between classical gradient descent and algorithm 1. In all simulation $N = 1000$, $p = 100$ and there is no outliers.

2.8 Conclusion

The paper introduces a new homogenity argument for learning problems with convex and Lipschitz losses. This argument allows to obtain estimation rates and oracle inequalities for ERM and minmax MOM estimators improving existing results. The ERM requires sub-Gaussian hypotheses on the class F with respect to the distribution of the design and a local Bernstein condition (see Theorem 2.1), both assumptions can be removed for minmax MOM estimators (see Theorem 2.5). The local Bernstein conditions provided in this article can be verified in several learning problems. In particular, it allows to derive optimal risk bounds in examples where analyses based on the small ball hypothesis fail. Minmax MOM estimators applied to convex and Lipschitz losses are efficient under weak assumptions on the outputs Y, under minimal L_2 assumptions on the class F with respect to the distribution of the design and the results are robust to the presence of few outliers in the dataset. A modification of these estimators can be implemented efficiently and confirm all these conclusions.

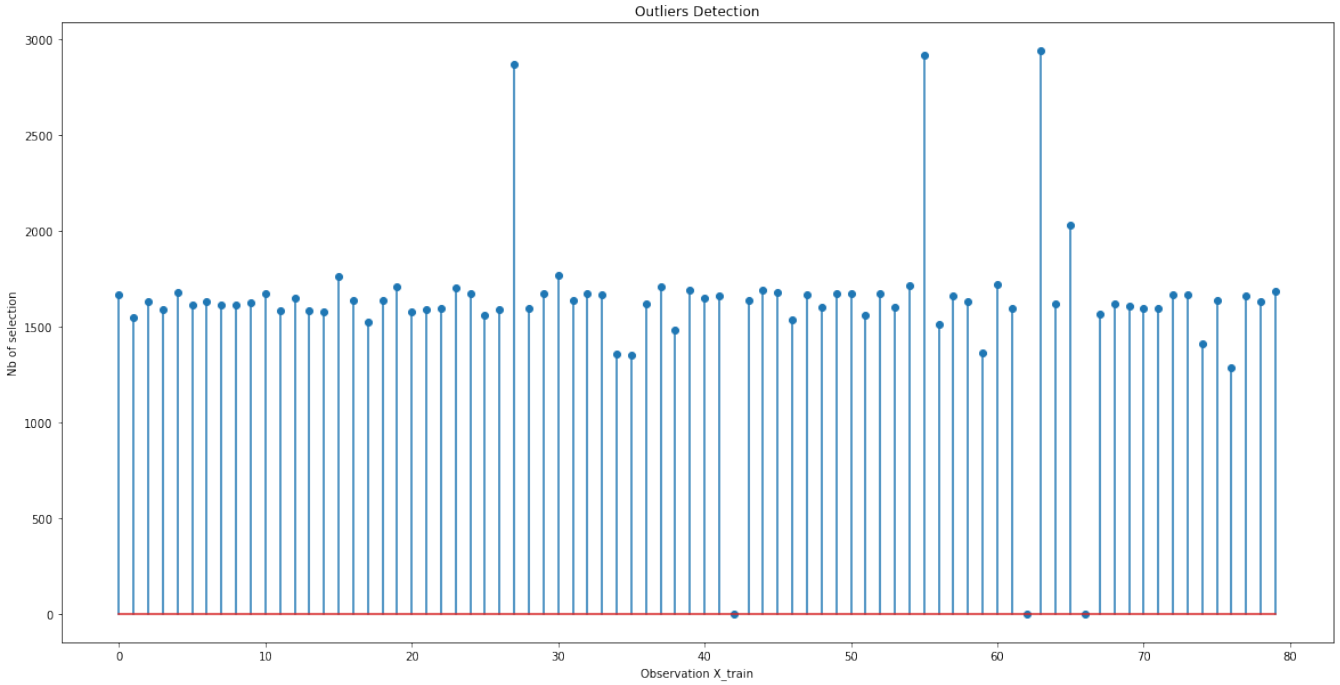


Figure 2.3: Outliers Detection Procedure for $N = 100$, $p = 10$ and outliers are $i = 42, 62, 66$

2.9 Proof of main Theorems

2.9.1 Proof of Theorem 2.1

The proof is split in two parts. First, we identify an event where the statistical behavior of the regularized estimator $\hat{f}^{ERM}$ can be controlled. Then, we prove that this event holds with probability at least (2.3). Introduce $\theta = 1/(2A)$ and define the following event:

$$\Omega := \left\{ \forall f \in F \cap (f^* + r_2(\theta)B_{L_2}), \quad \left|(P - P_N)\mathcal{L}_f\right| \leq \theta r_2^2(\theta) \right\}$$

where θ is a parameter appearing in the definition of r_2 in Definition 2.3.

Proposition 2.3. *On the event Ω, one has*

$$\|\hat{f}^{ERM} - f^*\|_{L_2} \leq r_2(\theta) \text{ and } P\mathcal{L}_{\hat{f}^{ERM}} \leq \theta r_2^2(\theta).$$

Proof. By construction, $\hat{f}^{ERM}$ satisfies $P_N\mathcal{L}_{\hat{f}^{ERM}} \leq 0$. Therefore, it is sufficient to show that, on Ω, if $\|f - f^*\|_{L_2} > r_2(\theta)$, then $P_N\mathcal{L}_f > 0$. Let $f \in F$ be such that $\|f - f^*\|_{L_2} > r_2(\theta)$. By convexity of F, there exists $f_0 \in F \cap (f^* + r_2(\theta)S_{L_2})$ and $\alpha > 1$ such that

$$f = f^* + \alpha(f_0 - f^*) \ . \tag{2.19}$$

For all $i \in \{1, \cdots, N\}$, let $\psi_i : \mathbb{R} \to \mathbb{R}$ be defined for all $u \in \mathbb{R}$ by

$$\psi_i(u) = \overline{\ell}(u + f^*(X_i), Y_i) - \overline{\ell}(f^*(X_i), Y_i). \tag{2.20}$$

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

$$P_N\mathcal{L}_f = \frac{1}{N}\sum_{i=1}^{N}\psi_i\big(f(X_i) - f^*(X_i)\big) = \frac{1}{N}\sum_{i=1}^{N}\psi_i(\alpha(f_0(X_i) - f^*(X_i)))$$
$$\geq \frac{\alpha}{N}\sum_{i=1}^{N}\psi_i((f_0(X_i) - f^*(X_i))) = \alpha P_N\mathcal{L}_{f_0}. \tag{2.21}$$

Until the end of the proof, the event Ω is assumed to hold. Since $f_0 \in F \cap (f^* + r_2(\theta)S_{L_2})$, $P_N\mathcal{L}_{f_0} \geq P\mathcal{L}_{f_0} - \theta r_2^2(\theta)$. Moreover, by Assumption 2.4, $P\mathcal{L}_{f_0} \geq A^{-1}\|f_0 - f^*\|_{L_2}^2 = A^{-1}r_2^2(\theta)$, thus

$$P_N\mathcal{L}_{f_0} \geq (A^{-1} - \theta)r_2^2(\theta). \tag{2.22}$$

From Eq. (2.21) and (2.22), $P_N\mathcal{L}_f > 0$ since $A^{-1} > \theta$. Therefore, $\|\hat{f}^{ERM} - f^*\|_{L_2} \leq r_2^2(\theta)$. This proves the L_2-bound.

Now, as $\|\hat{f}^{ERM} - f^*\|_{L_2} \leq r_2^2(\theta)$, $|(P - P_N)\mathcal{L}_{\hat{f}^{ERM}}| \leq \theta r_2^2(\theta)$. Since $P_N\mathcal{L}_{\hat{f}^{ERM}} \leq 0$,

$$P\mathcal{L}_{\hat{f}^{ERM}} = P_N\mathcal{L}_{\hat{f}^{ERM}} + (P - P_N)\mathcal{L}_{\hat{f}^{ERM}} \leq \theta r_2^2(\theta) \ .$$

This show the excess risk bound. ∎

Proposition 2.3 shows that $\hat{f}^{ERM}$ has the risk bounds given in Theorem 2.1 on the event Ω. To show that Ω holds with probability (2.3), recall the following results from (Alquier et al., 2019).

Lemma 2.2. *(Alquier et al., 2019) [Lemma 8.1] Grant Assumptions 2.1 and 2.3. Let $F' \subset F$ with finite L_2-diameter $d_{L_2}(F')$. For every $u > 0$, with probability at least $1 - 2\exp(-u^2)$,*

$$\sup_{f,g\in F'}|(P - P_N)(\mathcal{L}_f - \mathcal{L}_g)| \leq \frac{16L}{\sqrt{N}}\left(w(F') + ud_{L_2}(F')\right) \ .$$

It follows from Lemma 2.2 that for any $u > 0$, with probability larger that $1 - 2\exp(-u^2)$,

$$\sup_{f\in F\cap(f^*+r_2(\theta)B_{L_2})}\big|(P - P_N)\mathcal{L}_f\big| \leq \sup_{f,g\in F\cap(f^*+r_2(\theta)B_{L_2})}\big|(P - P_N)(\mathcal{L}_f - \mathcal{L}_g)\big|$$
$$\leq \frac{16L}{\sqrt{N}}\big(w((F - f^*) \cap r_2(\theta)B_{L_2}) + ud_{L_2}((F - f^*) \cap r_2(\theta)B_{L_2})\big)$$

where $d_{L_2}((F - f^*) \cap r_2(\theta)B_{L_2}) \leq r_2(\theta)$. By definition of the complexity parameter (see Eq. (2.3)), for $u = \theta\sqrt{N}r_2(\theta)/(64L)$, with probability at least

$$1 - 2\exp\big(-\theta^2 N r_2^2(\theta)/(16^3L^2)\big) \ , \tag{2.23}$$

for every f in $F \cap (f^* + r_2(\theta)B_{L_2})$,

$$\big|(P - P_N)\mathcal{L}_f\big| \leq \theta r_2^2(\theta). \tag{2.24}$$

Together with Proposition 2.3, this concludes the proof of Theorem 2.1.

2.9.2 Proof of Theorem 2.2

The proof is split in two parts. First, we identify an event Ω_K where the statistical properties of $\hat{f}$ from Theorem 2.2 can be established. Next, we prove that this event holds with probability (2.8). Let α, θ and γ be positive numbers to be chosen later. Define

$$C_{K,r} = \max\left(\frac{4L^2K}{\theta^2\alpha N}, \tilde{r}_2^2(\gamma)\right)$$

where the exact form of α, θ and γ are given in Equation (2.33). Set the event Ω_K to be such that

$$\Omega_K = \left\{\forall f \in F \cap \left(f^* + \sqrt{C_{K,r}}B_{L_2}\right), \exists J \subset \{1,\dots,K\} : |J| > K/2 \text{ and } \forall k \in J, |(P_{B_k} - P)\mathcal{L}_f| \leq \theta C_{K,r}\right\} . \tag{2.25}$$

Deterministic argument

The goal of this section is to show that, on the event Ω_K, $\|\hat{f} - f^*\|_{L_2}^2 \leq C_{K,r}$ and $P\mathcal{L}_{\hat{f}} \leq 2\theta C_{K,r}$.

Lemma 2.3. *If there exists $\eta > 0$ such that*

$$\sup_{f \in F \setminus \left(f^* + \sqrt{C_{K,r}}B_{L_2}\right)} MOM_K\left(\ell_{f^*} - \ell_f\right) < -\eta \quad \text{and} \quad \sup_{f \in F \cap \left(f^* + \sqrt{C_{K,r}}B_{L_2}\right)} MOM_K\left(\ell_{f^*} - \ell_f\right) \leq \eta \ , \tag{2.26}$$

then $\|\hat{f} - f^*\|_{L_2}^2 \leq C_{K,r}$.

Proof. Assume that (2.26) holds, then

$$\inf_{f \in F \setminus \left(f^* + \sqrt{C_{K,r}}B_{L_2}\right)} \mathrm{MOM}_K[\ell_f - \ell_{f^*}] > \eta \ . \tag{2.27}$$

Moreover, if $T_K(f) = \sup_{g \in F} \mathrm{MOM}_K[\ell_f - \ell_g]$ for all $f \in F$, then

$$T_K(f^*) = \sup_{f \in F \cap \left(f^* + \sqrt{C_{K,r}}B_{L_2}\right)} \mathrm{MOM}_K[\ell_{f^*} - \ell_f] \vee \sup_{f \in F \setminus \left(f^* + \sqrt{C_{K,r}}B_{L_2}\right)} \mathrm{MOM}_K[\ell_{f^*} - \ell_f] \leqslant \eta \ . \tag{2.28}$$

By definition of $\hat{f}$ and (2.28), $T_K(\hat{f}) \leqslant T_K(f^*) \leqslant \eta$. Moreover, by (2.27), any $f \in F \setminus \left(f^* + \sqrt{C_{K,r}}B_{L_2}\right)$ satisfies $T_K(f) \geqslant \mathrm{MOM}_K[\ell_f - \ell_{f^*}] > \eta$. Therefore $\hat{f} \in F \cap (f^* + \sqrt{C_{K,r}}B_{L_2})$. ∎

Lemma 2.4. *Grant Assumption 2.6 and assume that $\theta - A^{-1} < -\theta$. On the event Ω_K,* (2.26) *holds with $\eta = \theta C_{K,r}$.*

Proof. Let $f \in F$ be such that $\|f - f^*\|_{L_2} > C_{K,r}$. By convexity of F, there exists $f_0 \in F \cap \left(f^* + \sqrt{C_{K,r}}S_{L_2}\right)$ and $\alpha > 1$ such that $f = f^* + \alpha(f_0 - f^*)$. For all $i \in \{1,\dots,N\}$, let $\psi_i : \mathbb{R} \to \mathbb{R}$ be defined for all $u \in \mathbb{R}$ by

$$\psi_i(u) = \overline{\ell}(u + f^*(X_i), Y_i) - \overline{\ell}(f^*(X_i), Y_i). \tag{2.29}$$

The functions ψ_i are convex because $\overline{\ell}$ is and such that $\psi_i(0) = 0$, so $\alpha\psi_i(u) \leq \psi_i(\alpha u)$ for all $u \in \mathbb{R}$ and $\alpha \geq 1$. As $\psi_i(f(X_i) - f^*(X_i)) = \overline{\ell}(f(X_i), Y_i) - \overline{\ell}(f^*(X_i), Y_i)$, for any block B_k,

$$\begin{aligned} P_{B_k}\mathcal{L}_f &= \frac{1}{|B_k|}\sum_{i\in B_k}\psi_i\big(f(X_i) - f^*(X_i)\big) = \frac{1}{|B_k|}\sum_{i\in B_k}\psi_i(\alpha(f_0(X_i) - f^*(X_i))) \\ &\geq \frac{\alpha}{|B_k|}\sum_{i\in B_k}\psi_i((f_0(X_i) - f^*(X_i))) = \alpha P_{B_k}\mathcal{L}_{f_0}. \end{aligned} \tag{2.30}$$

As $f_0 \in F \cap (f^* + \sqrt{C_{K,r}}S_{L_2})$, on Ω_K, there are strictly more than $K/2$ blocks B_k where $P_{B_k}\mathcal{L}_{f_0} \geq P\mathcal{L}_{f_0} - \theta C_{K,r}$. Moreover, from Assumption 2.6, $P\mathcal{L}_{f_0} \geq A^{-1}\|f_0 - f^*\|_{L_2}^2 = A^{-1}C_{K,r}$. Therefore, on strictly more than $K/2$ blocks B_k,

$$P_{B_k}\mathcal{L}_{f_0} \geq (A^{-1} - \theta)C_{K,r}. \tag{2.31}$$

From Eq. (2.30) and (2.31), there are strictly more than $K/2$ blocks B_k where $P_{B_k}\mathcal{L}_f \geq (A^{-1} - \theta)C_{K,r}$. Therefore, on Ω_K, as $(\theta - A^{-1}) < -\theta$,

$$\sup_{f\in F\setminus\left(f^*+\sqrt{C_{K,r}}B_{L_2}\right)} \mathrm{MOM}_K\big(\ell_{f^*} - \ell_f\big) < (\theta - A^{-1})C_{K,r} < -\theta C_{K,r} \ .$$

In addition, on the event Ω_K, for all $f \in F \cap (f^* + \sqrt{C_{K,r}}B_{L_2})$, there are strictly more than $K/2$ blocks B_k where $|(P_{B_k} - P)\mathcal{L}_f| \leq \theta C_{K,r}$. Therefore

$$\mathrm{MOM}_K\big(\ell_{f^*} - \ell_f\big) \leq \theta C_{K,r} - P\mathcal{L}_f \leq \theta C_{K,r}.$$

■

Lemma 2.5. *Grant Assumption 2.6 and assume that $\theta - A^{-1} < -\theta$. On the event Ω_K, $P\mathcal{L}_{\hat{f}} \leq 2\theta C_{K,r}$.*

Proof. Assume that Ω_K holds. From Lemmas 2.3 and 2.4, $\|\hat{f} - f^*\|_{L_2} \leq \sqrt{C_{K,r}}$. Therefore, on strictly more than $K/2$ blocks B_k, $P\mathcal{L}_{\hat{f}} \leq P_{B_k}\mathcal{L}_{\hat{f}} + \theta C_{K,r}$. In addition, by definition of $\hat{f}$ and (2.28) (for $\eta = \theta C_{K,r}$),

$$MOM_K(\ell_{\hat{f}} - \ell_{f^*}) \leq \sup_{f\in F} MOM_K(\ell_{f^*} - \ell_f) \leq \theta C_{K,r}.$$

As a consequence, there exist at least $K/2$ blocks B_k where $P_{B_k}\mathcal{L}_{\hat{f}} \leq \theta C_{K,r}$. Therefore, there exists at least one block B_k where both $P\mathcal{L}_{\hat{f}} \leq P_{B_k}\mathcal{L}_{\hat{f}} + \theta C_{K,r}$ and $P_{B_k}\mathcal{L}_{\hat{f}} \leq \theta C_{K,r}$. Hence $P\mathcal{L}_{\hat{f}} \leq 2\theta C_{K,r}$.

■

Stochastic argument

This section shows that Ω_K holds with probability at least (2.8).

Proposition 2.4. *Grant Assumptions 2.1, 2.2, 2.5 and 2.6 and assume that* $(1-\beta)K \geq |\mathcal{O}|$. *Let* $x > 0$ *and assume that* $\beta(1-\alpha-x-8\gamma L/\theta) > 1/2$. *Then* Ω_K *holds with probability larger than* $1-\exp(-x^2\beta K/2)$.

Proof. Let $\mathcal{F} = F \cap \left(f^* + \sqrt{C_{K,r}}B_{L_2}\right)$ and set $\phi : t \in \mathbb{R} \to I\{t \geq 2\} + (t-1)I\{1 \leq t \leq 2\}$ so, for all $t \in \mathbb{R}$, $I\{t \geq 2\} \leq \phi(t) \leq I\{t \geq 1\}$. Let $W_k = ((X_i, Y_i))_{i \in B_k}$, $G_f(W_k) = (P_{B_k} - P)\mathcal{L}_f$. Let

$$z(f) = \sum_{k=1}^{K} I\{|G_f(W_k)| \leq \theta C_{K,r}\}.$$

Let $\mathcal{K}$ denote the set of indices of blocks which have not been corrupted by outliers, $\mathcal{K} = \{k \in \{1, \cdots, K\} : B_k \subset \mathcal{I}\}$ and let $f \in \mathcal{F}$. Basic algebraic manipulations show that

$$z(f) \geq |\mathcal{K}| - \sup_{f \in \mathcal{F}} \sum_{k \in \mathcal{K}} \left(\phi(2\theta^{-1}C_{K,r}^{-1}|G_f(W_k)|) - \mathbb{E}\phi(2\theta^{-1}C_{K,r}^{-1}|G_f(W_k)|)\right) - \sum_{k \in \mathcal{K}} \mathbb{E}\phi(2\theta^{-1}C_{K,r}^{-1}|G_f(W_k)|) \ .$$

By Assumptions 2.1 and 2.5, using that $C_{K,r}^2 \geq \|f - f^*\|_{L_2}^2 \left[(4L^2K)/(\theta^2\alpha N)\right]$,

$$\begin{aligned}\mathbb{E}\phi(2\theta^{-1}C_{K,r}^{-1}|G_f(W_k)|) &\leq \mathbb{P}\left(|G_f(W_k)| \geq \frac{\theta C_{K,r}}{2}\right) \leq \frac{4}{\theta^2 C_{K,r}^2}\mathbb{E}G_f(W_k)^2 = \frac{4}{\theta^2 C_{K,r}^2}\mathbb{V}ar(P_{B_k}\mathcal{L}_f) \\ &\leq \frac{4K^2}{\theta^2 C_{K,r}^2 N^2}\sum_{i \in B_k}\mathbb{E}[\mathcal{L}_f^2(X_i, Y_i)] \leq \frac{4L^2K}{\theta^2 C_{K,r}^2 N}\|f - f^*\|_{L_2}^2 \leq \alpha \ .\end{aligned}$$

Therefore,

$$z(f) \geq |\mathcal{K}|(1-\alpha) - \sup_{f \in \mathcal{F}} \sum_{k \in \mathcal{K}} \left(\phi(2\theta^{-1}C_{K,r}^{-1}|G_f(W_k)|) - \mathbb{E}\phi(2\theta^{-1}C_{K,r}^{-1}|G_f(W_k)|)\right) \ . \tag{2.32}$$

Using Mc Diarmid's inequality (Boucheron et al., 2013, Theorem 6.2), for all $x > 0$, with probability larger than $1 - \exp(-x^2|\mathcal{K}|/2)$,

$$\begin{aligned}&\sup_{f \in \mathcal{F}} \sum_{k \in \mathcal{K}} \left(\phi(2\theta^{-1}C_{K,r}^{-1}|G_f(W_k)|) - \mathbb{E}\phi(2\theta^{-1}C_{K,r}^{-1}|G_f(W_k)|)\right) \\ &\qquad \leq x|\mathcal{K}| + \mathbb{E}\sup_{f \in \mathcal{F}} \sum_{k \in \mathcal{K}} \left(\phi(2\theta^{-1}C_{K,r}^{-1}|G_f(W_k)|) - \mathbb{E}\phi(2\theta^{-1}\theta^{-1}C_{K,r}^{-1}|G_f(W_k)|)\right) \ .\end{aligned}$$

Let $\epsilon_1, \ldots, \epsilon_K$ denote independent Rademacher variables independent of the $(X_i, Y_i), i \in \mathcal{I}$. By Giné-Zinn symmetrization argument,

$$\begin{aligned}&\sup_{f \in \mathcal{F}} \sum_{k \in \mathcal{K}} \left(\phi(2\theta^{-1}C_{K,r}^{-1}|G_f(W_k)|) - \mathbb{E}\phi(2\theta^{-1}C_{K,r}^{-1}|G_f(W_k)|)\right) \\ &\qquad \leq x|\mathcal{K}| + 2\mathbb{E}\sup_{f \in \mathcal{F}} \sum_{k \in \mathcal{K}} \epsilon_k \phi(2\theta^{-1}C_{K,r}^{-1}|G_f(W_k)|)\end{aligned}$$

As ϕ is 1-Lipschitz with $\phi(0) = 0$, using the contraction lemma (Ledoux and Talagrand, 2013, Chapter 4),

$$\mathbb{E}\sup_{f\in\mathcal{F}}\sum_{k\in\mathcal{K}}\epsilon_k\phi(2\theta^{-1}C_{K,r}^{-1}|G_f(W_k)|) \le 2\mathbb{E}\sup_{f\in\mathcal{F}}\sum_{k\in\mathcal{K}}\epsilon_k\frac{G_f(W_k)}{\theta C_{K,r}} = 2\mathbb{E}\sup_{f\in\mathcal{F}}\sum_{k\in\mathcal{K}}\epsilon_k\frac{(P_{B_k}-P)\mathcal{L}_f}{\theta C_{K,r}}.$$

Let $(\sigma_i : i \in \cup_{k\in\mathcal{K}}B_k)$ be a family of independent Rademacher variables independent of $(\epsilon_k)_{k\in\mathcal{K}}$ and $(X_i, Y_i)_{i\in\mathcal{I}}$. It follows from the Giné-Zinn symmetrization argument that

$$\mathbb{E}\sup_{f\in\mathcal{F}}\sum_{k\in\mathcal{K}}\epsilon_k\frac{(P_{B_k}-P)\mathcal{L}_f}{C_{K,r}} \le 2\mathbb{E}\sup_{f\in\mathcal{F}}\frac{K}{N}\sum_{i\in\cup_{k\in\mathcal{K}}B_k}\sigma_i\frac{\mathcal{L}_f(X_i,Y_i)}{C_{K,r}}.$$

By the Lipschitz property of the loss, the contraction principle applies and

$$\mathbb{E}\sup_{f\in\mathcal{F}}\sum_{i\in\cup_{k\in\mathcal{K}}B_k}\sigma_i\frac{\mathcal{L}_f(X_i,Y_i)}{C_{K,r}} \le L\mathbb{E}\sup_{f\in\mathcal{F}}\sum_{i\in\cup_{k\in\mathcal{K}}B_k}\sigma_i\frac{(f-f^*)(X_i)}{C_{K,r}} \ .$$

To bound from above the right-hand side in the last inequality, consider two cases 1) $C_{K,r} = \tilde{r}_2^2(\gamma)$ or 2) $C_{K,r} = 4L^2K/(\alpha\theta^2N)$. In the first case, by definition of the complexity parameter $\tilde{r}_2(\gamma)$ in (2.6),

$$\mathbb{E}\sup_{f\in\mathcal{F}}\sum_{i\in\cup_{k\in\mathcal{K}}B_k}\sigma_i\frac{(f-f^*)(X_i)}{C_{K,r}} = \mathbb{E}\sup_{f\in F:\|f-f^*\|_{L_2}\le\tilde{r}_2(\gamma)}\frac{1}{\tilde{r}_2^2(\gamma)}\left|\sum_{i\in\cup_{k\in\mathcal{K}}B_k}\sigma_i(f-f^*)(X_i)\right| \le \frac{\gamma|\mathcal{K}|N}{K}.$$

In the second case,

$$\begin{aligned}
&\mathbb{E}\sup_{f\in\mathcal{F}}\sum_{i\in\cup_{k\in\mathcal{K}}B_k}\frac{\sigma_i(f-f^*)(X_i)}{C_{K,r}} \\
&\le \mathbb{E}\left[\sup_{\substack{f\in F:\\ \|f-f^*\|_{L_2}\le\tilde{r}_2(\gamma)}}\left|\sum_{i\in\cup_{k\in\mathcal{K}}B_k}\frac{\sigma_i(f-f^*)(X_i)}{\tilde{r}_2^2(\gamma)}\right| \vee \sup_{\substack{f\in F:\\ \tilde{r}_2(\gamma)\le\|f-f^*\|_{L_2}\le\sqrt{\frac{4L^2K}{\alpha\theta^2N}}}}\left|\sum_{i\in\cup_{k\in\mathcal{K}}B_k}\sigma_i\frac{(f-f^*)(X_i)}{\frac{4L^2K}{\alpha\theta^2N}}\right|\right] .
\end{aligned}$$

Let $f \in F$ be such that $\tilde{r}_2(\gamma) \le \|f-f^*\|_{L_2} \le \sqrt{[4L^2K]/[\alpha\theta^2N]}$; by convexity of F, there exists $f_0 \in F$ such that $\|f_0-f^*\|_{L_2} = \tilde{r}_2(\gamma)$ and $f = f^* + \alpha(f_0-f^*)$ with $\alpha = \|f-f^*\|_{L_2}/\tilde{r}_2(\gamma) \ge 1$. Therefore,

$$\left|\sum_{i\in\cup_{k\in\mathcal{K}}B_k}\sigma_i\frac{(f-f^*)(X_i)}{\frac{4L^2K}{\alpha\theta^2N}}\right| \le \frac{1}{\tilde{r}_2(\gamma)}\left|\sum_{i\in\cup_{k\in\mathcal{K}}B_k}\sigma_i\frac{(f-f^*)(X_i)}{\|f-f^*\|_{L_2}}\right| = \frac{1}{\tilde{r}_2^2(\gamma)}\left|\sum_{i\in\cup_{k\in\mathcal{K}}B_k}\sigma_i(f_0-f^*)(X_i)\right|$$

and so

$$\sup_{\substack{f\in F:\\ \tilde{r}_2(\gamma)\le\|f-f^*\|_{L_2}\le\sqrt{\frac{4L^2K}{\alpha\theta^2N}}}}\left|\sum_{i\in\cup_{k\in\mathcal{K}}B_k}\sigma_i\frac{(f-f^*)(X_i)}{\frac{4L^2K}{\alpha\theta^2N}}\right| \le \frac{1}{\tilde{r}_2^2(\gamma)}\sup_{\substack{f\in F:\\ \|f-f^*\|_{L_2}=\tilde{r}_2(\gamma)}}\left|\sum_{i\in\cup_{k\in\mathcal{K}}B_k}\sigma_i(f-f^*)(X_i)\right|.$$

By definition of $\tilde{r}_2(\gamma)$, it follows that

$$\mathbb{E}\sup_{f\in\mathcal{F}}\left|\sum_{i\in\cup_{k\in\mathcal{K}}B_k}\sigma_i\frac{(f-f^*)(X_i)}{C_{K,r}}\right| \le \frac{\gamma|\mathcal{K}|N}{K}.$$

Therefore, as $|\mathcal{K}| \geq K - |\mathcal{O}| \geq \beta K$, with probability larger than $1 - \exp(-x^2\beta K/2)$, for all $f \in F$ such that $\|f - f^*\|_{L_2} \leq \sqrt{C_{K,r}}$,

$$z(f) \geq |\mathcal{K}| \left(1 - \alpha - x - \frac{8\gamma L}{\theta}\right) > \frac{K}{2}. \tag{2.33}$$

■

End of the proof of Theorem 2.2

Theorem 2.2 follows from Lemmas 2.3, 2.4, 2.5 and Proposition 2.4 for the choice of constant

$$\theta = 1/(3A) \quad \alpha = 1/24, \quad x = 1/24, \quad \beta = 4/7 \text{ and } \gamma = 1/(575AL).$$

2.9.3 Proof of Theorem 2.3

Let $K \in \left[7|\mathcal{O}|/3, N\right]$ and consider the event Ω_K defined in (2.25). It follows from the proof of Lemmas 2.3 and 2.4 that $T_K(f^*) \leq \theta C_{K,r}$ on Ω_K. Setting $\theta = 1/(3A)$, on $\cap_{J=K}^N \Omega_J$, $f^* \in \hat{R}_J$ for all $J = K, \ldots, N$, so $\cap_{J=K}^N \hat{R}_J \neq \emptyset$. By definition of $\hat{K}$, it follows that $\hat{K} \leq K$ and by definition of $\tilde{f}$, $\tilde{f} \in \hat{R}_K$ which means that $T_K(\tilde{f}) \leq \theta C_{K,r}$. It is proved in Lemmas 2.3 and 2.4 that on Ω_K, if $f \in F$ satisfies $\|f - f^*\|_{L_2} \geq \sqrt{C_{K,r}}$ then $T_K(f) > \theta C_{K,r}$. Therefore, $\left\|\tilde{f} - f^*\right\|_{L_2} \leq \sqrt{C_{K,r}}$. On Ω_K, since $\left\|\tilde{f} - f^*\right\|_{L_2} \leq \sqrt{C_{K,r}}$, $P\mathcal{L}_{\tilde{f}} \leq 2\theta C_{K,r}$. Hence, on $\cap_{J=K}^N \Omega_J$, the conclusions of Theorem 2.3 hold. Finally, by Proposition 2.4,

$$\mathbb{P}\left[\cap_{J=K}^N \Omega_J\right] \geq 1 - \sum_{J=K}^{N} \exp(-K/2016) \geq 1 - 4\exp(-K/2016).$$

2.9.4 Proof of Theorem 2.4

The proof of Theorem 2.4 follows the same path as the one of Theorem 2.2. We only sketch the different arguments needed because of the localization by the excess loss and the lack of Bernstein condition.

Define the event Ω'_K in the same way as Ω_K in (2.25) where $C_{K,r}$ is replaced by $\bar{r}_2^2(\gamma)$ and the L_2 localization is replaced by the "excess loss localization":

$$\Omega'_K = \left\{\forall f \in (\mathcal{L}_F)_{\bar{r}_2^2(\gamma)}, \exists J \subset \{1, \ldots, K\} : |J| > K/2 \text{ and } \forall k \in J, |(P_{B_k} - P)\mathcal{L}_f| \leq (1/4)\bar{r}_2^2(\gamma)\right\} \tag{2.34}$$

where $(\mathcal{L}_F)_{\bar{r}_2^2(\gamma)} = \{f \in F : P\mathcal{L}_f \leq \bar{r}_2^2(\gamma)\}$. Our first goal is to show that on the event Ω'_K, $P\mathcal{L}_{\hat{f}} \leq (1/4)\bar{r}_2^2(\gamma)$. We will then handle $\mathbb{P}[\Omega'_K]$.

Lemma 2.6. *Grant Assumptions 2.1 and 2.2. For every $r \geq 0$, the set $(\mathcal{L}_F)_r := \{f \in F : P\mathcal{L}_f \leq r\}$ is convex and relatively closed to F in $L_1(\mu)$. Moreover, if $f \in F$ is such that $P\mathcal{L}_f > r$ then there exists $f_0 \in F$ and $(P\mathcal{L}_f/r) \geq \alpha > 1$ such that $(f - f^*) = \alpha(f_0 - f^*)$ and $P\mathcal{L}_{f_0} = r$.*

Proof. Let f and g be in $(\mathcal{L}_F)_r$ and $0 \leq \alpha \leq 1$. We have $\alpha f + (1-\alpha)g \in F$ because F is convex and for all $x \in \mathcal{X}$ and $y \in \mathbb{R}$, using the convexity of $u \to \bar{\ell}(u + f^*(x), y)$, we have

$$\begin{aligned}
&\ell_{\alpha f+(1-\alpha)g}(x,y) - \ell_{f^*}(x,y) = \bar{\ell}(\alpha(f-f^*)(x) + (1-\alpha)(g-f^*)(x) + f^*(x), y) - \bar{\ell}(f^*(x), y) \\
&\leq \alpha\big(\bar{\ell}((f-f^*)(x) + f^*(x), y) - \bar{\ell}(f^*(x), y)\big) + (1-\alpha)\big(\bar{\ell}((g-f^*)(x) + f^*(x), y) - \bar{\ell}(f^*(x), y)\big) \\
&= \alpha(\ell_f - \ell_{f^*}) + (1-\alpha)(\ell_g - \ell_{f^*})
\end{aligned}$$

and so $P\mathcal{L}_{\alpha f+(1-\alpha)g} \leq \alpha P\mathcal{L}_f + (1-\alpha)P\mathcal{L}_g$. Given that $P\mathcal{L}_f, P\mathcal{L}_g \leq r$ we also have $P\mathcal{L}_{\alpha f+(1-\alpha)g} \leq r$. Therefore, $\alpha f + (1-\alpha)g \in (\mathcal{L}_F)_r$ and $(\mathcal{L}_F)_r$ is convex.

For all $f, g \in F$, $|P\mathcal{L}_f - P\mathcal{L}_g| \leq \|f - f^*\|_{L_1(\mu)}$ so that $f \in F \to P\mathcal{L}_f$ is continuous onto F in $L_1(\mu)$ and therefore its level sets, such as $(\mathcal{L}_F)_r$, are relatively closed to F in $L_1(\mu)$.

Finally, let $f \in F$ be such that $P\mathcal{L}_f > r$. Define $\alpha_0 = \sup\{\alpha \geq 0 : f^* + \alpha(f - f^*) \in (\mathcal{L}_F)_r\}$. Note that $P\mathcal{L}_{f^*+\alpha(f-f^*)} \leq \alpha P\mathcal{L}_f = r$ for $\alpha = r/P\mathcal{L}_f$ so that $\alpha_0 \geq r/P\mathcal{L}_f$. Since $(\mathcal{L}_F)_r$ is relatively closed to F in $L_1(\mu)$, we have $f^* + \alpha_0(f - f^*) \in (\mathcal{L}_F)_r$ and in particular $\alpha_0 < 1$ otherwise, by convexity of $(\mathcal{L}_F)_r$, we would have $f \in (\mathcal{L}_F)_r$. Moreover, by maximality of α_0, $f_0 = f^* + \alpha_0(f - f^*)$ is such that $P\mathcal{L}_{f_0} = r$ and the results follows for $\alpha = \alpha_0^{-1}$. ∎

Lemma 2.7. *Grant Assumptions 2.1 and 2.2. On the event Ω'_K, $P\mathcal{L}_{\hat{f}} \leq \bar{r}_2^2(\gamma)$.*

Proof. Let $f \in F$ be such that $P\mathcal{L}_f > \bar{r}_2^2(\gamma)$. It follows from Lemma 2.6 that there exists $\alpha \geq 1$ and $f_0 \in F$ such that $P\mathcal{L}_{f_0} = \bar{r}_2^2(\gamma)$ and $f - f^* = \alpha(f_0 - f^*)$. According to (2.30), we have for every $k \in \{1, \dots, K\}$, $P_{B_k}\mathcal{L}_f \geq \alpha P_{B_k}\mathcal{L}_{f_0}$. Since $f_0 \in (\mathcal{L}_F)_{\bar{r}_2^2(\gamma)}$, on the event Ω'_K, there are strictly more than $K/2$ blocks B_k such that $P_{B_k}\mathcal{L}_{f_0} \geq P\mathcal{L}_{f_0} - (1/4)\bar{r}_2^2(\gamma) = (3/4)\bar{r}_2^2(\gamma)$ and so $P_{B_k}\mathcal{L}_f \geq (3/4)\bar{r}_2^2(\gamma)$. As a consequence, we have

$$\sup_{f \in F \setminus (\mathcal{L}_F)_{\bar{r}_2^2(\gamma)}} \mathrm{MOM}_K\big(\ell_{f^*} - \ell_f\big) \leq (-3/4)\bar{r}_2^2(\gamma) \ . \tag{2.35}$$

Moreover, on the event Ω'_K, for all $f \in (\mathcal{L}_F)_{\bar{r}_2^2(\gamma)}$, there are strictly more than $K/2$ blocks B_k such that $P_{B_k}(-\mathcal{L}_f) \leq (1/4)\bar{r}_2^2(\gamma) - P\mathcal{L}_f \leq (1/4)\bar{r}_2^2(\gamma)$. Therefore,

$$\sup_{f \in (\mathcal{L}_F)_{\bar{r}_2^2(\gamma)}} \mathrm{MOM}_K\big(\ell_{f^*} - \ell_f\big) \leq (1/4)\bar{r}_2^2(\gamma) \ . \tag{2.36}$$

We conclude from (2.35) and (2.36) that $\sup_{f \in F} \mathrm{MOM}_K\big(\ell_{f^*} - \ell_f\big) \leq (1/4)\bar{r}_2^2(\gamma)$ and that every $f \in F$ such that $P\mathcal{L}_f > \bar{r}_2^2(\gamma)$ satisfies $\mathrm{MOM}_K\big(\ell_f - \ell_{f^*}\big) \geq (3/4)\bar{r}_2^2(\gamma)$. But, by definition of $\hat{f}$, we have

$$\mathrm{MOM}_K\big(\ell_{\hat{f}} - \ell_{f^*}\big) \leq \sup_{f \in F} \mathrm{MOM}_K\big(\ell_{f^*} - \ell_f\big) \leq (1/4)\bar{r}_2^2(\gamma) \ .$$

Therefore, we necessarily have $P\mathcal{L}_{\hat{f}} \leq \bar{r}_2^2(\gamma)$. ∎

Now, we prove that Ω'_K is an exponentially large event using similar argument as in Proposition 2.4.

Proposition 2.5. *Grant Assumptions 2.1, 2.2 and 2.7 and assume that* $(1-\beta)K \geq |\mathcal{O}|$ *and* $\beta(1-1/12-32\gamma L) > 1/2$. *Then* Ω'_K *holds with probability larger than* $1-\exp(-\beta K/1152)$.

Sketch of proof. The proof of Proposition 2.5 follows the same line as the one of Proposition 2.4. Let us precise the main differences. We set $\mathcal{F}' = (\mathcal{L}_F)_{\bar{r}_2^2(\gamma)}$ and for all $f \in \mathcal{F}'$, $z'(f) = \sum_{k=1}^K I\{|G_f(W_k)| \leq (1/4)\bar{r}_2^2(\gamma)\}$ where $G_f(W_k)$ is the same quantity as in the proof of Proposition 2.5. Let us consider the contraction ϕ introduced in Proposition 2.5. By definition of $\bar{r}_2^2(\gamma)$ and $V_K(\cdot)$, we have

$$\begin{aligned}\mathbb{E}\phi(8(\bar{r}_2^2(\gamma))^{-1}|G_f(W_k)|) &\leq \mathbb{P}\left(|G_f(W_k)| \geq \frac{\bar{r}_2^2(\gamma)}{8}\right) \leq \frac{64}{(\bar{r}_2^2(\gamma))^2}\mathbb{E}G_f(W_k)^2 = \frac{64}{(\bar{r}_2^2(\gamma))^2}\mathbb{V}ar(P_{B_k}\mathcal{L}_f)\\ &\leq \frac{64K^2}{(\bar{r}_2^2(\gamma))^2N^2}\sum_{i\in B_k}\mathbb{V}ar_{P_i}(\mathcal{L}_f) \leq \frac{64K}{(\bar{r}_2^2(\gamma))^2N}\sup\{\mathbb{V}ar_{P_i}(\mathcal{L}_f) : f\in\mathcal{F}', i\in\mathcal{I}\}\\ &\leq \frac{64K}{(\bar{r}_2^2(\gamma))^2N}\sup\{\mathbb{V}ar_{P_i}(\mathcal{L}_f) : P\mathcal{L}_f \leq \bar{r}_2^2(\gamma), i\in\mathcal{I}\} \leq \frac{1}{24}\ .\end{aligned}$$

Using Mc Diarmid's inequality, the Giné-Zinn symmetrization argument and the contraction lemma twice and the Lipschitz property of the loss function, such as in the proof of Proposition 2.4, we obtain with probability larger than $1-\exp(-|\mathcal{K}|/1152)$, for all $f \in \mathcal{F}'$,

$$z(f) \geq |\mathcal{K}|(1-1/12) - \frac{32LK}{N}\mathbb{E}\sup_{f\in\mathcal{F}'}\frac{1}{\bar{r}_2^2(\gamma)}\left|\sum_{i\in\cup_{k\in\mathcal{K}}B_k}\sigma_i(f-f^*)(X_i)\right|. \tag{2.37}$$

Now, it remains to use the definition of $\bar{r}_2^2(\gamma)$ to bound the expected supremum in the right-hand side of (2.37) to get

$$\mathbb{E}\sup_{f\in\mathcal{F}'}\frac{1}{\bar{r}_2(\gamma)^2}\left|\sum_{i\in\cup_{k\in\mathcal{K}}B_k}\sigma_i(f-f^*)(X_i)\right| \leq \frac{\gamma|\mathcal{K}|N}{K}. \tag{2.38}$$

Proof of Theorem 2.4. The proof of Theorem 2.4 follows from Lemma 2.7 and Proposition 2.5 for $\beta = 4/7$ and $\gamma = 1/(768L)$.

2.10 Other proofs

2.10.1 Proof of Lemma 2.1

Proof. We have

$$\frac{1}{\sqrt{N}}\mathbb{E}\sup_{f\in F:\|f-f^*\|_{L_2}\leq r}\sum_{i=1}^N\sigma_i(f-f^*)(X_i) = \mathbb{E}\sup_{t\in\mathbb{R}^d:\mathbb{E}\left\langle t,X\right\rangle^2\leq r^2}\left\langle t, \frac{1}{\sqrt{N}}\sum_{i=1}^N\sigma_iX_i\right\rangle\ .$$

Let $\Sigma = \mathbb{E}X^TX$ denote the covariance matrix of X and consider its SVD, $\Sigma = QDQ^T$ where $Q = [Q_1|\cdots|Q_d] \in \mathbb{R}^{d\times d}$ is an orthogonal matrix and D is a diagonal $d\times d$ matrix with non-negative entries. For all $t\in\mathbb{R}^d$, we have $\mathbb{E}\big\langle X,t\big\rangle^2 = t^T\Sigma t = \sum_{j=1}^d d_j\big\langle t,Q_j\big\rangle^2$. Then

$$\mathbb{E}\sup_{t\in\mathbb{R}^d:\sqrt{\mathbb{E}\left\langle t,X\right\rangle^2}\le r}\big\langle t,\frac{1}{\sqrt{N}}\sum_{i=1}^N\sigma_iX_i\big\rangle = \mathbb{E}\sup_{t\in\mathbb{R}^d:\sqrt{\mathbb{E}\left\langle t,X\right\rangle^2}\le r}\big\langle\sum_{j=1}^d\langle t,Q_j\rangle Q_j,\frac{1}{\sqrt{N}}\sum_{i=1}^N\sigma_iX_i\big\rangle$$
$$= \mathbb{E}\sup_{t\in\mathbb{R}^d:\sqrt{\sum_{j=1}^d d_j\left\langle t,Q_j\right\rangle^2}\le r}\sum_{j=1:d_j\neq 0}^d\sqrt{d_j}\langle t,Q_j\rangle\big\langle\frac{Q_j}{\sqrt{d_j}},\frac{1}{\sqrt{N}}\sum_{i=1}^N\sigma_iX_i\big\rangle$$
$$\le r\mathbb{E}\sqrt{\sum_{j=1:d_j\neq0}^d\big\langle\frac{Q_j}{\sqrt{d_j}},\frac{1}{\sqrt{N}}\sum_{i=1}^N\sigma_iX_i\big\rangle^2} \le r\sqrt{\mathbb{E}\sum_{j=1:d_j\neq0}^d\big\langle\frac{Q_j}{\sqrt{d_j}},\frac{1}{\sqrt{N}}\sum_{i=1}^N\sigma_iX_i\big\rangle^2}\ .$$

Moreover, for any j such that $d_j\neq 0$,

$$\mathbb{E}\big\langle\frac{Q_j}{\sqrt{d_j}},\frac{1}{\sqrt{N}}\sum_{i=1}^N\sigma_iX_i\big\rangle^2 = \mathbb{E}\frac{1}{N}\sum_{k,l=1}^N\sigma_l\sigma_k\big\langle\frac{Q_j}{\sqrt{d_j}},X_k\big\rangle\big\langle\frac{Q_j}{\sqrt{d_j}},X_l\big\rangle = \frac{1}{N}\sum_{k=1}^N\mathbb{E}\big\langle\frac{Q_j}{\sqrt{d_j}},X_k\big\rangle^2$$
$$= \frac{1}{N}\sum_{k=1}^N\Big(\frac{Q_j}{\sqrt{d_j}}\Big)^T\mathbb{E}X_k^TX_k\Big(\frac{Q_j}{\sqrt{d_j}}\Big) = \frac{1}{N}\sum_{k=1}^N\Big(\frac{Q_j}{\sqrt{d_j}}\Big)^T\Sigma\Big(\frac{Q_j}{\sqrt{d_j}}\Big)$$

By orthonormality, $Q^TQ_j = e_j$ and $Q_j^TQ = e_j^T$, then, for any j such that $d_j\neq 0$,

$$\mathbb{E}\big\langle\frac{Q_j}{\sqrt{d_j}},\frac{1}{\sqrt{N}}\sum_{i=1}^N\sigma_iX_i\big\rangle^2 = \frac{1}{N}\sum_{k=1}^N\frac{1}{d_j}e_j^TDe_j = 1\ .$$

Finally, we obtain

$$\frac{1}{\sqrt{N}}\mathbb{E}\sup_{f\in F:\|f-f^*\|_{L_2}\le r}\sum_{i=1}^N\sigma_i(f-f^*)(X_i) \le r\sqrt{\sum_{j=1}^d\mathbf{1}_{\{d_j\neq0\}}} = r\sqrt{\mathrm{Rank}(\Sigma)}$$

and therefore the fixed point $\tilde r_2(\gamma)$ is such that

$$\tilde r_2(\gamma) = \inf\Big\{r>0,\forall J\in\mathcal{I}:|J|\ge N/2,\ \mathbb{E}\sup_{t\in\mathbb{R}^d:\sqrt{\mathbb{E}\left\langle t-t^*,X\right\rangle^2}\le r}\sum_{i\in J}\sigma_i\big\langle X_i,t-t^*\big\rangle\ \le r^2|J|\gamma\Big\}$$
$$\le \inf\Big\{r>0,\forall J\in\mathcal{I}:|J|\ge N/2,\quad r\sqrt{\mathrm{Rank}(\Sigma)}\le r^2\sqrt{|J|}\gamma\Big\} \le \sqrt{\frac{\mathrm{Rank}(\Sigma)}{2\gamma^2N}}\ .$$

■

2.10.2 Proofs of the results of Section 2.5

We begin this Section with a simple Lemma coming from the convexity of F.

Lemma 2.8. *For any $f \in F$,*

$$\lim_{t \to 0^+} \frac{R(f^* + t(f - f^*)) - R(f^*)}{t} \geq 0$$

where we recall that $R(f) = \mathbb{E}_{(X,Y)\sim P}[\ell_f(X, Y)]$.

Proof. Let $t \in (0, 1)$. By convexity of F, $f^* + t(f - f^*) \in F$ and $R(f^* + t(f - f^*)) - R(f^*) \geq 0$ because f^* minimizes the risk over F. ∎

Proof of Theorem 2.6

Let $r > 0$. Let $f \in F$ be such that $\|f - f^*\|_{L_2} \leq r$. For all $x \in \mathcal{X}$ denote by $F_{Y|X=x}$ the conditional c.d.f. of Y given $X = x$. We have

$$\begin{aligned}\mathbb{E}\Big[\ell_f(X, Y)|X = x\Big] &= (\tau - 1) \int \mathbf{1}_{y \leq f(x)}(y - f(x))F_{Y|X=x}(\mathrm{d}y) + \tau \int \mathbf{1}_{y > f(x)}(y - f(x))F_{Y|X=x}(\mathrm{d}y) \\ &= \int \mathbf{1}_{y > f(x)}(y - f(x))F_{Y|X=x}(\mathrm{d}y) + (\tau - 1) \int \mathbf{1}_{\mathbb{R}}(y - f(x))F_{Y|X=x}(\mathrm{d}y) \ .\end{aligned}$$

By Fubini's theorem,

$$\begin{aligned}&\int \mathbf{1}_{z \geq f(x)}(1 - F_{Y|X=x}(z))\mathrm{d}z = \int \mathbf{1}_{z \geq f(x)}\Big(1 - \mathbb{P}(Y \leq z|X = x)\Big)\mathrm{d}z = \int \mathbf{1}_{z \geq f(x)}\mathbb{E}[\mathbf{1}_{Y>z}|X = x]\mathrm{d}z \\ &= \int\int \mathbf{1}_{y>z\geq f(x)} f_{Y|X=x}(y)\mathrm{d}y\mathrm{d}z = \int \mathbf{1}_{y>f(x)}(y - f(x))f_{Y|X=x}(y)\mathrm{d}y \\ &= \int \mathbf{1}_{y>f(x)}(y - f(x))F_{Y|X=x}(\mathrm{d}y) \ .\end{aligned}$$

Therefore,

$$\begin{aligned}\mathbb{E}\Big[\ell_f(X, Y)|X = x\Big] &= \int \mathbf{1}_{y \geq f(x)}(1 - F_{Y|X=x}(y))\mathrm{d}y + (\tau - 1)\Big(\int_{\mathbb{R}} yF_{Y|X=x}(\mathrm{d}y) - f(x)\Big) \\ &= g(x, f(x)) + (\tau - 1)\int_{\mathbb{R}} yF_{Y|X=x}(\mathrm{d}y)\end{aligned}$$

where $g : (x, a) \in \mathcal{X} \times \mathbb{R} \to \int \mathbf{1}_{y \geq a}(1 - F_{Y|X=x}(y))\mathrm{d}y + (1 - \tau)a$. It follows that

$$P\mathcal{L}_f = \mathbb{E}[g(X, f(X)) - g(X, f^*(X))] \ . \tag{2.39}$$

Since for all $x \in \mathcal{X}$, $a \mapsto g(x, a)$ is twice differentiable, from a second order Taylor expansion we get

$$\begin{aligned}P\mathcal{L}_f = \mathbb{E}\Big[g(X, f(X)) - g(X, f^*(X))\Big] &= \mathbb{E}\Big[\frac{\partial g(X, a)}{\partial a}(f^*(X))(f(X) - f^*(X))\Big] \\ &+ \frac{1}{2}\int_{x \in \mathcal{X}} \frac{\partial^2 g(x, a)}{\partial a^2}(z_x)(f(x) - f^*(x))^2 dP_X(x)\end{aligned}$$

where for all $x \in \mathcal{X}$, z_x is some point in $\big[\min(f(x), f^*(x)), \max(f(x), f^*(x))\big]$. For the first order term, we have

$$\mathbb{E}\bigg[\frac{\partial g(X,a)}{\partial a}(f^*(X))(f(X)-f^*(X))\bigg] = \mathbb{E}\lim_{t\to 0^+}\frac{g(X, f^*(X)+t(f(X)-f^*(X))-g(X,f^*(X))}{t}.$$

For all $x \in \mathcal{X}$, we have $[g(x, f^*(x)+t(f(x)-f^*(x)))-g(x,f^*(x))]/t \leq (2-\tau)|f^*(x)-f(x)|$ which is integrable with respect to P_X. Thus, by the dominated convergence theorem, it is possible to interchange integral and limit and therefore using Lemma 2.8, we obtian

$$\begin{aligned}\mathbb{E}\bigg[\frac{\partial g(X,a)}{\partial a}(f^*(X))(f(X)-f^*(X))\bigg] &= \lim_{t\to 0^+}\mathbb{E}\frac{g(X, f^*(X)+t(f(X)-f^*(X))-g(X,f^*(X))}{t}\\ &= \lim_{t\to 0^+}\frac{R(f^*+t(f-f^*))-R(f^*)}{t} \geq 0.\end{aligned}$$

Given that for all $x \in \mathcal{X}$, $\frac{\partial^2 g(x,a)}{\partial a^2}(z) = f_{Y|X=x}(z)$ for all $z \in \mathbb{R}$ it follows that

$$P\mathcal{L}_f \geq \frac{1}{2}\int_{x\in\mathcal{X}} f_{Y|X=x}(z_x)(f(x)-f^*(x))^2 dP_X(x).$$

Consider $A = \{x \in \mathcal{X}, |f(x)-f^*(x)| \leq (\sqrt{2}C')^{(2+\varepsilon)/\varepsilon} r\}$. Given that $\|f-f^*\|_{L_2} \leq r$, by Markov's inequality, $P(X \in A) \geqslant 1 - 1/(\sqrt{2}C')^{(4+2\varepsilon)/\varepsilon}$. From Assumption 2.10 we get

$$\frac{2P\mathcal{L}_f}{\alpha} \geqslant \mathbb{E}[I_A(X)(f(X)-f^*(X))^2] = \|f-f^*\|_{L_2}^2 - \mathbb{E}[I_{A^c}(X)(f(X)-f^*(X))^2]\ . \tag{2.40}$$

By Holder and Markov's inequalities,

$$\mathbb{E}[I_{A^c}(X)(f(X)-f^*(X))^2] \leqslant \big(\mathbb{E}[I_{A^c}(X)]\big)^{\varepsilon/(2+\varepsilon)}\big(\mathbb{E}[(f(X)-f^*(X))^{2+\varepsilon}]\big)^{2/(2+\varepsilon)} \leqslant \frac{\|f-f^*\|_{L_{2+\varepsilon}}^2}{2(C')^2}\ .$$

By Assumption 2.9, it follows that $\mathbb{E}[I_{A^c}(X)(f(X)-f^*(X))^2] \leqslant \|f-f^*\|_{L_2}^2/2$ and we conclude with (2.40).

Proof of Theorem 2.7

Let $r > 0$. Let $f \in F$ be such that $\|f-f^*\|_{L_2} \leq r$. We have

$$P\mathcal{L}_f = \mathbb{E}_X\mathbb{E}\Big[\rho_H(Y-f(x)) - \rho_H(Y-f^*(x))|X=x\Big] = \mathbb{E}\big[g(X,f(X)) - g(X,f^*(X))\big]$$

where $g : (x,a) \in \mathcal{X}\times\mathbb{R} = \mathbb{E}[\rho_H(Y-a)|X=x]$. Let $F_{Y|X=x}$ denote the c.d.f. of Y given $X=x$. Since for all $x \in \mathcal{X}$, $a \mapsto g(x,a)$ is twice differentiable in its second argument (see Lemma 2.1 in (Elsener and van de Geer, 2018)), a second Taylor expansion yields

$$P\mathcal{L}_f = \mathbb{E}\bigg[\frac{\partial g(X,a)}{\partial a}(f^*(X))(f(X)-f^*(X))\bigg] + \frac{1}{2}\int_{x\in\mathcal{X}}(f(x)-f^*(x))^2\frac{\partial^2 g(x,a)}{\partial a^2}(z_x)dP_X(x)$$

where for all $x \in \mathcal{X}$, z_x is some point in $[\min(f(x), f^*(x)), \max(f(x), f^*(x))]$. By Lemma 2.8, with the same reasoning as the one in Section 2.10.2, we get

$$P\mathcal{L}_f \geq \frac{1}{2}\int_{x\in\mathcal{X}} (f(x) - f^*(x))^2 \frac{\partial^2 g(x,a)}{\partial a^2}(z_x) dP_X(x) \ .$$

Moreover, for all $z \in \mathbb{R}$,

$$\frac{\partial^2 g(x,a)}{\partial a^2}(z) = F_{Y|X=x}(z+\delta) - F_{Y|X=x}(z-\delta).$$

Now, let $A = \{x \in \mathcal{X} : |f(x) - f^*(x)| \leq (\sqrt{2}C')^{(2+\varepsilon)/\varepsilon} r\}$. It follows from Assumption 2.10 that $P\mathcal{L}_f \geq (\alpha/2)\mathbb{E}[(f(X) - f^*(X))^2 I_A(X)]$. Since $\|f - f^*\|_{L_2} \leq r$, by Markov's inequality, $P(X \in A) \geqslant 1 - 1/(\sqrt{2}C')^{(4+2\varepsilon)/\varepsilon}$. By Holder and Markov's inequalities,

$$\mathbb{E}[I_{A^c}(X)(f(X) - f^*(X))^2] \leqslant \big(\mathbb{E}[I_{A^c}(X)]\big)^{\varepsilon/(2+\varepsilon)} \big(\mathbb{E}[(f(X) - f^*(X))^{2+\varepsilon}]\big)^{2/(2+\varepsilon)} \leqslant \frac{\|f - f^*\|^2_{L_{2+\varepsilon}}}{2(C')^2} \ .$$

By Assumption 2.9, it follows that $\mathbb{E}[I_{A^c}(X)(f(X) - f^*(X))^2] \leqslant \frac{\|f-f^*\|^2_{L_2}}{2}$, which concludes the proof.

Proof of Theorem 2.8

Let $r > 0$. Let $f \in F$ be such that $\|f - f^*\|_{L_2} \leq r$. Let $\eta(x) = P(Y = 1|X = x)$. Write first that $P\mathcal{L}_f = \mathbb{E}\Big[g(X, f(X)) - g(X, f^*(X))\Big]$ where for all $x \in \mathcal{X}$ and $a \in \mathbb{R}$, $g(x,a) = \eta(x)\log(1 + \exp(-a)) + (1-\eta(x))\log(1+\exp(a))$. From Lemma 2.8 and the same reasoning as in Section 2.10.2 and 2.10.2 we get

$$P\mathcal{L}_f \geq \int_{x\in\mathcal{X}} \frac{\partial_2^2 g(x,a)}{\partial a^2}(z_x)\frac{(f(x) - f^*(x))^2}{2} dP_X(x) = \int_{x\in\mathcal{X}} \frac{e^{z_x}}{(1+e^{z_x})^2}\frac{(f(x) - f^*(x))^2}{2} dP_X(x)$$

for some $z_x \in [\min(f(x), f^*(x)), \max(f(x), f^*(x))]$. Now, let

$$A = \big\{x \in \mathcal{X} : |f^*(x)| \leq c_0, |f(x) - f^*(x)| \leq (2C')^{(2+\varepsilon)/\varepsilon} r\big\} \ .$$

On the event A we have

$$P\mathcal{L}_f \geq \frac{e^{-c_0 - (2C')^{(2+\varepsilon)/\varepsilon} r}}{2\big(1 + e^{c_0 + (2C')^{(2+\varepsilon)/\varepsilon} r}\big)^2} \mathbb{E}[I_A(X)(f(X) - f^*(X))^2]$$

Using the fact that $P(X \notin A) \leq P(|f^*(X)| > c_0) + P(|f(X) - f^*(X| > (2C')^{(2+\varepsilon)/\varepsilon} r) \leq 2/(2C')^{(4+\varepsilon)/\varepsilon}$, we conclude with Assumption 2.9 and the same analysis as in the two previous proofs.

Proof of Theorem 2.9

Let $r > 0$ such that $r(\sqrt{2}C')^{(2+\varepsilon)/\varepsilon} \leq 1$. Let f be in F such that $\|f - f^*\|_{L_2} \leq r$. For all x in $\mathcal{X}$ let us denote $\eta(x) = \mathbb{P}(Y = 1|X = x)$. It is easy to verify that the Bayes estimator (which is equal to the oracle) is defined as $f^*(x) = \text{sign}(2\eta(x) - 1)$. Consider the set $A = \{x \in \mathcal{X}, |f(x) - f^*(x)| \leq r(\sqrt{2}C')^{(2+\varepsilon)/\varepsilon}\}$. Since $\|f - f^*\|_{L_2} \leq r$, by Markov's inequality $\mathbb{P}(X \in A) \geq 1 - 1/(\sqrt{2}C')^{(4+2\varepsilon)/\varepsilon}$. Let x be in A. If $f^*(x) = -1$ (i.e $2\eta(x) \leq 1$) and $f(x) \leq f^*(x) = -1$ we obtain

$$\mathbb{E}\big[\ell_f(X,Y)|X = x\big] - \mathbb{E}\big[\ell_{f^*}(X,Y)|X = x\big] = \eta(x)(1 - f(x)) - \eta(x)(1 - f^*(x)) \geq \eta(x)\big(f(x) - f^*(x)\big)^2$$

where we used the fact that on A, $|f(x) - f^*(x)| \leq r(\sqrt{2}C')^{(2+\varepsilon)/\varepsilon} \leq 1$. Using the same analysis for the other cases we get that

$$\begin{aligned}\mathbb{E}\big[\ell_f(X,Y)|X = x\big] - \mathbb{E}\big[\ell_{f^*}(X,Y)|X = x\big] &\geq \min\big(\eta(x), 1 - \eta(x), |1 - 2\eta(x)|\big)\big(f(x) - f^*(x)\big)^2 \\ &\geq \alpha\big(f(x) - f^*(x)\big)^2\end{aligned}$$

Therefore,

$$\frac{P\mathcal{L}_f}{\alpha} \geqslant \mathbb{E}[I_A(X)(f(X) - f^*(X))^2] = \|f - f^*\|_{L_2}^2 - \mathbb{E}[I_{A^c}(X)(f(X) - f^*(X))^2] \ . \tag{2.41}$$

By Holder and Markov's inequalities,

$$\mathbb{E}[I_{A^c}(X)(f(X) - f^*(X))^2] \leqslant \big(\mathbb{E}[I_{A^c}(X)]\big)^{\varepsilon/(2+\varepsilon)}\big(\mathbb{E}[(f(X) - f^*(X))^{2+\varepsilon}]\big)^{2/(2+\varepsilon)} \leqslant \frac{\|f - f^*\|_{L_{2+\varepsilon}}^2}{2(C')^2} \ .$$

By Assumption 2.9, it follows that $\mathbb{E}[I_{A^c}(X)(f(X) - f^*(X))^2] \leqslant \frac{\|f-f^*\|_{L_2}^2}{2}$ and we conclude with (2.41).

Chapter 3

Robust high dimensional learning for Lipschitz and convex losses

In this chapter, we establish risk bounds for Regularized Empirical Risk Minimizers (RERM) when the loss is Lipschitz and convex and the regularization function is a norm. In a first part, we obtain these results in the i.i.d. setup under subgaussian assumptions on the design. In a second part, a more general framework where the design might have heavier tails and data may be corrupted by outliers both in the design and the response variables is considered. In this situation, RERM performs poorly in general. We analyse an alternative procedure based on median-of-means principles and called "minmax MOM". We show optimal subgaussian deviation rates for these estimators in the relaxed setting. The main results are meta-theorems allowing a wide-range of applications to various problems in learning theory. To show a non-exhaustive sample of these potential applications, it is applied to classification problems with logistic loss functions regularized by LASSO and SLOPE, to regression problems with Huber loss regularized by Group LASSO. Another advantage of the minmax MOM formulation is that it suggests a systematic way to slightly modify descent based algorithms used in high-dimensional statistics to make them robust to outliers (Lecué and Lerasle, 2019). We illustrate this principle in a Simulations section where a " minmax MOM" version of classical proximal descent algorithms are turned into robust to outliers algorithms.

3.1 Introduction

Regularized empirical risk minimizers (RERM) are standard estimators in high dimensional classification and regression problems. They are solutions of minimization problems of a regularized empirical risk functions for a given loss and regularization functions. In regression, the quadratic loss of linear functionals regularized by the ℓ_1-norm (LASSO) (Tibshirani, 1996) is probably the most famous example of RERM, see for example (Koltchinskii, 2011b; Bühlmann and van de Geer, 2011; Giraud, 2015) for overviews. Recent results and references, including more general regularization functions can be found, for example in (Lecué and Mendelson, 2018; Bellec et al., 2017; Bach et al., 2012; Bhaskar et al., 2013; Argyriou et al., 2013). RERM based on the quadratic loss function are highly unstable when data have heavy-tails or when the dataset has been corrupted by outliers. These problems have attracted a lot of attention in robust statistics, see for example (Huber and Ronchetti, 2011) for an overview. By considering alternative losses, one can efficiently solve these problems when heavy-tails or corruption happen in the output variable Y. There is a growing literature analyzing performance of some of these alternatives in learning theory. In regression problems, among others, one can mention the L_1 absolute loss (Shalev-Shwartz and Tewari, 2011), the Huber loss (Zhou et al., 2018; Elsener and van de Geer, 2018) and the quantile loss (Alquier et al., 2019) that is popular in finance and econometrics. In classification, besides the 0/1 loss function which is known to lead to computationally intractable RERM, the logistic loss and the hinge loss are among the most popular convex surrogates (Zhang, 2004; Bartlett et al., 2006). Quantile, L_1, Huber loss functions for regression and Logistic, Hinge loss functions for classification are all Lipschitz and convex loss functions (in their first variable, see Assumption 3.2 for a formal definition). This remark motivated (Alquier et al., 2019) to study systematically RERM based on Lipschitz loss functions. A remarkable feature of Lipschitz losses proved in (Alquier et al., 2019) is that optimal results can be proved with almost no assumption on the response variable Y.

This paper is built on the approach initiated in (Chinot et al., 2019b). Compared with (Alquier et al., 2019), the approach of (Chinot et al., 2019b) improves the results by deriving risk bounds depending on a localized complexity parameters rather than global ones and by considering a more flexible setting where a global Bernstein condition is relaxed into a local one, see Assumption 3.5 and the following discussion for details. The paper (Chinot et al., 2019b) only considers estimators that are not regularized and that can therefore only be efficient in small dimensional settings.

The first main result of this paper is a high dimensional extension of the results in (Chinot et al., 2019b) that is achieved by analyzing estimators (based on the empirical risk or a Median-of-Means version) regularized by a norm. The main results are two meta-theorem allowing to study a broad range of estimators including LASSO, SLOPE, group LASSO and their minmax MOM version. Section 3.6 provides applications of the main results to some examples among these.

While RERM is studied without assumption on the output variables, somehow strong, albeit classical, hypotheses are granted on the design X in our first main result. We assume actually in this

analysis subgaussian assumptions on the input variables as in (Alquier et al., 2019). The necessity of this assumption to derive optimal exponential deviation bounds for RERM is not surprising as RERM have downgraded performance when the design is heavy tailed (see (Mendelson, 2014) or (Chinot et al., 2019b) for instance).

In a second part, we study an alternative to RERM in a framework with less stringent assumptions on the data. These estimators are based on the Median-Of-Means (MOM) principle (Nemirovsky and Yudin, 1983; Birgé, 1984; Jerrum et al., 1986; Alon et al., 1999) and the minmax approach (Audibert and Catoni, 2011; Baraud et al., 2017). They are called minmax MOM estimators as in (Lecué and Lerasle, 2019). A non-regularized version of these estimators was analyzed in (Chinot et al., 2019b). The second main and most important result of the paper shows that minmax MOM estimators achieve optimal subgaussian deviation bounds in the relaxed setting where RERM perform poorly because of outliers and heavy-tailed data. This result is obtained under a local Bernstein condition as for the RERM. It allows to derive fast rates of convergence in a large set of applications where typically, subgaussian assumptions on the design X are replaced by moment assumptions. Minmax MOM estimators are then analysed without the local Bernstein condition. Oracle inequalities holding with exponentially large probability are proved in this case. Compared with results under Bernstein's assumption, an extra variance term appears in the convergence rate. This extra term typically would yield to slow rates of convergence in the applications, which are known to be minimax in the case where no Bernstein assumption holds. However, the variance term disappears under the Bernstein's condition, which shows that fast rates can be recovered from the general results. In addition, all results on minmax MOM estimators, both with or without Bernstein condition, are shown in the "$\mathcal{O} \cup \mathcal{I}$" framework – where $\mathcal{O}$ stands for "outliers" and $\mathcal{I}$ for "informative"– see Section 3.4.1 or (Lecué and Lerasle, 2017, 2019) for details. In this framework, all assumptions (such as the Bernstein's condition) are granted on "inliers" $(X_i, Y_i)_{i \in \mathcal{I}}$. These inliers may have different distributions but the oracles of these distributions should match. On the other hand, no assumption are granted on outliers $(X_i, Y_i)_{i \in \mathcal{O}}$, which is to the best of our knowledge the strongest form of aggressive/adversarial outliers (it includes, in particular, Huber's ϵ-contamination setup). The minmax MOM estimators perform well in this setting, it means that the accuracy of their predictions is not downgraded by the presence of outliers in the dataset. Mathematically, this robustness is not surprising as it is a byproduct of the median step used in the MOM principle. However, in practice, it is an important advantage of MOM estimators compared to RERM.

The main results on minmax MOM estimators are also meta-theorems that can be applied to the same examples as RERM. Each of these examples provide a new (to the best of our knowledge) estimator that reach performance that RERM could not typically achieve. For example, when the class of classifiers/regressors is the class of linear functions on $\mathbb{R}^p$, minmax MOM estimators have a risk bounded by the minimax rate with optimal exponential probability of deviation even if the inputs X only satisfy weak moment assumptions and/or have been corrupted by outliers. These applications are also discussed in Section 3.6.

Finally, in Section 3.7, we consider the modification of standard algorithms suggested by the minmax MOM formulation introduced in (Lecué and Lerasle, 2019) to construct robust algorithms.

The paper is organized as follows. Section 3.2 presents the formal setting. Section 3.3 presents results for RERM and Section 3.4 those for minmax MOM estimators under a local Bernstein condition and in Section 3.5 without this condition. Section 3.6 details several examples of applications of the main results. A short simulation study illustrating our theoretical findings is presented in Section 3.7. The proofs are postponed to Sections 3.9.1- 3.9.3.

3.2 Mathematical background and notations

Let $(\mathcal{Z}, \mathcal{A}, P)$ denote a probability space, where $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ is a product space such that $\mathcal{X}$ denotes a measurable space of inputs and $\mathcal{Y} \subset \mathbb{R}$ is the set of values taken by the outputs. Let $Z = (X, Y)$ denote a random variable taking values in $\mathcal{Z}$ with distribution P and let μ denote the marginal distribution of the design X.

Let $\overline{\mathcal{Y}} \subset \mathbb{R}$ denote a convex set such that $\mathcal{Y} \subset \overline{\mathcal{Y}}$ and let F denote a class of functions $f : \mathcal{X} \to \overline{\mathcal{Y}}$. The set $\overline{\mathcal{Y}}$ is typically the co,vex hull of $\mathcal{Y}$. As such, it will always contain $\mathcal{Y}$. Let $\ell : \overline{\mathcal{Y}} \times \mathcal{Y} \to \mathbb{R}$ denote a loss function such that $\ell(f(x), y)$ measures the error made when predicting y by $f(x)$. For any distribution Q on $\mathcal{Z}$ and any function $g : \mathcal{Z} \to \mathbb{R}$ for which it makes sense, let $Qg = \mathbb{E}_{Z\sim Q}[g(Z)]$ denote the expectation of the function g under the distribution Q and, for any $p \geqslant 1$, let $\|g\|_{L_p(Q)} := (Q[|g|^p])^{1/p}$ and $\|g\|_{L_p} := \|g\|_{L_p(P)}$. The risk of any $f \in F$ is given by $P\ell_f$, where $\ell_f(x, y) := \ell(f(x), y)$. The prediction of Y with minimal risk is given by $f^*(X)$, where f^*, called *oracle*, is defined as any function such that

$$f^* \in \operatorname*{argmin}_{f\in F} P\ell_f \ .$$

Hereafter, for simplicity, it is assumed that f^* exists and is uniquely defined. The oracle is unknown to the statistician that has only access to a dataset $(X_i, Y_i)_{i\in\{1,\dots,N\}}$ of random variables taking values in $\mathcal{X} \times \mathcal{Y}$. The goal is to build a data-driven estimator $\hat{f}$ of f^* that predicts almost as well as f^*. The quality of an estimator $\hat{f}$ is measured by the error rate $\|\hat{f} - f^*\|_{L_2}^2$ and the excess risk $P\mathcal{L}_{\hat{f}}$, where, respectively,

$$\|\hat{f} - f^*\|_{L_2}^2 = P[(\hat{f} - f)^2] = \mathbb{E}\left[\left(\hat{f}(X) - f^*(X)\right)^2 |(X_i, Y_i)_{i=1}^N\right] \text{ and } \mathcal{L}_{\hat{f}} := \ell_{\hat{f}} - \ell_{f^*} \ . \tag{3.1}$$

Let P_N denote the empirical measure i.e $P_N(A) = (1/N)\sum_{i=1}^N I(Z_i \in A)$ for all $A \in \mathcal{A}$. A natural candidate for the estimation of f^* is the Empirical Risk Minimizer (ERM) of (Vapnik and Červonenkis, 1971), see also (Vapnik, 1998) for an overview, which is defined by

$$\hat{f}^{ERM} \in \operatorname*{argmin}_{f\in F} P_N\ell_f \ . \tag{3.2}$$

The choice of F is a central issue: enlarging the space F deteriorates the quality of the oracle estimation but improves its predictive performance. It is possible to use large classes F without significantly altering the quality estimation if certain structural properties of the oracle f^* are known a priori from the statistician. In that case, a widely spread approach is to add to the empirical loss a regularization term promoting this structural property. In this paper, we consider this problem when the regularization term is a norm. Formally, let E be a linear space such that $F \subset E \subset L_2(\mu)$ and let $\|\cdot\| : E \mapsto \mathbb{R}^+$ denote a norm on E. For any $\lambda \geq 0$, the regularized ERM (RERM) is defined by

$$\hat{f}_\lambda^{RERM} \in \operatorname*{argmin}_{f \in F} P_N \ell_f^\lambda, \quad \text{where} \quad \ell_f^\lambda(x,y) = \ell_f(x,y) + \lambda\|f\| \ . \tag{3.3}$$

In regression, one can mention Thikonov regularization which promotes smoothness (Golub et al., 1999) and ℓ_1 regularization which promotes sparsity (Tibshirani, 1996). Likewise, for matrix reconstruction, the 1-Schatten norm S_1 promotes low rank solutions (see (Koltchinskii et al., 2011; Cai et al., 2016)).

In the remaining of the paper, the following notations will be used repeatedly: for any $r > 0$, let

$$rB_{L_2} = \{f \in L_2(\mu) : \|f\|_{L_2} \leqslant r\}, \quad rS_{L_2} = \{f \in L_2(\mu) : \|f\|_{L_2} = r\} \ .$$

Let $rB = \{f \in E : \|f\| \leq r\}$ and $rS = \{f \in E : \|f\| = r\}$. For any set H for which it makes sense, let $H + f^* = \{h + f^* : h \in H\}$, $H - f^* = \{h - f^* : h \in H\}$. Let $(e_i)_{i=1}^p$ be the canonical basis of $\mathbb{R}^p$. Let c denote an absolute constant whose value might change from line to line and let $c(A)$ denote a function depending on the parameters A whose value may also change from line to line.

3.3 Regularized ERM with Lipschitz and convex loss functions

This section presents and improves results from (Alquier et al., 2019). A local Bernstein assumption, holding in a neighborhood of the *oracle* f^* is introduced in the spirit of (Chinot et al., 2019b). This assumption does not imply boundedness of F in L^2-norm unlike the global Bernstein condition considered in (Alquier et al., 2019). New rates of convergence are obtained, depending on **localized** complexity parameters improving the global ones from (Alquier et al., 2019).

3.3.1 Main assumptions

We start with a set of assumptions sufficient to prove exponential deviation bounds for the error rate and excess risk of RERM for general convex and Lipschitz loss functions and for any regularization norm. In this section, we consider the classical i.i.d. assumption (we will relax this assumption in the next sections in order to consider corrupted databases).

Assumption 3.1. $(X_i, Y_i)_{i=1}^N$ *are independent and identically distributed with distribution* P.

All along the paper, we consider Lipschitz and convex loss functions.

Assumption 3.2. *There exists $L > 0$ such that, for any $y \in \mathcal{Y}$, $\ell(\cdot, y)$ is L-**Lipschitz** i.e for every f and g in F, $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, $|\ell(f(x), y) - \ell(g(x), y)| \leq L|f(x) - g(x)|$ and **convex** i.e for all $\alpha \in [0, 1]$, $\ell(\alpha f(x) + (1-\alpha)g(x), y) \leq \alpha\ell(f(x), y) + (1-\alpha)\ell(g(x), y)$.*

There are many examples of loss functions satisfying Assumption 3.2. The two examples studied in this work (see Section 3.6) are

- the **logistic loss function** defined for any $u \in \mathbb{R}$ and $y \in \mathcal{Y} = \{-1, 1\}$, by $\ell(u, y) = \log(1 + \exp(-yu))$. It satisfies Assumption 3.2 for $L = 1$.
- Tte **Huber loss function** with parameter $\delta > 0$ is defined for all $u, y \in \mathbb{R}$, by

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

  It satisfies Assumption 3.2 for $L = \delta$.

We will also assume that the functions class F is convex.

Assumption 3.3. *The class F is convex.*

In particular, Assumption 3.3 holds in the important case considered in high-dimensional statistics when F is the class of all linear functions indexed by $\mathbb{R}^p$, $F = \{\langle t, \cdot \rangle : t \in \mathbb{R}^p\}$. This example is studied in great details in Section 3.6.

RERM performs well when the empirical excess risk $f \in F \to P_N\mathcal{L}_f$ is uniformly concentrated around the excess risk $f \in F \to P\mathcal{L}_f$. This requires strong concentration properties of the class of random variables $\{\mathcal{L}_f(X) : f \in F\}$, which is implied by concentration properties of $\{(f - f^*)(X) : f \in F\}$ thanks to the Lipschitz assumption on the loss function. Here, we study RERM under a subgaussian assumption on the design. We first recall the definition of a subgaussian class of functions.

Definition 3.1. *A class F is called L_0-subgaussian (with respect to X), where $L_0 \geq 1$, when for all f in F and for all $\lambda > 1$, $\mathbb{E}\exp(\lambda|f(X)|/\|f\|_{L_2}) \leq \exp(\lambda^2 L_0^2/2)$.*

Assumption 3.4. *The class $F - f^*$ is L_0-subgaussian with respect to X.*

Assumptions 3.1-3.4 are also granted in (Alquier et al., 2019). In this setup, a natural way to measure the statistical complexity of the problem is via Gaussian mean widths (of some subsets of F). We recall the definition of this measure of complexity.

Definition 3.2. *Let $H \subset L_2(\mu)$ and $(G_h)_{h\in H}$ be the canonical centered Gaussian process indexed by H, with covariance structure given by $(\mathbb{E}(G_{h_1} - G_{h_2})^2)^{1/2} = (\mathbb{E}(h_1(X) - h_2(X))^2)^{1/2}$ for all $h_1, h_2 \in H$. The **Gaussian mean-width** of H is $w(H) = \mathbb{E}\sup_{h\in H} G_h$.*

Gaussian mean widths of various sets have been computed in (Amelunxen et al., 2014), (C Bellec, 2019), (Chatterjee and Goswami, 2019) or (Gordon et al., 2007) for example. Risk bounds for $\hat{f}_\lambda^{RERM}$ are driven by fixed point solutions of a Gaussian mean width of regularization balls $(F - f^*) \cap \rho B$, which measure the local complexity of F around f^* .

Definition 3.3. *For all $A > 0$, the **complexity function** is a non-decreasing function $r(A, \cdot)$, such that for every $\rho \geq 0$,*

$$r(A, \rho) \geq \inf\{r > 0 : 96AL_0Lw\big(F \cap (f^* + \rho B \cap rB_{L_2})\big) \leq r^2\sqrt{N}\} \ .$$

Here, L is the Lipschitz constant in Assumption 3.2 and L_0 is the subgaussian constant from Assumption 3.4.

The last tool and assumption comes from (Lecué and Mendelson, 2018). A key observation is that the regularization norm $\|\cdot\|$ promoting some sparsity structure has large subdifferentials at sparse functions (see, for instance, atomic norms in (Bhaskar et al., 2013)). The subdifferential of $\|\cdot\|$ in f is defined as

$$(\partial\|.\|)_f = \{z^* \in E^* \ : \ \|f + h\| - \|f\| \geq z^*(h) \ \text{ for every } h \in E\} \ , \tag{3.4}$$

where E^* is the dual space of the normed space $(E, \|\cdot\|)$. Let

$$\Gamma_{f^*}(\rho) = \bigcup_{f \in f^* + \frac{\rho}{20}B} (\partial\|\cdot\|)_f$$

be the union of all subdifferentials of the regularization norm $\|\cdot\|$ of functions f close to the oracle f^*. We expect $\Gamma_{f^*}(\rho)$ to be a "large" subset of the unit dual sphere of $\|\cdot\|$ when f^* is "sparse" – for the notion of sparsity associated with $\|\cdot\|$. This intuition is formalized in the following definition from (Lecué and Mendelson, 2018)

Definition 3.4 ((Lecué and Mendelson, 2018))**.** *For any $A > 0$ and $\rho > 0$, let*

$$H_{\rho,A} = \{f \in F \ : \ \|f^* - f\| = \rho \ \text{ and } \ \|f^* - f\|_{L_2} \leq r(A, \rho), \ldots, \ \}.$$

Let

$$\Delta(\rho, A) = \inf_{h \in H_{\rho,A}} \sup_{z^* \in \Gamma_{f^*}(\rho)} z^*(h - f^*) \ . \tag{3.5}$$

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

Any constant in $(0, 1)$ could replace $4/5$ in Definition 3.4 as can be seen from a close inspection of the proof of Theorem 3.1. If the norm $\|\cdot\|$ is "smooth" in f, the subdifferential of $\|\cdot\|$ in f is just the gradient of $\|\cdot\|$ in f. In that case, $(\partial\|\cdot\|)_f$ is not rich (it is a singleton) and the regularization norm has only a low "sparsity inducing power" unless the variety of gradients of $\|\cdot\|$ at f in the neighborhood $f^* + (\rho/20)B$ is rich enough (the latter case can be seen as $\|\cdot\|$ being "almost not

differentiable" in f^*). However, any norm has a subdifferential in 0 equal to the entire unit dual ball associated with $\|\cdot\|$. Therefore, when 0 belongs to $f^* + (\rho/20)B$, for example when $\rho \geq 20\|f^*\|$, the sparsity equation is satisfied since, in that case, $\Delta(\rho) = \rho$. We can use this fact to obtain "complexity dependent" rates of convergence – i.e. rates depending on $\|f^*\|$. In high-dimensional setups, we also look for statistical bounds depending on the sparsity of f^* enforced by $\|\cdot\|$ (see (Lecué and Mendelson, 2017, 2018) for details regarding the difference between "complexity and sparsity" dependent bounds). Hereafter, we focus on norms $\|\cdot\|$ promoting some sparsity structure and we establish sparsity dependent rates of convergence and sparse oracle inequalities in Section 3.6.

Margin assumptions (Mammen and Tsybakov, 1999; Tsybakov, 2004; van de Geer, 2016) such as the Bernstein conditions from (Bartlett and Mendelson, 2006a) have been widely used in statistics and learning theory to prove fast convergence rates of RERM. Here, we use a **local Bernstein condition** in the spirit of (Chinot et al., 2019b).

Assumption 3.5. *There exist constants $A > 0$ and ρ^* such that ρ^* satisfies the A-sparsity equation and for all $f \in F$ satisfying $\|f - f^*\|_{L_2} = r(A, \rho^*)$ and $\|f - f^*\| \leq \rho^*$, then $\|f - f^*\|_{L_2}^2 \leq AP\mathcal{L}_f$.*

Hereafter, whenever Assumption 3.5 is granted, we assume that the constant A is fixed satisfying this assumption and write $r(\rho)$ instead of $r(A, \rho)$. As explained in (Chinot et al., 2019b), the local Bernstein condition holds in examples where F is not bounded in L_2-norm. It allows to cover the class of all linear functions on $\mathbb{R}^d$ where the global Bernstein condition of (Alquier et al., 2019) – $\|f - f^*\|_{L_2}^2 \leqslant AP\mathcal{L}_f$ for all $f \in F$– does not hold.

Remark 3.1. *From Assumption 3.2 it follows that if the local Bernstein condition is granted as in Assumption 3.5 that is for all functions f in F such that $\|f - f^*\|_{L_2} = r(A, \rho^*)$ and $\|f - f^*\| \leq \rho^*$ (and if there exists such an f) then we necessary have $r(A, \rho^*) \leq AL$. Indeed, if there is an f in $F \cap (f^* + r(A, \rho^*)S_{L_2} \cap \rho^* B)$, it follows from the Lipschitz property of the loss function that*

$$r^2(A, \rho^*) = \|f - f^*\|_{L_2}^2 \leq AP\mathcal{L}_f \leq AL\|f - f^*\|_{L_2} = ALr(A, \rho^*)$$

and so $r(A, \rho^*) \leq AL$. The latter condition will be always satisfied as soon as N is large enough. For example, for the LASSO regularization, we recover from the latter restriction, the classical condition "$N \gtrsim s\log(ep/s)$" where s is the oracle's sparsity.

3.3.2 Main theorem for the RERM

The following theorem gives the main result on the statistical performance of RERM.

Theorem 3.1. *Grant Assumptions 3.1, 3.2, 3.3, 3.4. Suppose that Assumption 3.5 holds with $\rho = \rho^*$ satisfying the A-sparsity equation from Definition 3.4. With this value of A, let $r(\cdot) := r(A, \cdot)$ denote the complexity function from Definition 3.3. Assume that*

$$\frac{10}{21A}\frac{r^2(\rho^*)}{\rho^*} < \lambda < \frac{2}{3A}\frac{r^2(\rho^*)}{\rho^*} \quad . \tag{3.6}$$

Then, with probability larger than

$$1 - 2\exp\big(-c(A, L, L_0)r^2(\rho^*)N\big) \ , \tag{3.7}$$

the following bounds hold

$$\|\hat{f}_\lambda^{RERM} - f^*\| \le \rho^*, \quad \|\hat{f}_\lambda^{RERM} - f^*\|_{L_2} \le r(\rho^*) \text{ and } P\mathcal{L}_{\hat{f}_\lambda^{RERM}} \le \frac{r^2(\rho^*)}{A} \ .$$

Remark 3.2. *A remarkable feature of Theorem 3.1 is that it holds without assumption on Y. We do not even need Y to be in L_1 since one can always fix some $f_0 \in F$ and work with $\ell_f - \ell_{f_0}$ to define all the object. In that case we have $|\ell_f - \ell_{f_0}| \le L|f - f^0|$ and so $(\ell_f - \ell_{f_0})(Z) \in L^1$ when $F \subset L^1(\mu)$ even when $Y \notin L^1$. So we can define f^* such that $f^* \in \operatorname{argmin}_{f\in F} P(\ell_f - \ell_{f_0})$ with no assumption on Y. This is an important consequence of the Lipschitz property which has been widely used in robust statistics because it implies robustness to heavy-tailed noise without any strong technical difficulty.*

Remark 3.3. *Theorem 3.1 holds for subgaussian classes of functions F. As in (Alquier et al., 2019), it is possible to extend this result under boundedness assumptions.*

Theorem 3.1 improves (Alquier et al., 2019, Theorem 2.1) in two directions: First, the complexity function $r(\cdot)$ measures the (Gaussian mean width) complexity of the **local** set $(F - f^*) \cap \rho B \cap rB_{L_2}$ and not the global gaussian mean width of $(F - f^*) \cap \rho B$ such as in (Alquier et al., 2019). Second, Theorem 3.1 holds in a setting where F can be unbounded in L_2-norm. The proof of Theorem 3.1 is postponed to Section 3.9.1. The proof relies on the convexity of the loss function (and F) which allows to use an homogeneity argument as in (Chinot et al., 2019b) for Lipshitz and convex loss functions and in (Lecué and Mendelson, 2013) for the quadratic loss function, simplifying the peeling step of (Alquier et al., 2019). Theorem 3.1 is a general result which is applied in various applications in Section 3.6.

3.4 Minmax MOM estimators

Even if the results of Section 3.3 are interesting on their own (because the i.i.d. sub-gaussian framework is one of the most considered setup in Statistics and Learning theory), the setup considered in Section 3.3 can be restrictive in some applications. It does not cover more realistic situations where data are heavy-tailed and/or corrupted. In this section, we consider a more general setup beyond the i.i.d. subgaussian setup in order to cover these more realistic frameworks. The results from Section 3.3 will serve as benchmarks: we show that similar bounds can be achieved in a more realistic framework by alternative estimators. These estimators use the median-of-means principles instead of empirical means.

3.4.1 Definition

Recall the definition of MOM estimators of univariate means from (Alon et al., 1999; Jerrum et al., 1986; Nemirovsky and Yudin, 1983). Let $(B_k)_{k=1,\dots,K}$ denote a partition of $\{1,\dots,N\}$ into blocks B_k of equal size N/K (it is implicitly assumed that K divides N. An extension to blocks with almost equal size is possible (see (Minsker et al., 2019)). It is not considered here to simplify the presentation of the results, the extension is thus left to the interested reader). For any function $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ and $k \in \{1,\dots,K\}$, let $P_{B_k} f = (K/N) \sum_{i \in B_k} f(X_i, Y_i)$ denote the empirical mean on the block B_k. The MOM estimator based on this partition is the empirical median of the latter empirical means:

$$\mathrm{MOM}_K\big(f\big) = \mathrm{Med}(P_{B_1} f, \cdots, P_{B_K} f) \ . \tag{3.8}$$

The estimator $\mathrm{MOM}_K\big(f\big)$ of Pf achieves subgaussian deviation tails if $(f(X_i, Y_i))_{i=1}^N$ have 2 moments, see (Devroye et al., 2016). The number of blocks K is a tuning parameter of the procedure. The larger K, the more outliers are allowed. When $K = 1$, $\mathrm{MOM}_K\big(f\big)$ is the empirical mean, when $K = N$, it is the empirical median.

Building on ideas introduced in (Audibert and Catoni, 2011; Baraud et al., 2017), (Lecué and Lerasle, 2019) proposed the following strategy to use MOM estimators in learning problems. Since the *oracle* f^* is also solution of the following minmax problem

$$f^* = \operatorname*{argmin}_{f \in F} P\ell_f = \operatorname*{argmin}_{f \in F} \sup_{g \in F} P(\ell_f - \ell_g) \ ,$$

minmax MOM estimators are obtained by plugging MOM estimators of the unknown expectations $P(\ell_f - \ell_g)$ in this minmax formulation. Applying this principle to regularized procedures yields the following "minmax MOM version" of RERM that we study in this paper:

$$\hat{f}_{K,\lambda} \in \operatorname*{argmin}_{f \in F} \sup_{g \in F} \mathrm{MOM}_K\big(\ell_f - \ell_g\big) + \lambda\big(\|f\| - \|g\|\big) \ . \tag{3.9}$$

The linearity of the empirical process P_N is important to use localization techniques in the analysis of RERM to derive fast rates of convergence for these estimators improving upon the slow rates of (Vapnik, 1998), see (Tsybakov, 2004; Koltchinskii, 2011b) for example. The minmax reformulation comes from (Audibert and Catoni, 2011), it allows to overcome the lack of linearity of robust mean estimators and obtain fast rates of convergence for robust estimators based on nonlinear estimators of univariate expectations.

3.4.2 Assumptions and main results

To highlight robustness properties of minmax MOM estimators with respect to outliers in the dataset, their analysis is performed in the following framework. Let $\mathcal{I} \cup \mathcal{O}$ denote a partition of $\{1, \cdots, N\}$ that is unknown to the statistician. Data $(X_i, Y_i)_{i \in \mathcal{O}}$ are considered as outliers. **No assumption** on the distribution of these data is made, they can be dependent or adversarial. Data

$(X_i, Y_i)_{i\in\mathcal{I}}$ bring information on f^* and are called informative or inliers. Assumptions are made uniquely on these informative data (and not on the outliers). They have to induce the same L_2 geometries on F and the same excess risks.

Assumption 3.6. *$(X_i, Y_i)_{i\in\mathcal{I}}$ are independent and for all $i \in \mathcal{I}$: $P_i(f - f^*)^2 = P(f - f^*)^2$ and $P_i\mathcal{L}_f = P\mathcal{L}_f$.*

Assumption 3.6 holds in the i.i.d case, it also covers situations where informative data $(X_i, Y_i)_{i\in I}$ may have different distributions. It implies in particular that f^* is also the oracle in F w.r.t. all the distributions P_i for $i \in \mathcal{I}$.
Several quantities introduced to study RERM have to be modified to state the results for minmax MOM estimators. First, the complexity function is no longer based on Gaussian mean width, it is now defined as a fixed point of local Rademacher complexities (Koltchinskii, 2011a, 2006; Bartlett et al., 2002b, 2005). Let $(\sigma_i)_{i\in\mathcal{I}}$ denote i.i.d. Rademacher random variables (i.e. uniformly distributed on $\{-1, 1\}$), independent from $(X_i, Y_i)_{i\in\mathcal{I}}$. The **complexity function** $\rho \to r_2(\gamma, \rho)$ is a non-decreasing function such that for all $\rho > 0$

$$r_2(\gamma,\rho) \geq \inf\left\{ r>0 : \forall J \subset \mathcal{I} \text{ s.t } |J| \geqslant N/2, \quad \mathbb{E}\left\{ \sup_{f\in(F-f^*)\cap\rho B\cap rB_{L_2}} \left|\sum_{i\in J}\sigma_i f(X_i)\right| \right\} \leq \gamma r^2|J| \right\} . \tag{3.10}$$

As in Theorem 3.1, parameter $r_2(\gamma, \rho)$ measures the statistical complexity of the sub-model $F \cap (f^* + \rho B)$ locally in a L_2-neighborhood of f^*. It only involves the distribution of informative data and does not depend on the distribution of the outputs $(Y_i)_{i\in\mathcal{I}}$. The local Bernstein condition, Assumption 3.5, as well as the sparsity equation have now to be extended to this new definition of complexity. We start with the sparsity equation.

Definition 3.5. *For any $A > 0$ and $\rho > 0$, let*

$$C_{K,r}(\rho, A) = \max\left(r_2^2(\gamma,\rho), c(A, L)\frac{K}{N} \right) \tag{3.11}$$

and $\tilde{H}_{\rho,A} = \{f \in F \ : \ \|f^* - f\| = \rho \text{ and } \|f^* - f\|_{L_2} \leq \sqrt{C_{K,r}(\rho, A)}, \dots,\}$. Let

$$\tilde{\Delta}(\rho, A) = \inf_{h\in\tilde{H}_{\rho,A}} \sup_{z^*\in\Gamma_{f^*}(\rho)} z^*(h - f^*) \ . \tag{3.12}$$

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

The value of $c(A, L)$ in Definition 3.5 is made explicit in Section 3.9.2. To simplify the presentation we write $c(A, L)$ as it is an absolute constant depending only on A and L. With this definition in mind, one can extend the local Bernstein assumption.

Assumption 3.7. *There exist a constant $A > 0$ and ρ^* such that ρ^* satisfies the A-sparsity equation from Definition 3.5 and, for all $f \in F$ such that $\|f - f^*\|_{L_2}^2 = C_{K,r}(2\rho^*, A)$ and $\|f - f^*\| \leq 2\rho^*$, $\|f - f^*\|_{L_2}^2 \leqslant AP\mathcal{L}_f$.*

As in Assumption 3.5, the link between $\|f - f^*\|_{L_2}^2$ and the excess risk $P\mathcal{L}_f$ in Assumption 3.7 is only granted in a $L_2(\mu)$-sphere around the oracle f^* whose radius is proportional to the rate of convergence of the estimators (see Theorems 3.1 and 3.2). The local Bernstein assumption is somehow "minimal" since it is only granted on the smallest set of the form $F \cap (f^* + 2\rho^* B \cap r_2(\gamma, 2\rho^*)B_{L_2})$ centered in f^* that can be proved to contain $\hat{f}_{K,\lambda}$ (when K is such that $\sqrt{C_{K,r}(2\rho^*, A)} = r_2(\gamma, 2\rho^*)$).

Remark 3.4. *As in Remark 3.1 we necessary have $\sqrt{C_{K,r}(2\rho^*, A)} \leq AL$ under Assumption 3.7 and the Lipschitz assumption from Assumption 3.2. This is also this condition which requires a minimal number of observations to hold out of which we recover the classical conditions such as $N \gtrsim s\log(ep/s)$ when one wants to reconstruct a s-sparse vector.*

We are now in position to state our main result on the statistical performances of the regularized minmax MOM estimator.

Theorem 3.2. *Grant Assumptions 3.2, 3.3, 3.6 and 3.7 for ρ^* satisfying the A-sparsity equation from Definition 3.5. Let $K \geq 7|\mathcal{O}|/3$, $\gamma = 1/(6528L)$, and define*

$$\lambda = \frac{5}{17A}\frac{C_{K,r}(2\rho^*, A)}{\rho^*} \ .$$

Then, with probability larger than $1 - 2\exp(-cK)$, the minmax MOM estimator $\hat{f}_{K,\lambda}$ defined in (3.9) *satisfies*

$$\|\hat{f}_{K,\lambda} - f^*\| \leq 2\rho^*, \quad \|\hat{f}_{K,\lambda} - f^*\|_{L_2}^2 \leq C_{K,r}(2\rho^*, A) \quad \textit{and} \quad P\mathcal{L}_{\hat{f}_{K,\lambda}} \leq \frac{1}{A}C_{K,r}(2\rho^*, A) \ .$$

Suppose that $K = c(A, L)r_2^2(\gamma, 2\rho^*)N$, which is possible as long as $|\mathcal{O}| \leq c(A, L)Nr_2^2(\gamma, 2\rho^*)$. The L_2-estimation bound obtained in Theorem 3.2 is then $r_2^2(\gamma, 2\rho^*)$ and the probability that this bound holds is $1 - \exp(-c(A, L)Nr_2^2(\gamma, 2\rho^*))$. Up to absolute constants, regularized minmax MOM estimators achieve the same bounds as RERM with the same probability when the inlier data satisfy the subgaussian assumption as in the framework of Theorem 3.1. Indeed, in that case, a straightforward chaining argument shows that the Rademacher complexity from (3.10) is upper bounded by the Gaussian mean width. The difference with Theorem 3.1 is that the estimator depends on K. On the other hand, the results from Theorem 3.2 hold in a setting beyond the subgaussian assumption on F and the data may not be identically distributed and may have been corrupted by outliers. In Section 3.6.2, we consider an example where rate optimal bounds can be derived from this general result under weak moment assumptions while still achieving the same rate as in the sub-gaussian framework. It is also possible to adapt in a data-driven way to the best K and λ by using a Lepski's adaptation method such as in (Devroye et al., 2016; Lecué and Lerasle, 2017, 2019; Chinot et al., 2019b; Chinot, 2019b). This step is now well understood, it is not reproduced here. Theorem 3.2 is general result in the sense that it allows to handle many applications where a convex and Lipschitz loss function and a regularization norm are used (some examples are presented in Section 3.6).

3.5 Relaxing the Bernstein condition

In this section, we study minmax MOM estimators when the Bernstein assumption 3.7 is relaxed. The price to pay for this relaxation is that, on one hand, the L_2-risk is not controlled and on the other hand an extra variance term appears in the excess risk $P\mathcal{L}_{\hat{f}_K^\lambda}$. Nevertheless, under a slightly stronger local Bernstein's condition, the extra variance term can be controled and the bounds from Theorem 3.2 can be recovered. We consider the following assumption which is weaker than Assumption 3.6 since it does not require that the distribution of the X_i's, for $i \in \mathcal{I}$ induce the same L_2 structure as the one of $L_2(\mu)$.

Assumption 3.8. *$(X_i, Y_i)_{i\in\mathcal{I}}$ are independent and for all $i \in \mathcal{I}$, (X_i, Y_i) has distribution P_i, X_i has distribution μ_i. We assume that, for any $i \in \mathcal{I}$, $F \subset L_1(\mu_i)$ and $P_i\mathcal{L}_f = P\mathcal{L}_f$ for all $f \in F$.*

Since the local Bernstein Assumption 3.7 does not hold, the localization argument has to be modified. Instead of using the L_2-norm to define neighborhoods of f^* as in the previous section, we use the excess loss $f \in F \to P\mathcal{L}_f$ as proximity function defining the neighborhoods. The new fixed point is defined for all $\gamma, \rho > 0$ and $K \in \{1, \cdots, N\}$:

$$\bar{r}(\gamma,\rho) = \inf\left\{ r > 0 : \max\left(\frac{E(r,\rho)}{\gamma}, \sqrt{c}V_K(r,\rho)\right) \le r^2 \right\}, \quad \text{where} \tag{3.13}$$

$$E(r,\rho) = \sup_{J\subset\mathcal{I}:|J|\ge N/2} \mathbb{E} \sup_{f\in F: P\mathcal{L}_f\le r^2,\, \|f-f^*\|\le\rho} \left| \frac{1}{|J|}\sum_{i\in J} \sigma_i (f-f^*)(X_i) \right| ,$$

$$V_K(r,\rho) = \max_{i\in\mathcal{I}} \sup_{f\in F: P\mathcal{L}_f\le r^2,\, \|f-f^*\|\le\rho} \left(\sqrt{\mathbb{V}ar_{P_i}(\mathcal{L}_f)}\right) \sqrt{\frac{K}{N}} ,$$

and $(\sigma_i)_{i\in\mathcal{I}}$ are i.i.d. Rademacher random variables independent from $(X_i, Y_i)_{i\in\mathcal{I}}$. The value of c in Equation (3.13) can be found in Section 3.9.3. The main differences between $r_2(\gamma,\rho)$ in (3.10) and $\bar{r}(\gamma,\rho)$ in (3.13) are the extra variance V_K term and the L_2 localization which is replaced by an "excess of risk" localization. Under the local Bernstein Assumption 3.9 below, this extra variance term $V_K(r,\rho)$ becomes negligible in front of the complexity term $E(r,\rho)$. In that case, the fixed point $\bar{r}(\gamma, r)$ matches the $r_2(\gamma,\rho)$ used in Theorem 3.2. As in Section 3.4, the sparsity equation has to be modified according to this new definition of fixed point.

Definition 3.6. *For any $\rho > 0$, let*

$$\bar{H}_\rho = \{ f \in F \ : \ \|f^* - f\| = \rho \ \text{and} \ P\mathcal{L}_f \le \bar{r}^2(\gamma,\rho), \dots, \ \}. \tag{3.14}$$

Let

$$\bar{\Delta}(\rho) = \inf_{h\in\bar{H}_\rho} \sup_{z^*\in\Gamma_{f^*}(\rho)} z^*(h - f^*) \ . \tag{3.15}$$

*A real number $\rho > 0$ satisfies the **sparsity equation** if $\bar{\Delta}(\rho) \ge 4\rho/5$.*

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

Theorem 3.3. *Grant Assumptions 3.2, 3.3, 3.8 and assume that $|\mathcal{O}| \leq 3N/7$. Let ρ^* satisfying the sparsity equation from Definition 3.6. Let $\gamma = 1/(3840L)$ and $K \in \big[7|\mathcal{O}|/3, N\big]$. Define*

$$\lambda = \frac{11}{40}\frac{\bar{r}^2(\gamma, 2\rho^*)}{\rho^*}$$

The minmax MOM estimator $\hat{f}_{K,\lambda}$ defined in (3.9) satisfies, with probability at least $1-2\exp(-cK)$,

$$P\mathcal{L}_{\hat{f}_{K,\lambda}} \leq \bar{r}^2(\gamma, 2\rho^*) \quad \text{and} \quad \|\hat{f}_{K,\lambda} - f^*\| \leq 2\rho^* \ .$$

In Theorem 3.3, the only stochastic assumption is Assumption 3.8 which says that the inliers data are independent and define the same excess risk as (X, Y) over F. In particular, Theorem 3.3 does not assume anything on the outliers $(X_i, Y_i)_{i\in\mathcal{O}}$ nor on the outputs of the inliers $(Y_i)_{i\in\mathcal{I}}$ like in the previous section but it also does not require any other assumption than the existence of all the considered objects. It follows from Theorem 3.3 that all the difficulty of the problem is now contained in the computation of the local Rademacher complexities $E(r, \rho)$.

To conclude the section, let us show that Theorem 3.2 can be recovered from Theorem 3.3 under the following local Bernstein assumption which is slightly stronger than the one assumed in Theorem 3.3.

Assumption 3.9. *There exist a constant $\bar{A} > 0$ and ρ^* satisfying the sparsity equation from Definition 3.6 such that, for all $f \in F$, if $P\mathcal{L}_f \leqslant \bar{C}_{K,r}(\rho^*, \bar{A})$ and $\|f - f^*\| \leq 2\rho^*$, then $\|f - f^*\|_{L_2}^2 \leqslant \bar{A}P\mathcal{L}_f$, where*

$$\bar{C}_{K,r}(\rho, A) = \max\left(\frac{r_2^2(\gamma/A, 2\rho)}{\sqrt{A}}, c(A, L)\frac{K}{N}\right) \quad \text{and} \quad \gamma = 1/(3840L) \ . \tag{3.16}$$

Up to constants, $\bar{C}_{K,r}$ is equivalent to $C_{K,r}$ given in Definition 3.5. Assumption 3.9 is a condition on all functions $f \in F$ such that $P\mathcal{L}_f \leq \bar{C}_{K,r}(\rho^*, \bar{A})$ which is a slightly stronger condition than being in the L_2-sphere as in Assumption 3.7.

Theorem 3.4. *Grant Assumptions 3.2, 3.3, 3.6 and assume that $|\mathcal{O}| \leq 3N/7$. Assume that the local Bernstein condition Assumption 3.9 holds with ρ^* satisfying the $\bar{A}$-sparsity equation from Definition 3.6. Let $\gamma = 1/(3840L)$ and $K \in \big[7|\mathcal{O}|/3, N\big]$. Define*

$$\lambda = \frac{11}{40}\frac{\bar{r}^2(\gamma, 2\rho^*)}{\rho^*} \ .$$

The minmax MOM estimator $\hat{f}_{K,\lambda}$ defined in (3.9) satisfies, with probability at least $1-2\exp(-cK)$,

$$||\hat{f}_{K,\lambda} - f^*||_{L_2}^2 \leq \bar{C}_{K,r}(\rho^*, \bar{A}), \quad P\mathcal{L}_{\hat{f}_{K,\lambda}} \leq \bar{C}_{K,r}(\rho^*, \bar{A}) \quad \text{and} \quad \|\hat{f}_{K,\lambda} - f^*\| \leq 2\rho^* \ .$$

Theorem 3.4 is proved in Section 3.9.4.

Remark 3.5. *Under Assumption 3.9 and a slight modification in the constants, ρ^* satisfies the sparsity equation of Definition 3.6 if it verifies the sparsity equation of Definition 3.5.*

3.6 Applications

This section presents some applications of Theorem 3.2 to derive statistical properties of regularized minmax MOM estimators for various choices of loss functions and regularization norm. To check the assumptions of the Theorem 3.2, the following routine is applied:

1. Check Assumptions 3.2, 3.3, 3.6.
2. Compute the local rademacher complexity $r_2(\gamma, \rho)$.
3. Solve the sparsity equation from Definition 3.5: find ρ^* such that $\Delta(\rho^*, A) \geq 4\rho^*/5$.
4. Check the local Bernstein condition from Assumption 3.7.

In this section, we focus on high dimensional statistical problems with sparsity inducing regularization norms (Bach et al., 2012) such as the ℓ_1 norm (Tibshirani, 1996), the SLOPE norm (Bogdan et al., 2015), the group LASSO norm (Simon et al., 2013). We consider the class of linear functions $F = \{\langle t, \cdot \rangle : t \in \mathbb{R}^p\}$ indexed by $\mathbb{R}^p$. We denote by $t^* \in \mathbb{R}^p$ the vector such that $f^*(\cdot) = \langle t^*, \cdot \rangle$. We consider the logistic loss function for the LASSO and the SLOPE, with data $(X_i, Y_i)_{i=1}^N$ taking values in $\mathbb{R}^p \times \{-1, 1\}$ and the Huber loss function for the Group LASSO, with data $(X_i, Y_i)_{i=1}^N$ taking values in $\mathbb{R}^p \times \mathbb{R}$. In particular, the results of this section extend results on the logistic LASSO and logistic SLOPE from (Alquier et al., 2019) and present new results for the Group Lasso.

3.6.1 Preliminary tools and results

In this section, we recall some tools to check the Local Bernstein condition, compute the local Rademacher complexity and verify the sparsity equation.

Local Bernstein conditions for the logistic and Huber loss functions

In this section, we recall some results from (Chinot et al., 2019b) on the local Bernstein condition for the logistic and Huber loss functions.

For the logistic loss function (i.e. $\ell_f : (x, y) \in \mathbb{R}^p \times \{\pm 1\} \to \log(1 + \exp(-yf(x)))$), we first introduce the following assumption. Note that we do not use the full strength of the approach since we check the inequality $\|f - f^*\|_{L_2}^2 \leqslant AP\mathcal{L}_f$ for all $f \in F \cap (f^* + rB_{L_2})$ instead of just all functions in $F \cap (f^* + rS_{L_2} \cap \rho B)$.

Assumption 3.10. *Let $\varepsilon > 0$, there are constants C' and $c_0 > 0$ such that*

a) for all f in F, $\|f - f^*\|_{L_{2+\varepsilon}} \leq C'\|f - f^*\|_{L_2}$

b) $\mathbb{P}(|f^*(X)| \leq c_0) \geq 1 - 1/(2C')^{(4+2\varepsilon)/\varepsilon}$

Under Assumption 3.10, we check the Bernstein condition on the entire L_2-ball of radius r around f^*.

Proposition 3.1 ((Chinot et al., 2019b), Theorem 9)**.** *Grant Assumption 3.10. Let $r > 0$. The local Bernstein condition holds for the logistic loss function: for all $f \in F$ if $\|f - f^*\|_{L_2} \leqslant r$ then $\|f - f^*\|_{L_2}^2 \leqslant AP\mathcal{L}_f$ for*

$$A = \frac{\exp\big(-c_0 - r(2C')^{(2+\varepsilon)/\varepsilon}\big)}{2\Big(1 + \exp\big(c_0 + r(2C')^{(2+\varepsilon)/\varepsilon}\big)\Big)^2} \ .$$

Note that if r is larger than the order of a constant then A is no longer a constant and the convergence rates are deteriorated (see the link with Remark 3.1). So that we will assume that $r(2C')^{(2+\varepsilon)/\varepsilon} \leq c_0/2$ in order to keep A like an absolute constant. The price to pay for assuming this latter condition is on the number of observations: we will for instance recover the classical assumption $N \gtrsim s\log(ep/s)$ for the reconstruction of a s-sparse vector.

For the Huber loss function with parameter $\delta > 0$ (i.e. $\ell_f(x,y) = \rho_\delta(y - f(x))$ where $\rho_\delta(t) = t^2/2$ if $|t| \leq \delta$ and $\rho_\delta(t) = \delta|t| - \delta^2/2$ if $|t| \geq \delta$), we use the following result also borrowed from (Chinot et al., 2019b). Let us introduce the following assumption.

Assumption 3.11. *Let $\varepsilon > 0$ and let $F_{Y|X=x}$ be the conditional cumulative function of Y given $X = x$.*

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

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

Note that if r is larger than the order of a constant the point b) can be verified only if δ, the Lipschitz constant, is large enough and α is small enough. In that case, convergence rates would be degraded. To avoid this situation we assume that $r(\sqrt{2}C')^{(2+\varepsilon)/\varepsilon} \leq c$ where c is some absolute constant. In that case, δ and α can be considered like constants. Again the price we pay for that assumption will be on the number of observations such as the classical one $N \gtrsim s\log(ep/s)$ for the reconstruction of a s-sparse vector.

Proposition 3.2 ((Chinot et al., 2019b), Theorem 7)**.** *Grant Assumption 3.11 for $r > 0$. The Huber loss function with parameter $\delta > 0$ satisfies the Bernstein condition: for all $f \in F$, if $\|f - f^*\|_{L_2} \leq r$ then $(4/\alpha)P\mathcal{L}_f \geq \|f - f^*\|_{L_2}^2$.*

Local Rademacher complexities and Gaussian mean widths

The computation of $r_2(\gamma, \rho)$ may be involved, but can sometimes be reduced to the computation of Gaussian mean widths. A typical result in that direction is the one from (Mendelson, 2017). The results of (Mendelson, 2017) are based on the concepts of unconditional norm and isotropic random vectors.

Definition 3.7. *For a given vector $x = (x_i)_{i=1}^p$, let $(x_i^*)_{i=1}^p$ be the non-increasing rearrangement of $(|x_i|)_{i=1}^p$. The norm $\|\cdot\|$ in $\mathbb{R}^p$ is said κ-unconditional with respect to the canonical basis $(e_i)_{i=1}^p$ if, for every x in $\mathbb{R}^p$ and every permutation π of $\{1, \cdots, p\}$,*

$$\left\| \sum_{i=1}^p x_i e_i \right\| \leq \kappa \left\| \sum_{i=1}^p x_{\pi(i)} e_i \right\| ,$$

and, for any $y \in \mathbb{R}^p$ such that, for all $1 \leq i \leq p$, $x_i^* \leq y_i^*$, then

$$\left\| \sum_{i=1}^p x_i e_i \right\| \leq \kappa \left\| \sum_{i=1}^p y_i e_i \right\| .$$

Typical examples of κ-unconditional norms can be found in (Mendelson, 2017). In the following we use the fact that the dual norms of the ℓ_1 and SLOPE norms are 1-unconditional.

Definition 3.8. *A random vector X in $\mathbb{R}^p$ is isotropic if $\mathbb{E}[\langle t, X\rangle^2] = \|t\|_2^2$, for all $t \in \mathbb{R}^p$, where $\|\cdot\|_2$ is the Euclidean norm in $\mathbb{R}^p$.*

Recall the main result of (Mendelson, 2017).

Theorem 3.5. *(Mendelson, 2017, Theorem 1.6) Let C_0, κ and M be real numbers. Let $V \subset \mathbb{R}^p$ be such that $\sup_{v \in V} |\langle v, \cdot\rangle|$ is κ-unconditional with respect to $(e_i)_{i=1}^p$. Assume that $X \in \mathbb{R}^p$ is isotropic and satisfies, for all $1 \leq j \leq p$ and $1 \leq q \leq C_0 \log(p)$,*

$$\left\| \langle X, e_j \rangle \right\|_{L_q} \leq M\sqrt{q} \ . \tag{3.17}$$

Let $X_1, \ldots, X_N$ denote independent copies of X, then there exists a constant c_2 depending only on C_0 and M such that

$$\mathbb{E}\left\{ \sup_{v \in V} \sum_{i=1}^N \sigma_i \langle X_i, v\rangle \right\} \leq c_2 \kappa \sqrt{N} w(V)$$

where $w(V)$ is the Gaussian mean width of V.

Recall that a real valued random variable Z is L_0-subgaussian if and only if for all $q \geq 1$, $\|Z\|_{L_q} \leq c_0 L_0 \sqrt{q}$, for some absolute constant c_0, see Theorem 1.1.5 in (Chafaï et al., 2012). Hence, Theorem 3.5 shows that $C_0 \log(p)$ "subgaussian" moments for the coordinates of the design X are enough to upper bound the Rademacher complexity by the Gaussian mean width. Such a result is useful to show that minmax MOM estimators can achieve the same rate as the ERM (in the subgaussian framework) even when the data are heavy-tailed data.

Sub-differential of a norm

To solve the sparsity equation – find ρ^* such that $\tilde{\Delta}(\rho^*, A) \geq 4\rho^*/5$ – from Definition 3.5, we use the following classical result on the sub-differential of a norm: if $\|\cdot\|$ is a norm on $\mathbb{R}^p$, then, for all $t \in \mathbb{R}^p$, we have

$$(\partial \|\cdot\|)_t = \begin{cases} \{z^* \in S^* : \langle z^*, t\rangle = \|t\|\} & \text{if } t \neq 0 \\ B^* & \text{if } t = 0 \end{cases} . \tag{3.18}$$

Here, B^* is the unit ball of the dual norm associated with $\|\cdot\|$, i.e. $t \in \mathbb{R}^p \to \|t\|^* = \sup_{\|v\|\leq 1}\langle v, t\rangle$ and S^* is its unit sphere. In other words, when $t \neq 0$, the sub-differential of $\|\cdot\|$ in t is the set of all vectors z^* in the unit dual sphere S^* which are norming for t (i.e. z^* is such that $\langle z^*, t\rangle = \|t\|$). In particular, when $t \neq 0$, $(\partial\|\cdot\|)_t$ is a subset of the dual sphere S^*.

In the following, understanding the sub-differentials of the regularization norm is a key point for solving the sparsity equation. If one is only interested in proving "complexity" dependent bounds – which are bounds depending on $\|t^*\|$ and not on the sparsity of t^* – then one can simply take $\rho^* = 20\|t^*\|$. Actually, in this case, $0 \in \Gamma_{t^*}(\rho)$, so $\tilde{\Delta}(\rho^*, A) = \rho^* \geq 4\rho^*/5$ (because $B^* = (\partial\|\cdot\|)_0 = \Gamma_{t^*}(\rho)$ according to (3.18)). Therefore, understanding the sub-differential of the regularization norm matters when one wants to derive statistical bounds depending on the dimension of the low-dimensional structure that contains t^*. This is something expected since a norm has sparsity inducing power if its sub-differential is a "large" subset of the dual sphere at vectors having the sparse structure (see, for instance, the construction of atomic norms in (Bhaskar et al., 2013)).

We now have all the necessary tools to derive statistical bounds for many procedures by applying Theorem 3.2. In each example (given by a convex and Lipschitz loss function and a regularization norm), we just have to compute the complexity function r_2, solve a sparsity equation and check the local Bernstein condition.

3.6.2 The minmax MOM logistic LASSO procedure

When the dimension p of the problem is large and $\|t^*\|_0 = |\{i \in \{1, \cdots, p\} : t_i^* \neq 0\}|$ is small, it is possible to derive error rate depending on the size of the support of t^* instead of the dimension p by using a ℓ_1 regularization norm. It leads to the well-known LASSO estimators, see (Tibshirani, 1996; Bickel et al., 2009). For the logistic loss function, its minmax MOM formulation is the following. For a given $K \in \{1, \ldots, N\}$ and $\lambda > 0$, the minmax MOM logistic LASSO procedure is defined by

$$\hat{t}_{\lambda,K} \in \operatorname*{argmin}_{t\in\mathbb{R}^p} \sup_{\tilde{t}\in\mathbb{R}^p} \left(\mathrm{MOM}_K\big(\ell_t - \ell_{\tilde{t}}\big) + \lambda(\|t\|_1 - \|\tilde{t}\|_1) \right) ,$$

with the logistic loss function defined as $\ell_t(x, y) = \log(1 + \exp(-y\langle x, t\rangle))$ for all $t, x \in \mathbb{R}^p$ and $y \in \{\pm 1\}$, and with the ℓ_1 regularization norm defined for all $t \in \mathbb{R}^p$ by $\|t\|_1 = \sum_{i=1}^p |t_i|$.

We first compute the complexity function r_2. Theorem 3.5 can be applied to upper bound the Rademacher complexities from (3.10) in that case because the dual norm of ℓ_1-norm (i.e the ℓ_∞-norm) is 1-unconditional with respect to $(e_i)_{i=1}^p$. Then, if X is an isotropic random vector satisfying (3.17), Theorem 3.5 holds and

$$\mathbb{E} \sup_{t\in\rho B_1^p \cap r B_2^p} \left| \sum_{j\in J} \sigma_j \langle t, X_j\rangle \right| \leq c(C_0, M)\sqrt{|J|}\, w(\rho B_1^p \cap r B_2^p) ,$$

where B_1^p denote the unit ball of the ℓ_1 norm. From (Lecué and Mendelson, 2018, Lemma 5.3), we

have

$$w(\rho B_1^p \cap r B_2^p) \leq c \begin{cases} r\sqrt{p} & \text{if } r \leq \rho/\sqrt{p} \\ \rho\sqrt{\log(ep\min(r^2/\rho^2, 1))} & \text{if } r \geq \rho/\sqrt{p} \end{cases} . \tag{3.19}$$

Therefore, one can take

$$r_2^2(\gamma, \rho) = c(\gamma, C_0, M) \begin{cases} \frac{p}{N} & \text{if } N\rho^2 \geq c(\gamma, C_0, M)\gamma p^2 \\ \rho\sqrt{\frac{1}{N}\log\left(\frac{ep^2}{\rho^2 N}\right)} & \text{if } \log p \leq c(\gamma, C_0, M)N\rho^2 \leq c(\gamma, C_0, M)p^2 \\ \rho\sqrt{\frac{\log p}{N}} & \text{if } \log p \geq c(\gamma, C_0, M)N\rho^2. \end{cases} . \tag{3.20}$$

Let us turn to the local Bernstein assumption. We need to verify Assumption 3.10. Let $\varepsilon > 0$. If X is an isotropic random vector satisfying (3.17) and $C_0 \log(p) \geq 2 + \varepsilon$, where C_0 is the constant appearing in Theorem 3.5, then the point a) of Assumption 3.10 is verified with $C' = c(M, C_0)$. For any $x \in \mathbb{R}^p$, let us write $f^*(x) = \langle x, t^* \rangle$, where $t^* \in \mathbb{R}^p$. Let us assume that the oracle is such that

$$\mathbb{P}\big(|\langle X, t^* \rangle| \leq c_0\big) \geq 1 - \frac{1}{2(C')^{(4+2\varepsilon)/\varepsilon}}. \tag{3.21}$$

Therefore, if Equation (3.21) holds, the local Bernstein Assumption is verified for a constant A depending on M, C_0 and c_0 given in Proposition 3.1 (since the latter formula is rather complicated, we will keep the notation A all along this section).

Finally, let us turn to a solution to the sparsity equation for the ℓ_1^p norm . The result can be found in (Lecué and Mendelson, 2018).

Lemma 3.1. *(Lecué and Mendelson, 2018, Lemma 4.2) . Let us assume that X is isotropic. If the oracle t^* can be decomposed as $t^* = v + u$ with $u \in (\rho/20)B_1^p$ and $100s \leq \big(\rho/\sqrt{C_{K,r}(\rho, A)}\big)^2$ then $\Delta(\rho) \geq (4/5)\rho$, where $s = |supp(v)|$.*

Assume that t^* is a s-sparse vector, so Lemma 3.1 applies. We consider two cases depending on the values of K and $Nr_2^2(\gamma, \rho^*)$. When $C_{K,r}(\rho^*, A) = r_2^2(\gamma, \rho^*)$ – which holds when $K \leq c(c_0, C_0, M)Nr_2^2(\gamma, \rho^*)$ – Lemma 3.1 shows that $\rho^* = c(c_0, M, C_0)s\sqrt{\log(ep/s)/N}$ satisfies the sparsity equation. For these values, the value of r_2 given in (3.20) yields

$$r_2^2(\gamma, \rho^*) = c(c_0, M, C_0, \gamma)\frac{s\log(ep/s)}{N} .$$

Now, if $C_{K,r}(\rho, A) = c(A, L)K/N$ – which holds when $K \geq c(c_0, C_0, M)Nr_2^2(\gamma, \rho^*)$– we can take $\rho^* = c(c_0, M, C_0)\sqrt{sK/N}$. Therefore, Theorem 3.2 applies with

$$\rho^* = c(c_0, M, C_0)\max(s\sqrt{\log(ep/s)/N}, \sqrt{sK/N}) .$$

Finally from Remark 3.1, note that is necessary to have $N \geq c\log(ep/s)$, where $c > 0$ is an absolute constant in order to have A like a constant in Proposition 3.1.

Theorem 3.6. *Let $\varepsilon > 0$ and (X, Y) be a random variable taking values in $\mathbb{R}^p \times \{\pm 1\}$, where X is an isotropic random vector such that for all $1 \leq j \leq p$ and $1 \leq q \leq C_0 \log(p)$, $\left\|\langle X, e_j\rangle\right\|_{L_q} \leq M\sqrt{q}$ with $C_0 \log(p) \geq 2 + \varepsilon$. Let $f^* : x \in \mathbb{R}^p \mapsto \langle x, t^*\rangle$ be the oracle where $t^* \in \mathbb{R}^p$ is s-sparse. Assume also that the oracle satisfies (3.21). Assume that $(X, Y), (X_i, Y_i)_{i\in\mathcal{I}}$ are i.i.d distributed and $N \geq cs\log(ep/s)$. Let $K \geq 7|\mathcal{O}|/3$. With probability larger than $1 - 2\exp(-cK)$, the minmax MOM logistic LASSO estimator $\hat{t}_{\lambda,K}$ with*

$$\lambda = c(c_0, M, C_0) \max\left(\sqrt{\frac{\log(ep/s)}{N}}, \sqrt{\frac{K}{sN}}\right)$$

satisfies

$$\|\hat{t}_{\lambda,K} - t^*\|_1 \leq c(c_0, M, C_0) \max\left(s\sqrt{\frac{\log(ep/s)}{N}}, \sqrt{s}\sqrt{\frac{K}{N}}\right),$$
$$\|\hat{t}_{\lambda,K} - t^*\|_2^2 \leq c(c_0, M, C_0) \max\left(\frac{K}{N}, s\frac{\log(ep/s)}{N}\right) ,$$
$$P\mathcal{L}_{\hat{f}_{\lambda,K}} \leq c(c_0, M, C_0) \max\left(\frac{K}{N}, s\frac{\log(ep/s)}{N}\right) .$$

For $K \leq c(c_0, M, C_0)s\log(ep/s)$, the upper bound on the estimation risk and excess risk matches the minimax rates of convergence for s-sparse vectors in $\mathbb{R}^p$. It is also possible to adapt in a data-driven way to the best K and λ by using a Lepski's adaptation method such as in (Devroye et al., 2016; Lecué and Lerasle, 2017, 2019; Chinot et al., 2019b; Chinot, 2019b). This step is now well understood, it is not reproduced here.

3.6.3 The minmax MOM logistic SLOPE

In this section, we study the minmax MOM estimator with the logistic loss function and the SLOPE regularization norm. Given $\beta_1 \geq \beta_2 \geq \cdots \geq \beta_p > 0$, the SLOPE norm (see (Bogdan et al., 2015)) is defined for all $t \in \mathbb{R}^p$ by

$$\|t\|_{\text{SLOPE}} = \sum_{i=1}^{p} \beta_i t_i^* ,$$

where $(t_i^*)_{i=1}^p$ denotes the non-increasing re-arrangement of $(|t_i|)_{i=1}^p$. The SLOPE norm coincides with the ℓ_1 norm when $\beta_j = 1$ for all $j = 1, \cdots, p$.

Given $K \in \{1, \ldots, N\}$ and $\lambda > 0$, the minmax MOM logistic SLOPE procedure is

$$\hat{t}_{\lambda,K} \in \operatorname*{argmin}_{t\in\mathbb{R}^p} \sup_{\tilde{t}\in\mathbb{R}^p} \left(\text{MOM}_K\big(\ell_t - \ell_{\tilde{t}}\big) + \lambda(\|t\|_{\text{SLOPE}} - \|\tilde{t}\|_{\text{SLOPE}})\right) , \tag{3.22}$$

where $\ell_t : (x, y) \in \mathbb{R}^p \times \{-1, 1\} = \log(1 + \exp(-y\langle x, t\rangle))$ for all $t \in \mathbb{R}^p$.

Let us first compute the complexity function r_2. If $V \subset \mathbb{R}^p$ is closed under permutations and reflections (sign-changes)– which is the case for B^p_{SLOPE}, the unit ball of the SLOPE norm – then $\sup_{v\in V} |\langle \cdot, v\rangle|$ is 1-unconditional. Therefore, the dual norm of $\|\cdot\|_{SLOPE}$ is 1-unconditional and

Theorem 3.5 applies provided that X is isotropic and verifies (3.17). By (Lecué and Mendelson, 2018, Lemma 5.3), we have

$$\mathbb{E}\sup_{t\in\rho B^p_{\mathrm{SLOPE}}\cap rB^p_2}\left|\sum_{i\in J}\sigma_i\langle X_i,t\rangle\right| \le c(C_0,M)\sqrt{|J|}w(\rho B^p_{\mathrm{SLOPE}}\cap rB^p_2)$$
$$\le c(C_0,M)\sqrt{|J|}\begin{cases} r\sqrt{p} & \text{if } r\le\rho/\sqrt{p}\\ \rho & \text{if } r\ge\rho/\sqrt{p}\end{cases} \tag{3.23}$$

It follows that

$$r_2^2(\gamma,\rho)=c(C_0,\gamma,M)\begin{cases}\frac{p}{N} & \text{if } p\le c(C_0,\gamma,M)\rho\sqrt{N}\\ \frac{\rho}{\sqrt{N}} & \text{if } p\ge c(C_0,\gamma,M)\rho\sqrt{N}.\end{cases}$$

Let us turn to the local Bernstein Assumption. Since the loss function is the same as the one used in Section 3.6.2, the local Bernstein assumption holds if there exists $c_0>0$ such that

$$\mathbb{P}\big(|\langle X,t^*\rangle|\le c_0\big)\ge 1-\frac{1}{2(C')^{(2+2\varepsilon)/\varepsilon}} \tag{3.24}$$

where $C'=c(M,C_0)$ is a function of M and C_0 only. The constant A in the Bernstein condition depends on c_0, C_0 and M. As for the LASSO, since the formula of A is complicated (given in Proposition 3.1), we write A all along this section but we assume that $r_2(\gamma,\rho^*)(2C')^{(2+\varepsilon)/\varepsilon}\le c_0/2$ so that A can be considered like an absolute constant (depending only on c_0). This condition is equivalent to assuming $N\gtrsim s\log(ep/s)$.

A solution to the sparsity equation relative to the SLOPE norm can be found in (Lecué and Mendelson, 2018). We recall this result here.

Lemma 3.2. *(Lecué and Mendelson, 2018, Lemma 4.3) Let $1\le s\le p$ and set $\mathcal{B}_s=\sum_{i\le s}\beta_i/\sqrt{i}$. If t^* can be decomposed as $t^*=u+v$ with $u\in(\rho/20)B^p_{SLOPE}$ and v is s-sparse and if $40\mathcal{B}_s\le\rho/\sqrt{C_{K,r}(\rho,A)}$ then $\Delta(\rho)\ge 4\rho/5$.*

Assume that t^* is exactly s-sparse, so that Lemma 3.2 applies. We consider two cases depending on K. Consider the case where $K\le c(c_0,C_0,M)Nr_2^2(\gamma,\rho^*)$, so $\sqrt{C_{K,r}(\rho^*,A)}=r_2(\gamma,\rho^*)$. For $\beta_j=c\sqrt{\log(ep/j)}$, one may show that $\mathcal{B}_s=c\sqrt{s\log(ep/s)}$ (see (Bellec et al., 2018; Lecué and Mendelson, 2018)). From (3.23) and Lemma 3.2, it follows that we can choose

$$\rho^*=c(c_0,M,C_0)s\frac{\log(ep/s)}{\sqrt{N}}\quad\text{and thus}\quad r_2^2(\gamma,\rho^*)=c(c_0,M,C_0)\frac{s\log(ep/s)}{N}\ . \tag{3.25}$$

For $C_{K,r}(\rho,A)=c(c_0,M,C_0)K/N$ holding when $K\ge c(c_0,C_0,M)Nr_2^2(\gamma,\rho^*)$, we take $\rho^*=c(c_0,C_0,M)\sqrt{sK}$ satisfying the sparsity equation. We can therefore apply Theorem 3.2 for

$$\rho^*=c(c_0,M,C_0)\max(s\sqrt{\log(ep/s)/N},\sqrt{sK}/\sqrt{N})\ .$$

Theorem 3.7. *Let $\varepsilon>0$ and (X,Y) be random variable with values in $\mathbb{R}^p\times\{\pm1\}$ such that X is an isotropic random vector such that for all $1\le j\le p$ and $1\le q\le C_0\log(p)$, $\big\|\langle X,e_j\rangle\big\|_{L_q}\le M\sqrt{q}$*

with $C_0 \log(p) \geq 2 + \varepsilon$. *Let* $f^* : x \in \mathbb{R}^p \mapsto \langle x, t^* \rangle$ *be the oracle where* $t^* \in \mathbb{R}^p$ *is* s*-sparse. Assume also that the oracle satisfies* (3.21). *Assume that* $(X, Y), (X_i, Y_i)_{i \in \mathcal{I}}$ *are i.i.d and* $N \geq cs \log(ep/s)$. *Let* $K \geq 7|\mathcal{O}|/3$. *Let* $\hat{t}_{\lambda,K}$ *be the minmax MOM logistic Slope procedure introduced in* (3.22) *for the choice of weights* $\beta_j = \sqrt{\log(ep/j)}, j = 1, \ldots, p$ *and regularization parameter* $\lambda = c(c_0, M, C_0) \max(1/\sqrt{N}, \sqrt{K/(sN)})$. *With probability larger than* $1 - 2\exp(-cK)$,

$$\|\hat{t}_{\lambda,K} - t^*\|_{SLOPE} \leq c(c_0, M, C_0) \max\left(s\sqrt{\frac{\log(ep/s)}{N}}, \sqrt{s}\sqrt{\frac{K}{N}} \right),$$

$$\|\hat{t}_{\lambda,K} - t^*\|_2^2 \leq c(c_0, M, C_0) \max\left(\frac{K}{N}, s\frac{\log(ep/s)}{N} \right) ,$$

$$P\mathcal{L}_{\hat{t}_{\lambda,K}} \leq c(c_0, M, C_0) \max\left(\frac{K}{N}, s\frac{\log(ep/s)}{N} \right) .$$

For $K \leq c(c_0, M, C_0) s \log(ep/s)/N$, the parameter λ is independent from the unknown sparsity s and these bounds match the minimax rates of convergence over the class of s-sparse vectors in $\mathbb{R}^p$ without any restriction on s (Bellec et al., 2018). Ultimately, one can use a Lepski's adaptation method to chose in a data-driven way the number of blocks K as in (Lecué and Lerasle, 2019) to achieve these optimal rates without prior knowledge on the sparsity s.

3.6.4 The minmax MOM Huber Group-Lasso

In this section, we consider regression problems where $\mathcal{Y} = \mathbb{R}$. We consider group sparsity as notion of low-dimensionality for t^*. This setup is particularly useful when features (i.e. coordinates of X) are organized by blocks, as when one constructs dummy variables from a categorical variable.

The regularization norm used to induce this type of "structured sparsity" is called the Group LASSO (see, for example (Yang and Zou, 2015) and (Meier et al., 2008)). It is built as follows: let $G_1, \cdots, G_M$ be a partition of $\{1, \cdots, p\}$ and define, for any $t \in \mathbb{R}^p$

$$\|t\|_{\mathrm{GL}} = \sum_{k=1}^{M} \|t_{G_k}\|_2 \ . \tag{3.26}$$

Here, for all $k = 1, \ldots, M$, t_{G_k} denotes the orthogonal projection of t onto the linear $\mathrm{Span}(e_i, i \in G_k)$ – $(e_1, \ldots, e_p)$ being the canonical basis of $\mathbb{R}^p$.

The estimator we consider is the minmax MOM Huber Group-LASSO defined, for all $K \in \{1, \cdots, N\}$ and $\lambda > 0$, by

$$\hat{t}_{\lambda,K} \in \operatorname*{argmin}_{t \in \mathbb{R}^p} \sup_{\tilde{t} \in \mathbb{R}^p} \left(\mathrm{MOM}_K\big(\ell_t - \ell_{\tilde{t}}\big) + \lambda(\|t\|_{GL} - \|\tilde{t}\|_{GL}) \right) ,$$

where $t \in \mathbb{R}^p \to \ell_t$ is the Huber loss function with parameter $\delta > 0$ defined as

$$\ell_t(X_i, Y_i) = \begin{cases} \frac{1}{2}(Y_i - \langle X_i, t \rangle)^2 & \text{if } |Y_i - \langle X_i, t \rangle| \leq \delta \\ \delta|Y_i - \langle X_i, t \rangle| - \frac{\delta^2}{2} & \text{if } |Y_i - \langle X_i, t \rangle| > \delta \end{cases} .$$

In particular, it is a Lipschitz loss function with $L = \delta$. Estimation bounds and oracle inequalities satisfied by $\hat{t}_{\lambda,K}$ follow from Theorem 3.2 as long as we can compute the complexity function r_2, we verify the local Bernstein Assumption and we find a radius ρ^* satisfying the sparsity equation. We now handle these problems starting with the computation of the complexity function r_2.

The dual norm of $\|\cdot\|_{\mathrm{GL}}$ is $z \in \mathbb{R}^p \to \|z\|_{\mathrm{GL}}^* = \max_{1\le k\le M} \|z_{G_k}\|_2$, it is not κ-unconditional with respect to the canonical basis $(e_i)_{i=1}^p$ of $\mathbb{R}^p$ for some absolute constant κ, so Theorem 3.5 does not apply directly. Therefore, in order to avoid long and technical materials on the rearrangement of empirical means under weak moment assumptions for the computation of the local Rademacher complexity from (3.10), we simply assume that the design vectors $(X_i)_{i\in\mathcal{I}}$ are L_0-subgaussian and isotropic: for all $i \in \mathcal{I}$, all $t \in \mathbb{R}^p$ and all $q \ge 1$

$$\big\|\langle X_i, t\rangle\big\|_{L_q} \le L_0\sqrt{q}\,\big\|\langle X_i, t\rangle\big\|_{L_2} \text{ and } \big\|\langle X_i, t\rangle\big\|_{L_2} = \|t\|_2\,. \tag{3.27}$$

In that case, a direct chaining argument allows to bound Rademacher processes by the Gaussian processes (see (Talagrand, 2014) for chaining methods):

$$\mathbb{E}\sup_{t\in\rho B_{\mathrm{GL}}^p\cap rB_2^p}\left|\sum_{j\in J}\sigma_j\langle t, X_j\rangle\right| \le c(L_0)\sqrt{J}w(\rho B_{\mathrm{GL}}^p \cap rB_2^p)\ .$$

Here, B_{GL}^p is the unit ball of $\|\cdot\|_{\mathrm{GL}}$, $w(\rho B_{\mathrm{GL}}^p \cap rB_2^p)$ is the Gaussian mean width of the interpolated body $\rho B_{\mathrm{GL}}^p \cap rB_2^p$. It follows from the proof of Proposition 6.7 in (Bellec et al., 2017) that when the M groups $G_1, \dots, G_M$ are all of same size p/M we have

$$w(\rho B_{\mathrm{GL}}^p \cap rB_2^p) \le \begin{cases} c\rho\sqrt{\frac{p}{M} + \log\left(\frac{Mr^2}{\rho^2}\right)} & \text{if } 0 < \rho \le r\sqrt{M} \\ cr\sqrt{p} & \text{if } \rho \ge r\sqrt{M} \end{cases}\ .$$

This yields

$$r_2^2(\gamma,\rho) = c(\delta, L_0, \gamma)\begin{cases} \frac{\rho}{\sqrt{N}}\sqrt{\frac{p}{M} + \log\left(\frac{Mr^2}{\rho^2}\right)} & \text{if } 0 < c(\delta,L_0,\gamma)\frac{\rho}{r} \le \sqrt{M} \\ \frac{r}{\sqrt{N}}\sqrt{p} & \text{if } c(\delta,L_0,\gamma)\frac{\rho}{r} \ge \sqrt{M} \end{cases}\ . \tag{3.28}$$

Let us now turn to the local Bernstein Assumption. We need to verify Assumption 3.11. As we assumed that the design vectors $(X_i)_{i\in\mathcal{I}}$ are isotropic and L_0-subgaussian, it is clear that the point a) in Assumption 3.11 holds with $C' = L_0$. Let us take $\varepsilon = 2$ (another choice would only change the constant). For the point b), we assume that there exists $\alpha > 0$ such that, for all $x \in \mathcal{X}$ and all $z \in \mathbb{R}$ satisfying $|z - f^*(x)| \le 2L_0^2\sqrt{C_{K,r}(\rho, 4/\alpha)}$, $F_{Y|X=x}(z+\delta) - F_{Y|X=x}(z-\delta) \geqslant \alpha$. Under these conditons, the local Bernstein Assumption is verified for $A = 4/\alpha$ according to Proposition 3.2. We will assume that $C_{K,r}(\rho^*, 4/\alpha) \le c$ for some absolute constant c so that δ and α can be taken like absolute constant. Condition "$C_{K,r}(\rho^*, 4/\alpha) \le c$" is satisfied when $N \gtrsim cs\log(ep/s)$.

Finally, we turn to the sparsity equation. The following lemma is an extension of Lemma 3.1 to the Group Lasso norm.

Lemma 3.3. *Assume that X is isotropic. Assume that $t^* = u + v$ where $\|u\|_{GL} \leq \rho/20$ and v is group-sparse i.e $v_{G_k} = 0$ for all $k \notin I$ for some $I \subset \{1, \dots, M\}$. If $100|I| \leq (\rho/\sqrt{C_{K,r}(\rho, 4/\alpha)})^2$, then $\Delta(\rho) \geq 4\rho/5$.*

Proof. Let us define $r(\rho) := \sqrt{C_{K,r}(\rho, 4/\alpha)}$ and recall that

$$\tilde{\Delta}(\rho, 4/\alpha) = \inf_{w \in \rho S_{GL} \cap r(\rho) B_2^p} \sup_{z^* \in \Gamma_{t^*}(\rho)} \left\langle z^*, w \right\rangle \ .$$

Here, S_{GL} is the unit sphere of $\|\cdot\|_{GL}$ and $\Gamma_{t^*}(\rho)$ is the union of all sub-differentials $(\partial \|\cdot\|_{GL})_v$ for all $v \in t^* + (\rho/20)B_{GL}^p$. We want to find a condition on $\rho > 0$ insuring that $\tilde{\Delta}(\rho, 4/\alpha) \geq 4\rho/5$.

Let w be a vector in $\mathbb{R}^p$ such that $\|w\|_{GL} = \rho$ and $\|w\|_2 \leq r(\rho)$. We construct $z^* \in \mathbb{R}^p$ such that $z^*_{G_k} = w_{G_k}/\|w_{G_k}\|_2$ if $k \notin I$ (so that $\left\langle z^*_{G_k}, w_{G_k} \right\rangle = \|w_{G_k}\|_2$ for all $k \notin I$) and $z^*_{G_k} = v_{G_k}/\|v_{G_k}\|_2$ if $k \in I$ (so that $\left\langle z_{G_k}, v_{G_k} \right\rangle = \|v_{G_k}\|_2$ for all $k \in I$). We have $\left\|z^*_{G_k}\right\|_2 = 1$ for all $k \in [M]$, so $\|z^*\|^*_{GL} = 1$ (i.e. z^* is in the dual sphere of $\|\cdot\|_{GL}$) and $\left\langle z^*, v \right\rangle = \|v\|_{GL}$ (i.e. z^* is norming for v). Therefore, it follows from (3.18) that $z^* \in (\partial \|\cdot\|_{GL})_v$. Moreover, $\|u\|_{GL} \leq \rho/20$ hence $v \in t^* + (\rho/20)B_{GL}^p$ and so $z^* \in \Gamma_{t^*}(\rho)$. Furthermore, for this choice of sub-gradient z^*, we have

$$\begin{aligned}
\left\langle z^*, w \right\rangle &= \sum_{k \in I} \left\langle z^*_{G_k}, w_{G_k} \right\rangle + \sum_{k \notin I} \left\langle z^*_{G_k}, w_{G_k} \right\rangle \geq -\sum_{k \in I} \|w_{G_k}\|_2 + \sum_{k \notin I} \|w_{G_k}\|_2 \\
&= \sum_{k=1}^{M} \|w_{G_k}\|_2 - 2 \sum_{k \in I} \|w_{G_k}\|_2 \geq \rho - 2\sqrt{|I|} r(\rho) \ .
\end{aligned}$$

In the last inequality, we used that $\|w\|_{GL} = \rho$ and that

$$\sum_{k \in I} \|w_{G_k}\|_2 \leq \sqrt{|I|} \sqrt{\sum_{k \in I} \|w_{G_k}\|_2^2} \leq \sqrt{|I|}\, \|w\|_2 \leq \sqrt{|I|} r(\rho).$$

Then $\left\langle z^*, w \right\rangle \geq 4\rho/5$ when $\rho - 2\sqrt{|I|} r(\rho) \geq 4\rho/5$ which happens to be true when $100|I| \leq (\rho/r(\rho))^2$. ∎

Assume that t^* is exactly s-group sparse, so Lemma 3.3 applies. We consider two cases depending on the value of K. When $K \leq c(L_0, \alpha, \delta) N r_2^2(\gamma, \rho^*)$, $\sqrt{C_{K,r}(\rho^*, 4/\alpha)} = r_2(\gamma, \rho^*)$. By Lemma 3.3 and (3.28), it follows that (for equal size blocks), one can choose

$$\rho^* = c(L_0, \alpha, \delta) \frac{s}{\sqrt{N}} \sqrt{\frac{p}{M} + \log M} \quad \text{and thus} \quad r^2(\gamma, \rho^*) = c(L_0, \alpha, \delta) \frac{s}{N} \left(\frac{p}{M} + \log M \right) \ . \tag{3.29}$$

This result has a similar flavor as the one for the Lasso. The term $s' = sp/M$ equals *block sparsity* $\times$ *size of each blocks*, i.e to the total number of non-zero coordinates in t^*: $s' = \|t^*\|_0$. Replacing the sparsity s' by sp/M in Theorem 3.6, we would have obtained $\rho^* = c(L_0, \alpha, \delta)(sp/M)\sqrt{\log(p)/N}$ which is larger than the bound obtained for the Group Lasso in Equation (3.29). It is therefore better to induce the sparsity by blocks instead of just coordinate-wise when we are aware of such block-structured sparsity. In the other case, when $K \leq c(L_0, \alpha, \delta) N r_2^2(\gamma, \rho^*)$, we have $\sqrt{C_{K,r}(\rho^*, 4/\alpha)} =$

$c(L_0,\alpha,\delta)\sqrt{K/N}$ and so one can take $\rho^* = c(L_0,\alpha,\delta)\sqrt{sK/N}$. We can therefore apply Theorem 3.2 with

$$\rho^* = c(L_0,\alpha,\delta)\max\left(\frac{s}{\sqrt{N}}\sqrt{\frac{p}{M}+\log(M)}, \sqrt{s}\sqrt{\frac{K}{N}}\right) .$$

Theorem 3.8. *Let (X,Y) be a random variables with values in $\mathbb{R}^p\times\mathbb{R}$ such that $Y\in L_1$ and X is an isotropic and L_0-subgaussian random vector in $\mathbb{R}^p$. Assume that $(X,Y),(X_i,Y_i)_{i\in\mathcal{I}}$ are i.i.d. Let $f^*(\cdot)=\langle t^*,\cdot\rangle$ for some $t^*\in\mathbb{R}^p$ which is s-group sparse with respect to equal-size groups $(G_k)_{k=1}^M$. Let $K\geq 7|\mathcal{O}|/3$ and $N\geq cs(p/M+\log(M))$. Assume that there exists $\alpha>0$ such that, for all $x\in\mathbb{R}^p$ and all $z\in\mathbb{R}$ satisfying $|z-\langle t^*,x\rangle|\leq 2L_0^2\sqrt{C_{K,r}(2\rho^*,4/\alpha)}$, $F_{Y|X=x}(\delta+z)-F_{Y|X=x}(z-\delta)\geq\alpha$ (where $F_{Y|X=x}$ is the cumulative ditribution function of Y given $X=x$). With probability larger than $1-2\exp(-cK)$, the MOM Huber group-LASSO estimator $\hat{t}_{\lambda,K}$ for*

$$\lambda = c(L_0,\alpha,\delta)\max\left(\frac{1}{\sqrt{N}}\sqrt{\frac{p}{M}+\log M}, \sqrt{\frac{K}{sN}}\right)$$

satisfies

$$\|\hat{t}_{\lambda,K}-t^*\|_{GL}\leq c(L_0,\alpha,\delta)\max\left(\frac{s}{\sqrt{N}}\sqrt{\frac{p}{M}+\log(M)}, \sqrt{s}\sqrt{\frac{K}{N}}\right),$$
$$\|\hat{t}_{\lambda,K}-t^*\|_2^2\leq c(L_0,\alpha,\delta)\max\left(\frac{s}{N}\left(\frac{p}{M}+\log(M)\right), \frac{K}{N}\right),$$
$$P\mathcal{L}_{\hat{t}_{\lambda,K}}\leq c(L_0,\alpha,\delta)\max\left(\frac{s}{N}\left(\frac{p}{M}+\log(M)\right), \frac{K}{N}\right) .$$

For $K\leq c(L_0,\alpha,\delta)s(p/M+\log M)$, the regularization parameter λ is independent from the unknown group sparsity s (the choice of K can be done in data-driven way using either a Lepski method or a MOM cross validation as in (Lecué and Lerasle, 2019)). In the ideal i.i.d. setup (with no outliers), the same result holds for the RERM as we assumed that the class $F-f^*$ is L_0-subgaussian and for the choice of regularization parameter $\lambda = c(L_0,\alpha,\delta)(\sqrt{p/(NM)}+\sqrt{\log(M)/N})$. The minmax MOM estimator has the advantage to be robust up to $c(L_0,\alpha,\delta)s(p/M+\log M)$ outliers in the dataset.

3.7 Simulations

This section provides a simulation study to illustrate our theoretical findings. Minmax MOM estimators are approximated using an alternating proximal block gradient descent/ascent with a wisely chosen block of data as in (Lecué and Lerasle, 2019). At each iteration, the block on which the descent/ascent is performed is chosen according to its "centrality" (see algorithm 2 below). Two examples from high-dimensional statistics are considered 1) Logistic classification with a ℓ_1 penalization and 2) Huber regression with a Group-Lasso penalization.

3.7.1 Presentation of the algorithm

Let $\mathcal{X} = \mathbb{R}^p$ and let $F = \{\langle t, \cdot \rangle, t \in \mathbb{R}^p\}$. The oracle $f^* = \operatorname{argmin}_{f \in F} P\ell_f(X, Y)$ is such that $f^*(\cdot) = \langle t^*, \cdot \rangle$ for some $t^* \in \mathbb{R}^p$. The minmax MOM estimator is defined as

$$\hat{t}_{\lambda,K} \in \operatorname*{argmin}_{t \in \mathbb{R}^p} \sup_{\tilde{t} \in \mathbb{R}^p} MOM_K(\ell_t - \ell_{\tilde{t}}) + \lambda(\|t\| - \|\tilde{t}\|) \tag{3.30}$$

where ℓ is a convex and Lipschitz loss function and $\|\cdot\|$ is a norm in $\mathbb{R}^p$.

Following the idea of (Lecué and Lerasle, 2019), the minmax problem (3.30) is approximated by a proximal block gradient ascent-descent algorithm, see Algorithm 2. At each step, one considers the block of data realizing the median and perform an ascent/descent step onto this block. The regularization step is obtained via the proximal operator

$$\operatorname{prox}_{\lambda\|\cdot\|} : x \in \mathbb{R}^p \to \operatorname*{argmin}_{y \in \mathbb{R}^p} \left\{ \frac{1}{2}\|x - y\|_2^2 + \lambda\|y\| \right\} .$$

Algorithm 2: Proximal Descent-Ascent gradient method with median blocks

Input: A number of blocks K, initial points t_0 and $\tilde{t}_0$ in $\mathbb{R}^p$, two sequences of step sizes $(\eta_t)_t$ and $(\tilde{\eta}_t)_t$ and T a number of epochs

Output: An approximating solution of the minimax problem (3.30)

for $i = 1, \cdots, T$ **do**

  Construct a random equipartition $B_1 \sqcup \cdots \sqcup B_K$ of $\{1, \cdots, N\}$

  Find $k \in [K]$ such that $\mathrm{MOM}_K(\ell_{t_i} - \ell_{\tilde{t}_i}) = P_{B_k}(\ell_{t_i} - \ell_{\tilde{t}_i})$

  Update:

    $t_{i+1} = \operatorname{prox}_{\lambda\|\cdot\|}\big(t_i - \eta_i \nabla_t (t \to P_{B_k}\ell_t)_{|t=t_i}\big)$

    $\tilde{t}_{i+1} = \operatorname{prox}_{\lambda\|\cdot\|}\big(\tilde{t}_i - \tilde{\eta}_i \nabla_{\tilde{t}} (\tilde{t} \to P_{B_k}\ell_{\tilde{t}})_{|\tilde{t}=\tilde{t}_i}\big)$

end

To make the presentation simple in Algorithm 2, we have not perform any line search or any sophisticated stopping rule (see, (Lecué and Lerasle, 2019) for more involved line search and stopping rules in the setup of minmax MOM algorithms). To compare the statistical and robustness performances of the minmax MOM and RERM, we perform a proximal gradient descent to approximate the RERM, see Algorithm 3 below.

The number of blocks K is chosen by MOM cross-validation (see (Lecué and Lerasle, 2019) for more precision on that procedure). The sequences of stepsizes are constant along the algorithm $(\eta_t)_t := \eta$ and $(\tilde{\eta}_t)_t = \tilde{\eta}$ and are also chosen by MOM cross-validation.

Algorithm 3: Proximal gradient descent algorithm

Input: Initial points t_0 in $\mathbb{R}^p$ and a sequence of stepsizes $(\eta_t)_t$

Output: Approximating solution to the RERM estimator.

for $i = 1, \cdots, T$ **do**

$$t_{i+1} = \mathrm{prox}_{\lambda\|\cdot\|}\big(t_i - \eta_i \nabla_t (t \to P_N \ell_t)_{|t=t_i}\big)$$

end

3.7.2 Organisation of the results

In all simulations, the links between inputs and outputs are given in the regression and classification problems in $\mathbb{R}^p$ respectively by the following model:

$$\text{in regression: } Y = \big\langle X, t^* \big\rangle + \zeta; \qquad \text{in classification: } Y = \mathrm{sign}\big(\big\langle X, t^* \big\rangle + \zeta\big) \tag{3.31}$$

where the distribution of X and ζ depend on the considered framework:

- **First framework:** X is a standard Gaussian random vector in $\mathbb{R}^p$ and ζ is a real-valued standard Gaussian variable independent of X with variance σ^2.
- **Second framework:** X is a standard Gaussian random vector in $\mathbb{R}^p$ and $\zeta \sim \mathcal{T}(2)$ (student distribution with 2 degrees of freedom). This framework is used to verify the robustness w.r.t the noise.
- **Third framework:** $X = (x_1, \cdots, x_p)$ with $x_1, \ldots, x_p \overset{i.i.d.}{\sim} \mathcal{T}(2)$ and ζ is a real-valued standard Gaussian variable independent of X with variance σ^2. Here we want to test the robustness w.r.t heavy-tailed design $(X_i)_i$.
- **Fourth framework:** $X = (x_1, \cdots, x_p)$ with $x_1, \ldots, x_p \overset{i.i.d.}{\sim} \mathcal{T}(2)$ and $\zeta \sim \mathcal{T}(2)$. We also corrupt the database with $|\mathcal{O}|$ outliers which are such that for all $i \in \mathcal{O}$, $X_i = (10^5)_{i=1}^p$ and $Y = 1$. Here we verify the robustness w.r.t possible outliers in the dataset.

In a both first and second frameworks, the RERM and minmax MOM estimators are expected to perform well according to Theorem 3.1 and Theorem 3.2 even though the noise ζ can be heavy-tailed. In the third framework, the design vector X is no longer subgaussian, as a consequence Theorem 3.1 does not apply and we have no guarantee for the RERM. On the contrary, Theorem 3.2 provides statistical guarantees for the minmax MOM estimators. Nevertheless, it should also be noticed that the study of RERM under moment assumptions on the design can also be performed, see for instance (Lecué and Mendelson, 2017). In that case, the rates of convergence are still the same but the deviation is only polynomial whereas it is exponential for the minmax MOM estimators. Therefore, in the third example, we may expect similar performance for both estimators but with a larger variance in the results for the RERM. In the fourth framework, the database has been corrupted by outliers (in both outputs Y_i and inputs X_i); in that case, only minmax MOM estimators are expected to perform well as long as $|\mathcal{O}|$ is not too large compare with K, the number of blocks.

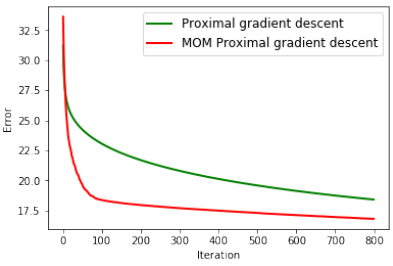
3.7.3 Sparse Logistic regression

Let ℓ denote the Logistic loss (i.e. $t \in \mathbb{R}^p \to \ell_t(x,y) = \log(1+\exp(-y\langle x,t\rangle)), \forall x \in \mathbb{R}^p, y \in \mathcal{Y} = \{\pm 1\}$), and let the ℓ_1 norm in $\mathbb{R}^p$ be the regularization norm. Figure 3.1 presents the results of our simulations for $N = 1000$, $p = 400$ and $s = 30$. In subfigures (a), (b) and (c) the error is the L_2 error, which is here $\left\|\hat{t}^T_{K,\lambda} - t^*\right\|_2$, between the output $\hat{t}^T_{K,\lambda}$ of the algorithm and the true $t^* \in \mathbb{R}^p$. In subfigure (d), an increasing number of outliers is added. The error rate is the proportion of misclassification on a test dataset. The stepsizes, the number of block and the parameteter of regularization are all chosen by MOM cross-validation (see (Lecué and Lerasle, 2019) for more details on the MOM cross-validation procedure) Subfigure (a) shows convergence of the error for both algorithms in the first framework. Similar performances are observed for both algorithms but Algorithm 2 converges faster than Algorithm 3. It may be because the computation of the gradient on a smaller batch of data in step **5** and **6** of Algorithm 2 is faster than the one on the entire database in step **2** of Algorithm 3 and that the choice of the median blocks at each descent/ascent step is particularly good in Algorithm 2. Subfigure (b) shows the results in the second framework. The convergence for the alternating gradient ascent/descent algorithm is a bit faster than the one from Algorithm 3, but the performances are the same. Subfigure (c) shows results in the third setup where ζ is Gaussian and the feature vector $X = (x_1, \cdots, x_p)$ is heavy-tailed, i.e. $x_1, \ldots, x_p$ are i.i.d. with $x_1 \sim \mathcal{T}(2)$ – a Student with degree 2. Minmax MOM estimators perform better than RERM. It highlights the fact that minmax MOM estimators have optimal subgaussian performance even without the sub-gaussian assumption on the design while RERM are expected to have downgraded statistical properties in heavy-tailed scenariis. Subfigure (d) shows result in the fourth setup where an increasing number of outliers is added in the dataset. Outliers are $X = (10^5)^p_1$ and $Y_i = 1$ a.s.. While RERM has deteriorated performance just after one outliers was added to the dataset, minmax MOM estimators maintains good performances up to 10% of outliers.

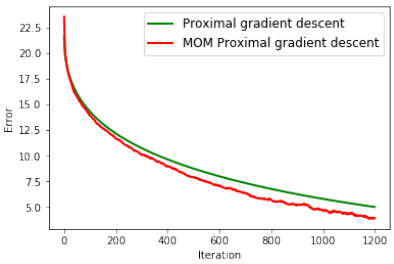
3.7.4 Huber regression with a Group Lasso penalty

Let ℓ denote the Huber loss function $t \in \mathbb{R}^d \to \ell_t(x,y) = (y - \langle x,t\rangle)^2/2$ if $|y - \langle x,t\rangle| \leq \delta$ and $\ell_t(x,y) = \delta|y - \langle x,t\rangle| - \delta^2/2$ other wise for all $x \in \mathbb{R}^p$ and $y \in \mathcal{Y} = \mathbb{R}$. Let $G_1, \cdots, G_M$ be a partition of $\{1, \cdots, p\}$, $\|t\| = \|t\|_{\mathrm{GL}} = \sum_{k=1}^M \|t_{G_k}\|_2$. Figure 3.1 presents the results of our simulation for $N = 1000$, $p = 400$ for 30 blocks with a block-sparsity parameter $s = 5$. In subfigures (a), (b) and (c), the error is the L_2-error between the output of the algorithm and the oracle t^* – which corresponds here to a ℓ_2^p estimation error, given that the design in all cases is isotropic. In subfigure (d) the prediction error on a (non-corrupted) test set of both the RERM and the minmax MOM estimators are depicted.

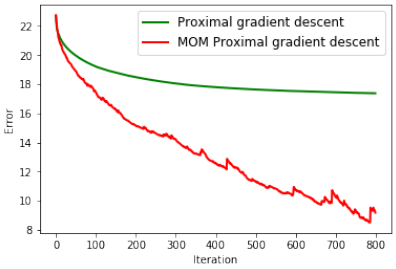
The conclusion are the same as for the Lasso Logistic regression: Algorithm 2 (regularized minmax MOM) has better performances than algorithm 3 (RERM) in case of heavy-tailed inliers and when outliers pollute the dataset while both are robust w.r.t heavy-tailed noise.



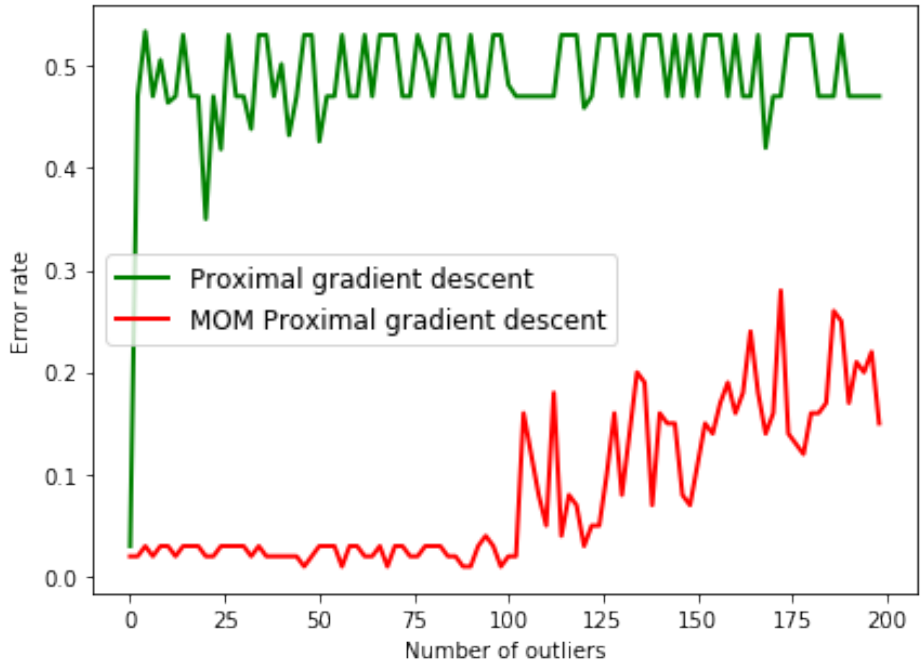
(a) Gaussian design and Gaussian noise.



(b) Heavy-tailed noise ζ (Student distribution of order 2) and standard Gaussian design.



(c) Gaussian noise and heavy-tailed design (Student distribution of order 2).

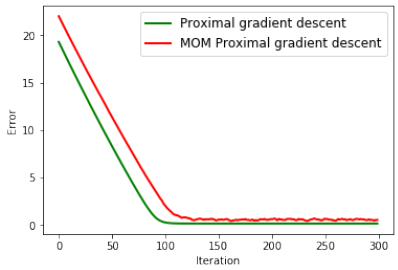


(d) Student of order 2 design and noise corrupted by outliers.

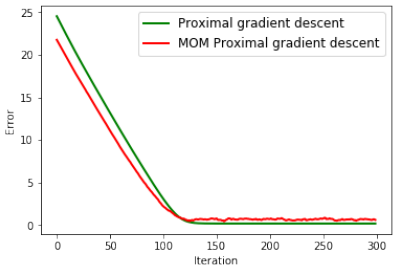
Figure 3.1: ℓ_2 estimation error rates of RERM and minmax MOM proximal descent algorithms (for the logistic loss and the ℓ_1 regularization norm) versus time in (a), (b) and (c) and versus number of outliers in (d) in the classification model (3.31) for $N = 1000$, $p = 400$ and $s = 30$.

3.8 Conclusion

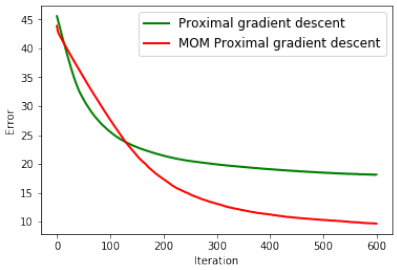
We obtain estimation and prediction results for RERM and regularized minmax MOM estimators for any Lipschitz and convex loss functions and for any regularization norm. When the norm has some sparsity inducing properties the statistical bounds depend on the dimension of the low-dimensional structure where the oracle belongs. We develop a systematic way to analyze both estimators by identifying three key idea 1) the local complexity function r_2 2) the sparsity equation 3) the local Bernstein condition. All these quantities and condition depend only on the structure and complexity of a local set around the oracle. This local set is ultimately proved to be the smallest set containing



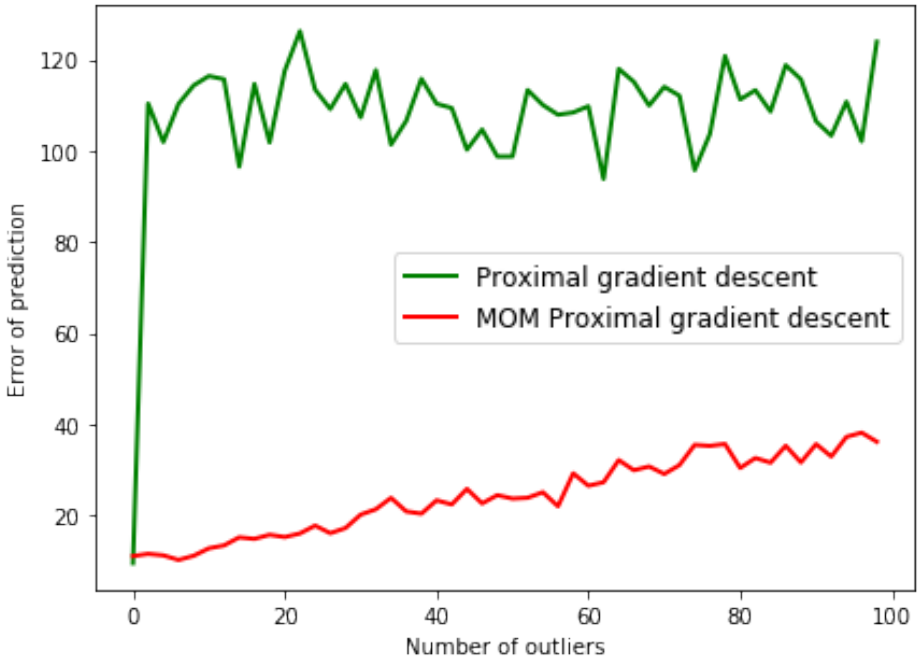
(a) Simulations from model (3.31) with standard Gaussian design and Gaussian noise



(b) Simulation with heavy-tailed noise ζ and standard Gaussian design



(c) Simulations with Gaussian noise heavy tailed design (Student distribution)



(d) Error of prediction in function of the number of outliers in the dataset

Figure 3.2: Results for the Huber regression with Group-Lasso penalization

our estimators. We show the versatility of our main meta-theorems on several applications covering two different loss functions and four sparsity inducing regularization norms. Some of them inducing highly structured sparsity concept such as the Group Lasso norm.

On top of these results, we show that the minmax MOM approach is robust to outliers and to heavy-tailed data and that the computation of the key objects such as the complexity functions r_2 and a radius ρ^* satisfying the sparsity equation can be done in this corrupted heavy-tailed scenario. Moreover, we show in a simulation section that they can be computed by a simple modification of existing proximal gradient descent algorithms by simply adding a selection step of the central block of data in these algorithms. The resulting algorithms are robust to heavy-tailed data and to few outliers (in both input and output variables) for the examples in Section 3.7.

3.9 Proof main theorems

3.9.1 Proof Theorem 3.1

All along this section we will write $r(\rho)$ for $r(A, \rho)$. Let $\theta = 1/(3A)$. The proof is divided into two parts. First, we identify an event where the RERM $\hat{f} := \hat{f}_\lambda^{RERM}$ is controlled. Then, we prove that this event holds with large probability. Let ρ^* satisfying the A-sparsity Equation from Definition 3.4 and let $\mathcal{B} = \rho^* B \cap r(\rho^*) B_{L_2}$ and consider

$$\Omega := \left\{ \forall f \in F \cap (f^* + \mathcal{B}), \quad \left| (P - P_N)\mathcal{L}_f \right| \leq \theta r^2(\rho^*) \right\} \ .$$

Proposition 3.3. *Let λ be as in* (3.6) *and let ρ^* satisfy the A- sparsity from Definition 3.4. On Ω, one has*

$$\|\hat{f} - f^*\| \leq \rho^*, \quad \|\hat{f} - f^*\|_{L_2} \leq r(\rho^*) \text{ and } P\mathcal{L}_{\hat{f}} \leq A^{-1} r^2(\rho^*) \ .$$

Proof. Prove first that $\hat{f} \in f^* + \mathcal{B}$. Recall that

$$\forall f \in F, \qquad \mathcal{L}_f^\lambda = \mathcal{L}_f + \lambda(\|f\| - \|f^*\|) \ .$$

Since $\hat{f}$ satisfies $P_N \mathcal{L}_{\hat{f}}^\lambda \leqslant 0$, it is sufficient to prove that $P_N \mathcal{L}_f^\lambda > 0$ for all $f \in F \backslash (f^* + \mathcal{B})$ to get the result. The proof relies on the following homogeneity argument. If $P_N \mathcal{L}_{f_0} > 0$ on the border of $f^* + \mathcal{B}$, then $P_N \mathcal{L}_f > 0$ for all $f \in F \setminus \{f^* + \mathcal{B}\}$.

Let $f \in F \setminus \{f^* + \mathcal{B}\}$. By convexity of $\{f^* + \mathcal{B}\} \cap F$, there exists $f_0 \in F$ and $\alpha > 1$ such that $f - f^* = \alpha(f_0 - f^*)$ and $f_0 \in \partial(f^* + \mathcal{B})$ where $\partial(f^* + \mathcal{B})$ denotes the border of $f^* + \mathcal{B}$.

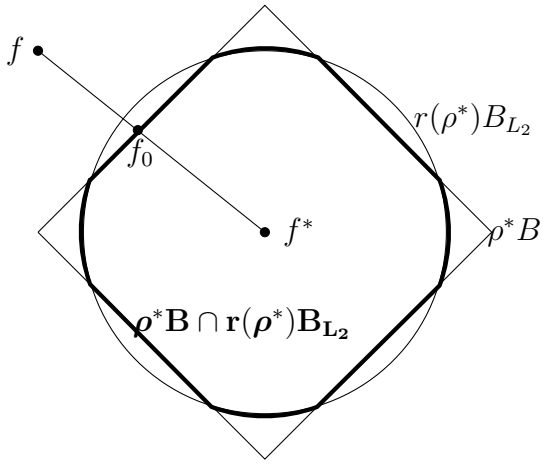


Figure 3.3: Construction of f_0.

For all $i \in \{1, \cdots, N\}$, let $\psi_i : \mathbb{R} \to \mathbb{R}$ be the random function defined for all $u \in \mathbb{R}$ by

$$\psi_i(u) = \ell(u + f^*(X_i), Y_i) - \ell(f^*(X_i), Y_i) \ . \tag{3.32}$$

By construction, for any i, $\psi_i(0) = 0$ and ψ_i is convex because ℓ is. Hence, $\alpha\psi_i(u) \leq \psi_i(\alpha u)$ for all $u \in \mathbb{R}$ and $\alpha \geq 1$. In addition, $\psi_i(f(X_i) - f^*(X_i)) = \ell(f(X_i), Y_i) - \ell(f^*(X_i), Y_i)$. Therefore,

$$\begin{aligned} P_N\mathcal{L}_f &= \frac{1}{N}\sum_{i=1}^{N}\psi_i\big(f(X_i) - f^*(X_i)\big) = \frac{1}{N}\sum_{i=1}^{N}\psi_i\big(\alpha(f_0(X_i) - f^*(X_i))\big) \\ &\geq \frac{\alpha}{N}\sum_{i=1}^{N}\psi_i(f_0(X_i) - f^*(X_i)) = \alpha P_N\mathcal{L}_{f_0} \ . \end{aligned} \tag{3.33}$$

For the regularization term, by the triangular inequality,

$$\|f\| - \|f^*\| = \|f^* + \alpha(f_0 - f^*)\| - \|f^*\| \geq \alpha(\|f_0\| - \|f^*\|) \ .$$

From the latter inequality, together with (3.33), it follows that

$$P_N\mathcal{L}_f^\lambda \geq \alpha P_N\mathcal{L}_{f_0}^\lambda \ . \tag{3.34}$$

As a consequence, if $P_N\mathcal{L}_{f_0}^\lambda > 0$ for all $f_0 \in F \cap \partial(f^* + \mathcal{B})$ then $P_N\mathcal{L}_f^\lambda > 0$ for all $f \in F\backslash(f^* + \mathcal{B})$.

In the remaining of the proof, assume that Ω holds and let $f_0 \in F \cap \partial(f^* + \mathcal{B})$. As $f_0 \in F \cap (f^* + \mathcal{B})$, on Ω,

$$|(P - P_N)\mathcal{L}_{f_0}| \leq \theta r^2(\rho^*) \ . \tag{3.35}$$

By definition of $\mathcal{B}$, as $f_0 \in \partial(f^* + \mathcal{B})$, either: 1) $\|f_0 - f^*\| = \rho^*$ and $\|f_0 - f^*\|_{L_2} \leq r(\rho^*)$ so $\alpha = \|f - f^*\|/\rho^*$ or 2) $\|f_0 - f^*\|_{L_2} = r(\rho^*)$ and $\|f_0 - f^*\| \leq \rho^*$ so $\alpha = \|f - f^*\|_{L_2}/r(\rho^*)$. We treat these cases independently.

Assume first that $\|f_0 - f^*\| = \rho^*$ and $\|f_0 - f^*\|_{L_2} \leq r(\rho^*)$. Let $v \in E$ be such that $\|f^* - v\| \leq \rho^*/20$ and $g \in \partial\|\cdot\|(v)$. We have

$$\begin{aligned} &\|f_0\| - \|f^*\| \geq \|f_0\| - \|v\| - \|f^* - v\| \geq \langle g, f_0 - v\rangle - \|f^* - v\| \\ &\geqslant \langle g, f_0 - f^*\rangle - 2\|f^* - v\| \geqslant \langle g, f_0 - f^*\rangle - \rho^*/10 \ . \end{aligned}$$

As the latter result holds for all $v \in f^* + (\rho^*/20)B$ and $g \in \partial\|\cdot\|(v)$, since $f_0 - f^* \in \rho^* S \cap r(\rho^*)B_{L_2}$, it yields

$$\|f_0\| - \|f^*\| \geq \Delta(\rho^*) - \rho^*/10 \geq 7\rho^*/10 \ . \tag{3.36}$$

Here, the last inequality holds because ρ^* satisfies the sparsity equation. Hence,

$$P_N\mathcal{L}_f^\lambda = P_N\mathcal{L}_f + \lambda(\|f\| - \|f^*\|) \geq \alpha(P_N\mathcal{L}_{f_0} + 7\lambda\rho^*/10) \ . \tag{3.37}$$

Thus, on Ω, since $\lambda > 10\theta r^2(\rho^*)^2/(7\rho^*)$,

$$P_N\mathcal{L}_{f_0} + 7\lambda\rho^*/10 = P\mathcal{L}_{f_0} + (P_N - P)\mathcal{L}_{f_0} + 7\lambda\rho^*/10 \geq -\theta r^2(\rho^*) + 7\lambda\rho^*/10 > 0 \ .$$

Assume now that $\|f_0 - f^*\|_{L_2} = r(\rho^*)$ and $\|f_0 - f^*\| \leq \rho^*$. By Assumption 3.5, on Ω,

$$\begin{aligned} P_N\mathcal{L}_f^\lambda &\geqslant P_N\mathcal{L}_{f_0} - \lambda\|f_0 - f^*\| \geqslant P\mathcal{L}_{f_0} + (P_N - P)\mathcal{L}_{f_0} - \lambda\rho^* \\ &\geqslant A^{-1}\|f_0 - f^*\|_{L_2}^2 - \theta r^2(\rho^*) - \lambda\rho^* \geqslant (A^{-1} - \theta)r^2(\rho^*) - \lambda\rho^* \ . \end{aligned}$$

From (3.6), $\lambda < (A^{-1} - \theta)r^2(\rho^*)^2/\rho^*$, thus $P_N\mathcal{L}_f^\lambda > 0$. Together with (3.37), this proves that $\hat{f} \in f^* + \mathcal{B}$. Now, on Ω, this implies that $|(P - P_N)\mathcal{L}_{\hat{f}}| \leq \theta r^2(\rho^*)$, so by definition of $\hat{f}$,

$$P\mathcal{L}_{\hat{f}} = P_N\mathcal{L}_{\hat{f}}^\lambda + (P - P_N)\mathcal{L}_{\hat{f}} + \lambda(\|f^*\| - \|\hat{f}\|) \leqslant \theta r^2(\rho^*) + \lambda\rho^* \leqslant A^{-1}r^2(\rho^*) \ .$$

■

To prove that Ω holds with large probability, the following result from (Alquier et al., 2019) is useful.

Lemma 3.4. *(Alquier et al., 2019, Lemma 9.1) Grant Assumptions 3.2 and 3.4. Let $F' \subset F$ denote a subset with finite L_2-diameter $d_{L_2}(F')$. For every $u > 0$, with probability at least $1 - 2\exp(-u^2)$*

$$\sup_{f,g\in F'} |(P - P_N)(\mathcal{L}_f - \mathcal{L}_g)| \leq \frac{16LL_0}{\sqrt{N}}\left(w(F') + ud_{L_2}(F')\right) \ .$$

It follows from Lemma 3.4 that for any $u > 0$, with probability larger that $1 - 2\exp(-u^2)$,

$$\begin{aligned} \sup_{f\in F\cap(f^*+\mathcal{B})} \left|(P - P_N)\mathcal{L}_f\right| &\leqslant \sup_{f,g\in F\cap(f^*+\mathcal{B})} \left|(P - P_N)(\mathcal{L}_f - \mathcal{L}_g)\right| \\ &\leqslant \frac{16LL_0}{\sqrt{N}}\big(w(F\cap(f^* + \mathcal{B})) + ud_{L_2}(F\cap(f^* + \mathcal{B}))\big) \ . \end{aligned}$$

It is clear that $d_{L_2}(F \cap (f^* + \mathcal{B})) \leqslant r(\rho^*)$. By definition of the complexity function (3.3), for $u = \theta\sqrt{N}r(\rho^*)/(32LL_0)$, we have with probability at least $1 - 2\exp\big(-\theta^2Nr^2(\rho^*)/(32LL_0)^2\big)$,

$$\forall f \in F \cap (f^* + \mathcal{B}), \qquad \left|(P - P_N)\mathcal{L}_f\right| \leq \theta r^2(\rho^*) \ .$$

3.9.2 Proof Theorem 3.2

All along the proof, the following notations will be used repeatedly.

$$\theta = \frac{1}{34A}, \quad \gamma = \theta/(192L) \quad \hat{f} = \hat{f}_{K,\lambda} \ .$$

The proof is divided into two parts. First, we identify an event where the minmax MOM estimator $\hat{f}$ is controlled. Then, we prove that this event holds with large probability. Let $K \geqslant 7|\mathcal{O}|/3$, and $\kappa \in \{1, 2\}$ let

$$C_{K,r,\kappa} = \max\left(\frac{96L^2K}{\theta^2N}, r_2^2(\gamma, \kappa\rho^*)\right) \quad \text{and} \quad \lambda = 10\theta\frac{C_{K,r,2}}{\rho^*} \ .$$

Let $\mathcal{B}_\kappa = \sqrt{C_{K,r,\kappa}}B_{L_2} \cap \kappa\rho^*B$. Consider the following event

$$\Omega_K = \left\{\forall\kappa \in \{1, 2\},\ \forall f \in F \cap f^* + \mathcal{B}_\kappa,\ \sum_{k=1}^K I\left(\left|(P_{B_k} - P)(\ell_f - \ell_{f^*})\right| \leq \theta C_{K,r,\kappa}\right) \geqslant \frac{K}{2}\right\} \quad (3.38)$$

Deterministic argument

Lemma 3.5. *$\hat{f} - f^* \in \mathcal{B}_\kappa$ if there exists $\eta > 0$ such that*

$$\sup_{f \in f^* + F \setminus \mathcal{B}_\kappa} MOM_K\big(\ell_{f^*} - \ell_f\big) + \lambda\big(\|f^*\| - \|f\|\big) < -\eta \ , \tag{3.39}$$

$$\sup_{f \in F} MOM_K\big(\ell_{f^*} - \ell_f\big) + \lambda\big(\|f^*\| - \|f\|\big) \leq \eta \ . \tag{3.40}$$

Proof. For any $f \in F$, denote by $S(f) = \sup_{g \in F} \mathrm{MOM}_K[\ell_f - \ell_g] + \lambda\big(\|f\| - \|g\|\big)$. If (3.39) holds, by homogeneity of MOM_K, any $f \in f^* + F \setminus \mathcal{B}_\kappa$ satisfies

$$S(f) \geqslant \inf_{f \in f^* + F \setminus \mathcal{B}_\kappa} \mathrm{MOM}_K[\ell_f - \ell_{f^*}] + \lambda\big(\|f\| - \|f^*\|\big) > \eta \ . \tag{3.41}$$

On the other hand, if (3.40) holds,

$$S(f^*) = \sup_{f \in F} \mathrm{MOM}_K[\ell_{f^*} - \ell_f] + \lambda\big(\|f^*\| - \|f\|\big) \leqslant \eta \ .$$

Thus, by definition of $\hat{f}$ and (3.40),

$$S(\hat{f}) \leqslant S(f^*) \leqslant \eta \ .$$

Therefore, if (3.39) and (3.40) hold, $\hat{f} \in f^* + \mathcal{B}_\kappa$. ∎

It remains to show that, on Ω_K, Equations (3.39) and (3.40) hold for $\kappa = 2$.

Let $\kappa \in \{1, 2\}$ and $f \in F \cap \mathcal{B}_\kappa$. On Ω_K, there exist more than $K/2$ blocks B_k such that

$$\Big| (P_{B_k} - P)(\ell_f - \ell_{f^*}) \Big| \leq \theta C_{K,r,\kappa} \ . \tag{3.42}$$

It follows that

$$\sup_{f \in f^* + F \cap \mathcal{B}_\kappa} \mathrm{MOM}_K\big(\ell_{f^*} - \ell_f\big) \leq \theta C_{K,r,\kappa}$$

In addition, $\|f\| - \|f^*\| \leq \kappa \rho^*$. Therefore, from the choice of λ, on Ω_K, one has

$$\sup_{f \in f^* + F \cap \mathcal{B}_\kappa} \mathrm{MOM}_K\big(\ell_{f^*} - \ell_f\big) + \lambda\big(\|f^*\| - \|f\|\big) \leq (1 + 10\kappa)\theta C_{K,r,\kappa} \ . \tag{3.43}$$

Assume that f belongs to $F \setminus \mathcal{B}_\kappa$. By convexity of F, there exists $f_0 \in f^* + F \cap \mathcal{B}_\kappa$ and $\alpha > 1$ such that

$$f = f^* + \alpha(f_0 - f^*) \ . \tag{3.44}$$

For all $i \in \{1, \cdots, N\}$, let $\psi_i : \mathbb{R} \to \mathbb{R}$ be the random function defined for all $u \in \mathbb{R}$ by

$$\psi_i(u) = \ell(u + f^*(X_i), Y_i) - \ell(f^*(X_i), Y_i) \ . \tag{3.45}$$

The functions ψ_i are convex and satisfy $\psi_i(0) = 0$. Thus $\alpha\psi_i(u) \leq \psi_i(\alpha u)$ for all $u \in \mathbb{R}$ and $\alpha > 1$ and $\psi_i(f(X_i) - f^*(X_i)) = \ell(f(X_i), Y_i) - \ell(f^*(X_i), Y_i)$. Hence, for any block B_k,

$$\begin{aligned} P_{B_k}\mathcal{L}_f &= \frac{1}{|B_k|}\sum_{i\in B_k}\psi_i\big(f(X_i) - f^*(X_i)\big) = \frac{1}{|B_k|}\sum_{i\in B_k}\psi_i\big(\alpha(f_0(X_i) - f^*(X_i))\big) \\ &\geq \frac{\alpha}{|B_k|}\sum_{i\in B_k}\psi_i\big(f_0(X_i) - f^*(X_i)\big) = \alpha P_{B_k}\mathcal{L}_{f_0} \ . \end{aligned} \tag{3.46}$$

By the triangular inequality,

$$\|f\| - \|f^*\| = \|f^* + \alpha(f_0 - f^*)\| - \|f^*\| \geq \alpha(\|f_0\| - \|f^*\|).$$

Together with (3.46), this yields, for all block B_k

$$P_{B_k}\mathcal{L}_f^\lambda \geq \alpha P_{B_k}\mathcal{L}_{f_0}^\lambda \ . \tag{3.47}$$

As $f_0 \in F \cap \mathcal{B}_\kappa$, on Ω_K,

$$|(P - P_{B_k})\mathcal{L}_{f_0}| \leq \theta C_{K,r,\kappa}. \tag{3.48}$$

As f_0 can be chosen in $\partial(f^* + \mathcal{B}_\kappa)$, either: 1) $\|f_0 - f^*\| = \kappa\rho^*$ and $\|f_0 - f^*\|_{L_2} \leq \sqrt{C_{K,r,\kappa}}$ or 2) $\|f_0 - f^*\|_{L_2} = \sqrt{C_{K,r,\kappa}}$ and $\|f_0 - f^*\| \leq \kappa\rho^*$.

Assume first that $\|f_0 - f^*\| = \kappa\rho^*$ and $\|f_0 - f^*\|_{L_2} \leq \sqrt{C_{K,r,\kappa}}$. Since the sparsity equation is satisfied for $\rho = \rho^*$, it is also satisfied for $\kappa\rho^*$. By (3.36),

$$\lambda\big(\,\|f_0\| - \|f^*\|\,\big) \geq 7\lambda\kappa\rho^*/10 = 7\kappa C_{K,r,2} \ . \tag{3.49}$$

Therefore, on Ω_K, there are more than $K/2$ blocks B_k where

$$P_{B_k}\mathcal{L}_f^\lambda \geq \alpha P_{B_k}\mathcal{L}_{f_0}^\lambda \geq \alpha\bigg(-\theta C_{K,r,\kappa} + \frac{7\kappa\lambda\rho^*}{10}\bigg) \geqslant \alpha(7\kappa - 1)\theta C_{K,r,2} \ . \tag{3.50}$$

It follows that

$$\mathrm{MOM}_K\big(\ell_f - \ell_{f^*}\big) + \lambda\big(\|f\| - \|f^*\|\big) \geqslant \alpha\theta\big(7\kappa C_{K,r,2} - C_{K,r,\kappa}\big)C_{K,r,2} \ . \tag{3.51}$$

Assume that $\|f_0 - f^*\|_{L_2} = \sqrt{C_{K,r,\kappa}}$ and $\|f_0 - f^*\| \leq \kappa\rho^*$. By Assumption 3.7, on Ω_K, there exist more than $K/2$ blocks B_k where

$$\begin{aligned} P_{B_k}\mathcal{L}_f^\lambda &\geqslant P_{B_k}\mathcal{L}_{f_0} - \lambda\,\|f_0 - f^*\| \geq P\mathcal{L}_{f_0} + (P_{B_k} - P)\mathcal{L}_{f_0} - \lambda\kappa\rho^* \\ &\geq A^{-1}\,\|f_0 - f^*\|_{L_2}^2 - \theta C_{K,r,\kappa} - \kappa\lambda\rho^* = \theta(33C_{K,r,\kappa} - 10\kappa C_{K,r,2}) \ . \end{aligned}$$

It follows that

$$\mathrm{MOM}_K\big(\ell_f - \ell_{f^*}\big) + \lambda\big(\|f\| - \|f^*\|\big) \geqslant \alpha\theta(33C_{K,r,\kappa} - 10\kappa C_{K,r,2}) \ . \tag{3.52}$$

From Equations (3.43), (3.51) and (3.52) with $\kappa = 1$, it follows that

$$\sup_{f\in F}\mathrm{MOM}_K\big(\ell_{f^*} - \ell_f\big) + \lambda\big(\|f^*\| - \|f\|\big) \leq 11\theta C_{K,r,2} \ . \tag{3.53}$$

Therefore, (3.40) holds with $\eta = 11\theta C_{K,r,2}$. Now, Equations (3.51) and (3.52) with $\kappa = 2$ yield

$$\sup_{f\in f^*+F\setminus\mathcal{B}_2} \mathrm{MOM}_K\big(\ell_{f^*}-\ell_f\big)+\lambda\big(\|f^*\|-\|f\|\big) \leqslant -13\alpha\theta C_{K,r,2} < -11\theta C_{K,r,2} \ .$$

Therefore, Equation (3.39) holds with $\eta = 11\theta C_{K,r,2}$. Overall, Lemma 3.5 shows that $\hat{f} \in \mathcal{B}_2$. On Ω_K, this implies that there exist more than $K/2$ blocks B_k where $P\mathcal{L}_{\hat{f}} \leq P_{B_k}\mathcal{L}_{\hat{f}} + \theta C_{K,r,2}$. In addition, by definition of $\hat{f}$ and (3.53),

$$\mathrm{MOM}_K\big(\ell_{\hat{f}}-\ell_{f^*}\big)+\lambda(\|\hat{f}\|-\|f^*\|) \leqslant \sup_{f\in F}\mathrm{MOM}_K\big(\ell_{f^*}-\ell_f\big)+\lambda(\|f^*\|-\|f\|) \leq 11\theta C_{K,r,2} \ .$$

This means that there exist at least $K/2$ blocks B_k where $P_{B_k}\mathcal{L}_{\hat{f}} + \lambda(\|\hat{f}\|-\|f^*\|) \leqslant 11\theta C_{K,r,2}$. As $\|\hat{f}\|-\|f^*\| \geqslant -\|\hat{f}-f^*\| \geqslant -2\rho^*$, on these blocks, $P_{B_k}\mathcal{L}_{\hat{f}} \leq 31\theta C_{K,r,2}$. Therefore, there exists at least one block B_k for which simultaneously $P\mathcal{L}_{\hat{f}} \leq P_{B_k}\mathcal{L}_{\hat{f}} + \theta C_{K,r,2}$ and $P_{B_k}\mathcal{L}_{\hat{f}} \leq 31\theta C_{K,r,2}$. This shows that $P\mathcal{L}_{\hat{f}} \leq 32\theta C_{K,r,2} \leq A^{-1}C_{K,r,2}$.

Control of the stochastic event

Proposition 3.4. *Grant Assumptions 3.2, 3.3, 3.6 and 3.7. Let $K \geq 7|\mathcal{O}|/3$. Then Ω_K holds with probability larger than $1-2\exp(-K/504)$.*

Proof. Let $\mathcal{F} = F \cap (f^* + \mathcal{B}_\kappa)$ and let $\phi(t) = \mathbb{1}\{t \geq 2\} + (t-1)\mathbb{1}\{1 \leq t \leq 2\}$. This function satisfies $\forall t \in \mathbb{R}^+ \quad \mathbb{1}\{t \geq 2\} \leq \phi(t) \leq \mathbb{1}\{t \geq 1\}$. Let $W_k = ((X_i, Y_i))_{i\in B_k}$ and, for any $f \in \mathcal{F}$, let $G_f(W_k) = (P_{B_k} - P)(\ell_f - \ell_{f^*})$. Let also $C_{K,r,\kappa} = \max\left(96L^2K/(\theta^2 N), r_2^2(\gamma, \kappa\rho^*)\right)$. For any $f \in \mathcal{F}$, let

$$z(f) = \sum_{k=1}^{K} \mathbb{1}\{|G_f(W_k)| \leq \theta C_{K,r,\kappa}\} \ .$$

Proposition 3.4 will be proved if $z(f) \geqslant K/2$ with probability larger than $1-e^{-K/504}$. Let $\mathcal{K}$ denote the set of indices of blocks which have not been corrupted by outliers, $\mathcal{K} = \{k \in \{1, \cdots, K\} : B_k \subset \mathcal{I}\}$, where we recall that $\mathcal{I}$ is the set of informative data. Basic algebraic manipulations show that

$$z(f) \geqslant |\mathcal{K}| - \sup_{f\in\mathcal{F}} \sum_{k\in\mathcal{K}} \left(\phi(2(\theta C_{K,r,\kappa})^{-1}|G_f(W_k)|) - \mathbb{E}\phi(2(\theta C_{K,r,\kappa})^{-1}|G_f(W_k)|)\right) - \sum_{k\in\mathcal{K}} \mathbb{E}\phi(2(\theta C_{K,r,\kappa})^{-1}|G_f(W_k)|) \ . \quad (3.54)$$

The last term in (3.54) can be bounded from below as follows. Let $f \in \mathcal{F}$ and $k \in \mathcal{K}$,

$$\mathbb{E}\phi(2(\theta C_{K,r,\kappa})^{-1}|G_f(W_k)|) \leqslant \mathbb{P}\left(|G_f(W_k)| \geq \frac{\theta C_{K,r,\kappa}}{2}\right) \leqslant \frac{4\mathbb{E}G_f(W_k)^2}{(\theta C_{K,r,\kappa})^2}$$
$$\leqslant \frac{4K^2}{\theta^2 C_{K,r,\kappa}^2 N^2} \sum_{i\in B_k} \mathbb{E}[(\ell_f - \ell_{f^*})^2(X_i, Y_i)] \leq \frac{4L^2K}{\theta^2 C_{K,r,\kappa}^2 N}\|f - f^*\|_{L_2}^2 \ .$$

The last inequality follows from Assumption 3.6. Since $\|f - f^*\|_{L_2} \leq \sqrt{C_{K,r,\kappa}}$,

$$\mathbb{E}\phi(2(\theta C_{K,r,\kappa})^{-1}|G_f(W_k)|) \leqslant \frac{4L^2K}{\theta^2 C_{K,r,\kappa}N} \ .$$

As $C_{K,r,\kappa} \geqslant 96L^2K/(\theta^2 N)$,

$$\mathbb{E}\phi(2(\theta C_{K,r,\kappa})^{-1}|G_f(W_k)|) \leq \frac{1}{24} \ .$$

Plugging this inequality in (3.54) yields

$$z(f) \geq |\mathcal{K}|(1 - \frac{1}{24}) - \sup_{f\in\mathcal{F}} \sum_{k\in\mathcal{K}} \left(\phi(2(\theta C_{K,r,\kappa})^{-1}|G_f(W_k)|) - \mathbb{E}\phi(2(\theta C_{K,r,\kappa})^{-1}|G_f(W_k)|) \right) \ . \quad (3.55)$$

Using the Mc Diarmid's inequality, with probability larger than $1 - \exp(-|\mathcal{K}|/288)$,we get

$$\begin{aligned} &\sup_{f\in\mathcal{F}} \sum_{k\in\mathcal{K}} \left(\phi(2(\theta C_{K,r,\kappa})^{-1}|G_f(W_k)|) - \mathbb{E}\phi(2(\theta C_{K,r,\kappa})^{-1}|G_f(W_k)|) \right) \\ &\quad \leq \frac{|\mathcal{K}|}{24} + \mathbb{E}\sup_{f\in\mathcal{F}} \sum_{k\in\mathcal{K}} \left(\phi(2(\theta C_{K,r,\kappa})^{-1}|G_f(W_k)|) - \mathbb{E}\phi(2(\theta C_{K,r,\kappa})^{-1}|G_f(W_k)|) \right) \ . \end{aligned}$$

By the symmetrization lemma, it follows that, with probability larger than $1 - \exp(-|\mathcal{K}|/288)$,

$$\begin{aligned} &\sup_{f\in\mathcal{F}} \sum_{k\in\mathcal{K}} \left(\phi(2(\theta C_{K,r,\kappa})^{-1}|G_f(W_k)|) - \mathbb{E}\phi(2(\theta C_{K,r,\kappa})^{-1}|G_f(W_k)|) \right) \\ &\qquad \leqslant \frac{|\mathcal{K}|}{24} + 2\mathbb{E}\sup_{f\in\mathcal{F}} \sum_{k\in\mathcal{K}} \sigma_k \phi(2(\theta C_{K,r,\kappa})^{-1}|G_f(W_k)|) \ . \end{aligned}$$

As ϕ is 1-Lipschitz with $\phi(0) = 0$, the contraction lemma from (Ledoux and Talagrand, 2013)and yields

$$\begin{aligned} &\sup_{f\in\mathcal{F}} \sum_{k\in\mathcal{K}} \left(\phi(2(\theta C_{K,r,\kappa})^{-1}|G_f(W_k)|) - \mathbb{E}\phi(2(\theta C_{K,r,\kappa})^{-1}|G_f(W_k)|) \right) \\ &\qquad \leqslant \frac{|\mathcal{K}|}{24} + \frac{4}{\theta}\mathbb{E}\sup_{f\in\mathcal{F}} \sum_{k\in\mathcal{K}} \sigma_k \frac{G_f(W_k)}{C_{K,r,\kappa}} \\ &\qquad = \frac{|\mathcal{K}|}{24} + \frac{4}{\theta}\mathbb{E}\sup_{f\in\mathcal{F}} \sum_{k\in\mathcal{K}} \sigma_k \frac{(P_{B_k} - P)(\ell_f - \ell_{f^*})}{C_{K,r,\kappa}} \end{aligned}$$

For any $k \in \mathcal{K}$, let $(\sigma_i)_{i\in B_k}$ independent from $(\sigma_k)_{k\in\mathcal{K}}$, $(X_i)_{i\in\mathcal{I}}$ and $(Y_i)_{i\in\mathcal{I}}$. The vectors $(\sigma_i\sigma_k(\ell_f - \ell_{f^*})(X_i, Y_i))_{i,f}$ and $(\sigma_i(\ell_f - \ell_{f^*})(X_i, Y_i))_{i,f}$ have the same distribution. Thus, by the symmetrization

and contraction lemmas, with probability larger than $1-\exp(-|\mathcal{K}|/288)$,

$$\begin{aligned}\sup_{f\in\mathcal{F}}\sum_{k\in\mathcal{K}}&\left(\phi(2C_{K,r,\kappa}^{-1}|G_f(W_k)|)-\mathbb{E}\phi(2C_{K,r,\kappa}^{-1}|G_f(W_k)|)\right)\\ &\leq \frac{|\mathcal{K}|}{24}+\frac{8}{\theta}\mathbb{E}\sup_{f\in\mathcal{F}}\sum_{k\in\mathcal{K}}\frac{1}{|B_k|}\sum_{i\in B_k}\sigma_i\frac{(\ell_f-\ell_{f^*})(X_i,Y_i)}{C_{K,r,\kappa}}\\ &= \frac{|\mathcal{K}|}{24}+\frac{8K}{\theta N}\mathbb{E}\sup_{f\in\mathcal{F}}\sum_{i\in\cup_{k\in\mathcal{K}}B_k}\sigma_i\frac{(\ell_f-\ell_{f^*})(X_i,Y_i)}{C_{K,r,\kappa}}\\ &\leq \frac{|\mathcal{K}|}{24}+\frac{8LK}{\theta N}\mathbb{E}\sup_{f\in\mathcal{F}}\left|\sum_{i\in\cup_{k\in\mathcal{K}}B_k}\sigma_i\frac{(f-f^*)(X_i)}{C_{K,r,\kappa}}\right| . \end{aligned}\tag{3.56}$$

Now either 1) $K\leq\theta^2r_2^2(\gamma,\kappa\rho^*)N/(96L^2)$ or 2) $K>\theta^2r_2^2(\gamma,\kappa\rho^*)N/(96L^2)$. Assume first that $K\leq\theta^2r_2^2(\gamma,\kappa\rho^*)N/(96L^2)$, so $C_{K,r,\kappa}=r_2^2(\gamma,\kappa\rho^*)$ and by definition of the complexity parameter

$$\mathbb{E}\sup_{f\in\mathcal{F}}\left|\sum_{i\in\cup_{k\in\mathcal{K}}B_k}\sigma_i\frac{(f-f^*)(X_i)}{C_{K,r,\kappa}}\right|=\mathbb{E}\sup_{f\in\mathcal{F}}\frac{1}{r_2^2(\gamma,\kappa\rho^*)}\left|\sum_{i\in\cup_{k\in\mathcal{K}}B_k}\sigma_i(f-f^*)(X_i)\right|\leq\frac{\gamma|\mathcal{K}|N}{K} .$$

If $K>\theta^2r_2^2(\gamma,\kappa\rho^*)N/(96L^2)$, $C_{K,r,\kappa}=96L^2K/(\theta^2N)$. Write $\mathcal{F}=\mathcal{F}_1\cup\mathcal{F}_2$, where

$$\mathcal{F}_1:=\{f\in\mathcal{F}:\|f-f^*\|_{L_2}\leqslant r_2(\gamma,\kappa\rho^*)\},\qquad\mathcal{F}_2=\mathcal{F}\setminus\mathcal{F}_1 .$$

Then,

$$\begin{aligned}\mathbb{E}\sup_{f\in\mathcal{F}}&\left|\sum_{i\in\cup_{k\in\mathcal{K}}B_k}\sigma_i\frac{(f-f^*)(X_i)}{C_{K,r,\kappa}}\right|\\ &=\frac{1}{C_{K,r,\kappa}}\mathbb{E}\left[\sup_{f\in\mathcal{F}_1}\left|\sum_{i\in\cup_{k\in\mathcal{K}}B_k}\sigma_i(f-f^*)(X_i)\right|\vee\sup_{f\in\mathcal{F}_2}\left|\sum_{i\in\cup_{k\in\mathcal{K}}B_k}\sigma_i(f-f^*)(X_i)\right|\right] .\end{aligned}$$

For any $f\in\mathcal{F}_2$, $g=f^*+(f-f^*)r_2(\gamma,\kappa\rho^*)/\sqrt{C_{K,r,\kappa}}\in\mathcal{F}_1$ and

$$\left|\sum_{i\in\cup_{k\in\mathcal{K}}B_k}\sigma_i(f-f^*)(X_i)\right|=\frac{\sqrt{C_{K,r,\kappa}}}{r_2(\gamma,\kappa\rho^*)}\left|\sum_{i\in\cup_{k\in\mathcal{K}}B_k}\sigma_i(g-f^*)(X_i)\right| .$$

It follows that

$$\sup_{f\in\mathcal{F}_2}\left|\sum_{i\in\cup_{k\in\mathcal{K}}B_k}\sigma_i(f-f^*)(X_i)\right|\leqslant\frac{\sqrt{C_{K,r,\kappa}}}{r_2(\gamma,\kappa\rho^*)}\sup_{f\in\mathcal{F}_1}\left|\sum_{i\in\cup_{k\in\mathcal{K}}B_k}\sigma_i(f-f^*)(X_i)\right| .$$

Hence,

$$\mathbb{E}\sup_{f\in\mathcal{F}}\left|\sum_{i\in\cup_{k\in\mathcal{K}}B_k}\sigma_i\frac{(f-f^*)(X_i)}{C_{K,r,\kappa}}\right|\leqslant\frac{1}{r_2(\gamma,\kappa\rho^*)\sqrt{C_{K,r,\kappa}}}\mathbb{E}\sup_{f\in\mathcal{F}_1}\left|\sum_{i\in\cup_{k\in\mathcal{K}}B_k}\sigma_i(f-f^*)(X_i)\right| .$$

By definition of r_2, this implies

$$\mathbb{E}\sup_{f\in\mathcal{F}}\left|\sum_{i\in\cup_{k\in\mathcal{K}}B_k}\sigma_i\frac{(f-f^*)(X_i)}{C_{K,r,\kappa}}\right|\leqslant\frac{r_2(\gamma,\kappa\rho^*)}{\sqrt{C_{K,r,\kappa}}}\frac{\gamma|\mathcal{K}|N}{K}\leqslant\frac{\gamma|\mathcal{K}|N}{K} .$$

Plugging this bound in (3.56) yields, with probability larger than $1 - e^{-|\mathcal{K}|/288}$

$$\sup_{f\in\mathcal{F}} \sum_{k\in\mathcal{K}} \left(\phi(2C^{-1}_{K,r,\kappa}|G_f(W_k)|) - \mathbb{E}\phi(2C^{-1}_{K,r,\kappa}|G_f(W_k)|) \right) \leqslant |\mathcal{K}| \left(\frac{1}{24} + \frac{8L\gamma}{\theta} \right) = \frac{|\mathcal{K}|}{12} \ .$$

Plugging this inequality into (3.55) shows that, with probability at least $1 - e^{-|\mathcal{K}|/288}$,

$$z(f) \geqslant \frac{7|\mathcal{K}|}{8} \ .$$

As $K \geqslant 7|\mathcal{O}|/3$, $|\mathcal{K}| \geqslant K - |\mathcal{O}| \geqslant 4K/7$, hence, $z(f) \geqslant K/2$ holds with probability at least $1 - e^{-K/504}$. Since it has to hold for any κ in $\{1, 2\}$, the final probablity is $1 - 2e^{-K/504}$. ∎

3.9.3 Proof Theorem 3.3

The proof is very similar to the one of Theorem 3.2. We only present the different arguments we use coming from the localization with the excess risk. The proof is split into two parts. First we identify an event $\bar{\Omega}_K$ in the same way is Ω_K in (3.38) where the L_2-localization is replaced by the excess risk localization. For $\kappa \in \{1, 2\}$ let $\mathcal{B}_\kappa = \{f \in E : \ P\mathcal{L}_f \leq \bar{r}^2(\gamma, \kappa\rho^*), \ \|f - f^*\| \leq \ \kappa\rho^*\}$ and

$$\bar{\Omega}_K = \left\{ \forall \kappa \in \{1, 2\}, \forall f \in F \cap \mathcal{B}_\kappa, \sum_{k=1}^{K} I\big\{ |(P_{B_k} - P)\mathcal{L}_f| \leq \frac{1}{20}\bar{r}^2(\gamma, 2\rho^*) \big\} \geq K/2 \right\}$$

Let us us the following notations,

$$\lambda = \frac{11\bar{r}^2(\gamma, 2\rho^*)}{40\rho^*}, \quad \hat{f} = \hat{f}^{\lambda}_K \quad \text{and} \quad \gamma = 1/3840L$$

Finally recal that the complexity parameter is defined as

$$\bar{r}(\gamma, \rho) = \inf \left\{ r > 0 : \max \left(\frac{E(r, \rho)}{\gamma}, \sqrt{384000} V_K(r, \rho) \right) \leq r^2 \right\}$$

where

$$E(r, \rho) = \sup_{J\subset\mathcal{I}:|J|\geq N/2} \mathbb{E} \sup_{f\in F: P\mathcal{L}_f \leq r^2, \, \|f-f^*\|\leq\rho} \left| \frac{1}{|J|} \sum_{i\in J} \sigma_i (f - f^*)(X_i) \right|$$

$$V_K(r, \rho) = \max_{i\in\mathcal{I}} \sup_{f\in F: P\mathcal{L}_f \leq r^2, \, \|f-f^*\|\leq\rho} \left(\sqrt{\mathbb{V}ar_{P_i}(\mathcal{L}_f)} \right) \sqrt{\frac{K}{N}}$$

First, we show that on the event $\bar{\Omega}_K$, $P\mathcal{L}_{\hat{f}} \leq \bar{r}^2(\gamma, 2\rho^*)$ and $\|f - f^*\| \leq 2\rho^*$. Then we will control the probability of $\bar{\Omega}_K$.

Lemma 3.6. *Grant Assumptions 3.2 and 3.3. Let ρ^* satisfy the sparsity equation from Definition 3.6. On the event $\bar{\Omega}_K$, $P\mathcal{L}_{\hat{f}} \leq \bar{r}^2(\gamma, 2\rho^*)$ and $\|f - f^*\| \leq 2\rho^*$.*

Proof. Let $f \in F\backslash\mathcal{B}_\kappa$. From Lemma 6 in (Chinot et al., 2019b) there exist $f_0 \in F$ and $\alpha > 0$ such that $f - f^* = \alpha(f_0 - f^*)$ and $f_0 \in \partial\mathcal{B}_\kappa$. By definition of $\mathcal{B}_\kappa$, either 1)$P\mathcal{L}_{f_0} = \bar{r}^2(\gamma, \kappa\rho^*)$ and $\|f_0 - f^*\| \leq \kappa\rho^*$ or 2) $P\mathcal{L}_{f_0} \leq \bar{r}^2(\gamma, \kappa\rho^*)$ and $\|f_0 - f^*\| = \kappa\rho^*$.

Assume that $P\mathcal{L}_{f_0} = \bar{r}^2(\gamma, \kappa\rho^*)$ and $\|f_0 - f^*\| \leq \kappa\rho^*$. On $\bar{\Omega}_K$, there exist at least $K/2$ blocks B_k such that $P_{B_k}\mathcal{L}_{f_0} \geq P\mathcal{L}_{f_0} - (1/20)\bar{r}^2(\gamma, \kappa\rho^*) = (19/20)\bar{r}^2(\gamma, \kappa\rho^*)$. It follows that, on at least $K/2$ blocks B_k

$$P_{B_k}\mathcal{L}_f^\lambda \geq \alpha P_{B_k}\mathcal{L}_{f_0}^\lambda = \alpha\big(P_{B_k}\mathcal{L}_{f_0} + \lambda(\|f_0\| - \|f^*\|)\big) \geq (19/20)\bar{r}^2(\gamma, \kappa\rho^*) - 11\kappa\bar{r}^2(\gamma, 2\rho^*)/40 \quad (3.57)$$

Assume that $P\mathcal{L}_{f_0} \leq \bar{r}^2(\gamma, \kappa\rho^*)$ and $\|f_0 - f^*\| = \kappa\rho^*$. From the sparsity equation defined in Definition 3.6 we get $\|f_0\| - \|f^*\| \geq 7\kappa\rho^*/10$. And on more than $K/2$ blocks B_k

$$P_{B_k}\mathcal{L}_f^\lambda \geq -(1/20)\bar{r}^2(\gamma, \kappa\rho^*) + 7\lambda\kappa\rho^*/10 = -(1/20)\bar{r}^2(\gamma, \kappa\rho^*) + 77\kappa\bar{r}^2(\gamma, 2\rho^*)/400 \quad (3.58)$$

Now let us consider $f \in F \cap \mathcal{B}_\kappa$. On $\bar{\Omega}_K$, there exist at least $K/2$ blocks B_k such that

$$P_{B_k}\mathcal{L}_f^\lambda \geq -(1/20)\bar{r}^2(\gamma, \kappa\rho^*) - \lambda\kappa\rho^* = -(1/20)\bar{r}^2(\gamma, \kappa\rho^*) - 11\kappa\bar{r}^2(\gamma, 2\rho^*)/40 \quad (3.59)$$

As Equations (3.57), (3.58) and (3.59) hold for more than $K/2$ blocks it follows for $\kappa = 1$ that

$$\sup_{f\in F} \mathrm{MOM}_K\big(\ell_{f^*} - \ell_f\big) + \lambda(\|f^*\| - \|f\|) \leq (13/40)\bar{r}^2(\gamma, 2\rho^*) \ . \quad (3.60)$$

From Equations (3.57), (3.58) and (3.59) with $\kappa = 2$ we get

$$\sup_{f\in F\backslash\mathcal{B}_2} \mathrm{MOM}_K\big(\ell_{f^*} - \ell_f\big) + \lambda(\|f^*\| - \|f\|) < (13/40)\bar{r}^2(\gamma, 2\rho^*) \ . \quad (3.61)$$

From Equations (3.60) and (3.61) and a slight modification of Lemma 3.5 it easy to see that on $\bar{\Omega}_K$, $P\mathcal{L}_{\hat{f}} \leq \bar{r}^2(\gamma, 2\rho^*)$ and $\|f - f^*\| \leq \rho^*$. ∎

Proposition 3.5. *Grant Assumptions 3.2, 3.3 and 3.8. Then $\bar{\Omega}_K$ holds with probability larger than $1 - 2\exp(-cK)$*

Sketch of proof. The proof of Proposition 3.5 follows the same line as the one of Proposition 3.4. Let us precise the main differences. For all $f \in F \cap \mathcal{B}_\kappa$ we set, $z'(f) = \sum_{k=1}^K I\{|G_f(W_k)| \leq (1/20)\bar{r}^2(\gamma, \kappa\rho^*)\}$ where $G_f(W_k)$ is the same quantity as in the proof of Proposition 3.4. Let us consider the contraction ϕ introduced in Proposition 3.4. By definition of $V_K(r)$ and $\bar{r}^2(\gamma, \kappa\rho^*)$ we have

$$\begin{aligned}
\mathbb{E}\phi(40|G_f(W_k)|/\bar{r}^2(\gamma, \kappa\rho^*)) &\leq \mathbb{P}\left(|G_f(W_k)| \geq \frac{\bar{r}^2(\gamma, \kappa\rho^*)}{40}\right) \leq \frac{(40)^2}{\bar{r}^4(\gamma, \kappa\rho^*)}\mathbb{E}G_f(W_k)^2 \\
&= \frac{(40)^2}{\bar{r}^4(\gamma, \kappa\rho^*)}\mathbb{V}ar(P_{B_k}\mathcal{L}_f) \leq \frac{(40)^2K^2}{\bar{r}^4(\gamma, \kappa\rho^*)N^2}\sum_{i\in B_k}\mathbb{V}ar_{P_i}(\mathcal{L}_f) \\
&\leq \frac{(40)^2K}{\bar{r}^4(\gamma, \kappa\rho^*)N}\sup\{\mathbb{V}ar_{P_i}(\mathcal{L}_f) : f \in F \cap \mathcal{B}_\kappa, i \in \mathcal{I}\} \leq 1/24 \ .
\end{aligned}$$

Using Mc Diarmid's inequality, the Giné-Zinn symmetrization argument and the contraction lemma twice and the Lipschitz property of the loss function, such as in the proof of Proposition 3.4, we obtain for all $x > 0$, with probability larger than $1 - \exp(-|\mathcal{K}|/288)$, for all $f \in \mathcal{F}'$,

$$z'(f) \geq 11|\mathcal{K}|/12 - \frac{160LK}{\theta N}\mathbb{E}\sup_{f\in F\cap\mathcal{B}_\kappa}\frac{1}{\bar{r}^2(\gamma,\kappa\rho^*)}\left|\sum_{i\in\cup_{k\in\mathcal{K}}B_k}\sigma_i(f-f^*)(X_i)\right|. \tag{3.62}$$

From the definition of $\bar{r}^2(\gamma,\kappa\rho^*)$ it follows that $\mathbb{E}\sup_{f\in F\cap\mathcal{B}_\kappa}\left|\sum_{i\in\cup_{k\in\mathcal{K}}B_k}\sigma_i(f-f^*)(X_i)\right| \leq \gamma\bar{r}^2(\gamma,\kappa\rho^*)$ and $z'(f) \geq |\mathcal{K}|(11/12 - 160L^2\gamma) = 7|\mathcal{K}|/8$. The rest of the proof is totally similar.

3.9.4 Proof of Theorem 3.4

From Assumption 3.2, it holds $V_K(r) \leq LV'_K(r)$, where for all $r > 0$,

$$V'_K(r) = \sqrt{K/N}\max_{i\in\mathcal{I}}\sup_{f\in F:P\mathcal{L}_f\leq r^2,\,\|f-f^*\|\leq\rho}\|f-f^*\|_{L_2} \ .$$

By Assumption 3.9,

$$\sqrt{c}V_K\left(\sqrt{384000}L\sqrt{\frac{\bar{A}K}{N}},2\rho^*\right) \leq 384000L^2\frac{\bar{A}K}{N} \ .$$

From the definition of $r_2^2(\gamma,2\rho^*)$ and Assumption 3.9, it follows

$$\frac{1}{\gamma}E\left(\frac{r_2(\gamma/\bar{A},2\rho^*)}{\sqrt{\bar{A}}}\right) \leq \frac{r_2^2(\gamma/\bar{A},2\rho^*)}{\bar{A}} \ .$$

Hence, $\bar{r}^2(\gamma,2\rho^*) \leq \max\left(r_2^2(\gamma/\bar{A},2\rho^*)/\sqrt{\bar{A}}, 384000L^2\bar{A}K/N\right)$ and the proof is complete.

Chapter 4

Robust learning and complexity dependent bounds for regularized problems

We study Regularized Empirical Risk Minimizers (RERM) and minmax Median-Of-Means (MOM) estimators where the regularization function $\phi(\cdot)$ is an even convex function. We obtain bounds on the L_2-estimation error and the excess risk that depend on $\phi(f^*)$, where f^* is the minimizer of the risk over a class F. The estimators are based on loss functions that are both Lipschitz and convex. Results for the RERM are derived under weak assumptions on the outputs and a sub-Gaussian assumption on the class $\{(f - f^*)(X), f \in F\}$. Similar results are shown for minmax MOM estimators in a close setting where outliers may have corrupted the dataset and where the class $\{(f - f^*)(X), f \in F\}$ is only supposed to satisfy weak moment assumptions, relaxing the sub-Gaussian and the i.i.d hypothesis necessary for RERM. The analysis of RERM and minmax MOM estimators with Lipschitz and convex loss functions is based on a weak local Bernstein Assumption. We obtain two "meta theorems" that we use to study linear estimators regularized by the Elastic Net. We also examine Support Vector Machines (SVM), where no sub-Gaussian assumption is required and when the target Y can be heavy-tailed, improving the existing literature.

4.1 Introduction

On one hand, real world data analysis problems require nonlinear methods to model complex dependencies between random variables. On the other hand, linear models are well-understood and easy to implement, even in high dimension (Bishop, 2006). Over the last two decades, learning with positive definite kernels have become very popular in machine learning (Shawe-Taylor et al., 2004; Schölkopf et al., 1999; Steinwart and Christmann, 2008). This popularity can be explained because kernel methods combine these advantages. Kernels can be used to model non linear dependencies, mapping them to a (usually high-dimensional) feature space. In this space, the estimation is linear. In this sense, kernel methods extend well-understood, linear statistical learning technics to real-world, complicated, structured, high-dimensional data based on a rigorous mathematical framework leading to practical modelling tools and algorithms. They have been used in many different fields such as finance (Chalup and Mitschele, 2008), biology (Schölkopf et al., 2004; Ben-Hur and Noble, 2005; Noble et al., 2004), econometric (Li and Racine, 2007), computer vision (Yang et al., 2000). Let (X, Y) be a random variable with distribution P and $\mathcal{H}_K$ a reproducing Kernel Hilbert Space (RKHS) associated to a positive definite kernel K. Kernel methods consist in computing f^* in $\mathcal{H}_K$ such that the *risk* $\mathcal{R}(f) := \mathbb{E}_{(X,Y)\sim P}[\ell(f(X), Y)]$ is minimized in f^*, where $\ell(f(X), Y)$ measures the error of predicting $f(X)$ while the true answer is Y. However, the distribution P is unknown and the minimization of the risk, necessary to compute f^*, is impossible in practice. To proceed, one is given a dataset $\mathcal{D} = (X_i, Y_i)_{i=1}^N$ of random variables. Using the dataset $\mathcal{D}$, kernel methods compute $\hat{f}_N^\lambda$ in $\mathcal{H}_K$ such that

$$\hat{f}_N^\lambda \in \operatorname*{argmin}_{f\in\mathcal{H}_K} \frac{1}{N}\sum_{i=1}^N \ell(f(X_i), Y_i) + \lambda\|f\|_{\mathcal{H}_K}^2 \ , \tag{4.1}$$

where $\|f\|_{\mathcal{H}_K}$ is the norm of f in $\mathcal{H}_K$ and $\lambda \geq 0$ is an hyper-parameter to be tuned. The regularization term $\lambda\|f\|_{\mathcal{H}_K}^2$ controls the smoothness of $\hat{f}_N^\lambda$ through the value of λ. This regularization term is introduced to avoid "overfitting" since kernels provide enough flexibility to fit training data exactly. The value of λ balance the bias and the variance of $\hat{f}_N^\lambda$. Theoritical properties of kernel methods have been widely studied (Shawe-Taylor et al., 2004; Schölkopf et al., 1999; Steinwart and Christmann, 2008). Non-asymptotic bounds on the $L_2(\mu)$-*error rate* $\|f^* - \hat{f}_N^\lambda\|_{L_2(\mu)}$, where μ denotes the marginal distribution of X, have been obtained for the quadratic loss function in (Mendelson et al., 2010; Smale and Zhou, 2007). These bounds depend on the decay of eigenvalues of the kernel (at the population level) and are obtained for bounded continuous kernels but under the restrictive assumption that the random variable $Y \in [-M, M]$ almost surely. In (Caponnetto and De Vito, 2007), also for the quadratic loss function, the authors do not assume that $|Y|$ is bounded but that $Y - f^*(X)$ admits a Laplace transform. In this paper, we recover the same error rates as (Mendelson et al., 2010; Caponnetto and De Vito, 2007) when the loss function ℓ is simultaneously Lipschitz and convex. We do not assume that Y is bounded or $Y - f^*(X)$ is light-tailed. Our analysis uses a new localization technique developed in (Chinot et al., 2019b) taking advantage of the convexity

of the loss function ℓ. Theorem 4.1 presents an informal result when ℓ is the absolute loss function.

Theorem 4.1 (Informal)**.** *Let K be a bounded kernel. Assume that $Y = f^*(X) + W$ with W a Cauchy random variable and $f^* \in \mathcal{H}_K$, the RKHS associated with K. With probability larger than $1 - \exp\big(-C_1 N^{p/(p+1)}\big)$, for a well chosen value of λ the estimator $\hat{f}$ associated to the absolute loss function defined in* (4.1) *satisifies:*

$$\|\hat{f}_N^{\lambda} - f^*\|_{L_2(\mu)}^2 \leq \frac{C_2}{N^{1/(1+p)}} \ ,$$

where C_1 and C_2 are functions of the kernel and $\|f^*\|_{\mathcal{H}_K}$. The value of $p \in (0,1)$ represents how fast the eigenvalues of the Kernel matrix decrease (see Section 4.4.2 for more precise arguments).

Theorem 4.1 deals with a Cauchy noise but many different distributions can be handled with our analysis (see Theorem 4.10). We obtain the same bounds as (Mendelson et al., 2010; Caponnetto and De Vito, 2007). This is a first important contribution of this work. Fast rates for Kernel methods are derived even when the noise is heavy-tailed. Note also that nothing is assumed on the design X.

Kernel methods belong to the more general class of regularized methods, widespread in statistics and machine learning. These procedures date back to Tikhonov (Golub et al., 1979), and have been widely used in non-parametric statistics (Marsh and Cormier, 2001; Huang et al., 2003) to smooth estimators. For example, the regularization $\phi(f) = \int (f'')^2$ for spline estimators promotes smoothness by imposing regularity on the estimate. In kernel methods, the norm of a function in the RKHS controls how fast the function varies with respect to the geometry defined by the kernel. Consequently, the norm of regularization $\|\cdot\|_{\mathcal{H}_K}$ is related with its degree of smoothness w.r.t. the metric defined by the kernel. Following the approach of (Chinot et al., 2019b), we present an analysis for RERM with loss functions that are simultaneously Lipchitz and convex. The penalization function is not assumed to be a norm. It is simply required to be an even convex function. We derive bounds on the L_2-error and the excess loss for these general procedures. As far as we know, the only article considering a generic analysis of the RERM (with the quadratic loss) with a convex penalization is (Lecué and Mendelson, 2017). However, their analysis does not hold for the square of a norm (see Assumption 5.1), which is a classical regularization methods in RKHS, see for instance (Steinwart and Christmann, 2008). By contrast, the new analysis presented in this paper covers many well-known methods such as kernel methods regularized by the square of a norm or the elastic net procedure (Zou and Hastie, 2005). The restriction here is that the loss function must be Lipschitz and convex. Both regression and classification problems can be addressed with our analysis.

Let $\mathcal{X}, \mathcal{Y}$ be two measurable spaces such that $\mathcal{Y} \subset \mathbb{R}$ and $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ be random variables with joint distribution P. Let μ be the marginal distribution of X. For E a linear subset of

$L_2(X)$, let $F \subset E$ be a class of measurable functions $f : \mathcal{X} \mapsto \bar{\mathcal{Y}}$ where $\bar{\mathcal{Y}} \subset \mathbb{R}$ is convex (we do not have necessarily $\mathcal{Y} = \bar{\mathcal{Y}}$ for classification problems). In the standard learning framework, one would like to identify the best approximation to Y using functions f in the class F. To do so, let ℓ be a loss function, $\ell : F \times \mathcal{X} \times \mathcal{Y} \mapsto \mathbb{R}$, $(f, x, y) \mapsto \ell_f(x, y) = \bar{\ell}(f(x), y)$ measuring the error made when predicting y by $f(x)$, for $\bar{\ell} : \bar{\mathcal{Y}} \times \mathcal{Y} \mapsto \mathbb{R}$. Let $f^* \in \operatorname{argmin}_{f \in F} R(f)$ where $R(f) := P\ell_f := \mathbb{E}_P[\ell_f(X, Y)]$. The *oracle* f^* provides the prediction of Y with minimal risk among functions in F. Obviously, the distribution P is unknown and minimizing the risk $R(f)$ over f in F is impossible in practice. Instead, one is given a dataset $\mathcal{D} = (X_i, Y_i)_{i=1}^N$ of random variables taking values in $\mathcal{X} \times \mathcal{Y}$. Using $\mathcal{D}$, the objective is to construct an estimator $\hat{f}_N$ such that the $L_2(\mu)$-**error rate**

$$\|\hat{f}_N - f^*\|^2_{L_2(\mu)} = \mathbb{E}\left[\left(\hat{f}_N(X) - f^*(X)\right)^2 | \mathcal{D}\right]$$

and the **excess risk**

$$P\mathcal{L}_{\hat{f}_N} := (P\ell_{\hat{f}_N} - P\ell_{f^*})|\mathcal{D} = \mathbb{E}_P\left[\bar{\ell}(\hat{f}_N(X), Y) - \bar{\ell}(f^*(X), Y)|\mathcal{D}\right]$$

are small. While $P\mathcal{L}_{\hat{f}_N}$ specifies the quality of prediction of the estimator $\hat{f}_N$, $\|\hat{f}_N - f^*\|_{L_2(\mu)}$ quantifies the $L_2(\mu)$ approximation of the *oracle* f^* by the estimator $\hat{f}_N$. These two quantities being random, the results are derived with exponentially large probability. All along the paper, the following geometric Assumption is also granted.

Assumption 4.1. *The class F is convex.*

Assumption 4.1 imposes a geometric structure on the class F. This assumption is essential to use our "projection trick" and derive our main results. For example Assumption 4.1 holds when F is a Hilbert space or the set of linear functionals in $\mathbb{R}^p$, $F = \{\langle t, \cdot \rangle : t \in \mathbb{R}^p\}$. As in (Chinot et al., 2019b), we consider Lipschitz and convex loss functions.

Assumption 4.2. *There exists $L > 0$ such that, for any $y \in \mathcal{Y}$, $\bar{\ell}(\cdot, y)$ is L-**Lipschitz** (see (4.2)) and **convex** i.e for all $\alpha \in [0, 1], (x, y) \in \mathcal{X} \times \mathcal{Y}$ and $f, g \in F$, $\bar{\ell}(\alpha f(x) + (1 - \alpha)g(x), y) \leq \alpha\bar{\ell}(f(x), y) + (1 - \alpha)\bar{\ell}(g(x), y)$*

Assumption 4.2 is satisfied in several examples, let us provide a short list of some of them.

- The **logistic loss** defined, for any $u \in \bar{\mathcal{Y}} = \mathbb{R}$ and $y \in \mathcal{Y} = \{-1, 1\}$, by $\ell(u, y) = \log(1 + \exp(-yu))$ satisfies Assumption 4.2 with $L = 1$.
- The **hinge loss** defined, for any $u \in \bar{\mathcal{Y}} = \mathbb{R}$ and $y \in \mathcal{Y} = \{-1, 1\}$, by $\ell(u, y) = \max(1 - uy, 0)$ satisfies Assumption 4.2 with $L = 1$.

In those examples, the sets $\mathcal{Y}$ and $\bar{\mathcal{Y}}$ are different. The fact that every function f in F maps to the convex set $\bar{\mathcal{Y}}$ is crucial for the computation of the estimator $\hat{f}_N$ in practice (Zhang, 2004; Bartlett et al., 2006)

- The **Huber loss** defined, for any $\delta > 0$, $u, y \in \mathcal{Y} = \bar{\mathcal{Y}} = \mathbb{R}$, by

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

  satisfies Assumption 4.2 with $L = \delta$.

- The **quantile loss** is defined, for any $\tau \in (0,1)$, $u, y \in \mathcal{Y} = \bar{\mathcal{Y}} = \mathbb{R}$, by $\ell(u,y) = \rho_\tau(u-y)$ where, for any $z \in \mathbb{R}$, $\rho_\tau(z) = z(\tau - I\{z \leq 0\})$. It satisfies Assumption 4.2 with $L = 1$. For $\tau = 1/2$, the quantile loss is the L_1 loss.

- The **Hinge loss for regression** is defined for any $u, y \in \mathcal{Y} = \bar{\mathcal{Y}} = \mathbb{R}$, by $\ell(u,y) = \max(y - u, 0)$. It satisfies Assumption 4.2 with $L = 1$. Note that the Hinge loss function is modified for regression problems.

Classical results on the RERM in learning theory consider the quadratic loss function (Mendelson, 2014; Lecué and Mendelson, 2017, 2018). In this case $\bar{\ell}(u,v) = (u-v)^2/2$ for any $(u,v) \in \bar{\mathcal{Y}} \times \mathcal{Y}$. The starting point of their analysis is the following multiplier/quadratic decomposition

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

for any f in F. While the quadratic process $f \mapsto (f(X)-f^*(X))^2$ does not depend on the target Y, the multiplier process $f \mapsto (f^*(X)-Y)(f(X)-f^*(X))$ depends on the "noise" $Y - f^*(X)$. It can only be controlled under some restriction on this "noise". For example, when $Y = g(X) + W$, where $g : \mathcal{X} \mapsto \mathbb{R}$ is a function in F and W is a random variable independent to X, we have $g = f^*$ and thus $Y - f^*(X) = W$. In this problem, bounding the multiplier process requires strong moment assumptions on the noise W (see Theorem 1.2 in (Mendelson, 2017)). If we replace the quadratic loss function by the absolute loss and if the noise is symmetric and independent to X we also have $f^* = g$. In this case, from the Lipschitz property,

$$\forall (x,y) \in \mathcal{X} \times \mathcal{Y} \text{ and } f, g \in F, \quad |\bar{\ell}(f(x),y) - \bar{\ell}(g(x),y)| \leq L|f(x)-g(x)| \quad \text{for } L > 0 \ , \qquad (4.2)$$

the multiplier process disappears. It becomes possible to handle heavy-tailed symmetric noise W. From (4.2), note also that the random variable Y does not need to be integrable. For instance, W can be a Cauchy distribution.

To get fast rates of convergence, our analysis is based on the following local Bernstein condition

$$\forall f \in F : \|f - f^*\|_{L_2(\mu)} = r \text{ and } \phi(f - f^*) \leq \rho, \quad AP\mathcal{L}_f \geq \|f - f^*\|^2_{L_2(\mu)}$$

where $r, \rho > 0$. In the sequel, we have respectively r and ρ of the order of the error rate and $\phi(f^*)$, where we recall that $\phi(\cdot)$ is the regularization function and f^* the oracle. This condition

states that the excess risk $f \mapsto P\mathcal{L}_f$ is $1/A$-strongly convex in a neighborhood of the oracle f^*. This new **local** Bernstein condition introduced in (Chinot et al., 2019b) is the cornerstone to obtain fast rates of convergence for settings where the noise may be heavy-tailed. Contrary to the analysis for the quadratic loss function, no Small Ball assumption is required (Mendelson, 2014; Lecué and Mendelson, 2017). In addition to handle heavy-tailed noise, the use of Lipschitz function significantly simplifies the proof since only one process has to be considered. The main argument of the proof is a new "projection trick" (see the sketch of proof in Section 4.2) making the proof simpler. For example, no peeling technic is required. To summarize, the **contributions** of our new analysis for the RERM are the following

- We consider very general convex regularization functions $\phi(\cdot)$.
- For Lipschitz and convex loss function, heavy-tailed noise can be handled.
- Our proof relies on a convex argument simple to understand.

The RERM are robust with respect to the noise of the problem as long as the loss function is Lipschitz. However a single outlier in the X_i may make the RERM really bad. In addition, the RERM performs well only when the empirical excess of risk $f \mapsto P_N\mathcal{L}_f$ uniformly concentrates around its expectation $f \mapsto P\mathcal{L}_f$. To do so, it is necessary to impose a strong concentration assumption on the class $\{\mathcal{L}_f(X,Y), f \in F\}$. From Assumption 4.2 it is implied by a concentration assumption on the class $\{(f-f^*)(X), f \in F\}$. Consequently, sub-Gaussian or boundedness assumptions are necessary on the class $\{(f-f^*)(X), f \in F\}$ to obtain an exponentially large confidence for RERM.

RERM serves as benchmark for more advanced estimators. In a second time, we study regularized minmax MOM-estimators introduced in (Lecué and Lerasle, 2019) for least-squares regression as an alternative to other MOM-based procedures (Lugosi and Mendelson, 2016; Lugosi et al., 2019a,b; Lecué and Lerasle, 2017). In the case of convex and Lipschitz loss functions, these estimators satisfy the following properties 1) as the RERM, they are efficient under weak assumptions on the noise 2) they achieve optimal rates of convergence under weak stochastic assumptions on the class $\{\mathcal{L}_f(X,Y), f \in F\}$ and 3) the rates are not downgraded by the presence of some outliers in the dataset. These results are not surprising since it has already been observed in (Lecué and Lerasle, 2019; Chinot et al., 2019b). Although attractive, mimmax MOM-estimators present some drawbacks. Their construction depends on the confidence level (through K). Under stronger moment assumptions, (Minsker, 2018) proposed a construction of MOM-based estimators independent to the confidence level. The implementation of MOM-based estimators is still an open question even if good empirical results have been obtained in (Lecué and Lerasle, 2019; Lecué et al., 2018; Chinot et al., 2019b).

The main theorems (for the RERM and the Minmax MOM estimators) are general and can be applied for different applications. In particular, we study 1) the Elastic net regularization for linear

estimators in $\mathbb{R}^p$ and 2) kernel methods in RKHS associated to a bounded kernel. In particular, we extend the results from (Mendelson et al., 2010; Smale and Zhou, 2007; Wu et al., 2006; Caponnetto and De Vito, 2007) for heavy-tailed noise.

To summarize, the **contributions of this paper** are the following:

- We obtain an analysis for the RERM for general convex regularization functions under weak assumptions on the noise. This analysis is based on a local Bernstein assumption and holds under a strong concentration assumption on the class $\{(f - f^*)(X), f \in F\}$.
- Under the same local Bernstein assumption, we study Minmax MOM estimators and show that 1) as the RERM, they are efficient under weak assumptions on the noise 2) they achieve optimal rates of convergence under weak stochastic assumptions on the class $\{(f-f^*)(X), f \in F\}$ and 3) the rates are not downgraded by the presence of some outliers in the dataset
- We apply this analysis to linear estimators regularized with elastic net.
- Under the same local Bernstein assumption, with a slightly different concentration argument, we study regularized learning problems in RKHS. The noise can be heavy-tailed and no sub-Gaussian on $\{(f - f^*)(X), f \in F\}$ is required to get fast rates of convergence.

The paper is organized as follow. In Section 4.2 and 4.3 we respectively present general results for RERM and minmax MOM estimators. Section 4.4 is devoted to the application of our main theorems for the problems of linear estimators regularized with elastic net and Support vector machines. Section 4.6.1- 4.7 gather the proofs of the main theorems.

Notations: In the remaining of the paper, the following notations will be used repeatedly. We will write L_2 instead of $L_2(\mu)$, let $r > 0$,

$$rB_{L_2} = \{f \in F : \|f(X)\|_{L_2(\mu)} \leqslant r\}, \quad rS_{L_2} = \{f \in F : \|f(X)\|_{L_2(\mu)} = r\} \ .$$

For any set H for which it makes sense, let $H+f^* = \{h+f^* \text{ s.t } h \in H\}$, $H-f^* = \{h-f^* \text{ s.t } h \in H\}$. The notations $a \vee b$ and $a \wedge b$, will denote respectively $\max(a, b)$ and $\min(a, b)$.

4.2 Regularized Empirical Risk Minimization (RERM)

All along this section, data $(X_i, Y_i)_{i=1}^N$ are **independent and identically distributed** with common distribution P. The unknown risks are estimated by their empirical counterparts, and the oracle is estimated by the *empirical risk minimizer* (ERM) (see (Koltchinskii, 2011b)), defined by

$$\hat{f}^{ERM} = \operatorname*{argmin}_{f \in F} P_N \ell_f := \frac{1}{N} \sum_{i=1}^{N} \bar{\ell}(f(X_i), Y_i) \ .$$

Clearly, if the class F is too small, there is no hope that $f^*(X)$ is close to Y. One has to consider large classes leading to large error rates. To bypass the fact that F may be very large, we can use

the classical approach of *regularization* where the penalization function emphasizes the belief we may have on the *oracle* f^*. It leads to the Regularized Empirical Risk Minimizer (RERM) defined as

$$\hat{f}_\lambda^{RERM} = \operatorname*{argmin}_{f\in F} P_N\ell_f + \lambda\|f\| \ , \tag{4.3}$$

where $\|\cdot\| : E \mapsto \mathbb{R}^+$ is a norm. However, the estimators $\hat{f}_\lambda^{RERM}$ defined in (4.3) are rather restrictive since it does not cover penalizations which are not a norm such as $\|f\|^2_{\mathcal{H}_K}$ (i.e the square of the norm in a reproducing Kernel Hilbert space) or the Elastic net procedure (see (Zou and Hastie, 2005)). To bypass this limitation, the estimator defined in Equation (4.3) will be replaced by

$$\hat{f}_\lambda^{\phi} = \operatorname*{argmin}_{f\in F} P_N\ell_f + \lambda\phi(f) := \operatorname*{argmin}_{f\in F} P_N\mathcal{L}_f^\lambda \tag{4.4}$$

where $\phi : E \mapsto \mathbb{R}^+$ is a function satisfying the following Assumption.

Assumption 4.3. *Let $\phi : E \mapsto \mathbb{R}^+$ be a real function such that*

- *ϕ is even, convex and $\phi(0) = 0$*
- *There exists a constant $\eta > 0$ such that for all $f, g \in F$*

$$\phi(f+g) \le \eta\big(\phi(f) + \phi(g)\big) \tag{4.5}$$

Assumption 4.3 holds for any norm but also for the square of a norm (with $\eta = 2$), the elastic net penalization (with $\eta = 2$) defined for any t in $\mathbb{R}^p$ as $\phi(t) = (1-\alpha)\|t\|_1 + \alpha\|t\|_2^2$, where $\alpha \in [0,1]$, $\|t\|_1 = \sum_{i=1}^p |t_i|$ and $\|t\|_2^2 = \sum_{i=1}^p t_i^2$. To control the L_2-error rates for the RERM, it is necessary to impose a **concentration assumption** on the class $\{\mathcal{L}_f(X,Y), f \in F\}$. From Assumption 4.2 it is implied by a concentration assumption on the class $\{(f-f^*)(X), f \in F\}$ (this assumption will be relaxed using MOM-type estimators in Section 4.3).

Definition 4.1. *A class F is called B sub-Gaussian (with respect to X) for some constant $B \ge 0$ when for all f in F and for all $\lambda > 1$*

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

Assumption 4.4. *The class $F - f^*$ is B sub-Gaussian.*

For example, when F is the class of linear functionals in $\mathbb{R}^p$, $F = \{\langle\cdot, t\rangle,\ t \in T\}$ for $T \subset \mathbb{R}^p$, $F - f^*$ is 1 sub-Gaussian if $X \sim \mathcal{N}(0,\Sigma)$ or if $X = (x_j)_{j=1}^p$ has independent coordinates that are 1 sub-Gaussian. In the sub-Gaussian framework, a natural way to measure the *statistical complexity* of the class of functions F is via the Gaussian mean-width that we introduce now.

Definition 4.2. *Let $H \subset L_2$ and $(G_h)_{h\in H}$ be the canonical centered Gaussian process indexed by H, with covariance structure*

$$\forall h_1, h_2 \in H, \qquad \big(\mathbb{E}(G_{h_1} - G_{h_2})^2\big)^{1/2} = \big(\mathbb{E}(h_1(X) - h_2(X))^2\big)^{1/2} \ .$$

*The **Gaussian mean-width** of H is $w(H) = \mathbb{E}\sup_{h\in H} G_h$.*

For example, when $F = \{\langle \cdot, t\rangle,\ t \in \mathbb{R}^p\}$, and $X \sim \mathcal{N}(0, \Sigma)$, $w(T) = \mathbb{E}\sup_{t\in T}\langle G, t\rangle$, where T is a subset of $\mathbb{R}^p$ and $G \sim \mathcal{N}(0, \Sigma)$. The Gaussian mean-width is closely related with metric complexities such as the entropy through the Sudakov's inequality, see Chapter 1 in (Chafaï et al., 2012) for precise inequalities.
Following ideas developed in (Lecué and Lerasle, 2017; Lecué and Mendelson, 2017, 2018; Mendelson, 2014), the complexity parameter driving the statistical behavior of the estimator $\hat{f}_\lambda^\phi$ is defined as a fixed point depending on the Gaussian mean-width:

Definition 4.3. *The complexity is measured via a non-decreasing function $r(\cdot)$ such that for every $A > 0$,*

$$r(A) = \inf\left\{ r > 0 :\ 32LBw\big(F \cap B^\phi_{\eta(4+2A^{-1})\phi(f^*)}(f^*) \cap (f^* + rB_{L_2})\big) \le (2A)^{-1}\sqrt{N}r^2 \right\}$$

where $B^\phi_\delta(g) = \{f \in F :\ \phi(f-g) \le \delta\}$, L is the Lipschitz constant of Assumption 4.2, B is the sub-Gaussian constant defined in Assumption 4.4 and η is defined in Assumption 4.3.

Note that when ϕ is a norm, $B^\phi_\delta(g)$ simply corresponds to the ball of regularization centered in g with radius δ. We are now in position to introduce the **local Bernstein condition** allowing to derive fast rates of convergence for heavy-tailed problem.

Assumption 4.5. *There exists a constant $A^* > 0$ such that for all $f \in F$ if $\|f - f^*\|_{L_2} = r(A^*)$ and $\phi(f - f^*) \le \eta(4 + 2(A^*)^{-1})\phi(f^*)$ then $\|f - f^*\|^2_{L_2} \leqslant A^* P\mathcal{L}_f$.*

In the sequel of this section we will write r^* instead of $r(A^*)$. Condition 4.5 states that $f \mapsto P\mathcal{L}_f$ is $1/A^*$-strongly convex in a subset of the L_2-sphere centered in f^* with radius r^*. As explained in (Chinot et al., 2019b), this local Bernstein condition holds in examples where F is not bounded in L_2-norm, and therefore, where the global Bernstein condition of (Alquier et al., 2019)($\|f - f^*\|^2_{L_2} \leqslant A^* P\mathcal{L}_f$ for all $f \in F$) does not hold. Assumption 4.5 replaces the small-ball Assumption (see (Mendelson, 2014) for instance) for learning problems with Lipschitz and convex loss functions. In (Chinot et al., 2019b), the authors consider non-regularized problems where the local Bernstein condition is required over the whole L_2-sphere of radius r^*. For regularized-procedure, this condition is required only for functions f in this L_2-sphere of radius r^* such that $\phi(f - f^*) \le \eta(4 + 2(A^*)^{-1})\phi(f^*)$. For instance, in the case of RKHS associated to a bounded kernel K, the condition $\phi(f - f^*) \le \rho$, for $\rho > 0$ implies that the function $f - f^*$ are bounded by $\sqrt{\rho\|K\|_\infty}$ (see Section 4.4.2). This localization with respect to the regularization norm is essential to verify the local Bernstein Assumption in practice and obtain fast rates of convergence (see Section 4.4.2). We are now in position to present the main theorems of this section.

Theorem 4.2. *Grant Assumptions 4.2, 4.1, 4.3, 4.4 and 4.5. With probability larger than*

$$1 - 2\exp\left(- \frac{N(r^*)^2}{4(32A^*LB)^2} \right) \tag{4.6}$$

for all regularization parameters $\lambda \geq \lambda_0 = (r^*)^2/\phi(f^*)$ *the estimator* $\hat{f}_\lambda^\phi$ *defined in Equation* (4.4) *satisfies*

$$\|\hat{f}_\lambda^\phi - f^*\|_{L_2} \leq (4 + 6A^*)\lambda \frac{\phi(f^*)}{r^*}$$
$$\text{and} \quad \phi(\hat{f}_\lambda^\phi - f^*) \leq (4 + 2/A^*)\eta\phi(f^*).$$

Remark 4.1. *Theorem 4.2 holds for an exponentially large probability* (4.6) *simultaneously for all* $\lambda \geq \lambda_0$. *As a consequence it can be used with a random choice of regularization parameter* $\hat{\lambda}$ *as long as* $\{\hat{\lambda} \geq \lambda_0\}$ *hold with large probability. For example, we could use a cross validation scheme to generate* $\hat{\lambda}$.

Note that for $\lambda = \lambda_0$, we obtain $\|\hat{f}_\lambda^\phi - f^*\|_{L_2} \leq (4 + 6A^*)r^*$, which is the minimax rate into the class $\{f \in F : \phi(f) \leq \phi(f^*)\}$ (see (Lecué and Mendelson, 2017)). Since we do not have access to $\phi(f^*)$, taking λ_0 is impossible. To bypass this issue we use a Lepski's adaptation method (see (Lepskii, 1992, 1993; Birgé, 2001)). To do so, the following assumption is required.

Assumption 4.6. *There exists* $M > 0$ *such that* $\phi(f^*) \leq M$.

Assumption 4.6 is natural since regularization procedures are used when one believes that $\phi(f^*)$ is small. Since Theorem 4.2 holds with the same probability for all $\lambda \geq \lambda_0$, one can choose M very large in the Lepski's method without deteriorating the probability of the event.
For $j = 1, \cdots, J = M + \lceil \log_2(M) \rceil$, let us define $\phi_j = 2^j/2^M$, $\phi_0 = 0$ and $\lambda_j = r_j^2/\phi_j$ where

$$r_j = \inf\left\{r > 0 : \ 32LBw\big(F \cap B^\phi_{\eta(4+2(A^*)^{-1})\phi_j}(f^*) \cap (f^* + rB_{L_2})\big) \leq (2A^*)^{-1}\sqrt{N}r^2\right\}$$

Moreover for all $\lambda > 0$ let us define

$$T_\lambda(f) = P_N(\ell_f - \ell_{\hat{f}_\lambda^\phi}) + \lambda\big(\phi(f) - \phi(\hat{f}_\lambda^\phi)\big), \quad \hat{R}_j = \{f \in F : \ T_{\lambda_j}(f) \leq \big((A^*)^{-1} + 2\big)\lambda_j\phi_j\}$$
$$k^* = \inf\{k \in \{1, \cdots, J\} : \ \cap_{j \geq k}^J \hat{R}_j \neq \emptyset\} \quad \text{and set} \quad \tilde{f} \in \cap_{j \geq k^*}^J \hat{R}_j \ .$$

Using the Lepski's method we are in position to state to following theorem.

Theorem 4.3. *Assumptions 4.2, 4.1, 4.3, 4.4, 4.5 and 4.6, with probability larger than*

$$1 - 2\exp\big(-\frac{N(r^*)^2}{4(64A^*LB(8 + 12A^*))^2}\big)$$

$$\|\tilde{f} - f^*\|_{L_2} \leq (8 + 12A^*)r^*, \quad \phi(\tilde{f} - f^*) \leq (4 + 2/A^*)\eta\phi(f^*)$$
$$\text{and} \quad P\mathcal{L}_{\tilde{f}} \leq (4 + 3/A^*)(r^*)^2 \ .$$

Note that such a procedure required the knowledge of A^* and M. Complete proofs of Theorem 4.3 and Theorem 4.2 are presentend in Section 4.6.1. Here we present a simple sketch of the proof of Theorem 4.2. Our proof relies on a homogeneity argument allowing to study the empirical

excess risk only in neighborhood around the oracle f^*.

Sketch of the proof : The main arguments are presented up to some constants depending on A^*, L and η. The proof is split into two parts. First, we identify a random event onto which the statistical behavior of $\hat{f}_\lambda^\phi$ can be studied using deterministic arguments. Next, we prove that this event holds with large probability. Here we will only focus on the deterministic argument (see Section 4.6.1 for the stochastic control).
Let $\mathcal{B}_\lambda = \{f \in F : \ \|f - f^*\|_{L_2} \leq \lambda\phi(f^*)/r^* \text{ and } \phi(f - f^*) \leq \phi(f^*)\}$ and the stochastic event is defined as

$$\Omega := \Big\{\text{for all } f \in F \cap (f^* + r^* B_{L_2}) \cap B^\phi_{\phi(f^*)}(f^*), \quad \big|(P - P_N)\mathcal{L}_f\big| \leq (r^*)^2\Big\}$$

By definition, the estimator $\hat{f}_\lambda^\phi$ satisfies $P_N\mathcal{L}^\lambda_{\hat{f}_\lambda^\phi} \leq 0$. Therefore, to prove Theorem 4.2 it is sufficient to show that on Ω, $P_N\mathcal{L}^\lambda_f > 0$ for all functions f in $F\backslash\mathcal{B}_\lambda$. The proof follows from an homogeneity argument saying that for all functions $f \in F\backslash\mathcal{B}_\lambda$, there exist f_0 in the frontier of $\mathcal{B}_\lambda$ and $\alpha \geq 1$ such that $P_N\mathcal{L}^\lambda_f \geq \alpha P_N\mathcal{L}^\lambda_{f_0}$. On the frontier of $\mathcal{B}_\lambda$, either we have 1) $\phi(f_0 - f^*) = \phi(f^*)$ and $\|f_0 - f^*\|_{L_2} \leq \lambda\phi(f^*)/r^*$ or 2) $\|f_0 - f^*\|_{L_2} = \lambda\phi(f^*)/r^*$ and $\phi(f_0 - f^*) \leq \phi(f^*)$.
The homogeneity argument linking the empirical excess risk of f to the one of f_0 is the following. For all $i \in \{1, \cdots, N\}$, let $\psi_i : \mathbb{R} \to \mathbb{R}$ be defined for all $u \in \mathbb{R}$ by

$$\psi_i(u) = \bar{\ell}(u + f^*(X_i), Y_i) - \bar{\ell}(f^*(X_i), Y_i). \tag{4.7}$$

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

$$\begin{aligned} P_N\mathcal{L}_f &= \frac{1}{N}\sum_{i=1}^N \psi_i\big(f(X_i) - f^*(X_i)\big) = \frac{1}{N}\sum_{i=1}^N \psi_i(\alpha(f_0(X_i) - f^*(X_i))) \\ &\geq \frac{\alpha}{N}\sum_{i=1}^N \psi_i((f_0(X_i) - f^*(X_i))) = \alpha P_N\mathcal{L}_{f_0}. \end{aligned} \tag{4.8}$$

For the regularization part, since $\alpha \geq 1$, the same homogeneity arguments holds.

$$\phi(f) - \phi(f^*) = \phi\big(f^* + \alpha(f_0 - f^*)\big) - \phi(f^*) \geq \alpha\big(\phi(f_0) - \phi(f^*)\big)$$

It remains to control $P_N\mathcal{L}^\lambda_{f_0}$ in the two cases 1) and 2). Up to technicalities, in case 1), we use Assumption 4.3 to showing that $\phi(f_0) - \phi(f^*) \geq \phi(f^*)$ (up to constants). Using the event Ω, we show that $P_N\mathcal{L}_{f_0} \geq -\theta\lambda\phi(f^*)$ for $\theta > 0$ small enough. In case 2), we use that $\phi(f_0) - \phi(f^*) \geq -\phi(f^*)$ and the local Bernstein Assumption 4.5 to prove that $P_N\mathcal{L}_{f_0} \geq \gamma\lambda\phi(f^*)$ for $\gamma > 0$ large enough which concludes the deterministic argument. ■

4.3 Robustness to outliers and heavy-tailed data via Minmax MOM estimators

In Section 4.2, we assumed that the class $\{(f-f^*)(X), f \in F\}$. is sub-Gaussian and that the data $(X_i, Y_i)_{i=1}^N$ are i.i.d with the same distribution P. In this section, we relax these assumptions using **minmax-MOM type estimators**. For any $i \in \{1, \cdots, N\}$, let P_i be the distribution of (X_i, Y_i). Let $\mathcal{I} \cup \mathcal{O}$ denote an unknown partition of $\{1, \cdots N\}$. The cardinality of $\mathcal{O}$ is denoted $|\mathcal{O}|$. Data $(X_i, Y_i)_{i\in\mathcal{O}}$ are considered as outliers. **No assumption** on the distribution P_i for $i \in \mathcal{O}$ is made and can be dependent or even adversarial. The informative random variables $(X_i, Y_i)_{i\in\mathcal{I}}$ satisfy:

Assumption 4.7. *The data $(X_i, Y_i)_{i\in\mathcal{I}}$ are independent and for all $i \in \mathcal{I}$: $P_i(f-f^*)^2(X_i) = P(f-f^*)^2(X)$ and $P_i\mathcal{L}_f = P\mathcal{L}_f$ where we recall that P is the distribution of (X, Y) .*

Assumption 4.7 holds in the i.i.d framework but it covers other situations where informative data $(X_i, Y_i)_{i\in\mathcal{I}}$ may not have the same distribution. It is only required to induce the same L_2-structure on the class F and the same excess risk.

Let $(B_s)_{s=1,\dots,S}$ denote a partition of $\{1, \dots, N\}$ into blocks B_s of equal size N/S (if N is not a multiple of S, just remove some data). Following (Lecué and Lerasle, 2019) the minmax MOM-estimators are defined as

$$\hat{f}_S^{\lambda} = \operatorname*{argmin}_{f\in F} \sup_{g\in F} MOM_S(\ell_f - \ell_g) + \lambda\big(\phi(f) - \phi(g)\big), \tag{4.9}$$

where $MOM_S(\ell_f - \ell_g) = \mathrm{Med}\big(P_{B_1}(\ell_f - \ell_g), \cdots, P_{B_S}(\ell_f - \ell_g)\big)$ with $P_{B_s}(\ell_f - \ell_g) = (1/|B_s|)\sum_{i\in B_s} \ell_f(X_i, Y_i) - \ell_g(X_i, Y_i)$.
Since we no longer consider the sub-Gaussian framework, we have to adapt the complexity parameter to this new setup. The complexity is measured via a function $\tilde{r}(\cdot)$ defined as

$$\tilde{r}(A) = \inf\Bigg\{ r > 0 : \forall J \subset \mathcal{I} : |J| \geqslant N/2, \\ \mathbb{E} \sup_{f\in F\cap(f^*+rB_{L_2})\cap B^{\phi}_{\eta(4+2A^{-1})\phi(f^*)}(f^*)} \left|\sum_{i\in J} \sigma_i(f-f^*)(X_i)\right| \leq (384AL)^{-1} r^2 |J| \Bigg\} \tag{4.10}$$

where $(\sigma_i)_{i=1}^N$ are i.i.d Rademacher random variables independent from $(X_i, Y_i)_{i\in\mathcal{I}}$.
This complexity function is very close to the one in the sub-Gaussian case from Section 4.2 expect that the Rademacher-complexity replaces the Gaussian mean-width. When the class $F - f^*$ is B-sub-Gaussian, a standard chaining argument (Talagrand, 2006) shows that $\tilde{r}(\cdot)$ and $r(\cdot)$ are equivalent. However, when only L_p conditions are granted on the class $F - f^*$, $\tilde{r}(\cdot)$ may be larger than $r(\cdot)$, see (Chinot et al., 2019b), for instance. It is also necessary to adapt the local Bernstein condition from Assumption 4.5 to the MOM-framework

Assumption 4.8. *There exists a constant $\tilde{A} > 0$ such that, for all f in F satisfying $\|f - f^*\|_{L_2} = \sqrt{C_{S,r}(\tilde{A})}$ and $\phi(f - f^*) \leq \eta(4 + 2/\tilde{A})\phi(f^*)$, then $\|f - f^*\|_{L_2}^2 \leqslant \tilde{A}P\mathcal{L}_f$ where*

$$C_{S,r}(A) = \max\left(\tilde{r}^2(A), 368A^2L^2\frac{S}{N}\right) \quad . \tag{4.11}$$

As Assumption 4.5, Assumption 4.8 is only granted on a subset of the L_2-sphere centered in the oracle f^* where the radius is proportional to the rate of convergence of the estimators. We are now in position to state our main results for the minmax MOM estimators.

Theorem 4.4. *Grant Assumptions 4.2, 4.1, 4.3, 4.7 and 4.8. Let $S \geq 7|\mathcal{O}|/3$, Then, with probability larger than $1 - 2\exp(-S/504)$, for any regularization parameter $\lambda > C_{S,r}(\tilde{A})/\phi(f^*)$, the estimator $\hat{f}_S^\lambda$ defined in Equation (4.9) satisfies*

$$\phi(\hat{f}_S^\lambda - f^*) \leq \eta(4 + 2/\tilde{A})\phi(f^*), \quad \|\hat{f}_S^\lambda - f^*\|_{L_2} \leq (4 + 6\tilde{A})\lambda\frac{\phi(f^*)}{\sqrt{C_{S,r}(\tilde{A})}}$$

It is also possible to use the Lepski's method to get an adaptive estimator as the one in Theorem 4.3. For the sake of brevity, we do not present this result here. There is a tradeoff between confidence and accuracy and an optimal choice of S would be $S \asymp \tilde{r}(\tilde{A})N$. In that case, $C_{S,r}(\tilde{A}) \asymp \tilde{r}(\tilde{A})$. For this value of S, the optimal λ is $\tilde{r}^2(\tilde{A})/\phi(f^*)$ and we would obtain $\|\hat{f}_S^\lambda - f^*\|_{L_2}^2 \lesssim C(\tilde{A})\tilde{r}(\tilde{A})$. With $S \asymp \tilde{r}(\tilde{A})N$ and $\lambda =\asymp \tilde{r}^2(\tilde{A})/\phi(f^*)$, we recover the same result as the one in the sub-Gaussian setting as long as Rademacher complexity and Gaussian-mean width are equivalent. We will see in Section 4.4.2 that it is the case for the precise example of RKHS associated to bounded kernel. Moreover, by construction, the estimator $\hat{f}_S^\lambda$ is robust to $3S/7$ outliers in the dataset.
Therefore, using minmax-MOM estimators, we have relaxed two strong Assumptions 1) the i.i.d setting and 2) the sub-Gaussian Assumption on the class $F - f^*$. Properly calibrated minmax-MOM estimators are not affected if the number of outliers is less than *number of observations $\times$ square of the optimal rate in the i.i.d setup* (when $S \asymp \tilde{r}(\tilde{A})N$ and $r(A) \asymp \tilde{r}(A)$).

4.4 Applications

Our results are very general and may be applied to various examples. To do so, it is necessary to:

- Verify Assumptions 4.2, 4.1 and 4.3.
- If the RERM is studied, check Assumption 4.4 and compute the Gaussian-mean-width $w\big(F \cap B^{\phi}_{\eta(4+2(A^*)^{-1})\phi_j}(f^*) \cap (f^* r B_{L_2})\big)$ to deduce $r(A)$ for every $A > 0$.
- If the minmax MOM-estimators is considerer, compute the Rademacher complexity to deduce $\tilde{r}(A)$ for every $A > 0$.
- Find A satisfying the local Bernstein condition (the L_2-radius depends on the estimator we consider).

As an illustration, we study in the sequel RERM and minmax MOM-estimators for linear estimators in $\mathbb{R}^p$ regularized by the elastic net and for regularized kernel methods. It turns out that the sub-Gaussian assumption over the class $F - f^*$ is not required by using the reproducing property of RKHS. Instead we develop another general analysis to study RERM in RKHS associated with bounded kernel (see Section 4.4.2).

4.4.1 Application to Elastic net with Huber loss function

In (Zou and Hastie, 2005), the authors noticed that the performance of the LASSO is not as good as the one of Ridge regression when the variables are highly correlated. Theoretically, it is now known that the covariance matrix of the design X must satisfy the Restricted Eigenvalue condition to obtain fast rates of convergence for the LASSO (Bellec et al., 2018; Bickel et al., 2009). To bypass this limitation, the authors introduced in (Zou and Hastie, 2005) the Elastic net regularization.

Regularized Empirical Risk Minimizers Let F be the class of linear functionals in $\mathbb{R}^p$, $F = \{\langle \cdot, t\rangle,\ t \in \mathbb{R}^p\}$ which satisfies Assumption 4.1. Let $(X_i, Y_i)_{i=1}^N$ be random variables valued in $\mathbb{R}^p \times \mathcal{Y}$. As the *oracle* is denoted f^*, we introduce t^* such that $f^*(\cdot) = \langle t^*, \cdot\rangle$. Let $\alpha \in [0,1]$, for any t in $\mathbb{R}^p$, the elastic net penalization is defined as

$$\phi(t) = (1-\alpha)\|t\|_1 + \alpha\|t\|_2^2 \ , \tag{4.12}$$

where $\|t\|_1 = \sum_{i=1}^p |t_i|$ and $\|t\|_2^2 = \sum_{i=1}^p t_i^2$. For $\alpha = 1$ and $\alpha = 0$ we recover respectively the ridge and the Lasso penalizations (these cases will not be studied in the sequel). Clearly ϕ defined in Equation (4.12) satisfies Assumption 4.3 with $\eta = 2$. Let $\bar{\ell}^\delta$ be the huber loss function with parameter $\delta > 0$ (which is δ-Lipschiz), the estimator RERM is defined as

$$\hat{t}_\lambda^{\delta,\alpha} \in \operatorname*{argmin}_{t\in\mathbb{R}^p} \frac{1}{N}\sum_{i=1}^N \bar{\ell}^\delta\big(\langle X_i, t\rangle, Y_i\big) + \lambda\big((1-\alpha)\|t\|_1 + \alpha\|t\|_2^2\big) \ . \tag{4.13}$$

Theorems 4.2 and 4.3 require the computation of the Gaussian mean-width $w\big(F \cap B_\rho^\phi(t^*) \cap (f^* + rB_{L_2})\big)$ for $r, \rho > 0$. To do so, let us assume that the design X is isotropic i.e for all $t \in \mathbb{R}^p$, $\mathbb{E}\langle X, t\rangle_{\mathbb{R}^p}^2 = \|t\|_2^2$. It means that the $L_2(\mu)$ norm coincides with the natural Euclidean structure on the space ℓ_2^p. Thus, for all $\rho, r > 0$, under the isotropic assumption, we have

$$w\big(F \cap B_\rho^\phi(t^*) \cap (f^* + rB_{L_2})\big) = w(B_\rho^\phi(0) \cap rB_2^p) = \mathbb{E}\sup_{t\in\mathbb{R}:\ (1-\alpha)\|t\|_1+\alpha\|t\|_2^2\le\rho,\ \|t\|_2\le r} \langle \mathbf{G}, t\rangle_{\mathbb{R}^p} \ , \tag{4.14}$$

where $\mathbf{G}$ is a standard Gaussian random vector in $\mathbb{R}^p$ and B_l^p denotes the unit ball in $(\mathbb{R}^p, \|\cdot\|_l)$, for $l \ge 0$. Let $\alpha \in (0,1)$. We have,

$$w(B_\rho^\phi(0) \cap rB_{L_2}) \le \min\left(w\big(\frac{\rho}{1-\alpha}B_1^p \cap rB_2^p\big), w\big(\min(r, \sqrt{\frac{\rho}{\alpha}})B_2^p\big)\right) \ . \tag{4.15}$$

Let us introduce

$$r_1^* = \inf\left\{r>0:\ 64\delta BA^* w\left(\frac{(8+4/A^*)\phi(f^*)}{1-\alpha}B_1^p \cap rB_2^p\right) \leq \sqrt{N}r^2\right\}\ .$$

$$r_2^* = \inf\left\{r>0:\ 64\delta BA^* w\left(\min\left(r, \sqrt{\frac{(8+4/A^*)\phi(f^*)}{\alpha}}\right)B_2^p\right)\right) \leq \sqrt{N}r^2\right\}\ .$$

From Equation (4.15) and the definition of r^* it is clear that $r^* \leq \min(r_1^*, r_2^*)$. Using the computations of $w\big(\rho B_1^p \cap rB_2^p\big)$ for all $r, \rho > 0$ presented in (Lecué and Mendelson, 2017), it follows that

$$(r_1^*)^2 = \begin{cases} \frac{(8+4/A^*)\phi(f^*)}{1-\alpha}\sqrt{\frac{64\delta BA^*}{\boldsymbol{N}}\log\left(\frac{e\boldsymbol{p}(1-\alpha)}{\sqrt{\boldsymbol{N}}(8+4/A^*)\phi(f^*)}\right)} & \text{if } \frac{(8+4/A^*)^2\phi^2(f^*)\boldsymbol{N}}{(1-\alpha)^2 64\delta BA^*} \leq \boldsymbol{p}^2 \\ \frac{64\delta BA^*\boldsymbol{p}}{\boldsymbol{N}} & \text{if } \frac{(8+4/A^*)^2\phi^2(f^*)\boldsymbol{N}}{(1-\alpha)^2 64\delta BA^*} \geq \boldsymbol{p}^2 \end{cases}$$

$$(r_2^*)^2 = \begin{cases} \frac{64\delta BA^*\boldsymbol{p}}{\boldsymbol{N}} & \text{if } \boldsymbol{N} \geq \frac{64\delta BA^*\alpha\boldsymbol{p}}{(8+4/A^*)\phi(f^*)} \\ \sqrt{\frac{64\delta B(8+4/A^*)\phi(f^*)\boldsymbol{p}}{\alpha\boldsymbol{N}}} & \text{if } \boldsymbol{N} \leq \frac{64\delta BA^*\alpha\boldsymbol{p}}{(8+4/A^*)\phi(f^*)} \end{cases}$$

For the sake of presentation, the dependence with respect to the dimension and the sample size is presented in bold. Since $r^* \leq \min(r_1^*, r_2^*)$, it is clear that r^* captures the best situation between the LASSO (complexity parameter r_1^*) and the Ridge regression (complexity parameter r_2^*).

To apply Theorems 4.2 and 4.3, it remains to verify the local Bernstein condition. Results on the local Bernstein Assumption (see Assumptions 4.5 and 4.8) can be found in (Chinot et al., 2019b) for the quantile and Huber losses for regression problems and for the logistic and the Hinge loss for classification. For the sake of brevity, we only present the results for the Huber loss function with parameter $\delta > 0$ (absolute loss function will be studied in Section 4.4.2). Note that δ must be of the order of a constant. Let us introduce the following assumption.

Assumption 4.9. *Let $r, \rho, \varepsilon > 0$.*

- *a) There exists $C' > 0$ such that, for all $f \in F$ such that $\|f - f^*\|_{L_2} = r$ and $\phi(f - f^*) \leq \rho$, $\|f - f^*\|_{L_{2+\varepsilon}} \leq C'\|f - f^*\|_{L_2}$.*
- *b) Let C' be the constant defined above. There exists $\gamma > 0$ such that, for all $x \in \mathcal{X}$ and for all z in $\mathbb{R}$ such that $|z - f^*(x)| \leq (\sqrt{2}C')^{(2+\varepsilon)/\varepsilon}r$, we have $F_{Y|X=x}(z+\delta) - F_{Y|X=x}(z-\delta) \geq \gamma$, where $F_{Y|X=x}$ is the conditional cumulative function of Y given $X = x$.*

When the class $F - f^*$ is 1-sub-Gaussian, it is clear that the point a) of Assumption 4.9 holds with an absolute constant C' for $\varepsilon = 2$ (see theorem 1.1.5 in (Chafaï et al., 2012)). For the point b), if $Y = \langle t, X\rangle + W$, where W is a symmetric random variable independent from X and $t \in \mathbb{R}^p$, we have $t^* = t$. In this case, the point b) holds if $F_W(\delta - 2(C')^2 r) - F_W(2(C')^2 r - \delta) \geq \gamma$, where F_W denotes the cdf of W. It simply means that the noise puts enough mass around 0. In particular, point b) holds when W is Cauchy. In this case, Y is not integrable and yet we are able to verify the Bernstein condition and derive fast rates of convergence.

Theorem 4.5 ((Chinot et al., 2019b)). *Grant Assumptions 4.9 (with parameter r, ρ and γ). Then, for all $f \in F$ satisfying $\|f - f^*\|_{L_2} = r$ and $\phi(f - f^*) \leq \rho$, $\|f - f^*\|_{L_2}^2 \leqslant (4/\gamma) P\mathcal{L}_f$.*

Note that in (Chinot et al., 2019b), the proof holds for any f in F such that $\|f - f^*\|_{L_2} = r$. The proof of Theorem 4.5 is exactly the same as the one in (Chinot et al., 2019b) with simple modifications taking into account the new localization with respect to the regularization.
We are now in position to state the main theorem for the elastic net procedure.

Theorem 4.6. *Let $r^* = \min(r_1^*, r_2^*)$. Let $(X_i, Y_i)_{i=1}^N$ be i.i.d random variables distributed as (X, Y) where $Y = \langle X, t^* \rangle + W$, where $t^* \in \mathbb{R}^p$, $X = (x_1, \cdots, x_p)$ is a sub-Gaussian random vector. Let us assume that the noise W is a symmetric random variable independent from X such that there exists $\gamma > 0$ for which $F_W(\delta - 2(C')^2 r^*) - F_W(2(C')^2 r^* - \delta) \geq \gamma$. Let $\lambda = (r^*)^2/\phi(f^*)$. With probability larger than $1 - 2\exp\big(-\frac{\gamma^2}{4(128)^2\delta^2} N (r^*)^2\big)$, the estimator $\hat{t}_\lambda^{\delta,\alpha}$ associated with the Huber loss function defined in Equation (4.13) satisfies*

$$\|\hat{t}_\lambda^{\delta,\alpha} - t^*\|_2 \leq (4 + 24/\gamma) r^* \quad \phi(\hat{t}_\lambda^{\delta,\alpha} - f^*) \leq (8 + \gamma)\phi(t^*) \quad \text{and} \quad P\mathcal{L}_{\hat{t}_\lambda^{\delta,\alpha}} \leq (4 + 3\gamma/4)(r^*)^2 \ .$$

In Theorem 4.6 we set $\lambda = (r^*)^2/\phi(f^*)$ which is evidently unknown. However it is possible to use Theorem 4.3 to get an adaptive estimator for the Elastic net achieving the same rates. When $1 - \alpha$ is close to 1 that is when the penalization ℓ_1 is dominant we have $r^* = r_1^*$ and we recover the result for the Lasso (see (Lecué and Mendelson, 2017)). When α is close to 1 the elastic net is almost equivalent to ridge regression and $r^* = r_2^*$. We recover the results for the ridge regression.
In Theorem 4.6 it is not clear if there exists γ such that $F_W(\delta - 2(C')^2 r^*) - F_W(2(C')^2 r^* - \delta) \geq \gamma$. It turns out that this condition is very weak. It simply means that the noise W puts enough mass around 0. For instance let W be a standard Cauchy distribution. The condition $F_W(\delta - 2(C')^2 r^*) - F_W(2(C')^2 r^* - \delta) \geq \gamma$ can be rewritten as $\delta - 2(C')^2 r^* \geq \tan(\gamma\pi/2)$. If $r^* \leq 1$ we can take $\gamma = 1$ and $\delta = 4(C')^2 + 1$. The condition $r^* \leq 1$ means that enough data are given to the statistican which corresponds to interesting learning problems. Consequently, even for non-integrable noise such as a Cauchy distribution we are able to derive fast rates of convergence.

Minmax MOM-estimators Now, let us turn to the robust minmax MOM-estimator associated with the Huber loss function for the elastic net procedure defined as

$$\hat{t}_{\lambda,S}^{\delta,\alpha} \in \operatorname*{argmin}_{t \in \mathbb{R}^p} \sup_{\tilde{t} \in \mathbb{R}^p} MOM_S\big(\ell_t^\delta - \ell_{\tilde{t}}^\delta\big) + \lambda\big((1-\alpha)(\|t\|_1 - \|\tilde{t}\|_1) + \alpha(\|t\|_2^2 - \|\tilde{t}\|_2^2)\big) \tag{4.16}$$

where ℓ^δ denotes the Huber loss function with parameter δ. To study these estimators, is necessary to compute the rademacher complexity given in the definition of $\tilde{r}(\cdot)$. From Theorem 1.6 in (Mendelson, 2017), it is possible to link Rademacher complexity and Gaussian mean-width for the Elastic-net regularization as long as X is isotropic (i.e for all t in $\mathbb{R}^p$, $\mathbb{E}\langle X, t\rangle^2 = \|t\|_2^2$) and satisfies

$$\forall 1 \leq q \leq c_1 \log(p), 1 \leq i \leq p, \quad \|\langle X, e_i \rangle\|_{L_q} \leq c_1 \sqrt{q} \ , \tag{4.17}$$

for $c_1, c_2 > 0$ two absolute constants and where $(e_i)_{i=1}^p$ denotes the canonical basis of $\mathbb{R}^p$. Since any a real valued random variable Z is L_0-sub-Gaussian if and only if for all $q \geq 1$, $\|Z\|_{L_q} \leq c_3 L_0 \sqrt{q}$, for $c_3 > 0$ an absolute constant, the condition (4.17) imposes "$c_1 \log(p)$ sub-Gaussian moments" on the design X. From Theorem 1.6 in (Mendelson, 2017), if condition (4.17) holds, we get $\tilde{r}(\tilde{A}) \leq c_4 r(\tilde{A})$ for $c_4 > 0$ an absolute constant and the following theorem holds:

Theorem 4.7. *Let $\tilde{r} = c_4 \min(r_1^*, r_2^*)$. Let (X, Y) be a random variable such that $Y = \langle X, t^* \rangle + W$, where $t^* \in \mathbb{R}^p$ and W a symmetric random variable independent from X such that there exists $\gamma > 0$ with $F_W(\delta - 2(C')^2\tilde{r}) - F_W(2(C')^2\tilde{r} - \delta) \geq \gamma$. X is assumed to be an isotropic random vector satisfying condition (4.17). Assume that $(X_i, Y_i)_{i \in \mathcal{I}}$ are independent and distributed as (X, Y). Let $S \geq 7|\mathcal{O}|/3$. With probability larger than $1 - \exp(-S/504)$, the estimators $\hat{t}^{\delta,\alpha}_{\lambda,S}$ defined in (4.16) with*

$$\lambda = \frac{\max\left((\tilde{r})^2, \frac{5588\delta^2}{\gamma^2}\frac{S}{N}\right)}{\phi(t^*)}$$

satisfies

$$\|\hat{t}^{\delta,\alpha}_{\lambda,S} - f^*\|_2^2 \leq (8+\gamma)^2 \max\left((\tilde{r})^2, \frac{5588\delta^2}{\gamma^2}\frac{S}{N}\right) \quad \phi(\hat{t}^{\delta,\alpha}_{\lambda,S} - f^*) \leq (4 + 3\gamma/4)\phi(t^*) \ .$$

When $S \lesssim N(\tilde{r})^2$, Theorem 4.7 improves Theorem 4.6 by relaxing the sub-Gaussian Assumption. Moreover, for $S \asymp N(\tilde{r})^2$ up to $3N(\tilde{r})^2/7$ outliers can be present in the dataset without affecting the error rate. Note also that it is possible to adapt the estimator in a data-driven way to the best S and λ by using a Lepski's adaptation as we have done in Theorem 4.3.

Remark 4.2. *In Theorems 4.6 and 4.7, we assumed that the design X is isotropic. This assumption is only used for the computation of the Gaussian mean-with of the intersection of the ℓ_1 ball with the ℓ_2 ball. Using the recent work from (C Bellec, 2019) it is possible to extend the result for more general covariance matrices.*

4.4.2 Application to RKHS

In this section, we consider regularization methods in some general Reproducing Kernel Hilbert Space (RKHS) (cf. (Steinwart and Christmann, 2008) for a specific analysis on RKHS). The regularization function $\phi(\cdot)$ is defined as $\phi(\cdot) = \|\cdot\|^2_{\mathcal{H}_K}$ where $\|\cdot\|_{\mathcal{H}_K}$ is the norm in the space $\mathcal{H}_K$ associated to a kernel K. This section is inspired from the work in (Alquier et al., 2019). The authors established convergence rates when $\phi(\cdot) = \|\cdot\|_{\mathcal{H}_k}$ and $F = RB_{\mathcal{H}_K}$, for $R > 0$, for classification problems under a much stronger global Margin assumption. We improve their work in many aspects 1) heavy-tailed noise can be handled, 2) the margin assumption is replaced by the weaker local Bernstein condition, 3) we can analyse the regularization $\phi(\cdot) = \|\cdot\|^2_{\mathcal{H}_K}$ and 4) there is no restriction on the class $F = \mathcal{H}_K$, we do not restrict F to be a regularization ball in $\mathcal{H}_K$.
Using Theorems 4.4 we derive explicit bounds on the error rates depending on $\|f^*\|_{\mathcal{H}_K}$ for the minmax-MOM estimators. For the RERM, we could use Theorem 4.2. However, it turns out that

the sub-Gaussian Assumption 4.4 on the class $F - f^*$ is complicated to verify for RKHS and the application of Theorem 4.2 may be tricky. Instead, we derive another analysis where no sub-Gaussian assumption is required. In the precise example of RKHS, our homogeneity argument implies that we can restrict ourselves to a bounded class of functions. As a consequence, we can use concentration tools such as Talagrand's inequality instead of results from the sub-Gaussian theory. Nothing has to be assumed on the design X.

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 $Y_i \in \mathcal{Y}$ where $\mathcal{Y} = \{-1, 1\}$ for binary classification problems and $\mathcal{Y} = \mathbb{R}$ for regression problems. 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 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 an 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 $c_1, \cdots, c_n \in \mathbb{R}$, $\sum_{i=1}^n \sum_{j=1}^n c_i c_j K(x_i, x_j) \geq 0$, then by Mercer's theorem there exists an orthonormal basis $(\phi_i)_{i=1}^\infty$ of $L_2(\mu)$ 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

$$\begin{aligned} T_K : L_2(\mu) &\to L_2(\mu) \\ (T_K f)(x) &= \int K(x, y) f(y) d\mu(y) \ . \end{aligned} \tag{4.18}$$

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 $L_2(\mu)$ 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

$$B_{\mathcal{H}_K} = \{ f_\beta(\cdot) = \big\langle \beta, \Phi(\cdot) \big\rangle_{\ell_2}, \ \|\beta\|_2 \leq 1 \} \ , \tag{4.19}$$

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 therefore a convex class of functions from $\mathcal{X}$ to $\mathbb{R}$ that can be used as a learning class F. Let the *oracle* f^* be defined as

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

Let f be in $\mathcal{H}_K$, by the reproducing property and Cauchy-Schwarz we have for all x, y in $\mathcal{X}$

$$|f(x) - f(y)| = \langle f, K_x - K_y \rangle \leq \|f\|_{\mathcal{H}_K} \|K_x - K_y\|_{\mathcal{H}_K} \ . \tag{4.20}$$

From Equation (4.20), 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 $\mathcal{X}$. Let $\bar{\ell}$ be any loss function satisfying Assumption 4.2, the estimators $\hat{f}^{\phi}_{\lambda}$ and $\hat{f}^{\phi}_{\lambda,S}$ defined respectively in Equation (4.4) and (4.9) are given by

$$\hat{f}^{\phi}_{\lambda} = \operatorname*{argmin}_{f \in \mathcal{H}_K} \frac{1}{N} \sum_{i=1}^{N} \bar{\ell}(f(X_i), Y_i) + \lambda \|f\|^2_{\mathcal{H}_K} \tag{4.21}$$

and

$$\hat{f}^{\phi}_{\lambda,S} = \operatorname*{argmin}_{f \in \mathcal{H}_K} \sup_{g \in \mathcal{H}_K} MOM_S(\ell_f - \ell_g) + \lambda\big(\|f\|^2_{\mathcal{H}_K} - \|g\|^2_{\mathcal{H}_K}\big). \tag{4.22}$$

It is clear that $\phi(\cdot) = \|\cdot\|^2_{\mathcal{H}_K}$ verifies Assumption 4.3 with $\eta = 2$ We establish oracle inequalities for $\hat{f}^{\phi}_{\lambda}$ and $\hat{f}^{\phi}_{\lambda,S}$ respectively defined in Equation (4.21) and (4.22) when the loss satisfies Assumption 4.2. In (Mendelson et al., 2010; Meister and Steinwart, 2016; Wu et al., 2006; Smale and Zhou, 2007) for the quadratic loss function and (Eberts et al., 2013; Farooq and Steinwart, 2019) for the pinball loss (which is Lipschitz), the authors establish error bounds for when the target Y is assumed to satisfy $Y \in [-M, M]$ almost surely which is a really strong Assumption. Our analysis applies when the target Y is unbounded and may even be heavy-tailed which is, as far as we know, a new result. In (Caponnetto and De Vito, 2007) the authors do not assume that the target Y is bounded. However, their analysis requires to control the Laplace transform of the noise $Y - f^*(X)$ (see Assumption 2 in (Caponnetto and De Vito, 2007)). As a consequence they cannot consider heavy-tailed noise. In (Eberts et al., 2013; Farooq and Steinwart, 2019) the authors are also interested in the approximation error of kernel methods and compare ourselves with their results is a complicated task. We obtain the same error rate as (Mendelson et al., 2010; Caponnetto and De Vito, 2007) when the eigenvalues of the integral operator T_K satisfies $\lambda_n \leq \beta n^{-1/p}$ for some $0 < p < 1$ and $\beta > 0$ an absolute constant when Y may be **unbounded and heavy-tailed**. The value of p is related with the smoothness of the space $\mathcal{H}_K$. Different kinds of spectrum could be analysis. 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 studied in (Caponnetto and De Vito, 2007; Mendelson et al., 2010) for instance.

New general analysis for the RERM

Since every RKHS are convex, Assumption 4.1 holds. Therefore, when the loss function satisfies Assumption 4.2, to use Theorem 4.2 it is necessary to verify Assumptions 4.4 and 4.5. However, it

turns out that the sub-Gaussian Assumption on the class $F - f^*$ cannot be verified in practice except for very precise Kernels. Our analysis (see Section 4.6.1) requires the sub-Gaussian Assumption to show that with an exponentially large probability for all f in F such that $\|f - f^*\|_{L_2} \leq r(A^*)$ and $\phi(f - f^*) \leq \eta(2 + 2/A^*)\phi(f^*)$:

$$\left|(P - P_N)\mathcal{L}_f\right| \leq \frac{r^2(A^*)}{2A^*} \ , \tag{4.23}$$

where A^* satisfies Assumption 4.5 and $r(\cdot)$ is the complexity parameter defined in Definition 4.3. However, when $F = \mathcal{H}_K$ we have $\{f \in F : \phi(f - f^*) \leq \eta(2+2/A^*)\phi(f^*)\} = \{f \in \mathcal{H}_K : \|f - f^*\|_{\mathcal{H}_K} \leq 2\sqrt{1 + 1/A^*}\|f^*\|_{\mathcal{H}_K}\}$. Moreover, from the reproducing property, for all $x \in \mathcal{X}$ and all f in $\mathcal{H}_K$ such that $\|f - f^*\|_{\mathcal{H}_K} \leq 2\sqrt{1 + 1/A^*}\|f^*\|_{\mathcal{H}_K}$ we have

$$\begin{aligned} |f(x) - f^*(x)| = \left\langle f - f^*, K_x \right\rangle_{\mathcal{H}_K} &\leq \|f - f^*\|_{\mathcal{H}_K}\|K_x\|_{\mathcal{H}_K} \\ &= \|f - f^*\|_{\mathcal{H}_K}\sqrt{K(x,x)} \leq 2\sqrt{(1 + 1/A^*)\|K\|_\infty}\|f^*\|_{\mathcal{H}_K} \end{aligned}$$

Therefore, when $F = \mathcal{H}_K$, for K a bounded Kernel, the control of (4.23) is over a bounded class of functions. As a consequence, the sub-Gaussian Assumption is no longer necessary. Instead we develop another analysis based of the Bousquet's version of Talagrand's inquality (Bousquet, 2002). Since no sub-Gaussian assumption is required we use another complexity parameter where the Rademacher complexity replaces the Gaussian mean-width.

$$\bar{r}(A) = \inf\left\{ r > 0, \quad \mathbb{E} \sup_{\substack{f \in F : \|f - f^*\|_{L_2} \leq r, \\ \|f - f^*\|_{\mathcal{H}_K} \leq 2\sqrt{2 + 1/A}\|f^*\|_{\mathcal{H}_K}}} \sum_{i=1}^{N} \sigma_i (f - f^*)(X_i) \leq \frac{N r^2}{64AL} \right\} \tag{4.24}$$

We also adapt the local Bernstein assumption to the Definition (4.24).

Assumption 4.10. *There exists a constant $\bar{A} \geq 1$ such that for all $f \in \mathcal{H}_K$ if $\|f - f^*\|_{L_2} = 2L\sqrt{(2 + 1/\bar{A})\|K\|_\infty}\|f^*\|_{\mathcal{H}_K}\bar{r}(\bar{A})$ and $\|f - f^*\|_{\mathcal{H}_K} \leq 2\sqrt{2 + 1/\bar{A}}\|f^*\|_{\mathcal{H}_K}$ then $\|f - f^*\|_{L_2}^2 \leqslant \bar{A}P\mathcal{L}_f$.*

Theorem 4.8. *Let $(X_i, Y_i)_{i=1}^N$ be i.i.d random variables with common distribution P. Let ℓ be a loss function satisfying Assumption 4.2 with $L \geq 1$. Let $\mathcal{H}_K$ be a RKHS associated to a bounded Kernel K. Grant Assumption 4.10 such that $\bar{A} \geq 1$. Let $U = 2L\sqrt{(2 + 1/\bar{A})\|K\|_\infty}$. With probability larger than*

$$1 - 2\exp\left(- \frac{N\bar{r}^2(\bar{A})}{64(L\bar{A})^2} \right)$$

for all regularization parameters $\lambda \geq \lambda_0 = \max(1, U\|f^*\|_{\mathcal{H}_K})\bar{r}^2(\bar{A})/\|f^*\|_{\mathcal{H}_K}^2$ the estimators $\hat{f}_\lambda^\phi$ defined in Equation (4.21) satisfies

$$\begin{aligned} &\|\hat{f}_\lambda^\phi - f^*\|_{L_2} \leq (4 + 6\bar{A})\lambda \frac{\|f^*\|_{\mathcal{H}_K}^2}{\max(1, \sqrt{U\|f^*\|_{\mathcal{H}_K}})\bar{r}(\bar{A})} \\ &\textit{and} \quad \|\hat{f}_\lambda^\phi - f^*\|_{\mathcal{H}_K} \leq (8 + 4/\bar{A})\|f^*\|_{\mathcal{H}_K}. \end{aligned}$$

The proof can be find in Section 4.6.3. Theorem 4.8 is similar to Theorem 4.2 for RKHS when the sub-Gaussian assumption is relaxed. By taking $\lambda = \lambda_0$ we get

$$\|\hat{f}_\lambda^\phi - f^*\|_{L_2} \leq (4 + 6\bar{A}) \max(1, \sqrt{U\|f^*\|_{\mathcal{H}_K}})\bar{r}(\bar{A}) \ .$$

When $\|f^*\|_{\mathcal{H}_K} \leq M$, we obtain the same bounds as the one in Theorem 4.2 (up to a constant depending on $\bar{A}$ and $\|K\|_\infty$) and a Lespki's procedure as in Theorem 4.3 yields to an adaptive estimator. Note that the assumption that $\|K\|_\infty < \infty$ is really weak since any continuous kernel on a compact space is bounded. Moreover many results in RKHS are derived for the Gaussian Kernel with is bounded by 1, (Farooq and Steinwart, 2019; Steinwart and Christmann, 2008).

Explicit bounds for the ERM and the minmax MOM estimators

To obtain explicit bounds in Theorems 4.4 and 4.8 it is necessary to calculate the complexity parameters $\bar{r}(\bar{A})$ and $\tilde{r}(\tilde{A})$. To do so, we have to compute the Rademacher complexity of the set $\{f \in \mathcal{H}_K : \ \|f - f^*\|_{\mathcal{H}_K}^2 \leq \rho, \ \|f - f^*\|_{L_2} \leq r\}$ for any $\rho, r > 0$. From Theorem 2.1 in (Mendelson, 2003), if K is a bounded kernel, then for all $\rho, r > 0$

$$\mathbb{E} \sup_{f \in \mathcal{H}_K \cap (f^* + rB_{L_2} \cap \rho B_{\mathcal{H}_K})} \frac{1}{\sqrt{N}} \left| \sum_{i=1}^N \sigma_i (f - f^*)(X_i) \right| \leq \sqrt{2}\|K\|_\infty \left(\sum_{k=1}^\infty \left(\rho^2 \lambda_k \wedge r^2 \right) \right)^{1/2}$$

Remark 4.3. *Since the feature map Φ defines an isometry between $\mathcal{H}_K$ and ℓ_2, the computation of the Gaussian mean-width of the set $\{f \in \mathcal{H}_K : \ \|f - f^*\|_{\mathcal{H}_K}^2 \leq \rho, \ \|f - f^*\|_{L_2} \leq r\}$ is equivalent to the computation of the Gaussian mean-width of an ellipsoid in ℓ_2. Consequently, it is easy to show that Rademacher complexity and Gaussian mean-width (and thus $\bar{r}(A)$ and $r(A)$) are equivalent.*

In the case where the eigenvalues $\lambda_k \leq \beta k^{-1/p}$ for all $k \in \mathbb{N}^*$ and $0 < p < 1$, where $\beta > 0$ is an absolute constant and $\rho/r \geq 1$, straightforward computations give

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

It follows that for any bounded kernel K such that the eigenvalues assoicated to T_K satisfy $\lambda_k \leq \beta k^{-1/p}$ for all $k \in \mathbb{N}^*$ and $0 < p < 1$ and $A > 0$

$$\tilde{r}^2(A) = C(A, \beta, L, p) \frac{\|f^*\|_{\mathcal{H}_K}^{(2p)/(p+1)}}{N^{1/(p+1)}} = 6\bar{r}^2(A)$$
$$\text{where} \quad C(A, \beta, L, p) = \left(384 A \beta L\right)^{2/(p+1)} \left(4(2 + 1/A)\right)^{2p/(p+1)}$$

Now, let us turn to Bernstein condition. We use the results from (Chinot et al., 2019b) where the local Bernstein condition has been extensively studied for many convex and Lipschitz loss functions. In Section 4.4.1 we studied the Huber loss function. Here, we consider the absolute loss (which is the quantile loss for $\tau = 1/2$). Let us present the Assumptions required to study the Bernstein condition for the quantile loss function.

Assumption 4.11. *Let $r, \rho, \varepsilon > 0$.*

- *a) There exists $C' > 0$ such that for all $f \in F$ such that $\|f - f^*\|_{L_2} = r$ and $\|f - f^*\|_{\mathcal{H}_K} \leq \rho$, $\|f - f^*\|_{L_{2+\varepsilon}} \leq C'\|f - f^*\|_{L_2}$*
- *b) Let C' be the constant defined above. There exists $\alpha > 0$ such that, for all $x \in \mathcal{X}$ and for all z in $\mathbb{R}$ such that $|z - f^*(x)| \leq (\sqrt{2}C')^{(2+\varepsilon)/\varepsilon} r$, we have $f_{Y|X=x}(z) \geq \gamma$, where $f_{Y|X=x}$ is the conditional density function of Y given $X = x$.*

Assumption 4.11 and 4.9 are very similar. When $Y = f^*(X) + W$, for f^* in $\mathcal{H}_K$ and W is a symmetric noise, condition b) simply means that the noise W puts enough mass around 0.

Theorem 4.9 ((Chinot et al., 2019b))**.** *Grant Assumptions 4.11 (with parameter r, ρ and γ). Then, for all $f \in F$ satisfying $\|f - f^*\|_{L_2} = r$ and $\|f - f^*\|_{\mathcal{H}_K} \leq \rho$, $\|f - f^*\|_{L_2}^2 \leqslant (4/\gamma) P\mathcal{L}_f$.*

For kernel methods, the point a) of Assumption 4.11 is a $L_{2+\varepsilon}/L_2$-norm equivalence which is only required in the ball defined by the norm in the RKHS. Let f in F such that $\|f - f^*\|_{\mathcal{H}_K} \leq \rho$ and $\|f - f^*\|_{L_2} = r$, we have

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

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

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

Therefore, the point a) holds with $C' = (\rho\|K\|_\infty / r)^{\varepsilon/(2+\varepsilon)}$. Let us turn to the point b). From the fact that $C' = (\rho\|K\|_\infty / r)^{\varepsilon/(2+\varepsilon)}$, we have $\sqrt{2C'}^{(2+\varepsilon)/\varepsilon} r = 2^{(2+\varepsilon)/2\varepsilon} \rho \|K\|_\infty$. For example, when $Y = g(X) + W$, where $g \in \mathcal{H}_K : \mathcal{X} \mapsto \mathbb{R}$ and W is symetric and independent from X, it is easy to see that $f^* = g$. In this case the second point of Assumption 4.11 can be rewritten as $f_W(z) \geq \gamma$ for all $z \in \mathbb{R}$ such that $|z| \leq 2^{(2+\varepsilon)/2\varepsilon} \rho \|K\|_\infty$, where f_W denotes the density function of W. It simply means that the noise puts enough mass around 0.

We are now in position to state our main Theorems in a RKHS associated with a bounded kernel when the absolute loss function is considered for the RERM and the minmax MOM estimators.

Theorem 4.10. *Let $\mathcal{X}$ be some measurable space and $K : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ be a positive definite bounded kernel where $\mathcal{H}_K$ denote its associated RKHS. Let $(\lambda_k)_{k=1}^\infty$ be the sequence of eigenvalues associated to T_K in $L_2(\mu)$ such that $\lambda_k \leq \beta k^{-1/p}$ for all $k \in \mathbb{N}^*$ and $0 < p < 1$, where $\beta > 0$ is an absolute constant. For any $x \in \mathcal{X}$, let $f_{Y|X=x}$ denote the conditional density function of Y given $X = x$. Let us assume that there exists $\gamma > 0$ such that, for all $x \in \mathcal{X}$ and for all z in $\mathbb{R}$ such that $|z - f^*(x)| \leq 2\sqrt{8+\gamma} \|f^*\|_{\mathcal{H}_K} \|K\|_\infty$, we have $f_{Y|X=x}(z) \geq \gamma$. Let $(X_i, Y_i)_{i=1}^N$ be i.i.d random variables distributed as (X, Y). Then with probability larger than*

$$1 - \exp\left(- \frac{\gamma C(4/\gamma, \beta, 1, p)}{256} N^{p/(p+1)} \|f^*\|_{\mathcal{H}_K}^{2p/(p+1)} \right),$$

when

$$\lambda = C(4/\gamma, \beta, 1, p)\max(1, (8+\gamma)\|K\|_\infty \|f^*\|_{\mathcal{H}_K}) \frac{\|f^*\|_{\mathcal{H}_K}^{2/(p+1)}}{N^{1/(1+p)}} \quad ,$$

the estimator $\hat{f}_\lambda^\phi$ *associated to the absolute loss function defined in Equation* (4.21) *satisfies*

$$\|\hat{f}_\lambda^\phi - f^*\|_{L_2}^2 \leq (4 + 3/(2\gamma))C(4/\gamma, \beta, 1, p)\max(1, (8+\gamma)\|K\|_\infty \|f^*\|_{\mathcal{H}_K}) \frac{\|f^*\|_{\mathcal{H}_K}^{2/(p+1)}}{N^{1/(1+p)}}$$
$$\text{and} \quad \|\hat{f}_\lambda^\phi - f^*\|_{\mathcal{H}_K} \leq (8+\gamma)\|f^*\|_{\mathcal{H}_K}$$

The error rate in Theorem 4.10 is the same as in (Mendelson et al., 2010). However **our analysis do not require that the target Y is bounded. It can even be heavy-tailed**. Note also that nothing is assumed on the design X.

Remark 4.4. *When $Y = f^*(X) + W$, where W is a standard Cauchy distribution, the condition $f_{Y|X=x}(z) \geq \gamma$ for z in $\mathbb{R}$ such that $|z - f^*(x)| \leq 2\sqrt{8+\gamma}\|f^*\|_{\mathcal{H}_K}\|K\|_\infty$ is satisfied as long as there exists $\gamma \in (0, 1]$ such that*

$$\frac{1}{\pi\big(1 + 4(8+\gamma)\|f^*\|_{\mathcal{H}_K}^2 \|K\|_\infty^2\big)} \geq \gamma$$

which holds for $\gamma = \min(1, 1/(\pi(1 + 36\|f^*\|^2\|K\|_\infty^2)))$. Consequently the analysis holds for heavy-tailed distribution.

Let us turn to the MOM-estimators.

Theorem 4.11. *Let $\mathcal{X}$ be some measurable space and $K : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ be a positive definite bounded kernel where $\mathcal{H}_K$ denote its associated RKHS. Let $(\lambda_k)_{k=1}^\infty$ be the sequence of eigenvalues associated to T_K in $L_2(\mu)$ such that $\lambda_k \leq \beta k^{-1/p}$ for all $k \in \mathbb{N}^*$ and $0 < p < 1$, where $\beta > 0$ is an absolute constant. For any $x \in \mathcal{X}$, let $f_{Y|X=x}$ denote the conditional density function of Y given $X = x$. Let us assume that there exist $\gamma > 0$ such that, for all $x \in \mathcal{X}$ and for all z in $\mathbb{R}$ such that $|z - f^*(x)| \leq 2\sqrt{8+\gamma}\|f^*\|_{\mathcal{H}_K}\|K\|_\infty$, we have $f_{Y|X=x}(z) \geq \gamma$. Let us assume that $(X_i, Y_i)_{i\in\mathcal{I}}$ are independent and distributed as (X, Y). Let $S \geq 7|\mathcal{O}|/3$. Let:*

$$C_{S,N} = \max\left(6C(4/\gamma, \beta, 1, p)\frac{\|f^*\|_{\mathcal{H}_K}^{(2p)/(p+1)}}{N^{1/(p+1)}}, \frac{13888}{\gamma}\frac{S}{N}\right)$$

Then with probability larger than $1 - \exp(-S/504)$ when

$$\lambda = \frac{C_{S,N}}{\|f^*\|_{\mathcal{H}_K^2}} \quad ,$$

the estimator $\hat{f}_{\lambda,S}^\phi$ associated to the absolute loss function defined in Equation (4.22) *satisfies*

$$\|\hat{f}_{\lambda,S}^\phi - f^*\|_{L_2}^2 \leq (4 + 3/(2\gamma))C_{S,N} \quad \text{and} \quad \|\hat{f}_{\lambda,S}^\phi - f^*\|_{\mathcal{H}_K} \leq (8+\gamma)\|f^*\|_{\mathcal{H}_K}$$

When $S \lesssim N^{p/(p+1)}\|f^*\|_{\mathcal{H}_K}^{(2p)/(p+1)}$ we recover the bounds from Theorem 4.10. However for the minmax MOM-estimators, up to $3S/7$ outliers can contaminate the dataset without deteriorated the error rate.

4.5 Conclusion

We have presented two general results for the RERM and minmax-MOM estimators describing the statistical properties of regularization in learning theory. For those two estimators we do not assume that the regularization is a norm which is, as far as we know a new general result for Lipschitz and convex loss functions. Under the local Bernstein Assumption, we can obtain rates of convergence depending on $\phi(f^*)$. Results for the RERM have been derived under the i.i.d and the sub-Gaussian Assumptions on the class $F - f^*$ while no concentration Assumption is required for minmax MOM-estimators. For MOM-estimators, a number of outliers smaller than *square of the rate of convergence in a non-contaminated setting* $\times$ *number of observations* does not deteriorate the learning procedure. We studied the particular example of SVM where no sub-Gaussian assumption on the class F is required and when the target Y may be heavy-tailed, widely improving the existing results in the literature.
There are a number of interesting directions in which this work can be extended. One relevant and closely related problem is to obtain sparsity bounds, i.e bounds depending on an underlying structure of the *oracle* f^* such as the sparsity or the rank of the *oracle* f^*. It has been partially done (under a really strong Assumption) in (Alquier et al., 2019; Chinot, 2019b) when the regularization function if a norm. However without this Assumption, the proofs no longer hold and a new analysis has to be developed.

4.6 Proof main theorems

4.6.1 Proof of Theorems 4.2, 4.3 RERM

In the remaining of the proof we shall use repeatedly the following notations

$$A = A^*, \quad \theta = \frac{1}{2A}, \quad \delta = \frac{2}{A} + 3 \quad \gamma = \frac{2}{A} + 2 \ .$$

Proof Theorem 4.2

Proof of Theorem 4.2 is split into two parts. First, we identify an event onto which the statistical behavior of the regularized estimator $\hat{f}_\lambda := \hat{f}^\phi_\lambda$ can be controlled using only deterministic arguments. Then, we prove that this event holds with a probability at least as large as the one in (4.6). Let us define $\rho^* = (2+\gamma)\eta\phi(f^*)$. We first introduce this event:

$$\Omega := \left\{ \text{for all } f \in F \cap (f^* + r^* B_{L_2}) \cap B^\phi_{\rho^*}(f^*), \quad \left|(P - P_N)\mathcal{L}_f\right| \leq \theta (r^*)^2 \right\}$$

where we recall that $r^* = r(A^*)$ and $B^\phi_{\rho^*}(f^*) = \{f \in F : \ \phi(f - f^*) \leq \rho^*\}$.

Lemma 4.1. *Let* $\lambda \geq (r^*)^2/\phi(f^*)$*, on the event* Ω *we have*

- *For all* $f \in F \backslash \mathcal{B}_\lambda$, $P_N \mathcal{L}^\lambda_f > 2(\theta+1)\lambda\phi(f^*)$

- *For all* $f \in F \cap \mathcal{B}_\lambda$, $P_N\mathcal{L}_f^\lambda \geq -2(\theta+1)\lambda\phi(f^*)$

Proposition 4.1. *Let* $\lambda \geq \lambda_0 := (r^*)^2/\phi(f^*)$*, on the event* Ω*, one has*

$$\phi(\hat{f}_\lambda - f^*) \leq \rho^*, \quad \|\hat{f}_\lambda - f^*\|_{L_2} \leq \lambda\frac{\delta\phi(f^*)}{(A^{-1}-\theta)r^*}$$

Proof. Let $\lambda \geq \lambda_0$, we denote $\mathcal{B}_\lambda = \Big(f^* + \big(\lambda\delta\phi(f^*)/((A^{-1}-\theta)r^*)\big)B_{L_2}\Big) \cap B^\phi_{\rho^*}(f^*)$. We want to prove that $\hat{f}_\lambda \in \mathcal{B}_\lambda$.We recall that the regularized empirical excess loss function is defined for all $f \in F$ by

$$P_N\mathcal{L}_f^\lambda = P_N\mathcal{L}_f + \lambda\big(\phi(f) - \phi(f^*)\big).$$

Since $\hat{f}_\lambda$ is such that $P_N\mathcal{L}^\lambda_{\hat{f}_\lambda} \leq 0$, it is enough to prove that $P_N\mathcal{L}_f^\lambda > 0$ for all $f \in F\backslash\mathcal{B}_\lambda$ to get that $\hat{f}_\lambda \in \mathcal{B}_\lambda$. In fact, for the adaptive procedure it will be necessary to use the results from Lemma 4.1 which is equivalent (up to the choice of the constants) to show than $P_N\mathcal{L}_f^\lambda > 0$ for all $f \in F\backslash\mathcal{B}_\lambda$. From Lemma 4.1 it follows immediately that $\phi(\hat{f}_\lambda - f^*) \leq \rho^*$ and $\|\hat{f}_\lambda - f^*\|_{L_2} \leq \lambda\frac{\delta\phi(f^*)}{(A^{-1}-\theta)r^*}$

■

Proof. **Lemma 4.1**

The proof follows from an homogeneity argument saying that if $P_N\mathcal{L}^\lambda_{f_0} > 2(\theta+1)\lambda\phi(f^*)$ on the border of $\mathcal{B}_\lambda$ then we also have $P_N\mathcal{L}_f^\lambda > 2(\theta+1)\lambda\phi(f^*)$ for all $f \in F$ outside $\mathcal{B}_\lambda$. Inside $\mathcal{B}_\lambda$ the arguments are similar.

Let f in F be outside of $\mathcal{B}_\lambda$. By convexity of F, there exists $f_0 \in F$ and $\alpha > 1$ such that $f - f^* = \alpha(f_0 - f^*)$ and $f_0 \in \partial\mathcal{B}_\lambda$ where we denote by $\partial\mathcal{B}_\lambda$ the border of $\mathcal{B}_\lambda$. By definition, we either have: 1) $\phi(f_0 - f^*) = \rho^*$ and $\|f_0 - f^*\|_{L_2} \leq (\lambda\delta\phi(f^*))/((A^{-1}-\theta)r^*)$ in that case, α is such that $1 \leq \alpha \leq \phi(f-f^*)/\rho^*$ (see Lemma 4.5 in Section 4.7) or 2) $\|f_0 - f^*\|_{L_2} = (\lambda\delta\phi(f^*))/((A^{-1}-\theta)r^*)$ and $\phi(f_0 - f^*) \leq \rho^*$ and, in that case, $\alpha = \|f - f^*\|_{L_2}/\big((\lambda\delta\phi(f^*))/((A^{-1}-\theta)r^*)\big)$. We will treat the two cases independently.

Let us first explain the role of the convexity of the loss function by writing down an homogeneity argument linking the empirical excess risk of f to the one of f_0. For all $i \in \{1, \cdots, N\}$, let $\psi_i : \mathbb{R} \to \mathbb{R}$ be defined for all $u \in \mathbb{R}$ by

$$\psi_i(u) = \bar{\ell}(u + f^*(X_i), Y_i) - \bar{\ell}(f^*(X_i), Y_i). \tag{4.25}$$

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

$$\begin{aligned} P_N\mathcal{L}_f &= \frac{1}{N}\sum_{i=1}^N \psi_i\big(f(X_i) - f^*(X_i)\big) = \frac{1}{N}\sum_{i=1}^N \psi_i(\alpha(f_0(X_i) - f^*(X_i))) \\ &\geq \frac{\alpha}{N}\sum_{i=1}^N \psi_i((f_0(X_i) - f^*(X_i))) = \alpha P_N\mathcal{L}_{f_0}. \end{aligned} \tag{4.26}$$

For the regularization part the same homogeneity arguments holds.

$$\phi(f) - \phi(f^*) = \phi\big(f^* + \alpha(f_0 - f^*)\big) - \phi(f^*) \geq \alpha\big(\phi(f_0) - \phi(f^*)\big)$$

where we used Lemma 4.6 (see Section 4.7). Therefore

$$P_N\mathcal{L}_f^\lambda \geq \alpha P_N\mathcal{L}_{f_0}^\lambda$$

Let us now place ourselves on the event Ω up to the end of the proof and let $f_0 \in F \cap \partial\mathcal{B}_\lambda$. We explore two cases depending on the localization of f_0 on the border of $\mathcal{B}_\lambda$: 1) $\phi(f_0 - f^*) = \rho^*$ and $\|f_0 - f^*\|_{L_2} \leq (\lambda\delta\phi(f^*))/((A^{-1}-\theta)r^*)$ which is the case where the regularization part helps to show that $P_N\mathcal{L}_{f_0}^\lambda > 2(\theta+1)\lambda\phi(f^*)$ or 2) $\|f_0 - f^*\|_{L_2} = (\lambda\delta\phi(f^*))/((A^{-1}-\theta)r^*)$ and $\phi(f_0-f^*) \leq \rho^*$ which is where the Bernstein's condition helps. We consider the first case which is when $\phi(f_0 - f^*) = \rho^*$

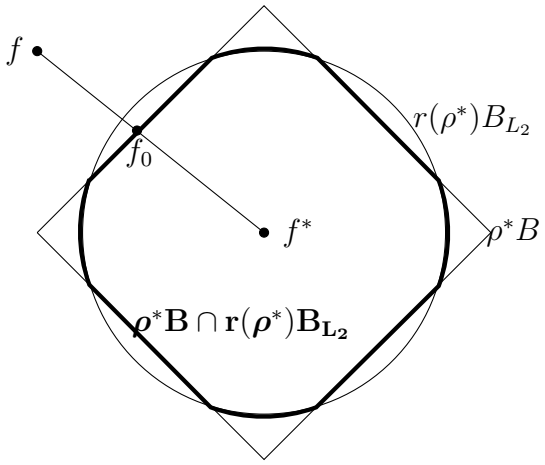


Figure 4.1: Construction of f_0.

and $\|f_0 - f^*\|_{L_2} \leq (\lambda\delta\phi(f^*))/((A^{-1} - \theta)r^*)$. There are two cases, either $\|f_0 - f^*\|_{L_2} \leq r^*$ or $\|f_0 - f^*\|_{L_2} \geq r^*$. In both cases, from the fact that $\phi(f_0 - f^*) \leq \eta\big(\phi(f_0) + \phi(f^*)\big)$ we have $\phi(f_0) - \phi(f^*) \geq \gamma\phi(f^*)$. If $\|f_0 - f^*\|_{L_2} \leq r^*$, on Ω we have $|(P - P_N)\mathcal{L}_{f_0}| \leq \theta(r^*)^2$ and we get

$$\begin{aligned} P_N\mathcal{L}_f^\lambda = P_N\mathcal{L}_f + \lambda\left(\phi(f) - \phi(f^*)\right) &\geq \alpha\big(P_N\mathcal{L}_{f_0} + \lambda\gamma\phi(f^*)\ \big) \geq \alpha\big(-\theta(r^*)^2 + \gamma\lambda\phi(f^*)\ \big) \\ &\geq (-\theta + \gamma)\lambda\phi(f^*) > 2(\theta+1)\lambda\phi(f^*) \end{aligned}$$

where we used the facts that $\lambda \geq (r^*)^2/\phi(f^*)$ and $P\mathcal{L}_{f_0} \geq 0$. If $r^* \leq \|f_0 - f^*\|_{L_2} \leq \lambda\delta\phi(f^*)/((A^{-1}-\theta)r^*)$ we use the same projection trick. Let $\alpha_1 = \|f_0 - f^*\|_{L_2}/r^*$ and set f_1 in F be such that $f_0 - f^* = \alpha_1(f_1 - f^*)$. We have $\|f_1 - f^*\|_{L_2} = r^*$ and $\phi(f_1 - f^*) \leq \rho^*$. Therefore on Ω we have

$$P_N\mathcal{L}_f^\lambda \geq \alpha\big(P_N\mathcal{L}_{f_0} + \gamma\lambda\phi(f^*)\ \big) \geq \alpha\big(\alpha_1 P_N\mathcal{L}_{f_1} + \gamma\lambda\phi(f^*)\ \big) \geq \gamma\lambda\phi(f^*) > 2(\theta+1)\lambda\phi(f^*)$$

Since, on Ω, $P_N\mathcal{L}_{f_1} \geq P\mathcal{L}_{f_1} - \theta(r^*)^2 \geq A^{-1}\|f_1 - f^*\|_{L_2} - \theta(r^*)^2 = (A^{-1} - \theta)(r^*)^2 > 0$ where we used Assumption 4.5.

We now turn to the second case where $\|f_0 - f^*\|_{L_2} = \lambda\delta\phi(f^*)/((A^{-1}-\theta)r^*)$ and $\phi(f_0-f^*) \leq \rho^*$. Remember that in this case $\alpha = \|f - f^*\|_{L_2} / \big((\lambda\delta\phi(f^*))/((A^{-1} - \theta)r^*)\big)$. The regularization part no longer helps. However, by the Bernstein Assumption 4.5 and using the same projection trick we get

$$\begin{aligned} P_N\mathcal{L}_f \geq \frac{\|f - f^*\|_{L_2}}{(\lambda\delta\phi(f^*))/((A^{-1} - \theta)r^*)} P_N\mathcal{L}_{f_0} &\geq \frac{\|f - f^*\|_{L_2}}{(\lambda\delta\phi(f^*))/((A^{-1} - \theta)r^*)} \frac{\|f_0 - f^*\|_{L_2}}{r^*} P_N\mathcal{L}_{f_1} \\ &\geq \frac{\|f - f^*\|_{L_2}}{r^*}(A^{-1} - \theta)(r^*)^2 \end{aligned}$$

where f_1 is such that $f_0 - f^* = \big(\|f_0 - f^*\|_{L_2}/(r^*)\big)(f_1 - f^*)$. We have $\|f_1 - f^*\|_{L_2} = r^*$ and $\phi(f_1 - f^*) \leq \rho^*$. Since $\|f - f^*\|_{L_2} \geq \lambda\delta\phi(f^*)/((A^{-1} - \theta)r^*)$, we finally get

$$P_N\mathcal{L}_f^\lambda \geq \frac{\|f - f^*\|_{L_2}}{r^*}(A^{-1} - \theta)(r^*)^2 - \lambda\phi(f^*) \geq (\delta - 1)\lambda\phi(f^*) > 2(\theta + 1)\lambda\phi(f^*)$$

We conclude the proof by studying $P_N\mathcal{L}_f^\lambda$ for $f \in F \cap \mathcal{B}_\lambda$. One more time there are two cases, either $\|f - f^*\|_{L_2} \leq r^*$ or $\|f - f^*\|_{L_2} \geq r^*$. In the first case, since $P\mathcal{L}_{f_0}$, on Ω we get that

$$P_N\mathcal{L}_f^\lambda \geq -\theta(r^*)^2 - \lambda\phi(f^*) \geq -(\theta + 1)\lambda\phi(f^*)$$

For $\|f - f^*\|_{L_2} \geq r^*$ using the projection trick, there exists $\alpha \geq 1$ such that $P_N\mathcal{L}_f \geq \alpha P_N\mathcal{L}_{f_0}$ where f_0 satisfies $\|f_0 - f^*\|_{L_2} = r^*$ and $\phi(f_0 - f^*) \leq \rho^*$. Therefore on Ω, using Assumption 4.5, we get $P_N\mathcal{L}_f \geq \alpha(A^{-1} - \theta)(r^*)^2 \geq -\theta\lambda\phi(f^*)$. Finally in that case

$$P_N\mathcal{L}_f^\lambda \geq -(\theta + 1)\lambda\phi(f^*)$$

■

Next, we prove that Ω holds with large probability. To that end, we use the results from (Alquier et al., 2019).

Lemma 4.2. *(Alquier et al., 2019) Assume that Assumption 4.2 and Assumption 4.4 hold. Let $F' \subset F$ then for every $u > 0$, with probability at least $1 - 2\exp(-u^2)$*

$$\sup_{f,g \in F'} |(P - P_N)(\mathcal{L}_f - \mathcal{L}_g)| \leq \frac{16LB}{\sqrt{N}} \left(w(F') + u d_{L_2}(F')\right)$$

where d_{L_2} is the L_2 metric, $d_{L_2}(F')$ is the L_2 diameter of F'.

It follows from Lemma 4.2 that for any $u > 0$, with probability larger that $1 - 2\exp(-u^2)$,

$$\begin{aligned} &\sup_{f \in F \cap (f^* + r^* B_{L_2}) \cap B_{\rho^*}^\phi(f^*)} \big|(P - P_N)\mathcal{L}_f\big| \leq \sup_{f,g \in F \cap (f^* + r^* B_{L_2}) \cap B_{\rho^*}^\phi(f^*)} \big|(P - P_N)(\mathcal{L}_f - \mathcal{L}_g)\big| \\ &\leq \frac{16LB}{\sqrt{N}} \left(w\big(F \cap (f^* + r^* B_{L_2}) \cap B_{\rho^*}^\phi(f^*)\big) + u d_{L_2}\big(F \cap (f^* + r^* B_{L_2}) \cap B_{\rho^*}^\phi(f^*)\big) \right) . \end{aligned}$$

We have $d_{L_2}\big(F\cap(f^*+r^*B_{L_2})\cap B^{\phi}_{\rho^*}(f^*)\big)\leq r^*$ and $w\big(F\cap(f^*+r^*B_{L_2})\cap B^{\phi}_{\rho^*}(f^*)\big)=w\big(F\cap r^*B_{L_2}\cap B^{\phi}_{\rho^*}(0)\big)$, By definition of the complexity parameter (see Equation (4.3)), for $u=\theta\sqrt{N}r^*/(32LB)$, with probability at least

$$1-2\exp\big(-\theta^2N(r^*)^2/(32^2L^2B^2)\big) \tag{4.27}$$

for every f in $F\cap(f^*+r^*B_{L_2})\cap B^{\phi}_{\rho^*}(f^*)$,

$$\big|(P-P_N)\mathcal{L}_f\big|\leq\theta(r^*)^2 \tag{4.28}$$

Proof Theorem 4.3

In this section we work on the event

$$\tilde{\Omega}:=\left\{\text{for all } f\in F\cap\left(f^*+\frac{2\delta}{A^{-1}-\theta}r^*B_{L_2}\right)\cap B^{\phi}_{\rho^*}(f^*),\quad \big|(P-P_N)\mathcal{L}_f\big|\leq\theta(r^*)^2\right\}$$

Using the same proof as the one for Ω, it easy to show that $\tilde{\Omega}$ holds with probability larger than

$$1-2\exp\left(-\frac{\big(\theta(A^{-1}-\theta)\big)^2N(r^*)^2}{(64LB\delta)^2}\right)$$

Note that $\Omega\subset\tilde{\Omega}$ and then Lemma 4.1 still holds.
Let us assume that $(\lambda_j)_{j=0}^J=(r_j^2/\phi_j)_{j=0}^J$ is non increasing. From the choice of $(\phi_j)_{j=0}^J$, there exists $\tilde{k}$ such that $\phi_{\tilde{k}}\leq\phi(f^*)\leq 2\phi_{\tilde{k}}$. Note that if $(\lambda_j)_{j=0}^J$ is non decreasing, it is enough to use the same proof with $\tilde{k}$ such that $(1/2)\phi_{\tilde{k}}\leq\phi(f^*)\leq\phi_{\tilde{k}}$.
Moreover, from Lemma 4.1, for all $\lambda\geq\lambda_0$, $T_\lambda(f^*)=-P_N\mathcal{L}^{\lambda}_{\hat{f}_\lambda}\leq(\theta+1)\lambda\phi(f^*)\leq 2(\theta+1)\lambda\phi_{\tilde{k}}$. Since $\phi_{\tilde{k}}\leq\phi(f^*)$ it follows that $\lambda_{\tilde{k}}\geq\lambda_0$. And finally

$$P_N\mathcal{L}^{\lambda}_{\hat{f}_\lambda}\leq 2(\theta+1)\phi_{\tilde{k}}\lambda_{\tilde{k}}\leq 2(\theta+1)\phi_k\lambda_k \text{ for all } k\geq\tilde{k} \tag{4.29}$$

From the definition of k^* and Equation (4.29) it follows that $k^*\leq\tilde{k}$ and thus, $\tilde{f}\in\hat{R}_{\tilde{k}}$. As a consequence, $P_N\mathcal{L}^{\lambda_{\tilde{k}}}_{\tilde{f}}\leq T_{\lambda_{\tilde{k}}}(\tilde{f})$ and we get

$$P_N\mathcal{L}^{\lambda_{\tilde{k}}}_{\tilde{f}}\leq 2(\theta+1)\lambda_{\tilde{k}}\phi_{\tilde{k}}\leq 2(\theta+1)\lambda_{\tilde{k}}\phi(f^*)$$

From Lemma 4.1 it follows that $\tilde{f}$ satisfies $\|\tilde{f}-f^*\|_{L_2}\leq\lambda_{\tilde{k}}\delta\phi(f^*)/((A^{-1}-\theta)r^*)\leq 2\lambda_{\tilde{k}}\delta\phi_{\tilde{k}}/((A^{-1}-\theta)r^*)\leq\big(2\delta/(A^{-1}-\theta)\big)r^*$ and $\phi(\tilde{f}-f^*)\leq\eta(2+\gamma)\phi(f^*)$.

We finish this section by showing a *oracle inequality* for $\tilde{f}$. From the fact that $\|\tilde{f}-f^*\|_{L_2}\leq\big(2\delta/(A^{-1}-\theta)\big)r^*$ and $\phi(\tilde{f}-f^*)\leq\eta(2+\gamma)\phi(f^*)$, it follows, on $\tilde{\Omega}$ that $(P-P_N)\mathcal{L}_{\tilde{f}}\leq\theta(r^*)^2$. For all $\lambda>0$

$$P\mathcal{L}_{\tilde{f}}=P_N\mathcal{L}_{\tilde{f}}+(P-P_N)\mathcal{L}_{\tilde{f}}\leq P_N\mathcal{L}^{\lambda}_{\tilde{f}}+\lambda\big(\phi(f^*)-\phi(\tilde{f})\big)+\theta(r^*)^2\leq P_N\mathcal{L}^{\lambda}_{\tilde{f}}+\lambda\phi(f^*)+\theta(r^*)^2\ .$$

In particular for $\lambda=\lambda_{\tilde{k}}$ one has $P_N\mathcal{L}^{\lambda_{\tilde{k}}}_{\tilde{f}}\leq 2(\theta+2)\phi_{\tilde{k}}\lambda_{\tilde{k}}\leq 2(\theta+1)(r^*)^2$ and $\lambda_{\tilde{k}}\phi(f^*)\leq 2(r^*)^2$. Finally

$$P\mathcal{L}_{\tilde{f}}\leq(4+3\theta)(r^*)^2$$

4.6.2 Proof Theorem 4.4 minmax MOM estimators

Let $\tilde{r}$ and $C_{S,r}$ design respectively $\tilde{r}(\tilde{A})$ and $C_{s,r}(\tilde{A})$. Moreover, all along the proof, the following notations will be used repeatedly.

$$A = \tilde{A}, \quad \theta = \frac{1}{2A}, \quad \delta = \frac{2}{A} + 3 \quad \gamma = \frac{2}{A} + 2, \quad \mu = \frac{\theta}{192L} \ .$$

The proof is divided into two parts. First, we identify an event where the minmax MOM estimators $\hat{f}_S^{\lambda} := \hat{f}_S$ is controlled. Then, we prove that this event holds with large probability. Let $S \geqslant 7|\mathcal{O}|/3$, and

$$C_{s,r} = \max\left(\frac{96L^2S}{\theta^2 N}, \tilde{r}^2\right) \quad \text{and} \quad \rho^* = \eta(2+\gamma)\phi(f^*)$$

Let $\mathcal{B}_{\lambda,S} = \{f \in E : \ \|f - f^*\|_{L_2} \leq \frac{\delta}{A^{-1}-\theta}\frac{\lambda\phi(f^*)}{\sqrt{C_{s,r}}} \quad \text{and} \quad \phi(f^* - f^*) \leq \rho^*\}$. Consider the following event

$$\Omega_S = \left\{\forall f \in F \cap \sqrt{C_{S,r}}B_{L_2} \cap B_{\rho^*}^{\phi}(f^*), \qquad \sum_{s=1}^{S} I\left(\left|(P_{B_s} - P)(\ell_f - \ell_{f^*})\right| \leq \theta C_{s,r}\right) \geqslant \frac{S}{2}\right\} \ . \quad (4.30)$$

Deterministic argument

Lemma 4.3. *$\hat{f}_S \in \mathcal{B}_{\lambda,S}$ if the following inequalities holds*

$$\sup_{f\in F\setminus\mathcal{B}_{\lambda,s}} MOM_S(\ell_{f^*} - \ell_f) + \lambda\big(\phi(f^*) - \phi(f)\big) \leq -2(\theta+1)\lambda\phi(f^*) \ , \quad (4.31)$$

$$\sup_{f\in F\cap\mathcal{B}_{\lambda,S}} MOM_S(\ell_{f^*} - \ell_f) + \lambda\big(\phi(f^*) - \phi(f)\big) \leq (\theta+1)\lambda\phi(f^*) \ . \quad (4.32)$$

Proof. For any $f \in F$, denote by $S(f) = \sup_{g\in F} MOM_S(\ell_f - \ell_g) + \lambda\big(\phi(f) - \phi(g)\big)$. If (4.31) holds, by homogeneity of MOM_S, any $f \in F\setminus\mathcal{B}_{\lambda,S}$ satisfies

$$S(f) \geqslant MOM_S(\ell_f - \ell_{f^*}) + \lambda\big(\phi(f) - \phi(f^*)\big) > 2(\theta+1)\lambda\phi(f^*) \ .$$

On the other hand, if (4.32) and (4.31) hold,

$$S(f^*) = \sup_{f\in F} MOM_S(\ell_{f^*} - \ell_f) + \lambda\big(\phi(f^*) - \phi(f)\big) \leqslant (\theta+1)\lambda\phi(f^*) \ .$$

Thus, by definition of $\hat{f}_S$ and (4.32),

$$S(\hat{f}_S) \leqslant S(f^*) \leqslant (\theta+1)\lambda\phi(f^*) \ .$$

Therefore, if (4.31) and (4.32) hold, $\hat{f}_s \in \mathcal{B}_{\lambda,S}$. ∎

Lemma 4.4. *For all $S \geq 7|\mathcal{O}|/3$ and $\lambda \geq C_{S,r}/\phi(f^*)$, inequalities* (4.31) *and* (4.32) *holds on Ω_S.*

Proof. The arguments are exactly the same as the one in the proof of Lemma 4.1. For all functions $f \in F \backslash \mathcal{B}_{\lambda,S}$ and for each block B_s there exist $\alpha \geq 1$ and $f_0 \in F$ in the border of $\mathcal{B}_{\lambda,S}$ such that $P_{B_s}\mathcal{L}_f \geq \alpha P_{B_s}\mathcal{L}_{f_0}$. We present here only one case (the others are trivial applications of the arguments in the proof of Lemma 4.1). In the case where $\phi(f_0 - f^*) = \rho^*$ and $\sqrt{C_{S,r}} \leq \|f_0 - f^*\|_{L_2} \leq (\lambda\delta\phi(f^*))/((A^{-1} - \theta)\sqrt{C_{S,r}})$. We still have $\lambda\big(\phi(f_0) - \phi((f^*)\big) \geq \lambda\gamma\phi(f^*)$. Using the projection trick, there exists $\alpha_1 > 1$ such that on each block B_s, $P_{B_s}\mathcal{L}_{f_0} \geq \alpha_1 P_{B_s}\mathcal{L}_{f_1}$ for f_1 such that $\|f_1 - f^*\|_{L_2} = \sqrt{C_{S,r}}$ and $\phi(f_1 - f^*) \leq \rho^*$ and then, on the event Ω_S, one more than $S/2$ blocks B_s

$$P_{B_s}\mathcal{L}_f^\lambda \geq \alpha\big(P_{B_s}\mathcal{L}_{f_0} + \gamma\lambda\phi(f^*)\ \big) \geq \alpha\big(\alpha_1 P_{B_s}\mathcal{L}_{f_1} + \gamma\lambda\phi(f^*)\ \big) \geq \gamma\lambda\phi(f^*) > 2(\theta+1)\lambda\phi(f^*) \quad (4.33)$$

where we used the fact that on Ω_S, there are at least $S/2$ blocks B_s such that, $P_{B_s}\mathcal{L}_{f_1} \geq P\mathcal{L}_{f_1} - \theta C_{S,r} \geq A^{-1}\|f_1 - f^*\|_{L_2}^2 - \theta C_{S,r} = (A^{-1} - \theta)C_{S,r} > 0$ and Assumption 4.8.
As Equation (4.33) holds on more than $S/2$ blocks we get that

$$MOM_S(\ell_f - \ell_{f^*}) + \lambda\big(\phi(f) - \phi(f^*)\big) \geq 2(\theta+1)\lambda\phi(f^*)$$

From the same arguments as the one in the proof of Lemma 4.1 we finally obtain

$$\sup_{f \in F\backslash\mathcal{B}_{\lambda,S}} MOM_S(\ell_{f^*} - \ell_f) + \lambda\big(\phi(f^*) - \phi(f)\big) < -2(\theta+1)\lambda\phi(f^*)\ ,$$
$$\sup_{f \in F\cap\mathcal{B}_{\lambda,S}} MOM_S(\ell_{f^*} - \ell_f) + \lambda\big(\phi(f^*) - \phi(f)\big) \leq (\theta+1)\lambda\phi(f^*)$$

which concludes to proof. ∎

Control of the stochastic event

Contrary to the deterministic argument, the control of the stochastic event is very different from the one for the RERM.

Proposition 4.2. *Grant Assumptions 4.2, 4.1, 4.3, 4.7 and 4.8. Let $S \geq 7|\mathcal{O}|/3$. Then Ω_S holds with probability larger than $1 - 2\exp(-S/504)$.*

Proof. Let $\mathcal{F} = \{f \in F :\ \|f - f^*\|_{L_2} \leq \sqrt{C_{S,r}},\ \phi(f - f^*) \leq \rho^*\}$ and let $h(t) = I\{t \geq 2\} + (t-1)I\{1 \leq t \leq 2\}$. This function satisfies $\forall t \in \mathbb{R}^+,\ I\{t \geq 2\} \leq h(t) \leq I\{t \geq 1\}$. Let $W_s = ((X_i, Y_i))_{i \in B_s}$ and, for any $f \in \mathcal{F}$, let $G_f(W_s) = (P_{B_s} - P)(\ell_f - \ell_{f^*})$. Let also $C_{S,r} = \max\left(96L^2S/(\theta^2 N), \tilde{r}^2\right)$. For any $f \in \mathcal{F}$, let

$$z(f) = \sum_{s=1}^{S} I\{|G_f(W_s)| \leq \theta C_{S,r}\}\ .$$

Proposition 4.2 will be proved if $\mathbb{P}\big(z(f) \geq S/2\big) \geqslant 1 - e^{-S/504}$. Let $\mathcal{S}$ denote the set of indices of blocks which have not been corrupted by outliers, $\mathcal{S} = \{s \in \{1, \cdots, S\} : B_s \subset \mathcal{I}\}$. Basic algebraic manipulations show that

$$z(f) \geqslant |\mathcal{S}| - \sup_{f \in \mathcal{F}} \sum_{s \in \mathcal{S}} \Big(h\big(2(\theta C_{S,r})^{-1} |G_f(W_s)|\big) - \mathbb{E} h\big(2(\theta C_{S,r})^{-1} |G_f(W_s)|\big) \Big) - \sum_{s \in \mathcal{S}} \mathbb{E} h\big(2(\theta C_{S,r})^{-1} |G_f(W_s)|\big) \ . \quad (4.34)$$

The last term in (4.34) can be bounded from below since for all $f \in \mathcal{F}$ and $s \in \mathcal{S}$,

$$\mathbb{E} h\big(2(\theta C_{S,r})^{-1} |G_f(W_s)|\big) \leqslant \mathbb{P}\Big(|G_f(W_s)| \geq \frac{\theta C_{S,r}}{2} \Big) \leqslant \frac{4 \mathbb{E} G_f(W_s)^2}{(\theta C_{S,r})^2} \leqslant \frac{4S^2}{\theta^2 C_{S,r}^2 N^2} \sum_{i \in B_s} \mathbb{E}[(\ell_f - \ell_{f^*})^2(X_i, Y_i)] \leq \frac{4L^2 S}{\theta^2 C_{S,r}^2 N} \|f - f^*\|_{L_2}^2 \ .$$

The last inequality follows from Assumption 4.7. Since $\|f - f^*\|_{L_2} \leq \sqrt{C_{S,r}}$,

$$\mathbb{E} h\big(2(\theta C_{S,r})^{-1} |G_f(W_s)|\big) \leqslant \frac{4L^2 S}{\theta^2 C_{S,r} N} \ .$$

As $C_{S,r} \geqslant 96 L^2 S/(\theta^2 N)$,

$$\mathbb{E} h\big(2(\theta C_{S,r})^{-1} |G_f(W_s)|\big) \leq \frac{1}{24} \ .$$

Plugging this inequality in (4.34) yields

$$z(f) \geq |\mathcal{S}|(1 - \frac{1}{24}) - \sup_{f \in \mathcal{F}} \sum_{s \in \mathcal{S}} \Big(h\big(2(\theta C_{S,r})^{-1} |G_f(W_s)|\big) - \mathbb{E} h\big(2(\theta C_{S,r})^{-1} |G_f(W_s)|\big) \Big) \ . \quad (4.35)$$

Using the Mc Diarmid's inequality, with probability larger than $1 - \exp(-|\mathcal{S}|/288)$ we get

$$\sup_{f \in \mathcal{F}} \sum_{s \in \mathcal{S}} \Big(h\big(2(\theta C_{S,r})^{-1} |G_f(W_s)|\big) - \mathbb{E} h\big(2(\theta C_{S,r})^{-1} |G_f(W_s)|\big) \Big) \leq \frac{|\mathcal{S}|}{24} + \mathbb{E} \sup_{f \in \mathcal{F}} \sum_{s \in \mathcal{S}} \Big(h\big(2(\theta C_{S,r})^{-1} |G_f(W_s)|\big) - \mathbb{E} h\big(2(\theta C_{S,r})^{-1} |G_f(W_s)|\big) \Big) \ .$$

By the symmetrization lemma, it follows that

$$\sup_{f \in \mathcal{F}} \sum_{s \in \mathcal{S}} \Big(h\big(2(\theta C_{S,r})^{-1} |G_f(W_s)|\big) - \mathbb{E} h\big(2(\theta C_{S,r})^{-1} |G_f(W_s)|\big) \Big) \leqslant \frac{|\mathcal{S}|}{24} + 2 \mathbb{E} \sup_{f \in \mathcal{F}} \sum_{s \in \mathcal{S}} \sigma_k h\big(2(\theta C_{S,r})^{-1} |G_f(W_s)|\big) \ .$$

As ϕ is 1-Lipschitz with $\phi(0) = 0$, the contraction Lemma from (Ledoux and Talagrand, 2013) and yields

$$\begin{aligned}\sup_{f\in\mathcal{F}}\sum_{s\in\mathcal{S}}\Big(h\big(2(\theta C_{S,r})^{-1}|G_f(W_s)|\big)-\mathbb{E}h\big(2(\theta C_{S,r})^{-1}|G_f(W_s)|\big)\Big)&\\ \leqslant \frac{|\mathcal{S}|}{24}+\frac{4}{\theta}\mathbb{E}\sup_{f\in\mathcal{F}}\sum_{s\in\mathcal{S}}\sigma_s\frac{G_f(W_s)}{C_{S,r}}&\\ =\frac{|\mathcal{S}|}{24}+\frac{4}{\theta}\mathbb{E}\sup_{f\in\mathcal{F}}\sum_{s\in\mathcal{S}}\sigma_s\frac{(P_{B_s}-P)(\ell_f-\ell_{f^*})}{C_{S,r}}&\end{aligned}$$

For any $s \in \mathcal{S}$, let $(\sigma_i)_{i\in B_s}$ independent from $(\sigma_s)_{s\in\mathcal{S}}$, $(X_i)_{i\in\mathcal{I}}$ and $(Y_i)_{i\in\mathcal{I}}$. The vectors $(\sigma_i\sigma_s(\ell_f - \ell_{f^*})(X_i, Y_i))_{i,f}$ and $(\sigma_i(\ell_f-\ell_{f^*})(X_i, Y_i))_{i,f}$ have the same distribution. Thus, by the symmetrization and contraction lemmas, with probability larger than $1 - \exp(-|\mathcal{S}|/288)$,

$$\begin{aligned}\sup_{f\in\mathcal{F}}\sum_{s\in\mathcal{S}}\Big(h\big(2C_{S,r}^{-1}|G_f(W_k)|\big)-\mathbb{E}h\big(2C_{S,r}^{-1}|G_f(W_s)|\big)\Big)&\\ \leq\frac{|\mathcal{S}|}{24}+\frac{8}{\theta}\mathbb{E}\sup_{f\in\mathcal{F}}\sum_{s\in\mathcal{S}}\frac{1}{|B_s|}\sum_{i\in B_s}\sigma_i\frac{(\ell_f-\ell_{f^*})(X_i,Y_i)}{C_{S,r}}&\\ =\frac{|\mathcal{S}|}{24}+\frac{8S}{\theta N}\mathbb{E}\sup_{f\in\mathcal{F}}\sum_{i\in\cup_{s\in\mathcal{S}}B_s}\sigma_i\frac{(\ell_f-\ell_{f^*})(X_i,Y_i)}{C_{S,r}}&\\ \leq\frac{|\mathcal{S}|}{24}+\frac{8LS}{\theta N}\mathbb{E}\sup_{f\in\mathcal{F}}\Big|\sum_{i\in\cup_{s\in\mathcal{S}}B_s}\sigma_i\frac{(f-f^*)(X_i)}{C_{S,r}}\Big|\ .&\qquad(4.36)\end{aligned}$$

Now either 1) $S \leq \theta^2\tilde{r}^2N/(96L^2)$ or 2) $S > \theta^2\tilde{r}^2N/(96L^2)$. Assume first that $S \leq \theta^2\tilde{r}^2N/(96L^2)$, so $C_{S,r} = \tilde{r}^2$ and by definition of the complexity parameter

$$\mathbb{E}\sup_{f\in\mathcal{F}}\Big|\sum_{i\in\cup_{s\in\mathcal{S}}B_s}\sigma_i\frac{(f-f^*)(X_i)}{C_{S,r}}\Big|=\mathbb{E}\sup_{f\in\mathcal{F}}\ \frac{1}{\tilde{r}^2}\Big|\sum_{i\in\cup_{s\in\mathcal{S}}B_s}\sigma_i(f-f^*)(X_i)\Big|\leq\frac{\mu|\mathcal{S}|N}{S}\ .$$

If $S > \theta^2\tilde{r}^2N/(96L^2)$, $C_{S,r} = 96L^2S/(\theta^2N)$. Then,

$$\begin{aligned}\mathbb{E}\sup_{f\in\mathcal{F}}\Big|\sum_{i\in\cup_{k\in\mathcal{S}}B_s}\sigma_i\frac{(f-f^*)(X_i)}{C_{S,r}}\Big|&\\ \leq\mathbb{E}\Bigg[\frac{1}{\tilde{r}^2}\sup_{f\in F\cap B^{\phi}_{\rho^*}(f^*)\cap\left(f^*+\tilde{r}B_{L_2}\right)}\Big|\sum_{i\in\cup_{s\in\mathcal{S}}B_s}\sigma_i(f-f^*)(X_i)\Big|&\\ \vee\sup_{f\in F\cap B^{\phi}_{\rho^*}(f^*):\,\tilde{r}\leq\|f-f^*\|_{L_2}\leq\sqrt{96L^2S/(\theta^2N)}}\Big|\sum_{i\in\cup_{s\in\mathcal{S}}B_s}\sigma_i\frac{(f-f^*)(X_i)}{96L^2S/(\theta^2N)}\Big|\Bigg]&\end{aligned}$$

By an homogeneity argument we obtain

$$\begin{aligned}
&\sup_{f\in F\cap B^{\phi}_{\rho^*}(f^*):\, \tilde{r}\le\|f-f^*\|_{L_2}\le\sqrt{96L^2S/(\theta^2N)}} \left| \sum_{i\in\cup_{s\in\mathcal{S}}B_s} \sigma_i \frac{(f-f^*)(X_i)}{96L^2S/(\theta^2N)} \right| \Bigg] \\
&\le \frac{1}{\tilde{r}} \sup_{f\in F\cap B^{\phi}_{\rho^*}(f^*):\, \tilde{r}\le\|f-f^*\|_{L_2}\le\sqrt{96L^2S/(\theta^2N)}} \left| \sum_{i\in\cup_{s\in\mathcal{S}}B_s} \sigma_i \frac{(f-f^*)(X_i)}{\|f-f^*\|} \right| \\
&\le \frac{1}{\tilde{r}^2} \sup_{f\in F\cap B^{\phi}_{\rho^*}(f^*):\, \|f-f^*\|_{L_2}=\tilde{r}} \left| \sum_{i\in\cup_{s\in\mathcal{S}}B_s} \sigma_i (f-f^*)(X_i) \right|
\end{aligned}$$

Finally, in the second case 2) we also have

$$\mathbb{E}\sup_{f\in\mathcal{F}} \left| \sum_{i\in\cup_{s\in\mathcal{S}}B_s} \sigma_i \frac{(f-f^*)(X_i)}{\max(\frac{4L^2S}{\alpha\theta^2N}, \tilde{r}^2)} \right| \le \frac{\mu|\mathcal{S}|N}{S}$$

Plugging this bound in (4.36) yields, with probability larger than $1-e^{-|\mathcal{S}|/288}$

$$\sup_{f\in\mathcal{F}} \sum_{s\in\mathcal{S}} \left(h\big(2C^{-1}_{S,r}|G_f(W_s)|\big) - \mathbb{E}h\big(2C^{-1}_{S,r}|G_f(W_s)|\big) \right) \leqslant |\mathcal{S}| \left(\frac{1}{24} + \frac{8L\mu}{\theta} \right) = \frac{|\mathcal{S}|}{12} \ .$$

Plugging this inequality into (4.35) shows that, with probability at least $1-e^{-|\mathcal{S}|/288}$,

$$z(f) \geqslant \frac{7|\mathcal{S}|}{8} \ .$$

As $S \geqslant 7|\mathcal{O}|/3$, $|\mathcal{S}| \geqslant S - |\mathcal{O}| \geqslant 4S/7$, hence, $z(f) \geqslant S/2$ holds with probability at least $1-e^{-S/504}$.

■

4.6.3 Proof Theorem 4.8

As for the proof of Theorem 4.2 presented in Section 4.6.1 the proof is split into two parts. While we develop another stochastic argument the deterministic part from Proposition 4.1 is exactly the same.

In the example of RKHS, the sub-Gaussian Assumption is not necessary. Instead the tools from bounded class of functions such as the Bousquet's inequality that we recall here can be used.

Theorem 4.12 (Theorem 2.6, (Koltchinskii, 2011a)). *Let $\mathcal{F}$ be a class of functions bounded by M. For all $t>0$, with probability larger than $1-\exp(-t)$*

$$\sup_{f\in\mathcal{F}} |(P_N-P)f| \le \mathbb{E}\sup_{f\in\mathcal{F}} |(P_N-P)f| + \sqrt{2\frac{t}{N}\left(\sup_{f\in\mathcal{F}} Pf^2 + 2M\mathbb{E}\sup_{f\in\mathcal{F}} |(P_N-P)f| \right)} + \frac{tM}{3N} \quad (4.37)$$

Let us define

$$\Omega := \Bigg\{ \forall f\in F : \|f-f^*\|_{L_2} \le \max(1, \sqrt{U\|f^*\|_{\mathcal{H}_K}})\bar{r}(\bar{A}),\ \ \|f-f^*\|^2_{\mathcal{H}_K} \le 4(2+1/\bar{A})\|f^*\|^2_{\mathcal{H}_K}, \\ \big|(P-P_N)\mathcal{L}_f\big| \le \frac{\max(1, U\|f^*\|_{\mathcal{H}_K})\bar{r}^2(\bar{A})}{2\bar{A}} \Bigg\}$$

where we recall that $U = 2L\sqrt{(2+1/\bar{A})\|K\|_\infty}$. By taking $r^* = \max(1, \sqrt{U\|f^*\|_{\mathcal{H}_K}})\bar{r}(\bar{A})$ in the proof of Proposition 4.1 it is clear that the deterministic argument is exactly the same.
Let us show that Ω holds with probability larger than $1 - \exp\big(-(N\bar{r}^2(\bar{A}))/(64(\bar{A}L)^2)\big)$. Let $\mathcal{F} = \{f \in \mathcal{H}_K,\ \|f - f^*\|_{L_2} \leq \max(1, \sqrt{U\|f^*\|_{\mathcal{H}_K}})\bar{r}(\bar{A}),\ \|f - f^*\|^2_{\mathcal{H}_K} \leq \rho^*\}$. From Assumption 4.2 for all $x, y \in \mathcal{X} \times \mathcal{Y}$ and $f \in \mathcal{F}$

$$|(\ell_f - \ell_{f^*})(x,y)| \leq L|f(x) - f^*(x)| \leq \max(1, U\|f^*\|_{\mathcal{H}_K})$$

We can Therefore use Theorem 4.12 with $M = \max(1, U\|f^*\|_{\mathcal{H}_K})$. From the definition of $\mathcal{F}$ it follows that $\sup_{f\in\mathcal{F}} P(\ell_f - \ell_{f^*})^2 \leq L^2 \max(1, U\|f^*\|_{\mathcal{H}_K})\bar{r}^2(\bar{A})$. Let $(\sigma_i)_{i=1}^N$ be i.i.d Rademacher random variables independent from $(X_i, Y_i)_{i=1}$, from the symmetrization and contraction Lemmas (Ledoux and Talagrand, 2013) we get

$$\mathbb{E}\sup_{f\in\mathcal{F}} |(P_N - P)\mathcal{L}_f| \leq 4L\mathbb{E}\sup_{f\in\mathcal{F}}\ \frac{1}{N}\sum_{i=1}^N \sigma_i(f - f^*)(X_i) \leq \max(1, U\|f^*\|_{\mathcal{H}_K})\frac{\bar{r}^2(\bar{A})}{16\bar{A}}$$

where we used the Definition 4.24 of $r(\cdot)$. For any $t > 0$, it follows from Theorem 4.12 that for any function f in $\mathcal{F}$

$$\begin{aligned}|(P_N - P)\mathcal{L}_f| \leq& \max(1, U\|f^*\|_{\mathcal{H}_K})\frac{\bar{r}^2(\bar{A})}{16\bar{A}} + \frac{\max(1, U\|f^*\|_{\mathcal{H}_K})t}{3N} \\ &+ \sqrt{\frac{2t}{N}\max(1, U^2\|f^*\|^2_{\mathcal{H}_K})\bar{r}^2(\bar{A})(L^2 + \frac{1}{8\bar{A}})}\ .\end{aligned}$$

Take $t = N\bar{r}^2(\bar{A})/(64(L\bar{A})^2)$ and use the fact that $\bar{A}, L \geq 1$ conclude the proof.

4.7 Supplementary lemmas

Lemma 4.5. *Let $\gamma > 0$ and f in F such that $\phi(f - f^*) \geq \gamma$. Then, there exist f_0 in F and $1 \leq \alpha \leq \phi(f - f^*)/\gamma$ such that $f = f^* + \alpha(f_0 - f^*)$ and $\phi(f_0 - f^*) = \gamma$*

Proof. Let $\alpha_0 = \sup\{\alpha > 0,\ \phi\big(\alpha(f - f^*)\big) \leq \gamma\}$. For $\alpha = \gamma/\phi(f - f^*) \leq 1$ we have $\phi\big(\alpha(f - f^*)\big) \leq \alpha\phi(f - f^*) = \gamma$ so that $\alpha_0 \geq \gamma/\phi(f - f^*)$. By convexity of F, $f_0 := f^* + \alpha_0(f - f^*) \in F$ and $\alpha_0 \leq 1$ otherwise, by convexity of ϕ we would have $\alpha_0\phi(f - f^*) \leq \phi\big(\alpha_0(f - f^*)\big) \leq \gamma$. Moreover, by maximality of α_0, f_0 is such that $\phi\big(\alpha(f - f^*)\big) = \phi(f_0 - f^*) = \gamma$. The result follows for $\alpha = \alpha_0^{-1}$ ■

Lemma 4.6. *Let $f : \mathbb{R} \mapsto \mathbb{R}$ be a convex function. Then for all $\lambda \geq 1$ and x, y in $\mathbb{R}$:*

$$f(\lambda x + (1-\lambda)y) \geq \lambda f(x) + (1-\lambda)f(y) \tag{4.38}$$

Proof. Let $\lambda \geq 1$, by convexity of f, for all x, y in $\mathbb{R}$:

$$f\left(\frac{1}{\lambda}x + (1 - \frac{1}{\lambda})y\right) \leq \frac{1}{\lambda}f(x) + (1 - \frac{1}{\lambda})f(y)$$

It suffice to take $x = \lambda x + (1-\lambda)y$ to get the result. ■

Chapter 5

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

In this chapter, 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 may contaminate the labels. In that case, we show that the L_2-error rate is bounded by $r_N + L|\mathcal{O}|/N$, where N is the total number of observations and r_N is the L_2-error rate in the non-contaminated setting. When r_N is minimax-rate-optimal in a non-contaminated setting, the rate $r_N + L|\mathcal{O}|/N$ 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. For instance, we present results for Huber's M-estimators (without penalization or regularized by the ℓ_1-norm) and for general regularized learning problems in reproducing kernel Hilbert spaces.

5.1 Introdution

Let $(X_i, Y_i)_{i=1,\cdots,N}$ be random variables taking values in $\mathcal{X} \times \mathbb{R}$, where $\mathcal{X}$ is a measurable space. Given a new input $X \in \mathcal{X}$, one wants to predict its associated label $Y \in \mathbb{R}$. To proceed, we consider (X, Y) as a random variable valued in $\mathcal{X} \times \mathbb{R}$ and given a class of predictors F of functions $f : \mathcal{X} \mapsto \mathbb{R}$, the goal is to predict/approximate the oracle f^* defined as

$$f^* \in \operatorname*{argmin}_{f \in F} \mathbb{E}\left[\ell(f(X), Y)\right] \quad ,$$

where $\ell(f(X), Y)$ measures the error of predicting $f(X)$ while the true label is Y. To estimate/approximate the function f^*, we use the dataset $(X_i, Y_i)_{i=1,\cdots,N}$. Regularized empirical risk minimization is the most widespread strategy in machine learning to estimate f^*. There exists an extensive literature on its generalization capabilities (Vapnik, 1998; Koltchinskii et al., 2006; Koltchinskii, 2011b; Lecué and Mendelson, 2018; Chinot et al., 2019b). However, in the recent years, many papers 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)) may deteriorate the performances of RERM. Consequently, RERM is in general, not robust to outliers. However, what happens if only the labels $(Y_i)_{i=1,\cdots,N}$ are contaminated ? In (Dalalyan and Thompson, 2019); 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, $Y_i = \left\langle X_i, t^* \right\rangle + \epsilon_i$, where X_i is a Gaussian random vector in $\mathbb{R}^p$ with a covariance matrix satisfying the Restricted Eigenvalue condition (Van De Geer et al., 2009) and t^* is s-sparse. For non-contaminated data they suppose that $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$, while it can be anything when malicious outliers contaminate the sample. 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

$$\sigma \sqrt{s \frac{\log(p)}{N}} + \frac{|\mathcal{O}|}{N} \tag{5.1}$$

where $|\mathcal{O}|$ is the number of outliers contaminating the labels. Consequently, they showed that RERM associated with the Huber loss function is minimax-rate-optimal when $|\mathcal{O}|$ malicious outliers corrupt the labels.

5.1.1 Setting

Let $(\Omega, \mathcal{A}, \mathbb{P})$ 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

funtion $f : \mathcal{X} \times \mathcal{Y} \mapsto \mathbb{R}$ we write $Qf = \mathbb{E}_{(X,Y)\sim P}[f(X,Y)]$. 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 $\mathcal{X} \times \mathcal{Y}$. In this paper, we consider a setup where $|\mathcal{O}|$ outputs may 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 $\mathcal{O}$ the set of **outliers**. It is assumed that:

Assumption 5.1. *$(X_i, Y_i)_{i\in\mathcal{I}}$ 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 even be adversial outliers making the learning as hard as possible. The goal is, without knowing the partition $\mathcal{I} \cup \mathcal{O}$, to use the informative data $(X_i, Y_i)_{i\in\mathcal{I}}$ to construct an estimator $\hat{f}$ that approximates/estimates the oracle f^*. A way of measuring the quality of an estimator is via the **error rate** $\|\hat{f} - f\|_{L_2(\mu)}$ or **the excess risk** $P\mathcal{L}_{\hat{f}} := P\ell_{\hat{f}} - P\ell_{f^*}$. We assume the following:

Assumption 5.2. *The class F is convex.*

A natural idea to construct robust estimators when the labels might be contaminated is to consider Lipschitz loss functions (Huber, 1992; Huber and Ronchetti, 2011). Moreover, for computational purposes we will also focus on convex loss functions (van de Geer, 2016).

Assumption 5.3. *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

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

where $\lambda > 0$ is a tuning parameter and $\|\cdot\|$ is a norm. Under Assumptions 5.2 and 5.3 the ERM and RERM are computable using tools from convex optimization.

5.1.2 Our contributions

As exposed in (Dalalyan and Thompson, 2019), in a setting where $|\mathcal{O}|$ outliers contaminate only the labels, RERM with the Huber loss function is minimax-optimal for the sparse-regression problem when the noise and design of non-contaminated data are both Gaussian. It leads to the following question:

1. *Are the RERM optimal for other loss functions and other regresssion problems than the sparse-regression problem when malicious outliers corrupt the labels ?*

Based on previous works (Chinot et al., 2019b; Chinot, 2019b; Chinot et al., 2019a; Alquier et al., 2019), we study ERM and RERM for regression problems when the penalization is a norm and the loss function is simultaneously convex and Lipschitz and show that:

> In a framework where $|\mathcal{O}|$ outliers may contaminate the labels, with weak assumptions on the noise, the excess risk and the square of the error rate for both ERM and RERM can be bounded by
> $$r_N^2 + L^2 \frac{|\mathcal{O}|^2}{N^2} \tag{5.2}$$
> where N is the total number of observations, L is the Lipschitz constant from Assumption 5.3 and r_N is the error rate in a non-contaminated setting.

When the proportion of outliers $|\mathcal{O}|/N$ is smaller than the error rate normalized by the Lipschitz constant r_N/L, 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 5.3.3), non-parametric regression (see Section 5.3.4) and matrix trace regression (using the results from (Alquier et al., 2019)).

The next question one may ask is the following:

2. *Is the general bound* (5.2) *minimax-rate-optimal when* $|\mathcal{O}|$ *malicious outliers may have corrupted the labels ?*

To answer question 2, we use the results from (Chen et al., 2018). The authors established a general minimax theory for the ε-contamination model defined as $\mathbb{P}_{(\varepsilon,\theta,Q)} = (1-\varepsilon)P_\theta + \varepsilon Q$ 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 $Y = f_\theta(X) + \epsilon$, $\theta \in \Theta$, in Section 5.6, we show that the lower minimax bounds for regression problems in the ε-contamination model are the same when

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

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

When the bound (5.2) 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.

In particular, we recover and generalize the results from (Dalalyan and Thompson, 2019) when the noise of non-contaminated data is not necessarily Gaussian but may be heavy-tailed.

The results are derived under the local Bernstein condition introduced in (Chinot et al., 2019b). This condition enables to obtain fast rates of convergence 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 may be heavy-tailed. In these cases, the error rates are respectively $\sqrt{Tr(\Sigma)/N} + |\mathcal{O}|/N$ and $\sqrt{s\log(p)/N} + |\mathcal{O}|/N$, where Σ is the covariance matrix of the design X. We also study learning problems in general reproducing Kernel Hilbert Space (RKHS). We derive error rates depending on the spectrum of the integral operator as in (Smale and Zhou, 2007; Mendelson et al., 2010; Caponnetto and De Vito, 2007) without assumption on the design and when the noise has heavy tails (see section 5.3.3).

5.1.3 Related Litterature

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 (Huber, 1992; Huber and Ronchetti, 2011; Huber et al., 1967) , John Tukey (Tukey, 1960, 1962) and Frank Hampel (Hampel, 1971, 1974). The most classical alternatives to least-squares estimators are M-estimators which consist in replacing the quadratic loss function by another one, less sensitive to outliers (Maronna, 1976; Yohai and Maronna, 1979).

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 robust statistics, it is still a very active research area. Surprisingly, optimal bounds for heavy-tailed data have been obtained only recently (Lugosi et al., 2019b). The estimator in (Lugosi et al., 2019b) cannot be computed in practice. Using SDP, (Hopkins, 2018) obtained optimal bounds achievable in polynomial time. In recent works, still using SDP, (Lecué and Depersin, 2019) designed an algorithm computable in nearly linear time, while (Lei et al., 2019) developed the first tractable optimal algorithm not based on the SDP. In the meantime, another recent trend in robust statistics is 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 of the estimation error measured in Euclidean norm scales as $\sqrt{p/N} + |\mathcal{O}|/N$. In (Chen et al., 2018), the authors developed a general analysis for the ε-contamination model. In (Chen et al., 2016), the same authors proposed an optimal estimator when $|\mathcal{O}|$ outliers with the same distribution contaminate the data. In (Diakonikolas et al., 2019b), 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 (Cheng et al., 2019; Diakonikolas et al., 2019a; Liu et al., 2018; Bhatia et al., 2015). 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 (Dalalyan and Thompson, 2019), 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 and more general.

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 (Lecué and Lerasle, 2019; Lecué et al., 2018; Chinot et al., 2019b). 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 $\sqrt{s\log(p)/N} + L\sqrt{|\mathcal{O}|/N}$ (see (Chinot et al., 2019b)) while the optimal dependence with respect to the number of outliers is $\sqrt{s\log(p)/N} + L|\mathcal{O}|/N$. 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 (Minsker et al., 2015), (Chen et al., 2017) designed a robust gradient descent scheme. More recently, (Nazin et al., 2019) showed that a simple truncation of the gradient enhances the robutness of the stochastic mirror descent algorithm.

The paper is organized as follows. In Section 5.2, 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 5.3 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 5.5 presents simple simulations to illustrate our theoritical findings. In section 5.6, 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 5.7 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 5.8.

Notations All along the paper, for any f in F, $\|f\|_{L_2}$ will be written instead of $\|f\|_{L_2(\mu)}$ where $\|f\|^2_{L_2(\mu)} = \int f^2 d\mu$. The letter c will denote an absolute constant. 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 $H+f^* = \{h+f^* \text{ s.t } h \in H\}$, $H-f^* = \{h-f^* \text{ s.t } h \in H\}$.

5.2 Non-regularized procedures

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

$$\hat{f}_N = \arg\min_{f \in F} \frac{1}{N} \sum_{i=1}^{N} \ell(f(X_i), Y_i) \ . \tag{5.3}$$

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 locallly bounded. We derive fast rates of convergence under very weak assumptions.

5.2.1 General results in the sub-Gaussian framework

The ERM performs well when the empirical excess risk $f \mapsto P_N\mathcal{L}_f$ uniformly concentrates around its expectation $f \mapsto P\mathcal{L}_f$. Thus, it is necessary to impose a strong concentration assumption on the class $\{\mathcal{L}_f(X, Y), f \in F\}$. From assumption 5.3 it is implied by a concentration assumption on the class $\{(f - f^*)(X), f \in F\}$.

Assumption 5.4. *The class $F - f^*$ is B sub-Gaussian i.e for all $f \in F$ and all $\lambda > 0$*

$$\mathbb{E} \exp(\lambda (f - f^*)(X) / \|f - f^*\|_{L_2}) \leq \exp(\lambda^2 B^2 / 2) \ .$$

See (Lecué and Mendelson, 2013) for many examples of sub-Gaussian classes. In this context, we use the Gaussian mean-width as a measure of the complexity of the class function F that we introduce here

Definition 5.1. *Let $H \subset L_2(\mu)$. Let $(G_h)_{h \in H}$ be the canonical centered Gaussian process indexed by H (in particular, the covariance structure of $(G_h)_{h \in H}$ is given by $(\mathbb{E}(G_{h_1} - G_{h_2})^2)^{1/2} = (\mathbb{E}(h_1(X) - h_2(X))^2)^{1/2}$ for all $h_1, h_2 \in H$). The **Gaussian mean-width** of H is $w(H) = \mathbb{E} \sup_{h \in H} G_h$.*

For example, when $F = \{\langle t, \cdot \rangle, t \in T\}$ and the covariance matrix of X is Σ, we have $w(F) = \mathbb{E} \sup_{t \in T} \langle t, \mathbf{G} \rangle$, where $\mathbf{G} \sim \mathcal{N}(0, \Sigma)$. Similarly to (Lecué and Mendelson, 2018; Chinot et al., 2019b,a; Alquier et al., 2019), the error rate and the excess risk are driven by fixed point solutions of a Gaussian mean-width:

Definition 5.2. *Let B_{L_2} denote the unit ball induced by $L_2(\mu)$. The* ***complexity parameter*** *$r_{\mathcal{I}}(\cdot)$ is defined as*

$$r_{\mathcal{I}}(A) = \inf\{r > 0 : ALB(L+1)w(F \cap (f^* + rB_{L_2})) \leq cr^2\sqrt{|\mathcal{I}|}\}$$

where $c > 0$ denotes an absolute constant, L is the Lipschitz constant from assumption 5.3 and B is the sub-Gaussian constant from assumption 5.4.

To obtain fast rates of convergence it is necessary to impose assumptions on the distribution P. For instance, the margin assumptions (Mammen and Tsybakov, 1999; Tsybakov, 2004; van de Geer, 2016) and the Bernstein conditions from (Bartlett and Mendelson, 2006a) have been widely used in statistics and learning theory to prove fast convergence rates for the ERM. In the spirit of (Chinot et al., 2019b) we introduce a weaker **local Bernstein assumption**.

Assumption 5.5. *Let $r(\cdot)$ be a complexity parameter s.t for all $A > 0$, $r(A) \geq r_{\mathcal{I}}(A)$. There exists a constant $A > 0$ such that for all $f \in F$ if $\|f - f^*\|_{L_2} = r(A)$ we have $\|f - f^*\|_{L_2}^2 \leq AP\mathcal{L}_f$.*

Note that assumption 5.5 holds locally around the oracle f^*. The smallest radius corresponds to $r_{\mathcal{I}}(A)$. The bigger $r(\cdot)$ the stronger assumption 5.5 is. Assumption 5.5 has been extensively studied in (Chinot et al., 2019b,a) 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 5.1. *Let $\mathcal{I} \cup \mathcal{O}$ be a partition of $\{1, \cdots, N\}$ where $|\mathcal{O}| \leq |\mathcal{I}|$. Let $r(\cdot)$ be a complexity parameter such that for all $A > 0$, $r(A) \geq r_{\mathcal{I}}(A)$. Grant Assumptions 5.1, 5.3 with $L \geq 1$, 5.2, 5.4 and 5.5 with $r(\cdot)$ for $A \geq 1$. As long as $|\mathcal{O}| < |\mathcal{I}|r(A)/(2AL)$, with probability larger that $1 - 2\exp\big(-c|\mathcal{I}|r^2(A)/(ALB(1+L))\big)$, the estimator $\hat{f}_N$ defined in Equation (5.3) satisfies*

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

The partition $\mathcal{I} \cup \mathcal{O}$ is unknown: no one knows which observations are outliers. In Theorem 5.1, we can always take $r(A) = \max(r_{\mathcal{I}}(A), 2AL|\mathcal{O}|/|\mathcal{I}|)$. With such a choice of complexity parameter, we necessarily have $|\mathcal{O}| < (|\mathcal{I}|r(A))/(2AL)$ and with probability larger that

$$1 - 2\exp\left(-\frac{c}{ALB(L+1)}\max\left(|\mathcal{I}|r_{\mathcal{I}}^2(A), \frac{|\mathcal{O}|^2}{|\mathcal{I}|}\right)\right)$$

the estimator $\hat{f}_N$ defined in Equation (5.3) satisfies

$$\|\hat{f}_N - f^*\|_{L_2} \leq cAL\left(r_{\mathcal{I}}(A) + \frac{|\mathcal{O}|}{N}\right) .$$

Theorem 5.1 holds if the local Bernstein condition 5.5 is satisfied for all functions f in F such that $\|f - f^*\|_{L_2} = cAL(r_{\mathcal{I}}(A) + |\mathcal{O}|/N)$, that is on an L_2-sphere with a radius equal to the rate of convergence. The bound on the error rate can be decomposed as the sum of the error rate in the

non-contaminated setting and the proportion of outliers $|\mathcal{O}|/N$. As long as the proportion of outliers is smaller than the error rate in the non-contaminated setting, the error rate remains constant. On the other hand, when the proportion of outliers exceeds the error rate in the non-contaminated setting, the error rate in the contaminated setting becomes linear with respect to the proportion of outliers. When $r_{\mathcal{I}}$ is minimax optimal in a non-contaminated setting, we obtain that the ERM is minimax optimal when less that $Nr_{\mathcal{I}}$ outliers contaminate the labels. In Section 5.2.3, we show that this dependence with respect to the number of outliers is minimax optimal for linear regression in $\mathbb{R}^p$.

5.2.2 General results in the bounded framework

In Section 5.2.1 we considered sub-Gaussian class of functions to derive fast rates of convergence. In this section, we derive a general result when the **localized** class $F - f^*$ is bounded (localized around the oracle f^* with respect to the $L_2(\mu)$-norm, see Assumption 5.6). Since the Gaussian mean-width no longer appears naturally, it is necessary to define a new measure of the complexity of the class F. A way to measure the complexity a class of functions F is via **Rademacher complexities** (Koltchinskii et al., 2006; Koltchinskii, 2011b).

Definition 5.3. *The **complexity parameter** in the bounded setting $r^b_{\mathcal{I}}(\cdot)$ is defined as*

$$r^b_{\mathcal{I}}(A) = \inf\left\{ r > 0 : \mathbb{E} \sup_{f \in F \cap (f^* + rB_{L_2})} \sum_{i \in \mathcal{I}} \sigma_i \big(f - f^*\big)(X_i) \leq \frac{|\mathcal{I}| r^2}{32 A (L+1) L} \right\}$$

where $(\sigma_i)_{i\in\mathcal{I}}$ are i.i.d Rademacher random variables independent to $(X_i)_{i\in\mathcal{I}}$, L is the Lipschitz constant from assumption 5.3 and B_{L_2} denote the unit ball with respect to $L_2(\mu)$.

To obtain fast rates, we need to adapt the local Bernstein condition to this new complexity parameter and introduce the local boundedness assumption

Assumption 5.6. *Let $r^b(\cdot)$ be a complexity parameter such that for every $A > 0$, $r^b(A) \geq r^b_{\mathcal{I}}(A)$. There exist constants $A \geq 1$, $M > 0$ such that for all $f \in F$ if $\|f - f^*\|_{L_2} = \max(1, \sqrt{LM}) r^b(A)$ we have*

$$\|f - f^*\|^2_{L_2} \leq AP\mathcal{L}_f \quad \textit{and} \quad \forall x \in \mathcal{X}, |(f - f^*)(x)| \leq M \tag{5.4}$$

The second part of Equation (5.4) requires L_∞-boundedness only in the L_2-neighborhood around the oracle f^* where the radius is proportional to the rate of convergence $r^b(A)$. For example, let us consider the case when $F = \{\langle t, \cdot \rangle, t \in \mathbb{R}^p\}$ and X is isotropic (i.e $\mathbb{E}\langle X, t\rangle^2 = \|t\|_2^2$ for all $t \in \mathbb{R}^p$). Let $f(\cdot) = \langle t, \cdot \rangle$ be such that $\|f - f^*\|_{L_2} = \|t - t^*\|_2 \leq \max(1, \sqrt{LM}) r^b(A)$ and $|(f - f^*)(x)| = |\langle t - t^*, x\rangle| \leq \|t - t^*\|_2 \|x\|_2 \leq \|x\|_2 \max(1, \sqrt{LM}) r^b(A)$. Without loss of generality we can assume that $M \geq 1$ and the condition becomes, there exists $M \geq 1$ such that for all x in $\mathcal{X} \subset \mathbb{R}^p$, $\|x\|_2^2 \leq M/(L(r^b(A)^2)$. Simple computations (see (Koltchinskii et al., 2006)) show that when $r^b(A) = r^b_{\mathcal{I}}(A)$, the complexity parameter $r^b(A)$ is of the order $\sqrt{p/|\mathcal{I}|}$ and the condition

become $\|x\|_2^2 \leq (M|\mathcal{I}|)/(pL)$. The more informative data we have, the larger the euclidean radius of $\mathcal{X}$ can be.
Assumption 5.6 is local around the oracle f^*. The smallest radius corresponds to $\max(1, \sqrt{LM})r_{\mathcal{I}}^b(A)$. The bigger $r^b(\cdot)$ the stronger assumption 5.5 is. We are now in position to state the main theorem for the ERM in the bounded setting.

Theorem 5.2. *Let $\mathcal{I} \cup \mathcal{O}$ be a partition of $\{1, \cdots, N\}$ where $|\mathcal{O}| \leq |\mathcal{I}|$. Let $r^b(\cdot)$ be a complexity parameter such that for all $A > 0$, $r^b(A) \geq r_{\mathcal{I}}^b(A)$. Grant Assumptions 5.1, 5.3 with $L \geq 1$, 5.2 and 5.6 with $r^b(\cdot)$ for $A \geq 1$ and $M > 0$. As long as $|\mathcal{O}| < (|\mathcal{I}|r(A))/(2AL)$, with probability larger than $1 - 2\exp\big(-c|\mathcal{I}|r^2(A)/(L+1)^2A^2)\big)$, the estimator $\hat{f}_N$ defined in Equation (5.3) satisfies*

$$\|\hat{f}_N - f^*\|_{L_2(\mu)} \leq \max(1, \sqrt{LM})r^b(A) \quad \textit{and} \quad P\mathcal{L}_{\hat{f}_N} \leq \max(1, LM)\frac{r^2(A)}{A}$$

In Theorem 5.2 we can always take $r^b(A) = \max(r_{\mathcal{I}}^b(A), 2AL|\mathcal{O}|/|\mathcal{I}|)$. With such a choice of $r^b(\cdot)$ we necessarily have $|\mathcal{O}| < (|\mathcal{I}|r^b(A))/(2AL)$ and with probability larger that

$$1 - 2\exp\left(-\frac{c}{A^2(L+1)^2}\max\left(|\mathcal{I}|(r_{\mathcal{I}}^b(A))^2, \frac{|\mathcal{O}|^2}{|\mathcal{I}|}\right)\right)$$

the estimator $\hat{f}_N$ defined in Equation (5.3) satisfies

$$\|\hat{f}_N - f^*\|_{L_2(\mu)} \leq cAL\max(1, \sqrt{LM})\left(r_{\mathcal{I}}^b(A) + \frac{|\mathcal{O}|}{N}\right) \ .$$

As in the sub-Gaussian setting there is a tradeoff between confidence and accuracy. When the number of outliers is smaller than $Nr_{\mathcal{I}}^b(A)$, confidence and accuracy are constant. When $|\mathcal{O}|$ becomes larger than the threshold $Nr_{\mathcal{I}}^b(A)$ the confidence is improved while the accuracy is deteriorated. The conclusion is the same as in the bounded case. The error rate in the contaminated setting is the maximum between the error rate in the non-contaminated setting and the proportion of outliers.

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

To put into perspective the results obtained in Sections 5.2.1, we apply Theorem 5.1 for linear regression in $\mathbb{R}^p$. For the sake of brevity we do no present the result for Theorem 5.2. In the vocabulary of Section 5.1, the class F of predictors is defined as $F = \{\langle t, \cdot \rangle, t \in \mathbb{R}^p\}$ which satisfies assumption 5.2. Let $(X_i, Y_i)_{i=1}^N$ be random variables defined by the following linear model:

$$Y_i = \langle X_i, t^* \rangle + \epsilon_i \ , \tag{5.5}$$

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 centered and independent to X_i. For the moment, nothing more

is assumed for $(\epsilon_i)_{i\in\mathcal{I}}$. It is clear that assumption 5.1 holds. The Empirical Risk Minimizer with the Huber loss function is defined as

$$\hat{t}_N^\delta = \underset{t\in\mathbb{R}^p}{\operatorname{argmin}} \frac{1}{N}\sum_{i=1}^N \ell^\delta(\langle X_i, t\rangle, Y_i) \tag{5.6}$$

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

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

which satisfies assumption 5.3 for $L = \delta$. All along this section, δ will be considered as a **constant** (i.e independent to the sample size N and the dimension p). Let $t, v \in \mathbb{R}^p$ such that $f(\cdot) = \langle t, \cdot\rangle$ and $g(\cdot) = \langle v, \cdot\rangle$. Since $\mu = N(0,\Sigma)$, we have $\|f-g\|_{L_2}^2 = \mathbb{E}\langle t-v, X_1\rangle^2 = (t-v)^T\Sigma(t-v)$ and $\lambda(f(X_1)-g(X_1))/\|f-g\|_{L_2} = \big(\lambda/(t-v)^T\Sigma(t-v)\big)(t-v)^T X_1 \sim \mathcal{N}(0,\lambda^2)$ and assumption 5.4 holds with $B = 1$. To apply Theorem 5.1, it remains to study the local Bernstein assumption for the Huber loss function. We recall the result from (Chinot et al., 2019b). Let us introduce the following assumption.

Assumption 5.7. *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 $|z - f^*(x)| \le (\sqrt{2}(C'))^{(2+\varepsilon)/\varepsilon} r$, $F_{Y|X=x}(z+\delta) - F_{Y|X=x}(z-\delta) \geqslant \alpha$.

Proposition 5.1 ((Chinot et al., 2019b),Theorem 7)**.** *Grant assumption 5.7. The Huber loss function with parameter $\delta > 0$ satisfies the Bernstein condition for $A = 4/\alpha$: for all $f \in F$, if $\|f-f^*\|_{L_2} = r$ then $(4/\alpha)P\mathcal{L}_f \ge \|f-f^*\|_{L_2}^2$.*

Since $\mu = \mathcal{N}(0,\Sigma)$, the point a) holds with $C' = 3$. Moreover, from the model (5.5), the point b) can be rewritten as: for all $x \in \mathbb{R}^p$, for all $z \in \mathbb{R}$ such that $|z - \langle x, t^*\rangle| \le 18r$,

$$\mathbb{P}\Big(z - \delta \le \langle x, t^*\rangle + \epsilon \le z + \epsilon\Big) = F_\epsilon(z+\delta-\langle x,t^*\rangle) - F_\epsilon(z-\delta-\langle x,t^*\rangle) \ge \alpha$$

which is satisfied if

$$F_\epsilon(\delta - 18r) - F_\epsilon(18r - \delta) \ge \alpha \tag{5.7}$$

where F_ϵ denotes the cumulative distribution of ϵ distributed as ϵ_i for any $i \in \mathcal{I}$. The sufficient condition (5.7) implies that the noise puts enough mass around zero. To finish, we need to compute complexity parameter $r_{\mathcal{I}}(4/\alpha)$. For an absolute constant $c > 0$, well-known computations (see (Talagrand, 2014)) give:

$$w(F \cap (f^* + rB_{L_2(\mu)})) \le r\sqrt{Tr(\Sigma)} \quad \text{and} \quad r_{\mathcal{I}}(4/\alpha) = c\frac{\delta(1+\delta)}{\alpha}\sqrt{\frac{Tr(\Sigma)}{N}}$$

where we used the fact that $|\mathcal{I}| \geq N/2$ and $L = \delta$.

We are now in position to apply Theorem 5.1 for Huber's M-estimator in $\mathbb{R}^p$ with $r(4/\alpha) = c\frac{\delta}{\alpha}\max\left((1+\delta)\sqrt{Tr(\Sigma)/N}, |\mathcal{O}|/N\right)$.

Theorem 5.3. *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 $i \in \{1, \cdots, N\}$*

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

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

$$F_\epsilon\left(\delta - c\frac{\delta}{\alpha}\max\left((1+\delta)\sqrt{\frac{Tr(\Sigma)}{N}}, \frac{|\mathcal{O}|}{N}\right)\right) - F_\epsilon\left(c\frac{\delta}{\alpha}\max\left((1+\delta)\sqrt{\frac{Tr(\Sigma)}{N}}, \frac{|\mathcal{O}|}{N}\right) - \delta\right) \geq \alpha \quad (5.8)$$

where F_ϵ denotes the cdf of ϵ distributed as ϵ_i for i in $\mathcal{I}$, δ is the hyperparameter of the Huber loss function. Nothing is assumed on $(\epsilon_i)_{i\in\mathcal{O}}$. Then with probability larger than

$$1 - 2\exp\left(-c\frac{\delta}{\alpha(1+\delta)}\max\left((1+\delta)^2 Tr(\Sigma), \frac{|\mathcal{O}|^2}{N}\right)\right) \ , \quad (5.9)$$

the estimator $\hat{t}_N^\delta$ defined in Equation (5.6) satisfies

$$\|\Sigma^{1/2}(\hat{t}_N^\delta - t^*)\|_2 \leq c\frac{\delta(1+\delta)}{\alpha}\max\left(\sqrt{\frac{Tr(\Sigma)}{N}}, \frac{|\mathcal{O}|}{N}\right)$$
$$\text{and} \quad P\mathcal{L}_{\hat{t}_N^\delta} \leq c\frac{\delta^2(1+\delta)^2}{\alpha}\max\left(\frac{Tr(\Sigma)}{N}, \frac{|\mathcal{O}|^2}{N^2}\right)$$

In Theorem 5.3 there is no assumption on $|\mathcal{O}|$ as long as $|\mathcal{O}| \leq |\mathcal{I}|$. There are two situations: 1) the number of outliers $|\mathcal{O}|$ is smaller than $\sqrt{Tr(\Sigma)N}$. We obtain the optimal rate of convergence $\sqrt{Tr(\Sigma)/N}$ for linear regression in $\mathbb{R}^p$ with an exponentially large probability, 2) the number of outliers exceeds $\sqrt{Tr(\Sigma)N}$. In this case, the error rate and the excess risk are deteriorated but the confidence is improved. According to (Chen et al., 2018), this rate is minimax optimal in the ε-contamination model for $\varepsilon = |\mathcal{O}|/N$. It follows that Theorem 5.3 is **minimax-optimal** for the problem of linear regression in $\mathbb{R}^p$ when malicious outliers contaminate the labels (Chen et al., 2018).
In Section 5.5, we run simple simulations to illustrate the linear dependence between the error rate and the proportion of outliers.

Theorem 5.3 handles many different distributions for the noise as long as Equation (5.8) is satisfied. It is not necessary to impose that the noise is sub-Gaussian neither integrable. For instance, when $\epsilon \sim C(1)$ is a standard Cauchy distribution, for all $t \in \mathbb{R}$, we have $F_\epsilon(t) = 1/2 + \arctan(t)/\pi$. With straightforward computations, Equation (5.7) can be rewritten as

$$18r \leq \delta - \tan(\frac{\pi}{2}\alpha) \quad (5.10)$$

From Equation (5.10), Equation (5.8) is satisfied if

$$c\frac{\delta}{\alpha}\max\left((1+\delta)\sqrt{\frac{Tr(\Sigma)}{N}},\frac{|\mathcal{O}|}{N}\right)\leq\delta-\tan(\frac{\pi}{2}\alpha)$$

Let us fix $\delta>0$ to be a quantity independent of the dimension p and the number of observations N. Take $\alpha=2\arctan(\delta/2)/\pi$. When $\sqrt{N}\geq c\sqrt{p}(1+\delta)/\alpha$ and $|\mathcal{O}|\leq c\alpha N$ the condition defined in Equation (5.8) holds and the local Bernstein condition 5.5 is verified for $A=4/\alpha$. We get the following corollary.

Corollary 5.1. *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 variables with $X_1\sim\mathcal{N}(0,\Sigma)$ and for all $i\in\{1,\cdots,N\}$*

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

where $(\epsilon_i)_{i\in I}$ are i.i.d standard Cauchy random variables independent to $(X_i)_{i\in\mathcal{I}}$. Consider the Huber loss function with a parameter $\delta>0$. Assume that $\sqrt{N}\geq c\sqrt{Tr(\Sigma)}(1+\delta)/\arctan(\delta/2)$ and $|\mathcal{O}|\leq c\arctan(\delta/2)N$. Then with probability larger than

$$1-2\exp\left(-c\frac{\delta}{(1+\delta)\arctan(\delta/2)}\max\left((1+\delta)^2Tr(\Sigma),\frac{|\mathcal{O}|^2}{N}\right)\right)\ , \tag{5.11}$$

the estimator $\hat{t}_N^\delta$ defined in Equation (5.6) satisfies

$$\|\Sigma^{1/2}(\hat{t}_N^\delta-t^*)\|_2\leq c\frac{\delta(1+\delta)}{\arctan(\delta/2)}\max\left(\sqrt{\frac{Tr(\Sigma)}{N}},\frac{|\mathcal{O}|}{N}\right)$$
$$\text{and}\quad P\mathcal{L}_{\hat{t}_N^\delta}\leq c\frac{\delta^2(1+\delta)^2}{\arctan(\delta/2)}\max\left(\frac{Tr(\Sigma)}{N},\frac{|\mathcal{O}|^2}{N^2}\right)\ .$$

5.3 High dimensional setting

In Section 5.2 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. The regularization term highlights the belief the statistician may have on the oracle f^*. More formally, let $F\subset E\subset L_2(\mu)$ 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

$$\hat{f}_N^\lambda=\operatorname*{argmin}_{f\in F}\frac{1}{N}\sum_{i=1}^N\ell(f(X_i),Y_i)+\lambda\|f\| \tag{5.12}$$

For high dimensional statistics, it is possible to impose a low dimensional structure. For instance, the use of the ℓ_1 norm promotes sparsity (Tibshirani, 1996) for regression and classification problems

in $\mathbb{R}^p$ while the 1-Schatten norm promotes low rank solutions for matrix reconstructions. Up to some technicalities the main result for the RERM is the same as the one in Section 5.2: the excess risk and the square of the error rate will be of the order

$$r_N^2 + \frac{|\mathcal{O}|^2}{N^2}$$

where r_N denote the (sparse or low-dimensional) error rate in the non-contaminated setting. As long as the proportion of outliers is smaller than the error rate the RERM behaves as if there was no contamination.

5.3.1 General result in the sub-Gaussian framework

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

Definition 5.4. *Let B be the unit ball induced by the regularization norm $\|\cdot\|$. The **complexity parameter** $\tilde{r}_{\mathcal{I}}(\cdot,\cdot)$ is defined as*

$$\tilde{r}_{\mathcal{I}}(A,\rho) = \inf\{r>0 : cALB(L+1)w(F\cap(f^*+rB_{L_2(\mu)}\cap\rho B)) \le r^2\sqrt{|\mathcal{I}|}\}$$

where $c>0$ denotes an absolute constant, L is the Lipschitz constant from assumption 5.3 and B, the sub-Gaussian constant from assumption 5.4.

The main difference between $r_{\mathcal{I}}(A)$ from Definition 5.2 and $\tilde{r}_{\mathcal{I}}(A,\rho)$ is that $\tilde{r}_{\mathcal{I}}(A,\rho)$ measures the local complexity of $F\cap(f^*+\rho B)$ whereas $r_{\mathcal{I}}(A)$ measures the local complexity of the entire set F around f^*. The regularization shifts the estimator towards a neighborhood of the oracle f^* with respect to the regularization norm.
To deal with the regularization part, we use the tools from (Lecué and Mendelson, 2018). The idea is the following: the ℓ_1 norm induces sparsity properties because it has large subdifferentials at sparse vectors. Therefore to obtain "sparsity depedendent bounds", i.e bounds depending on the unknown sparsity of the oracle f^*, a natural tool is to look at the size of the subdifferential of $\|\cdot\|$ in f^* where we recall that the subdifferential of $\|\cdot\|$ in f is defined as

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

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

$$(\partial\,\|\cdot\|)_f = \begin{cases} \{z^*\in S^* : z^*(f)=\|f\|\} & \text{if } f\neq 0 \\ B^* & \text{if } f=0 \end{cases} \tag{5.13}$$

where B^* is the unit ball of the dual norm associated with $\|\cdot\|$, i.e. $z^*\in E^*\to\|z^*\|^* = \sup_{\|f\|\le 1} z^*(f)$ and S^* is its unit sphere. In other words, when $f\neq 0$, the subdifferential of $\|\cdot\|$ in f is the set of

all vectors z^* in the unit dual sphere S^* which are norming for f. For any $\rho > 0$, let

$$\Gamma_{f^*}(\rho) = \bigcup_{f \in F : \|f - f^*\| \leq \rho/20} (\partial \|\cdot\|)_f \ .$$

Instead of looking at the subdifferential of $\|\cdot\|$ exactly in f^* we consider 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 (Lecué and Mendelson, 2018).

Definition 5.5. *Let $\tilde{r}(\cdot, \cdot)$ such that for any $A > 0$ and $\rho > 0$, $\tilde{r}(A, \rho) \geq r_{\mathcal{I}}(A, \rho)$. For any $A, \rho > 0$, set*

$$H_{\rho, A, \tilde{r}} = \{f \in F \ : \ \|f^* - f\| = \rho \ \text{ and } \ \|f^* - f\|_{L_2} \leq \tilde{r}(A, \rho)\} \ ,$$

and define

$$\Delta(\rho, A, \tilde{r}) = \inf_{h \in H_{\rho, A, \tilde{r}}} \sup_{z^* \in \Gamma_{f^*}(\rho)} z^*(h - f^*) \ . \tag{5.14}$$

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

The constant $4/5$ in Definition 5.5 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" by taking ρ^* function of the unknown sparsity (see Section 5.3.3 for a more explicit example or (Chinot et al., 2019a; Lecué and Mendelson, 2018) for many other illustrations).

Remark 5.1. *It can also induce "norm dependent bounds", i.e bounds depending on the norm of the oracle $\|f^*\|$. By taking $\rho^* = 20\|f^*\|$, we get that $0 \in \{f \in F : \|f - f^*\| \leq \rho^*/20\}$ and from Equation (5.13) it follows that $\Gamma_{f^*}(20\|f^*\|) = B^*$ and $\Delta(20\|f^*\|, A, \tilde{r}) = \rho^*$. In other words, the sparsity equation is always satisified for $\rho^* = 20\|f^*\|$ (see Section 5.3.4 for examples)*

Finally, we adapt the local Bernstein assumption to this new framework.

Assumption 5.8. *Let $\tilde{r}(\cdot, \cdot)$ be such that for all $A, \rho > 0$, $\tilde{r}(A, \rho) \geq \tilde{r}_{\mathcal{I}}(A, \rho)$. There exist $A > 0$ and ρ^* satisfying the $A, \tilde{r}$-sparsity equation from Definition 5.5 such that for all $f \in F : \|f - f^*\|_{L_2(\mu)} = \tilde{r}(A, \rho^*)$ and $\|f - f^*\| \leq \rho^*$ we have $\|f - f^*\|_{L_2(\mu)}^2 \leq AP\mathcal{L}_f$.*

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

Theorem 5.4. *Let $\mathcal{I} \cup \mathcal{O}$ denote a partition of $\{1, \cdots, N\}$ such that $|\mathcal{O}| \leq |\mathcal{I}|$. Let $\tilde{r}(\cdot, \cdot)$ be such that for all $A, \rho > 0$, $\tilde{r}(A, \rho) \geq \tilde{r}_{\mathcal{I}}(A, \rho)$. Grant Assumptions 5.1, 5.3, 5.2, 5.4. Suppose that assumption 5.8 holds with $\rho = \rho^*$ satisfying the $A, \tilde{r}$-sparsity equation from Definition 5.5. Set:*

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

As long as $|\mathcal{O}| < c|\mathcal{I}|\tilde{r}(A,\rho^*)/(AL)$*, with probability larger that*

$$1 - 2\exp\bigg(- c\frac{|\mathcal{I}|\tilde{r}^2(A,\rho^*)}{ABL(L+1)}\bigg),$$

the estimator $\hat{f}_N^\lambda$ *defined in Equation* (5.12) *satisfies*

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

By taking $\tilde{r}(A,\rho^*) = c\max(\tilde{r}_{\mathcal{I}}(A,\rho^*), AL|\mathcal{O}|/|\mathcal{I}|)$, the condition $|\mathcal{O}| < c|\mathcal{I}|\tilde{r}(A,\rho^*)/(AL)$ is necessarily satisfied and, with exponentially large probability, we get

$$\|\hat{f}_N^\lambda - f^*\|_{L_2(\mu)} \leq cAL\bigg(\tilde{r}_{\mathcal{I}}(A,\rho^*) + \frac{|\mathcal{O}|}{N}\bigg) \ .$$

The error rate can be decomposed as the sum of the error rate in the non-contaminated setting and the proportion of outliers $|\mathcal{O}|/N$. Theorem 5.4 is a "meta" theorem in the sense that it can used for many practical problems. We use Theorem 5.4 for ℓ_1-penalized Huber's M-estimator in Section 5.3.3. It is also possible to use Theorem 5.4 for many other convex and Lipschitz loss functions and regularization norms as it is done in (Chinot et al., 2019a). It can also be used for matrix reconstruction problems by penalizing with the 1-Schatten norm (Lecué and Mendelson, 2018).

General routine to apply Theorem 5.4 This small paragraph explains how in practice we can use Theorem 5.4.

1. Verify assumptions 5.1, 5.3, 5.2, 5.4.

2. Compute the localized Gaussian mean width $w\big(F\cap(f^*+rB_{L_2}\cap\rho B)\big)$ for any $r,\rho>0$. Deduce the value of $\tilde{r}_{\mathcal{I}}(A,\rho)$ for any $A,\rho>0$.

3. Choose a new complexity parameter such that for every $A,\rho>0$, $\tilde{r}(A,\rho)\geq\tilde{r}_{\mathcal{I}}(A,\rho)$. For instance, to derive results in the contaminated setting we will take $\tilde{r}(A,\rho) = c\max(\tilde{r}_{\mathcal{I}}(A,\rho), AL|\mathcal{O}|/N)$. From the computation of $\tilde{r}_{\mathcal{I}}(A,\rho)$ deduce the closed form of $\tilde{r}(A,\rho)$.

4. For a fixed constant $A>0$, find $\rho^*>0$ satisfying the $A,\tilde{r}$- sparsity equation, where $\tilde{r}(\cdot,\cdot)$ is the complexity parameter chosen in the previous step.

5. From the value of ρ^*, compute $\tilde{r}(A,\rho^*)$ for any $A>0$.

6. Find a constant $A>0$ verifying Assumption 5.8.

5.3.2 General result in the local bounded framework

In Section 5.3.1, we established a meta theorem to analyze the RERM when the class $F - f^*$ is sub-Gaussian. In this section, we provide another meta theorem when the class $F - f^*$ is locally bounded. Contrary to the main result in the non-regularized case, the neighborhood is now defined with respect to the $L_2(\mu)$ norm **and the regularization norm**.

Definition 5.6. *Let B be the unit ball induced by the regularization norm $\|\cdot\|$. The **complexity parameter** $\tilde{r}^b_{\mathcal{I}}(\cdot,\cdot)$ is defined as*

$$\tilde{r}^b_{\mathcal{I}}(A,\rho) = \inf\left\{ r > 0 : \mathbb{E} \sup_{f \in F(f^* + rB_{L_2} \cap \rho B)} \sum_{i\in\mathcal{I}} \sigma_i (f - f^*)(X_i) \leq \frac{cr^2|\mathcal{I}|}{AL(L+1)} \right\}$$

where $(\sigma_i)_{i\in\mathcal{I}}$ are i.i.d Rademacher random variables independent to $(X_i)_{i\in\mathcal{I}}$, $c > 0$ denotes an absolute constant and L is the Lipschitz constant from assumption 5.3.

Now, adapt the sparsity equation and the local Bernstein condition to this new complexity parameter.

Definition 5.7. *Let $\tilde{r}^b(\cdot,\cdot)$ such that for any $A, \rho > 0$ and, $\tilde{r}^b(A,\rho) \geq r^b_{\mathcal{I}}(A,\rho)$. For any $A, \rho, M > 0$, set*

$$H_{\rho,A,M,\tilde{r}^b} = \{ f \in F \ : \ \|f^* - f\| = \rho \ \text{ and } \ \|f^* - f\|_{L_2} \leq \max(1, \sqrt{LM})\tilde{r}^b(A,\rho) \} \ ,$$

and define

$$\Delta(\rho, A, \tilde{r}^b, M) = \inf_{h \in H_{\rho,A,M,\tilde{r}^b}} \sup_{z^* \in \Gamma_{f^*}(\rho)} z^*(h - f^*) \ . \tag{5.15}$$

*A real number $\rho > 0$ satisfies the $A, M, \tilde{r}^b$-**sparsity equation** if $\Delta(\rho, A, M, \tilde{r}^b) \geq 4\rho/5$.*

Finally, the following assumption imposes boundedness and a Bernstein condition in the small neighborhood around the oracle f^*.

Assumption 5.9. *Let $\tilde{r}^b(\cdot,\cdot)$ be such that for all $A, \rho > 0$, $\tilde{r}^b(A,\rho) \geq \tilde{r}^b_{\mathcal{I}}(A,\rho)$. There exist $A, M > 0$ and ρ^* satisfying the $A, M, \tilde{r}^b$-sparsity equation from Definition 5.7 such that for all $f \in F : \|f - f^*\|_{L_2} = \max(1, \sqrt{LM})\tilde{r}^b(A,\rho^*)$ and $\|f - f^*\| \leq \rho^*$ we have:*

$$\|f - f^*\|^2_{L_2} \leq AP\mathcal{L}_f \quad \text{and} \quad \forall x \in \mathcal{X}, |(f - f^*)(x)| \leq M$$

Assumption 5.8 generalizes the local Bernstein condition and the local boundedness assumption to the regularized case. In this setting, the neighborhood around the oracle f^* can be much smaller than in the non-regularized setting. In particular in Section 5.3.4, the localization with respect to the norm in the RKHS imposes local boundedness of $F - f^*$.
We are now in position to state the main theorem of this section.

Theorem 5.5. *Let $\mathcal{I} \cup \mathcal{O}$ denote a partition of $\{1, \cdots, N\}$ such that $|\mathcal{O}| \leq |\mathcal{I}|$. Let $\tilde{r}^b(\cdot, \cdot)$ be such that for all $A, \rho > 0$, $\tilde{r}^b(A, \rho) \geq \tilde{r}^b_{\mathcal{I}}(A, \rho)$. Grant Assumptions 5.1, 5.3 with $L \geq 1$, 5.2. Suppose that assumption 5.8 holds with $\rho = \rho^*$ satisfying the $A, M, \tilde{r}$-sparsity equation from Definition 5.5 with $A \geq 1$. Set:*

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

As long as $|\mathcal{O}| < c|\mathcal{I}|\tilde{r}^b(A, \rho^*)/(AL)$, with probability larger that

$$1 - 2\exp\bigg(- c\frac{|\mathcal{I}|(\tilde{r}^b(A, \rho^*))^2}{A^2(L+1)^2}\bigg),$$

the estimator $\hat{f}^{\lambda}_N$ defined in Equation (5.12) satisfies

$$\|\hat{f}^{\lambda}_N - f^*\|_{L_2} \leq \max(1, \sqrt{LM})\tilde{r}^b(A, \rho^*) \quad , \quad \|\hat{f}^{\lambda}_N - f^*\| \leq \rho^*$$
$$\text{and} \quad P\mathcal{L}_{\hat{f}^{\lambda}_N} \leq c\max(1, LM)\frac{(\tilde{r}^b(A, \rho^*))^2}{A} \quad .$$

By taking $\tilde{r}^b(A, \rho^*) = c\max(\tilde{r}^b_{\mathcal{I}}(A, \rho^*), AL|\mathcal{O}|/|\mathcal{I}|)$, the condition $|\mathcal{O}| < c|\mathcal{I}|\tilde{r}^b(A, \rho^*)/(AL)$ is necessarily satisfied and we get

$$\|\hat{f}^{\lambda}_N - f^*\|_{L_2} \leq cAL\bigg(\tilde{r}^b_{\mathcal{I}}(A, \rho^*) + \frac{|\mathcal{O}|}{N}\bigg) \quad .$$

The error rate can be decomposed as the sum of the error rate in the non-contaminated setting and the proportion of outliers $|\mathcal{O}|/N$. Theorem 5.5 is a "meta" theorem in the sense that it can used for many practical problems.

General routine to apply Theorem 5.5 This small paragraph explains how in practice we can use Theorem 5.5.

1. Verify assumptions 5.1, 5.3, 5.2.
2. Compute the localized Rademacher complexity localized on $F \cap (f^* + rB_{L_2} \cap \rho B)$ for any $r, \rho > 0$. Deduce the value of $\tilde{r}^b_{\mathcal{I}}(A, \rho)$ for any $A, \rho > 0$.
3. Choose a new complexity parameter such that for every $A, \rho > 0$, $\tilde{r}^b(A, \rho) \geq \tilde{r}^b_{\mathcal{I}}(A, \rho)$. For instance, to derive results in the contaminated setting we will take $\tilde{r}^b(A, \rho) = c\max(\tilde{r}^b_{\mathcal{I}}(A, \rho), AL|\mathcal{O}|/N)$. From the computation of $\tilde{r}^b_{\mathcal{I}}(A, \rho)$ deduce the closed form of $\tilde{r}^b(A, \rho)$.
4. For fixed constants $A, M > 0$, find $\rho^* > 0$ satisfying the $A, M, \tilde{r}^b$- sparsity equation, where $\tilde{r}^b(\cdot, \cdot)$ is the complexity parameter chosen in the previous step.
5. From the value of ρ^*, compute $\tilde{r}(A, \rho^*)$ for any $A > 0$.
6. Find the constants $A, M > 0$ verifying Assumption 5.9.

The main difference with the application of Theorem 5.4 in the sub-Gaussian setting is that we no longer have Assumption 5.4. However it is necessary to verify that the class $F - f^*$ is locally bounded by a constant M.

5.3.3 Application to ℓ_1-penalized Huber's M-estimator with sub-Gaussian design

In this section we use the routine of Theorem 5.4 to the study of ℓ_1-penalized Huber's M-estimator when the design X is supposed to be Gaussian.
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 i.e $\|t^*\|_0 := \sum_{i=1}^p \mathbb{I}\{t_i^* \neq 0\} \leq s$. ℓ_1-penalized Huber's M-estimator is defined as

$$\hat{t}_N^{\delta,\lambda} = \underset{t\in\mathbb{R}^p}{\operatorname{argmin}} \frac{1}{N} \sum_{i=1}^N \ell^\delta(\langle X_i, t \rangle, Y_i) + \lambda \|t\|_1 \tag{5.16}$$

where $\ell^\delta(\cdot, \cdot)$ is the Huber loss function.
Step 1: Under such assumptions, it is clear that Assumptions 5.1, 5.2, 5.3 with $L = \delta$, 5.4 with $B = 1$ are verified. All along this section δ will be considered as a constant.
Step 2: Let us turn to the second step, i.e the computation of the local Gaussian-mean width. Since X is isotropic i.e $\mathbb{E}\langle X, t\rangle^2 = \|t\|_2^2$ for every $t \in \mathbb{R}^p$, we have $w\big(F \cap (f^* + rB_{L_2} \cap \rho B)\big) = w(rB_2^p \cap \rho B_1^p)$ for every $r, \rho > 0$, where B_q^p denotes the ℓ_q ball in $\mathbb{R}^p$ for $q > 0$. Well-known computations give (see (Vershynin, 2018) for example)

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

and consequently,

$$\tilde{r}_{\mathcal{I}}^2(A, \rho) = cA\delta(1 + \delta)\rho\sqrt{\frac{\log(p)}{N}}, \ \ .$$

Step 3 : For any $A, \rho > 0$ let us define $\tilde{r}(A, \rho) = c \max(\tilde{r}_{\mathcal{I}}(A, \rho), A\delta|\mathcal{O}|/|\mathcal{I}|)$. From step 2, since $|\mathcal{I}| \geq N/2$, we easily get:

$$\tilde{r}^2(A, \rho) = cA\delta \max\left((1 + \delta)\rho\sqrt{\frac{\log(p)}{N}}, \frac{|\mathcal{O}|^2}{N^2}\right) \ .$$

Step 4 : To verify the $A, \tilde{r}$-sparsity equation from Definition 5.5 for the ℓ_1 norm and compute ρ^* we use a result from (Lecué and Mendelson, 2018).

Lemma 5.1. *(Lecué and Mendelson, 2018, Lemma 4.2) . Let 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 $100s \leq \big(\rho/\big(\tilde{r}(A, \rho)\big)\big)^2$ then $\Delta(A, \rho, \tilde{r}) \geq (4/5)\rho$.*

Lemma 5.1 implies that the $A, \tilde{r}$-sparsity equation is satisfied by $\rho^* > 0$ if the sparsity s is smaller than $\left(\rho^*/\left(\tilde{r}(A,\rho^*)\right)\right)^2$. Since $\tilde{r}(A,\rho)$ is the maximum of two quantities, we consider two cases depending on the value of $|\mathcal{O}|$. When $\tilde{r}(A,\rho) = \tilde{r}_{\mathcal{I}}(A,\rho)$, which holds when $|\mathcal{O}| \leq |\mathcal{I}|\tilde{r}_{\mathcal{I}}(A,\rho)/(A\delta)$, Lemma 5.1 shows that $\rho^* = c\sqrt{s}\tilde{r}_{\mathcal{I}}(A,\rho^*)$ satisfies the $A, \tilde{r}$-sparsity equation. In this case, straightforward computations give

$$\rho^* = cA\delta(1+\delta)s\sqrt{\frac{\log(p)}{N}} \quad \text{and} \quad \tilde{r}^2_{\mathcal{I}}(A,\rho^*) = c\left(A\delta(\delta+1)\right)^2 s\frac{\log(p)}{N} \ .$$

In the second case, when $\tilde{r}(A,\rho) = A\delta|\mathcal{O}|/|\mathcal{I}|$ which holds when $|\mathcal{O}| \geq |\mathcal{I}|\tilde{r}_{\mathcal{I}}(A,\rho^*)/(A\delta)$ we get that

$$\rho^* = A\delta\sqrt{s}\frac{|\mathcal{O}|}{N}$$

satisfies the $A, \tilde{r}$-sparsity equation. Consequently

$$\rho^* = cA\delta \max\left((\delta+1)s\sqrt{\frac{\log(p)}{N}}, \sqrt{s}\frac{|\mathcal{O}|}{N}\right),$$

satisfies the $A, \tilde{r}$-sparsity equation.

Step 5: From step 4, Theorem 5.4 can be used with

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

Step 6 : We use Proposition 5.1 to show that the local Bernstein condition holds for functions f in $f^* + \tilde{r}(A,\rho^*)S_{L_2} \cap \rho^* B \subset f^* + \tilde{r}(A,\rho^*)S_{L_2}$. Since $X \sim \mathcal{N}(0, I_p)$, the point a) in Assumption 5.7 is verified. Moreover, the point b) in Assumption 5.7 holds and the local Bernstein condition is verified with $A = 4/\alpha$ if $\alpha > 0$ satisfies

$$F_\epsilon\left(\delta - c\tilde{r}(4/\alpha,\rho^*)\right) - F_\epsilon\left(c\tilde{r}(4/\alpha,\rho^*) - \delta\right) \geq \alpha \ , \tag{5.17}$$

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 5.6. *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 $i \in \{1, \cdots, N\}$*

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

where t^* is s-sparse and $(\epsilon_i)_{i\in I}$ are i.i.d centered random variables independent to $(X_i)_{i\in\mathcal{I}}$ such that there exists $\alpha > 0$ such that

$$F_\epsilon\left(\delta - c\frac{\delta}{\alpha}\max\left((\delta+1)\sqrt{s\frac{\log(p)}{N}}, \frac{|\mathcal{O}|}{N}\right)\right) - F_\epsilon\left(c\frac{\delta}{\alpha}\max\left((\delta+1)\sqrt{s\frac{\log(p)}{N}}, \frac{|\mathcal{O}|}{N}\right) - \delta\right) \geq \alpha \tag{5.18}$$

where F_ϵ denotes the cdf of ϵ where ϵ is distributed as ϵ_i for i in $\mathcal{I}$, δ is the hyperparameter of the Huber loss function. Nothing is assumed on $(\epsilon_i)_{i\in\mathcal{O}}$. Set

$$\lambda = c\frac{\delta}{\alpha}\max\left((\delta+1)\sqrt{\frac{\log(p)}{N}}, \frac{|\mathcal{O}|}{\sqrt{s}N}\right) .$$

Then with probability larger than

$$1 - 2\exp\left(-c\frac{\delta}{\alpha(1+\delta)}\max\left((\delta+1)^2 s\log(p), \frac{|\mathcal{O}|^2}{N}\right)\right) \tag{5.19}$$

the estimator $\hat{t}_N^{\delta,\lambda}$ defined in Equation (5.16) satisfies

$$\begin{aligned}
&\|\hat{t}_N^{\delta,\lambda} - t^*\|_2 \leq \frac{\delta}{\alpha}\max\left((\delta+1)\sqrt{s\frac{\log(p)}{N}}, \frac{|\mathcal{O}|}{N}\right)\\
&P\mathcal{L}_{\hat{t}_N^{\delta,\lambda}} \leq \frac{\delta^2}{\alpha}\max\left((\delta+1)^2 s\frac{\log(p)}{N}, \frac{|\mathcal{O}|^2}{N^2}\right)\\
&\textit{and} \quad \|\hat{t}_N^{\delta,\lambda} - t^*\|_1 \leq c\frac{\delta}{\alpha}\max\left((\delta+1)s\sqrt{\frac{\log(p)}{N}}, \sqrt{s}\frac{|\mathcal{O}|}{N}\right)
\end{aligned}$$

Let us analyze the two different cases. 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. We obtain the (nearly) minimax-optimal rate in sparse linear regression in $\mathbb{R}^p$ with an exponentially large probability (Bellec et al., 2018; Lecué and Mendelson, 2018; Dalalyan et al., 2017). Using more involved computations and taking a regularization parameter λ depending on the unknown sparsity we can get 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 (but the confidence is improved) and becomes linear with respect to the proportion of outliers $|\mathcal{O}|/N$. From (Chen et al., 2018), this error rate is minimax optimal (up to a logarithmic term) in the ε-contamination problem when $\varepsilon = |\mathcal{O}|/N$. It follows that Theorem 5.6 is **minimax-optimal** (up to a logarithmic term) when $|\mathcal{O}|$ malicious outliers contaminate the labels.
In Section 5.5, we run simple simulations to illustrate the linear dependence between the error rate and the proportion of outliers.

Remark 5.2. *In Theorem 5.6 we assumed that $\mu = \mathcal{N}(0, I_p)$ to apply Lemma 5.1 and compute the local Gaussian-mean width. It is possible to generalize the result to Gaussian random vectors with covariance matrices Σ verifying $RE(s, 9)$ (Van De Geer et al., 2009), for s being 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 5.6 modulo an extra term $1/\kappa$ in front of the error rate (see Section 5.7 for a precise result).*

In Theorem 5.6, there is no restriction on the noise as long as there exists $\alpha > 0$ such that Equation (5.18) holds. For example when ϵ is a standard Cauchy random variable, Equation (5.18) can be rewritten as

$$c\frac{\delta}{\alpha}\max\left((\delta+1)\sqrt{s\frac{\log(p)}{N}}, \frac{|\mathcal{O}|}{N}\right) \leq \delta - \tan\left(\frac{\pi\alpha}{2}\right) \tag{5.20}$$

Let $\delta > 0$ be a constant (independent to p, s, N) and take $\alpha = (2/\pi)\arctan(\delta/2)$. Equation (5.20) is equivalent to

$$c\max\left((\delta+1)\sqrt{s\frac{\log(p)}{N}}, \frac{|\mathcal{O}|}{N}\right) \leq \alpha = \frac{2}{\pi}\arctan\left(\frac{\delta}{2}\right)$$

which holds as long as $N \geq c(\delta+1)\sqrt{s\log(p)}/\arctan(\delta/2)$ and $|\mathcal{O}| \leq c\arctan(\delta/2)N$ and the local Bernstein condition holds for $A = 4/\alpha = 2\pi/(\arctan(\delta/2))$.

5.3.4 Application to RKHS with the huber loss function

This section is mainly inspired from the work (Alquier et al., 2019). We present another example of application of our main results. In particular, we use the routine associated with Theorem 5.5 for the problem of learning in a reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}_K$ (Steinwart and Christmann, 2008) associated to a positive definite kernel K. We improve the results of (Alquier et al., 2019) in two points 1) we can take $F = \mathcal{H}_K$ while in (Alquier et al., 2019), the authors restrict themselves to the case $F = RB_{\mathcal{H}_K}$, for $R > 0$, where $B_{\mathcal{H}_K}$ denotes the unit ball of $\mathcal{H}_K$ and 2) the bayes rule (i.e the minimizer of the risk over all measurable functions) does not have to belong to $RB_{\mathcal{H}_K}$ and no margin assumption (Audibert et al., 2007) is required.

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 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 $c_1, \cdots, c_n \in \mathbb{R}$, $\sum_{i=1}^n\sum_{j=1}^n c_i c_j K(x_i, x_j) \geq 0$, then by Mercer's theorem there exists an orthonormal basis $(\phi_i)_{i=1}^\infty$ of $L_2(\mu)$ 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

$$\begin{aligned} T_K : L_2(\mu) &\to L_2(\mu) \\ (T_K f)(x) &= \int K(x,y)f(y)d\mu(y) \end{aligned} \tag{5.21}$$

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 $L_2(\mu)$ 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

$$B_{\mathcal{H}_K} = \{f_\beta(\cdot) = \big\langle \beta, \Phi(\cdot) \big\rangle_{\ell_2},\ \|\beta\|_2 \leq 1\} \tag{5.22}$$

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 therefore 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

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

where ℓ^δ is the Huber loss function. Let f be in $\mathcal{H}_K$, by the reproducing property and the Cauchy-Schwarz inequality we have for all x, y in $\mathcal{X}$

$$|f(x) - f(y)| = |\big\langle f, K_x - K_y \big\rangle| \leq \|f\|_{\mathcal{H}_K} \|K_x - K_y\|_{\mathcal{H}_K} \tag{5.23}$$

From Equation (5.23), 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 $\mathcal{X}$. The estimators $\hat{f}_N^{\delta,\lambda}$ we study in this section is defined as

$$\hat{f}_N^{\delta,\lambda} = \operatorname*{argmin}_{f\in\mathcal{H}_K} \frac{1}{N} \sum_{i=1}^{N} \ell^\delta(f(X_i), Y_i) + \lambda \|f\|_{\mathcal{H}_K} \tag{5.24}$$

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

Assumption 5.10. *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 5.10, the value of p is related with the smoothness of the space $\mathcal{H}_K$. Different kinds of spectra could be analysis. 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 (Caponnetto and De Vito, 2007; Mendelson et al., 2010) to obtain fast rates of convergence.

Let us use the routine to apply Theorem 5.5.

Step 1: Since every Reprocucible Kernel Hilbert space is convex, it is clear that assumptions 5.1, 5.2, 5.3 with $L = \delta$ are verified.

Step 2: From Theorem 2.1 in (Mendelson, 2003), if K is a bounded kernel, then for all $\rho, r > 0$

$$\mathbb{E} \sup_{f \in \mathcal{H}_K \cap (f^* + rB_{L_2} \cap \rho B_{\mathcal{H}_K})} \frac{1}{\sqrt{N}} \left| \sum_{i=1}^{N} \sigma_i (f - f^*)(X_i) \right| \leq \sqrt{2} \|K\|_\infty \left(\sum_{k=1}^{\infty} \left(\rho^2 \lambda_k \wedge r^2 \right) \right)^{1/2} .$$

Under assumption 5.10, 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 $A, \rho > 0$

$$\tilde{r}^b_{\mathcal{I}}(A, \rho) = c \big(A\delta(\delta + 1) \|K\|_\infty \big)^{1/(p+1)} \frac{\rho^{p/(p+1)}}{N^{1/(2(p+1))}}$$

Step 3: For any $A, \rho > 0$, let us define $\tilde{r}^b(A, \rho) = c \max \left(\tilde{r}^b_{\mathcal{I}}(A, \rho), A\delta |\mathcal{O}|/|\mathcal{I}| \right)$. From step 2, since $|\mathcal{I}| \geq N/2$, we easily get

$$\tilde{r}^b(A, \rho) = c \max \left(\big(A\delta(\delta + 1) \|K\|_\infty \big)^{1/(p+1)} \frac{\rho^{p/(p+1)}}{N^{1/(2(p+1))}}, A\delta \frac{|\mathcal{O}|}{N} \right)$$

Step 4: Let $A, M > 0$. From Remark 5.1, $\rho^* = 20 \|f^*\|_{\mathcal{H}_K}$ satisfies the $A, M, \tilde{r}^b$-sparsity equation.

Step 5: From step 4, we easily get

$$\tilde{r}^b(A, \rho^*) = c \max \left(\big(A\delta(\delta + 1) \|K\|_\infty \big)^{1/(p+1)} \frac{\|f^*\|^{p/(p+1)}_{\mathcal{H}_K}}{N^{1/(2(p+1))}}, A\delta \frac{|\mathcal{O}|}{N} \right)$$

Step 6: In assumption 5.9 there are two conditions to verify 1) the local Bernstein and 2) the local boundedness. Let us begin by the local Bernstein condition. We use the localized version of Theorem 5.1.

Assumption 5.11. *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 verifying $\|f - f^*\| \leq \rho$ and $\|f - f^*\|_{L_2} = r$ 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 $|z - f^*(x)| \leq (\sqrt{2}(C'))^{(2+\varepsilon)/\varepsilon} r$, $F_{Y|X=x}(z + \delta) - F_{Y|X=x}(z - \delta) \geqslant \alpha$.

The only difference with Assumption 5.7 is that the point a) is only required for functions f in F such that $\|f - f^*\| \leq \rho$.

Proposition 5.2. *Grant assumption 5.7. The Huber loss function with parameter $\delta > 0$ satisfies the Bernstein condition for $A = 4/\alpha$: for all $f \in F$, if $\|f - f^*\|_{L_2} = r$ and $\|f - f^*\| \leq \rho$ then $(4/\alpha)P\mathcal{L}_f \geq \|f - f^*\|_{L_2}^2$.*

Proposition 5.2 is a simple refinement of Proposition 5.1. 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

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

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

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

Therefore, the point a) holds with $C' = (\rho \|K\|_\infty / r)^{\varepsilon/(2+\varepsilon)}$. Let us turn to the point b) of assumption 5.11. From the fact that $C' = (\rho \|K\|_\infty / r)^{\varepsilon/(2+\varepsilon)}$, we have $(\sqrt{2}C')^{(2+\varepsilon)/\varepsilon} r = 2^{(2+\varepsilon)/2\varepsilon} \rho \|K\|_\infty$ and the point b) can be rewritten as, there exists $\alpha > 0$ such that

$$F_\epsilon(\delta - c\rho\|K\|_\infty) - F_\epsilon(c\rho\|K\|_\infty - \delta) \geq \alpha \tag{5.25}$$

where F_ϵ denotes the cdf of ϵ distributed as ϵ_i for $i \in \mathcal{I}$. Equation (5.25), simply means that the noise ϵ puts enough mass around 0. In our problem we have $\rho = \rho^* = c\|f^*\|_{\mathcal{H}_K}$ and Equation (5.25) becomes,

$$F_\epsilon(\delta - c\|f^*\|_{\mathcal{H}_K}\|K\|_\infty) - F_\epsilon(c\|f^*\|_{\mathcal{H}_K}\|K\|_\infty - \delta) \geq \alpha$$

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 $|f(x) - f^*(x)| \leq \|K\|_\infty \rho^*$. As a consequence, in our setting, $M = c\|K\|_\infty \|f^*\|_{\mathcal{H}_K}$ satisfies the local boundedness assumption.

We are now in position to state our main theorem for regularized learning in RKHS with the Huber loss function.

Theorem 5.7. *Let $\mathcal{H}_K$ be a reproducing kernel Hilbert space associated with a bounded kernel K. 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 $(\epsilon_i)_{i \in I}$ are i.i.d symmetric random variables independent to $(X_i)_{i \in \mathcal{I}}$ such that there exists $\alpha > 0$ such that

$$F_\epsilon\big(\delta - c\|f^*\|_{\mathcal{H}_K}\|K\|_\infty\big) - F_\epsilon\big(c\|f^*\|_{\mathcal{H}_K}\|K\|_\infty - \delta\big) \geq \alpha \tag{5.26}$$

where F_ϵ denotes the cdf of ϵ where ϵ is distributed as ϵ_i for i in $\mathcal{I}$, δ is the hyperparameter of the Huber loss function. Nothing is assumed on $(\epsilon_i)_{i\in\mathcal{O}}$. Grant assumption 5.10 and let

$$\lambda = c\frac{\alpha}{\|f^*\|_{\mathcal{K}_K}} \max\left(\left(\frac{\delta(\delta+1)}{\alpha}\|K\|_\infty\right)^{2/(p+1)} \frac{\|f^*\|_{\mathcal{H}_K}^{(2p)/(p+1)}}{N^{1/(p+1)}}, \frac{\delta^2}{\alpha^2}\frac{|\mathcal{O}|^2}{N^2}\right) .$$

Then with probability larger than

$$1 - 2\exp\left(-c\frac{\alpha^2}{(1+\delta)^2}\max\left(\left(\frac{\delta(\delta+1)}{\alpha}\|K\|_\infty\right)^{2/(p+1)} \|f^*\|_{\mathcal{H}_K}^{(2p)/(p+1)} N^{p/(p+1)}, \frac{\delta^2}{\alpha^2}\frac{|\mathcal{O}|^2}{N}\right)\right)$$

the estimator $\hat{f}_N^{\delta,\lambda}$ defined in Equation (5.24) satisfies

$$\|\hat{f}_N^{\delta,\lambda} - f^*\|_2^2 \le c\max\left(1, \delta\|f^*\|_{\mathcal{H}_K}\|K\|_\infty\right)\max\left(\left(\frac{\delta(\delta+1)}{\alpha}\|K\|_\infty\right)^{2/(p+1)} \frac{\|f^*\|_{\mathcal{H}_K}^{(2p)/(p+1)}}{N^{1/(p+1)}}, \frac{\delta^2}{\alpha^2}\frac{|\mathcal{O}|^2}{N^2}\right)$$

$$P\mathcal{L}_{\hat{f}_N^{\delta,\lambda}} \le c\alpha\max\left(1, \delta\|f^*\|_{\mathcal{H}_K}\|K\|_\infty\right)\max\left(\left(\frac{\delta(\delta+1)}{\alpha}\|K\|_\infty\right)^{2/(p+1)} \frac{\|f^*\|_{\mathcal{H}_K}^{(2p)/(p+1)}}{N^{1/(p+1)}}, \frac{\delta^2}{\alpha^2}\frac{|\mathcal{O}|^2}{N^2}\right)$$

and $\quad \|\hat{f}_N^{\delta,\lambda} - f^*\|_{\mathcal{H}_K} \le c\|f^*\|_{\mathcal{H}_K}$

Theorem 5.7 holds with no assumption on the design X. When $|\mathcal{O}| \le (\delta/\alpha)Nr_{\mathcal{I}}^b(4\alpha, \|f^*\|_{\mathcal{H}_K})$ we recover the same rates as (Smale and Zhou, 2007; Mendelson et al., 2010) even when the target Y is heavy-tailed. In (Smale and Zhou, 2007; Mendelson et al., 2010) the authors assume that Y is bounded while in (Caponnetto and De Vito, 2007) the noise is assumed to be light-tailed. When $|\mathcal{O}| \ge (\delta/\alpha)Nr_{\mathcal{I}}^b(4\alpha, \|f^*\|_{\mathcal{H}_K})$ the error rate is deteriorated and becomes linear with respect to the proportion of outliers.
It is assumed that the noise is symmetric and satisfies Equation (5.26). When the noise ϵ is a standard Cauchy random variable Equation (5.26) can be rewritten as

$$c\|f^*\|_{\mathcal{H}_K}\|K\|_\infty \le \delta - \tan\left(\frac{\alpha\pi}{2}\right)$$

which holds for $\delta = c\|f^*\|_{\mathcal{H}_K}\|K\|_\infty$ and $\alpha = \arctan(\delta/2)$. When $\delta, \|K\|_\infty$ and $\|f^*\|_{\mathcal{H}_K}$ are seen as constants, the error rate is of order $N^{-1/(p+1)}$. Depending on the value of p we obtained fast rates of convergence for regularized Kernel methods. The faster the spectrum of T_K decreases the faster the rates of convergence.

5.4 Conclusion and perspectives

We have presented general analyses to study ERM and RERM when a number $|\mathcal{O}|$ of outliers may 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 obtain minimax-optimal rate of convergence for these two examples when $|\mathcal{O}|$

malicious outliers corrupt the labels. We also 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 analysed as it has be done in (Chinot et al., 2019a) such as Group Lasso, SLOPE ... 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. However in such setting ERM and RERM are known to be sub-optimal which is why such results have not been presented in this paper.

5.5 Simulations

In this section, we present simple simulations to illustrate our theoritical 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 distribution 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 M-Huber's estimator defined as

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

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

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

Note that other loss functions could be considered as the absolute loss function, or more generally, any quantile loss function. According to Theorem 5.3, we have

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

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 hyperparameter of the problem is δ. For the sake of simplicity we took $\delta = 1$ for all our simulations. We see on Figure 5.1 that no matter the noise, the error rate is proportional to the proportion of outliers which is in adequation with our theoritical findings.

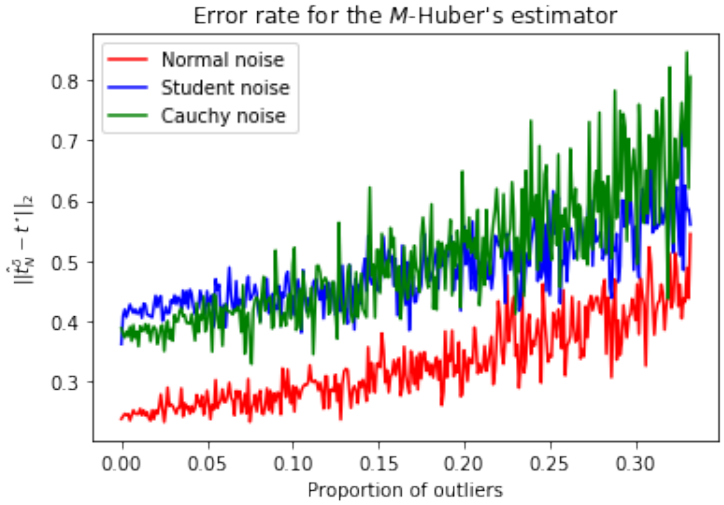


Figure 5.1: Error rate for the M-Huber's estimator ($p = 50$ and $N = 1000$)

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

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

where $\ell^\delta : \mathbb{R} \times \mathbb{R} \mapsto \mathbb{R}^+$ is the Huber loss function and $\lambda > 0$ is a hyperparameter. According to Theorem 5.6 we have

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

where $c > 0$ is an absolute constant. We ran our simulations with $N = 1000$ and $p = 1000$ and $s = 50$. The hyperparameters of the problem are δ and λ. For the sake of simplicity we took $\delta = 1$ and $\lambda = 10^{-3}$ for all our simulations. We see on Figure 5.2 that no matter the noise, the error rate is proportional to the proportion of outliers which is in adequation with our theoritical findings. The fact that the error rate may be large comes to the fact that we did not optimize the value of λ.

5.6 Lower bound minimax risk in regression where only the labels are contaminated

This section is built on the work (Chen et al., 2018) where the authors establish a general minimax theory for the ε-contamination model defined as $\mathbb{P}_{(\varepsilon,\theta,Q)} = (1-\varepsilon)P_\theta + \varepsilon Q$ given a general statistical

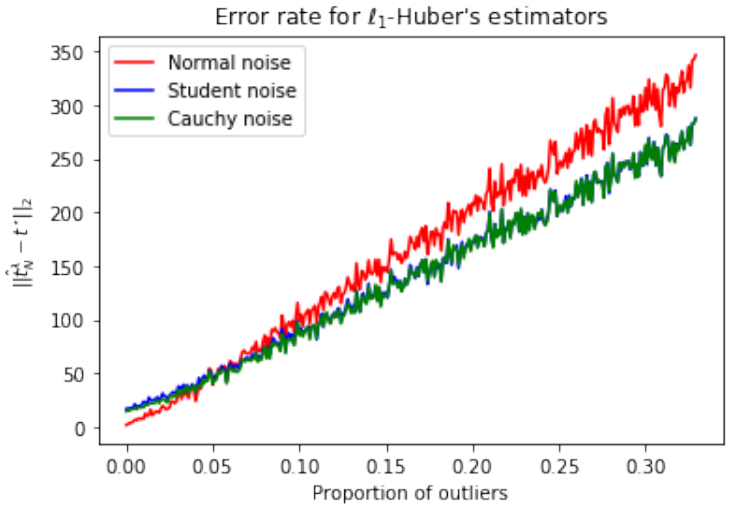


Figure 5.2: Error rate for ℓ_1 penalized M-Huber's estimator ($p = 1000$ and $N = 1000$ and $s = 50$)

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 $\{\mathbb{P}_{(\varepsilon,\theta,Q)}, \theta \in \Theta, Q\}$ depends on the modulus of continuity defined as:

$$w(\varepsilon, \Theta) = \sup \left\{ L(\theta_1, \theta_2) : TV(P_{\theta_1}, P_{\theta_2}) \leq \frac{\varepsilon}{1-\varepsilon}, \theta_1, \theta_2 \in \Theta \right\} \tag{5.27}$$

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 sigma-algebra onto which P_{θ_1} and P_{θ_2} are defined.

Theorem 5.8 (Theorem 5.1 (Chen et al., 2018))**.** *Suppose there is some $\mathcal{M}(0)$ such that for $\varepsilon = 0$*

$$\inf_{\hat{\theta}} \sup_{\theta \in \Theta} \sup_{Q} \mathbb{P}_{(\varepsilon,\theta,Q)} \left(L(\theta, \hat{\theta}) \geq \mathcal{M}(\varepsilon) \right) \geq c \tag{5.28}$$

holds. Then, for any $\varepsilon \in [0, 1]$ (5.28) 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 illusrate Theorem 5.8, 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 (Chen et al., 2016), the authors consider a setting when both the design X and the response variable in the model can be contaminated i.e $(X_1, Y_1), \cdots, (X_N, Y_N) \sim (1-\varepsilon)P_\theta + \varepsilon Q$, whith $P_\theta = P(X)P(Y|X)$, $P(X) = \mathcal{N}(0, \Sigma)$ 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

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

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 5.9. *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* $\varepsilon = 0$

$$\inf_{\hat{\theta}} \sup_{R_{\theta,\varepsilon} \in \mathcal{P}_{\theta,\varepsilon}, \theta \in \Theta} R_{\theta,\varepsilon}\Big(L(\theta, \hat{\theta}) \geq \mathcal{M}(\varepsilon) \Big) \geq c \tag{5.29}$$

holds. Then For any $\varepsilon \in [0,1]$ (5.29) *holds for* $\mathcal{M}(\varepsilon) = c\big(\mathcal{M}(0) \vee w(\varepsilon, \Theta)\big)$

Theorem 5.9 states that the minimax optimal rates for regression problems in the ε-contamination model are the same when

- Both the design X and the response variable Y are contaminated.
- Only the response variable $Y$is contaminated.

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

$$p_{\theta_i}(x,y) = \frac{dP^{\theta_i}_{(X,Y)}}{d\big(P^{\theta_1}_{(X,Y)} + P^{\theta_2}_{(X,Y)}\big)}(x,y) \tag{5.30}$$

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 (Chen et al., 2018) we have $R_{\theta_1,\varepsilon} \in \mathcal{P}_{\theta_1,\varepsilon}$ and $R_{\theta_2,\varepsilon} \in \mathcal{P}_{\theta_2,\varepsilon}$.
To conclude, it remains to show that $R_{\theta_1,\varepsilon} = R_{\theta_2,\varepsilon}$. For any $(x,y) \in \mathcal{X}\times\mathcal{Y}$. Straightforward computations give

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

■

5.7 ℓ_1-penalized Huber's M-estimator with non-isotropic design

In this section, we relax the isotropic assumption on the design X. Recal that a random variable X is isotropic if for every $t \in \mathbb{R}^p$, $\mathbb{E}\big\langle X, t\big\rangle^2 = \|t\|_2^2$. Instead, we consider covariance matrices satisfying a Resticted 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 5.6 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 5.1 and the computation of the Gaussian mean-width is more involved.

Lemma 5.2. *Let 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 $100s \leq \left(\kappa\rho/r\right)^2$ then:*

$$\Delta(\rho) = \inf_{w \in \rho S_1 \cap r B_{L_2}} \sup_{z^* \in \Gamma_{t^*}(\rho)} \left\langle z^*, w \right\rangle \geq 4\rho/5 \ .$$

The difference with Lemma 5.1 is the term κ coming from the RE condition.

Proof. To solve the sparsity equation – find ρ^* such that $\Delta(\rho) \geq (4/5)\rho$ – , we use the following classical result on the sub-differential of a norm: if $\|\cdot\|$ is a norm on $\mathbb{R}^p$, then, for all $t \in \mathbb{R}^p$, we have

$$(\partial \|\cdot\|)_t = \begin{cases} \{z^* \in S^* : \left\langle z^*, t \right\rangle = \|t\|\} & \text{if } t \neq 0 \\ B^* & \text{if } t = 0 \end{cases} \ . \tag{5.31}$$

Here, B^* is the unit ball of the dual norm associated with $\|\cdot\|$, i.e. $t \in \mathbb{R}^p \to \|t\|^* = \sup_{\|v\| \leq 1} \left\langle v, t \right\rangle$ and S^* is its unit sphere. In other words, when $t \neq 0$, the sub-differential of $\|\cdot\|$ in t is the set of all vectors z^* in the unit dual sphere S^* which are norming for t (i.e. z^* is such that $\left\langle z^*, t \right\rangle = \|t\|$). In particular, when $t \neq 0$, $(\partial \|\cdot\|)_t$ is a subset of the dual sphere S^*.

Since $F = \{\left\langle t, \cdot \right\rangle, t \in \mathbb{R}^p\}$, $\|f\|_{L_2} = \|\left\langle t, X \right\rangle\|_{L_2} = \|\Sigma^{1/2} t\|_2$. 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 $i \in I$, $z_i = sign(t_i^*)$, and for every $i \in I^c$, $z_I = sign(w_i)$. It is clear that z is norming for t^* i.e $\left\langle z, t^* \right\rangle = \|t^*\|_1$ and $z \in S_1^* = S_\infty$ and

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

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

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

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

$$\left\langle z, w \right\rangle \geq \rho - 2\|P_I w\|_1 \geq \rho - 2\sqrt{s}\|P_I w\|_2 \geq \rho - \frac{2}{\kappa}\sqrt{s}\|\Sigma^{1/2} w\|_2 \geq \rho - \frac{2}{\kappa}\sqrt{s} r \geq 4\rho/5$$

■

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 5.3 (Proposition 1 (C Bellec, 2019))**.** *Let $p \geq 1$ and $M \geq 2$. Let T be the convex hull of M points in $\mathbb{R}^p$ and assume that $T \subset B_2^p$. Let $\mathbf{G} \sim \mathcal{N}(0, I_p)$. Then for all $s > 0$,*

$$\mathbb{E} \sup_{t \in sB_2^p \cap T} \left\langle t, \mathbf{G} \right\rangle = w(sB_2^p \cap T) \leq 4\sqrt{\log_+(4eM(s^2 \wedge 1))},$$

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

$$w\big(F \cap (f^* + rB_{L_2} \cap \rho B_1^p\big) = \mathbb{E} \sup_{t \in \mathbb{R}^p : \|\Sigma^{1/2}\|_2 \leq r, \|t\|_1 \leq \rho} \big\langle \Sigma^{1/2} t, \mathbf{G} \big\rangle$$

where $\mathbf{G} \sim \mathcal{N}(0, I_p)$. If Σ is assumed to be inversible, we get

$$w\big(F \cap (f^* + rB_{L_2} \cap \rho B_1^p\big) = w\big(rB_2^p \cap \rho \Sigma^{1/2} B_1^p\big) = w\big(rB_2^p \cap \rho T\big)$$

where $T := \Sigma^{1/2} B_1^p$ is the convex hull of $(\pm \Sigma^{1/2} e_i)_{i=1}^p$. To apply Proposition 5.3 it is necessary to assume that for every $i = 1, \cdots, p$, $\Sigma^{1/2} e_i \in B_2^p$ which holds when $\Sigma_{i,i} \leq 1$ and we get

Proposition 5.4. *Let $F = \{\langle t, \cdot \rangle, t \in \mathbb{R}^p\}$ and assume that, Σ, the covariance matrix of X is invertible and satisfies $\Sigma_{i,i} \leq 1$ for every $i = 1, \cdots, p$. Then, for every $r, \rho > 0$*

$$w\big(F \cap (f^* + rB_{L_2} \cap \rho B_1^p\big) \leq 4\rho \sqrt{\log_+(8ep((r/\rho)^2 \wedge 1))}$$

Straightforward computations (see (Lecué and Mendelson, 2018) for instance) show that s Steps 3,4,5,6 in Section 5.3.3 are not modified and the following theorem extends Theorem 5.6 for a non-isotropic design:

Theorem 5.10. *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 is invertible, 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 $i \in \{1, \cdots, N\}$*

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

where t^* is s-sparse and $(\epsilon_i)_{i \in I}$ are i.i.d random variables independent to $(X_i)_{i \in \mathcal{I}}$ such that there exists $\alpha > 0$ such that

$$F_\epsilon\bigg(\delta - c\frac{\delta}{\kappa\alpha} \max\bigg((\delta+1)\sqrt{s\frac{\log(p)}{N}}, \frac{|\mathcal{O}|}{N}\bigg)\bigg) - F_\epsilon\bigg(c\frac{\delta}{\kappa\alpha} \max\bigg((\delta+1)\sqrt{s\frac{\log(p)}{N}}, \frac{|\mathcal{O}|}{N}\bigg) - \delta\bigg) \geq \alpha \quad (5.32)$$

where F_ϵ denotes the cdf of ϵ where ϵ is distributed as ϵ_i for i in $\mathcal{I}$, δ is the hyperparameter of the Huber loss function. Nothing is assumed on $(\epsilon_i)_{i \in \mathcal{O}}$. Set

$$\lambda = c\frac{\delta}{\alpha} \max\bigg((\delta + 1)\sqrt{\frac{\log(p)}{N}}, \frac{|\mathcal{O}|}{\sqrt{s}N}\bigg) \ .$$

Then with probability larger than

$$1 - 2\exp\bigg(- \frac{\delta}{\kappa^2 \alpha(1+\delta)} \max\bigg((\delta+1)^2 s \log(p), \frac{|\mathcal{O}|^2}{N}\bigg)\bigg) \quad (5.33)$$

the estimator $\hat{t}_N^{\delta,\lambda}$ defined in Equation (5.16) satisfies

$$\|\hat{t}_N^{\delta,\lambda} - t^*\|_2 \leq \frac{\delta}{\kappa\alpha} \max\bigg((\delta+1)\sqrt{s\frac{\log(p)}{N}}, \frac{|\mathcal{O}|}{N}\bigg)$$

$$P\mathcal{L}_{\hat{t}_N^{\delta,\lambda}} \leq \frac{\delta^2}{\kappa^2\alpha} \max\bigg((\delta+1)^2 s\frac{\log(p)}{N}, \frac{|\mathcal{O}|^2}{N^2}\bigg)$$

$$\text{and} \quad \|\hat{t}_N^{\delta,\lambda} - t^*\|_1 \leq c\frac{\delta}{\kappa^2\alpha} \max\bigg((\delta+1)s\sqrt{\frac{\log(p)}{N}}, \sqrt{s}\frac{|\mathcal{O}|}{N}\bigg)$$

We recover the main result from (Dalalyan and Thompson, 2019) as a special case of our main theorem. However, we do not assume that the noise is Gaussian. It can be heavy-tailed. It mainly generalizes their results.

Remark 5.3. *When $|\mathcal{O}| \leq (\delta+1)\sqrt{s\log(p)N}$, the regularization parameter λ does not depend on the unknown sparsity s. 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 depend on the sparisty s.*

5.8 Proofs main Theorems

5.8.1 Proof Theorem 5.1

Let $r(\cdot)$ be such that for all $A>0$: $r(A) \geq r_{\mathcal{I}}(A)$ and let A satisfying assumption 5.5 with $r(\cdot)$. The proof is split into two parts. First we identify a stochastic argument holding with large probability. Then we show on that event that $\|\hat{f}_N - f^*\|_{L_2(\mu)} \leq r(A)$. Finally, at the very end of the proof we show that $P\mathcal{L}_{\hat{f}_N} \leq r^2(A)/A$.

Stochastic arguments First we identifiate the stochastic event onto which the proof easily follows. Let,

$$\Omega_{\mathcal{I}} = \left\{ \forall f \in F : \|f-f^*\|_{L_2(\mu)} \leq r(A) : \left| \big(P - P_{\mathcal{I}}\big)\big(\ell_f - \ell_{f^*}\big) \right| \leq \frac{1}{2A(1+L)} r^2(A) \right\} \tag{5.34}$$

$$\Omega_{\mathcal{O}} = \left\{ \forall f \in F : \|f-f^*\|_{L_2(\mu)} \leq r(A) : \left| \big(P - P_{\mathcal{O}}\big)|f-f^*| \right| \leq \frac{1}{2A(1+L)} \sqrt{\frac{|\mathcal{I}|}{|\mathcal{O}|}} r^2(A) \right\} \tag{5.35}$$

where for any $K \subset \{1,\cdots,N\}$, $g : \mathcal{X}\times\mathcal{Y} \mapsto \mathbb{R}$, $P_K g = 1/(|K|)\sum_{i\in K} g(X_i,Y_i)$. Finally let us define $\Omega = \Omega_{\mathcal{I}} \cap \Omega_{\mathcal{O}}$.

Lemma 5.3. *Grant Assumptions 5.1, 5.3, 5.2, 5.4 and 5.5 with $r(\cdot)$. Then there exists an absolute constant $c>0$ such the event Ω holds with probability larger than*

$$1 - 2\exp\big(-c|\mathcal{I}|r^2(A)/(LBA(L+1))\big)$$

The proof of Lemma 5.3 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 $(X_t)_{t\in T}$ denote a stochastic process indexed by a pseudo metric space (T,d) satisfying the following Lipschitz condition

$$\text{for all } t,s\in T,\quad \|X_t-X_s\|_{\psi_2}\leq d(t,s) \tag{5.36}$$

For such a process it is possible to control the deviation of $\sup_{t\in T} X_t$ in terms of the geometry of (T,d) trough the Talagrand's γ-functionals.

Theorem 5.11 ((Ledoux and Talagrand, 2013), Theorem 11.13)**.** *Let $(X_t)_{t\in T}$ be a random process in $L_1(\Omega,\mathcal{A},\mathbb{P})$ indexed by a pseudo metric space (T,d) such that for all measurable sets A in Ω*

$$\int_A |X_s-X_t|d\mathbb{P}\leq d(s,t)\mathbb{P}(A)\psi_2^{-1}\left(\frac{1}{\mathbb{P}(A)}\right)\ , \tag{5.37}$$

then, there exists an absolute constant $c>0$ such that for all $u>0$

$$\mathbb{P}\left(\sup_{s,t\in T}|X_t-X_s|\geq c(\gamma_2+u)\right)\leq\left(\psi_2(u/D(T))\right)^{-1}$$

where γ_2 is the majorizing measure integral $\gamma(T,d,\psi_2)$ and $D(T)$ is the diameter of (T,d).

First note that Equation (5.36) implies Equation (5.37). By Jensen inequality and the definition of $\|\cdot\|_{\psi_2}$ we get

$$\begin{aligned}\int_A |X_s-X_t|d\mathbb{P} &= d(s,t)\mathbb{P}(A)\int_A \psi_2^{-1}\circ\psi_2\left(\frac{|X_s-X_t|}{d(s,t)}\right)\frac{d\mathbb{P}}{\mathbb{P}(A)}\\ &\leq d(s,t)\mathbb{P}(A)\psi_2^{-1}\left(\frac{1}{\mathbb{P}(A)}\mathbb{E}\psi_2\left(\frac{|X_s-X_t|}{d(s,t)}\right)\right)\\ &\leq d(s,t)\mathbb{P}(A)\psi_2^{-1}\left(\frac{1}{\mathbb{P}(A)}\right)\end{aligned}$$

Moreover, from the Majorizing Measure Theorem (Talagrand, 2006)[Theorem 2.1.1], when T is a subset of $L_2(\mu)$ and $d(s,t)=\sqrt{\mathbb{E}(X_s-X_t)^2}$ we have $c_1 w(T)\leq\gamma_2(T)\leq c_2 w(T)$ for $c_1,c_2>0$ two absolute constants and $w(T)$ is the Gaussian mean-width of T defined in Definition 5.1. The corollary follows:

Corollary 5.2. *Let $\tilde{F}\subset L_2(\mu)$ such that $(X_f)_{f\in\tilde{F}}$ is stochastic process indexded by $\tilde{F}$ satisfying for any $f,g\in\tilde{F}$: $\|X_f-X_g\|_{\psi_2}\leq L\|f-g\|_{L_2(\mu)}$. Then, for any $u\geq\log(2)$, with probability larger than $1-\exp(u^2)$*

$$\sup_{f,g\in\tilde{F}}|X_f-X_g|\leq cL\big(w(\tilde{F})+uD_{L_2(\mu)}(\tilde{F})\big)$$

where $c>0$ is an absolute constant, $w(\tilde{F})$ is the Gaussian mean-width of $\tilde{F}$ and $D_{L_2(\mu)}(\tilde{F})$ its $L_2(\mu)$-diameter.

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

Lemma 5.4 ((Chafaï et al., 2012), 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*

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

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

Lemma 5.5 ((Chafaï et al., 2012), 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 $\lambda \geq c_1/K$*

$$\mathbb{E}\exp(\lambda|Z|) \leq c_3 \exp(c_2\lambda^2 K^2) \tag{5.38}$$

We are now in position to prove Lemma 5.3.

Proof. First we prove that $\Omega_{\mathcal{I}}$ holds with probability larger than $\exp\big(-c|\mathcal{I}|r^2(A)/(ALB(1+L))\big)$. Let $\tilde{F} = \{f \in F : \|f - f^*\|_{L_2(\mu)} \leq r(A)\}$. Let us assume that for any f, g in $\tilde{F}$, the following condition holds

$$\|\big(P - P_{\mathcal{I}}\big)\big(\ell_f - \ell_g\big)\|_{\psi_2} \leq c(LB/\sqrt{|\mathcal{I}|})\|f - g\|_{L_2(\mu)} \tag{5.39}$$

then, from Corollary 5.2, for any $u \geq \log(2)$, there exists an absolute constant $c > 0$ such that with probability larger that $1 - \exp(u^2)$

$$\begin{aligned}
\sup_{f\in\tilde{F}}\Big|\big(P - P_{\mathcal{I}}\big)\big(\ell_f - \ell_{f^*}\big)\Big| &\leq \sup_{f,g\in\tilde{F}}\Big|\big(P - P_{\mathcal{I}}\big)\big(\ell_f - \ell_g\big)\Big| \\
&\leq c\frac{LB}{\sqrt{|\mathcal{I}|}}\big(w(\tilde{F}) + uD_{L_2(\mu)}(\tilde{F})\big) \\
&\leq c\frac{LB}{\sqrt{|\mathcal{I}|}}\Big(w\big(F \cap (f^* + r(A)B_{L_2(\mu)})\big) + ur(A)\Big)
\end{aligned}$$

As $r(A) \geq r_{\mathcal{I}}(A)$ it follows that $w\big(F \cap (f^* + r(A)B_{L_2(\mu)})\big) \leq \sqrt{|\mathcal{I}|}r^2(A)/(ABL(L+1))$.By taking $u = c\sqrt{|\mathcal{I}|}r(A)/(ABL(L+1))$ we obtain the result. With the same reasoning if we assume that

$$\big\|\big(P - P_{\mathcal{O}}\big)|f - g|\big\|_{\psi_2} \leq c(BL)/\sqrt{|\mathcal{O}|})\|f - g\|_{L_2(\mu)} \ , \tag{5.40}$$

then, with probability larger that $1 - \exp\big(-c|\mathcal{I}|r^2(A)/(ABL(L+1))\big)$:

$$\sup_{f\in\tilde{F}}\Big|\big(P - P_{\mathcal{O}}\big)|f - f^*|\Big| \leq \frac{1}{2A(L+1)}\sqrt{\frac{|\mathcal{I}|}{|\mathcal{O}|}}r^2(A)$$

To finish the proof it remains to show that Equations (5.39) and (5.40) hold. From Lemma 5.4 we get

$$\begin{aligned}\|\big(P - P_{\mathcal{I}}\big)(\ell_f - \ell_g)\|_{\psi_2} &\leq 16\bigg(\sum_{i\in\mathcal{I}} \frac{\|(\ell_f - \ell_g)(X_i, Y_i) - \mathbb{E}(\ell_f - \ell_g)(X_i, Y_i)\|_{\psi_2}^2}{|\mathcal{I}|^2}\bigg)^{1/2} \\ &= \frac{16}{\sqrt{|\mathcal{I}|}}\|(\ell_f - \ell_g)(X, Y) - \mathbb{E}(\ell_f - \ell_g)(X, Y)\|_{\psi_2}\end{aligned}$$

Thus, it remains to show that $\|(\ell_f - \ell_g)(X, Y) - \mathbb{E}(\ell_f - \ell_g)(X, Y)\|_{\psi_2} \leq cLB\|f - g\|_{L_2(\mu)}$ for $c > 0$ an absolute constant. To do so, we use Lemma 5.5. Let $\lambda \geq cLB/(\|f - g\|_{L_2(\mu)})$. From the symmetrization principle (Lemma 6.3 in (Ledoux and Talagrand, 2013)) and the contraction principle (Theorem 2.2 in (Koltchinskii, 2011b)) 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 variation independent to (X, Y). From assumption 5.4, we get

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

which concludes the proof for $\Omega_{\mathcal{I}}$ with Lemma 5.5. For $\Omega_{\mathcal{O}}$, since $L \geq 1$ we have

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

which also concludes the proof for $\Omega_{\mathcal{O}}$. ∎

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 $\|\hat{f}_N - f^*\|_{L_2(\mu)} \leq r(A)$ it is sufficient to show that for all functions $f \in F$ such that $\|f - f^*\|_{L_2(\mu)} \geq r(A)$ we have $P_N\mathcal{L}_f > 0$. Let f in F such that $\|f - f^*\|_{L_2(\mu)} \geq r(A)$. By convexity of F there exists a function f_1 such that $\|f_1 - f^*\|_{L_2(\mu)} = r(A)$ for which

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

where $\alpha = \big(\|f - f^*\|_{L_2(\mu)}/r(A)\big) \geq 1$. 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 5.3. 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:

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

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 $f_1 \in F \cap (f^* + r(A)S_{L_2(\mu)})$, where $S_{L_2(\mu)}$ denotes the unit sphere induced by $L_2(\mu)$. 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 (5.34)) it follows that

$$P_{\mathcal{I}}\mathcal{L}_{f_1} \geq P\mathcal{L}_{f_1} - \frac{1}{2A(1+L)}r^2(A) \geq \left(A^{-1} - \frac{1}{2A(1+L)}\right)r^2(A) \tag{5.41}$$

where we used assumption 5.5. Moreover, from assumption 5.3, 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 (5.35)), we get

$$P_{\mathcal{O}}\mathcal{L}_{f_1} \geq -L\|f_1 - f^*\|_{L_1} - \frac{L}{2A(1+L)}\sqrt{\frac{|\mathcal{I}|}{|\mathcal{O}|}}r^2(A) \geq -L\|f_1 - f^*\|_{L_2} - \frac{L}{2A(1+L)}\sqrt{\frac{|\mathcal{I}|}{|\mathcal{O}|}}r^2(A)$$
$$= -Lr(A) - \frac{L}{2A(1+L)}\sqrt{\frac{|\mathcal{I}|}{|\mathcal{O}|}}r^2(A) \ . \tag{5.42}$$

Since $|\mathcal{O}| < |\mathcal{I}|$, from Equations (5.41), (5.42) it follows

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

as long as $|\mathcal{O}| < (1/2AL)|\mathcal{I}|r(A)$. It concludes the proof for the error rate.
We finish the proof by establishing the result for the excess risk. Since $\|\hat{f}_N - f^*\|_{L_2(\mu)} \leq r(A)$, on

$\Omega_{\mathcal{I}}$ we have

$$\begin{aligned}
P\mathcal{L}_{\hat{f}_N} &\leq P_{\mathcal{I}}\mathcal{L}_{\hat{f}_N} + \frac{1}{2A(1+L)}r^2(A) = \frac{N}{|\mathcal{I}|}P_N\mathcal{L}_{\hat{f}_N} - \frac{|\mathcal{O}|}{|\mathcal{I}|}P_{\mathcal{O}}\mathcal{L}_{\hat{f}_N} + \frac{1}{2A(1+L)}r^2(A) \\
&\leq -\frac{|\mathcal{O}|}{|\mathcal{I}|}P_{\mathcal{O}}\mathcal{L}_{\hat{f}_N} + \frac{1}{2A(1+L)}r^2(A) \\
&\leq L\frac{|\mathcal{O}|}{|\mathcal{I}|}P_{\mathcal{O}}|\hat{f}_N - f^*| + \frac{1}{2A(1+L)}r^2(A) \\
&\leq L\frac{|\mathcal{O}|}{|\mathcal{I}|}\left(\|\hat{f}_N - f^*\|_{L_2(\mu)} + \frac{1}{2A(1+L)}\sqrt{\frac{|\mathcal{I}|}{|\mathcal{O}|}}r^2(A)\right) + \frac{1}{2A(1+L)}r^2(A) \\
&\leq L\frac{|\mathcal{O}|}{|\mathcal{I}|}\left(r(A) + \frac{1}{2A(1+L)}\sqrt{\frac{|\mathcal{I}|}{|\mathcal{O}|}}r^2(A)\right) + \frac{1}{2A(1+L)}r^2(A) \\
&\leq L\frac{|\mathcal{O}|}{|\mathcal{I}|}r(A) + \frac{1}{2A}r^2(A) \\
&< \frac{1}{A}r^2(A)
\end{aligned}$$

where we used the fact that $P_N\mathcal{L}_{\hat{f}_N} \leq 0$, that we work on $\Omega_{\mathcal{O}}$ and the inequality $|\mathcal{O}| < (1/2AL)|\mathcal{I}|r(A)$.

5.8.2 Proof Theorem 5.2

The proof is very similar to the one of Theorem 5.1. We present only the stochastic argument. The deterministic argument can be simply obtained by reproducing line by line the proof of Theorem 5.1.

Theorem 5.12 (Theorem 2.6, (Koltchinskii, 2011a)). *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

$$\begin{aligned}
\Omega := \Big\{ &\forall f \in F : \|f - f^*\|_{L_2} \leq \max(1, \sqrt{LM})r^b(A), \\
&\left|(P - P_{\mathcal{I}})\mathcal{L}_f\right| \leq \frac{\max(1, LM)(r^b(A))^2}{2A(L+1)} \\
&\text{and} \quad \left|(P - P_{\mathcal{O}})|f - f^*|\right| \leq \frac{\max(1, LM)(r^b(A))^2}{2A(L+1)}\Big\}
\end{aligned}$$

Lemma 5.6. *Grant Assumptions 5.1, 5.3, 5.2 and 5.6 with the complexity parameter $r^b(\cdot)$. Then, the event Ω holds with probability larger than*

$$1 - 2\exp\left(-\frac{|\mathcal{I}|(r^b(A))^2}{36A^2(L+1)^2}\right)$$

Proof. Let $\mathcal{F} = \{f \in F,\ \|f - f^*\|_{L_2} \leq \max(1, \sqrt{LM}) r^b(A)\}$. Let $(\sigma_i)_{i=1}^N$ be i.i.d Rademacher random variables independent to $(X_i, Y_i)_{i=1}$, from the symmetrization and contraction Lemmas (see (Ledoux and Talagrand, 2013)) we get

$$\mathbb{E} \sup_{f \in \mathcal{F}} |(P_{\mathcal{I}} - P)\mathcal{L}_f| \leq 4L\mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \sigma_i (f - f^*)(X_i) \leq \max(1, LM) \frac{(r^b(A))^2}{8A(L+1)}$$

where we used the Definition 5.3 of $r^b_{\mathcal{I}}(\cdot)$ and the fact that $r^b(A) \geq r^b_{\mathcal{I}}(A)$ for all $A > 0$. From Assumption 5.6, any function f in $\mathcal{F}$, $|\mathcal{L}_f(x, y)| \leq LM$ for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$. For any $t > 0$, it follows from Theorem 5.12 that for any function f in $\mathcal{F}$

$$|(P_{\mathcal{I}} - P)\mathcal{L}_f| \leq \max(1, LM) \frac{(r^b(A))^2}{8A(L+1)} + \frac{LMt}{N} + \sqrt{\frac{2t}{|\mathcal{I}|} \left(\max(1, LM)(r^b(A))^2 + 2LM \max(1, LM) \frac{(r^b(A))^2}{8A(L+1)} \right)} .$$

Since $A, L \geq 1$, taking $t = (|\mathcal{I}|(r^b(A))^2)(36A^2(L+1)^2)$ concludes the proof for the informative data $\mathcal{I}$. For the outliers $\mathcal{O}$, we used the same arguments since from Assumption 5.6, any function f in $\mathcal{F}$, $|f(x) - f^*(x)| \leq M$ for all $x \in \mathcal{X}$.

■

5.8.3 Proof Theorem 5.4

Let $\tilde{r}(\cdot, \cdot)$ such that for all $A, \rho > 0$, $\tilde{r}(A, \rho) \geq \tilde{r}_{\mathcal{I}}(A, \rho)$ and let ρ^* satisfying the $A, \tilde{r}$-sparsity equation with A verifying assumption 5.8

The proof is split into two parts and is very similar as the one of Theorem 5.1. First we identify a stochastic argument holding with large probability. Then, we show on that event that $\|\hat{f}^{\lambda}_N - f^*\|_{L_2(\mu)} \leq \tilde{r}(A, \rho^*)$ and $\|\hat{f}^{\lambda}_N - f^*\| \leq \rho^*$. Then, at the very end of the proof we will control the excess risk $P\mathcal{L}\hat{f}^{\lambda}_N$ where $\hat{f}^{\lambda}_N$ is defined in equation (5.12). Let us fix $\lambda = 41\tilde{r}^2(A, \rho^*)/(112A\rho^*)$.

Stochastic arguments The stochastic part is the same as the one in the proof of Theorem 5.1 where a localization with respect to the regularization norm is added. First we identifiate the stochastic event onto which the proof easily follows. Let,

$$\Omega_{\mathcal{I}} = \Big\{ \forall f \in F \cap \big(f^* + \rho^* B \cap \tilde{r}(A, \rho^*) B_{L_2(\mu)}\big) : \Big| \big(P - P_{\mathcal{I}}\big)\big(\ell_f - \ell_{f^*}\big) \Big| \leq \frac{1}{4A(1+L)} \tilde{r}^2(A, \rho^*) \Big\} \tag{5.43}$$

$$\Omega_{\mathcal{O}} = \Big\{ \forall f \in F \cap \big(f^* + \rho^* B \cap \tilde{r}(A, \rho^*) B_{L_2(\mu)}\big) : \Big| \big(P - P_{\mathcal{O}}\big)|f - f^*| \Big| \leq \frac{1}{4A(1+L)} \sqrt{\frac{|\mathcal{I}|}{|\mathcal{O}|}} \tilde{r}^2(A, \rho^*) \Big\} , \tag{5.44}$$

where we recall that B is the unit ball induced by the regularization norm $\|\cdot\|$. Finally, set $\Omega = \Omega_{\mathcal{I}} \cap \Omega_{\mathcal{O}}$

Lemma 5.7. *Grant Assumptions 5.1, 5.3, 5.2, 5.4 and 5.8 with $\tilde{r}(\cdot,\cdot)$. Then the event Ω holds with probability larger than*

$$1 - 2\exp\left(- c\frac{|\mathcal{I}|\tilde{r}^2(A,\rho^*)}{LBA(L+1)}\right) \tag{5.45}$$

Proof. The proof is exactlty 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 $\tilde{r}_{\mathcal{I}}(A,\rho)$ from Equation (5.4). ∎

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

$$P_N\mathcal{L}_f^\lambda = P_N(\ell_f - \ell_{f^*}) + \lambda(\|f\| - \|f^*\|) \tag{5.46}$$

Let $\mathcal{B} = \rho^* B \cap \tilde{r}(A,\rho^*)B_{L_2(\mu)}$. From the definition of $\hat{f}_N^\lambda$, we have $P_N\mathcal{L}_{\hat{f}_N^\lambda}^\lambda \leq 0$. To show that $\hat{f}_N^\lambda \in F \cap \big(f^* + \mathcal{B}\big)$ it is sufficient to show that for all functions $f \in F\backslash\big(f^* + \mathcal{B}\big)$ we have $P_N\mathcal{L}_f^\lambda > 0$. Let f in $F\backslash\big(f^* + \mathcal{B}\big)$. 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 $f_1 \in \partial(f^* + \mathcal{B})$ where $\partial(f^* + \mathcal{B})$ denotes the border of $f^* + \mathcal{B}$. Using the same convex argument as the one in the proof of Theorem 5.1 we obtain:

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

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 $f_1 \in F \cap (f^* + \mathcal{B})$. By definition of $\mathcal{B}$, there are two different cases: 1) $\|f_1 - f^*\| = \rho^*$ and $\|f_1 - f^*\|_{L_2} \leq \tilde{r}(A,\rho^*)$ and 2) $\|f_1 - f^*\| \leq \rho^*$ and $\|f_1 - f^*\|_{L_2} = \tilde{r}(A,\rho^*)$. In the first case 1), the sparsity equation will help us to show that $P_N\mathcal{L}_{f_1}^\lambda > 0$ while in case 2) it will be the local Bernstein condition. Let us begin by the case where $\|f_1 - f^*\| = \rho^*$ and $\|f_1 - f^*\|_{L_2} \leq \tilde{r}(A,\rho^*)$.

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

On $\Omega_{\mathcal{I}}$ (see Equation (5.43)) it follows that

$$P_{\mathcal{I}}\mathcal{L}_{f_1} \geq P\mathcal{L}_{f_1} - \frac{1}{4A(1+L)}\tilde{r}^2(A,\rho^*) \geq -\frac{1}{4A(1+L)}\tilde{r}^2(A,\rho^*) \tag{5.47}$$

Moreover, from assumption 5.3 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 (5.44)), we get

$$-LP_{\mathcal{O}}\mathcal{L}_{f_1} \geq -L\tilde{r}(A,\rho^*) - \frac{L}{4A(1+L)}\sqrt{\frac{|\mathcal{I}|}{|\mathcal{O}|}}\tilde{r}^2(A,\rho^*) \ . \tag{5.48}$$

Since $|\mathcal{O}| \leq |\mathcal{I}|$, from Equations (5.47), (5.48) it follows

$$P_N\mathcal{L}_{f_1} \geq -\frac{1}{4A}\frac{|\mathcal{I}|}{N}\tilde{r}^2(A,\rho^*) - \frac{|\mathcal{O}|L}{N}\tilde{r}(A,\rho^*)$$

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

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

As the latter result holds for all $v \in f^* + (\rho^*/20)B$ and $g \in \partial\|\cdot\|(v)$, since $f_1 - f^* \in \rho^* S \cap \tilde{r}(A,\rho^*)B_{L_2(\mu)}$, we get

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

Here, the last inequality holds because ρ^* satisfies the sparsity equation. Finally we have

$$P_N\mathcal{L}^{\lambda}_{f_1} \geq -\frac{1}{4A}\frac{|\mathcal{I}|}{N}\tilde{r}^2(A,\rho^*) - \frac{|\mathcal{O}|L}{N}\tilde{r}(A,\rho^*) + \frac{7\lambda\rho^*}{10}$$

From the choice of $\lambda = 41\tilde{r}^2(A,\rho^*)/(112A\rho^*) \geq 41|\mathcal{I}|\tilde{r}^2(A,\rho^*)/(112AN\rho^*)$ we get

$$P_N\mathcal{L}^{\lambda}_{f_1} \geq \frac{1}{160A}\frac{|\mathcal{I}|}{N}\tilde{r}^2(A,\rho^*) - \frac{|\mathcal{O}|L}{N}\tilde{r}(A,\rho^*) > 0$$

when $|\mathcal{O}| < 1/(160AL)|\mathcal{I}|\tilde{r}(A,\rho^*)$.

Let us turn to the second case 2) $\|f_1 - f^*\| \leq \rho^*$ and $\|f_1 - f^*\|_{L_2(\mu)} = \tilde{r}(A,\rho^*)$. On $\Omega_{\mathcal{I}}$ (see Equation (5.43)) and from assumption 5.8 it follows that

$$P_{\mathcal{I}}\mathcal{L}_{f_1} \geq \left(\frac{1}{A} - \frac{1}{4A(1+L)}\right)\tilde{r}^2(A,\rho^*) \ .$$

With the same reasoning as the one in case 1) we get

$$P_{\mathcal{O}}\mathcal{L}_{f_1} \geq -L\frac{|\mathcal{O}|}{N}\tilde{r}(A,\rho^*) - \frac{L}{4A(1+L)}\sqrt{\frac{|\mathcal{I}|}{|\mathcal{O}|}}\tilde{r}^2(A,\rho^*) \ .$$

As $|\mathcal{O}| \leq |\mathcal{I}|$ and $\|f_1\| - \|f^*\| \geq -\|f_1 - f^*\| \geq -\rho^*$, it follows that

$$P_N\mathcal{L}^{\lambda}_{f_1} \geq \frac{3}{4A}\frac{|\mathcal{I}|}{N}\tilde{r}^2(A,\rho^*) - \lambda\rho^* - L\frac{|\mathcal{O}|}{N}\tilde{r}(A,\rho^*) \ .$$

Since $|\mathcal{I}| \geq N/2$ we get $\lambda < 82|\mathcal{I}|\tilde{r}^2(A,\rho^*)/(112AN\rho^*)$ and thus

$$P_N\mathcal{L}^{\lambda}_{f_1} \geq \frac{1}{56A}\tilde{r}^2(A,\rho^*) - L\frac{|\mathcal{O}|}{N}\tilde{r}(A,\rho^*) > 0 \ .$$

when $|\mathcal{O}| < 1/(56AL)|\mathcal{I}|\tilde{r}(A,\rho^*)$

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

$$P\mathcal{L}_{\hat{f}^{\lambda}_N} \leq P_{\mathcal{I}}\mathcal{L}_{\hat{f}^{\lambda}_N} + \frac{1}{4A(1+L)}\tilde{r}^2(A,\rho^*)$$

Moreover we have

$$\begin{aligned}
P_{\mathcal{I}}\mathcal{L}_{\hat{f}^{\lambda}_N} &= \frac{N}{|\mathcal{I}|}P_N\mathcal{L}_{\hat{f}^{\lambda}_N} - \frac{|\mathcal{O}|}{|\mathcal{I}|}P_{\mathcal{O}}\mathcal{L}_{\hat{f}^{\lambda}_N} = \frac{N}{|\mathcal{I}|}P_N\mathcal{L}^{\lambda}_{\hat{f}^{\lambda}_N} + \lambda\frac{N}{|\mathcal{I}|}(\|f^*\| - \|\hat{f}^{\lambda}_N\|) - \frac{|\mathcal{O}|}{|\mathcal{I}|}P_{\mathcal{O}}\mathcal{L}_{\hat{f}^{\lambda}_N} \\
&\leq 2\lambda\rho^* + L\frac{|\mathcal{O}|}{|\mathcal{I}|}P_{\mathcal{O}}|\hat{f}^{\lambda}_N - f^*| \\
&\leq 2\lambda\rho^* + L\frac{|\mathcal{O}|}{|\mathcal{I}|}\left(\|\hat{f}^{\lambda}_N - f^*\|_{L_2(\mu)} + \frac{1}{4A(1+L)}\sqrt{\frac{|\mathcal{I}|}{|\mathcal{O}|}}\tilde{r}^2(A,\rho^*)\right) \\
&\leq \left(\frac{82}{112A} + \frac{L}{4A(1+L)}\right)\tilde{r}^2(A,\rho^*) + L\frac{|\mathcal{O}|}{|\mathcal{I}|}\tilde{r}(A,\rho^*) \\
&< \left(\frac{82}{112A} + \frac{L}{4A(1+L)} + \frac{1}{160A}\right)\tilde{r}^2(A,\rho^*)
\end{aligned}$$

where we used the fact that $P_N\mathcal{L}^{\lambda}_{\hat{f}^{\lambda}_N} \leq 0$ and the inequality $|\mathcal{O}| < 1/(160AL)|\mathcal{I}|\tilde{r}(A,\rho^*)$.

5.8.4 Proof Theorem 5.5

The proof consists in taking the stochastic argument from the proof of Theorem 5.2 (and adding the localization with respect to the regularization norm) and the deterministic argument from the proof of Theorem 5.4

Chapter 6

Benign overfitting in the large deviation regime

In this chapter, we investigate the benign overfitting phenomenon in the large deviation regime where the bounds on the prediction risk hold with probability $1 - e^{-\zeta n}$, for some absolute constant ζ. We prove that these bounds can converge to 0 for the quadratic loss. We obtain this result by a new analysis of the interpolating estimator with minimal Euclidean norm, relying on a preliminary localization of this estimator with respect to the Euclidean norm. This new analysis complements and strengthens particular cases obtained in (Bartlett et al., 2019) for the square loss and is extended to other loss functions. To illustrate this, we also provide excess risk bounds for the Huber and absolute losses, two widely spread losses in robust statistics.

6.1 Introduction

In this paper, we consider Gaussian regression problems where one observes a dataset D_n of i.i.d. random vectors (x_i, y_i), $i \in \{1, \dots, n\}$ such that $y_i = \langle x_i, \beta^* \rangle + \xi_i$, where $\beta^* \in \mathbb{R}^p$ is an unknown vector, $x \sim \mathcal{N}(0, \Sigma) \in \mathbb{R}^p$ and $\xi \sim \mathcal{N}(0, \sigma^2) \in \mathbb{R}$ are independent random variables. Defining the matrix $\boldsymbol{X}$ with lines x_i^T and the vector $\boldsymbol{Y} = (y_1, \dots, y_n)^T \in \mathbb{R}^n$, the set of least-squares estimators is defined by

$$\hat{\beta} \in \operatorname*{argmin}_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} (y_i - \langle x_i, \beta \rangle)^2 = \operatorname*{argmin}_{\beta \in \mathbb{R}^p} \|\boldsymbol{X}\beta^* - \boldsymbol{Y}\|_2^2 \ .$$

The solutions of this problem are $\hat{\beta} = \boldsymbol{X}^g \boldsymbol{Y}$, where $\boldsymbol{X}^g$ is any pseudo-inverse of $\boldsymbol{X}$. When the dimension p of β is smaller than n, the least-squares estimator is typically unique and has a risk of order $O(\sigma^2 p/n)$, which deteriorates with the dimension p. This deterioration is unavoidable in general, a phenomenon known as the "curse of dimensionality" in statistical textbooks.

To bypass this issue, statisticians have focused on situations where β^* satisfies some sparsity conditions, meaning that it belongs, or is close, to a known set $\mathcal{S}$ of small dimensional subspaces $S \subset \mathbb{R}^p$. In many of these situations, least-squares estimators can be improved, by considering minimizers of regularized least-squares criteria of the form $\|\boldsymbol{X}\beta - \boldsymbol{Y}\|_2^2 + \Omega(\beta)$. Several examples of such procedures have been studied in the literature. Among the most popular ones, one can mention ridge regression (Hoerl and Kennard, 1970; Casella, 1980), the LASSO (Tibshirani, 1996; Van de Geer et al., 2008; Bickel et al., 2009) and the elastic net (Zou and Hastie, 2005; De Mol et al., 2009). Regularization ensures that both the prediction risk

$$\mathbb{E}[\langle x, \hat{\beta} - \beta^* \rangle^2 | D_n] = (\hat{\beta} - \beta^*)^T \Sigma (\hat{\beta} - \beta^*) = \|\Sigma^{1/2}(\hat{\beta} - \beta^*)\|_2^2$$

and the estimation risk $\|\hat{\beta} - \beta^*\|_2^2 = (\hat{\beta} - \beta^*)^T(\hat{\beta} - \beta^*)$ are controlled. These results hold even if $p \geqslant n$ provided that β^* is close to a linear subspace $S \subset \mathbb{R}^p$ with dimension $s < n$.

When the dimension $p \geqslant n$, the set of least-squares estimators is typically infinite. Actually, the matrix $\boldsymbol{X}$ in this case has typically full rank (and a non trivial kernel) and any solution in the set $\{\boldsymbol{X}^g \boldsymbol{Y}\}$, where $\boldsymbol{X}^g$ describes all pseudo-inverses of $\boldsymbol{X}$ satisfy $\boldsymbol{X}\boldsymbol{X}^g\boldsymbol{Y} = \boldsymbol{Y}$. In other words, in large dimension, least-squares estimators interpolate data. This kind of behavior is typically undesirable in statistics, as the estimators clearly overfit the observed dataset, and have usually poor generalization abilities. However, and perhaps counter-intuitively, it turns out that, when the dimension p is large in front of n, the risk of prediction can become smaller for some of these solutions. This interesting phenomenon has given rise to a rapidly growing literature these last months, see (Belkin et al., 2019a,b, 2018a,b; Bunea et al., 2020; Feldman, 2019; Liang and Rakhlin, 2018; Mei and Montanari, 2019). This success is not surprising as many algorithms in machine learning require to fit a huge number of parameters with a smaller number of data. The most famous examples are neural networks for which it has been repeatedly observed empirically that enlarging the network, hence, the number of parameters, may help to improve prediction performance (Advani

and Saxe, 2017; Belkin et al., 2019a; Zhang et al., 2016). Of course, linear regression is much simpler than neural networks and the results proved here are not sufficient to explain the amazing prediction properties of these algorithms, but it is interesting to understand when and how high dimension helps prediction, at least in this simpler example. Moreover, several recent works have shown that the analysis of linear models can be relevant for over-parametrized neural networks. A reason is that, when neural networks are trained by gradient descent properly initialized, they are well approximated by a linear model in a Hilbert space. This method is known as *neural tangent kernel* approach (Jacot et al., 2018; Bietti and Mairal, 2019; Arora et al., 2019; Lee et al., 2019). Understanding the generalization of over-parametrized linear models could therefore be seen as a first step in the direction of understanding deep learning.

In this paper, we consider more precisely the problem of (Bartlett et al., 2019) where the least-squares solution with minimal Euclidean norm is analysed. It is well known that this solution is $\hat{\beta} = \boldsymbol{X}^+\boldsymbol{Y}$, where $\boldsymbol{X}^+$ is the Moore-Penrose pseudo inverse of $\boldsymbol{X}$. Our main results complement those in (Bartlett et al., 2019) in the following sense. First, our results are derived in the large deviation regime, meaning that they hold with probability $1 - e^{-\zeta n}$, for some absolute constant ζ. This regime is considered in (Bartlett et al., 2019) but the bounds there don't converge to 0 as $n \to \infty$. On the contrary, our bounds can converge to 0 under proper assumptions on the spectrum of the covariance matrix $\Sigma = \mathbb{E}[xx^T]$. These assumptions involve the rest of the series of singular values of the matrix Σ, $r_{k^*}(\Sigma) = \sum_{k=k^*}^{p} \lambda_i(\Sigma)$ for a well chosen index k^* as in (Bartlett et al., 2019). The index k^* in our result is typically slightly larger than the one in (Bartlett et al., 2019) by a logarithmic factor, see (6.4) for a definition of k^* and the discussion at the end of Section 6.3.1 for a precise comparison between the k^* in a particular example. Besides considering the large deviation regime, our new bounds improve those of (Bartlett et al., 2019) in typical examples where benign overfitting occurs, see the discussion following Corollary 6.1. These improvements are made possible by a new analysis of the estimator $\hat{\beta}$, that relies on preliminary results showing that dimension may help to localize this estimator with respect to the estimation norm $\|\hat{\beta} - \beta\|_2$, see Theorem 6.3. This localization allows, for example, to prove rates of convergence that can be as fast as $1/n$ for this estimator, while the bounds in (Bartlett et al., 2019) only allow to reach $1/\sqrt{n}$. Our bounds exhibit a phase transition of the rates of convergence when the signal to noise ratio $\mathrm{SNR} = \|\beta^*\|^2/\sigma^2$ becomes larger than a threshold $t = n/r_{k^*}(\Sigma)$ (this threshold typically grows to infinity in the examples). When $\mathrm{SNR} > t$, the prediction risk of the estimator satisfies, in the large deviation regime, $\|\Sigma^{1/2}(\hat{\beta} - \beta^*)\|_2^2 \lesssim \|\beta^*\|^2 \mathrm{Tr}(\Sigma)/n$. This rate can be exponentially better than the one in (Bartlett et al., 2019) for some spectrum of the covariance matrix Σ, even if it holds with probability $1 - e^{-\zeta n}$ in our result and with constant probability in (Bartlett et al., 2019) (see the example following Corollary 6.1). On the other hand, when the SNR is too low, $\mathrm{SNR} \leq t$, these rates deteriorate into $\|\Sigma^{1/2}(\hat{\beta} - \beta^*)\|_2^2 \lesssim \sigma^2 + k^*/n$. In this case, our rates improve those of (Bartlett et al., 2019) which are always larger than $\sigma^2 k^*$ in the large deviation regime and actually met the optimal rate σ^2 as proved in (Lecué and Mendelson, 2013, Theorem A').

Besides the least-squares loss, our new strategy can be easily applied to analyse the excess risk of interpolating estimators with respect to other loss functions. This extension was mentioned as a relevant conjecture in (Bartlett et al., 2019). We illustrate this by providing a short analysis of the excess risk of $\hat{\beta}$ with respect to the Huber loss and the absolute loss, two widely spread methods in robust statistics. The bounds obtained on the excess risk of $\hat{\beta}$ with respect to these losses involve the same quantities as for the quadratic loss. They are gathered in Theorem 6.2.

The remainder of the paper is decomposed as follows. Section 6.2 sets the main notations and recall the construction of the estimator $\hat{\beta}$. Section 6.3 gathers the main results of the paper, the upper bounds on the excess risk of the estimator β with respect to the quadratic, absolute and Huber losses. The proofs of these results are gathered in Section 6.4.

6.2 Setting

Let $(x, y), (x_i, y_i)_{i\in\{1,\dots,n\}}$ denote i.i.d random vectors generated according to the following Gaussian linear model,

$$y = x^T\beta^* + \xi \ , \tag{6.1}$$

where $\beta^* \in \mathbb{R}^p$ is the signal of interest, the design x is a Gaussian vector $x \sim \mathcal{N}(0, \Sigma) \in \mathbb{R}^p$ and the noise ξ is a Gaussian random variable $\xi \sim \mathcal{N}(0, \sigma^2)$, independent of x. Let $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ denote the matrix with lines $x_1^T, \cdots, x_n^T$. Let $\boldsymbol{Y} = (y_1, \cdots, y_n)^T \in \mathbb{R}^n$ and $\boldsymbol{\xi} = (\xi_1, \dots, \xi_n)^T$. Using these notations, the dataset $D_n = \{(x_1, y_1), \dots, (x_n, y_n)\}$ can be represented in the matrix form as

$$\boldsymbol{Y} = \boldsymbol{X}\beta^* + \boldsymbol{\xi} \ .$$

The set of interpolating vectors $H_n \subset \mathbb{R}^p$ is defined as $H_n = \{\beta \in \mathbb{R}^p : \boldsymbol{X}\beta = \boldsymbol{Y}\}$. We analyse the estimator defined as the interpolating vector with minimal Euclidean norm, that is

$$\hat{\beta} = \operatorname*{argmin}_{\beta\in H_n} \|\beta\|_2 \ , \tag{6.2}$$

where $\|\cdot\|_2$ denotes the Euclidean norm in $\mathbb{R}^p$. This estimator is defined only when the set H_n is non-empty. In general, this occurs only when $\boldsymbol{X}$ has full rank n, which holds almost surely when the dimension p is larger than the number of observations n, provided that Σ has rank at least n. In the following, we assume therefore that $p \geq 4n$ and that Σ has rank at least n. The constant 4 has no particular meaning here, it could be replaced by any constant strictly larger than 1 without affecting the results.

Our main results give upper bounds on the prediction loss of $\hat{\beta}$. Let $\ell : \mathbb{R} \times \mathbb{R} \mapsto \mathbb{R}^+$ denotes a loss function such that $\ell(y, y) = 0$ for all $y \in \mathbb{R}$ and $\ell(y, y') > 0$ if $y \neq y'$. It is also assumed that the function $y \mapsto \ell(y, y')$ is convex for any $y \in \mathbb{R}$. In the first part of the paper, ℓ will be the square loss $\ell(y, y') = (y - y')^2$. Other losses will be considered in Section 6.3.2. For any $\beta \in \mathbb{R}^p$ and any

$(u,v) \in \mathbb{R}^p \times \mathbb{R}$, let $\ell_\beta(u,v) = \ell(\langle u, \beta \rangle, v)$ and let $\mathcal{L}_\beta(u,v) = \ell_\beta(u,v) - \ell_{\beta^*}(u,v)$. For any function $f : \mathbb{R}^p \times \mathbb{R} \to \mathbb{R}$, let $Pf = \mathbb{E}[f(x,y)]$. The excess risk is then defined as:

$$\mathbb{E}\left[\ell\left(\langle x, \hat{\beta}\rangle, y\right) - \ell\left(\langle x, \beta^*\rangle, y\right) | D_n\right] = P(\ell_{\hat{\beta}} - \ell_{\beta^*}) = P\mathcal{L}_{\hat{\beta}} \ . \tag{6.3}$$

As usual, the expectation is taken over the random variables (x,y) only, so the excess risk is a random variable. In this paper, we provide risk bounds for the estimator $\hat{\beta}$ that hold in the large deviation regime. This means that we build deterministic upper bounds r_n on $P\mathcal{L}_{\hat{\beta}}$ such that $\mathbb{P}(P\mathcal{L}_{\hat{\beta}} > r_n) \leqslant -\exp(\zeta n)$, for some absolute constant ζ.

For any symmetric matrix $A \in \mathbb{R}^{n\times n}$, we denote by $\lambda_1(A) \geq \cdots \geq \lambda_n(A)$ its eigenvalues in the non-increasing order and by $r_k(A) = \sum_{i=k}^n \lambda_i(A)$. More generally, for any matrix $B \in \mathbb{R}^{n\times p}$, we denote by $\sigma_1(B) \geq \cdots \geq \sigma_{min}(B) > 0$, its positive singular values in the non-increasing order. The operator norm of B is denoted by $\|B\| = \sigma_1(B)$. For any symmetric positive semi-definite matrix A, let $\|\beta\|_A = \sqrt{\beta^T A \beta}$. Let $S(r)$ (resp. $S_A(r)$) denote the sphere in $\mathbb{R}^p$ with radius r with respect to the Euclidean norm $\|\cdot\|_2$ (resp. with respect to the semi-norm $\|\cdot\|_A$). Define similarly $B(r)$ and $B_A(r)$ to be the balls with radius r. Let also, for any subset $\mathcal{B}$ of $\mathbb{R}^p$, denote by $\beta + \mathcal{B} = \{u \in \mathbb{R}^p : \exists v \in \mathcal{B} \text{ such that } u = \beta + v\}$. All along the paper, c and ζ denote absolute positive constants. Typically, ζ denotes a small constant while c denotes a large one.

6.3 Main results

This section provides our main contributions. Prediction bounds for the square loss are provided in Section 6.3.1 and for other loss functions in Section 6.3.2.

6.3.1 Prediction with least-squares loss

The following theorem is the main result of this paper.

Theorem 6.1. *Let*

$$k^* = \inf\left\{k \in \{1,\cdots,p\} : \frac{r_k(\Sigma)}{\lambda_k(\Sigma)} \geq 32n\log\left(1 + \frac{44}{3}\sqrt{\frac{p\|\Sigma\|}{r_k(\Sigma)}}\right)\right\} \ . \tag{6.4}$$

Let $\zeta > 0$ be an absolute constant. Define the parameter v, the estimation rate ρ and the prediction rate r^* by

$$v = \frac{r_{k^*}(\Sigma)}{32n\lambda_{k^*}(\Sigma)}, \qquad \rho = \|\beta^*\|_2 + \sigma\sqrt{\frac{32n}{r_{k^*}(\Sigma)}} \ , \tag{6.5}$$

$$r^* = \inf\left\{r > 0 : \sum_{i=1}^p r^2 \wedge \lambda_i(\Sigma)\rho^2 \leq \zeta n r^2\right\} \ . \tag{6.6}$$

If $k^* \leq cn$, for $c > 0$ an absolute constant, then, with probability larger than $1 - 7e^{-(v \wedge \zeta)n}$, the estimator $\hat{\beta}$ defined in Equation (6.2) satisfies

$$\|\hat{\beta} - \beta^*\|_2 \leq \rho \qquad \|\Sigma^{1/2}(\hat{\beta} - \beta^*)\|_2 \leq r^* \ .$$

Theorem 6.1 is proved in Section 6.4.2. The estimation bound ρ does not converge to 0, which is not surprising in our high dimensional setting, in absence of sparsity assumption. However, it is interesting to see that it may decrease, up to a certain threshold, with the dimension p. In particular, when the signal to noise ratio $\|\beta^*\|^2/\sigma^2$ is larger that the threshold $n/r_{k^*}(\Sigma)$, $\|\hat{\beta} - \beta^*\|_2$ is at most of order $\|\beta^*\|_2$ when the dimension is large enough.

To discuss the prediction bounds, it is useful to give the following corollary, whose proof is a direct consequence of Theorem 6.1 left as an exercise. The corollary shows a phase transition in the rates of convergence when the signal to noise ratio $\text{SNR} = \|\beta^*\|^2/\sigma^2$ becomes larger than the threshold $t = n/r_{k^*}(\Sigma)$.

Corollary 6.1. *Grant the assumptions and notations of Theorem 6.1,*

- *If the signal to noise ratio is large enough, $\|\beta^*\|_2^2/\sigma^2 \geq n/r_{k^*}(\Sigma)$, the estimator $\hat{\beta}$ defined in Equation (6.2) satisfies, with probability larger than $1 - 7e^{-(v \wedge \zeta)n}$,*

$$\|\hat{\beta} - \beta^*\|_2 \lesssim \|\beta^*\|_2, \qquad \|\Sigma^{1/2}(\hat{\beta} - \beta^*)\|_2^2 \lesssim \|\beta^*\|_2^2 \frac{\text{Tr}(\Sigma)}{n} \ .$$

- *On the other hand, if the signal to noise ratio is too small, $\|\beta^*\|_2^2/\sigma^2 \leq n/r_{k^*}(\Sigma)$, then, the estimator $\hat{\beta}$ defined in Equation (6.2) satisfies, with probability larger than $1 - 7e^{-(v \wedge \zeta)n}$,*

$$\|\hat{\beta} - \beta^*\|_2 \lesssim \sigma \sqrt{\frac{n}{r_{k^*}(\Sigma)}}, \qquad \|\Sigma^{1/2}(\hat{\beta} - \beta^*)\|_2^2 \lesssim \left(\sigma^2 + \frac{k^*}{n}\right) \ .$$

Corollary 6.1 can be used to compare our results with those in (Bartlett et al., 2019).

1. The upper bounds in Corollary 6.1 hold with probability larger than $1 - \exp(-\zeta n)$ and may converge to 0 while those in (Bartlett et al., 2019) are always larger than a constant at these confidence levels.

2. For high signal to noise ratios, $\text{SNR} = \|\beta^*\|_2^2/\sigma^2 > t = n/r_{k^*}(\Sigma)$, Corollary 6.1 improves the results provided in (Bartlett et al., 2019), since the main term in this case here is $\|\beta^*\|_2^2 \text{Tr}(\Sigma)/n$ while it is $\|\beta^*\|_2^2 \sqrt{\text{Tr}(\Sigma)/n}$ in this paper.

3. For small signal to noise ratios, $\text{SNR} < t$, our rates are of order $\sigma^2 + k^*/n$, which improve the result of (Bartlett et al., 2019) at confidence levels $e^{-\zeta n}$. An interesting feature of the results in (Bartlett et al., 2019) is that it provides upper bounds that can converge to 0 at smaller confidence levels. On the other hand, (Lecué and Mendelson, 2013, Theorem A') shows that σ^2 is the optimal rate that can hold with probability larger than $1 - \exp(-\zeta n)$.

4. The parameter k^* in Theorem 6.1 is slightly larger in general than the one in (Bartlett et al., 2019), since they only require that $r_{k^*}(\Sigma)/\lambda_{k^*}(\Sigma) > cn$ while we have an extra logarithmic factor in the definition (6.4).

To illustrate the upper bounds, (Bartlett et al., 2019) provide several examples of "benign matrices" where the different quantities of interest in Theorem 6.1 can easily be computed. We compute the quantities appearing in one these examples now.

Assume that there exist $\epsilon = o(1)$ and $\tau = \Omega(1)$ such that, for any k,

$$\lambda_k(\Sigma) = e^{-k/\tau} + \epsilon, \qquad \text{with} \qquad \tau \log(1/\epsilon) < n, \qquad p = cn \log(1/\epsilon) \ .$$

In this case, for any k and $\gamma = \tau/(1 - e^{-\tau})$,

$$\frac{r_k}{\lambda_k} = \frac{(p-k)\epsilon + \gamma(e^{-k/\tau} - e^{-p/\tau})}{e^{-k/\tau} + \epsilon} \ ,$$
$$\frac{p\|\Sigma\|}{r_k(\Sigma)} = \frac{p}{(p-k)\epsilon + \gamma(e^{-k/\tau} - e^{-p/\tau})} \ .$$

Therefore, for $k = \tau \log(1/\epsilon) < p/2$ and c large enough,

$$\frac{r_k}{\lambda_k} \geq \frac{p\epsilon/2 + \gamma\epsilon}{2\epsilon} \geq \frac{p}{4} \geq 32n \log\left(1 + \frac{44}{3}\sqrt{\frac{2}{\epsilon}}\right) \geq 32n \log\left(1 + \frac{44}{3}\sqrt{\frac{p\|\Sigma\|}{r_k(\Sigma)}}\right) \ .$$

Hence, $k^* \leq \tau \log(1/\epsilon) < n$. Moreover, $r_{k^*}(\Sigma) = \Theta(p\epsilon) = \Theta(n\epsilon \log(1/\epsilon))$ so the threshold t for the SNR ratio is $t = \Omega(1/(\epsilon \log(1/\epsilon)))$. This threshold therefore grow to infinity if $\epsilon \to 0$. As $\mathrm{Tr}(\Sigma) \leq p\epsilon + \tau$ and the parameter $v \gtrsim p\epsilon/(n\epsilon) \gtrsim 1$, Corollary 6.1 shows in this example that, if $\|\beta^*\|^2/\sigma^2 \geq 1/(\epsilon \log(1/\epsilon))$, with probability larger than $1 - e^{-\zeta n}$,

$$\|\Sigma^{1/2}(\hat{\beta} - \beta^*)\|_2^2 \lesssim \|\beta^*\|_2^2 \frac{p\epsilon + \tau}{n} = \|\beta^*\|_2^2 \left(\epsilon \log(1/\epsilon) + \frac{\tau}{n}\right) \ .$$

Our rates of convergence in this example can therefore be, up to logarithmic factors as fast as $\epsilon \vee (1/n)$, while (Bartlett et al., 2019, Theorem 6) gives in this setting a rate $(1/\log(1/\epsilon)) \vee (1/n)$ that is exponentially slower. In addition, let us recall that Corollary 6.1 here shows that the rate $\epsilon \vee (1/n)$ holds with probability $1 - e^{-\zeta n}$ while (Bartlett et al., 2019, Theorem 6) only shows that the logarithmically slower rate $(1/\log(1/\epsilon)) \vee (1/n)$ holds with constant probability.

6.3.2 Extension to other loss functions

The purpose of this section is to show that the analysis developed to prove the main theorem can be easily extended and that the excess risk of $\hat{\beta}$ with respect to other loss functions can be controlled with the same arguments. To illustrate this general principle, we consider two losses, namely, the Huber and absolute losses. Both losses have been used repeatedly in robust statistics. Formally, let $\ell : \mathbb{R} \times \mathbb{R} \mapsto \mathbb{R}^+$ denote one of the following convex loss function:

- The **Huber loss** is defined, for any $u, y \in \mathbb{R}$, by

$$\ell(u,y) = \varphi_H(u-y), \qquad \text{where} \qquad \varphi_H(u) = \begin{cases} \frac{1}{2}u^2 & \text{if } |u| \le \delta \\ \delta|u| - \delta^2/2 & \text{if } |u| > \delta \end{cases} .$$

  Notice that φ_H is $\delta-$Lipschitz.

- The **absolute loss** is defined, for any $u, y \in \mathbb{R}$, by $\ell(u,y) = \varphi_A(u-y)$, where $\varphi_A(u) = |u|$ is 1-Lipschitz.

For both losses, recall that, for any $\beta \in \mathbb{R}^p$, by $P\mathcal{L}_\beta = P[\ell_\beta - \ell_{\beta^*}]$, with $\ell_\beta(x,y) = \ell(\langle x, \beta \rangle, y)$.

Theorem 6.2. *There exist absolute constants ζ, c, c_2 such that the following holds. Let k^*, ρ, v r^* be defined as in Theorem 6.1. If $k^* \le cn$, then*

- *if ℓ is the Huber loss with $\delta = c_2\sigma$, with probability larger $1 - 10e^{-(\zeta \wedge v)n}$,*

$$P\mathcal{L}_{\hat{\beta}} \le c(r^*)^2 \ ,$$

- *if ℓ is the absolute loss, with probability larger $1 - 8e^{-(\zeta \wedge v)n}$,*

$$P\mathcal{L}_{\hat{\beta}} \le cr^* \ .$$

Remark 6.1. *Theorem 6.2 is proved in Section 6.4.3. It shows that the excess risk for the Huber loss is of the same order as the one for the square loss. It is the square root of these rate for the absolute loss. Both results are expected as the same phenomenon appear in small dimension also, see for example (Chinot et al., 2019b).*

6.4 Proofs of the main results

The remaining of the paper is devoted to the proofs of the main results. Section 6.4.1 (resp. 6.4.2) shows the estimation bound (resp. the prediction bounds) in Theorem 6.1.

6.4.1 Proof of the estimation bound of Theorem 6.1

The following theorem establishes the bound on the estimation error in Theorem 6.1. In the following section, this preliminary estimate will be used to "localize" the analysis of the prediction risk of $\hat{\beta}$. This approach is now classical in statistical learning, it has been applied successfully, for example, in (Koltchinskii and Mendelson, 2015; Mendelson, 2014, 2016, 2017).

Theorem 6.3. *There exist absolute constants c and ζ such that the following holds. Let k^*, v and ρ be defined as in Theorem 6.1. If $k^* \le cn$, the estimator $\hat{\beta}$ defined in Equation (6.2) satisfies*

$$\mathbb{P}(\|\hat{\beta} - \beta^*\|_2 \le \rho) \ge 1 - 4\exp(-(v \wedge 1/5)n) \ . \tag{6.7}$$

Proof of Theorem 6.3. The proof starts with the following lemma.

Lemma 6.1. *With probability conditionally on $\boldsymbol{X}$ larger than $1 - e^{-n/2}$,*

$$\|\hat{\beta} - \beta^*\|_2 \leq \|\beta^*\|_2 + 2\sigma\sqrt{\frac{n}{\sigma_n^2(\boldsymbol{X})}} \ . \tag{6.8}$$

Proof of Lemma 6.1. Classical results of linear algebra show that

$$\hat{\beta} = \boldsymbol{X}^+\boldsymbol{Y} = \boldsymbol{X}^+\boldsymbol{X}\beta^* + \boldsymbol{X}^+\boldsymbol{\xi} \ ,$$

where $\boldsymbol{X}^+$ denotes the Moore-Penrose pseudo-inverse of $\boldsymbol{X}$. Therefore,

$$\|\hat{\beta} - \beta^*\|_2 = \|(\boldsymbol{X}^+\boldsymbol{X} - I_p)\beta^* - \boldsymbol{X}^+\boldsymbol{\xi}\|_2 \leq \|\beta^*\|_2 + \|\boldsymbol{X}^+\boldsymbol{\xi}\|_2 \ , \tag{6.9}$$

where the last inequality follows from the triangular inequality and the fact that $\boldsymbol{X}^+\boldsymbol{X} - I_p$ is the projection matrix onto the null-space of $\boldsymbol{X}$. Since $\|\boldsymbol{X}^+\boldsymbol{\xi}\|_2 \leq \|\boldsymbol{X}^+\|\|\boldsymbol{\xi}\|_2$, the function $\boldsymbol{\xi} \mapsto \|\boldsymbol{X}^+\boldsymbol{\xi}\|_2$ is $\|\boldsymbol{X}^+\|$-Lipschitz with respect to the Euclidean norm. From Borell's Gaussian concentration inequality, with probability conditionally on $\boldsymbol{X}$ larger than $1 - \exp(-n/2)$,

$$\|\boldsymbol{X}^+\boldsymbol{\xi}\|_2 \leq \mathbb{E}[\|\boldsymbol{X}^+\boldsymbol{\xi}\|_2|\boldsymbol{X}] + \sigma\|\boldsymbol{X}^+\|\sqrt{n} \ . \tag{6.10}$$

Since $\operatorname{rank}(\boldsymbol{X}) \leq n$, $\|\boldsymbol{X}^+\| \leq \sigma_n^{-1}(\boldsymbol{X})$. Similarly, $\operatorname{rank}\big((\boldsymbol{X}^+)^T\boldsymbol{X}^+\big) \leq \operatorname{rank}(\boldsymbol{X}^+) \leq n$. Therefore, writing $\mathbb{E}[\cdot]$ for $\mathbb{E}[\cdot|\boldsymbol{X}]$,

$$\begin{aligned}
\mathbb{E}\|\boldsymbol{X}^+\boldsymbol{\xi}\|_2 &\leq \big(\mathbb{E}\|\boldsymbol{X}^+\boldsymbol{\xi}\|_2^2\big)^{1/2} \\
&= \big(\mathbb{E}\boldsymbol{\xi}^T(\boldsymbol{X}^+)^T\boldsymbol{X}^+\boldsymbol{\xi}\big)^{1/2} = \sigma\big(\operatorname{Tr}\big((\boldsymbol{X}^+)^T\boldsymbol{X}^+\big)\big)^{1/2} \\
&= \sigma\bigg(\sum_{i=1}^n \lambda_i\big((\boldsymbol{X}^+)^T\boldsymbol{X}^+\big)\bigg)^{1/2} = \sigma\bigg(\sum_{i=1}^n \sigma_i^2\big(\boldsymbol{X}^+\big)\bigg)^{1/2} \\
&= \sigma\bigg(\sum_{i=1}^n \sigma_i^{-2}\big(\boldsymbol{X}\big)\bigg)^{1/2} \leq \sigma\sqrt{\frac{n}{\sigma_n^2(\boldsymbol{X})}} \ .
\end{aligned}$$

Plugging (6.10) and this bound on $\mathbb{E}[\|\boldsymbol{X}^+\boldsymbol{\xi}\|_2|\boldsymbol{X}]$ into (6.9) concludes the proof. ∎

Lemma 6.1 provides a random bound on the estimation error of $\hat{\beta}$. To prove Theorem 6.3, it remains to bound from below, with high probability, the smallest eigenvalue $\sigma_n^2(\boldsymbol{X})$ of $\boldsymbol{X}\boldsymbol{X}^T$. This control is obtained in the following lemma.

Lemma 6.2. *With probability larger than $1 - 2\exp(-p/18) - \exp(-nv)$, we have*

$$\sigma_n(\boldsymbol{X}) \geq \sqrt{\frac{r_{k^*}(\Sigma)}{8}} \ .$$

Proof. The matrix $\boldsymbol{X}^T$ is distributed as $\Sigma^{1/2}G$, where $G \in \mathbb{R}^{p\times n}$ is a random matrix with i.i.d standard Gaussian variables, hence $\sigma_n(\boldsymbol{X}) = \sigma_n(\boldsymbol{X}^T)$ is distributed as $\sigma_n(\Sigma^{1/2}Gx)$. From the Courant-Fischer-Weyl min-max principle, we have

$$\sigma_n(\Sigma^{1/2}Gx) = \min_{x\in S^{n-1}} \|\Sigma^{1/2}Gx\|_2 \ .$$

Let $x \in S^{n-1}$ and $\Lambda = \mathrm{diag}(\lambda_1(\Sigma), \cdots, \lambda_p(\Sigma))$. By the spectral theorem, there exists an orthogonal matrix P such that $\|\Sigma^{1/2}Gx\|_2^2 = \|P\Lambda^{1/2}P^TGx\|_2^2$. Hence, by rotation invariance of Gaussian random vectors, $\|\Sigma^{1/2}Gx\|_2^2$ is distributed as $\|\Lambda^{1/2}Gx\|_2^2$, that is, as $\|x\|_2^2 \sum_{i=1}^p \lambda_i(\Sigma) g_i^2$, where $g_1, \cdots, g_p$ are i.i.d standard Gaussian random variables. As $x \in S^{n-1}$, $\|\Sigma^{1/2}Gx\|_2^2$ is distributed as $\sum_{i=1}^p \lambda_i(\Sigma) g_i^2$. Clearly

$$\sum_{i=1}^{p} \lambda_i(\Sigma) g_i^2 \geqslant \sum_{i=k^*}^{p} \lambda_i(\Sigma) g_i^2 \ .$$

Elementary computations show that, for any i, $\lambda_i(\Sigma) g_i^2$ is sub-exponential (see Definition 6.1) with parameters $(2\sqrt{\lambda_i(\Sigma)}, 4\lambda_i(\Sigma))$. As these variables are independent, by Proposition 6.1, $\sum_{i=k^*}^p \lambda_i(\Sigma) g_i^2$ is sub-exponential with parameters $\big(2\sqrt{r_{k^*}(\Sigma)}, 4\lambda_{k^*}(\Sigma)\big)$. Therefore, by Proposition 6.2, with probability $1 - \exp(-2nv)$,

$$\|\Lambda^{1/2}Gx\|_2^2 \geq \frac{1}{2} r_{k^*}(\Sigma) \ . \tag{6.11}$$

Equation (6.11) holds for any fixed x in the unit sphere S^{n-1}. To obtain uniform deviations, let us introduce an ϵ-net Γ_ε of S^{n-1}. For any $x \in S^{n-1}$, there exists $y \in \Gamma_\varepsilon$ such that $\|x - y\|_2 \leq \varepsilon$. Thus,

$$\|\Sigma^{1/2}Gx\|_2 \geq \|\Sigma^{1/2}Gy\|_2 - \|\Sigma^{1/2}G(x-y)\|_2 \geq \|\Sigma^{1/2}Gy\|_2 - \varepsilon\|\Sigma^{1/2}G\| \ .$$

Since the operator norm is sub-multiplicative, $\|\Sigma^{1/2}G\| \leq \sqrt{\|\Sigma\|}\|G\|$. To bound the operator norm $\|G\|$, we use the following result.

Theorem 6.4. *(Vershynin, 2010)[Theorem 5.35]. Let $p \geq n$ and let G denote a $p \times n$ matrix with independent standard Gaussian entries. For every $0 < \delta \leq 1$, with probability at least $1 - \delta$:*

$$\sqrt{p} - \sqrt{n} - \sqrt{2\log(2/\delta)} \leq \sigma_{min}(G) \leq \sigma_1(G) \leq \sqrt{p} + \sqrt{n} + \sqrt{2\log(2/\delta)} \ . \tag{6.12}$$

From Theorem 6.4, with probability larger that $1 - 2\exp(-p/18)$,

$$\|G\| \leq \sqrt{p} + \sqrt{n} + \sqrt{\frac{2p}{18}} \leqslant \sqrt{p}\left(1 + \frac{1}{2} + \frac{1}{3}\right) = \frac{11\sqrt{p}}{6} \ .$$

It follows that

$$\min_{x \in S^{n-1}} \|\Sigma^{1/2}Gx\|_2 \geq \min_{y \in \Gamma_\varepsilon} \|\Sigma^{1/2}Gy\|_2 - \frac{11}{6}\varepsilon\sqrt{p\|\Sigma\|} \ . \tag{6.13}$$

Hence, for

$$\varepsilon = \frac{6}{44}\sqrt{\frac{r_{k^*}(\Sigma)}{p\|\Sigma\|}} \ ,$$

we have

$$\min_{x \in S^{n-1}} \|\Sigma^{1/2}Gx\|_2 \geq \min_{y \in \Gamma_\varepsilon} \|\Sigma^{1/2}Gy\|_2 - \frac{\sqrt{r_{k^*}(\Sigma)}}{4} \ . \tag{6.14}$$

Taking a union bound in (6.11), we get that, for this value of ε, with probability at least $1 - \exp\big(-2nv + \log(|\Gamma_\varepsilon|)\big)$,

$$\min_{x \in S^{n-1}} \|\Sigma^{1/2}Gx\|_2 \geq \sqrt{r_{k^*}(\Sigma)}\left(\frac{1}{\sqrt{2}} - \frac{1}{4}\right) \geqslant \sqrt{\frac{r_{k^*}(\Sigma)}{8}} \ .$$

A standard volume argument shows that, for every $\varepsilon > 0$, $|\Gamma_\varepsilon| \leq (1 + 2/\varepsilon)^n$. Therefore, the probability estimate is bounded from bellow by

$$1 - \exp\left(-2nv + n \log\left(1 + \frac{44}{3}\sqrt{\frac{p\|\Sigma\|}{r_{k^*}(\Sigma)}} \right) \right) \ .$$

By definition of k^*, this probability is bounded from bellow by

$$1 - \exp(-nv) \ .$$

This concludes the proof of Lemma 6.2. ■

Theorem 6.3 then follows directly from Lemmas 6.1 and 6.2.

■

6.4.2 Proof of the prediction bound in Theorem 6.1

Let $\beta^* + B(\rho) = \{\beta \in \mathbb{R}^p : \|\beta - \beta^*\|_2 \leq \rho\}$. Let $P_n\mathcal{L}_\beta := n^{-1}\sum_{i=1}^n (\ell_\beta(x_i, y_i) - \ell_{\beta^*}(x_i, y_i))$ denote the empirical excess-risk. The proof starts with the following elementary result.

Lemma 6.3. *With probability larger than* $1 - \exp(-n/16)$, $P_n\mathcal{L}_{\hat\beta} \leq -(1/2)\sigma^2$. *Moreover, for any* r^*, *let* $\Omega_{r^*,\rho}$ *denote the following event*

$$\Omega_{r^*,\rho} = \{\forall \beta \in \mathbb{R}^p \text{ such that } \beta - \beta^* \in B(\rho) \setminus B_\Sigma(r^*),\ P_n\mathcal{L}_\beta > -(1/2)\sigma^2\} \ .$$

On the event

$$\Omega_{r^*,\delta} \cap \{\hat\beta - \beta^* \in B(\rho)\} \cap \{P_n\mathcal{L}_{\hat\beta} \leq -(1/2)\sigma^2\} \ ,$$

$\hat\beta - \beta \in B_\Sigma(r^*)$, *that is*

$$\|\Sigma^{1/2}(\hat\beta - \beta^*)\|_2 \leq r^* \ .$$

Proof. Since $\hat\beta \in H_n$, $\langle x_i, \hat\beta\rangle = y_i$ for any $i \in \{1, \ldots, n\}$, so $P_n\ell_{\hat\beta} = 0$ and

$$P_n\mathcal{L}_{\hat\beta} = P_n(\ell_{\hat\beta} - \ell_{\beta^*}) = -P_n\ell_{\beta^*} = -\frac{1}{n}\sum_{i=1}^n \xi_i^2 \ .$$

Since $\xi_i \sim \mathcal{N}(0, \sigma^2)$, from Proposition 6.1, $\sum_{i=1}^n \xi_i^2$ is sub-exponential with parameters $(2\sigma\sqrt{n}, 4\sigma^2)$ and from Proposition 6.2, with probability larger than $1 - \exp(-n/16)$,

$$P_n\mathcal{L}_{\hat\beta} = -\frac{1}{n}\sum_{i=1}^n \xi_i^2 \leq -(1/2)\sigma^2 \ . \tag{6.15}$$

On $\Omega_{r^*,\rho}$ all β such that $\|\beta - \beta^*\|_2 \leqslant \rho$ and $\|\beta - \beta^*\|_\Sigma > r^*$ satisfy $P_n\mathcal{L}_\beta > -(1/2)\sigma^2$. Therefore, on $\Omega_{r^*,\rho}$, if $\|\hat\beta - \beta^*\|_2 \leqslant \rho$ and $P_n\mathcal{L}_{\hat\beta} \leq -(1/2)\sigma^2$, $\hat\beta$ cannot satisfy $\|\hat\beta - \beta^*\|_\Sigma > r^*$. Hence,

$$\{\hat\beta - \beta^* \in B_\Sigma(r^*)\} \supset \Omega_{r^*,\rho} \cap \{\hat\beta - \beta^* \in B(\rho)\} \cap \{P_n\mathcal{L}_{\hat\beta} \leq -(1/2)\sigma^2\} \ .$$

■

By Lemma 6.3, to bound the excess risk of $\hat{\beta}$, it is sufficient to show that r^* defined in (6.6) is such that, with high probability

$$\inf_{\beta:\beta-\beta^*\in B(\rho)\setminus B_\Sigma(r^*)}\{P_n\mathcal{L}_\beta\} > -(1/2)\sigma^2 \ . \tag{6.16}$$

Theorem 6.5. *There exists an absolute constant ζ such that, with probability larger than $1-2e^{-\zeta n}$,*

$$\inf_{\beta:\beta-\beta^*\in B(\rho)\setminus B_\Sigma(r^*)}\{P_n\mathcal{L}_\beta\} > -(1/2)\sigma^2 \ ,$$

where r^* is the complexity parameter defined in (6.6)*.*

By Lemma 6.3 and Theorem 6.3, this means that, with probability larger than $1-6e^{-\zeta n}-e^{-\upsilon n}$,

$$\|\hat{\beta}-\beta^*\|_2 \le \rho, \qquad \|\Sigma^{1/2}(\hat{\beta}-\beta^*)\|_2 \le r^* \ . \tag{6.17}$$

Proof of Theorem 6.5. Let $\beta \in \beta^* + B(\rho)\setminus B_\Sigma(r^*)$ and denote by $r = \|\Sigma^{1/2}(\beta-\beta^*)\|_2$, so $r > r^*$ and

$$\beta-\beta^* \in H_{r,\rho} = B(\rho)\cap S_\Sigma(r) \ .$$

Recall that, for any $\beta\in\mathbb{R}^p$, as

$$\langle x_i,\beta\rangle - y_i = \langle x_i,\beta-\beta^*\rangle + \langle x_i,\beta^*\rangle - y_i = \langle x_i,\beta-\beta^*\rangle - \xi_i \ ,$$

we have

$$\begin{aligned} P_n\mathcal{L}_\beta &= \frac{1}{n}\sum_{i=1}^n \big(\langle x_i,\beta\rangle - y_i\big)^2 - \big(\langle x_i,\beta^*\rangle - y_i\big)^2 \\ &= \frac{1}{n}\sum_{i=1}^n \langle x_i,\beta-\beta^*\rangle^2 - \frac{2}{n}\sum_{i=1}^n \xi_i\langle x_i,\beta-\beta^*\rangle \ . \end{aligned} \tag{6.18}$$

Write now $\alpha = r^*/r \in (0,1)$ and $\beta_0 = \beta^* + \alpha(\beta-\beta^*)$, so

$$P_n\mathcal{L}_\beta = \alpha^{-2}\frac{1}{n}\sum_{i=1}^n \langle x_i,\beta_0-\beta^*\rangle^2 - \alpha^{-1}\frac{2}{n}\sum_{i=1}^n \xi_i\langle x_i,\beta_0-\beta^*\rangle \ . \tag{6.19}$$

By definition, $\|\Sigma^{1/2}(\beta_0-\beta^*)\|_2 = r^*$ and $\|\beta_0-\beta^*\| \le \alpha\rho$, that is, $\beta_0-\beta^* \in H_{r^*,\alpha\rho} = S_\Sigma(r^*)\cap B(\alpha\rho)$. Define then

$$\begin{aligned} Q_{r,\rho} &= \sup_{\beta-\beta^*\in H_{r,\rho}}\left|\frac{1}{n}\sum_{i=1}^n \langle x_i,\beta-\beta^*\rangle^2 - \mathbb{E}\langle x_i,\beta-\beta^*\rangle^2\right| \ , \\ M_{r,\rho} &= \sup_{\beta-\beta^*\in H_{r,\rho}}\left|\frac{2}{n}\sum_{i=1}^n \xi_i\langle x_i,\beta-\beta^*\rangle\right| \ . \end{aligned}$$

By (6.19), we have thus

$$\begin{aligned} \inf_{\beta\in\beta^*+H_{r,\rho}} P_n\mathcal{L}_\beta &\ge \alpha^{-2}\Big[(r^*)^2 - Q_{r^*,\alpha\rho}\Big] - 2M_{r^*,\alpha\rho}\alpha^{-1} \\ &\ge \alpha^{-2}\Big[(r^*)^2 - Q_{r^*,\rho}\Big] - 2M_{r^*,\rho}\alpha^{-1} \ . \end{aligned} \tag{6.20}$$

It remains to bound the quadratic process $Q_{r^*,\rho}$ and the multiplier process $M_{r^*,\rho}$. This control is based on the Gaussian width of the sets $H_{r^*,\rho}$. Recall that the Gaussian width of a subset $H \subset \mathbb{R}^p$ is defined by

$$w^*(H) = \mathbb{E}\big[\sup_{h\in H}\langle G, h\rangle\big], \qquad \text{where} \qquad G \sim \mathcal{N}(0, I_p) \ .$$

The useful controls are provided in the following lemma, whose proof is postponed to Section 6.5.2.

Lemma 6.4. *Let $r, \rho \geq 0$ and $\delta, \eta \in (0,1)$. There exists an absolute constant c such that, with probability larger than $1-\delta$,*

$$Q_{r,\rho} \leq c\big[\mathcal{C}_{r,\rho}^2 + r\mathcal{C}_{r,\rho} + r^2(\mathcal{D}_{\delta,n} \vee \mathcal{D}_{\delta,n}^2)\big] \ ,$$

where the complexity $\mathcal{C}_{r,\rho} = w^*\big(\Sigma^{1/2}H_{r,\rho}\big)/\sqrt{n}$ and $\mathcal{D}_{\delta,n} = \sqrt{\log(1/\delta)/n}$. Moreover, there exists an other absolute constant c such that, with probability larger than $1-\eta$,

$$M_{r,\rho} \leq c\sigma\big[\mathcal{C}_{r,\rho} + r\mathcal{D}_{\eta,n}\big] \ .$$

We apply Lemma 6.4 with $\eta = \delta = e^{-\zeta^2 n}$, $r = r^*$ and ρ. We have $\mathcal{D}_{\delta,n}^2 \leqslant \mathcal{D}_{\delta,n} = \zeta < 1$. It shows that $\mathbb{P}(\Omega^*) \geqslant 1 - 2e^{-\zeta^2 n}$, where

$$\Omega^* = \{Q_{r^*,\rho} \leq c\big[\mathcal{C}_{r^*,\rho}^2 + r^*\mathcal{C}_{r^*,\rho} + \zeta(r^*)^2\big]\} \cap \{M_{r^*,\rho} \leq c\sigma\big[\mathcal{C}_{r^*,\rho} + r^*\zeta\big]\} \ .$$

Moreover, from Equation (6.20) and the fact that $\alpha = r^*/r$, on Ω^*

$$\inf_{\beta\in\beta^*+H_{r,\rho}} P_n\mathcal{L}_\beta \geq \big[r^2(1-c\zeta) - c\big(\frac{\mathcal{C}_{r^*,\rho}^2}{\alpha^2} + r\frac{\mathcal{C}_{r^*,\rho}}{\alpha}\big) - 2c\sigma\big(\frac{\mathcal{C}_{r^*,\rho}}{\alpha} + r\zeta\big)\big] \ . \tag{6.21}$$

It remains to bound the Gaussian width $w^*(\Sigma^{1/2}H_{r,\rho})$ to bound the complexity $\mathcal{C}_{r^*,\rho}$. This control is provided in the following lemma, whose proof is provided in Section 6.5.3.

Lemma 6.5. *Let $r, \rho \geq 0$. Then,*

$$w^*(\Sigma^{1/2}H_{r,\rho}) = \sqrt{2W_{r,\rho}}, \qquad \textit{where} \qquad W_{r,\rho} = \sum_{i=1}^p r^2 \wedge \lambda_i(\Sigma)\rho^2 \ .$$

From Lemma 6.5,

$$w^*\big(\Sigma^{-1/2}H_{r^*,\rho}\big) \leq c\sqrt{W_{r^*,\rho}} \ .$$

The choice of r^* ensures that

$$W_{r^*,\rho} \leq n(\zeta r^*)^2 \qquad \text{so} \qquad \frac{\mathcal{C}_{r^*,\rho}}{\alpha} \leq \zeta r \ .$$

Plugging this inequality into (6.21) shows that, on Ω^*,

$$\inf_{\beta\in\beta^*+H_{r,\rho}} P_n\mathcal{L}_\beta \geq r^2(1-3c\zeta) - 4c\zeta\sigma r \ .$$

The inequality $ab \leq (a^2+b^2)/2$ with $a = 4cr\zeta$ and $b = \sigma$ shows that, on Ω^*,

$$\inf_{\beta\in\beta^*+H_{r,\rho}} P_n\mathcal{L}_\beta \geq r^2(1-3c\zeta-8c^2\zeta^2) - \frac{\sigma^2}{2} \ .$$

Choosing ζ sufficiently small concludes the proof. ∎

6.4.3 Proof of Theorem 6.2

The proof is based on the following lemma, whose proof can be found in (Alquier et al., 2019) and (Chinot, 2019a) for example.

Lemma 6.6. *Assume that $\ell(u,y) = \rho(u-y)$, where ρ is L-Lipschitz. There exists an absolute constant c such that, for any positive r, ρ, with probability larger than $1-\eta$,*

$$\sup_{\beta \in H_{r,\rho}} |(P_n - P)(\ell_\beta - \ell_{\beta^*})| \leq \frac{cL}{\sqrt{n}}\big(w^*(H_{r,\rho}) + \sqrt{\log(1/\eta)}r\big) \ .$$

From Theorem 6.1, with probability larger than $1-7e^{-(\zeta\wedge v)n}$, $\|\hat{\beta}-\beta^*\| \leq \rho$ and $\|\Sigma^{1/2}(\hat{\beta}-\beta^*)\|_2 \leq r^*$. Consequently, from Lemmas 6.5 and 6.6, with probability larger than $1-\eta-7e^{-(\zeta\wedge v)n}$,

$$P\mathcal{L}_{\hat{\beta}} \leq P_n\mathcal{L}_{\hat{\beta}} + \frac{cL}{\sqrt{n}}\left(\sum_{i=1}^p (r^*)^2 \wedge \lambda_i(\Sigma)\rho^2\right)^{1/2} + cLr^*\sqrt{\frac{\log(1/\eta)}{n}} \ .$$

By definition of r^* this implies that

$$P\mathcal{L}_{\hat{\beta}} \leqslant P_n\mathcal{L}_{\hat{\beta}} + cLr^*\left(\zeta + \sqrt{\frac{\log(1/\eta)}{n}}\right) \ . \tag{6.22}$$

For the absolute loss function $P_n\mathcal{L}_{\hat{\beta}} \leq 0$ and $L=1$, so the proof is complete by taking $\eta = e^{-n}$.

For the Huber loss function, $L = c_2\sigma$ and $P_n\mathcal{L}_{\hat{\beta}} = -P_n\ell_{\beta^*} = -(1/n)\sum_{i=1}^p \rho_H(\xi_i)$. Moreover,

$$P_n\ell_{\beta^*} \geq \frac{1}{n}\sum_{i=1}^p \xi_i^2 \mathbf{1}\{|\xi_i| \leq c_2\sigma\} = \frac{1}{n}\sum_{i=1}^p \xi_i^2 - \frac{1}{n}\sum_{i=1}^p \xi_i^2\mathbf{1}\{|\xi_i| \geq c_2\sigma\} \ .$$

By (6.15), with probability larger than $1-\exp(-n/16)$, $(1/n)\sum_{i=1}^p \xi_i^2 \geq \sigma^2/2$. Similar arguments show that there exists an absolute constant ζ such that, with probability $1-\exp(-\zeta^2 n)$,

$$\frac{1}{n}\sum_{i=1}^p \xi_i^2\mathbf{1}\{|\xi_i| \geq c_2\sigma\} \leq \sigma^2/6 + \mathbb{E}[\xi^2\mathbf{1}\{|\xi| \geq c_2\sigma\}] \ .$$

Moreover, from Cauchy-Schwarz and Markov inequalities,

$$\mathbb{E}[\xi^2\mathbf{1}\{|\xi| \geq c_2\sigma\}] \leq \sqrt{3}\sigma^2\sqrt{\mathbb{P}(|\xi| \geq c_2\sigma)} \leq \frac{\sqrt{3}\sigma^2}{c_2} \ .$$

It follows that, with probability $1-2e^{-\zeta n}$, if $c_2 = 6\sqrt{3}$,

$$P_n\mathcal{L}_{\hat{\beta}} = -P_n\ell_{\beta^*} \leq -\sigma^2\left(\frac{1}{2} - \frac{1}{6} - \frac{\sqrt{3}}{c_2}\right) = -\frac{\sigma^2}{6} \ .$$

Plugging this estimate into (6.22) yields, with probability $1-10e^{-(\zeta\wedge v)n}$,

$$P\mathcal{L}_{\hat{\beta}} \leq -\frac{\sigma^2}{12} + 2c\zeta\sigma r^* \ .$$

The proof is complete since

$$2c\zeta\sigma r^* = 2(\sqrt{12}c\zeta r^*)\frac{\sigma}{\sqrt{12}} \leqslant 12c^2\zeta^2(r^*)^2 + \frac{\sigma^2}{12} \ .$$

6.5 Supplementary material

6.5.1 Sub-exponential random variables: definitions and properties

The following definition and propositions can be found in (Wainwright, 2019).

Definition 6.1. *A random variable X with mean $\mathbb{E}[X] = \mu$ is called sub-exponential with non-negative parameters (ν, b) if*

$$\mathbb{E}\big[e^{\lambda(X-\mu)}\big] \leq e^{\nu^2\lambda^2/2} \quad \text{for all} \quad |\lambda| \leq 1/b \ . \tag{6.23}$$

Proposition 6.1. *Let $X_1, \cdots, X_n$ be independent random variables such that X_i is sub-exponential with parameters (ν_i, b_i). Then $Y = \sum_{i=1}^n X_i$ is sub-exponential with parameters $\big((\sum_{i=1}^n \nu_i^2)^{1/2}, \max_{i=1,\cdots,n} b_i\big)$.*

Proposition 6.2 (Sub-exponential tail bound)**.** *Suppose that X is sub-exponential with parameters (ν, b). Then*

$$\mathbb{P}\big(|X-\mu| \geq t\big) \leq \begin{cases} 2e^{-t^2/(2\nu^2)} & \text{if } 0 < t \leq \nu^2/b \ , \\ 2e^{-t/(2b)} & \text{if } t \geq \nu^2/b \ . \end{cases} \tag{6.24}$$

6.5.2 Proof Lemma 6.4

The proof of the control of the quadratic process follows from (Dirksen et al., 2015, Theorem 5.5) and the majorizing measure theorem (see (Talagrand, 2014, Theorem 2.4.1)).
Let $(X_t)_{t\in T}$ be a stochastic process indexed by a set T of n-tuples $t = (t_1, \ldots, t_n)$. Let us assume that the random variables $X_{t_i} : \Omega_i \mapsto \mathbb{R}$ are sub-Gaussian. For every $t \in T$, let

$$A_t = \frac{1}{n}\sum_{i=1}^n \big(X_{t_i}^2 - \mathbb{E}X_{t_i}^2\big) \ . \tag{6.25}$$

Define on T the pseudo-distance d_{ψ_2}, by

$$d_{\psi_2}(t,s) = \max_{i=1,\cdots,n} \|X_{t_i} - X_{s_i}\|_{\psi_2} \ , \tag{6.26}$$

where, for any real random variable X, $\|X\|_{\psi_2} = \inf\{C > 0 : \mathbb{E}\exp(|X|^2/C^2) \leq 2\}$. The radius associated to T is defined as

$$\Delta_{\psi_2}(T) = \sup_{t\in T} \max_{i=1,\cdots,n} \|X_{t_i}\|_{\psi_2} \ . \tag{6.27}$$

Theorem 6.6 (Theorem 5.5 in (Dirksen et al., 2015))**.** *Let $(A_t)_{t\in T}$ be the process of averages defined in (6.25). There exists an absolute constant $c > 0$ such that, for any δ in $(0,1)$, with probability larger than $1-\delta$,*

$$\sup_{t\in T} A_t \leq c\left[\frac{\gamma_2^2(T, d_{\psi_2})}{n} + \Delta_{\psi_2}(T)\frac{\gamma_2(T, d_{\psi_2})}{\sqrt{n}} + K\frac{\log(1/\delta)}{n} + M\sqrt{\frac{\log(1/\delta)}{n}}\right] \ , \tag{6.28}$$

where the definition of γ_2 can be found in (Talagrand, 2014, Definition 2.2.19),

$$K = \sup_{t\in T} \max_{i=1,\cdots,n} \|X_{t_i}\|_{\psi_2}^2 \quad \text{and} \quad M = \sup_{t\in T}\left(\frac{1}{n}\sum_{i=1}^n \|X_{t_i}\|_{\psi_2}^4\right)^{1/2} \ .$$

To apply Theorem 6.6 to bound $Q_{r,\rho}$, let $T = \{(\langle x_1, \beta\rangle, \dots, \langle x_n, \beta\rangle), \beta \in H_{r,\rho}\}$ and, for any $t = (\langle x_1, \beta\rangle, \dots, \langle x_n, \beta\rangle) \in T$, let

$$X_{t_i} = \langle x_i, \beta\rangle, \qquad \text{so} \qquad Q_{r,\rho} = \sup_{t\in T} A_t \ .$$

For any $i = 1, \dots, n$, $X_{t_i} = \langle x_i, \beta\rangle \sim \mathcal{N}(0, \|\Sigma^{1/2}\beta\|_2^2) = \mathcal{N}(0, r^2)$. Therefore, $\|X_{t_i}\|_{\psi_2} = r$, for any $t \in T$ and any $i = \{1, \dots, n\}$, so $\Delta_{\psi_2}^2(T) = K = M = r^2$. Moreover, in our case $d_{\psi_2} = \|\cdot\|_\Sigma$ and from the majorizing measure theorem, see (Talagrand, 2014, Theorem 2.4.1), there exists an absolute constant $c > 0$ such that $\gamma_2(T, d_{\psi_2}) \le w^*(\Sigma^{1/2}H_{r,\rho})$, so, by Theorem 6.6, with probability $1 - \delta$

$$Q_{r,\rho} \le c\big[\mathcal{C}_{r,\rho}^2 + r\mathcal{C}_{r,\rho} + r^2(\mathcal{D}_{\delta,n} \vee \mathcal{D}_{\delta,n}^2)\big] \ .$$

Let us turn to the control of the multiplier process $M_{r,\rho}$. Since the noise ξ is Gaussian with variance σ^2, independent of x, by (Mendelson, 2016, Corollary 1.10), there exists an absolute constant c such that, for any δ in $(0, 1)$, with probability larger than $1 - \delta$,

$$nM_{r,\rho} \le c\sqrt{n}\sigma\big(w^*(\Sigma^{1/2}H_{r,\rho}) + r\sqrt{\log(1/\delta)}\big) \ .$$

6.5.3 Proof of Lemma 6.5

$$w^*(\Sigma^{1/2}H_{r,\rho}) = \mathbb{E}\sup_{t\in\Sigma^{1/2}H_{r,\rho}} \langle G, t\rangle \ ,$$

where $G \sim \mathcal{N}(0, I_p)$, and

$$\begin{aligned}
\Sigma^{1/2}H_{r,\rho} &= \{\Sigma^{1/2}t \in \mathbb{R}^p : \|t\| \le \rho, \|\Sigma^{1/2}t\|_2 = r\} \\
&= \{t \in \mathbb{R}^p : \|\Sigma^{-1/2}t\| \le \rho, \|t\|_2 = r\} \\
&= \left\{t \in \mathbb{R}^p : \sum_{i=1}^p \frac{t_i^2}{\lambda_i(\Sigma)\rho^2} \leqslant 1, \ \sum_{i=1}^p \frac{t_i^2}{r^2} \leqslant 1\right\} \\
&\subset \left\{t \in \mathbb{R}^p : \sum_{i=1}^p \frac{t_i^2}{\lambda_i(\Sigma)\rho^2 \wedge r^2} \leqslant 2\right\} \ .
\end{aligned}$$

The Gaussian mean-width of an ellipsoid is given by (Talagrand, 2014, Proposition 2.5.1) and it follows that

$$w^*(\Sigma^{1/2}H_{r,\rho}) \le \sqrt{2}\big(\sum_{i=1}^p \lambda_i(\Sigma)\rho^2 \wedge r^2\big)^{1/2} \ .$$

Bibliography

Advani, M. S. and Saxe, A. M. (2017). High-dimensional dynamics of generalization error in neural networks. *arXiv preprint arXiv:1710.03667*.

Alistarh, D., Allen-Zhu, Z., and Li, J. (2018). Byzantine stochastic gradient descent. In *Advances in Neural Information Processing Systems*, pages 4613–4623.

Alon, N., Matias, Y., and Szegedy, M. (1999). The space complexity of approximating the frequency moments. *J. Comput. System Sci.*, 58(1, part 2):137–147. Twenty-eighth Annual ACM Symposium on the Theory of Computing (Philadelphia, PA, 1996).

Alquier, P., Cottet, V., Lecué, G., et al. (2019). Estimation bounds and sharp oracle inequalities of regularized procedures with lipschitz loss functions. *The Annals of Statistics*, 47(4):2117–2144.

Amelunxen, D., Lotz, M., McCoy, M. B., and Tropp, J. A. (2014). Living on the edge: phase transitions in convex programs with random data. *Inf. Inference*, 3(3):224–294.

Argyriou, A., Baldassarre, L., Micchelli, C. A., and Pontil, M. (2013). On sparsity inducing regularization methods for machine learning. In *Empirical inference*, pages 205–216. Springer, Heidelberg.

Aronszajn, N. (1950). Theory of reproducing kernels. *Transactions of the American mathematical society*, 68(3):337–404.

Arora, S., Du, S. S., Hu, W., Li, Z., Salakhutdinov, R. R., and Wang, R. (2019). On exact computation with an infinitely wide neural net. In *Advances in Neural Information Processing Systems*, pages 8139–8148.

Audibert, J.-Y. and Catoni, O. (2011). Robust linear least squares regression. *Ann. Statist.*, 39(5):2766–2794.

Audibert, J.-Y., Tsybakov, A. B., et al. (2007). Fast learning rates for plug-in classifiers. *The Annals of statistics*, 35(2):608–633.

Bach, F., Jenatton, R., Mairal, J., and Obozinski, G. (2012). Structured sparsity through convex optimization. *Statist. Sci.*, 27(4):450–468.

Baraud, Y., Birgé, L., and Sart, M. (2017). A new method for estimation and model selection: ρ-estimation. *Invent. Math.*, 207(2):425–517.

Bartlett, P. L. (1998). The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network. *IEEE transactions on Information Theory*, 44(2):525–536.

Bartlett, P. L., Boucheron, S., and Lugosi, G. (2002a). Model selection and error estimation. *Machine Learning*, 48(1-3):85–113.

Bartlett, P. L., Bousquet, O., and Mendelson, S. (2002b). Localized Rademacher complexities. In *Computational learning theory (Sydney, 2002)*, volume 2375 of *Lecture Notes in Comput. Sci.*, pages 44–58. Springer, Berlin.

Bartlett, P. L., Bousquet, O., and Mendelson, S. (2005). Local Rademacher complexities. *Ann. Statist.*, 33(4):1497–1537.

Bartlett, P. L., Jordan, M. I., and McAuliffe, J. D. (2006). Convexity, classification, and risk bounds. *J. Amer. Statist. Assoc.*, 101(473):138–156.

Bartlett, P. L., Long, P. M., Lugosi, G., and Tsigler, A. (2019). Benign overfitting in linear regression. *arXiv preprint arXiv:1906.11300*.

Bartlett, P. L. and Mendelson, S. (2002). Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482.

Bartlett, P. L. and Mendelson, S. (2006a). Empirical minimization. *Probab. Theory Related Fields*, 135(3):311–334.

Bartlett, P. L. and Mendelson, S. (2006b). Local rademacher complexities and empirical minimization. *Annals of Statistics*, 34.

Belkin, M., Hsu, D., Ma, S., and Mandal, S. (2019a). Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854.

Belkin, M., Hsu, D., and Xu, J. (2019b). Two models of double descent for weak features. *arXiv preprint arXiv:1903.07571*.

Belkin, M., Hsu, D. J., and Mitra, P. (2018a). Overfitting or perfect fitting? risk bounds for classification and regression rules that interpolate. In *Advances in neural information processing systems*, pages 2300–2311.

Belkin, M., Rakhlin, A., and Tsybakov, A. B. (2018b). Does data interpolation contradict statistical optimality? *arXiv preprint arXiv:1806.09471*.

Bellec, P. C., Lecué, G., and Tsybakov, A. B. (2017). Towards the study of least squares estimators with convex penalty. *In Séminaire et Congrès, number 31. Société mathématique de France*.

Bellec, P. C., Lecué, G., Tsybakov, A. B., et al. (2018). Slope meets lasso: improved oracle bounds and optimality. *The Annals of Statistics*, 46(6B):3603–3642.

Ben-Hur, A. and Noble, W. S. (2005). Kernel methods for predicting protein–protein interactions. *Bioinformatics*, 21(suppl_1):i38–i46.

Bhaskar, B. N., Tang, G., and Recht, B. (2013). Atomic norm denoising with applications to line spectral estimation. *IEEE Trans. Signal Process.*, 61(23):5987–5999.

Bhatia, K., Jain, P., and Kar, P. (2015). Robust regression via hard thresholding. In *Advances in Neural Information Processing Systems*, pages 721–729.

Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2009). Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.*, 37(4):1705–1732.

Bietti, A. and Mairal, J. (2019). On the inductive bias of neural tangent kernels. In *Advances in Neural Information Processing Systems*, pages 12873–12884.

Birgé, L. (1984). Stabilité et instabilité du risque minimax pour des variables indépendantes équidistribuées. *Ann. Inst. H. Poincaré Probab. Statist.*, 20(3):201–223.

Birgé, L. (2001). An alternative point of view on lepski's method. *Lecture Notes-Monograph Series*, pages 113–133.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. springer.

Bogdan, M. g., van den Berg, E., Sabatti, C., Su, W., and Candès, E. J. (2015). SLOPE—adaptive variable selection via convex optimization. *Ann. Appl. Stat.*, 9(3):1103–1140.

Boucheron, S., Bousquet, O., and Lugosi, G. (2005). Theory of classification: a survey of some recent advances. *ESAIM Probab. Stat.*, 9:323–375.

Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration inequalities*. Oxford University Press, Oxford. A nonasymptotic theory of independence, With a foreword by Michel Ledoux.

Bousquet, O. (2002). A bennett concentration inequality and its application to suprema of empirical processes. *Comptes Rendus Mathematique*, 334(6):495–500.

Bousquet, O., Koltchinskii, V., and Panchenko, D. (2002). Some local measures of complexity of convex hulls and generalization bounds. In *International Conference on Computational Learning Theory*, pages 59–73. Springer.

Boyd, S., Boyd, S. P., and Vandenberghe, L. (2004). *Convex optimization*. Cambridge university press.

Brownlees, C., Joly, E., Lugosi, G., et al. (2015). Empirical risk minimization for heavy-tailed losses. *The Annals of Statistics*, 43(6):2507–2536.

Bubeck, S. (2015). Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357.

Bühlmann, P. and van de Geer, S. (2011). *Statistics for high-dimensional data*. Springer Series in Statistics. Springer, Heidelberg. Methods, theory and applications.

Bunea, F., Strimas-Mackey, S., and Wegkamp, M. (2020). Interpolation under latent factor regression models. *arXiv preprint arXiv:2002.02525*.

Bunea, F., Tsybakov, A., Wegkamp, M., et al. (2007). Sparsity oracle inequalities for the lasso. *Electronic Journal of Statistics*, 1:169–194.

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

Cai, T. T., Ren, Z., Zhou, H. H., et al. (2016). Estimating structured high-dimensional covariance and precision matrices: Optimal rates and adaptive estimation. *Electronic Journal of Statistics*, 10(1):1–59.

Candes, E. J. and Plan, Y. (2011). Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *IEEE Transactions on Information Theory*, 57(4):2342–2359.

Caponnetto, A. and De Vito, E. (2007). Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368.

Casella, G. (1980). Minimax ridge regression estimation. *The Annals of Statistics*, pages 1036–1056.

Catoni, O. (2012). Challenging the empirical mean and empirical variance: a deviation study. *Ann. Inst. Henri Poincaré Probab. Stat.*, 48(4):1148–1185.

Chafaï, D., Guédon, O., Lecué, G., and Pajor, A. (2012). *Interactions between compressed sensing random matrices and high dimensional geometry*. Citeseer.

Chalup, S. K. and Mitschele, A. (2008). Kernel methods in finance. In *Handbook on information technology in finance*, pages 655–687. Springer.

Chatterjee, S. and Goswami, S. (2019). New risk bounds for 2d total variation denoising. *arXiv preprint arXiv:1902.01215*.

Chen, M., Gao, C., Ren, Z., et al. (2016). A general decision theory for huber's epsilon-contamination model. *Electronic Journal of Statistics*, 10(2):3752–3774.

Chen, M., Gao, C., Ren, Z., et al. (2018). Robust covariance and scatter matrix estimation under Huber's contamination model. *The Annals of Statistics*, 46(5):1932–1960.

Chen, Y., Su, L., and Xu, J. (2017). Distributed statistical machine learning in adversarial settings: Byzantine gradient descent. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 1(2):1–25.

Cheng, Y., Diakonikolas, I., and Ge, R. (2019). 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.

Cherapanamjeri, Y., Flammarion, N., and Bartlett, P. L. (2019). Fast mean estimation with sub-gaussian rates. In Beygelzimer, A. and Hsu, D., editors, *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 786–806, Phoenix, USA. PMLR.

Chinot, G. (2019a). ERM and RERM are optimal estimators for regression problems when malicious outliers corrupt the labels. *arXiv preprint arXiv:1910.10923*.

Chinot, G. (2019b). Robust learning and complexity dependent bounds for regularized problems. *arXiv preprint arXiv:1902.02238*.

Chinot, G., Lecué, G., and Lerasle, M. (2019a). Robust high dimensional learning for lipschitz and convex losses. *arXiv:1905.04281*.

Chinot, G., Lecué, G., and Lerasle, M. (2019b). Robust statistical learning with lipschitz and convex loss functions. *Probability Theory and Related Fields*.

Dalalyan, A. and Thompson, P. (2019). 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.

Dalalyan, A. S., Hebiri, M., Lederer, J., et al. (2017). On the prediction performance of the lasso. *Bernoulli*, 23(1):552–581.

De Mol, C., De Vito, E., and Rosasco, L. (2009). Elastic-net regularization in learning theory. *Journal of Complexity*, 25(2):201–230.

Depersin, J. and Lecué, G. (2019). Robust subgaussian estimation of a mean vector in nearly linear time. *arXiv preprint arXiv:1906.03058*.

Devroye, L., Györfi, L., and Lugosi, G. (2013). *A probabilistic theory of pattern recognition*, volume 31. Springer Science & Business Media.

Devroye, L., Lerasle, M., Lugosi, G., Oliveira, R. I., et al. (2016). Sub-gaussian mean estimators. *The Annals of Statistics*, 44(6):2695–2725.

Diakonikolas, I., Kamath, G., Kane, D., Li, J., Moitra, A., and Stewart, A. (2019a). Robust estimators in high-dimensions without the computational intractability. *SIAM Journal on Computing*, 48(2):742–864.

Diakonikolas, I., Kong, W., and Stewart, A. (2019b). 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.

Dirksen, S. et al. (2015). Tail bounds via generic chaining. *Electronic Journal of Probability*, 20.

Donoho, D. L. and Huber, P. J. (1983). The notion of breakdown point. *A festschrift for Erich L. Lehmann*, 157184.

Eberts, M., Steinwart, I., et al. (2013). Optimal regression rates for svms using gaussian kernels. *Electronic Journal of Statistics*, 7:1–42.

Elsener, A. and van de Geer, S. (2018). Robust low-rank matrix estimation. *Ann. Statist.*, 46(6B):3481–3509.

Farooq, M. and Steinwart, I. (2019). Learning rates for kernel-based expectile regression. *Machine Learning*, 108(2):203–227.

Feldman, V. (2019). Does learning require memorization? a short tale about a long tail. *arXiv preprint arXiv:1906.05271*.

Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The elements of statistical learning*, volume 1. Springer series in statistics New York.

Giné, E., Koltchinskii, V., et al. (2006). Concentration inequalities and asymptotic results for ratio type empirical processes. *The Annals of Probability*, 34(3):1143–1216.

Giraud, C. (2015). *Introduction to high-dimensional statistics*, volume 139 of *Monographs on Statistics and Applied Probability*. CRC Press, Boca Raton, FL.

Golub, G. H., Hansen, P. C., and O'Leary, D. P. (1999). Tikhonov regularization and total least squares. *SIAM Journal on Matrix Analysis and Applications*, 21(1):185–194.

Golub, G. H., Heath, M., and Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2):215–223.

Gordon, Y., Litvak, A. E., Mendelson, S., and Pajor, A. (2007). Gaussian averages of interpolated bodies and applications to approximate reconstruction. *Journal of Approximation Theory*, 149(1):59–73.

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

Hampel, F. R. (1971). A general qualitative definition of robustness. *The Annals of Mathematical Statistics*, pages 1887–1896.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the american statistical association*, 69(346):383–393.

Han, Q. and Wellner, J. A. (2017). Convergence rates of least squares regression estimators with heavy-tailed errors.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

Holland, M. J. and Ikeda, K. (2017). Robust regression using biased objectives. *Machine Learning*, 106(9-10):1643–1679.

Holmes, R. B. (2012). *Geometric functional analysis and its applications*, volume 24. Springer Science & Business Media.

Hopkins, S. B. (2018). Sub-gaussian mean estimation in polynomial time. *arXiv preprint arXiv:1809.07425*.

Huang, J. Z. et al. (2003). Local asymptotics for polynomial spline regression. *The Annals of Statistics*, 31(5):1600–1635.

Huber, P. J. (1992). Robust estimation of a location parameter. *Breakthroughs in statistics*, pages 492–518.

Huber, P. J. et al. (1967). 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.

Huber, P. J. and Ronchetti, E. (2011). Robust statistics. In *International Encyclopedia of Statistical Science*, pages 1248–1251. Springer.

Jacot, A., Gabriel, F., and Hongler, C. (2018). Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in neural information processing systems*, pages 8571–8580.

Jerrum, M. R., Valiant, L. G., and Vazirani, V. V. (1986). Random generation of combinatorial structures from a uniform distribution. *Theoretical Computer Science*, 43:169–188.

Klein, T. (2002). Une inégalité de concentration à gauche pour les processus empiriques. *Comptes Rendus Mathematique*, 334(6):501–504.

Klein, T., Rio, E., et al. (2005). Concentration around the mean for maxima of empirical processes. *The Annals of Probability*, 33(3):1060–1077.

Koltchinskii, V. (2001). Rademacher penalties and structural risk minimization. *IEEE Transactions on Information Theory*, 47(5):1902–1914.

Koltchinskii, V. (2006). Local Rademacher complexities and oracle inequalities in risk minimization. *Ann. Statist.*, 34(6):2593–2656.

Koltchinskii, V. (2011a). Empirical and rademacher processes. In *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems*, pages 17–32. Springer.

Koltchinskii, V. (2011b). *Oracle inequalities in empirical risk minimization and sparse recovery problems*, volume 2033 of *Lecture Notes in Mathematics*. Springer, Heidelberg. 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].

Koltchinskii, V. et al. (2006). Local rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, 34(6):2593–2656.

Koltchinskii, V., Lounici, K., and Tsybakov, A. B. (2011). Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *Ann. Statist.*, 39(5):2302–2329.

Koltchinskii, V. and Mendelson, S. (2015). Bounding the smallest singular value of a random matrix without concentration. *Int. Math. Res. Not. IMRN*, (23):12991–13008.

Koltchinskii, V. and Panchenko, D. (2000). Rademacher processes and bounding the risk of function learning. In *High dimensional probability II*, pages 443–457. Springer.

Koltchinskii, V., Panchenko, D., et al. (2002). Empirical margin distributions and bounding the generalization error of combined classifiers. *The Annals of Statistics*, 30(1):1–50.

Lecué, G. and Depersin, J. (2019). Robust subgaussian estimation of a mean vector in nearly linear time. *arXiv preprint arXiv:1906.03058*.

Lecué, G. and Lerasle, M. (2017). Learning from mom's principles: Le cam's approach. *To appear in Stochastic processes and their applications*.

Lecué, G. and Lerasle, M. (2019). Robust machine learning by median-of-means: theory and practice. to appear in ann. statist. *Annals of Statistics*.

Lecué, G., Lerasle, M., and Mathieu, T. (2018). Robust classification via mom minimization. *arXiv preprint arXiv:1808.03106*.

Lecué, G. and Mendelson, S. (2013). Learning subgaussian classes: Upper and minimax bounds. *Topics in Learning Theory - Societe Mathématique de France*.

Lecué, G. and Mendelson, S. (2016). Performance of empirical risk minimization in linear aggregation. *Bernoulli*, 22(3):1520–1534.

Lecué, G. and Mendelson, S. (2017). Regularization and the small-ball method II: complexity dependent error rates. *J. Mach. Learn. Res.*, 18:Paper No. 146, 48.

Lecué, G. and Mendelson, S. (2018). Regularization and the small-ball method I: Sparse recovery. *Ann. Statist.*, 46(2):611–641.

Ledoux, M. (2001). *The concentration of measure phenomenon*, volume 89 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, RI.

Ledoux, M. and Talagrand, M. (2013). *Probability in Banach Spaces: isoperimetry and processes*. Springer Science & Business Media.

Lee, J., Xiao, L., Schoenholz, S., Bahri, Y., Novak, R., Sohl-Dickstein, J., and Pennington, J. (2019). Wide neural networks of any depth evolve as linear models under gradient descent. In *Advances in neural information processing systems*, pages 8570–8581.

Lee, W. S., Bartlett, P. L., and Williamson, R. C. (1998). The importance of convexity in learning with squared loss. *IEEE Transactions on Information Theory*, 44(5):1974–1980.

Lei, Z., Luh, K., Venkat, P., and Zhang, F. (2019). A fast spectral algorithm for mean estimation with sub-gaussian rates. *arXiv preprint arXiv:1908.04468*.

Lepskii (1992). Asymptotically minimax adaptive estimation. i: Upper bounds. optimally adaptive estimates. *Theory of Probability & Its Applications*, 36(4):682–697.

Lepskii (1993). Asymptotically minimax adaptive estimation. ii. schemes without optimal adaptation: Adaptive estimators. *Theory of Probability & Its Applications*, 37(3):433–448.

Lerasle, M. and Oliveira, R. I. (2011). Robust empirical mean estimators. *arXiv preprint arXiv:1112.3914*.

Li, Q. and Racine, J. S. (2007). *Nonparametric econometrics: theory and practice*. Princeton University Press.

Liang, T. and Rakhlin, A. (2018). Just interpolate: Kernel" ridgeless" regression can generalize. *arXiv preprint arXiv:1808.00387*.

Liu, L., Shen, Y., Li, T., and Caramanis, C. (2018). High dimensional robust sparse regression. *arXiv preprint arXiv:1805.11643*.

Loh, P.-L. et al. (2017). Statistical consistency and asymptotic normality for high-dimensional robust m-estimators. *The Annals of Statistics*, 45(2):866–896.

Loh, P.-L. and Wainwright, M. J. (2015). Regularized m-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *The Journal of Machine Learning Research*, 16(1):559–616.

Lugosi, G. and Mendelson, S. (2016). Risk minimization by median-of-means tournaments. *To appear in JEMS*.

Lugosi, G., Mendelson, S., et al. (2019a). Regularization, sparse recovery, and median-of-means tournaments. *Bernoulli*, 25(3):2075–2106.

Lugosi, G., Mendelson, S., et al. (2019b). Sub-gaussian estimators of the mean of a random vector. *The annals of statistics*, 47(2):783–794.

Lugosi, G., Wegkamp, M., et al. (2004). Complexity regularization via localized random penalties. *The Annals of Statistics*, 32(4):1679–1697.

Mammen, E. and Tsybakov, A. B. (1999). Smooth discrimination analysis. *Ann. Statist.*, 27(6):1808–1829.

Maronna, R. A. (1976). Robust m-estimators of multivariate location and scatter. *The annals of statistics*, pages 51–67.

Marsh, L. C. and Cormier, D. R. (2001). *Spline regression models*, volume 137. Sage.

Massart, P. (2000). Some applications of concentration inequalities to statistics. In *Annales de la Faculté des sciences de Toulouse: Mathématiques*, volume 9, pages 245–303.

Massart, P. (2007). *Concentration inequalities and model selection*, volume 6. Springer.

Mei, S. and Montanari, A. (2019). The generalization error of random features regression: Precise asymptotics and double descent curve. *arXiv preprint arXiv:1908.05355*.

Meier, L., Van De Geer, S., and Bühlmann, P. (2008). The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):53–71.

Meister, M. and Steinwart, I. (2016). Optimal learning rates for localized svms. *The Journal of Machine Learning Research*, 17(1):6722–6765.

Mendelson, S. (2001). Learning relatively small classes. In *International Conference on Computational Learning Theory*, pages 273–288. Springer.

Mendelson, S. (2002). Rademacher averages and phase transitions in glivenko-cantelli classes. *IEEE transactions on Information Theory*, 48(1):251–263.

Mendelson, S. (2003). On the performance of kernel classes. *Journal of Machine Learning Research*, 4(Oct):759–771.

Mendelson, S. (2014). Learning without concentration. In *Conference on Learning Theory*, pages 25–39.

Mendelson, S. (2015). Learning without concentration. *J. ACM*, 62(3):Art. 21, 25.

Mendelson, S. (2016). Upper bounds on product and multiplier empirical processes. *Stochastic Processes and their Applications*, 126(12):3652–3680.

Mendelson, S. (2017). On multiplier processes under weak moment assumptions. In *Geometric aspects of functional analysis*, volume 2169 of *Lecture Notes in Math.*, pages 301–318. Springer, Cham.

Mendelson, S., Neeman, J., et al. (2010). Regularization in kernel learning. *The Annals of Statistics*, 38(1):526–565.

Mendelson, S., Pajor, A., and Tomczak-Jaegermann, N. (2007). Reconstruction and subgaussian operators in asymptotic geometric analysis. *Geom. Funct. Anal.*, 17(4):1248–1282.

Minsker, S. (2018). Uniform bounds for robust mean estimators. *arXiv preprint arXiv:1812.03523*.

Minsker, S. et al. (2015). Geometric median and robust estimation in banach spaces. *Bernoulli*, 21(4):2308–2335.

Minsker, S. et al. (2018). Sub-gaussian estimators of the mean of a random matrix with heavy-tailed entries. *The Annals of Statistics*, 46(6A):2871–2903.

Minsker, S. et al. (2019). Distributed statistical estimation and rates of convergence in normal approximation. *Electronic Journal of Statistics*, 13(2):5213–5252.

Minsker, S. and Mathieu, T. (2019). Excess risk bounds in robust empirical risk minimization. *arXiv preprint arXiv:1910.07485*.

Nazin, A. V., Nemirovsky, A., Tsybakov, A. B., and Juditsky, A. (2019). Algorithms of robust stochastic optimization based on mirror descent method. *Automation and Remote Control*, 80(9):1607–1627.

Nemirovsky, A. S. and Yudin, D. B. (1983). *Problem complexity and method efficiency in optimization*. A Wiley-Interscience Publication. John Wiley & Sons, Inc., New York. Translated from the Russian and with a preface by E. R. Dawson, Wiley-Interscience Series in Discrete Mathematics.

Noble, W. S. et al. (2004). Support vector machine applications in computational biology. *Kernel methods in computational biology*, 71:92.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Pollard, D. et al. (1989). Asymptotics via empirical processes. *Statistical science*, 4(4):341–354.

Prasad, A., Suggala, A. S., Balakrishnan, S., and Ravikumar, P. (2018). Robust estimation via robust gradient estimation. *arXiv preprint arXiv:1802.06485*.

Recht, B., Fazel, M., and Parrilo, P. A. (2010). Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501.

Rohde, A., Tsybakov, A. B., et al. (2011). Estimation of high-dimensional low-rank matrices. *The Annals of Statistics*, 39(2):887–930.

Saumard, A. (2018). On optimality of empirical risk minimization in linear aggregation. *Bernoulli*, 24(3):2176–2203.

Schmidt, M., Roux, N. L., and Bach, F. R. (2011). Convergence rates of inexact proximal-gradient methods for convex optimization. In *Advances in neural information processing systems*, pages 1458–1466.

Schölkopf, B., Burges, C. J., Smola, A. J., et al. (1999). *Advances in kernel methods: support vector learning*. MIT press.

Scholkopf, B. and Smola, A. J. (2001). *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press.

Schölkopf, B., Tsuda, K., and Vert, J.-P. (2004). *Support vector machine applications in computational biology*. MIT press.

Shalev-Shwartz, S. and Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. Cambridge university press.

Shalev-Shwartz, S. and Tewari, A. (2011). Stochastic methods for l1-regularized loss minimization. *Journal of Machine Learning Research*, 12(Jun):1865–1892.

Shawe-Taylor, J., Cristianini, N., et al. (2004). *Kernel methods for pattern analysis*. Cambridge university press.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2013). A sparse-group lasso. *Journal of Computational and Graphical Statistics*, 22(2):231–245.

Smale, S. and Zhou, D.-X. (2007). Learning theory estimates via integral operators and their approximations. *Constructive approximation*, 26(2):153–172.

Steinwart, I. and Christmann, A. (2008). *Support vector machines*. Springer Science & Business Media.

Talagrand, M. (1996). New concentration inequalities in product spaces. *Inventiones mathematicae*, 126(3):505–563.

Talagrand, M. (2006). *The generic chaining: upper and lower bounds of stochastic processes*. Springer Science & Business Media.

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

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

Tsybakov, A. B. (2003). Optimal rates of aggregation. In *Learning theory and kernel machines*, pages 303–313. Springer.

Tsybakov, A. B. (2004). Optimal aggregation of classifiers in statistical learning. *Ann. Statist.*, 32(1):135–166.

Tsybakov, A. B. (2008). *Introduction to nonparametric estimation*. Springer Science & Business Media.

Tukey, J. W. (1960). A survey of sampling from contaminated distributions. *Contributions to probability and statistics*, pages 448–485.

Tukey, J. W. (1962). The future of data analysis. *The annals of mathematical statistics*, 33(1):1–67.

van de Geer, S. (2016). *Estimation and testing under sparsity*, volume 2159 of *Lecture Notes in Mathematics*. Springer, [Cham]. 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].

Van de Geer, S. A. (2000). *Applications of empirical process theory*, volume 91. Cambridge University Press Cambridge.

Van De Geer, S. A., Bühlmann, P., et al. (2009). On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics*, 3:1360–1392.

Van de Geer, S. A. et al. (2008). High-dimensional generalized linear models and the lasso. *The Annals of Statistics*, 36(2):614–645.

Van Der Vaart, A. W. and Wellner, J. A. (1996). Weak convergence. In *Weak convergence and empirical processes*, pages 16–28. Springer.

Vapnik, V. (1998). *Statistical learning theory*, volume 1. Wiley New York.

Vapnik, V. N. (2000). *The nature of statistical learning theory*. Statistics for Engineering and Information Science. Springer-Verlag, New York, second edition.

Vapnik, V. N. and Červonenkis, A. J. (1971). The uniform convergence of frequencies of the appearance of events to their probabilities. *Teor. Verojatnost. i Primenen.*, 16:264–279.

Vershynin, R. (2010). Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*.

Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press.

Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press.

Wu, Q., Ying, Y., and Zhou, D.-X. (2006). Learning rates of least-square regularized regression. *Foundations of Computational Mathematics*, 6(2):171–192.

Yang, M.-H., Ahuja, N., and Kriegman, D. (2000). Face recognition using kernel eigenfaces. In *Proceedings 2000 International Conference on Image Processing (Cat. No. 00CH37101)*, volume 1, pages 37–40. IEEE.

Yang, Y. and Zou, H. (2015). A fast unified algorithm for solving group-lasso penalize learning problems. *Statistics and Computing*, 25(6):1129–1141.

Yin, D., Chen, Y., Ramchandran, K., and Bartlett, P. (2018). Byzantine-robust distributed learning: Towards optimal statistical rates. *arXiv preprint arXiv:1803.01498*.

Yohai, V. J. and Maronna, R. A. (1979). Asymptotic behavior of m-estimators for the linear model. *The Annals of Statistics*, pages 258–268.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. (2016). Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*.

Zhang, T. (2004). Statistical behavior and consistency of classification methods based on convex risk minimization. *Ann. Statist.*, 32(1):56–85.

Zhou, W.-X., Bose, K., Fan, J., and Liu, H. (2018). A new perspective on robust M-estimation: finite sample theory and applications to dependence-adjusted multiple testing. *Ann. Statist.*, 46(5):1904–1931.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320.

ECOLE
DOCTORALE
DE MATHEMATIQUES
HADAMARD

Titre : Méthodes de localisation et applications à l'apprentissage robuste et à l'interpolation

Mots clés : Statistique, Machine learning, Robustesse

Résumé : Cette thèse de doctorat est centrée sur l'apprentissage supervisé. L'objectif principal est l'utilisation de méthodes de localisation pour obtenir des vitesses rapides de convergence, c'est-à-dire, des vitesse de l'ordre $\mathcal{O}(1/n)$, où n est le nombre d'observations. Ces vitesses ne sont pas toujours atteignables. Il faut imposer des contraintes sur la variance du problème comme une condition de Bernstein ou de marge. Plus particulièrement, dans cette thèse nous tentons d'établir des vitesses rapides de convergences pour des problèmes de robustesse et d'interpolation.
On dit qu'un estimateur est robuste si ce dernier présente certaines garanties théoriques, sous le moins d'hypothèses possibles. Cette problématique de robustesse devient de plus en plus populaire. La raison principale est que dans l'ère actuelle du "big data", les données sont très souvent corrompues. Ainsi, construire des estimateurs fiables dans cette situation est essentiel. Dans cette thèse nous montrons que le fameux minimiseur du risque empirique (regularisé) associé à une fonction de perte Lipschitz est robuste à des bruits à queues lourde ainsi qu'a des outliers dans les labels. En revanche si la classe de prédicteurs est à queue lourde, cet estimateur n'est pas fiable. Dans ce cas, nous construisons des estimateurs appelé estimateur minmax-MOM, optimal lorsque les données sont à queues lourdes et possiblement corrompues.
En apprentissage statistique, on dit qu'un estimateur interpole, lorsque ce dernier prédit parfaitement sur un jeu d'entrainement. En grande dimension, certains estimateurs interpolant les données peuvent être bons. En particulier, cette thèse nous étudions le modèle linéaire Gaussien en grande dimension et montrons que l'estimateur interpolant les données de plus petite norme est consistant et atteint même des vitesses rapides.

Title : Localization methods with applications to robust learning and interpolation

Keywords : Statistics, Machine learning, Robustness

Abstract : This PhD thesis deals with supervized machine learning and statistics. The main goal is to use localization techniques to derive fast rates of convergence, with a particular focus on robust learning and interpolation problems.
Localization methods aim to analyze localized properties of an estimator to obtain fast rates of convergence, that is rates of order $\mathcal{O}(1/n)$, where n is the number of observations. Under assumptions, such as the Bernstein condition, such rates are attainable.
A robust estimator is an estimator with good theoretical guarantees, under as few assumptions as possible. This question is getting more and more popular in the current era of big data. Large dataset are very likely to be corrupted and one would like to build reliable estimators in such a setting. We show that the well-known regularized empirical risk minimizer (RERM) with Lipschitz-loss function is robust with respect to heavy-tailed noise and outliers in the label. When the class of predictor is heavy-tailed, RERM is not reliable. In this setting, we show that minmax Median of Means estimators can be a solution. By construction minmax-MOM estimators are also robust to an adversarial contamination.
Interpolation problems study learning procedure with zero training error. Surprisingly, in large dimension, interpolating the data does not necessarily implies over-fitting. We study a high dimensional Gaussian linear model and show that sometimes the over-fitting may be benign.

Institut Polytechnique de Paris
91120 Palaiseau, France