Interplay between concentration, complexity and geometry in learning theory with applications to high dimensional data analysis

Guillaume Lecué

Université Paris-Est Marne-la-vallée

Habilitation à diriger des recherches

Spécialité : **Mathématiques**

soutenue par

Guillaume Lecué

Interplay between concentration, complexity and geometry in learning theory with applications to high dimensional data analysis.

Rapporteurs :	Pr.	Peter	**Bartlett**	UC Berkeley
	Pr.	Sara	**van de Geer**	ETH Zürich
	Pr.	Pascal	**Massart**	Université Paris 11 - Orsay

Soutenue publiquement le **8 Décembre 2011** devant le jury composé de

Pr. Lucien	**Birgé**	Université Paris 6 - Pierre et Marie Curie	Président
Pr. Stéphane	**Boucheron**	Université Paris 7 - Diderot	Examinateur
Pr. Djalil	**Chafaï**	Université Paris-Est Marne-la-vallée	Examinateur
Pr. Olivier	**Guédon**	Université Paris-Est Marne-la-vallée	Examinateur
Pr. Pascal	**Massart**	Université Paris 11 - Orsay	Rapporteur
Pr. Alain	**Pajor**	Université Paris-Est Marne-la-vallée	Directeur
Pr. Alexandre	**Tsybakov**	Université Paris 6 – ENSAE	Examinateur

Remerciements - Acknowledgement

Shahar Mendelson was the advisor of my post-doc that started in Australia and finished in Israël in 2007. I have learned so many beautifull things in Mathematics from Shahar that my acknowledgements go first and foremost to him. I really enjoyed all the discussions we had at the Purple Pickle in ANU, along the Lac Burley Griffith or at the Koko Black in Canberra, on the road to Pebly beach or Merimbula, at Greg's cafe at the Technion, in the streets of Paris,...(and many other places which do not sound very serious to work!). Since then I have been visiting Shahar dozens of times in Israël and Australia and I hope we will keep it like that for a long time. Thus once again, I wish to use the opportunity of writting my Habillitation to thank Shahar for everything I learned from him.

Je tiens à remercier vivement Alain Pajor pour m'avoir accueilli au LAMA et pour m'avoir présenté le problème du Compressed Sensing et en particulier la condition RIP pour les matrices de Fourier aléatoires. Je tiens aussi à remercier Alain pour toutes les discussions informelles qu'on a eu durant ces trois dernières années passées à Marne. J'ai aussi beaucoup appris de Djalil Chafaï et Olivier Guédon au cours de ces mêmes années. Je tiens à les en remercier.

J'exprime toute ma reconnaissance à Sara van de Geer, Peter Bartlett et Pacal Massart pour avoir accepté de rapporter ma thèse d'Habilitation.

Je remercie Karine Bertin et Sara van de Geer pour m'avoir invité à sejourner à l'ETH de Zürich et l'Universidad de Valparaíso.

Je remercie Lucien Birgé, Stéphane Boucheron, Djalil Chafaï, Olivier Guédon, Pascal Massart, Shahar Mendelson, Alain Pajor et Alexandre Tsybakov pour m'avoir fait l'immense honneur de faire partie de mon jury de thèse.

Je remercie Sylvain Arlot, Jean-Yves Audibert et Stéphane Gaïffas pour leurs commentaires sur ce manuscrit.

Je remercie chaleureusement mes collaborateurs, Christophe Chesneau, Stéphane Gaïffas, Karine Bertin, Shahar Mendelson et Charles Mitchell avec lesquels j'espère concrétiser encore d'autres projets.

Je tiens à remercier Christiane, Florence, Laurent et tous mes collègues chercheurs du LAMA.

Je termine par un grand remerciement à ma famille, mes amis, et Gaëlle; leur soutien et leurs encouragements constants ont permis à cette Habilitation de voir le jour.

Contents

Chapter 1

Introduction

In this document I present the works I undertook since the end of my Ph.D. I started my Ph.D in September 2004 at the Laboratoire de Probabilités et Modèles Aléatoires of Université Paris 6. I was then hired in October 2007 by the CNRS and spent my first two years at the Laboratoire d'Analyse, Topologie et Probabilité in Marseille. In 2009, I moved to the Laboratoire d'Analyse et Mathématiques Appliquées at the Université Paris-Est Marne-la-vallée. I will also use the opportunity of writing this manuscript to add some remarks and extensions to these works.

My main research interests are in learning theory and their applications to high-dimensional data analysis. The problem usually starts with a set of observations of the form input/output. The problem is to understand the interplay between the inputs and the outputs. In particular, given a new input, we want to associate, in a wise way, an output to this new input which is in compliance with what has been observed so far. Very complex real world systems can be understood in this paradigm and this is the reason why this theory has become so popular in many different fields of application like bioinformatics, speech, text and image recognition, computer vision, finance, energy and transport supply, etc..

Randomness comes into play at the step of the modelling of these data. It is common to treat these observations as a familly of n random variables $(X_1, Y_1), \dots, (X_n, Y_n)$ representing the data at hand. The input variables $X_1, \dots, X_n$ can take their values in very complicated spaces in practice. In theory, we denote by $\mathcal{X}$ the space where these variables take their values. The output variables $Y_1, \dots, Y_n$ are in general real numbers and sometimes even just binary labels with values for instance in $\{-1, 1\}$. Then, it is also common to assume that the variables $(X_1, Y_1), \dots, (X_n, Y_n)$ are independent and identically distributed. This assumption is well suited for modelling data coming at once: sometimes called the "batch setup". Of course, at this very first stage there are other ways of modelling these data. For instance, the input data can be deterministic: in this case, we speak about "deterministic design" or the data may be acquired one after the other: the "on line setup". In this case, the i.i.d. assumption does not hold anymore. We can think about many other ways of modelling a family of input/output data, but in all my works I focused on the i.i.d. setup with a random design.

Therefore, we start our mathematical problem with n i.i.d. random variables $(X_1, Y_1), \dots, (X_n, Y_n)$ with values in $\mathcal{X} \times \mathbb{R}$. The aim is to use mathematical tools to answer our concrete question concerning the interplay between the input and output data. There are many different ways of using these data to construct procedures. We thus want to be able to compare procedures. This is the role played by the loss function ℓ and the associated risk function $R(\cdot)$. We will be interested in the construction of procedures having a risk as small as possible. In particular, we are interested in oracle inequalities.

A big part of my work has been to prove oracle inequalities for particular procedures and

to prove their optimality thanks to lower bounds. Among these procedures are aggregation procedures which are studied in Chapter 2. The problem of aggregation is of particular interest since classical procedures (empirical risk minimization and its regularized and penalized versions) do not work in the aggregation context. Only very few optimal aggregation procedures have been constructed so far and the reason is that the geometrical aspect of the problem is of first importance. This point is at the heart of Chapter 2.

On the contrary, there are oracle inequalities which can be established without taking any special care of the geometry of the problem and for which the classical procedures work very well. Such oracle inequalities are shown in Chapter 3. It is interesting to note that very general oracle inequalities can be established for the three classical procedures: empirical risk minimization, regularized empirical risk minimization and penalized empirical risk minimization. In particular, in Chapter 3, we provide some ways of constructing regularizing and penalty functions.

The general oracle inequalities obtained in Chapter 3 are then applied in different problems like ℓ_1-regularization, S_1-regularization, Cross-validation procedures, selection of variables, etc..

Even though we introduced the problem of Learning theory for data of the form input/output, the area of action of Learning theory is much broader. Most of the results of this manuscript concern a model generalizing the input/output problems. We introduce now this general model.

1.1 A general model, examples and notations

1.1.1 A general model and some examples

Let $\mathcal{Z}$ be a space endowed with a probability measure P and let Z and $Z_1, \dots, Z_n$ be $n+1$ independent random variables with values in $\mathcal{Z}$, distributed according to P; from the statistical point of view, $\mathcal{D} = (Z_1, \dots, Z_n)$ is the set of given data. Let $\mathcal{F}$ be a linear space and define a loss function

$$\ell : \mathcal{F} \times \mathcal{Z} \longrightarrow \mathbb{R}$$

which associates a real number $\ell(f, z)$ to any element $f \in \mathcal{F}$ and point $z \in \mathcal{Z}$. For any $f \in \mathcal{F}$, we denote by ℓ_f the loss function $\ell(f, \cdot)$ associated with f. For any $f \in \mathcal{F}$, we assume that $\ell_f(Z)$ is a random variable (when $\mathbb{R}$ is endowed with the Borel algebra) and set $R(f) = \mathbb{E}\ell_f(Z)$ to be the risk of f. The risk of any statistic $\widehat{f}_n(\cdot) = \widehat{f}_n(\cdot, \mathcal{D})$ with values in $\mathcal{F}$ is defined by

$$R(\widehat{f}_n) = \mathbb{E}\big[\ell_{\widehat{f}_n}(Z)|\mathcal{D}\big] = \mathbb{E}\big[\ell(\widehat{f}_n, Z)|\mathcal{D}\big].$$

If the infimum

$$R^* = \inf_{f \in \mathcal{F}} R(f)$$

over all f in $\mathcal{F}$ is achieved, we write f^* for some choice of such a minimizer in $\mathcal{F}$. In this manuscript, we assume that $\inf_{f\in\mathcal{F}} R(f)$ is achieved — otherwise we can replace f^* by f_n^*, an element in $\mathcal{F}$ satisfying $R(f_n^*) \leq \inf_{f\in\mathcal{F}} R(f) + n^{-1}$, and still obtain the same results up to a n^{-1} additive term.

This model is best illustrated by its three key examples: regression, density estimation and classification.

The regression model

Take $\mathcal{Z} = \mathcal{X} \times \mathbb{R}$, where $(\mathcal{X}, \mathcal{A})$ is a measurable space and let $Z = (X, Y)$ be a random pair on $\mathcal{Z}$. Denote by P_X the probability distribution of X.

For the first example, take $\mathcal{F} = L^2(\mathcal{X}, \mathcal{A}, P_X)$. Assume that $\mathbb{E}|Y|^2 < \infty$ and define the regression function of Y given X for P_X-almost every $x \in \mathcal{X}$ by $\eta(x) = \mathbb{E}[Y|X = x]$. The square loss function is defined for any $(x, y) \in \mathcal{X} \times \mathbb{R}$ and $f \in \mathcal{F}$ by $\ell(f, (x, y)) = (y - f(x))^2$. The square risk is

$$R(f) = \mathbb{E}[\ell(f, (X, Y))] = \|\eta - f\|^2_{L^2(P^X)} + \mathbb{E}\left[\zeta^2\right],$$

where $\zeta = Y - \eta(X)$ is usually called the noise. In particular, $f^* = \eta$ is a minimizer of $R(\cdot)$ over $\mathcal{F}$ and the minimum achievable risk is $R^* = \mathbb{E}[\zeta^2]$.

For the second example, we consider $\mathcal{F} = L^1(\mathcal{X}, \mathcal{A}, P_X)$. Assume that $\mathbb{E}|Y| < \infty$ and define a conditional median function of Y given X for P_X-almost every $x \in \mathcal{X}$ by

$$m(x) \in \{m \in \mathbb{R} : \mathbb{P}[Y \leq m|X = x] \geq 1/2 \text{ and } \mathbb{P}[Y \geq m|X = x] \geq 1/2\}.$$

The L_1-loss function is defined for any $(x, y) \in \mathcal{X} \times \mathbb{R}$ and $f \in \mathcal{F}$ by $\ell(f, (x, y)) = |y - f(x)|$. The L_1-risk is

$$R(f) = \mathbb{E}[\ell(f, (X, Y))] = \mathbb{E}|Y - f(X)|.$$

For this example, any conditional median function $m(\cdot) = f^*(\cdot)$ is a minimizer of $R(\cdot)$ over $\mathcal{F}$.

The density estimation model

Let $(\mathcal{Z}, \mathcal{T}, \mu)$ be a measured space and take Z to be a random variable with values in $\mathcal{Z}$. We assume that the probability distribution P of Z is absolutely continuous with respect to μ and denote by f^* one version of its density. Consider $\mathcal{F}$ the set of all density functions on $(\mathcal{Z}, \mathcal{T}, \mu)$, i.e., the set of all $\mathcal{T}$-measurable functions $f : \mathcal{Z} \to \mathbb{R}_+$ that integrate to 1. We consider the loss function $\ell(f, z) = -\log f(z)$ for any $z \in \mathcal{Z}$ and $f \in \mathcal{F}$. The corresponding risk computes as

$$R(f) = \mathbb{E}[\ell(f, Z)] = K(f^*|f) - \int_{\mathcal{Z}} \log(f^*(z)) dP(z)$$

where $K(f^*|f) = \int_{\mathcal{Z}} \log\big(f^*(z)/f(z)\big) dP(z)$ is the Kullback-Leibler divergence between f^* and f. Thus f^* is a minimizer of $R(\cdot)$ over $\mathcal{F}$ and the minimum achievable risk is $R^* = -\int_{\mathcal{Z}} \log(f^*(z)) dP(z)$.

Instead of using the Kullback-Leibler loss, one can use the quadratic loss. The corresponding loss function is $\ell(f, z) = \int_{\mathcal{Z}} f^2 d\mu - 2f(z)$ for any $z \in \mathcal{Z}$ and $f \in \mathcal{F}$. Using this loss function, the risk of any $f \in \mathcal{F}$ works out as

$$R(f) = \mathbb{E}[\ell(f, Z)] = ||f^* - f||^2_{L^2(\mu)} - \int_{\mathcal{Z}} (f^*(z))^2 d\mu(z).$$

Thus the density function f^* is a minimizer of $R(\cdot)$ over $\mathcal{F}$ and the corresponding minimal risk is $R^* = -\int_{\mathcal{Z}} (f^*(z))^2 d\mu(z)$.

The classification model

Let $(\mathcal{X}, \mathcal{A})$ be a measurable space. We assume that the space $\mathcal{Z} = \mathcal{X} \times \{-1, 1\}$ is endowed with an unknown probability measure P, and consider a random pair $Z = (X, Y)$ which takes on values in $\mathcal{Z}$ and whose probability distribution is P. Denote by $\mathcal{F}$ the set of all measurable functions from $\mathcal{X}$ to $\mathbb{R}$, and furthermore let ϕ be a function from $\mathbb{R}$ to $\mathbb{R}$. For any $f \in \mathcal{F}$ consider the $\phi-$risk, $R(f) = \mathbb{E}[\ell(f, (X, Y))]$, where the loss function is given by $\ell(f, (x, y)) = \phi(yf(x))$ for any $(x, y) \in \mathcal{X} \times \{-1, 1\}$. In many situations, a minimizer f^* of the $\phi-$risk $R(\cdot)$ over $\mathcal{F}$ (or the sign of f^*, if the latter takes on arbitrary real values) is equal to the Bayes rule

$f^*(x) = \text{Sign}(2\eta(x) - 1), \forall x \in \mathcal{X}$, where $\eta(x) = \mathbb{P}(Y = 1|X = x)$ (cf. [121] and [14]). Classical examples of function ϕ are

$x \longrightarrow \mathbb{1}_{(x \leq 0)}$	classical loss or $0-1$ loss
$x \longrightarrow \max(0, 1-x)$	hinge loss (SVM loss)
$x \longrightarrow \log_2(1 + \exp(-x))$	logit-boosting loss
$x \longrightarrow \exp(-x)$	exponential boosting loss
$x \longrightarrow (1-x)^2$	squared loss
$x \longrightarrow \max(0, 1-x)^2$	2-norm soft margin loss

1.1.2 General notations

A subset $F \subset \mathcal{F}$ is called a *model*. Given a model F and a loss function ℓ, the *loss functions class* or *loss class* is the set

$$\ell_F = \{\ell_f : f \in F\} \text{ where } \ell_f = \ell(f, \cdot).$$

In what follows, we assume that the minimal risk over F is always achieved. Such a minimizer is called an *oracle*. We chose an oracle $f_F^* \in \text{argmin}_{f \in F} R(f)$. For any $f \in F$, the *excess loss of* f is the function $\mathcal{L}_f = \ell_f - \ell_{f_F^*}$ and the set of all the excess loss functions is called the *excess loss functions class* or the *excess loss class* and is defined by

$$\mathcal{L}_F = \{\mathcal{L}_f : f \in F\} \text{ where } \mathcal{L}_f = \ell_f - \ell_{f_F^*}.$$

We also consider another excess loss class. In this case, loss functions are compared with the loss function of the best element f^* in $\mathcal{F}$. For a given $f \in F$, we denote the *excess loss function with respect to* f^* by $\mathcal{E}_f = \ell_f - \ell_{f^*}$ and the *excess loss functions class with respect to* f^* is defined by

$$\mathcal{E}_F = \{\mathcal{E}_f : f \in F\} \text{ where } \mathcal{E}_f = \ell_f - \ell_{f^*}.$$

Throughout the manuscript, we denote absolute constants or constants that depend on other parameters by c, C, c_1, c_2, etc., (and, of course, we will specify when a constant is absolute and when it depends on other parameters). The values of these constants may change from line to line. The notation $x \sim y$ (resp. $x \lesssim y$) means that there exist absolute constants $0 < c < C$ such that $cy \leq x \leq Cy$ (resp. $x \leq Cy$). If $b > 0$ is a parameter then $x \lesssim_b y$ means that $x \leq C(b)y$ for some constant $C(b)$ depending only on b. We denote by ℓ_p^d the space $\mathbb{R}^d$ endowed with the ℓ_p norm $\|x\|_{\ell_p^d} = \left(\sum_j |x_j|^p\right)^{1/p}$. The unit ball there is denoted by B_p^d and the unit Euclidean sphere in $\mathbb{R}^d$ is S^{d-1}.

.

1.2 Oracle inequalities and classical procedures

Oracle inequalities are in Learning Theory as important as rates of convergence in Statistics. Given a Learning setup, many authors have been working on proving Oracle inequalities for some particular procedures. Trying to prove "optimal" oracle inequalities is one of the main topics of this work as well. The concept of optimality of oracle inequalities is close in nature to the one of minimax rate of convergence in Statistics and a precise definition will be given below. We first start to recall the classical setups and problems treated in Learning theory.

1.2.1 Three setups and three problems

Problems in Learning theory usually start with a set of n data and a large linear space $\mathcal{F}$. If one is able to determine some subset $F \subset \mathcal{F}$ of "small complexity" such that the best element in F may be of some special interest for the learning problem we want to solve (for instance, a good predictor of Y if the set of data are like input/output) then one may consider what we call the "Model setup". But, it is not always possible to find some subset $F \subset \mathcal{F}$ of small complexity such that an oracle in F may be interesting for the learning problem we have in mind. Either because the property we are looking for (smoothness or low-dimensionality, etc.) cannot be characterized by a set F of small complexity. In this circumstance, one may introduce some criterion function crit : $\mathcal{F} \to \mathbb{R}$ characterizing the property we have in mind. One of the main issue in this setup is to construct, from this criterion function, a function reg : $\mathcal{F} \to \mathbb{R}$ "regularizing" the empirical risk (this is the purpose of Section 3.5). This is what is called the "regularization setup" (cf. Section 1.2.3 for more details on the motivation behind this setup). Either, the model F is too large so that classical procedures like the ERM may fail due to a phenomenon called the "over-fitting" (cf. Section 1.2.4 for more details on this phenomenon). In this case, we usually write F as an increasing sequence $\mathcal{M}$ of sub-models with increasing complexity and we look for a way to "penalize large models". Since large models are the ones for which the ERM may perform badly. This framework is called the "Model Selection setup". To summarize, there are mainly three different setups in learning theory with different statistical motivations behind each one of them:

- **Model setup:** we are given a model $F \subset \mathcal{F}$. Usually, the set F is not too complex compared to the number of observations. The best element in F is of special interest. A natural candidate is thus the ERM over F. The study of the ERM in this setup is carried out in Section 2.2 and Section 3.3.

- **Regularization setup:** We are given a criterion function crit : $\mathcal{F} \to \mathbb{R}^+$ characterizing some property (low-dimensional structure or smoothness, etc.) or having some special implementation properties. We want to use this criterion function to regularize the empirical risk by constructing what is called a regularizing function reg : $\mathcal{F} \to \mathbb{R}$. In this setup, natural candidates are the regularized ERM procedures. These procedures are studied in Section 3.5.

- **Model Selection setup:** We are given a family $\mathcal{M}$ of models. We want to construct some function pen : $\mathcal{M} \to \mathbb{R}$ penalizing large models. In this setup, the natural procedure is the penalized estimator. It is studied in Section 3.6.

For each of the three setups, we consider three problems: non-exact prediction, exact prediction and estimation. As an introduction, let us present these three problems in the Model setup in which we are given a model $F \subset \mathcal{F}$.

- **Exact prediction problem:** construct a procedure having a risk close to $\inf_{f \in F} R(f)$ in expectation or in deviation. For this problem, it is important to note that we want to compare the risk of estimators to the exact minimal risk $\inf_{f \in F} R(f)$ with a leading constant 1. This is the problem treated in Chapter 2 in aggregation theory in a broad sense.

- **Non-exact prediction problem:** construct a procedure having a risk close to $(1 + \epsilon) \inf_{f \in F} R(f)$ for some $\epsilon > 0$ — still in expectation or deviation. This problem is studied in Chapter 3. In particular, it is interesting to underline the difference with the exact

prediction problem for which the leading constant has to be 1 whereas for the non-exact prediction problem, the leading constant is $1+\epsilon$ strictly greater than 1.

- **Estimation problem:** construct a procedure $\widehat{f}_n$ having an excess risk $R(\widehat{f}_n) - R(f^*)$ with respect to the best element $f^* \in \mathcal{F}$ as close as possible to $(1+\epsilon)\inf_{f\in F}\big(R(f) - R(f^*)\big)$ for some $\epsilon > 0$ with high probability or in expectation.

These three problems can as well be considered in the two other Regularization and Model Selection setups (cf. Section 1.2.3 and Section 1.2.4). Given one of the three setups, we will see that depending on the problem considered, different assumptions will be introduced (the Bernstein condition on ℓ_F or $\mathcal{L}_F$ or the Margin assumption on $\mathcal{E}_F$), different residual terms will be obtained and different way of regularizing and penalizing will come out of our study in Chapter 3. From a mathematical point of view, the study of the three problems in the three different setups will be understood through nine different inequalities which are called *oracle inequalities*. We consider three types of oracle inequalities depending on the setup: "oracle inequalities on models" (when we are given a model F in the model setup) — these inequalities are also called risk bounds —, "regularized oracle inequalities" (when we are given a criterion function crit : $\mathcal{F} \to \mathbb{R}^+$ in the regularization setup) and "penalized oracle inequalities" (when we are given a family $\mathcal{M}$ of models in the Model Selection setup). For each one of these oracle inequalities, there are three type of oracle inequalities depending on the goal we pursue: non-exact prediction, exact prediction or estimation.

All the results in Chapter 2 deal with the exact prediction problem in the model setup. In Chapter 3, we study the three classical procedures in the different setups: empirical risk minimization, regularized empirical risk minimization and penalized estimators, in the context of the three different problems. We introduce now the different oracle inequalities and the classical procedures for the three setups and the three problems.

1.2.2 Model setup: oracle inequalities and Empirical Risk Minimization

Let $F \subset \mathcal{F}$ be a model. In Learning theory, one wants to assume as little as possible on the class F, or on the probability measure P. The aim is to construct procedures $\widehat{f}_n$ such that, for some $\epsilon \geq 0$, with high probability,

$$R(\widehat{f}_n) \leq (1+\epsilon)\inf_{f\in F} R(f) + r_n(F), \tag{1.2.1}$$

for the exact and non-exact prediction problems, or, for the estimation problem,

$$R(\widehat{f}_n) - R(f^*) \leq (1+\epsilon)\inf_{f\in F}\big(R(f) - R(f^*)\big) + r_n(F). \tag{1.2.2}$$

The role of the *residual term* (or *rate*) $r_n(F)$ — which may depend on the probability distribution P (like the variance of the noise or some uniform bound) — is to capture the "complexity" of the problem, and the hope is to make it as small as possible.

When $r_n(F)$ tends to zero as n tends to infinity, Inequality (1.2.1) is called an *oracle inequality*. When $\epsilon = 0$, we say that $\widehat{f}_n$ satisfies an *exact oracle inequality* (the term *sharp* oracle inequality has been also used) and when $\epsilon > 0$ it satisfies a *non-exact oracle inequality*. Note that inequalities (1.2.1) and (1.2.2) have been also called risk bounds since they can be seen as a bound of the type "bias term + variance term". In the present manuscript, a best element $f_F^* \in \operatorname{argmin}_{f\in F} R(f)$ is called an oracle that is the reason why we call inequalities (1.2.1) and (1.2.2) oracle inequalities.

A natural algorithm in this setup is the *empirical risk minimization procedure* (ERM) (terminology due to [114]), in which the *empirical risk* functional

$$f \longmapsto R_n(f) = \frac{1}{n}\sum_{i=1}^{n} \ell_f(Z_i) = \frac{1}{n}\sum_{i=1}^{n} \ell(f, Z_i)$$

is minimized and produces

$$\widehat{f}_n^{ERM} \in \operatorname*{argmin}_{f\in F} R_n(f). \tag{1.2.3}$$

Note that when $R_n(\cdot)$ does not achieve its infimum over F or in case of ties, we define $\widehat{f}_n^{ERM}$ to be an element in F such that $R(\widehat{f}_n^{ERM}) \leq \inf_{f\in F} R(f) + 1/n$. This algorithm has been extensively studied, for instance, in [57, 15, 77]. We will also study this algorithm in Chapter 2 and Chapter 3.

1.2.3 Regularization setup: regularized Oracle inequalities and Regularized Empirical Risk Minimization

Now, we turn to the study of regularized empirical risk minimization procedures and to the introduction of regularized oracle inequalities. Usually a model F is chosen or constructed according to the belief that an oracle f_F^* in F is close, in some sense, to some minimizer f^* of the risk function in some larger class of functions $\mathcal{F}$ (for example, in the regression model with respect to the square loss and for $\mathcal{F} = L^2(P_X)$, f^* is the regression function of Y given X). Hence, by choosing a particular model $F \subset \mathcal{F}$, it implicitly means that we believe f^* to be close to F in some sense. It is not always possible to construct a class F that captures a property f^* is believed to have (e.g., a low-dimensional structure or some smoothness properties). In such situation, we are not given any model F (usually the set $\mathcal{F}$ is too large to be called a model), but a functional $\mathrm{crit} : \mathcal{F} \longrightarrow \mathbb{R}^+$, called a *criterion*, that characterizes each function according to its level of compliance with the desired property (roughly speaking, the smaller the criterion, the "closer" to the property). For instance, when $\mathcal{F}$ is a reproducing kernel Hilbert space (for a detailed exposition on RKHS and SVM we refer to [97]) one can take $\mathrm{crit}(f) = \|f\|_{\mathcal{F}}$ where $\|\cdot\|_{\mathcal{F}}$ is the reproducing norm over $\mathcal{F}$ (for some RKHS, if $\|f\|_{\mathcal{F}}$ is small then f is smooth) or when $\mathcal{F}$ is the set of all linear functionals in $\mathbb{R}^d$: $\mathcal{F} = \{f(\cdot) = \langle \cdot, \beta\rangle : \beta \in \mathbb{R}^d\}$ then one may choose $\mathrm{crit}(\beta) = \|\beta\|_{\ell_p}$ for some $p \in [0, \infty]$ (for $p = 0$, $\|\beta\|_{\ell_0}$ is the size of the support of β, thus a small criterion in this case means that β belongs to a small dimensional space). Instead of considering the ERM over the too large class $\mathcal{F}$, we want to construct a procedure having both good empirical performances and a small criterion. One idea, that we will not develop here, is to minimize the empirical risk over the set $F_r = \{f \in \mathcal{F} : \mathrm{crit}(f) \leq r\}$ (cf., for instance, [102, 16]) and try to find a data-dependent way of choosing the radius r. Another popular idea is to regularize the empirical risk: consider a non-decreasing function of the criterion called a *regularizing function* and denoted by $\mathrm{reg} : \mathcal{F} \longrightarrow \mathbb{R}^+$ (the choice of reg in function of crit depends on the complexity of the family of sets $(F_r)_{r\geq 0}$ and is the main purpose of Section 3.5 below) and construct

$$\widehat{f}_n^{RERM} \in \operatorname*{argmin}_{f\in\mathcal{F}} \big(R_n(f) + \mathrm{reg}(f)\big). \tag{1.2.4}$$

If the infimum of $f \to R_n(f) + \mathrm{reg}(f)$ over $\mathcal{F}$ is not attained, we can consider any function in $\mathcal{F}$ approximating this infimum up to a $1/n$ error term. But for simplicity, we will assume that the infimum is achieved. The procedure (1.2.4) is called *regularized empirical risk minimization procedure*. Regularized ERM procedures have been introduced to select functions with

additional properties, like smoothness (for instance, SVM estimators in [97]) or an underlying low-dimensional structure (e.g. the LASSO estimator) or having some particular computational interest (like the ℓ_1-norm being the convex relaxation of the ℓ_0 function).

The calibration of the regularizing function in terms of the criterion function is the main subject of Section 3.5. Intuitively, the regularizing function is an increasing function of the criterion. One way of constructing a "good" regularizing function in terms of a criterion (so that the regularized ERM satisfy some oracle inequalities) is by considering the complexity of the family of classes $\left(\ell_{F_r}\right)_{r\geq 0}$, $\left(\mathcal{L}_{F_r}\right)_{r\geq 0}$ or $\left(\mathcal{E}_{F_r}\right)_{r\geq 0}$ (depending on the problem we want to solve). Once the regularizing function is constructed (usually as $\mathrm{reg}(f) = h(\mathrm{crit}(f)), \forall f \in \mathcal{F}$ where h is an increasing function), we are interested in constructing estimators $\widehat{f}_n$ realizing the best possible trade-off between the risk and the regularizing function over $\mathcal{F}$: there exists some $\epsilon \geq 0$ such that with high probability

$$R(\widehat{f}_n) + \mathrm{reg}(\widehat{f}_n) \leq (1+\epsilon) \inf_{f\in\mathcal{F}} \big(R(f) + \mathrm{reg}(f)\big) \tag{1.2.5}$$

for the exact and non-exact prediction problem, or, for the estimation problem,

$$R(\widehat{f}_n) - R(f^*) + \mathrm{reg}(\widehat{f}_n) \leq (1+\epsilon) \inf_{f\in\mathcal{F}} \big(R(f) - R(f^*) + \mathrm{reg}(f)\big). \tag{1.2.6}$$

Using the same terminology as in (1.2.1), Inequality (1.2.5) is called a *regularized oracle inequality*. When $\epsilon = 0$, (1.2.5) is called an *exact regularized oracle inequality* and when $\epsilon > 0$, (1.2.5) is called a *non-exact regularized oracle inequality*. Such oracle inequalities are proved in Chapter 3.

1.2.4 Model Selection setup: penalized oracle inequalities and penalized estimators

We recall the setup of Model Selection as introduced in [76]. We are given a collection $\mathcal{M}$ of models. For every model $m \in \mathcal{M}$, an ERM procedure is constructed:

$$\widehat{f}_m \in \operatorname*{argmin}_{f\in m} R_n(f). \tag{1.2.7}$$

We know that for "large" or "complex" models m the ERM $\widehat{f}_m$ will have some tendency to stick to the data. When the data are corrupted by noise this is not very good to stick to them because in such a case, the procedure will have poor generalization capabilities. This phenomenon is called "over-fitting" and penalty functions have been introduced to circumvent this drawback of the ERM.

The role of the penalty function is to penalize large models and then to choose a model realizing a trade-off between empirical performances and the complexity measured by the penalty. Therefore, finding the "right" way to penalize models has been an important topic for many years and still is. This problem is studied in Section 3.6. Once a penalty function $\mathrm{pen} : \mathcal{M} \to \mathbb{R}^+$ has been constructed and after running the different ERM over all models $m \in \mathcal{M}$, the second step is now to select a model $\widehat{m}$ by

$$\widehat{m} \in \operatorname*{argmin}_{m\in\mathcal{M}} \Big(R_n(\widehat{f}_m) + \mathrm{pen}(m)\Big). \tag{1.2.8}$$

The *penalized estimator* studied in Model Selection is $\widehat{f}_{\widehat{m}}$. Once again, we assume that the infima in (1.2.7) and (1.2.8) are achieved. The aim is thus to prove penalized oracle inequalities:

for some $\epsilon > 0$, with high probability,

$$R(\widehat{f}_{\widehat{m}}) \leq (1+\epsilon) \inf_{m \in \mathcal{M}} \Big(\inf_{f \in m} R(f) + \mathrm{pen}'(m) \Big), \tag{1.2.9}$$

where pen' is proportional to pen up to some $1/n$ order additive terms. Note that, most of the oracle inequalities in Model Selection have been obtained for the estimation risk: with high probability,

$$R(\widehat{f}_{\widehat{m}}) - R(f^*) \leq (1+\epsilon) \inf_{m \in \mathcal{M}} \Big(\inf_{f \in m} R(f) - R(f^*) + \mathrm{pen}'(m) \Big). \tag{1.2.10}$$

Penalized estimators are studied in Section 3.6.

1.3 Margin assumption and Bernstein condition

The Margin assumption has been introduced by [107, 75] in a statistical setup and the Bernstein condition has been introduced in the Learning theory setup by [15]. We first recall the definition of the Margin assumption:

Definition 1.3.1 ([107]) *We say that the triple $(\mathcal{F}, \ell, P)$ satisfies the **Margin assumption** with parameters (β, B) for some $0 < \beta \leq 1$ and $B \geq 1$ when there exists $f^* \in \mathcal{F}$ such that $R(f^*) = \min_{f \in \mathcal{F}} R(f)$ and for every $f \in \mathcal{F}$, $\mathbb{E}\left(\ell_f - \ell_{f^*}\right)^2 \leq B\left(R(f) - R(f^*)\right)^\beta$. The parameter β is called the **Margin parameter** and B the **Margin constant**.*

Then, we recall the definition of the Bernstein condition:

Definition 1.3.2 ([15]) *We say that the triple (F, ℓ, P) satisfies the **Bernstein condition**, or that F or $\mathcal{L}_F$ satisfies the Bernstein condition, with parameters (β, B) for some $0 < \beta \leq 1$ and $B \geq 1$ when there exists $f_F^* \in F$ such that $R(f_F^*) = \min_{f \in F} R(f)$ and for every $f \in F$, $\mathbb{E}\big(\ell_f - \ell_{f_F^*}\big)^2 \leq B\left(R(f) - R(f_F^*)\right)^\beta$. The parameter β is called the **Bernstein parameter** and B the **Bernstein constant**.*

Note that the only formal difference between the two definitions is that for the Margin assumption, we compare the loss functions ℓ_f, for any $f \in \mathcal{F}$, with the loss function ℓ_{f^*} where f^* is a risk minimizer over $\mathcal{F}$ and, for the Bernstein condition, ℓ_f is compared with $\ell_{f_F^*}$ for any $f \in F$ where f_F^* is a risk minimizer over F. This difference makes the two assumptions actually very different in nature. The Margin assumption is a "statistical" assumption, measuring how good is the statistical problem whereas the Bernstein condition is a "geometrical" assumption measuring how good is the geometry of the system (F, ℓ, P). In the case of input/output data with respect to the square loss function, this is the relative position of Y and $\{f(X) : f \in F\}$ in $L_2(\Omega, \mathcal{A}, \mathbb{P})$ which is the key geometrical aspect of the Learning problem which is characterizes by the Bernstein condition.

As an example, consider the *bounded regression model with respect to the square loss.* This is the regression model defined in Section 1.1.1 (in particular $\mathcal{F} = L_2(P_X)$ and f^* is the regression function of Y given X) where it is further assumed that for some $b > 0$ and for a model $F \subset \mathcal{F}$, we have

$$|Y|, \sup_{f \in F} |f(X)| \leq b \text{ and } \ell_f(x,y) = (y - f(x))^2, \forall f \in F, (x,y) \in \mathcal{X} \times \mathbb{R}. \tag{1.3.1}$$

In this model, the Margin assumption is satisfied with the best possible Margin parameter $\beta = 1$ and Margin constant $B = (4b)^2$ since for any $f \in F$,

$$\begin{aligned}\big(\ell_f(x,y) - \ell_{f^*}(x,y)\big)^2 &= \big((y-f(x))^2 - (y-f^*(x))^2\big)^2 \\ &= (2y - f(x) - f^*(x))^2 (f^*(x) - f(x))^2 \le (4b)^2 (f^*(x) - f(x))^2\end{aligned}$$

and $R(f) - R(f^*) = \mathbb{E}(Y - f(X))^2 - \mathbb{E}(Y - f^*(X))^2 = \mathbb{E}(f(X) - f^*(X))^2$. On the other side, the Bernstein condition may have a very bad Bernstein constant B of the order of $\sqrt{n}$ (meaning that this condition does not help at all). To see this phenomenon, the set of *multiple minimizer*

$$N(F, \ell, X) = \big\{Y \in L_2(\Omega, \mathcal{A}, \mathbb{P}) : \mathrm{Card}\{\ell \in \ell_F : \mathbb{E}\ell(X,Y) = \min_{\ell \in \ell_F} \mathbb{E}\ell(X,Y)\} \ge 2\big\} \tag{1.3.2}$$

plays an important role as noticed in [80]. To be more precise, we see Y, X and $f(X)$ for all $f \in F$ as random variables defined on the "Kolmogorov probability space" $(\Omega, \mathcal{A}, \mathbb{P})$. Thus, $N(F, \ell, X)$ is the set of all the measurable outputs $Y : \Omega \to \mathbb{R}$ such that there are at least two oracles f_1 and f_2 in F for the loss function ℓ and the random variable $X : \Omega \to \mathbb{R}$ such that $\ell_{f_1} \neq \ell_{f_2}$ in $L_2(\mathcal{X} \times \mathbb{R}, \sigma((X,Y)), P)$. If the set $N(F, \ell, X)$ where ℓ is the square loss, is not empty and the output Y is $n^{-1/2}$-close to this set (cf. Figure 1.1) then this configuration may be unfavorable to the Bernstein condition. As an example, take $Y \equiv 0$ and define X by $\mathbb{P}[X = 1] = 1/2 - n^{-1/2}$ and $\mathbb{P}[X = -1] = 1/2 + n^{-1/2}$. Let $f_1 = \mathbb{1}_{[0,1]}$ and $f_2 = \mathbb{1}_{[-1,0]}$, and consider the dictionary $F = \{f_1, f_2\}$. It is easy to verify that the best function in F (the oracle) with respect to the quadratic risk is f_1 and that the excess loss function of f_2, $\mathcal{L}_2 = f_2^2 - f_1^2 = f_2 - f_1$, satisfies that

$$\mathcal{L}_2(X) = -X, \quad \mathbb{E}\mathcal{L}_2(X) = 2n^{-1/2} \quad \text{and } \sigma^2 = \mathbb{E}\big(\mathcal{L}_2(X) - \mathbb{E}\mathcal{L}_2(X)\big)^2 = 1 - 4/n. \tag{1.3.3}$$

For this example, the Bernstein constant is $B = \mathbb{E}(f_1 - f_2)^2 / \mathbb{E}\mathcal{L}_2 = \sqrt{n}/2$. The set $N(F, \ell, X)$ of multiple minimizer where ℓ is the quadratic loss, F and X are defined above is the set of all the real-valued random variables $Y \in L_2(\Omega, \mathcal{A}, X)$ such that if $f^*(X) = \mathbb{E}[Y|X]$ then $\mathbb{E}\big[f^*(X)(f_1(X) - f_2(X))\big] = 1/\sqrt{n}$. In particular the distance of $Y \equiv 0$ to this set is equal to $1/\sqrt{n}$. That is the reason why the Bernstein constant behaves like $\sqrt{n}$ and, we will see later, that the ERM performs badly in this context.

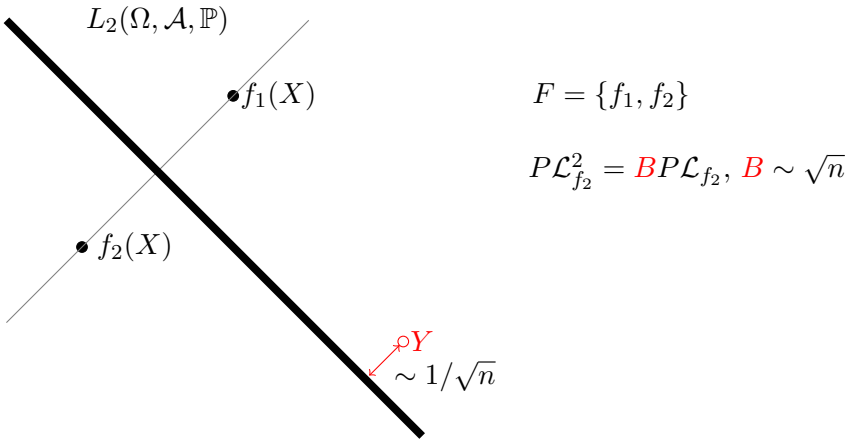

Figure 1.1: A bad geometrical situation for the Bernstein condition.

Therefore, simple examples exist for which the Margin assumption holds but not the Bernstein condition. Nevertheless, the two assumptions take their roots in the same idea: a control of the variance term by the expectation improves the concentration of the empirical mean and the level above which the empirical and actual mean are comparable. Indeed, in the case of a bounded random variable ζ (the study of a family of random variables indexed by a model will be studied in Subsection 3.2 and leads to the idea of localization), the Bernstein inequality yields, for any $x > 0$, with probability greater than $1 - 2\exp(-x)$,

$$\Big|\frac{1}{n}\sum_{i=1}^{n}\zeta_i - \mathbb{E}\zeta\Big| \leq K\sigma(\zeta)\sqrt{\frac{x}{n}} + K\left\|\zeta\right\|_\infty \frac{x}{n} \tag{1.3.4}$$

where $K > 0$ is an absolute constant, $\sigma(\zeta) = \big(\mathbb{E}(\zeta - \mathbb{E}\zeta)^2\big)^{1/2}$ is the standard deviation of ζ and $\zeta_1, \ldots, \zeta_n$ are n i.i.d. copies of ζ. If one has $\mathbb{E}\zeta^2 \leq B\mathbb{E}\zeta$ then it follows from (1.3.4) that for any $x > 0$, with probability greater than $1 - 2\exp(-x)$, if $\mathbb{E}\zeta \geq 4K(4B + \|\zeta\|_\infty)(x/n)$ then

$$(1/2)\mathbb{E}\zeta \leq \frac{1}{n}\sum_{i=1}^{n}\zeta_i \leq (3/2)\mathbb{E}\zeta \tag{1.3.5}$$

and if $\mathbb{E}\zeta \leq 4K(4B + \|\zeta\|_\infty)(x/n)$, then

$$\Big|\frac{1}{n}\sum_{i=1}^{n}\zeta_i - \mathbb{E}\zeta\Big| \leq \big(8KB + 2K\left\|\zeta\right\|_\infty\big)\frac{x}{n} \tag{1.3.6}$$

Then, under the assumption $\mathbb{E}\zeta^2 \leq B\mathbb{E}\zeta$ there is a phase transition regarding the expectation $\mathbb{E}\zeta$ at level of the order of $1/n$ such that above this level, the empirical mean and the actual mean are comparable (1.3.5) and below this level the empirical mean and the actual mean are both of the order of $1/n$ (1.3.6). These two properties are of particular interest from a statistical point of view when applied to the random variables $\zeta = \ell_f(Z) - \ell_{f^*}(Z)$ (when one wants to obtain estimation results and when the Margin assumption holds, see Theorem 3.2.4), or $\zeta = \ell_f(Z) - \ell_{f_F^*}(Z)$ (when one wants exact prediction results and when the Bernstein condition holds, see Theorem 3.2.3) or even $\zeta = \ell_f(Z)$ (when one wants non-exact prediction results, see Theorem 3.2.2). Indeed, for such variables the isomorphic property (1.3.5) means that the risk or the excess risk are comparable to their empirical version (meaning that what is observed is actually the truth up to some multiplying constants) and the property (1.3.6) means that the empirical (excess) risk and actual (excess) risk are both of a small $1/n$ order.

Although the idea behind the Margin assumption and the Bernstein condition are similar, they are, in fact, very different in nature, and have been also introduced in the context of different types of problems.

In the "Statistical framework" (cf. Section 1.4 for more details on what is called "Statistical framework"), one is given a model F with an upper bound on its complexity (whatever way of measuring the complexity is chosen) and an unknown target $f^* \in \mathcal{F}$, which is the minimizer of the risk over the entire set $\mathcal{F}$. In this framework, one usually assumes that f^* belongs to F and the aim is to construct an estimator $\widehat{f} = \widehat{f}(\cdot, \mathcal{D})$ for which the estimation risk $R(\widehat{f}) - R(f^*)$ or any other "distance" between $\widehat{f}$ and f^* tends to zero quickly as the sample size tends to infinity. In this framework, the Margin assumption can improve this rate of convergence thanks to a better concentration of empirical means of $\ell(f, \cdot) - \ell(f^*, \cdot)$ around its mean [107]. The Margin assumption for $\beta = 1$ compares the performance of each $f \in F$ to the *best possible element in* $\mathcal{F}$, but it has nothing to do with the geometric structure of F. The Margin is determined

for every f separately, because f^* does not depend on the choice of F at all. The role of the Margin assumption is best seen in the classification model with respect to the $0-1$ loss (i.e. $\ell_f(x,y) = \mathbb{1}_{f(x)\neq y}$). Indeed, for $\eta(x) = \mathbb{P}[Y=1|X=x]$ and $f^*(x) = \mathbb{1}_{\eta(x)\geq 1/2}$ defined for any $x \in \mathcal{X}$, it has been proved in [107] and [22] that the following are equivalent:

1. there exists $B > 0$ such that $\mathbb{E}\big(\ell_f - \ell_{f^*}\big)^2 \leq B\big(\mathbb{E}(\ell_f - \ell_{f^*})\big)^\beta$ for any $\{0,1\}-$valued measurable function f,
2. there exists $C > 0$ such that for any $t \geq 0$, $\mathbb{P}[|2\eta(X) - 1| \leq t] \leq Ct^{\beta/(1-\beta)}$.

Therefore, the Margin assumption over the set of all $\{0,1\}-$valued measurable functions characterizes the behavior of η around $1/2$ which determines the quality of the classification problem. In this setup, the Margin assumption is "statistical" in nature.

In the "Learning theory framework" (cf. Section 1.4 for more details on the difference between the "Statistical framework" and the "Learning theory framework"), we do not assume that f^* belongs to F. The aim is to construct a statistic $\widehat{f}_n$ whose risk is as close as possible to that of the best element $f_F^* \in F$. By assuming that the excess loss class $\mathcal{L}_F$ satisfies the Bernstein condition, one can improve the error rate (see, e.g., [82, 15]). On the other hand, the Bernstein condition involves a lot of geometry of the function class F, because f_F^* might change significantly by adding a single function to F or by removing one. In fact, the difficulty of "learning theory" problems is determined by the trade-off between concentration and complexity, *and* the geometry of the given class, since one measures the performance of the learning algorithm relative to the best *in the class*. Assuming that $f^* \in F$, as is usually done in classical statistics, exempts one from the need to consider the geometry of F, but we do not have that freedom in the Learning theory framework. The Bernstein condition should be seen as a tool characterizing the geometry of the Learning problem form which concentration properties can be derived. This is thus a condition characterizing the interplay between geometry and concentration for a given Learning problem. For instance, in the bounded regression model with respect to the square loss function, when the model is convex then the Bernstein condition is satisfied:

Proposition 1.3.3 *Let F be a convex set of $\mathcal{F}$. Consider the bounded regression model with respect to the square loss (1.3.1). Let $f_F^* \in \operatorname{argmin}_{f\in F} R(f)$ and, for any $f \in F$, set $\mathcal{L}_f = \ell_f - \ell_{f_F^*}$. We have*

$$P\mathcal{L}_f^2 \leq (4b)^2 P\mathcal{L}_f.$$

Proof. By convexity, we have for any $f \in F, \mathbb{E}\big((Y - f_F^*(X))(f(X) - f_F^*(X))\big) \leq 0$. Therefore, for any $f \in F$,

$$\begin{aligned} P\mathcal{L}_f &= \|Y - f(X)\|_2^2 - \|Y - f_F^*(X)\|_2^2 \\ &= -2\mathbb{E}\big((Y - f_F^*(X))(f(X) - f_F^*(X))\big) + \|f_F^* - f\|_2^2 \geq \|f_F^* - f\|_2^2 . \end{aligned}$$

The proof follows from the fact that for any $f \in F$,

$$P\mathcal{L}_f^2 = \mathbb{E}(2Y - f(X) - f_F^*(X))^2 (f(X) - f_F^*(X))^2 \leq (4b)^2 \|f_F^* - f\|_2^2 .$$

■

Note that an interesting condition has been introduced in [61] in the same spirit as the Bernstein condition and which is of particular interest in the case where the model F has multiple minimizers:

$$\forall f \in F, \exists f_F^* \in F(0) : \mathbb{E}\big(\ell_f - \ell_{f_F^*}\big)^2 \leq B\big(\mathbb{E}(\ell_f - \ell_{f_F^*})\big)^\beta, \tag{1.3.7}$$

where $F(0) = \{f \in F : R(f) = \min_{f \in F} R(f)\}$ is the set of oracles in F. It is mentioned in [61] that the same results obtained in [61] under the Bernstein condition for the ERM can be obtained under (1.3.7) as well. The advantage of this condition can be seen on the two functions case $F = \{f_1, f_2\}$ and for a target Y in the set of multiple minimizers $N(F, \ell, X)$. In this situation, the Bernstein condition does not hold since $R(f_1) = R(f_2)$ and $\mathbb{E}(\ell_{f_1} - \ell_{f_2}) > 0$ (since $\ell_{f_1} \neq \ell_{f_2}$ in $L_2(\mathcal{Z}, \sigma(Z), P)$) but the ERM $\widehat{f}_n$ satisfies a very good oracle inequality since with probability one $R(\widehat{f}_n) = \min_{f \in F} R(f)$. Thus in this particular example, the Bernstein condition does not reflect the quality of prediction of the ERM (since it does not hold) whereas (1.3.7) does (since $F(0) = F$ — any element in F is an oracle).

Finally, note that in [3, P10], a "local margin assumption" was introduced: there exists $f^* \in \mathcal{F}$ such that $R(f^*) = \min_{f \in \mathcal{F}} R(f)$ and for every $f \in F$, $\mathbb{E}\left(\ell_f - \ell_{f^*}\right)^2 \leq B\left(R(f) - R(f^*)\right)^\beta$ for some $0 < \beta \leq 1$ and $B \geq 1$. This condition is weaker than the "global" margin assumption of Definition 1.3.1 which requires that all the functions in $\mathcal{F}$ satisfies this condition whereas the local margin assumption requires this property only on the model F.

1.4 Some differences between Statistics and Learning Theory

In this last introductory section, we would like to underline some points which have already been discussed before concerning some differences between the Statistical framework and the Learning theory framework. Both problems start with the same batch of data. But the analysis of these data may be different if one adopts the Statistical or the Learning point of view.

For instance, consider a set of data $\mathcal{D}$ of the form input/output $(X_1, Y_1), \ldots, (X_n, Y_n)$. Given a new input X, we want to be able to predict the associated output Y in agreement with what has been observed so far (here, we assume that (X, Y) is distributed like the (X_i, Y_i)'s). In particular, we want to construct a function $\widehat{f}_n : \mathcal{X} \times \left(\mathcal{X} \times \mathbb{R}\right)^n \to \mathbb{R}$ such that $\widehat{f}_n(X, \mathcal{D})$ is close to Y in some sense.

In Statistics, we start with the remark that the best $\sigma(X)$-measurable function approaching Y in L_2 is the regression function of Y given X denoted by f^*. Thus, instead of predicting the output Y one should estimate $f^*(X)$ first. Because the best way to predict Y given X is by $f^*(X)$. The statistician is thus looking for a function $\widehat{f}_n(\cdot, \mathcal{D})$ close to f^* in $L_2(P_X)$ (other measure of proximity can also be considered, like $L_p(P_X)$ or $L_p(\lambda)$ where λ is some Lebesgue measure, or pointwise risk, etc.). We can now translate the prediction problem as the following problem: estimate f^* from noisy point-wise observations $Y_i = f^*(X_i) + \epsilon_i, i = 1, \ldots, n$ of f^* where $\epsilon_i = Y_i - f^*(X_i)$ is such that $\mathbb{E}[\epsilon_i | X_i] = 0, i = 1, \ldots, n$. Of course, even in the free noise setup $\epsilon_i = 0$ — meaning that Y is a function of X — there is no hope to estimate in L_2 the function f^* just from $f^*(X_1), \ldots, f^*(X_n)$. Therefore, if one wants to estimate f^* we have somehow to know more about f^* than just $f^* \in L_2(P_X)$. This is the point where the Statistician assumes that f^* has some property. In particular, that f^* belongs to some functions space $F \subset \mathcal{F}$ called a model. Many different models have been studied in Statistics and, for many classes F, we know how to construct optimal procedures achieving, up to some multiplying constant, the minimax rate of convergence over F defined by:

$$\inf_{\widehat{f}_n} \sup_{f^* \in F} \mathbb{E}_{f^*} \left\| \widehat{f}_n - f^* \right\|^2_{L_2(P_X)} \tag{1.4.1}$$

where the infimum is taken over all statistic $\widehat{f}_n$ constructed from n observations $(X_1, Y_1), \ldots, (X_n, Y_n)$ and $\mathbb{E}_{f^*}$ denotes the expectation with respect to the data $\mathcal{D}$ when $\mathbb{E}[Y_i | X_i] = f^*(X_i), i = 1, \ldots, n$.

The main difference between Statistics and Learning theory is that, in Learning theory, we don't assume that f^* belongs to some particular space F or has some particular properties: we want to assume as little as possible on the way the data have been generated.

In learning theory, we are given data and a model $F \subset \mathcal{F}$. Sometimes, instead of being given a model F, we are given a criterion $\text{crit} : \mathcal{F} \to \mathbb{R}$ (from which we want to construct some regularizing function to "regularize" the empirical risk) or a family $\mathcal{M}$ of models (for which we want to construct some penalty function to "penalize the empirical risk"). But for the purpose of this section, let us consider the case where we are given a model F — what we called the Model setup in Section 1.2.1. The model F may have nothing to do with the data but somehow we think that a best element $f_F^* \in \text{argmin}_{f \in F} R(f)$ in F (for a given risk function $R(\cdot)$ and where we assume that the infimum of $R(\cdot)$ over F is achieved) will provide a good prediction of Y by $f_F^*(X)$. We do not assume that f^* is in F or even close to F. We only want to construct a procedure $\widehat{f}_n$ having a prediction risk comparable to the minimal risk over F. That is why we are interested in oracle inequalities. That is for some $\epsilon \geq 0$, inequalities of the form

$$R(\widehat{f}_n) \leq (1+\epsilon) \min_{f \in F} R(f) + r_n(F)$$

or

$$R(\widehat{f}_n) - R(f^*) \leq (1+\epsilon) \min_{f \in F} \big(R(f) - R(f^*)\big) + r_n(F)$$

that hold with large probability. The role played by $r_n(F)$ is the same as the one played by the rates of convergence in Statistics. In particular, we can define optimal residual terms $r_n(F)$ in the same spirit as minimax rate of convergence (cf. for instance Definition 2.1.1 and Definition 2.1.2 in Chapter 2).

In particular, in oracle inequalities the bias term ("distance" of the best element in the model to the truth) is left untouched. Somehow, only the stochastic term is analyzed in Learning theory. Unlike in Statistics, there is no need for approximation theory of functions spaces in Learning theory (usually once we assume $f^* \in F$ in Statistics, we try to approximate f^* by some finite dimensional objects or when F is very large the Statistician will try to write F as an increasing sequence of models with increasing complexity). On the other side there is no need for geometry in Statistics since the target is suppose to be inside the model thus somehow we don't have to look at F from "outside". Moreover most of the models in Statistics are convex — except in classification. Meaning that models in Statistics have already a good geometry in general.

That is the reason why even though the Margin assumption and the Bernstein condition look similar they are in fact different. Because they have been introduced in different context and they are related to different aspect of the models. Nevertheless, when we study the problem of aggregation (which is a typical problem in Learning theory) under the Margin assumption (which is an assumption in Statistics), the quantity $R(f_F^*) - R(f^*)$ measuring the "distance" of f^* to the model F drives the residual term. Meaning that when f^* gets closer to F then the Bernstein condition gets "closer" to the Margin assumption and then classical residual terms under the Bernstein condition can be recovered in this "mixed setup" (a Learning theory problem under the Statistical Margin assumption).

From a technical point of view, this two theories share common tools in concentration and complexity theory. But, they also have their own background. In Statistics, many tools from approximation theory have been used to analyze the "bias term". In Leaning theory, the role played by the geometry aspect is of first importance. This is this aspect of Learning theory that is underlined in Chapter 2.

Chapter 2

The trade-off complexity/geometry in aggregation

Given a finite set $F \subset \mathcal{F}$, the problem of aggregation is to construct an estimator whose risk is as close as possible to the risk of the best element in F. Formally, we want to construct procedures $\widehat{f}_n$ such that with large probability

$$R(\widehat{f}_n) \leq \min_{f \in F} R(f) + r_n(F)$$

or in expectation

$$\mathbb{E}R(\widehat{f}_n) \leq \min_{f \in F} R(f) + r_n(F)$$

where, like before, $r_n(F)$ is called the residual term or the rate of aggregation that we want as small as possible. We are thus interested in proving exact oracle inequalities for finite models.

For this paradigm, it is possible to define an optimal rate of aggregation: this is the smallest price that one has to pay to mimic, in expectation or deviation, the best element in a function class F of cardinality M from n observations. A natural candidate is the ERM over F. Our first result is to exhibit the geometrical reasons why this procedure does not work for the aggregation problem. Then, after understanding the role played by the geometry in this problem, it is somehow natural to consider the ERM over the convex hull of F as a potential optimal aggregation procedure. We will show that this is still not the case: the ERM over the convex hull is sub-optimal for the aggregation problem. This result will follow from the study of the complexity of the intersection bodies $B_1^M \cap \sqrt{r}\mathcal{S}^{M-1}$ when $1/M \leq r \leq 1$. In particular, the ERM over the convex hull of F fails to achieve the optimal rate of aggregation for some complexity reasons.

Starting our analysis of the aggregation problem with these two facts: the ERM over F fails for some geometrical reason and the ERM over the convex hull of F fails for some complexity reason, it comes out that there is some sort of trade-off between the geometry and the complexity in the aggregation problem. Roughly speaking the set F has a very good complexity but very poor geometrical structure (this is a finite set), on the other side, the convex hull of F has a very good geometry (this is a convex set thus when working with a 2-convex loss function, we can hope for some gain in the approximation term) but a poor complexity: taking the convex hull of a set may increase drastically its complexity. Therefore, we will consider a procedure realizing some sort of optimal trade-off between complexity and geometry in the aggregation problem. It will come out an optimal aggregation procedure. We will see that the optimal aggregation procedure of J.-Y. Audibert in [7] is also based on this idea of "increasing the geometrical properties of F" without "increasing its complexity by too much".

Then we study a classical aggregation procedure: the aggregate with exponential weights. We will show that this procedure is suboptimal both in deviation and expectation for low temperatures (temperature smaller than a constant).

We will then say some word about the aggregation problem under the Margin assumption and the Bernstein condition.

We will end up this chapter with the study of the ERM over the convex hull for the Convex aggregation problem where the point here is to construct procedures doing as good as the best element in the convex hull of F.

2.1 The aggregation problem

The aggregation problem is a problem in Learning theory where we consider finite models. Note that in the "PAC-Bayesian" community, infinite models endowed with an a priori probability measure are also considered as aggregation problems but here our point of view is different and we will only consider finite model F. The problem is to construct a procedure having a risk as close as possible to the risk of the best element in F. This problem of aggregation is sometime called Model Selection aggregation or (MS) aggregation (cf. [8, 28, 42, 47, 53, P14, 103, 107, 118, 20, 25, 40, 51, 57, P11, 119, 120, 117, 39, 54]). There are other aggregation problems like the Convex aggregation problem (also called the (C) aggregation) where one wants to mimic the best element in the convex hull of F (cf. [5, 24, 25, 28, 53, 103, 120]) or the Linear aggregation problem (sometimes called the (L) aggregation) where one wants to mimic the best element in the linear span of F (cf. [28, 47, 57, 103]). For these problems, it is possible to define optimal aggregation procedures and optimal rates of aggregation in the same spirit as minimax procedures and minimax rates of convergence in statistics. The optimal rates of aggregation for the three problems have been obtained in [103] in the Gaussian regression model with respect to the square loss function. Those rates are now used as benchmarks for the aggregation problem in the sense that if a procedure achieves one of these rates then it is an optimal aggregation procedure. A formal definition of the concept of optimality in aggregation is now recalled in the bounded setup. As will be explained later, we consider two definitions of optimality: one in expectation and the other in deviation. We first start with the definition of optimality in expectation.

Definition 2.1.1 ([103]) *Let $b > 0$. We say that $(\psi_n(M))_{n,M\in\mathbb{N}^*}$ is an **optimal rate of aggregation in expectation** when there exists two positive constants c_0 and c_1 depending only on b for which the following holds for any $n \in \mathbb{N}^*$ and $M \in \mathbb{N}^*$:*

1. *there exists an aggregation procedure $\tilde{f}_n$ such that for any set $F \subset \mathcal{F}$ of cardinality M and any random variable Z satisfying $|\ell(f, Z)| \leq b$ a.s. for all $f \in F$, one has*

$$\mathbb{E}R(\tilde{f}_n) \leq \min_{f\in F} R(f) + c_0\psi_n(M), \tag{2.1.1}$$

2. *for any aggregation procedure $\bar{f}_n$ there exists a set $F \subset \mathcal{F}$ of size M and a random variable Z such that $|\ell(f, Z)| \leq b$ a.s. for all $f \in F$ and*

$$\mathbb{E}R(\bar{f}_n) \geq \min_{f\in F} R(f) + c_1\psi_n(M).$$

*A procedure $\tilde{f}_n$ satisfying (2.1.1) is called an **optimal aggregation procedure in expectation**.*

An aggregation procedure $\tilde{f}_n$ is a procedure having a dictionary F and a set of data $\mathcal{D}$ for arguments $\tilde{f}_n(\cdot, F, \mathcal{D}) = \tilde{f}_n(\cdot)$. In our setup, one can show (cf. [103]) that, in general, an optimal rate of aggregation in expectation is lower bounded by $(\log M)/n$. Hence, procedures satisfying an exact oracle inequality like (2.1.1) — that is an oracle inequality with a factor one in front of $\min_{f\in F} R(f)$ — with the residual term $\psi_n(M) = (\log M)/n$ are said to be optimal. There are very few aggregation procedures that have been proved to achieve this optimal rate. Some of them will be recalled in great details in the following like exponential aggregating schemes studied in [39, 6, 118, 7, 54], the "empirical star algorithm" introduced in [7] and the "preselection/convexification algorithm" defined in [P14].

Unlike other problems, the aggregation problem has different properties depending on one wishes to obtain results in expectation or in deviation. Because some procedures are optimal in expectation but on a constant probability event they can perform poorly (and thus are suboptimal in deviation). This surprising fact was first noticed in [7] and underlines the role of convexity and more generally of the geometry in the aggregation problem. We study this phenomenon in details in Section 2.3 but for now, we recall the definition of the optimality in deviation for the aggregation problem in the bounded case.

Definition 2.1.2 ([P14]) *Let $b > 0$. We say that $(\psi_n(M))_{n,M\in\mathbb{N}^*}$ is an optimal rate of aggregation in deviation with confidence $0 < \delta < 1/2$ if there exists three positive constants $c_1(\delta), c_2$ and c_3 for which the following holds for any $n \in \mathbb{N}^*$ and $M \in \mathbb{N}^*$:*

- *there exists an aggregation procedure $\tilde{f}_n$ such that for any set $F \subset \mathcal{F}$ of cardinality M and any random variable Z satisfying $|\ell(f, Z)| \leq b$ a.s. for all $f \in F$, one has, with $P^{\otimes n}$-probability at least $1 - \delta$,*

$$R(\tilde{f}_n) \leq \min_{f\in F} R(f) + c_1(\delta)\psi_n(M), \tag{2.1.2}$$

- *for any procedure $\bar{f}_n$, there exists a set F of cardinality M and a random variable Z satisfying $|\ell(f, Z)| \leq b$ a.s. such that with $P^{\otimes n}$-probability at least c_2,*

$$R(\bar{f}) \geq \min_{f\in F} R(f) + c_3\psi_n(M).$$

*A procedure $\tilde{f}_n$ is an **optimal aggregation procedure in deviation with confidence** δ if it satisfies (2.1.2).*

Note that optimal rates of aggregation are defined both in expectation and deviation up to some absolute multiplying constants.

These two definitions of optimality can be easily adapted to the two other Convex and Linear aggregation problems. Let us consider for a moment the problem of aggregation in deviation. In a general manner the three aggregation problems can be stated as follows: construct a procedure $\tilde{f}$ such that with high probability

$$R(\tilde{f}) \leq C \min_{f\in\Delta(F)} R(f) + \psi_n^{\Delta(F)}(M) \tag{2.1.3}$$

with $C = 1$ and $\Delta(F)$ is either F, or $\mathrm{conv}(F)$ or $\mathrm{span}(F)$. It is worth mentioning that the leading constant C in (2.1.3) should be equal to one in the aggregation setup for at least two reasons. First, there are deep mathematical differences in the analysis leading to exact oracle inequalities ($C = 1$) and non-exact oracle inequalities ($C > 1$). In particular, the geometry of

the set $\Delta(F)$ is of first importance to obtain exact oracle inequalities whereas non-exact oracle inequalities are mainly based on complexity and concentration arguments. In Chapter 3, we study non-exact oracle inequalities and we underline the differences between exact and non-exact oracle inequalities. Second, an exact oracle inequality for the prediction risk $R(\cdot)$ leads to an exact oracle inequality for the estimation risk; namely, for the regression model with respect to the square loss, by subtracting the risk of f^*, it follows from an exact oracle inequality on the risk that with high probability

$$\mathbb{E}\big[(\tilde{f}(X) - f^*(X))^2 | \mathcal{D}\big] \leq \min_{f \in \Delta(F)} \mathbb{E}\big[(f(X) - f^*(X))^2\big] + \psi_n^{\Delta(F)}(M),$$

where f^* denotes the regression function of Y given X. Such an oracle inequality, which estimates the regression function, cannot follow from a non-exact oracle inequality. In other words, exact oracle inequalities can provide prediction and estimation results whereas non-exact oracle inequalities only provide prediction results. More details on the difference between exact and non-exact oracle inequalities can be found in Chapter 3.

Following Definition 2.1.1, one can define the *optimal rates of the (MS), (C) and (L) aggregation* problems, respectively denoted by $\psi_n^{(MS)}(M)$, $\psi_n^{(C)}(M)$ and $\psi_n^{(L)}(M)$ (see, for example, [103]). For the square loss, it has been proved in [103] (see also [53] and [120] for the (C) aggregation problem) that

$$\psi_n^{(MS)}(M) = \frac{\log M}{n}, \psi_n^{(C)}(M) = \begin{cases} \frac{M}{n} & \text{if } M \leq \sqrt{n} \\ \sqrt{\frac{1}{n}\log\left(\frac{eM}{\sqrt{n}}\right)} & \text{if } M > \sqrt{n} \end{cases} \text{ and } \psi_n^{(L)}(M) = \frac{M}{n}.$$

Note that these rates obtained in [103] hold in expectation. Nevertheless, lower bounds in deviation follow from the arguments of [103] for the three aggregation problems with the same rates $\psi_n^{(MS)}(M)$, $\psi_n^{(C)}(M)$ and $\psi_n^{(L)}(M)$. In other words, there exist two absolute constants $c_0, c_1 > 0$ such that for any sample cardinality $n \geq 1$, any cardinality of dictionary $M \geq 1$ and any aggregation procedure $\bar{f}_n$, there exists a dictionary F of size M such that with probability larger than c_0,

$$R(\bar{f}_n) \geq \min_{f \in \Delta(F)} R(f) + c_1 \psi_n^{\Delta(F)}(M), \tag{2.1.4}$$

where the residual term $\psi_n^{\Delta(F)}(M)$ is $\psi_n^{(MS)}(M)$ (resp. $\psi_n^{(C)}(M)$ or $\psi_n^{(L)}(M)$) when $\Delta(F) = F$ (resp. $\Delta(F) = \mathrm{conv}(F)$ or $\Delta(F) = \mathrm{span}(F)$). Procedures achieving these rates in deviation have been constructed for the (MS) aggregation problem ([7] and [P14]) and the (L) aggregation problem [57]. So far, there is no example of a procedure that achieves the rate of aggregation $\psi_n^{(C)}(M)$ with a high exponential probability bound for the (C) aggregation problem (a result in deviation follows from the result in expectation from [103] by Markov inequality but only with a polynomial probability deviation). In Section 2.8, we construct an optimal aggregation procedure in deviation (with exponential bound) for the Convex aggregation problem with respect to the square loss function.

In what follows, we study the (MS) aggregation problem in the bounded regression model with respect to 2-convex risk function for the upper bounds. All the lower bounds will be proved for this model and with respect to the square loss function. The last subsection is devoted to the problem of (C) aggregation in the bounded regression model with respect to the square loss function.

2.2 On the suboptimality of the ERM

In this section, we study lower bounds for the empirical risk minimization algorithm over general models F which does not have to be finite. We want to understand the geometrical and complexity reasons why the ERM is suboptimal for the aggregation problem. In particular, we construct bad geometrical/complexity configurations for which the ERM performs poorly. This study is performed in the noiseless setup sometimes called the *function learning* problem, in which one observes n independent random variables $X_1, \dots, X_n$ distributed according to P_X, and the values $T(X_1), \dots, T(X_n)$ of an unknown target function T.

The goal is to construct a procedure that uses the data $\mathcal{D} = (X_i, T(X_i))_{1 \le i \le n}$ with a *risk* as close as possible to the best one in F; that is, we want to construct a statistic $\widehat{\widehat{f}}_n$ satisfying that for every n, with high $P_X^{\otimes n}$-probability

$$R(\widehat{f}_n) \le \inf_{f \in F} R(f) + r_n(F), \tag{2.2.1}$$

where the risk of f is defined by $R(f) = \mathbb{E}\ell(f(X), T(X))$ and $\ell : \mathbb{R}^2 \to \mathbb{R}$ is the loss function that measures the pointwise error between T and f. The residue $r_n(F)$ somehow captures the difficulty of the learning problem given by the triple (F, ℓ, P_X) from different point of view: complexity, geometry and concentration.

It is well known (see, for example, [115]) that if the class F is not too large, e.g., if it satisfies some kind of uniform Central Limit Theorem, T is bounded by 1 and ℓ is reasonable, there are upper bounds on $r_n(F)$ that are of the form $\sqrt{\mathrm{Comp}(F)/n}$, where $\mathrm{Comp}(F)$ is a complexity term that is independent of n. The algorithm that is used to produce the function $\widehat{f}_n$ is the ERM over F.

There is a well developed theory on ways in which the complexity term may be controlled, using various parameters associated with the geometry of the class (cf. [113] [115] [46] [101] and references therein). It turns out that this type of error rate, $\sim 1/\sqrt{n}$, is very pessimistic in many cases. In fact, if the class is small enough, then under the Bernstein condition, (cf. Definition 1.3.2), $r_n(F)$ can be much smaller - of the order of $\mathrm{Comp}(F)/n$.

In this section, we focus on "small classes" F in which empirical risk minimization performs poorly despite the size of the class. It has been shown in [80] that under mild assumptions on ℓ and F, if there is more than a single function in

$$V = \{\ell(f, T) : \mathbb{E}\ell(f, T) = \inf_{f \in F} \mathbb{E}\ell(f, T)\},$$

then the following holds: for every n large enough there will be a perturbation T_n of T (with respect to the L_∞ norm), for which $\mathbb{E}\ell(\cdot, T_n)$ has a unique minimizer in F, but the empirical risk minimization algorithm preforms poorly trying to predict T_n on samples of cardinality n. To be more exact, relative to the target T_n, with $P_X^{\otimes n}$-probability at least $1/12$,

$$R(\widehat{f}) \ge \inf_{f \in F} R(f) + \frac{c}{\sqrt{n}}, \tag{2.2.2}$$

where c is a constant depending only on F.

Although it is reasonable to expect that the larger the set V is, the more likely it is that the empirical risk minimization algorithm will perform poorly, it does not follow from the analysis in [80]. Therefore, our goal here is to provide a bound on the constant c in (2.2.2) that does take into account of the complexity of the set of minimizers V.

Just like in [80], our method of analysis can be applied to a wide variety of losses. However, for the sake of simplicity we will only present here what is arguably the most important case — in

which the risk is measured relative to the squared-loss that will be denoted by $\ell(x,y)=(x-y)^2$ in this section.

To explain our result we need several definitions from empirical processes theory. Other standard notions we require from the theory of Gaussian processes can be found in [46].

For every set $F \subset L_2(\mathcal{X}, P_X)$ let $\{G_f : f \in F\}$ be the canonical Gaussian process indexed by F (that is, with the covariance structure $\mathbb{E}G_tG_s = \langle s,t\rangle$) and set $H(F) = \mathbb{E}\sup_{f\in F} G_f$ - the expectation of the supremum of the Gaussian process indexed by F. Also, for every integer n and δ let

$$\mathrm{osc}_n(F,\delta) = \mathbb{E} \sup_{\{f,h\in F:\|f-h\|\le\delta\}} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n g_i(f-h)(X_i) \right|,$$

where $(g_i)_{i=1}^n$ are standard, independent Gaussian random variables and $(X_i)_{i=1}^n$ are independent, distributed according to P_X. It is known that if F is a class consisting of uniformly bounded functions then it is a P_X-Donsker class if and only if for every $\delta>0$, $\mathrm{osc}_n(F,\delta)$ tends to 0 as n tends to infinity (cf. [46], p.301). Given $f \in F$, we consider the oscillation in a ball around f

$$\mathrm{osc}_n(F,f,\delta) = \mathbb{E} \sup_{\{h\in F:\|f-h\|\le\delta\}} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n g_i(f-h)(X_i) \right|.$$

The quantity $\mathrm{osc}_n(F, f_F^*, \delta)$ is a natural upper bound for some local complexity measure of the problem we study here.

Let V be as above — the set of loss functions $\ell(f,T)$ that minimize the risk in F, select $f_F^* \in F$ for which $\ell(f_F^*,T) \in V$ and consider the following subset of excess loss functions

$$Q = \{\ell(f,T) - \ell(f_F^*,T) : \ell(f,T) \in V\}.$$

It turns out that the desired constant in (2.2.2) can be bounded from below by two parameters: the expectation of the supremum of the canonical Gaussian process indexed by Q and the oscillation around f_F^*. In particular, if Q is a rich set and one of the minimizers of $f \to \mathbb{E}\ell(f,T)$ is isolated, then for any n large enough the error of the empirical risk minimizer with respect to a wisely selected target (denoted by T_{λ_n} in what follows) which is a perturbation of T will be at least $\sim H(Q)/\sqrt{n}$. The core idea of this section is that a small wisely chosen pertubation of a target function T with multiple oracles (functions achieving $\min_{f\in F}\mathbb{E}\ell(f,T)$) is badly estimated by the empirical risk minimization procedure (for more discussion on this fact, we refer the reader to [80]).

Although the general philosophy of the proof presented here is similar to the proof from [80], it is much simpler and, in fact, it seems that the method used in the proof from [80] cannot be directly extended to obtain the sharper estimate on the constant as we do here. Naturally, this result recovers the previous estimates on lower bounds for the empirical risk minimization algorithm from [68, 57, P12, 80].

Finally before stating the main result of this section, a word about particular notation. We recall that if $\mathbb{E}\ell(\cdot,T)$ has a unique minimizer in F we denote it by f_F^*. If the minimizer is not unique, we fix one function in the set of minimizers and denote it by f_F^*. For every $f \in F$, let $\mathcal{L}(f) = \ell(f,T) - \ell(f_F^*,T)$ be the excess loss function associated with the target T. For every $0<\lambda\le 1$ set $T_\lambda = (1-\lambda)T + \lambda f_F^*$ and denote $\mathcal{L}_\lambda(f) = \ell(f,T_\lambda) - \ell(f_F^*,T_\lambda)$. It is standard to verify (cf. [80]) that f_F^* is a minimizer of $\mathbb{E}\ell(\cdot,T_\lambda)$, and that under mild convexity assumptions on ℓ that clearly hold if ℓ is the squared loss, it is the unique minimizer in F of $f \to \mathbb{E}\ell(f,T_\lambda)$. If $X_1,\dots,X_n$ is an independent sample selected according to P_X, set $P_nf = n^{-1}\sum_{i=1}^n f(X_i)$ and let $Pf = \mathbb{E}f(X)$. Thus, $\mathbb{E}\sup_{f\in F}|(P_n-P)f|$ is the expectation of the supremum of the

empirical process indexed by F. Finally, when the target function is T_λ, we denote the function produced by the empirical risk minimization algorithm by $\widehat{f}_\lambda$ — which is one element of the set $\operatorname{argmin}_{f\in F} P_n\ell(f, T_\lambda)$. Finally, if E is a normed space we denote its unit ball by $B(E)$, the inner product of $L_2(P_X)$ will be denoted by $\langle\cdot,\cdot\rangle$ and the corresponding norm by $\|\cdot\|$.

Theorem 2.2.1 ([P15]) *Let $F \subset L_2(P_X) \cap B(L_\infty)$ be P_X-Donsker (cf. [46]) and assume that $T \in B(L_\infty)$. Set ℓ to be the squared loss and put $Q = \{\mathcal{L}(f) : f \in F,\ \mathbb{E}\mathcal{L}(f) = 0\}$. There exist some absolute constants C_1 and C_2 and an integer $N(F)$ for which the following holds. For every $n \geq N(F)$, with $P_X^{\otimes n}$-probability at least C_1,*

$$\mathbb{E}\mathcal{L}_{\lambda_n}(\widehat{f}_{\lambda_n}) \geq C_2 \frac{H(Q)}{\sqrt{n}} \delta^2 \|T - f_F^*\|,$$

where δ satisfies that for every integer $n \geq N(F)$, $\operatorname{osc}_n(F, f_F^*, \delta) \leq C_2 H(Q)/\sqrt{n}$ and $\lambda_n = C_2 H(Q)/\sqrt{n}$.

Thus, two parameters control the behavior of the constant in (2.2.2). The complexity of the set of excess loss functions of the oracles in F with respect to T and the parameter δ. When one of the oracles f_F^* in F with respect to T is isolated then one can take δ as an absolute constant. This leads to a lower bound of the order of $H(Q)/\sqrt{n}$ which is optimal in the sense that an upper bound can be obtained of the order of $H(Q_0)/\sqrt{n}$ for some set Q_0 such that $Q \subset Q_0 \subset \mathcal{L}_F$ (cf. for instance [15] or [57]) . In other settings the lower bound obtained in theorem 2.2.1 may fail to match exactly with an upper bound. For instance, in settings where the oscillation function $\operatorname{osc}_n(F, f_F^*, \cdot)$ of all the oracles f_F^* in F with respect to T decrease to zero very slowly and at the same convergence rate, the factor δ^2 should deteriorate the lower bound whereas it seems that it should not appear in the lower bound. Finally, the noiseless model considered here is the worst case scenario to prove the lower bound. Indeed, adding some noise to the target function would increase the lower bound.

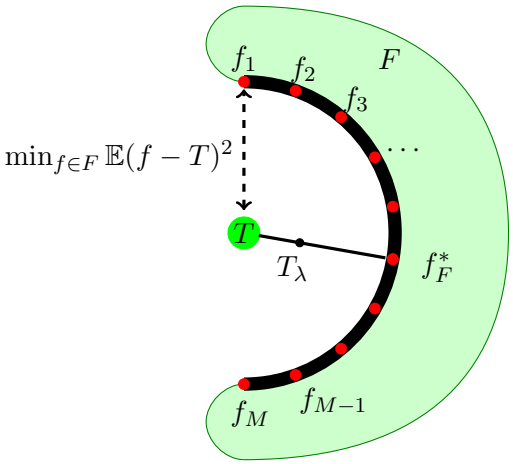

Figure 2.1: A bad geometrical situation for the ERM when the target function is T_λ.

Therefore a bad configuration for the ERM is given in Figure 2.1 when the target function is $T_\lambda = (1-\lambda)T + \lambda f_F^*$ for some well-chosen λ. In this situation, there is only one oracle but many functions in F which are far from the oracle but having a risk larger than the minimal one only

by a term of the order of $1/\sqrt{n}$. In some sort, the ERM is going to be misleading by all these "approximately" oracles. This will result in increasing the residual term of the ERM each time one of this approximately oracles is wrongly chosen by the ERM. Therefore, if all the complexity of the class lies in the set of almost oracles then this geometrical/complexity configuration is very disadvantageous for the ERM. In particular, in the aggregation problem if all the elements in F are at equal distance to some target T (with the distance from T to F being constant) then a small perturbation T_λ of this target for $\lambda = \lambda_n = \sqrt{(\log M)/n}$ will result in a bad aggregation configuration for the ERM and it follows from Theorem 2.2.1 that with $P_X^{\otimes n}$-probability at least C_1, the ERM $\widehat{f}_{\lambda_n}$ when the target function is T_{λ_n} satisfies

$$R(\widehat{f}_{\lambda_n}) \geq \min_{f\in F} R(f) + C_2\sqrt{\frac{\log M}{n}}.$$

This explains why the ERM procedure is suboptimal for the (MS) aggregation problem.

Note that in the two functions case (i.e. dictionary of size two), a simple argument shows that the ERM is suboptimal. Consider the example introduced in (1.3.3). The oracle is f_1 such that $R(f_1) = 1/2 - 1/\sqrt{n} < 1/2 + 1/\sqrt{n} = R(f_2)$. Therefore, every time the ERM chooses f_2, its risk is larger than the risk of the oracle by the quantity $2/\sqrt{n}$. This happens when $R_n(f_2) < R_n(f_1)$. On the other side, it follows from Berry-Esséen that for any $t \in \mathbb{R}$,

$$\left|\mathbb{P}\Big[R_n(f_2) - R_n(f_1) \leq R(f_2) - R(f_1) + \frac{\sigma t}{\sqrt{n}}\Big] - \mathbb{P}[g \leq t]\right| \leq \frac{c_0\mathbb{E}|X|^3}{\sqrt{n}}$$

where g is a standard real Gaussian random variable and $\sigma^2 = \mathbb{E}(X - \mathbb{E}X)^2 = 1 - 4/n$. In particular, there exists some absolute constant $n_0 \in \mathbb{N}$ such that when $n \geq n_0$, with probability greater then $\mathbb{P}[g \leq -3] - c_0/\sqrt{n} \geq c_1 > 0$,

$$R_n(f_2) - R_n(f_1) \leq R(f_2) - R(f_1) + \frac{\sigma t}{\sqrt{n}} \leq \frac{2}{\sqrt{n}} - \frac{3}{\sqrt{n}} < 0.$$

Therefore, with probability greater than c_1, the ERM is somehow misleading and chooses f_2 which implies that $R(\widehat{f}_n^{ERM}) = R(f_2) = \min_{f\in F} R(f) + 2/\sqrt{n}$. This proves a $1/\sqrt{n}$ lower bound for the ERM showing that the ERM cannot achieves the optimal $1/n$ rate in the context of a two functions class.

2.3 Improving the geometry by taking the convex hull?

In this section, our goal is to explain the role of convexity in the aggregation problem and that, even though the ERM over the convex hull seems a natural candidate to be an optimal aggregation procedure, we prove that this is not the case.

We first start with a remark from [7] showing that the *progressive mixture rule* is an optimal aggregation procedure in expectation but not in deviation. This procedure is defined for a dictionary F and a temperature parameter $T > 0$ by

$$\bar{f}_n = \frac{1}{n}\sum_{k=1}^{n} \tilde{f}_k^{AEW} \tag{2.3.1}$$

where f_k^{AEW} is the aggregate with exponential weights constructed on the first k observations. The aggregate with exponential weights is defined for a dictionary $F = \{f_1, \ldots, f_M\}$ and n

observations by

$$\tilde{f}_n^{AEW} = \sum_{j=1}^{M} \widehat{\theta}_j f_j \text{ where } \theta_j = \frac{\exp\big(-\frac{n}{T} R_n(f_j)\big)}{\sum_{k=1}^{M} \exp\big(-\frac{n}{T} R_n(f_k)\big)}. \tag{2.3.2}$$

The procedures $\tilde{f}_k^{AEW}$ for $1 \leq k \leq n$ are thus the same as in (2.3.2) but with the empirical risk $R_n(\cdot)$ constructed on n observations replaced by the empirical risk $R_k(\cdot)$ constructed on the first k observations for all $1 \leq k \leq n$. It has been proved by several authors [39, 54, 117, 119] that the progressive mixture rule is such that for T large enough and under some convexity assumption on the risk function,

$$\mathbb{E}R(\bar{f}_n) \leq \min_{f \in F} R(f) + c_0 \frac{T \log M}{n}. \tag{2.3.3}$$

This proves that $\bar{f}_n$ is an optimal aggregation procedure in expectation. But at the same time, [7] proves that $\bar{f}_n$ is suboptimal in deviation since under some regularity condition on the loss function, it is proved that for a dictionary of cardinality two and for some particular probability distribution P, with constant $P^{\otimes n}$-probability,

$$R(\bar{f}_n) \geq \min_{f \in F} R(f) + \frac{c_1}{\sqrt{n}}. \tag{2.3.4}$$

This phenomenon is very unusual in statistics since results in deviation can be derived from results in expectation by using Markov inequality in general. But for the aggregation problem, this particular aspect of the problem is due to the fact that aggregation procedures are allowed to take values outside of the class F — for instance in the convex hull of F. This is in particular the case of the progressive mixture rule $\bar{f}_n$. In such situations, the random variable $R(\bar{f}_n) - \min_{f \in F} R(f)$ may take negative values. Meaning that the aggregate does actually better than the oracle. Such a gain is due to convexity properties of the problem: convexity of the convex hull of F and of the risk function. In the simple two functions class example of Figure 1.1, we can already see that there is a large part of the convex hull of F where aggregation procedures taking values in this part will actually do better than the oracle itself. This is the case for the progressive mixture rule: even if there is a constant probability event on which $\bar{f}_n$ does worse than the oracle by a residual term of the order of $1/\sqrt{n}$ there is also a large event on which it is does much better than the oracle, better enough to compensate the loss of (2.3.4) and to be finally optimal in expectation as in (2.3.3). Therefore, convexity of the convex hull and of the loss function are key points in constructing optimal aggregation procedures.

We want to keep this idea that we can gain by taking aggregates with values in the convex hull of F when the loss function is convex. This is even worse than that for the aggregation problem, since it is proved in [54] (cf. also the counter-example in page 81 of [39] in density estimation with respect to the KL-loss in the two functions case) in the Gaussian regression model with respect to the square loss that any aggregation procedure taking its values only in F are necessarily suboptimal: there exists two absolute positive constants c_0 and c_1 such that for any aggregation procedure $\widehat{f}_n$ with values in F (called a *selector* in [54]) their exists a dictionary F of size M and a probability measure P for (X, Y) such that with $P^{\otimes n}$-probability greater than c_0

$$R(\widehat{f}_n) \geq \min_{f \in F} R(f) + c_1 \sqrt{\frac{\log M}{n}}. \tag{2.3.5}$$

Note that the result in [54] is given in expectation but the same argument holds for a result in deviation as well (like in (2.3.5)). In other words, we have to look at aggregation procedures

that take values in larger set than F and for which we can hope some gain due to convexity. A natural procedure that may come to mind is thus the ERM over the convex hull:

$$\widehat{f}^{ERM-C} \in \operatorname*{argmin}_{f\in \mathrm{conv}(F)} R_n(f) \tag{2.3.6}$$

where the convex hull of F is defined for $\Lambda = \{\lambda \in \mathbb{R}^M : \lambda_j \geq 0 \text{ and } \|\lambda\|_1 = 1\}$ by

$$\mathrm{conv}(F) = \{f_\lambda : \lambda \in \Lambda\} \text{ where } f_\lambda = \sum_{j=1}^M \lambda_j f_j. \tag{2.3.7}$$

It is tempting to believe that $\widehat{f}^{ERM-C}$ is indeed an optimal aggregation procedure and this question was asked to us by P. Massart: Is the ERM procedure over $\mathrm{conv}(F)$ an optimal aggregation procedure for the (MS) aggregation problem? To see why it is tempting to believe that $\widehat{f}^{ERM-C}$ is indeed optimal, we consider the square loss function $\ell(f,(x,y)) = (y-f(x))^2$ and a noiseless target function $T(X) : \Omega \to \mathbb{R}$ (the function learning setup). Set $f_F^* \in \mathrm{argmin}_{f\in F} R(f)$ and observe that f_F^* minimizes the $L_2(P_X)$ distance between T and F since $R(f) = \mathbb{E}(f(X)-T(X))^2$. The motivation to consider the ERM $\widehat{f}^{ERM-C}$ over $\mathcal{C} = \mathrm{conv}(F)$ is natural, since one can expect that $\min_{h\in\mathcal{C}} \|h-T\|_{L_2(P_X)} = \|f_{\mathcal{C}}^* - T\|_{L_2(P_X)}$ is much smaller than $\|f_F^* - T\|_{L_2(P_X)}$ (i.e. there should be some gain in the approximation term). Moreover, it is reasonable to think that empirical risk minimization performed in $\mathcal{C}$ has a relatively small error rate, which we denote by $c_1(\delta)\Psi(n,M)$, compared to the gain in the approximation term. Therefore, the ERM $\widehat{f}^{ERM-C}$ in $\mathcal{C}$ is such that with probability greater than $1-\delta$

$$\begin{aligned}\|\widehat{f}^{ERM-C} - T\|_{L_2(P_X)}^2 &\leq \|f_{\mathcal{C}}^* - T\|_{L_2(P_X)}^2 + c_1(\delta)\Psi(n,M)\\ &\leq \|f_F^* - T\|_{L_2(P_X)}^2 + c_1(\delta)\Psi(n,M) - \left(\|f_F^* - T\|_{L_2(P_X)}^2 - \|f_{\mathcal{C}}^* - T\|_{L_2(P_X)}^2\right),\end{aligned}$$

and the hope is that the gain in the approximation error

$$\|f_F^* - T\|_{L_2(P_X)}^2 - \|f_{\mathcal{C}}^* - T\|_{L_2(P_X)}^2$$

is far more significant than $\Psi(n,M)$, leading to a "fast" aggregation rate.

Although this approach is tempting, it has serious flaws. First of all, it turns out that the statistical error of empirical minimization in a convex hull of M well chosen functions may be as bad as $1/\sqrt{n}$ (see Theorem 2.3.1 below). Second, it is possible to construct such a class and a target for which $\|f_F^* - T\|_{L_2(P_X)} = \|f_{\mathcal{C}}^* - T\|_{L_2(P_X)}$, and thus, there is no gain in the approximation error by passing to the convex hull.

The class we shall construct is $F = \{0, \pm\phi_1, \cdots, \pm\phi_M\}$ where $(\phi_i)_{i=1}^{M+1}$ is a specific orthonormal family of $L_2([0,1], \sigma(X), P_X)$ and the target Y is $\phi_{M+1}(X)$, implying that $f_F^* = f_{\mathcal{C}}^* = 0$. For this choice of (F,Y,X) one can show for instance that $\Psi(n,M) \geq c_1/\sqrt{n}$ when $M = \sqrt{n}$ and for a suitable absolute constant c_1 (see Theorem 2.3.1 below). And, since there is no gain in the approximation term, the resulting convergence rate will be of the order of $1/\sqrt{n}$ for the aggregation of $\sqrt{n}$ functions which is suboptimal since the optimal rate of aggregation in this case is of the order of $(\log n)/n$. We now formulate this result showing that the ERM over the convex hull is a sub-optimal aggregation procedure in the bounded regression model with respect to the square loss function.

Theorem 2.3.1 ([P14, P18]) *There exist two absolute positive constants c_0 and c_1 for which the following holds. For any integer n and M such that $\log M \leq c_0 n^{1/3}$, there exists a dictionary*

F of cardinality M and a probability distribution P for (X,Y) such that, with $P^{\otimes n}$-probability greater than $3/4$

$$R(\tilde{f}^{ERM-C}) \geq \min_{f\in F} R(f) + c_2\psi_n(M),$$

where $\psi_n(M) = M/n$ when $M \leq \sqrt{n}$ and $\left(n\log(eM/\sqrt{n})\right)^{-1/2}$ when $M > \sqrt{n}$.

Note that the residual term $\psi_n(M)$ of Theorem 2.3.1 is much larger than the optimal rate $\psi_n^{(MS)}(M) = (\log M)/n$ for the (MS) aggregation problem. It shows that ERM in the convex hull satisfies a much stronger lower bound than the one mentioned in (2.1.4) that holds for any algorithm. This result is of particular importance since at a first glance it was conjectured that $\tilde{f}^{ERM-C}$ could be an optimal aggregation procedure for the (MS) aggregation problem. Theorem 2.3.1 shows that this is not the case (unless when M is like a constant since an upper bound for $\tilde{f}^{ERM-C}$ is proved in Subsection 2.8 with a residual term M/n for $M \leq \sqrt{n}$ which is of the same order as $\log M/n$ when M is a constant). Therefore, the ERM over $\mathrm{conv}(F)$ is not an optimal aggregation procedure for the (MS) aggregation problem.

The proof of Theorem 2.3.1 requires two separate arguments (as in the proofs of the lower bounds in [120] and [103]). The case $M \leq \sqrt{n}$ is easier than the other case $M > \sqrt{n}$. Some hint for this proof are given in Subsection 2.8 where it is in particular showed that the rate $\psi_n(M)$ is indeed optimal for the special example of Theorem 2.3.1.

Finally to answer the question of this subsection: taking the convex hull indeed improves the geometry of the aggregation problem but it increases so much its complexity that the resulting aggregation procedure $\widehat{f}^{ERM-C}$ is not optimal. Therefore, we keep this idea of taking the convex hull since it can provide a gain in the approximation term but in the next subsection we will perform it on a relative small subset of F which is empirically selected.

2.4 A good trade-off geometry/complexity: the convex hull of the set of almost ERM

Fortunately, not all is lost as far as using empirical risk minimization in a convex hull, but one has to be more careful in selecting the set in which it is performed. The key point is to identify situations in which there is a significant gain in the approximation error by passing to the convex hull.

Assume that there are at least two functions in F that almost minimize the risk function $R(\cdot)$ in F in the non-noisy function learning setup. For the square loss function, this means that there are at least two functions in F almost minimizing the L_2 distance between the target T and F. Also, assume that these two functions are relatively "far away" from each other in L_2. By the parallelogram equality (or by a 2-convexity argument for a more general loss function), if f_1 and f_2 are "almost minimizers" then

$$\begin{aligned}\left\|\frac{f_1+f_2}{2} - T\right\|^2_{L_2(P_X)} &\leq \frac{1}{2}\|f_1 - T\|^2_{L_2(P_X)} + \frac{1}{2}\|f_2 - T\|^2_{L_2(P_X)} - \frac{1}{4}\|f_1 - f_2\|^2_{L_2(P_X)} \\ &\approx \|f_F^* - T\|^2_{L_2(P_X)} - \frac{1}{4}\|f_1 - f_2\|^2_{L_2(P_X)}.\end{aligned}$$

Thus, if F_1 is the set of all the almost minimizers in F of the distance to T and the diameter of F_1 is large (to be precise, larger than $c\sqrt{(\log M)/n}$), the approximation error in the convex hull of F_1 is significantly smaller than in F. On the other hand, one can show that if the diameter of F_1 is smaller that $c\sqrt{(\log M)/n}$, the empirical risk minimization algorithm in $\mathrm{conv}(F_1)$ has a fast

error rate (because one has a very strong control on the variances of the various loss functions associated with this set — the diameter being an upper bound for these variances). Therefore, in both cases — but for two completely different reasons — if $\tilde{f}_n$ is the empirical minimizer performed in the convex hull of F_1 then $\|\tilde{f}_n - T\|^2_{L_2(P_X)} \leq \|f^*_F - T\|^2_{L_2(P_X)} + c(\delta)(\log M)/n$, with probability greater than $1-\delta$.

Naturally, using F_1 is not realistic because it is impossible to identify the set of almost true minimizers of the risk in F using the given data. However, it turns out that one can replace F_1 with a set that can be determined empirically and has similar properties to F_1. We now introduce such a set. Let $x > 0$ be the confidence bound that we want to achieve. For simplicity, we consider a sample $\mathcal{D} = (X_i, Y_i)_{i=1}^{2n}$ of size $2n$. We split $\mathcal{D}$ into two sub-samples, $\mathcal{D}_1 = (X_i, Y_i)_{i=1}^{n}$ and $\mathcal{D}_2 = (X_i, Y_i)_{i=n+1}^{2n}$. We use $\mathcal{D}_1$ to define a random subset of F:

$$\widehat{F}_1 = \left\{ f \in F : R_n(f) \leq R_n(\widehat{f}) + C_1 \max\left\{ \alpha \|\widehat{f} - f\|_{L_2^n}, \alpha^2 \right\} \right\}, \tag{2.4.1}$$

where C_1 is a constant to be named later that depends only on the loss function ℓ and b, $R_n(f) = n^{-1}\sum_{i=1}^n \ell(f,(X_i,Y_i))$ is the empirical risk constructed over $\mathcal{D}_1$, $\widehat{f}$ is a minimizer of the empirical risk $R_n(\cdot)$ in F, L_2^n is the L_2 space endowed by the random empirical measure $n^{-1}\sum_{i=1}^n \delta_{X_i}$ and $\alpha = ((x+\log M)/n)^{1/2}$. To make the exposition of our results easier to follow, we avoided presenting the computation of explicit values of constants. Our analysis showed that one can take $C_1 = 4\|\ell\|_{\rm lip}(1+9b)$ — which, of course, is not likely to be the optimal choice of C_1. The set $\widehat{F}_1$ is an empirical approximating set of the set of almost minimizers of the risk in F. We hope that taking the convex hull of this set will increase the complexity of the set $\widehat{F}_1$ only when there is a gain in the approximation term.

Once constructed the intermediate set $F \subset \mathrm{conv}(\widehat{F}_1) \subset \mathrm{conv}(F)$, the second step in the algorithm is performed using the second part $\mathcal{D}_2$ of the sample $\mathcal{D}$. The algorithm produces the empirical risk minimizer (relative to $\mathcal{D}_2$) in the convex hull of $\widehat{F}_1$:

$$\tilde{f} \in \operatorname*{argmin}_{h \in \mathrm{conv}(\widehat{F}_1)} \frac{1}{n} \sum_{i=n+1}^{2n} \ell(h,(X_i,Y_i)). \tag{2.4.2}$$

Note that considering only the "significant" part of a given class (like we do by using the subset $\widehat{F}_1 \subset F$) is an idea that already appeared, for example, in [74]. In that article, the authors used this idea to construct a sharp data-dependent penalty function which outperforms most of the well known data-dependent penalties like local Rademacher penalties (see [61] and reference therein) that are usually computed over the entire class. However, this type of "random subset" is different from the one we introduce here. Usually, the random subset consists of functions for which the empirical risk is smaller than the empirical risk of the empirical risk minimizer plus a sample-dependent complexity term; this complexity term does not depend on each $f \in F$, but rather, on the entire set. Here, in place of the complexity term we use a "function-dependent" additive term: the empirical L_2^n distance between the function and an empirical risk minimizer.

The argument we have in mind heavily depends on the convexity of the loss function since we hope that some gain may follow from the approximation term. Such an approximation term exists "only" when the class is convex (this is the case of $\mathrm{conv}(\widehat{F}_1)$) and the loss function is convex. Actually, we need more than just convexity for the loss function since we need to be able to quantify the gain in the approximation term. That is why we recall now the definition of 2-convexity.

Definition 2.4.1 *[14] Let $\phi : \mathbb{R} \to \mathbb{R}$ and set $\Psi_\phi : L_2(\mathcal{X} \times \mathbb{R}, P) \to \mathbb{R}$ by $\Psi_\phi(f) = \mathbb{E}\phi(f(X,Y))$. The **modulus of convexity** of Ψ_ϕ is the function δ_ϕ defined by*

$$\delta_\phi(\epsilon) = \inf_{\substack{f,g \in L_2(P) \\ \|f-g\|_2 \geq \varepsilon}} \left\{ \frac{\Psi_\phi(f) + \Psi_\phi(g)}{2} - \Psi_\phi\Big(\frac{f+g}{2}\Big) \right\}. \tag{2.4.3}$$

*We say that Ψ_ϕ is **uniformly convex** with respect to the $L_2(P)$ norm if δ_ϕ is positive for every $\varepsilon > 0$. We say that Ψ_ϕ is **2-convex** when there exists some absolute constant $c_\phi > 0$ such that $\delta_\phi(\varepsilon) \geq c_\phi \varepsilon^2$ for any $\varepsilon > 0$.*

For instance, if $\phi(x) = x^2$ then for every $f \in L_2(P)$, $\Psi_\phi(f) = \|f\|^2_{L_2(P)}$. Thus, using the parallelogram equality, for every $\varepsilon > 0$, $\delta_\phi(\varepsilon) = \varepsilon^2/4$. Note that the assumption that $\delta_\phi(\varepsilon) \geq c_\phi \varepsilon^2$ for every $\varepsilon > 0$ is a quantitative way of ensuring that the functional $\Psi_\phi : L_2(P) \longmapsto \mathbb{R}$ enjoys some convexity properties that are close to the parallelogram equality satisfied by the quadratic function risk $f \longmapsto \|f\|^2_{L_2(P)}$.

Assumption 2.4.1 *Assume that the risk function can be written as $R(f) = \mathbb{E}\ell(f(X), Y)$ for any $f \in L_2(P_X)$ where $\ell : \mathbb{R}^2 \to \mathbb{R}^+$ is a Lipschitz function on $[-b, b]^2$ with a Lipschitz constant $\|\ell\|_{\text{lip}}$. Assume further that there exists a convex function $\phi : \mathbb{R} \to \mathbb{R}^+$ such that for any $f, g \in L_2(P)$, $\mathbb{E}_P \ell(f, g) = \mathbb{E}_P \phi(f - g)$ and that the function $\Psi_\phi : f \to \mathbb{E}_P \phi(f)$ is 2-convex with respect to $L_2(P)$.*

In particular, if $\ell(x, y) = (x - y)^2$ then $\delta_\phi(\varepsilon) \geq \varepsilon^2/4$ and so the quadratic risk satisfies Assumption 2.4.1. For example, if $\ell(x, y) = |x - y|^p$ for $1 < p \leq 2$ then $\phi(x) = |x|^p$ and $\delta_\phi(\varepsilon) \geq [(p-1)/4]\varepsilon^2$ (cf. [93]).

Now, we are in position to state the result on the optimality in deviation of the procedure introduced in (2.4.2) for the problem of (MS) aggregation.

Theorem 2.4.2 ([P14]) *For every b and $\|\ell\|_{\text{lip}}$ there exists a constant c_1, depending only on b and $\|\ell\|_{\text{lip}}$, for which the following holds. For any $x > 0$, every class F of M functions, any target Y (all bounded by b) and any loss ℓ satisfying Assumption 2.4.1, the empirical risk minimizer $\tilde{f}$ over the convex hull of $\widehat{F}_1$ satisfies, with $P^{\otimes 2n}$-probability at least $1 - 2\exp(-x)$,*

$$R(\tilde{f}) \leq \min_{f \in F} R(f) + c_1(1 + x)\frac{\log M}{n},$$

Remark 2.4.3 *Note that the definition of the set $\widehat{F}_1$, and thus the algorithm, depends on the confidence x one is interested in through the factor α. Thus $\tilde{f}$ also depends on the confidence.*

Theorem 2.4.2 and the fact that $(\log M)/n$ is the best rate one can hope for the problem of (MS) aggregation proves that the procedure introduced in (2.4.2) is an optimal aggregation procedure in deviation with confidence δ for any $0 < \delta < 1$ and one can take $c_1(\delta) = c_1(1 + \log(2/\delta))$ for the constant introduced in Definition 2.1.2.

The idea of the proof is based on the study of two cases. Either the diameter of $\widehat{F}_1$ is small, then the ERM over $\text{conv}(\widehat{F}_1)$ will preform very well for some concentration reasons: the variance term will be small because the diameter is small (this is for instance the situation in the example considered in Theorem 2.3.1 of Subsection 2.3 for which the ERM over $\text{conv}(F)$ is suboptimal). Or the diameter of $\widehat{F}_1$ is large, and there will be a major gain in the approximation error by considering functions in the convex hull of $\widehat{F}_1$ (this is for instance the situation considered in Figure 2.1 or Figure 1.1 for which the ERM over F is suboptimal).

To conclude, the set $\mathrm{conv}(\widehat{F}_1)$ realizes some kind of optimal trade-off between complexity and geometry since the complexity of the set $\mathrm{conv}(\widehat{F}_1)$ is much bigger than the one of $\widehat{F}_1$ only in situations where there is a huge gain in the approximation term due to the geometry of $\mathrm{conv}(\widehat{F}_1)$. Whereas in the other situation where complexities of both $\mathrm{conv}(\widehat{F}_1)$ and $\widehat{F}_1$ are comparable then somehow the geometrical situation of the relative position of Y compared to F is good enough so that even the ERM on F could be optimal and that there is no need to look for some gain due to convexity.

$$\tilde{f} \in \mathrm{Arg\,min}_{f \in \mathrm{Conv}(\widehat{F}_1)} R_n(f)$$

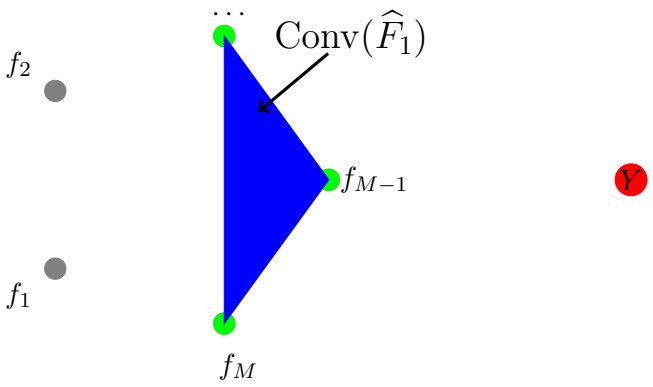

Figure 2.2: Improving the geometry of F without increasing its complexity by too much. Taking the ERM over the convex hull of almost ERM provides an optimal aggregation procedure in deviation.

2.5 Other optimal aggregation procedures

A careful inspection of the proof of Theorem 2.4.2 shows that we don't really need to take the entire convex hull of $\widehat{F}_1$ and that the convexity argument used to prove Theorem 2.4.2 is used only for one segment in $\mathrm{conv}(\widehat{F}_1)$ having a $L_2(P)$ diameter comparable to the diameter of $\widehat{F}_1$. Thus we can consider other "convexification" of the set $\widehat{F}_1$ and then minimize the empirical risk on this set to obtain an optimal aggregation procedure. For instance, in [P5] we propose to minimize the empirical risk over the set of all the segments in $\widehat{F}_1$ or the star-shaped hull of $\widehat{F}_1$ in the ERM $\widehat{f}$. We end up with three optimal aggregation procedures which can be implemented following the steps:

(0. Initialization) Choose a confidence level $x > 0$ and define

$$\alpha = \alpha_{n,M}(x) = b\sqrt{\frac{\log M + x}{n}}.$$

(1. Splitting) Split the sample $\mathcal{D} = (X_i, Y_i)_{i=1}^{2n}$ into $\mathcal{D}_1 = (X_i, Y_i)_{i=1}^{n}$ and $\mathcal{D}_2 = (X_i, Y_i)_{i=n+1}^{2n}$.

(2. Preselection) Use $\mathcal{D}_1$ to define a random subset of F :

$$\widehat{F}_1 = \Big\{ f \in F : R_{n,1}(f) \le R_{n,1}(\widehat{f}_{n,1}) + c \max\left(\alpha \left\| \widehat{f}_{n,1} - f \right\|_{n,1}, \alpha^2\right) \Big\},$$

where $\|f\|_{n,1}^2 = n^{-1}\sum_{i=1}^n f(X_i)^2$, $R_{n,1}(f) = n^{-1}\sum_{i=1}^n (f(X_i)-Y_i)^2$, $\widehat{f}_{n,1} \in \operatorname{argmin}_{f\in F} R_{n,1}(f)$.

(3. Aggregation) Choose $\widehat{\mathcal{F}}$ as one of the following sets:

$$\widehat{\mathcal{F}} = \operatorname{conv}(\widehat{F}_1) = \text{ the convex hull of } \widehat{F}_1$$
$$\widehat{\mathcal{F}} = \operatorname{seg}(\widehat{F}_1) = \text{ the segments between the all functions in } \widehat{F}_1$$
$$\widehat{\mathcal{F}} = \operatorname{star}(\widehat{f}_{n,1}, \widehat{F}_1) = \text{ the segments between } \widehat{f}_{n,1} \text{ with the elements of } \widehat{F}_1,$$

and return the ERM relative to $\mathcal{D}_2$:

$$\tilde{f} \in \operatorname*{argmin}_{g\in\widehat{\mathcal{F}}} R_{n,2}(g),$$

where $R_{n,2}(f) = n^{-1}\sum_{i=n+1}^{2n}(f(X_i) - Y_i)^2$.

These algorithms are illustrated in Figures 2.3 and are optimal aggregation procedures as it is stated in the following theorem.

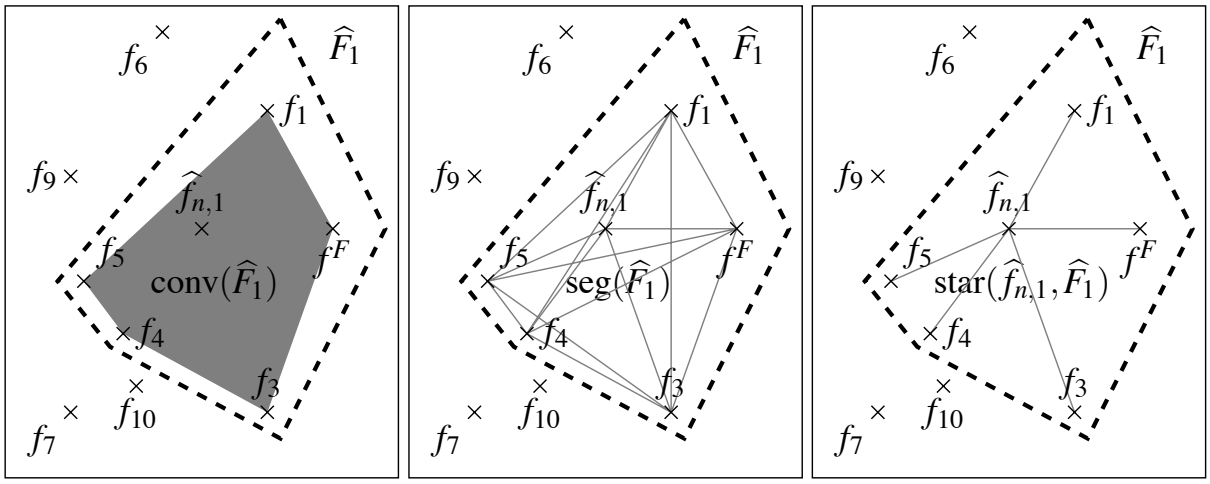

Figure 2.3: Aggregation algorithms: ERM over $\operatorname{conv}(\widehat{F}_1)$, $\operatorname{seg}(\widehat{F}_1)$, or $\operatorname{star}(\widehat{f}_{n,1}, \widehat{F}_1)$.

Theorem 2.5.1 ([P5]) *For every b and $\|\ell\|_{\text{lip}}$ there exists a constant c_1, depending only on b and $\|\ell\|_{\text{lip}}$, for which the following holds. For any $x > 0$, any class F of M functions, any target Y all bounded by b and any loss ℓ satisfying Assumption 2.4.1, the three ERM $\tilde{f}$ over $\operatorname{conv}(\widehat{F}_1)$ or $\operatorname{seg}(\widehat{F}_1)$ or $\operatorname{star}(\widehat{f}_{n,1}, \widehat{F}_1)$ as defined in the previous algorithm satisfy, with $P^{\otimes 2n}$-probability at least $1 - 2\exp(-x)$,*

$$R(\tilde{f}) \le \min_{f\in F} R(f) + c_1(1+x)\frac{\log M}{n},$$

Another optimal aggregation procedure was constructed in [7] which is based on the same idea of improving the geometry of F without increasing its complexity by too much. This procedure is called the empirical star algorithm and is constructed in two steps. First, run the ERM over F: $\widehat{f}_n \in \operatorname{argmin}_{f\in F} R_n(f)$. Then consider the star-shaped hull $\operatorname{star}(F, \widehat{f}_n) = \cup_{f\in F}[f, \widehat{f}_n]$ of all the segments $[f, \widehat{f}_n]$ for $f \in F$ and run the ERM over $\operatorname{star}(F, \widehat{f}_n)$:

$$\tilde{f}_n \in \operatorname*{argmin}_{f\in \operatorname{star}(F,\widehat{f}_n)} R_n(f). \tag{2.5.1}$$

It is proved in [7] that in the bounded regression model with respect to the square loss, for any dictionary F of cardinality M and any $x > 0$, with probability greater than $1 - \exp(-x)$, the empirical star algorithm $\tilde{f}_n$ satisfies

$$R(\tilde{f}_n) \le \min_{f\in F} R(f) + c_0 \frac{x + \log M}{n}.$$

This procedure shares the same idea as the three other optimal aggregation procedures introduced previously: obtaining some gain in the approximation term thanks to a convexity argument by "improving the geometry of F" without increasing the complexity of the set F by too much, which is the case since the complexity of a star-shaped hull of a set is "comparable" to the complexity of the set itself. Indeed, a key point in the proof of [7] is the parallelogram identity satisfied by the square risk which provides a way of quantifying the gain in the approximation term. This argument is in the same spirit as the 2-convexity assumption which is one way of considering risk functions satisfying some kind of parallelogram identity/inequality. Finally, it is interesting to note that the procedure (2.5.1) does not require to fit any constant whereas the three aggregation procedures that we proposed require to fit a constant in the preselection step that is for the construction of $\widehat{F}_1$.

2.6 Suboptimality of the aggregate with exponential weights for low temperatures

It is now well understood that to have any chance of constructing optimal aggregation procedures, one has to consider aggregation procedures taking values in larger sets than F, and the most natural set that may come to mind is the convex hull of F. The aggregate with exponential weights (AEW for short) takes its values in the convex hull of F and has been a very popular candidate for an optimal procedure. It was one of the first procedures to be studied in aggregation theory [54, 8, P11, 71, 39, 6, 118, 42]. The AEW was defined in (2.3.2) but for the reader convenience we recall here its definition

$$\tilde{f}_n^{AEW} = \sum_{j=1}^{M} \widehat{\theta}_j f_j \quad \text{where} \quad \widehat{\theta}_j = \frac{\exp\left(-\frac{n}{T} R_n(f_j)\right)}{\sum_{k=1}^{M} \exp\left(-\frac{n}{T} R_n(f_k)\right)} \tag{2.6.1}$$

for the dictionary $F = \{f_1, \dots, f_M\}$. The parameter $T > 0$ is called the *temperature*[1].

So far, there has been mainly three results surrounding the problem of the optimality of the AEW. First, the progressive mixture rule is optimal in expectation for T larger than some parameters of the model (see [39], [117], [119], [54], [8] or [9]) and under some convexity assumption

[1] This terminology comes from Thermodynamics, since the weights $(\widehat{\theta}_1, \dots, \widehat{\theta}_M)$ can be seen as a Gibbs measure with temperature T on the dictionary F.

on the loss function (cf. [9] for more details and for other procedures related to the progressive mixture rule). This procedure was defined in (2.3.1) and its optimality in expectation is recalled in (2.3.3).

Second, the optimality in expectation of the AEW was obtained in [42] for the regression model $Y_i = f(x_i) + \epsilon_i$ with a deterministic design $x_1, \dots, x_n \in \mathcal{X}$ with respect to the risk $\|g - f\|_n^2 = n^{-1} \sum_{i=1}^n (g(x_i) - f(x_i))^2$ (with its empirical version being $R_n(g) = n^{-1} \sum_{i=1}^n (Y_i - g(x_i))^2$); that is, it was shown that for $T \geq c \max(b, \sigma^2)$ (where σ^2 is the variance of the noise ϵ),

$$\mathbb{E} \left\| \tilde{f}_n^{AEW} - f \right\|_n^2 \leq \min_{g \in F} \|g - f\|_n^2 + \frac{T \log M}{n+1}. \tag{2.6.2}$$

Third, in [2], [6] and [40], the authors proved that in the high temperature regime, the AEW can achieve the optimal rate $(\log M)/n$ under the Bernstein condition both in expectation and in deviation.

Despite its long history, there has been no result on the optimality (or suboptimality) of the AEW in the regression model with random design in the general case (when the dictionary does not necessarily satisfy the Bernstein condition). In this section, we address this issue for the low temperatures regime by proving the following:

- AEW is suboptimal for low temperatures $T \leq c_1$ (where c_1 is an absolute positive constant), both in expectation and in probability, for the quadratic loss function and a dictionary of cardinality 2 (Theorem 2.6.1);
- AEW is suboptimal in probability for some large dictionaries (of cardinality $M \sim \sqrt{n \log n}$) and small temperatures $T \leq c_1$ (Theorem 2.6.2).

Theorem 2.6.1 ([P17]) *There exists absolute positive constants $c_0, \dots, c_5$ for which the following holds. For any integer $n \geq c_0$, there are random variables (X, Y) and a dictionary $F = \{f_1, f_2\}$ such that $(Y - f_i(X))^2 \leq 1$ almost surely for $i = 1, 2$, for which the quadratic risk of the AEW satisfies*

1. *if $T \leq c_1$ and n is odd then*

$$\mathbb{E} R(\tilde{f}_n^{AEW}) \geq \min_{f \in F} R(f) + \frac{c_2}{\sqrt{n}};$$

2. *if $T \leq c_3 \sqrt{n}/\log n$, then with probability greater than c_4,*

$$R(\tilde{f}_n^{AEW}) \geq \min_{f \in F} R(f) + \frac{c_5}{\sqrt{n}}.$$

Theorem 2.6.1 proves that AEW is suboptimal in expectation in the low temperature regime, and suboptimal in probability in both low and high temperature regimes since it is possible to construct procedures that achieve the rate C/n with large probability (cf. [7, P14]) and in expectation (cf. [39, 117, 119, 54, 8, 7]) in the same setup as in Theorem 2.6.1. Note that the problem of the optimality in probability of the progressive mixture rule (and other related procedures) was studied in [7]. Indeed, in [7], it is proved that, for a loss function ℓ satisfying some convexity and regularity assumption (for instance, the quadratic loss used in Theorem 2.6.1) the progressive mixture rule $\bar{f}_n$ defined in (2.3.1) is such that for any temperature parameter, with probability greater than an absolute constant $c_0 > 0$, $R(\bar{f}_n) \geq \min_{f \in F} R(f) + c_1 n^{-1/2}$.

We recall that suboptimality in probability does not imply suboptimality in expectation for the aggregation problem, nor vice-versa. This property of the aggregation problem was first noticed in [7] where the progressive mixture rule (and other related aggregation procedures) was proved to be suboptimal in probability for dictionaries of cardinality two, whereas it was known to be optimal in expectation (cf. [39], [117], [119] or [54]). This peculiarity of the problem of aggregation comes from the fact that an aggregate $\widehat{f}$ is not restricted to take values only in the set F and therefore $R(\widehat{f}) - \min_{f\in F} R(f)$ can take negative values. In [7], it is shown that, for the progressive mixture rule $\bar{f}_n$, in average these negative values do compensate larger values but there is still an event of constant probability on which $R(\bar{f}_n) - \min_{f\in F} R(f)$ takes values greater than $C/\sqrt{n}$.

Another consequence of the lower bounds stated in Theorem 2.6.1 is that AEW cannot be an optimal aggregation procedure both in expectation and probability for low temperatures for the problems of Convex and Linear aggregation, since, we have

$$\min_{f\in F} R(f) \geq \min_{f\in \mathrm{conv}(F)} R(f) \geq \min_{f\in \mathrm{span}(F)} R(f).$$

Also, the optimal rates of aggregation for the Convex and Linear aggregation problems for dictionaries of cardinality two are of the order of n^{-1} (see [103, 57, P18]), while the residual terms obtained in Theorem 2.6.1 are of the order of $n^{-1/2}$ for such a dictionary. Hence, AEW is suboptimal for these two other aggregation problems for low temperatures.

The proof of Theorem 2.6.1 shows that a dictionary consisting of two functions is enough to give the lower bound in expectation in the low temperature regime and in probability in both regimes (small temperatures regime $0 \leq T \leq c_1$ and large temperatures regime $c_1 \leq T \leq c_3\sqrt{n}/\log n$). It is based on the same counter-example introduced in (1.3.3) and Figure 1.1. In the following theorem, we study the behavior of AEW for larger dictionaries. To our knowledge, negative results on the behavior of exponential weights based aggregation procedures are not known for dictionaries with more than two functions, and what we show is that the behavior of the AEW deteriorates, in some sense, as the cardinality of the dictionary grows.

Theorem 2.6.2 ([P17]) *There exists an integer n_0 and absolute constants c_1 and c_2 for which the following holds. For every $n \geq n_0$ there are random variables (X, Y) and a dictionary $F = \{f_1, \cdots, f_M\}$ of cardinality $M = \lceil c_1\sqrt{n\log n}\rceil$ for which the quadratic loss function of any element in F is bounded by 2 almost surely, and for every $0 < \alpha \leq 1/2$, if $T \leq c_2\alpha$, then with probability at least $1 - c_3(\alpha)n^{\alpha-1/2}$,*

$$R(\tilde{f}_n^{AEW}) \geq \min_{f\in F} R(f) + c_4(\alpha)\sqrt{\frac{\log M}{n}}.$$

Moreover, if $f_F^* \in F$ denotes the optimal function in F with respect to the quadratic loss (the oracle), then there exists $f_j \neq f_F^*$ whose excess risk is larger than $c_5(\alpha)n^{-1/2}$ and for which the weight of f_j in the AEW procedure satisfies $\widehat{\theta}_j \geq 1 - n^{-c_6(\alpha)/T}$.

Theorem 2.6.2 implies that the AEW procedure might cause the weights to concentrate around a "bad" element in the dictionary (that is, an element whose risk is larger than the best in the class by at least $\sim n^{-1/2}$) with high probability. In particular, Theorem 2.6.2 gives additional evidence that the AEW procedure is suboptimal for low temperatures.

The analysis of the behavior of AEW for dictionary of cardinality larger than two is considerably harder than the two-function case, and it requires some results on rearrangement of independent random variables which are almost Gaussian.

2.7 The aggregation problem under the Margin Assumption and the Bernstein condition

Fortunately, not all is lost as far as optimality results for AEW go. In this section, we show that under the Bernstein condition, AEW can achieve fast rates (rates faster than $1/\sqrt{n}$); and the same holds for the ERM over F. In fact, both procedures can even adapt to the "real complexity" of the dictionary.

Intuitively, a good aggregation scheme should be able to ignore the elements in the dictionary whose risk is far from the optimal risk in F, or at least the impact of such elements on the function produced by the aggregation procedure should be small. Hence, a good procedure is one whose residual term is of the order of ψ/n, where ψ is a complexity measure that is determined only by the complexity of the set of "almost minimizers" in the dictionary.

By using the PAC-Bayesian approach, it was shown in [2], [6] and [40] that in the high temperature regime (T greater than a constant), AEW can adapt to the real complexity of the dictionary assuming that the class satisfies the Bernstein condition.

The Bernstein condition is very natural in the context of ERM because it has two consequences. Firstly, the empirical excess risk has better concentration properties around the excess risk, and secondly, the complexity of the subset of F consisting of almost minimizers is smaller under this condition. As a consequence, if the class $\mathcal{L}_F$ is a (β, B)-Bernstein class for $0 < \beta \leq 1$, then the ERM algorithm can achieve fast rates (see, for example [15], and references therein). As the results below show, the same is true for AEW. Indeed, under a Bernstein condition, it was proved in ([2], [6] or [40]) that if $R(\cdot)$ is a convex risk function and if F is such that $|\ell(f, Z)| \leq b$ almost surely for any $f \in F$ then for every $T \geq c_1 \max\{b, B\}$ and $x > 0$, with probability greater than $1 - 2\exp(-x)$,

$$R(\tilde{f}_n^{AEW}) \leq \min_{f \in F} R(f) + \frac{Tc_2}{n}\Big(x + \log\Big(\sum_{f \in F} \exp\big(-(n/2T)(R(f) - R(f_F^*))\big)\Big)\Big). \tag{2.7.1}$$

Although the PAC-Bayesian approach cannot be used to obtain (2.7.1) in the low temperature regime ($T \leq c_1 \max\{b, B\}$), such a result is not surprising. Indeed, since fast error rates for the ERM are to be expected when the underlying excess loss functions class satisfies the Bernstein condition and since AEW converges to the ERM when the temperature T tends to zero, it is likely that for "small values" of T, AEW inherits some of the properties of ERM, for example, fast rates under a Bernstein condition. This is what we show in the following theorem.

Before formulating Theorem 2.7.1, let us introduce the following measure of complexity. For every $r > 0$, let

$$\begin{aligned}\psi(r) =& \log(|\{f \in F : R(f) - R(f_F^*) \leq r\}| + 1) \\ &+ \sum_{j=1}^{\infty} 2^{-j} \log(|\{f \in F : 2^{j-1} r < R(f) - R(f_F^*) \leq 2^j r\}| + 1),\end{aligned}$$

where $|A|$ denotes the cardinality of the set A. Observe that $\psi(r)$ is a weighted sum of the number of elements in F that assigns smaller and smaller weights to functions whose excess risk is relatively large.

Theorem 2.7.1 ([P17]) *There exists absolute constants c_0, c_1, c_2 and c_3 for which the following holds. Let $F \subset \mathcal{F}$ be a finite model such that $|\ell(f, Z)| \leq b$ a.s. for any $f \in F$ and the excess loss class $\mathcal{L}_F$ is a $(1, B)$-Bernstein class with respect to Z. If the risk function $R(\cdot)$ is convex and*

if $T \leq c_0 \max\{b, B\}$*, then for every* $x > 0$*, with probability at least* $1 - 2\exp(-x)$*, the function* $\tilde{f}_n^{AEW}$ *produced by the AEW algorithm satisfies*

$$R(\tilde{f}_n^{AEW}) \leq R(f_F^*) + c_1(b+B)\frac{x+\psi(\theta)}{n},$$

where $\theta = c_2(b+B)(\log|F|)/n$*. In particular, we have*

$$\mathbb{E}R(\tilde{f}_n^{AEW}) \leq R(f_F^*) + c_3(b+B)\frac{\psi(\theta)}{n}.$$

In other words, the scaling factor θ we use is proportional to $(b+B)(\log|F|)/n$, and if the class is regular (in the sense that the complexity of F is well spread and not concentrated just around one point), $\psi(\theta)$ is roughly the cardinality of the elements in F whose risk is at most $\sim (b+B)(\log|F|)/n$.

Observe that for every $r > 0$, $\psi(r) \leq c\log|F|$ for a suitable absolute constant c. Therefore, if T is reasonably small — below a level proportional to $\max\{B, b\}$, the resulting aggregation rate is the optimal one, proportional to $(b+B)(x+\log M)/n$ with probability of $1-2\exp(-x)$, and proportional to $(b+B)(\log M)/n$ in expectation.

Although the residual terms in Theorem 2.7.1 and in (2.7.1) are not the same, they are comparable. Indeed, the contribution of each element in F in the residual term depends exponentially on its excess risk.

Theorem 2.7.1 together with the result for high temperatures from [2], [6] and [40] shows that the AEW is an optimal aggregation procedure under the Bernstein condition as long as $T = \mathcal{O}(1)$ when M and n tend to infinity. In general, the residual term one obtains is of the order of $((T+1)\log M)/n$ and it can be proved that the optimal rate of aggregation under the Bernstein condition is proportional to $(\log M)/n$ by using the classical tools in [108].

Let us mention that in the proof of Theorem 2.7.1 we have restricted ourselves to the Berntein parameter $\beta = 1$ simply to make the presentation as simple as possible. A very similar result holds if one assumes a Bernstein condition for any $0 < \beta \leq 1$.

Therefore, much better rates of aggregation can be achieved by the AEW and the ERM (Theorem 2.7.1 holds for $T = 0$ in which case the AEW is the ERM) than the $1/\sqrt{n}$ rates obtained in the lower bounds of Theorem 2.6.1 and Theorem 2.2.1. The scenario is completely different under the Margin assumption. Indeed, in the same general model of Subsection 1.1.1 and Theorem 2.7.1, the following result holds.

Theorem 2.7.2 ([P3, P10]) *Let* $F \subset \mathcal{F}$ *be a finite model and* Z *be a random variable on* $\mathcal{Z}$ *with probability distribution denoted by* P*. Assume that* (F, ℓ, P) *satisfies the "local" Margin assumption with parameters* (β, B) *for some* $0 < \beta \leq 1$ *and* $B > 0$ *and* $|\ell(f, Z) - \ell(f^*, Z)| \leq b$ *a.s. for any* $f \in F$*. Then there exists* c_0 *depending only on* β, B *and* b *such that the risk of the ERM over* F *satisfies*

$$\mathbb{E}R(\widehat{f}_n^{ERM}) \leq \min_{f\in F} R(f) + c_0 \max\left(\sqrt{\frac{\min_{f\in F}\left(R(f)-R(f^*)\right)^{\beta}\log|F|}{n}}, \left(\frac{\log|F|}{n}\right)^{\frac{1}{2-\beta}}\right).$$

In [P10], there are some examples for which the residual term of the oracle inequality satisfied by the ERM in Theorem 2.7.2 is optimal. In particular, it is interesting to note that this residual term depends on the approximation properties of f^* by F through the term $\min_{f\in F}\left(R(f)-R(f^*)\right) = R(f_F^*) - R(f^*)$. In particular, even if the Margin parameter β equals

to one then if $R(f_F^*)$ is greater than $R(f^*)$ by a constant then the aggregation rate can be as bad as $\sqrt{(\log M)/n}$. Therefore, there is no gain by assuming the Margin assumption if the approximation properties of the model F are bad (that is when f^* is badly approximated by the model F).

The aggregation problem is a problem in Learning theory where the model is finite. Thus, if one wants to make an assumption connecting the variance to the expectation, which improves the concentration properties of the empirical (excess) risk to the actual (excess) risk then the natural assumption is the Bernstein condition and not the Margin assumption. Margin assumption is more an assumption in Statistics than an assumption in Learning theory. Nevertheless, since aggregation procedures have been used for some statistical purpose like constructing adaptive procedures it can be useful to analyze aggregation procedures under the Margin assumption. An interesting fact that comes out of this analysis is that in general the Margin assumption does not help in the aggregation setup unless the best minimizer $f^* \in \mathcal{F}$ is close to the model F (or in general when f^* has a risk close to $R(f_F^*)$). In this situation, the Learning problem and the statistical problem become similar, the Margin assumption becomes closer to the Bernstein condition and thus assuming the Margin assumption helps. Otherwise the Margin assumption has no effect on the aggregation problem in general.

2.8 ERM for the convex aggregation problem

We finish this chapter with the problem of Convex aggregation and the study of the ERM over the convex hull $\mathrm{conv}(F)$ introduced in (2.3.6) by

$$\widehat{f}^{ERM-C} \in \operatorname*{argmin}_{f \in \mathrm{conv}(F)} R_n(f).$$

The problem of Convex aggregation is very different from the problem of (MS) aggregation since the class $\mathrm{conv}(F)$ has a very good geometry (this is a convex set) whereas F does not have any geometrical property in general. Therefore, one can hope that running the ERM in this situation can provide an aggregation procedure which achieves the optimal rate of convex aggregation (cf. [103])

$$\psi_n^{(C)}(M) \sim \begin{cases} \frac{M}{n} & \text{if } M \le \sqrt{n} \\ \sqrt{\frac{1}{n}\log\left(\frac{eM}{\sqrt{n}}\right)} & \text{if } M > \sqrt{n}. \end{cases} \tag{2.8.1}$$

Note that the rates obtained in [103] hold in expectation and their optimality holds only in that context (see also the lower bounds in the (C) aggregation context in [53] and [120]). It is worth mentioning that the rate $\psi_n^{(C)}(M)$ was achieved in [103] (in expectation) in the Gaussian regression model with a known variance and a known marginal distribution of the design. In [25], the authors were able to remove these assumptions at a price of an extra $\log n$ factor for $1 \le M \le \sqrt{n}$ (results are still in expectation). Nevertheless, it is not hard to prove that there exists a procedure achieving the rate (2.8.1) with exponentially large probability.

To construct an optimal aggregation procedure in deviation for the convex aggregation problem, we follow the idea of [103]: when $M \le \sqrt{n}$ consider the ERM over the linear span of F; when $M \ge \sqrt{n}$ consider an optimal aggregation procedure for the (MS) aggregation problem and run this procedure on a "good" finite approximating set $\mathcal{C}'$ of the convex hull $\mathrm{conv}(F)$. The construction of $\mathcal{C}'$ follows from the empirical method of Carl-Maurey: $\mathcal{C}'$ is the set of all the convex combination of the elements $f_1, \dots, f_M$ of F having coefficients equal to an integer multiples of $1/m$ where $m = \left\lceil \left(n/\log\left(eM/\sqrt{n}\right)\right)^{1/2} \right\rceil$. Now, consider an optimal aggregation

procedure $\tilde{f}_n$ in deviation for the (MS) aggregation problem as defined in Subsection 2.5 for instance and run it over the dictionary $\mathcal{C}'$. Finally, we consider the aggregation procedure

$$\bar{f}_n = \begin{cases} \widehat{f}_n^{ERM-L} \in \operatorname{argmin}_{f\in \operatorname{span}(F)} R_n(f) \text{ if } M \leq \sqrt{n}, \\ \tilde{f}_n \text{ an optimal aggregation procedure over } \mathcal{C}' \text{ when } M > \sqrt{n}. \end{cases} \tag{2.8.2}$$

The following result shows that this procedure is an optimal aggregation procedure in deviation with an exponential deviation bound. A proof of this result can be found in Chapter 6.

Theorem 2.8.1 *For every b there exists a constant c_1, depending only on b, for which the following holds. For any $x > 0$, every class F of M functions, any target Y (all bounded by b) and for the quadratic loss, the procedure $\bar{f}_n$ defined in (2.8.2) satisfies, with $P^{\otimes 2n}$-probability at least $1 - 2\exp(-x)$,*

$$R(\bar{f}_n) \leq \min_{f\in \operatorname{conv}(F)} R(f) + c_1 \max\left(\psi_n^{(C)}(M), \frac{x}{n}\right),$$

This result also proves that the optimal rate of (C) aggregation in deviation is $\psi_n^{(C)}(M)$. But the procedure $\bar{f}_n$ used to achieve this rate cannot be used in practice since it requires to aggregate an exponential number of elements. It would be much easier and somehow more natural to prove that the ERM over the convex hull $\operatorname{conv}(F)$ is an optimal aggregation procedure for the convex aggregation problem. Moreover, another motivation comes from what is known about ERM in the context of the three aggregation schemes introduced in Subsection 2.1. It follows from Theorem 2.2.1 that the ERM in F is, in general, a suboptimal aggregation procedure for the (MS) aggregation problem. Concerning the (L) aggregation problem, ERM in the linear span of F is an optimal procedure (cf. [57]). Therefore, studying the performances of ERM in the convex hull of F in the context of (C) aggregation can be seen as an "intermediate" problem which may deserve some attention.

The performances of ERM in the convex hull have been studied for an infinite dictionary in [24], in which estimates on its performance have been obtained in terms of the metric entropy of F. The resulting upper bounds were conjectured to be suboptimal in the case of a finite dictionary, since they provide an upper bound M/n for every n and M. And indeed, we establish the following upper bound on the risk of $\tilde{f}^{ERM-C}$ as a (C)-aggregation procedure:

Theorem 2.8.2 ([P18]) *For every $b > 0$ there is a constant $c_1(b)$ and an absolute constant c_2 for which the following holds. Let n and M be integers which satisfy that $\log M \leq c_1(b)\sqrt{n}$. In the bounded regression model with respect to the square loss function (i.e. for any couple (X, Y) and any finite dictionary F of cardinality M such that $|Y|, \sup_{f\in F} |f(X)| \leq b$), for any $x > 0$, with probability greater than $1 - \exp(-x)$,*

$$R(\tilde{f}^{ERM-C}) \leq \min_{f\in \operatorname{conv}(F)} R(f) + c_2 b^2 \max\left[\min\left(\frac{M}{n}, \sqrt{\frac{\log M}{n}}\right), \frac{x}{n}\right].$$

Although Theorem 2.8.2 is new, it is probably known to experts, and its proof is based on what is now, rather standard machinery (cf. for instance [57] or the upcoming Vladimir Koltchinskii's Saint Flour Lecture notes).

Note that the residual term of Theorem 2.8.2 behaves like $\psi_n^{(C)}(M)$ except for values of M for which $n^{1/2} < M \leq c(\epsilon)n^{1/2+\epsilon}$ for $\epsilon > 0$. Although there is a gap in this range in the general case, under the additional assumption that the dictionary is orthogonal, this gap can be removed.

Theorem 2.8.3 ([P18]) *Under the assumptions of Theorem 2.8.2, if $F = \{f_1, \dots, f_M\}$ is such that $\mathbb{E}f_i(X)f_j(X) = 0$ for any $i \neq j \in \{1, \dots, M\}$, then $\tilde{f}^{ERM-C}$ achieves the rate $\psi_n^{(C)}(M)$: for any $x > 0$, with probability greater than $1 - \exp(-x)$*

$$R(\tilde{f}^{ERM-C}) \leq \min_{f \in \mathrm{conv}(F)} R(f) + c_2 b^2 \max\left[\psi_n^{(C)}(M), \frac{x}{n}\right].$$

Removing the gap in the general case is likely to be a much harder problem, although we believe that the orthogonal case is the "worst" one.

Combining Theorem 2.3.1 with Theorem 2.8.2, it follows that up to some logarithmic terms, the rate $\psi_n^{(C)}(M)$ is the same rate of aggregation of $\tilde{f}^{ERM-C}$ for the (MS) and (C) aggregation problems. In particular, it achieves the optimal rate $\psi_n^{(C)}(M)$ for the (C) aggregation problem, up to a logarithmic factor that appears when $\sqrt{n} < M \leq c_1(\epsilon)n^{1/2+\epsilon}$. However, it is far from the optimal rate $(\log M)/n$ for the (MS) aggregation problem.

Chapter 3

Oracle inequalities for ERM, regularized ERM and penalized estimators

In the previous chapter, the geometrical aspect of the aggregation problem was of first importance. This leads us to consider procedures trying "to improve the geometry of the model" because the classical empirical risk minimization procedure or its penalized or regularized versions fail to satisfy exact oracle inequalities with the optimal residual term. In this section, we are interested in these classical procedures: empirical risk minimization procedure and its regularized and penalized versions. We want to understand the aspects of the problem driving the residual terms of the oracle inequalities satisfied by these procedures. In particular, these procedures somehow do not adapt to the geometry of the problem compared to the procedures constructed in the previous section. Consequently, and contrary to the procedures introduced in Chapter 2, some Margin/Bernstein conditions will be "required" to make these classical procedures satisfying oracle inequalities with fast rates.

In particular, we will see that they are some general situations where the Margin/Bernstein condition is trivially satisfied so that the ERM, regularized ERM and penalized estimators satisfy oracle inequalities with fast decreasing residual terms. This is the case when we compare their risk to $(1+\epsilon)\inf_{f\in F} R(f)$ for some $\epsilon > 0$. This kind of oracle inequalities are called non-exact oracle inequalities. In particular, it is much easier to obtain non-exact oracle inequalities with fast residual terms for the classical ERM based procedures than constructing procedures having a risk as close as possible to the the best possible risk $\inf_{f\in F} R(f)$ as we did in the previous chapter. At a first glance, the difference does not look that important but in fact, we will see in this chapter that this far from being the case.

Another important part of this chapter is devoted to the construction of regularizing and penalty functions. Given a family of models or a criterion, we will see how to construct penalty and regularizing functions in terms of some isomorphic properties of some functions classes. These oracle inequalities will be applied in Chapter 4 to concrete examples.

Finally, we study the ERM, regularized ERM and penalized estimators for the three different problems introduced in Section 1.2.1: non-exact prediction problem, exact prediction problem and estimation problem. We will see that depending on the goal pursued, different assumptions will be introduced, different residual terms will be obtained and different regularizing and penalty functions will come out of our study. One central aspect of our work relies on the isomorphic profile of some functions classes.

3.1 Isomorphic profile of functions classes

If we knew the value of the risk function $R(\cdot)$ over F then the problem would be done since a minimizer over F of the true risk would be an optimal procedure for the three problems introduced in Section 1.2.1. But we don't know the value of $R(\cdot)$ over F, we only have access to an empirical version of this function: the empirical risk $R_n(f) = n^{-1}\sum_{i=1}^n \ell(f, Z_i), \forall f \in F$. Therefore, a natural problem is to relate the empirical risk to the actual risk uniformly over F. This may follow from a uniform control over F of the deviation of the empirical risk around the actual risk: find a sharp bound on the quantity

$$\sup_{f\in F} \big|R_n(f) - R(f)\big| = \|P - P_n\|_{\ell_F} = \sup_{\ell\in\ell_F} \big|(P_n - P)\ell\big|. \tag{3.1.1}$$

In particular, since the aim is to minimize the risk over F a natural procedure is the ERM over F. A bound on the quantity (3.1.1) provides an excess risk bound for the ERM:

$$\begin{aligned}
&R(\widehat{f}_n^{ERM}) - \inf_{f\in F} R(f)\\
&= R(\widehat{f}_n^{ERM}) - R_n(\widehat{f}_n^{ERM}) + R_n(\widehat{f}_n^{ERM}) - R_n(f_F^*) + R_n(f_F^*) - R(f_F^*)\\
&\le 2\sup_{f\in F} \big|R_n(f) - R(f)\big|.
\end{aligned}$$

This is the strategy developed in [115]. It usually provides excess risk bounds for the ERM of the order of $1/\sqrt{n}$ under some complexity assumptions on F. Since then some refinement have improved this bound. In particular, better risk bounds of the order of $1/n$ have followed from the localization argument under some Margin/Bernstein condition. This approach will be detailed in the next section.

Another idea based on the isomorphy between the empirical and the actual structures has been introduced in [15]. An excess risk bound for the ERM follows from the following isomorphic property: for some $0 < \eta < 1$, there exists $r_\eta^* > 0$ such that, with high probability, for all $f \in F$, if $P\mathcal{L}_f \ge r_\eta^*$ then

$$(1-\eta)P\mathcal{L}_f \le P_n\mathcal{L}_f \le (1+\eta)P\mathcal{L}_f \tag{3.1.2}$$

where $\mathcal{L}_f = \ell_f - \ell_{f_F^*}, \forall f \in F$. Indeed, it follows from (3.1.2) that the ERM $\widehat{f}_n^{ERM}$ over F has an excess risk such that, with high probability,

$$R(\widehat{f}_n^{ERM}) - \inf_{f\in F} R(f) = P\mathcal{L}_{\widehat{f}_n^{ERM}} \le (1+\eta)P_n\mathcal{L}_{\widehat{f}_n^{ERM}} + r_\eta^* \le r_\eta^*. \tag{3.1.3}$$

Therefore, finding the smallest "level of isomorphy" between the empirical and the actual excess risk functions is of particular interest to derive excess risk bounds for the ERM. We introduce the quantity: for any $0 < \eta < 1$,

$$r^*(\mathcal{L}_F)_\eta = \inf\left(r > 0 : \forall \mathcal{L} \in \mathcal{L}_F, \begin{array}{c} P\mathcal{L} > r \Rightarrow (1-\eta)P\mathcal{L} \le P_n\mathcal{L} \le (1+\eta)P\mathcal{L} \\ P\mathcal{L} \le r \Rightarrow \big|P_n\mathcal{L} - P\mathcal{L}\big| \le \eta r \end{array}\right),$$

where $\mathcal{L}_F = \{\mathcal{L}_f : f \in F\}$ is the excess loss functions class indexed by F. The quantity $r^*(\mathcal{L}_F)_\eta$ is called the **isomorphic profile of** $\mathcal{L}_F$. In the next section, we will see how to derive bound on $r^*(\mathcal{L}_F)_\eta$ that hold with high probability. Note that for exact oracle inequalities the parameter $0 < \eta < 1$ does not play no important role and one should take in general $\eta = 1/2$.

So far, we have been interested only in the first problem: comparing the risk of the ERM to $\inf_{f\in F} R(f)$. It appeared that $r^*(\mathcal{L}_F)_\eta$ bounds the difference $R(\widehat{f}_n^{ERM}) - \inf_{f\in F} R(f)$. If we

are interested in the two other problems — non-exact prediction and estimation — then other quantities are of interest.

Consider the isomorphic profile of the loss functions class $\ell_F = \{\ell_f : f \in F\}$:

$$r^*(\ell_F)_\eta = \inf\left(r > 0 : \forall \ell \in \ell_F, \begin{array}{c} P\ell > r \Rightarrow (1-\eta)P\ell \leq P_n\ell \leq (1+\eta)P\ell \\ P\ell \leq r \Rightarrow \left|P_n\ell - P\ell\right| \leq \eta r \end{array}\right).$$

Then the risk of the ERM over F is such that

$$\begin{aligned} R(\widehat{f}_n^{ERM}) - \frac{1+\eta}{1-\eta}R(f_F^*) &\leq \frac{P_n\ell_{\widehat{f}_n^{ERM}}}{1-\eta} + \eta r^*(\ell_F)_\eta - \frac{1+\eta}{1-\eta}\Big(\frac{P_n\ell_{f_F^*}}{1+\eta} - \eta r^*(\ell_F)_\eta\Big) \\ &\leq (1-\eta)^{-1}\big(R_n(\widehat{f}_n^{ERM}) - R_n(f_F^*)\big) + \frac{2\eta}{1-\eta}r^*(\ell_F)_\eta \leq \frac{2\eta}{1-\eta}r^*(\ell_F)_\eta. \end{aligned} \tag{3.1.4}$$

Consider the excess loss functions class $\mathcal{E}_F = \{\ell_f - \ell_{f^*} : f \in F\}$ with respect to f^* and its isomorphic profile

$$r^*(\mathcal{E}_F)_\eta = \inf\left(r > 0 : \forall \mathcal{E} \in \mathcal{E}_F, \begin{array}{c} P\mathcal{E} > r \Rightarrow (1-\eta)P\mathcal{E} \leq P_n\mathcal{E} \leq (1+\eta)P\mathcal{E} \\ P\mathcal{E} \leq r \Rightarrow \left|P\mathcal{E} - P_n\mathcal{E}\right| \leq \eta r \end{array}\right).$$

Then the excess risk of the ERM over F with respect to f^* is such that

$$R(\widehat{f}_n^{ERM}) - R(f^*) \leq (1-\eta)^{-1}\big(R_n(\widehat{f}_n^{ERM}) - R_n(f^*)\big) + \eta r^*(\mathcal{E}_F)_\eta \tag{3.1.5}$$

$$\leq (1-\eta)^{-1}\big(R_n(f_F^*) - R_n(f^*)\big) + \eta r^*(\mathcal{E}_F)_\eta \leq \frac{1+\eta}{1-\eta}\big(R(f_F^*) - R(f^*)\big) + \frac{\eta(2-\eta)}{1-\eta}r^*(\mathcal{E}_F)_\eta. \tag{3.1.6}$$

Therefore the isomorphic profile of the loss functions class ℓ_F and of the two excess loss functions classes $\mathcal{L}_F$ (with respect to f_F^*) and $\mathcal{E}_F$ (with respect to f^*) are quantities bounding the difference $R(\widehat{f}_n^{ERM}) - (1+\epsilon)\inf_{f\in F} R(f)$ for some $\epsilon > 0$, or $R(\widehat{f}_n^{ERM}) - \inf_{f\in F} R(f)$ or $R(\widehat{f}_n^{ERM}) - R(f^*) - (1+\epsilon)\inf_{f\in F}\big(R(f) - R(f^*)\big)$ respectively. The isomorphic profile is thus some sort of general tool to study the behaviour of the ERM for the three different problems introduced in Section 1.2.1. We thus consider the following definition regarding the isomorphic profile of a functions class.

Definition 3.1.1 ([15]) *Let $\mathcal{Z}$ be a space endowed with a probability measure P and let $Z_1, \dots, Z_n$ be n independent random variables with values in $\mathcal{Z}$, distributed according to P. Let $\mathbf{F}$ be a class of real-valued measurable functions defined on $\mathcal{Z}$ and $0 < \eta < 1$. The **isomorphic profile of** $\mathbf{F}$ is defined by*

$$r^*(\mathbf{F})_\eta = \inf\big(r > 0 : \forall f \in \mathbf{F}, \left|P_n f - Pf\right| \leq \eta \max(Pf, r)\big)$$

In some circumstances, the isomorphic profile of a function class can be equal to 0, meaning that the empirical structure and the actual structure are $(1+\eta)$-isomorphic over the entire set $\mathbf{F}$. This is in particular the case in the Compressed Sensing setup. This property is called the Restricted Isometry Property and is studied in Section 3.8 from the Learning theory point of view. In general, the isomorphic profile is between quantities of the order of $\mathrm{comp}(\mathbf{F})/n$ and $\sqrt{\mathrm{comp}(\mathbf{F})/n}$ where $\mathrm{comp}(\mathbf{F})$ is some complexity measure of $\mathbf{F}$. An isomorphic profile of the order of $\mathrm{comp}(\mathbf{F})/n$ results in an oracle inequality with a residual term of the same order. This kind of rates are called **fast rates**. In general, any rate faster than $1/\sqrt{n}$ is called a fast rate. On the other side, rates slower than $1/\sqrt{n}$ are called **slow rates**.

In what follows, we study the isomorphic profile of the classes ℓ_F, $\mathcal{L}_F$ and $\mathcal{E}_F$. As we have seen so far, bounds on these quantities provide risk bounds for the ERM for the three problems. It appears that the isomorphic profile of other classes of functions will be useful to derive oracle inequalities for regularized ERM and penalized ERM (or penalized estimators). In the next section, we obtain a general bound for the isomorphic profile of any functions class $\mathbf{F}$ thanks to the localization argument.

3.2 Isomorphism, localization and Margin/Bernstein conditions

3.2.1 Isomorphic profile of loss and excess loss functions classes

The isomorphic profile of a functions class measures the "level" above which empirical means and actual means are equivalent. This notion was introduced in [15]. Although it is not necessary, if one wishes the isomorphic property to hold with exponential probability, one can use a high probability deviation bound on the supremum of the localized process. A standard way (though not the only way) of obtaining such a result is through Talagrand concentration inequality [100] applied to localizations of the functions class, combined with a good control of the variance in terms of the expectation (a Margin/Bernstein condition). When applied to the excess loss class $\mathcal{L}_F$ (respectively to $\mathcal{E}_F$), this argument leads to exact oracle inequalities for the ERM (see for example, [84, 16]) for the exact prediction problem (resp. oracle inequalities for the ERM for the estimation problem). If we are interested in non-exact oracle inequalities (for the non-exact prediction problem), we study the isomorphic properties of the loss functions class ℓ_F.

High probability bounds on the isomorphic profile of functions classes can be derived from Talagrand concentration inequality [100]. Since we would like to avoid the assumption that the classes ℓ_F, $\mathcal{L}_F$ or $\mathcal{E}_F$ consist of uniformly bounded functions, an important part of our analysis is the following ψ_1 version of Talagrand inequality [1].

To state this result, we need the following notation. Let G be a class of measurable real-valued functions defined on $\mathcal{Z}$. The supremum of the empirical process indexed by G is denoted by

$$\|P - P_n\|_G = \sup_{g\in G} |(P - P_n)g| \tag{3.2.1}$$

where for every $g \in G$ we set $Pg = \mathbb{E}g(Z)$ and $P_n g = n^{-1}\sum_{i=1}^n g(Z_i)$. Recall that for every $\alpha \geq 1$, the ψ_α norm of $g(Z)$ is

$$\|g(Z)\|_{\psi_\alpha} = \inf\Big(c > 0 : \mathbb{E}\exp\left(|g(Z)|^\alpha / c^\alpha\right) \leq 2\Big).$$

We control the supremum (3.2.1) using the quantities

$$\sigma(G) = \sup_{g\in G}\sqrt{Pg^2} \text{ and } b_n(G) = \Big\| \max_{1\leq i\leq n} \sup_{g\in G} |g(Z_i)| \Big\|_{\psi_1}.$$

Note that for a bounded class G, one has $b_n(G) \leq \sup_{g\in G}\|g\|_\infty$ and in the sub-exponential case, $b_n(G) \lesssim (\log en)\left\|\sup_{g\in G} g\right\|_{\psi_1}$ (this follows from Pisier inequality). Throughout the following chapters, we also use the notation $b_n(g) = \|\max_{1\leq i\leq n} g(Z_i)\|_{\psi_1}$ and for any pseudo-norm $\|\cdot\|$ on $L_2(P)$, we denote by $\mathrm{diam}(G, \|\cdot\|) = \sup_{g\in G}\|g\|$ the diameter of G with respect to $\|\cdot\|$.

Theorem 3.2.1 ([1]) *There exists an absolute constant $K > 0$ for which the following holds. Let $Z_1, \ldots, Z_n$ be n i.i.d. random variables with values in a space $\mathcal{Z}$ and let G be a countable class*

of real-valued measurable functions defined on $\mathcal{Z}$. For every $x > 0$ and $\alpha > 0$, with probability greater than $1 - 4\exp(-x)$,

$$\|P - P_n\|_G \leq (1+\alpha)\mathbb{E}\|P - P_n\|_G + K\sigma(G)\sqrt{\frac{x}{n}} + K(1+\alpha^{-1})b_n(G)\frac{x}{n}.$$

Theorem 3.2.1 can be extended to classes G satisfying some separability property like condition (M) in [77]: *there exists $G' \subset G$ a countable set such that for any $g \in G$ there exists a sequence $(g_k)_{k\in\mathbb{N}}$ of elements in G' such that for any $z \in \mathcal{Z}$, $(g_k(z))_{k\in\mathbb{N}}$ tends to $g(z)$ when k tends to infinity.* We apply Theorem 3.2.1 in this context and it will be implicitly assumed that every time we use Theorem 3.2.1, the separability condition (M) in [77] holds.

To obtain the desired risk or excess risk bounds, we study empirical processes indexed by sets associated with a functions class G, namely, the star-shaped hull of G around zero and the localized subsets for different levels $\lambda \geq 0$ defined by

$$V(G) = \{\theta g : 0 \leq \theta \leq 1, g \in G\} \text{ and } V(G)_\lambda = \{h \in V(G) : Ph \leq \lambda\}.$$

In particular, given a model F and a loss function ℓ, we consider empirical processes indexed by localized star-shaped hull of the loss functions class ℓ_F and of the two excess loss functions classes $\mathcal{L}_F$ and $\mathcal{E}_F$ defined by

$$\ell_F = \{\ell_f : f \in F\}, \quad \mathcal{L}_F = \{\ell_f - \ell_{f_F^*} : f \in F\} \text{ and } \mathcal{E}_F = \{\ell_f - \ell_{f^*} : f \in F\}$$

respectively (assuming that an oracle f_F^* exists in F and a risk minimizer f^* exists in $\mathcal{F}$). In particular, Theorem 3.2.1 will be applied to the localized sets $V(\ell_F)_\lambda$ (resp. $V(\mathcal{L}_F)_\lambda$ and $V(\mathcal{E}_F)_\lambda$) to get non-exact (resp. exact and estimation) oracle inequalities for the ERM algorithm (cf. Section 3.3), to a family $(V(\ell_{F_r})_\lambda)_{r\geq 0}$ (resp. $(V(\mathcal{L}_{F_r})_\lambda)_{r\geq 0}$ and $(V(\mathcal{E}_{F_r})_\lambda)_{r\geq 0}$) to get non-exact (resp. exact and estimation) regularized oracle inequalities for regularized ERM procedures (cf. Section 3.5) and to a family $(V(\ell_m)_\lambda)_{m\in\mathcal{M}}$ (resp. $(V(\mathcal{L}_m)_\lambda)_{m\in\mathcal{M}}$ and $(V(\mathcal{E}_m)_\lambda)_{m\in\mathcal{M}}$) to get non-exact (resp. exact and estimation) model selection oracle inequalities for penalized estimators (cf. Section 3.6).

Observe that Theorem 3.2.1 requires that the envelope function $\sup_{g\in G}|g|$ is bounded in $L_{\psi_1}(P)$ (i.e. sub-exponential), but since $\|\max_{1\leq i\leq n}\zeta_i\|_{\psi_1} \lesssim \|\zeta\|_{\psi_1}\log n$ for identically distributed variables $\zeta_1, \ldots, \zeta_n, \zeta$, it follows that $b_n(\ell_F)$ is not much larger than $\|\sup_{g\in G} g(Z)\|_{\psi_1}$. However, this condition can be a major drawback. For instance, if the set G consists of linear functions indexed by the Euclidean sphere $\mathcal{S}^{d-1}$, and Z is the standard Gaussian vector of $\mathbb{R}^d$, the resulting envelope function is bounded in $\psi_1(P)$, but its norm is of the order of $\sqrt{d}$. Therefore, for high-dimensional statistical problems, where d can be much larger than n, the envelop function of the model may have a bad ψ_1-behaviour. Nevertheless, under some condition on the ψ_1-norm of the envelop of ℓ_F, we can obtain the following upper bound on the isomorphic profile of ℓ_F.

Theorem 3.2.2 ([P16]) *Let $F \subset \mathcal{F}$ be a model and assume that there exists $B_n \geq 0$ such that for every $f \in F$, $P\ell_f^2 \leq B_n P\ell_f + B_n^2/n$. If $0 < \eta < 1/2$ and $\lambda_\eta^* > 0$ satisfy that*

$$\mathbb{E}\|P_n - P\|_{V(\ell_F)_{\lambda_\eta^*}} \leq (\eta/4)\lambda_\eta^*,$$

then for every $x > 0$, with probability larger than $1 - 4e^{-x}$, for every $f \in F$

- *if $P\ell_f > \rho_n(x)$ then $(1-\eta)P\ell_f \leq P_n\ell_f \leq (1+\eta)P\ell_f$,*

- *if* $P\ell_f \le \rho_n(x)$ *then* $\left|P\ell_f - P_n\ell_f\right| \le \eta\rho_n(x)$.

where, for K *the constant appearing in Theorem 3.2.1,*

$$\rho_n(x) = \max\Big(\lambda_\eta^*, \frac{\big(4Kb_n(\ell_F) + (6K)^2 B_n/\eta\big)(x+1)}{n\eta}\Big).$$

Proof. The proof follows the ideas from [15]. Fix $\lambda > 0$ and $x > 0$, and note that by Theorem 3.2.1, with probability larger than $1 - 4\exp(-x)$,

$$\|P - P_n\|_{V(\ell_F)_\lambda} \le 2\mathbb{E}\,\|P - P_n\|_{V(\ell_F)_\lambda} + K\sigma(V(\ell_F)_\lambda)\sqrt{\frac{x}{n}} + Kb_n(V(\ell_F)_\lambda)\frac{x}{n}. \tag{3.2.2}$$

Clearly, we have $b_n(V(\ell_F)_\lambda) \le b_n(\ell_F)$ and

$$\sigma^2(V(\ell_F)_\lambda) = \sup\Big(P(\alpha\ell_f)^2 : 0 \le \alpha \le 1, f \in F, P(\alpha\ell_f) \le \lambda\Big) \le B_n\lambda + B_n^2/n.$$

Moreover, since $V(\ell_F)$ is star-shaped, $\lambda > 0 \to \phi(\lambda) = \mathbb{E}\,\|P - P_n\|_{V(\ell_F)_\lambda}/\lambda$ is non-increasing, and since $\phi(\lambda_\eta^*) \le \eta/4$ and $\rho_n(x) \ge \lambda_\eta^*$ then

$$\mathbb{E}\,\|P - P_n\|_{V(\ell_F)_{\rho_n(x)}} \le (\eta/4)\rho_n(x).$$

Combined with (3.2.2), there exists an event $\Omega_0(x)$ of probability greater than $1 - 4\exp(-x)$, and on $\Omega_0(x)$,

$$\begin{aligned}\|P - P_n\|_{V(\ell_F)_{\rho_n(x)}} &\le (\eta/2)\rho_n(x) + K\sqrt{\frac{(B_n\rho_n(x) + B_n^2/n)x}{n}} + K\frac{b_n(\ell_F)x}{n}\\ &\le \eta\rho_n(x).\end{aligned}$$

Hence, on $\Omega_0(x)$, if $g \in V(\ell_F)$ satisfies that $Pg \le \rho_n(x)$, then $|Pg - P_ng| \le \eta\rho_n(x)$. Moreover, if $P\ell_f = \beta > \rho_n(x)$, then $g = \rho_n(x)\ell_f/\beta \in V(\ell_F)_{\rho_n(x)}$; hence $|Pg - P_ng| \le \eta\rho_n(x)$, and so $(1-\eta)P\ell_f \le P_n\ell_f \le (1+\eta)P\ell_f$. ∎

In particular, it follows from Theorem 3.2.2, that the isomorphic profile of the loss function class ℓ_F under the condition $P\ell_f^2 \le B_nP\ell_f + B_n^2/n, \forall f \in F$ is less than $\rho_n(x)$ with probability greater than $1 - 4\exp(-x)$. Below, the condition "$P\ell_f^2 \le B_nP\ell_f + B_n^2/n, \forall f \in F$" is called the Bernstein condition for the loss functions class ℓ_F. Compared with the classical Bernstein condition for $\mathcal{L}_F$ or the Margin assumption for $\mathcal{E}_F$, this condition is satisfied in very general situations (cf. Lemma 3.2.5 below).

Results similar to the one of Theorem 3.2.2 hold for the excess loss functions classes $\mathcal{L}_F$ and $\mathcal{E}_F$ under the Bernstein condition and the Margin assumption. But since these conditions are far from being trivially satisfied, we state these results depending on the Bernstein and Margin parameters $0 < \beta \le 1$. We first start with a result on the isomorphic profile of the excess loss class $\mathcal{L}_F$ under the Bernstein condition "$P\mathcal{L}_f^2 \le B_n\big(P\mathcal{L}_f\big)^\beta + B_n^2/n, \forall f \in F$" (note that the extra B_n^2/n term does not play any role since in most of the cases there will be no such extra term — as in Definition 1.3.2 — nevertheless, the following result holds with this extra term).

Theorem 3.2.3 *Let* $F \subset \mathcal{F}$ *be a model. Assume that there exists* $0 < \beta \le 1$ *and* $B_n \ge 0$ *such that for every* $f \in F$, $P\mathcal{L}_f^2 \le B_n\big(P\mathcal{L}_f\big)^\beta + B_n^2/n$. *If* $0 < \eta < 1/2$ *and* $\mu_\eta^* > 0$ *satisfy that*

$$\mathbb{E}\|P_n - P\|_{V(\mathcal{L}_F)_{\mu_\eta^*}} \le (\eta/4)\mu_\eta^*,$$

then for every $x > 0$, with probability larger than $1 - 4e^{-x}$, the isomorphic profile of $\mathcal{L}_F$ is such that

$$r^*(\mathcal{L}_F)_\eta \leq \max\Big(\mu^*_\eta, \frac{16}{\eta}\Big(\frac{xB_nK^2}{n}\Big(\frac{4}{\eta}\Big)^\beta\Big)^{\frac{1}{2-\beta}} + \frac{4KB_n\sqrt{x}}{\eta n} + \frac{4Kb_n(\mathcal{L}_F)x}{\eta n}\Big).$$

Finally, we turn to a result on the isomorphic profile of $\mathcal{E}_F$ under the Margin assumption " $P\mathcal{E}_f^2 \leq B_n\big(P\mathcal{E}_f\big)^\beta + B_n^2/n, \forall f \in F$". Like previously, there is an extra additive term B_n^2/n compared to the classical definition of the Margin assumption introduced in Definition 1.3.1. But this term does not play any role.

Theorem 3.2.4 *Let $F \subset \mathcal{F}$ be a model. Assume that there exists $0 < \beta \leq 1$ and $B_n \geq 0$ such that for every $f \in F$, $P\mathcal{E}_f^2 \leq B_n\big(P\mathcal{E}_f\big)^\beta + B_n^2/n$ where $\mathcal{E}_f = \ell_f - \ell_{f^*}, \forall f \in F$. If $0 < \eta < 1/2$ and $\nu^*_\eta > 0$ satisfy that*

$$\mathbb{E}\|P_n - P\|_{V(\mathcal{E}_F)_{\nu^*_\eta}} \leq (\eta/4)\nu^*_\eta,$$

then for every $x > 0$, with probability larger than $1 - 4e^{-x}$, the isomorphic profile of $\mathcal{E}_F$ is such that

$$r^*(\mathcal{E}_F)_\eta \leq \max\Big(\nu^*_\eta, \frac{16}{\eta}\Big(\frac{xB_nK^2}{n}\Big(\frac{4}{\eta}\Big)^\beta\Big)^{\frac{1}{2-\beta}} + \frac{4KB_n\sqrt{x}}{\eta n} + \frac{4Kb_n(\mathcal{E}_F)x}{\eta n}\Big).$$

The proofs of the Theorem 3.2.3 and Theorem 3.2.4 are provided in Chapter 6.

3.2.2 Localization and the Margin/Bernstein condition

In Theorem 3.2.2, the isomorphic profile of ℓ_F relies on a Bernstein-type condition: for every $f \in F$,

$$P\ell_f^2 \leq B_nP\ell_f + B_n^2/n. \tag{3.2.3}$$

A similar assumption on the excess loss functions $\mathcal{L}_f = \ell_f - \ell_{f^*_F}, \forall f \in F$ is a key point in Theorem 3.2.3: for some $0 < \beta \leq 1$,

$$P\mathcal{L}_f^2 \leq B_n\big(P\mathcal{L}_f\big)^\beta + B_n^2/n. \tag{3.2.4}$$

It is up to an extra B_n^2/n term, the Bernstein condition of Definition 1.3.2. Similarly, the same type of assumption is considered in Theorem 3.2.4: for some $0 < \beta \leq 1$,

$$P\mathcal{E}_f^2 \leq B_n\big(P\mathcal{E}_f\big)^\beta + B_n^2/n \tag{3.2.5}$$

where $\mathcal{E}_f = \ell_f - \ell_{f^*}$ is the excess loss function of f with respect to f^*. This is up to an extra B_n^2/n term, the Margin assumption introduced in Definition 1.3.1.

As explained in Section 1.3 for one single variable, this type of assumption plays a key role in the concentration properties of the empirical mean around its actual mean. In the case of an empirical process indexed by a functions class $\big((P - P_n)g\big)_{g\in G}$, this type of property together with the localization argument yield fast concentration rates of the supremum $\|P - P_n\|_G$ around $(1 + \alpha)\mathbb{E}\|P - P_n\|_G$ for some $\alpha > 0$ and for some well-chosen localized sets G.

In contrast to the two other Margin/Bernstein conditions in (3.2.4) and (3.2.5), Assumption (3.2.3) is trivially satisfied when the loss functions are positive and uniformly bounded: if $0 \leq \ell_f \leq B$ then $P\ell_f^2 \leq BP\ell_f$. It also turns out that (3.2.3) does not require any "global" structural assumption on F and is trivially verified if class members have sub-exponential tails.

Lemma 3.2.5 ([P16]) *Let X be a non-negative sub-exponential random variable (i.e. $\|X\|_{\psi_1} < \infty$). Then for every $z \geq 1$,*

$$\mathbb{E}X^2 \leq \log(ez)\,\|X\|_{\psi_1}\,\mathbb{E}X + \frac{\big(4 + 6\log^2(ez)\,\|X\|^2_{\psi_1}\big)}{ez}.$$

In particular, for any function f such that $\ell_f \geq 0$ and $\|\ell_f\|_{\psi_1} \leq D$ for some $D \geq 1$, it follows from Lemma 3.2.5 that, for every $n \geq 1$,

$$\mathbb{E}\ell_f^2 \leq \big(c_0 D\log(en)\big)\mathbb{E}\ell_f + \frac{\big(c_0 D\log(en)\big)^2}{n}.$$

This may be very different for the Bernstein condition. For instance, let us recall a result from [82]. Consider the functional learning problem where one observes a target T at some random points $X_1,\dots,X_n$. Let $1 < p < \infty$ and consider the L_p-loss function $\ell_f(x,y) = |y - f(x)|^p$ and the L_p-risk $R(f) = \mathbb{E}\ell_f(X,T(X)) = \mathbb{E}|f(X) - T(X)|^p$. Let F be a convex compact set of $L_p(P_X)$ such that $\sup_{f\in F}\|F\|_\infty \leq 1$. We consider the excess loss function $\mathcal{L}_f = \ell_f - \ell_{f_F^*}, \forall f \in F$ where $f_F^* \in \operatorname{argmin}_{f\in F} R(f)$. Then for any target T bounded by 1, one has $\mathbb{E}\mathcal{L}_f \leq c(p)\big(\mathbb{E}\mathcal{L}_f\big)^{\beta_p}$ for any $f \in F$ where $\beta_p = \min\big(p/2, 2/p\big)$. This means that the excess loss class $\mathcal{L}_F$ satisfies the Bernstein condition with Bernstein parameter β_p. Since β_p can take any value in $(0,1)$, there exists models and loss functions for which there is a non-trivial Bernstein parameter.

One can wonder why the Margin/Bernstein condition have become so useful in Learning theory and what is the role played by such assumptions on oracle inequalities. For one single variable, we have seen the role of such assumptions on the sub-gaussian term in the Bernstein inequality in Section 1.3. For an empirical process indexed by a functions class, this property is useful when used together with the localization on a star-shaped function class in Talagrand concentration inequality. Indeed, in Talagrand inequality (cf. for instance Theorem 3.2.1), the dominant term in the residue of the deviation inequality of $\|P - P_n\|_G$ with respect to $(1+\alpha)\,\|P - P_n\|_G$ is the subgaussian term:

$$\sigma(G)\sqrt{\frac{x}{n}}. \tag{3.2.6}$$

If no effort is made on this term then the residual term in Talagrand inequality will be of the order of $1/\sqrt{n}$ then yielding oracle inequalities with slow rates. One way of improving this rate is to apply Talagrand inequality to functions g in G such that Pg^2 is small. Therefore, this leads to consider the functions classes indexed by some $\lambda > 0$,

$$G^\lambda = \{g \in G : Pg^2 \leq \lambda\}.$$

The classes $G^\lambda, \lambda \geq 0$ are called the localized sets of G. Applying Talagrand inequalty to the localized set G^λ yields a deviation inequality with a residual term of the order of $\sqrt{\lambda x/n} \leq \lambda + x/n$. Therefore, if $\mathbb{E}\,\|P - P_n\|_{G^\lambda}$ is also of the order of λ then we get with high probability

$$\|P - P_n\|_{G^\lambda} \leq c_0 \max\big(\lambda, x/n\big) = \rho_n(x). \tag{3.2.7}$$

Finding the smallest λ^* such that $\mathbb{E}\,\|P - P_n\|_{G^{\lambda^*}} \leq c_1\lambda^*$ makes the bounds in (3.2.7) even better. That is the reason why fixed points like λ^* play such a central role in this approach.

The bound (3.2.7) for $\lambda = \lambda^*$ provides some concentration result for any function in G^{λ^*}: $\forall g \in G, Pg^2 \leq \lambda^* \Rightarrow |Pg - P_n g| \leq \rho_n(x)$. But since G is star-shaped in zero, any function g

in G such that $Pg^2 > \lambda^*$ can be scaled down to an other function $f = \lambda^* g/\sqrt{Pg^2}$ in G^{λ^*} for which we know some concentration property. It follows that above the level λ^*, the empirical and actual structure are isomorphic. That is the reason why star-shaped classes play such an important role in our approach: it allows to derive properties for the entire set G only from properties on the localized subsets G^λ.

In general, the localized sets $G_\lambda = \{g \in G : Pg \leq \lambda\}$ are "less complex" than the sets $G^\lambda = \{g \in G : Pg^2 \leq \lambda\}$ for $0 < \lambda < 1$. But under the Bernstein condition with Bernstein parameter $\beta = 1$ (i.e. $Pg^2 \leq BPg, \forall g \in G$), the set G^λ has a complexity comparable to the one of G_λ. This yields a smaller fixed point λ^*. This explains the role of the Margin/Bernstein condition in our approach: classes satisfying a Margin/Bernstein condition have smaller localized sets.

Fixed points have been used in Learning theory and Statistics for almost 20 years. Such examples can be found in [21] in terms of the bracketing entropy of the localized models. Other bounds of this type can be found in [109]. Below, the residual term of each oracle inequality depends on a fixed point characterizing the complexity of the model. The fixed points ε_*^2 (cf. [77] or (3.3.2) below) and $\delta_n(x)$ (cf. [57] or (3.3.4) below) are the residual terms of the oracle inequalities satisfied by the ERM in those papers. There are two main ideas to explain the introduction of fixed points in Learning theory. In [57, 62], *iterative localization* of the excess risk of the ERM converges to the fixed point $\delta_n(x)$ (up to some multiplying constant depending on the number of iterations). In [15] and in this document, fixed points are used to characterize the level above which the actual and the empirical structures are isomorphic. For instance in the case of the excess loss functions class $\mathcal{L}_F$ with Bernstein parameter $0 < \beta \leq 1$: with high probability, $\forall f \in F$ s.t. $P\mathcal{L}_f \geq \max\left(\mu^*_{1/2}, c_0 n^{-1/(2-\beta)}\right), (1/2)P_n\mathcal{L}_f \leq P\mathcal{L}_f \leq (3/2)P_n\mathcal{L}_f$. We refer the reader to those papers for more details on the interpretation of fixed points in Learning theory.

Finally, it should be remarked that Margin/Bernstein condition, localization, fixed points and Talagrand concentration inequality are sufficient tools and conditions which allow to get fast residual terms in some oracle inequalities for the ERM and its regularized and penalized versions. They are not by any mean necessary conditions and tools to be used to get those fast rates or to prove oracle inequalities in general (cf. for instance [109] or Section 3.10). A direct approach to study the ERM like in the second part of [15] or Theorem 3.9.1 may provide better bounds than the one following some fixed point argument.

3.2.3 Some bounds on $\mathbb{E}\|P - P_n\|_H$

Let H be the loss functions class ℓ_F or the excess loss functions classes $\mathcal{L}_F$ and $\mathcal{E}_F$ associated with a model F. To obtain oracle inequalities for the ERM, we want to compute the fixed point of the empirical process indexed by the localized sets $V(H)_\lambda$, that is, for some $c_0 < 1$, we want to find a small λ^* for which

$$\mathbb{E}\|P - P_n\|_{V(H)_{\lambda^*}} \leq c_0 \lambda^*. \tag{3.2.8}$$

First note that the complexity of the star-shaped hull $V(H)$ is not far from the one of H itself. Actually, a bound on the expectation of the supremum of the empirical process indexed by $V(H)_\lambda$ follows from a bound on the expectation of the supremum of the empirical process indexed by the localized sets $H_\mu = \{h \in H : \mathbb{E}h \leq \mu\}$ for different levels $\mu \in \{2^i\lambda : i \in \mathbb{N}\}$. This follows from the peeling argument of [16]:

$$V(H)_\lambda \subset \bigcup_{i=0}^{\infty} \{\theta h : 0 \leq \theta \leq 2^{-i}, h \in H_{2^{i+1}\lambda}\}.$$

Therefore, for $R^* = \inf_{h\in H} \mathbb{E}h$,

$$\mathbb{E}\,\|P - P_n\|_{V(H)_\lambda} \leq \sum_{\{i:2^{i+1}\lambda \geq R^*\}} 2^{-i}\mathbb{E}\,\|P - P_n\|_{H_{2^{i+1}\lambda}}, \tag{3.2.9}$$

because for values of i such that $2^{i+1}\lambda < R^*$ the sets $H_{2^{i+1}\lambda}$ are empty. Now, it remains to bound $\mathbb{E}\,\|P - P_n\|_{H_\mu}$ for any $\mu > 0$.

Let us mention that a naive attempt to control these empirical processes using a contraction argument is likely to fail, and will result in slow rates even in very simple cases (for example, a regression model with a bounded design). We refer to [48, 83, 85] for more details.

The bounds obtained below on $\mathbb{E}\,\|P - P_n\|_{H_\mu}$ are expressed in terms of a random metric complexity of H, which is based on the structure of coordinate projections $P_\sigma H$. These random sets are defined for every sample $\sigma = (X_1, \cdots, X_n)$ by

$$P_\sigma H = \{(h(X_1), \ldots, h(X_n)) : h \in H\}.$$

The complexity of these random sets will be measured via a metric complexity called the γ_2-functional. This is the natural complexity measure coming out of the generic chaining mechanism compared to the classical chaining method leading to the Dudley entropy integral.

Definition 3.2.6 ([101]) *Let (T, d) be a semi-metric space. An admissible sequence of T is a sequence $(T_s)_{s\in\mathbb{N}}$ of subsets of T such that $|T_0| \leq 1$ and $|T_s| \leq 2^{2^s}$ for any $s \geq 1$. We define*

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

where the infimum is taken over all admissible sequences $(T_s)_{s\in\mathbb{N}}$ of T.

We refer the reader to [101] for an extensive survey on chaining methods and on the γ_2-functionals. In particular, one can bound the γ_2-functional using Dudley entropy integral

$$\gamma_2(T, d) \lesssim \int_0^{\mathrm{diam}(T,d)} \sqrt{\log N(T, d, \epsilon)}\, d\epsilon \tag{3.2.10}$$

where $N(T, d, \epsilon)$ is the minimal number of balls with respect to d of radius ϵ needed to cover T, and $\mathrm{diam}(T, d)$ is the diameter of the metric space (T, d). We will use the γ_2-functional to state our theoretical bounds because there are examples for which there is a gain at using the γ_2-functional over the Dudley entropy integral (2-convex bodies for instance). Nevertheless, in all our concrete applications, we use the bound (3.2.10) since the computation is easier and the loss is at most logarithmic. But, if one cares about logarithmic terms then the γ_2-functional should be preferred to the Dudley entropy integral in general.

Now, we turn to some concrete examples where H is the loss functions class in the regression model with respect to the L_q-loss for some $q \geq 2$. For any real-valued measurable function f defined on $\mathcal{X}$, the L_q-loss function of f is $\ell_f^{(q)}(x, y) = |y - f(x)|^q$. In this case, the L_q-loss functions class localized at some level μ is $(\ell_F^{(q)})_\mu = \{\ell_f^{(q)} : f \in F, \mathbb{E}\ell_f^{(q)} \leq \mu\}$. The following result is a combination of a truncation argument and Rudelson's L_∞^n method. To formulate it, set $M = \left\|\sup_{\ell\in(\ell_F^{(q)})_\mu} |\ell|\right\|_{\psi_1}$, for any $A \subset \mathbb{R}^d$, let $\widetilde{A} = A \cup -A$, and if $F^{(\mu)} = \{f \in F : P\ell_f^{(q)} \leq \mu\}$, put $U_n = \mathbb{E}\gamma_2^2\left(\widetilde{P_\sigma F^{(\mu)}}, \ell_\infty^n\right)$.

Proposition 3.2.7 ([P16]) *For every $q \geq 2$, there exists a constant c_0 depending only on q for which the following holds. Let $F \subset \mathcal{F}$ be a functions class. For any $\mu > 0$, we have*

1. *if $q = 2$ then $\mathbb{E}\,\|P - P_n\|_{(\ell_F^{(q)})_\mu} \leq c_0 \max\left[\sqrt{\mu\frac{U_n}{n}}, \frac{U_n}{n}\right]$,*
2. *if $q > 2$ then*

$$\mathbb{E}\,\|P - P_n\|_{(\ell_F^{(q)})_\mu} \leq c_0 \max\left[\sqrt{\mu\frac{U_n}{n}}\sqrt{\big(M\log n\big)^{(q-2)/q}}, \frac{U_n}{n}\big(M\log n\big)^{(q-2)/q}, \frac{M\log n}{n}\right].$$

An example of the computation of the complexity term U_n can be found in [P16] for the calibration of the regularizing function in terms of the ℓ_1-norm (used as a criterion function in this example). Consider the family of models $(F_r)_{r\geq 0}$ associated with the ℓ_1-criterion $F_r = \{f_\beta : \|\beta\|_1 \leq r\}$, where $f_\beta(x) = \langle x, \beta\rangle$ is a linear functional on $\mathbb{R}^d$. The following result is used to prove Theorem 4.1.1 below.

Proposition 3.2.8 ([P16]) *There exists an absolute constant c_0 for which the following holds. For every μ and $r \geq 0$, and every $\sigma = (X_1, \ldots, X_n)$,*

$$\gamma_2\Big(\widetilde{P_\sigma F_r^{(\mu)}}, \ell_\infty^n\Big) \leq c_0 r\big(\max_{1\leq i\leq n}\|X_i\|_{\ell_\infty^d}\big)(\log d)\log\Big(\frac{\sqrt{n}}{\log d}\Big).$$

Moreover, if there exists some constant c_d (which may depends only on d) such that $\Big\|\|X\|_{\ell_\infty^d}\Big\|_{\psi_2} \leq c_d$ then $\Big(\mathbb{E}\gamma_2^2\Big(\widetilde{P_\sigma F_r^{(\mu)}}, \ell_\infty^n\Big)\Big)^{1/2} \leq c_0 r c_d (\log n)^{3/2}(\log d)$.

The proof of the first part of the claim is rather standard and has appeared in one form or another in several places (for example, see [16]). It follows from (3.2.10) and the Carl-Maurey empirical method. The second part is an immediate corollary of the first one combined with Pisier inequality.

Another example of the computation of U_n can be found in [P6] for the computation of the complexity of the Schatten balls $rB(S_1), \forall r \geq 0$ (cf. Section 4.3 for the definition of Schatten norms). The following result is used to prove Theorem 4.2.1 below.

Proposition 3.2.9 ([P6]) *There exists an absolute constant $c_0 > 0$ such that if there exists some constant c_{mT} (which may depend only on the product mT) satisfying $\Big\|\|X\|_{S_2}\Big\|_{\psi_2} \leq c_{mT}$, then $\big(\mathbb{E}\gamma_2^2(rB(S_1), \|\cdot\|_{\infty,n})\big)^{1/2} \leq c_0 c_{mT} r\log n$.*

Random complexities like the quantity U_n have been used in Empirical Process theory for many years. For instance, let us recall a result due to Giné and Zinn (cf, e.g. Theorem 3.5 in [115]). Let G be a set of measurable functions from $\mathcal{Z}$ to $\mathbb{R}$ and Z be a random variable on $\mathcal{Z}$ satisfying $\sup_{g\in G}|\mathbb{E}g(Z)| \leq c_0$. Under the appropriate conditions of measurability of G the following are equivalent:

1. $\|P - P_n\|_G \to 0$ a.s..
2. $\mathbb{E}\sup_{g\in G}|g(Z)| < \infty$ and for any $\theta > 0$ and $\epsilon > 0$, when n tends to infinity,

$$\frac{\mathbb{E}\log N(P_\sigma G(\theta), \ell_\infty^n, \epsilon)}{n} \longrightarrow 0,$$

where $G(\theta) = \{t_\theta(g) : g \in G\}$ and t_θ is a threshold function defined by $t_\theta(x)$ equals θ when $x \geq \theta$, equals x when $|x| \leq \theta$ and $-\theta$ when $x \leq -\theta$ for any $x \in \mathbb{R}$ and $\theta > 0$.

In particular, it follows from the key theorem in Learning theory (cf. Section 3.4 in [115]) that, given a model F and a random variable Z, the ERM over F is strictly consistent (cf. definition of strict consistency in Section 3.2 in [115]) if $G = \{\ell_f : f \in F\}$ and Z satisfy one of the previous two points.

3.3 Oracle inequalities for the ERM

It follows from the risk bounds of Section 3.1 in terms of the isomorphic profile of $\mathcal{L}_F$, ℓ_F and $\mathcal{E}_F$ and the bounds on the isomorphic profile of these functions classes in Section 3.2.1 the following risks bounds in terms of the fixed points λ^*_η, μ^*_η and ν^*_η. We start with a non-exact oracle inequality for the ERM for the non-exact prediction problem. The proof follows from (3.1.4) and Theorem 3.2.2.

Theorem 3.3.1 ([P16]) *There exists an absolute constant $c_0 > 0$ for which the following holds. Let $F \subset \mathcal{F}$ be a model assume that there exists $B_n \geq 0$ such that for every $f \in F$, $P\ell_f^2 \leq B_n P\ell_f + B_n^2/n$. Let $0 < \eta < 1$, set $\lambda^*_\eta > 0$ for which*

$$\mathbb{E}\|P_n - P\|_{V(\ell_F)_{\lambda^*_\eta}} \leq (\eta/4)\lambda^*_\eta.$$

Then, for every $x > 0$, with probability greater than $1 - 4\exp(-x)$,

$$R(\widehat{f}_n^{ERM}) \leq \frac{1+\eta}{1-\eta} \inf_{f\in F} R(f) + \frac{2\eta}{1-\eta} \max\Big(\lambda^*_\eta, c_0 \frac{(b_n(\ell_F) + B_n/\eta)(x+1)}{n\eta}\Big).$$

Although the formulation of Theorem 3.3.1 requires that $P\ell^2 \leq B_n P\ell + B_n^2/n, \forall \ell \in \ell_F$, we have seen in Lemma 3.2.5 that if ℓ is non-negative, this condition is trivially satisfied and one may take $B_n \sim \mathrm{diam}(\ell_F, \psi_1)\log(n)$. This type of condition is far from being trivially satisfied for the excess loss functions classes $\mathcal{L}_F = \{\ell_f - \ell_{f^*_F} : f \in F\}$ and $\mathcal{E}_F = \{\ell_f - \ell_{f^*} : f \in F\}$, which is one of the major difference between the problem of non-exact prediction on one side and the problems of exact prediction and estimation on the other side. The classical Bernstein condition for $\mathcal{L}_F$ and the Margin assumption for $\mathcal{E}_F$ are usually not trivially satisfied. That is the reason why, the two following results depends on the Bernstein and Margin parameter $0 < \beta \leq 1$. We start with an exact oracle inequality for the ERM for the exact prediction problem which was obtained in [15] under some boundedness assumption that we extend to the case where the envelope $\sup_{\mathcal{L}\in\mathcal{L}_F} \mathcal{L}$ is sub-exponential. The proof follows from (3.1.3) and Theorem 3.2.3.

Theorem 3.3.2 ([15]) *There exists an absolute constant $c_0 > 0$ for which the following holds. Let $F \subset \mathcal{F}$ be a model and assume that there exists $0 < \beta \leq 1$ and $B_n \geq 0$ such that for every $f \in F$, $P\mathcal{L}_f^2 \leq B_n\big(P\mathcal{L}_f\big)^\beta + B_n^2/n$. Let $0 < \eta < 1$, set $\mu^*_\eta > 0$ for which*

$$\mathbb{E}\|P_n - P\|_{V(\mathcal{L}_F)_{\mu^*_\eta}} \leq (\eta/4)\mu^*_\eta.$$

Then, for every $x > 0$, with probability greater than $1 - 4\exp(-x)$,

$$R(\widehat{f}_n^{ERM}) \leq \inf_{f\in F} R(f) + \max\Big(\mu^*_\eta, \frac{16}{\eta}\Big(\frac{xB_nK^2}{n}\Big(\frac{4}{\eta}\Big)^\beta\Big)^{\frac{1}{2-\beta}} + \frac{4KB_n\sqrt{x}}{\eta n} + \frac{4Kb_n(\mathcal{L}_F)x}{\eta n}\Big).$$

Then, we state an oracle inequality for the ERM for the estimation prediction problem under the Margin assumption. The proof follows from (3.1.5) and Theorem 3.2.4.

Theorem 3.3.3 *There exists an absolute constant $c_0 > 0$ for which the following holds. Let $F \subset \mathcal{F}$ be a model and assume that there exists $0 < \beta \leq 1$ and $B_n \geq 0$ such that for every $f \in F$, $P\mathcal{E}_f^2 \leq B_n\big(P\mathcal{E}_f\big)^\beta + B_n^2/n$. Let $0 < \eta < 1$, set $\nu_\eta^* > 0$ for which*

$$\mathbb{E}\|P_n - P\|_{V(\mathcal{E}_F)_{\nu_\eta^*}} \leq (\eta/4)\nu_\eta^*.$$

Then, for every $x > 0$, with probability greater than $1 - 4\exp(-x)$,

$$R(\widehat{f}_n^{ERM}) - R(f^*) \leq \frac{1+\eta}{1-\eta}\inf_{f\in F}\big(R(f) - R(f^*)\big) + \frac{\eta(2-\eta)}{1-\eta}\max\Big(\nu_\eta^*, \frac{16}{\eta}\Big(\frac{xB_nK^2}{n}\Big(\frac{4}{\eta}\Big)^\beta\Big)^{\frac{1}{2-\beta}} + \frac{4KB_n\sqrt{x}}{\eta n} + \frac{4Kb_n(\mathcal{E}_F)x}{\eta n}\Big).$$

Theorem 3.3.2 and Theorem 3.3.3 share strong similarities with the main result of [77] (cf. Theorem 2 in [77]): if for every $f \in F$, $\|\ell_f\|_\infty \leq 1$ and $\mathbb{E}\mathcal{L}_f^2 \leq B\big(\mathbb{E}\mathcal{L}_f\big)^\beta$ for some $0 \leq \beta \leq 1$ then for every $x \geq 1$, with probability greater than $1 - \exp(-x)$,

$$R(\widehat{f}_n^{ERM}) \leq \inf_{f\in F} R(f) + c_0 x\varepsilon_*^2 \tag{3.3.1}$$

where ε_* is the unique solution of the equation

$$\sqrt{n}\varepsilon_*^2 = \phi(\sqrt{B}\varepsilon_*^\beta) \tag{3.3.2}$$

where $\phi(\lambda) \geq \sqrt{n}\mathbb{E}\sup_{f,g\in F, P(\ell_f-\ell_g)^2\leq\lambda^2}(P - P_n)(\ell_f - \ell_g)$ for any λ such that $\phi(\lambda) \leq \sqrt{n}\lambda^2$ and ϕ is non decreasing, continuous, $\phi(1) \geq 1$ and $x \to \phi(x)/x$ is non increasing.

As an application in Learning theory for the $0-1$ loss function $\ell(f,(x,y)) = \mathbb{1}_{f(x)\neq y}$, an exact oracle inequality for the ERM over a class F of VC dimension $V \leq n$ (cf. [115] or [77] for more details) is derived in [77] with a residual term $r_n(F) \sim \epsilon_*^2$ of the order of $\big(V\log(enB^{1/\beta}/V)/n\big)^{1/(2-\beta)}$ (a similar result can be derived from Theorem 3.3.2 as well). In the same situation, for every $f \in F$, $\mathbb{E}\ell_f^2 \leq \mathbb{E}\ell_f$, therefore, it follows from Theorem 3.3.1 and the argument used to obtain Equation (29) in [77] (or Example 3 in [57]) combined with the peeling argument (3.2.9) that for every $x \geq 1$ and $0 < \eta < 1/2$, with probability greater than $1 - 8\exp(-x)$,

$$R(\widehat{f}_n^{ERM}) \leq (1+2\eta)\inf_{f\in F} R(f) + c_0\frac{xV\log\big(en/V\big)}{\eta^2 n}. \tag{3.3.3}$$

The residual term ϵ_*^2 obtained in [77] is optimal and heavily depends on the parameter β, it ranges between $\sqrt{V/n}$ and V/n (up to a logarithmic factor). In particular, it can be as bad as the *square root* of the residual term of the non-exact oracle inequality (3.3.3) in the same situation. The main difference between the two results is that the condition "$\forall f \in F, \mathbb{E}\ell_f^2 \leq \mathbb{E}\ell_f$" is always satisfied whereas the condition "$\forall f \in F, \mathbb{E}\mathcal{L}_f^2 \leq B\big(\mathbb{E}\mathcal{L}_f\big)^\beta$" depends on the geometry of the system (F, Y) (relative position of Y with respect to F). It is interesting to note that the residual term in (3.3.3) is always a fast rate even for hard classification problem such that $\mathbb{P}[Y = 1|X] = 1/2$. On one hand, this means that the non-exact prediction problem in classification is completely blind to the geometry of the model and to its statistical estimation property. On the other hand, the exact prediction problem is on the contrary heavily dependent of the geometry of (F, Y) and the estimation problem heavily depends on the Margin parameter.

Another related result is the one in [57] where (among other results) an exact oracle inequality is proved for the ERM with a residual term $\delta_n(x)$ expressed in terms of $\phi_n(\delta) =$

$\mathbb{E}\sup_{f,g\in F(\delta)}|(P-P_n)(\ell_f-\ell_g)|$ where $F(\delta)=\{f\in F: P\mathcal{L}_f\leq\delta\}$ and $D(\delta)=\sup_{f,g\in F(\delta)}\sqrt{P(\ell_f-\ell_g)^2}$: (cf. [57] for the general formulation):

$$\delta_n(x)=\operatorname{argmin}\Big(\delta>0:\phi_n(\delta)+\sqrt{\frac{2x}{n}\big(D(\delta)^2+2\phi_n(\delta)\big)}+\frac{x}{2n}\leq c_0\delta\Big). \tag{3.3.4}$$

In [77, 57, 15] the risk bounds were obtained under the boundedness assumption $\sup_{f\in F}\|\ell_f\|_\infty\leq c_0$ which has been a major drawback in the analysis of procedures in Learning theory by using tools from empirical processes (like Talagrand's extension of Bennett's inequality for the supremum of random process or the contraction principle, cf. [66]). In particular, these results do not apply to the Gaussian regression model. The approach that we developed in [P16] provides a slight improvement since risk bounds hold if the envelop $\sup_{f\in F}\ell_f$ is sub-exponential. This is in particular the case for the Gaussian regression model with respect to the square loss. From a technical point of view, we bypassed the boundedness assumption thanks to a result of [1] extending Talagrand concentration inequality to functions classes with a sub-exponential envelop and through upper bounds on the expectation of the supremum of the empirical process using a truncation argument and Rudelson's method for the truncated part. Moreover, the results in [77, 57, 15] and Theorem 3.3.2 are exact oracle inequalities relying on the properties of the excess loss functions class $\mathcal{L}_F$ and the result in Theorem 3.3.3 depends on the properties of $\mathcal{E}_F$, whereas Theorem 3.3.1 provides a non-exact oracle inequality for the non-exact prediction problem for the ERM relying on the property of the loss functions class ℓ_F which is "surprisingly" much simpler. More details on the difference between exact and non-exact oracle inequalities are provided in the next section.

3.4 Differences between exact and non-exact oracle inequalities for the prediction problem

One motivation in obtaining non-exact oracle inequalities (Equation (1.2.1) with $\epsilon>0$ or Theorem 3.3.1) is the observation that one can obtain such an inequality for the ERM with a residual term $r_n(F)$ of the order of $1/n$, while the best residual term achievable by the ERM in an exact oracle inequality (Equation (1.2.1) for $\epsilon=0$ or Theorem 3.3.2) will only be of the order of $1/\sqrt{n}$ for the same problem.

For example, consider the simple case of a finite model F of cardinality M (cf. the (MS) aggregation problem in Chapter 2) and the bounded regression model with the quadratic loss function (that is $Z=(X,Y)\in\mathcal{X}\times\mathbb{R}$ with $|Y|,\max_{f\in F}|f(X)|\leq C$ for some absolute constant C and $\ell(f,(X,Y))=(Y-f(X))^2$). It can be verified that for every $x>0$, with probability greater than $1-4\exp(-x)$, $\widehat{f}_n^{ERM}$ satisfies a non-exact oracle inequality with a residual term proportional to $(x+\log M)/(\epsilon n)$. On the other hand, it is known [67, 81, P15] that in the same setup, there are finite models for which, with probability greater than a positive constant, $\widehat{f}_n^{ERM}$ cannot satisfy an exact oracle inequality with a residual term better than $c_0\sqrt{(\log M)/n}$ (cf. Theorem 2.2.1). Thus, it is possible to establish two optimal oracle inequalities (i.e. oracle inequalities with a non-improvable residual term $r_n(F)$ up to some multiplying constants) for the same procedure — one exact and the other one non-exact — with two very different residual terms: one being the square of the other one.

This huge difference in the residual terms of the exact and the non-exact oracle inequalities was already pointed out in Section 3.3 for the classification problem over VC classes. The goal of this section is to describe the difference between the analysis used in [15] or Theorem 3.3.2 to obtain exact oracle inequalities for the ERM, and the one used in Theorem 3.3.1 to establish

non-exact oracle inequalities for the ERM. Our aim is to indicate why one may get faster rates for non-exact inequalities than for exact ones for the same problem. One should stress that this is not, by any means, a proof that it is impossible to get exact oracle inequalities with fast rates (there are in fact examples in which the ERM satisfies exact oracle inequalities with fast rates: the Linear aggregation problem, [57]). It is not even a proof that the localization method presented here is sharp. However, we believe that this explanation will help to shed some light on the differences between the two types of inequalities.

First note that the price to pay to obtain better rates for non-exact oracle inequalities compared to the one obtained for exact oracle inequalities is that exact oracle inequalities are somehow more "valuable" from a statistical point of view. Indeed, if the regression model with the quadratic loss is considered then it follows from an exact oracle inequality on the prediction risk (Equation (1.2.1) for $\epsilon = 0$), an other exact oracle inequality but for the estimation risk: $\left\| \widehat{f}_n^{ERM} - f^* \right\|_{L_2}^2 \leq \inf_{f \in F} \|f - f^*\|_{L_2}^2 + r_n(F)$ where f^* is the regression function of Y given X and $\|\cdot\|_{L_2}$ is the L_2-norm with respect to the marginal distribution of X. Such a result cannot follow from a non-exact oracle inequality on the prediction risk (Equation (1.2.1) for $\epsilon > 0$). In other words, exact oracle inequalities for the prediction risk $R(\cdot)$ provide both prediction and estimation results (prediction of the output Y and estimation of the regression function f^*) whereas non-exact oracle inequalities for the prediction risk provide only prediction results. The point in studying non-exact oracle inequalities is that if we don't really have to compare the risk $R(\widehat{f}_n)$ of some estimator $\widehat{f}_n$ with $\inf_{f \in F} R(f)$ then it may be advantageous to compare it with $(1 + \epsilon) \inf_{f \in F} R(f)$. Because, in this case, the residual term can be much smaller. This is in particular the situation, when one wants to construct adaptive prediction procedures. Even though the two problems (exact and non-exact prediction) seem rather close they are already different from a statistical point of view.

Now, let's turn to the mathematical differences behind the two problems. Roughly put, and as indicated in Section 3.2.2, localization arguments are based on two main aspects:

1. A Bernstein type condition, the essence of which is that it allows one to "replace" the localized sets $G^\lambda = \{g \in G : Pg^2 \leq \lambda\}$ by the "less complex" localized sets $G_\lambda = \{g \in G : Pg \leq \lambda\}$.

2. The fixed point of the empirical process indexed by the localized star-shaped hull of the loss functions class $V(\ell_F)_\lambda$ (for non-exact inequalities) or of the excess loss functions class $V(\mathcal{L}_F)_\lambda$ (for exact ones).

Although the two aspects seem similar for the exact and non-exact cases, they are very different when dealing with $\mathcal{L}_F$ rather than ℓ_F. Indeed, for non-exact oracle inequality the Margin/Bernstein condition (3.2.3) is almost trivially satisfied and requires no special properties on the learning problem (F, ℓ, P) — as long as the functions in ℓ_F have well behaved tails. As such, it is an individual property of every class members (cf. Lemma 3.2.5). On the other hand, the Bernstein condition (3.2.4) required for the exact oracle inequality obtained in Theorem 3.3.2 and in [15] is deeply connected to the geometry of the problem (see, for example, [82] or Section 1.3). In particular, this explains the gap that we observed in the finite model example ((MS) aggregation) introducing this section. In that case, the class is a finite set of functions and the set $N(F, \ell, X)$ of multiple minimizer (cf. Section 1.3 for more details) is not empty. Thus, one can find a set F and a target Y in a "bad" position (cf. Figure 1.1), leading to an excess loss class $\mathcal{L}_F$ with a trivial Bernstein constant (i.e. of the order of $\sqrt{n}$). On the other hand, in this example, regardless of the choice of Y, the Bernstein constant of ℓ_F is like a constant. Let us

mention that when the gap between exact and non-exact oracle inequalities is only due to the fact that the excess loss class $\mathcal{L}_F$ does not satisfy the Bernstein condition (or that the Bernstein constant is like $\sqrt{n}$), it is likely that in this setup ERM, regularized ERM and penalized estimators satisfy suboptimal exact oracle inequalities [67, 81, P15]. In particular, when slow rates are due to a lack of convexity of F (which is closely related to a bad Bernstein constant of $\mathcal{L}_F$ for 2-convex loss functions), one can consider procedures which "improve the geometry" of the model (for instance, the "empirical star algorithm" of [7] or the "pre-selection-convexification" method in [P14] - we refer to Chapter 2 for more details).

The second mathematical aspect of the problem is the fixed point of the localized empirical process. Although the complexity of the sets $\mathcal{L}_F$ and ℓ_F seem similar from a metric point of view ($\mathcal{L}_F$ is just a shift of ℓ_F) the localized star-shaped hulls $(V(\mathcal{L}_F))_\lambda$ and $(V(\ell_F))_\lambda$ are rather different. Since there are many ways of bounding the supremum of the empirical process indexed by these localized sets, let us show the difference for one of these methods — based on the random projection of the classes, and for the sake of simplicity, we will only consider the regression model with respect to the square loss function. Using this method of analysis at hand, the dominant term of the bound on $\mathbb{E}\|P-P_n\|_{V(\ell_F)_\mu}$ (for the loss functions class) which was obtained in Proposition 3.2.7 is

$$\sqrt{\mu}\sqrt{\frac{\mathbb{E}\gamma_2^2\Big(\widetilde{P_\sigma F^{(\mu)}},\ell_\infty^n\Big)}{n}}. \tag{3.4.1}$$

A similar bound can be obtained for $\mathbb{E}\|P-P_n\|_{V(\mathcal{L}_F)_\mu}$ in [84] and [16], in which the dominant term is

$$\sqrt{\Big(\inf_{f\in F} R(f)+\mu\Big)}\sqrt{\frac{\mathbb{E}\gamma_2^2\Big(\widetilde{P_\sigma F^{(\mu)}},\ell_\infty^n\Big)}{n}}. \tag{3.4.2}$$

If this bound is sharp (and it is in many cases), and since $R^*=\inf_{f\in F}R(f)$ is in general a non-zero constant, the fixed point μ_η^* in Theorem 3.3.2 or [15] is of the order of $\sqrt{\mathbb{E}\gamma_2^2\Big(\widetilde{P_\sigma F^{(\mu^*)}},\ell_\infty^n\Big)/n}$ resulting in a exact oracle inequality with a slow rate larger than $1/\sqrt{n}$. In contrast, the fixed point in the non-exact case of Theorem 3.3.1 is $\lambda_\eta^*\sim\mathbb{E}\gamma_2^2\Big(\widetilde{P_\sigma F^{(\lambda_\eta^*)}},\ell_\infty^n\Big)/n$ which is of the order of $1/n$ (up to logarithmic factors) when the complexity $\mathbb{E}\gamma_2^2\Big(\widetilde{P_\sigma F},\ell_\infty^n\Big)$ is "reasonable".

The reason for this gap comes from the observation that functions in the star hull of ℓ_F whose expectation is smaller than R^* are only "scaled down" versions of functions from ℓ_F. In fact, the "complexity" of the localized sets below the level of R^* can already be seen at the level R^*. Hence, the empirical process those sets index (when scaled properly), becomes smaller with λ.

In contrast, because there are functions $\mathcal{L}_f$ that can have an arbitrarily small expectation, the complexity of the localized subsets of the star hull of $\mathcal{L}_F$ (normalized properly), can even increase as λ decreases. This happens in very simple situations; for example, even in regression relative to the model B_1^M (close to the Convex aggregation problem or in general when the model F is the convex hull of orthonormal functions), if $R^*\neq 0$, the complexity of the localized sets $(\mathcal{L}_{B_1^M})_\mu$ remains almost stable and starts to decrease only at a very "low" level $\mu\leq 1/M$. This is the reason for the phase transition in the error rate ($\sim\max\{\sqrt{(\log M)/n}, M/n\}$) that one encounters in that problem (cf. Theorem 2.8.2). The first term is due to the fact that the complexity of the localized sets $(\mathcal{L}_{B_1^M})_\mu$ does not change as $1/M\leq\mu\leq 1$ decreases — up to

some critical level $\mu = 1/M$, while the second captures what happens when the localized sets begin to "shrink" at levels $0 < \mu \leq 1/M$. Whereas for any $\mu < R^*$, the localized sets $(\ell_{B_1^M})_\mu$ are empty. Thus in many situations the fixed point λ_η^* associated with $\ell_{B_1^M}$ is very small — it can even be equal to zero — but, for the same situation, $\left(\mathcal{L}_{B_1^M}\right)_\mu$ is non empty and can still be a complex set for small values of μ. A concrete example of this phenomenon is treated in Section 4.5 for the Convex aggregation problem.

3.5 Oracle inequalities for regularized ERM

In this section, we study regularized empirical risk minimization procedures for the three problems (non-exact prediction, exact prediction and estimation). The study of regularized ERM procedures is motivated in Section 1.2.3.

The next results are oracle inequalities for Regularized ERM procedures. Before stating these results, one has to say a word on the way the regularizing function $\mathrm{reg}(\cdot)$ and the criterion $\mathrm{crit}(\cdot)$ are related (cf. Section 1.2.3 for more details). We recall that we are given a criterion function $\mathrm{crit} : \mathcal{F} \to \mathbb{R}^+$ that characterizes each element in $\mathcal{F}$ to its level of compliance with a desired property or having some particular computational interest. The choice of $\mathrm{reg}(\cdot)$ in function of $\mathrm{crit}(\cdot)$ such that the regularized ERM $\widehat{f}_n^{RERM} \in \mathrm{argmin}_{f\in\mathcal{F}} \left(R_n(f) + \mathrm{reg}(f)\right)$ satisfies some oracle inequalities is driven by the complexity of the sequence $(F_r)_{r\geq 0}$ of models

$$F_r = \{f \in \mathcal{F} : \mathrm{crit}(f) \leq r\}.$$

We first start with a non-exact oracle inequality for a regularized ERM procedure for the non-exact prediction problem. In this context, for any $r \geq 0$, the complexity of F_r is measured by $\lambda_\eta^*(r)$ defined for some $0 < \eta < 1/2$ such that

$$\mathbb{E}\|P_n - P\|_{V(\ell_{F_r})_{\lambda_\eta^*(r)}} \leq (\eta/4)\lambda_\eta^*(r). \tag{3.5.1}$$

For any $r \geq 0$, $\lambda_\eta^*(r)$ is usually the dominant term in the isomorphic profile of ℓ_{F_r} above which the empirical and the actual structures are equivalent: with high probability, every $\ell \in \ell_{F_r}$ for which $P\ell \geq \lambda_\eta^*(r)$, satisfies that $(1-\eta)P\ell \leq P_n\ell \leq (1+\eta)P\ell$. Thus, $r \to \lambda_\eta^*(r)$ captures the "isomorphic profile" of the collection $(\ell_{F_r})_{r\geq 0}$. Roughly speaking, the regularizing function will be like $\mathrm{reg}(f) = \lambda_\eta^*(\mathrm{crit}(f)), \forall f \in \mathcal{F}$ (up to some multiplying constants and second order terms; cf. (3.5.4) for the exact definition of reg).

For technical reasons, we also introduce an auxiliary function α_n, defined as follows: if there exists $C_n > 0$ such that $\forall f \in \mathcal{F}, \mathrm{crit}(f) \leq C_n$ then take $\alpha_n \equiv C_n$, otherwise if $r \to \lambda_\eta^*(r)$ tends to infinity when r tends to infinity and if there exists $K_1 > 0$ such that for every $(r, x) \in \mathbb{R}_+ \times \mathbb{R}_+^*$, $2\rho_n^\ell(r, x) \leq \rho_n^\ell(K_1(r+1), x)$ (where ρ_n^ℓ is defined in Theorem 3.5.1 below) then, let f_0 be any function in $\cup_{r\geq 0} F_r$ (for instance, when $0 \in \cup_{r\geq 0} F_r$, take $f_0 \equiv 0$) and define for every $x > 0$ and $0 < \eta < 1/2$,

$$\begin{aligned} \alpha_n(\eta, x) \geq \max \Big[& K_1(\mathrm{crit}(f_0) + 2), \\ & (\lambda_\eta^*)^{-1}\big(2(1+\eta)(3R(f_0) + 2K'(b_n(\ell_{f_0}) + B_n(\mathrm{crit}(f_0)))((x+1)/n))\big)\Big], \end{aligned} \tag{3.5.2}$$

where $(\lambda_\eta^*)^{-1}$ is the generalized inverse function of λ_η^* (i.e. $(\lambda_\eta^*)^{-1}(y) = \sup\big(r > 0 : \lambda_\eta^*(r) \leq y\big), \forall y > 0$), b_n and B_n are functions introduced in Theorem 3.5.1 below and K_1, K' are absolute constants. Fortunately, α_n usually has little impact on the resulting rates. For instance, in the main applications of Chapter 4, we have $\log \alpha_n(\eta, x) \lesssim_\eta \log(x+n)$.

Theorem 3.5.1 ([P16]) *There exist absolute positive constants c_0 and c_1 for which the following holds. Assume that for every $f \in \mathcal{F}, \ell_f(Z) \geq 0$ a.s. and that there are non-decreasing functions ϕ_n and B_n such that for every $r \geq 0$ and every $f \in F_r$,*

$$b_n(\ell_{F_r}) \leq \phi_n(r) \text{ and } P\ell_f^2 \leq B_n(r)P\ell_f + B_n^2(r)/n.$$

Let $0 < \eta < 1/2$ and assume that there exists some function ρ_n^ℓ increasing in its first argument such that for every $(r,x) \in \mathbb{R}_+ \times \mathbb{R}_+^*$,

$$\rho_n^\ell(r,x) \geq \max\Big(\lambda_\eta^*(r), c_0\frac{(\phi_n(r)+B_n(r)/\eta)(x+1)}{n\eta}\Big).$$

Denote $F = \cup_{r\geq 0}F_r$ and take $x > 0$. We set

$$\widehat{f}_n^{RERM} \in \operatorname*{argmin}_{f\in F}\Big(R_n(f) + \frac{2}{1+\eta}\rho_n^\ell(\mathrm{crit}(f)+1, x+\log\alpha_n(\eta,x))\Big). \tag{3.5.3}$$

Then, with probability greater than $1-12\exp(-x)$,

$$\begin{aligned} &R(\widehat{f}_n^{RERM}) + \rho_n^\ell(\mathrm{crit}(\widehat{f}_n^{RERM})+1, x+\log\alpha_n(\eta,x))\Big) \\ &\leq \inf_{f\in F}\Big[(1+2\eta)R(f) + 2\rho_n^\ell(\mathrm{crit}(f)+1, x+\log\alpha_n(\eta,x)) \\ &\quad + c_1\frac{(b_n(\ell_f)+B_n(\mathrm{crit}(f))/\eta)(x+1)}{n\eta}\Big]. \end{aligned}$$

Like in Theorem 3.3.1, the condition $P\ell^2 \leq B_n(r)P\ell + B_n^2(r)/n, \forall \ell \in \ell_{F_r}$ holds when ℓ is non-negative and ψ_1 for some function B_n such that $B_n(r) \lesssim \mathrm{diam}(\ell_{F_r}, \psi_1)\log(n)$, and thus, unlike the two other situations (exact prediction and estimation problems), the "geometry" and statistical properties of the family $(F_r)_{r\geq 0}$ do not play a crucial role to obtain non-exact regularized oracle inequalities.

The choice of the regularizing function in terms of the criterion is now made explicit: for every $f \in \mathcal{F}$,

$$\mathrm{reg}(f) = \frac{2}{1+\eta}\rho_n^\ell(\mathrm{crit}(f)+1, x+\log\alpha_n(\eta,x)). \tag{3.5.4}$$

This choice of the regularizing function is the one suggested by our method to solve the non-exact prediction problem.

The regularizing function may be different for the exact prediction problem. We introduce $r \to \mu_{1/2}^*(r)$ such that for any $r \geq 0$,

$$\mathbb{E}\|P_n - P\|_{V(\mathcal{L}_{F_r})_{\mu_{1/2}^*(r)}} \leq (1/8)\mu_{1/2}^*(r). \tag{3.5.5}$$

Note that for exact oracle inequalities, we take $\eta = 1/2$ since this parameter does not play any role for the exact prediction problem. Like in Theorem 3.5.1, we also consider an auxiliary function β_n defined by: if there exists $C_n > 0$ such that $\forall f \in \mathcal{F}, \mathrm{crit}(f) \leq C_n$ then take $\beta_n \equiv C_n$, otherwise if $r \to \mu_{1/2}^*(r)$ tends to infinity when r tends to infinity and if there exists $K_1 > 0$ such that for every $(r,x) \in \mathbb{R}_+ \times \mathbb{R}_+^*$, $2\rho_n^{\mathcal{L}}(r,x) \leq \rho_n^{\mathcal{L}}(K_1(r+1),x)$ (where $\rho_n^{\mathcal{L}}$ is defined in Theorem 3.5.3 below) then, let f_0 be any function in $\cup_{r\geq 0}F_r$ (for instance, when $0 \in \cup_{r\geq 0}F_r$, take $f_0 \equiv 0$) and define for every $x > 0$,

$$\begin{aligned} \beta_n(x) \geq \max\Big[&K_1(\mathrm{crit}(f_0)+2), (\mu_{1/2}^*)^{-1}\Big(3\Big(3R(f_0) \\ &+ 256\Big(\frac{xB_n(\mathrm{crit}(f_0))K^2}{n}\Big)^{\frac{1}{2-\beta}} + \frac{8KB_n(\mathrm{crit}(f_0))\sqrt{x}}{n} + \frac{8Kb_n(\mathcal{L}_{f_0})x}{n}\Big)\Big)\Big], \end{aligned} \tag{3.5.6}$$

where $(\mu^*_{1/2})^{-1}$ is the generalized inverse function of $\mu^*_{1/2}$ (i.e. $(\mu^*_{1/2})^{-1}(y) = \sup\big(r > 0 : \mu^*_{1/2}(r) \leq y\big), \forall y > 0$), b_n and B_n are functions introduced in Theorem 3.5.3 below and K_1, K' are absolute constants. Fortunately, β_n usually has little impact on the resulting rates. For instance, in the main applications of Chapter 4, we have $\log \beta_n(x) \lesssim \log(x+n)$. We also consider the following regularity assumption on the family of models $(F_r)_{r\geq 0}$ which was introduced in [16] and [84].

Definition 3.5.2 (Definition 2.4 in [84], [16]) *We say that a family $(F_r)_{r\geq 0}$ of subsets of $\mathcal{F}$ is an **ordered, parametrized hierarchy of $\mathcal{F}$ with respect to the loss function ℓ and the probability distribution** P when the following conditions are satisfied:*

1. *$(F_r)_{r\geq 0}$ is non-decreasing (that is $s \leq t \Rightarrow F_s \subseteq F_t$);*
2. *for any $r \geq 0$, there exists a unique element $f^*_r \in F_r$ such that $R(f^*_r) = \inf(R(f) : f \in F_r)$; we consider the excess loss function associated with the class F_r: for any $f \in F_r$, $\mathcal{L}_{r,f}(\cdot) = Q(\cdot, f) - Q(\cdot, f^*_r)$;*
3. *the map $r \longmapsto R(f^*_r)$ is continuous;*
4. *for every $r_0 \geq 0$, $\cap_{r\geq r_0} F_r = F_{r_0}$.*

The following result is a slight modification of a result from [84] and [16]. A sketch of the proof of this result can be found in Chapter 6.

Theorem 3.5.3 *[Theorem 2.5, [84], [16]] There exist absolute positive constants c_0 and c_1, c_2, c_3 for which the following holds. Assume that for every $f \in \mathcal{F}, \ell_f(Z) \geq 0$ a.s. and that there exists $0 < \beta \leq 1$ and two non-decreasing functions ϕ_n and B_n such that for every $r \geq 0$ and every $f \in F_r$,*

$$b_n(\mathcal{L}_{F_r}) \leq \phi_n(r) \text{ and } P\mathcal{L}^2_{r,f} \leq B_n(r)\big(P\mathcal{L}_{r,f}\big)^\beta + B^2_n(r)/n.$$

Assume that $(F_r)_{r\geq 0}$ is an ordered, parametrized hierarchy of $\mathcal{F}$ and denote $F = \cup_{r\geq 0} F_r$. Consider a continuous function $\rho^{\mathcal{L}}_n$ that is increasing in both argument such that for every $(r, x) \in \mathbb{R}_+ \times \mathbb{R}^*_+$,

$$\rho^{\mathcal{L}}_n(r, x) \geq \max\Big(\mu^*_{1/2}(r), 256\Big(\frac{xB_n(r)K^2}{n}\Big)^{\frac{1}{2-\beta}} + \frac{8KB_n(r)\sqrt{x}}{n} + \frac{8K\phi_n(r)x}{n}\Big).$$

Let $x > 0$ and for $\theta(x) = x + c_0 \log\big(1 + R(f^*_0)/\rho^{\mathcal{L}}_n(0, x + c_1) + \log \beta_n(x)\big)$ define

$$\widehat{f}^{RERM}_n \in \operatorname*{argmin}_{f\in F}\Big(R_n(f) + c_2\rho^{\mathcal{L}}_n(2(\mathrm{crit}(f) + 1), \theta(x))\Big). \tag{3.5.7}$$

Then, with probability greater than $1 - 5\exp(-x)$,

$$R(\widehat{f}^{RERM}_n) \leq \inf_{f\in F}\big[R(f) + c_3\rho^{\mathcal{L}}_n(2(\mathrm{crit}(f) + 1), \theta(x))\big].$$

Therefore, in the context of the exact-prediction problem, the choice of the regularizing function in terms of the criterion is given, for every $f \in \mathcal{F}$, by

$$\mathrm{reg}(f) = c_0\rho^{\mathcal{L}}_n(2(\mathrm{crit}(f) + 1), \theta(x)). \tag{3.5.8}$$

An application of this result is given in Section 4.3. Note that when the criterion function takes values in a countable set then we don't need the family $(F_r)_{r\geq 0}$ to be an ordered, parametrized

hierarchy. This can be useful when models complexity is measured by their dimension, VC dimension or in general by any discrete complexity measure.

To see the difference between the properties of regularized ERM for the exact prediction problem and the non-exact prediction problem, it is interesting to compare the exact oracle inequalities for regularized ERM procedures obtained in [16, P6, 84] and Theorem 3.5.3 with Theorem 3.5.1. Indeed, Theorem 3.5.1 implies that a possible way of regularizing to obtain non-exact regularized oracle inequalities is roughly by the regularizing function $f \in \mathcal{F} \to \lambda_\eta^*(\mathrm{crit}(f))$. On the other hand, for exact regularized oracle inequalities, the resulting regularizing function in [16, P6, 84] and Theorem 3.5.3 is roughly $f \in \mathcal{F} \to \max\Big(\mu_{1/2}^*(\mathrm{crit}(f)), \big(B_n(\mathrm{crit}(f))x/n\big)^{1/(2-\beta)}\Big)$. Therefore, there are mainly two differences between the two different ways of regularizing. The first difference comes from the residual terms of the deviation inequalities of the supremum of the empirical process indexed by the two families $\big(\ell_{F_r}\big)_{r\geq 0}$ and $\big(\mathcal{L}_{F_r}\big)_{r\geq 0}$. In the "loss function case", this residual term is of the order of x/n. This is due to the fact that the Margin/Bernstein condition (3.2.3) is almost always satisfied in this case. On the other side, the residual term in the case of the excess loss class is of the order of $\big(x/n\big)^{1/(2-\beta)}$ depending on the Bernstein parameter $0 \leq \beta \leq 1$. The second difference comes from the complexity aspect of the two problems through the fixed point functions $r \to \lambda_\eta^*(r)$ and $r \to \mu_{1/2}^*(r)$. Although the two isomorphic profiles seem similar, $\lambda_\eta^*(\cdot)$ can be the square of $\mu_{1/2}^*(\cdot)$ in some examples (cf. the discussion in Section 3.4 and the example in Section 4.5). In those cases, one has to "regularize more" to obtain an exact oracle inequality than for a non-exact one (cf. Section 3.4 and 4.5 below for more details).

A similar result can be obtained for the estimation problem. Once again, we introduce a fixed point function $r \to \nu_\eta^*(r)$ and an auxiliary function γ_n to obtain an oracle inequality for a regularized ERM procedure for the estimation problem. We introduce $r \to \nu_\eta^*(r)$ such that for any $r \geq 0$,

$$\mathbb{E}\|P_n - P\|_{V(\mathcal{E}_{F_r})_{\nu_\eta^*(r)}} \leq (\eta/4)\nu_\eta^*(r). \tag{3.5.9}$$

Like in Theorem 3.5.1 and 3.5.3, we also consider an auxiliary function γ_n defined by: if there exists $C_n > 0$ such that $\forall f \in \mathcal{F}, \mathrm{crit}(f) \leq C_n$ then take $\gamma_n \equiv C_n$, otherwise if $r \to \nu_\eta^*(r)$ tends to infinity when r tends to infinity and if there exists $K_1 > 0$ such that for every $(r,x) \in \mathbb{R}_+ \times \mathbb{R}_+^*$, $2\rho_n^{\mathcal{E}}(r,x) \leq \rho_n^{\mathcal{E}}(K_1(r+1),x)$ (where $\rho_n^{\mathcal{E}}$ is defined in Theorem 3.5.4 below) then, let f_0 be any function in $\cup_{r\geq 0}F_r$ (for instance, when $0 \in \cup_{r\geq 0}F_r$, take $f_0 \equiv 0$) and define for every $x > 0$ and $0 < \eta < 1/2$,

$$\gamma_n(\eta,x) \geq \max\Big[K_1(\mathrm{crit}(f_0)+2), (\nu_\eta^*)^{-1}\Big(2(1+\eta)\Big(3R(f_0) + 4\Big(\frac{B_n(\mathrm{crit}(f_0))K^2x}{n}\Big)^{\frac{1}{2-\beta}} + \frac{KB_n(\mathrm{crit}(f_0))\sqrt{x}}{n} + \frac{2Kb_n(\mathcal{E}_{f_0})x}{n}\Big)\Big)\Big],$$

where $(\nu_\eta^*)^{-1}$ is the generalized inverse function of ν_η^* (i.e. $(\nu_\eta^*)^{-1}(y) = \sup\big(r > 0 : \nu_\eta^*(r) \leq y\big), \forall y > 0$), b_n and B_n are functions introduced in Theorem 3.5.4 below and K_1, K' are absolute constants. Fortunately, γ_n usually has little impact on the resulting rates. A proof of the following result can be found in Chapter 5.

Theorem 3.5.4 *There exist an absolute positive constant c_0 for which the following holds. Assume that for every $f \in \mathcal{F}, \ell_f(Z) \geq 0$ a.s. and that there exists $0 < \beta \leq 1$ and two non-decreasing functions ϕ_n and B_n such that for every $r \geq 0$ and every $f \in F_r$,*

$$b_n(\mathcal{E}_{F_r}) \leq \phi_n(r) \text{ and } P\mathcal{E}_f^2 \leq B_n(r)\big(P\mathcal{E}_f\big)^\beta + B_n^2(r)/n.$$

Let $0 < \eta < 1/2$ and assume that there exists some function $\rho_n^{\mathcal{E}}$ increasing in its first argument such that for every $(r,x) \in \mathbb{R}_+ \times \mathbb{R}_+^*$,

$$\rho_n^{\mathcal{E}}(r,x) \geq \max\Big(\nu_\eta^*(r), \frac{16}{\eta}\Big(\frac{xB_n(r)K^2}{n}\Big(\frac{4}{\eta}\Big)^\beta\Big)^{\frac{1}{2-\beta}} + \frac{4KB_n(r)\sqrt{x}}{\eta n} + \frac{4K\phi_n(r)x}{\eta n}\Big).$$

Denote $F = \cup_{r\geq 0} F_r$ and let $x > 0$. We set

$$\widehat{f}_n^{RERM} \in \operatorname*{argmin}_{f\in F}\Big(R_n(f) + \frac{2}{1+\eta}\rho_n^{\mathcal{E}}(\mathrm{crit}(f)+1, x + \log\gamma_n(\eta,x))\Big). \tag{3.5.10}$$

Then, with probability greater than $1 - 12\exp(-x)$,

$$\begin{aligned} &R(\widehat{f}_n^{RERM}) - R(f^*) + \rho_n^{\mathcal{E}}(\mathrm{crit}(\widehat{f}_n^{RERM})+1, x+\log\gamma_n(\eta,x)) \\ &\leq \inf_{f\in F}\Big[(1+2\eta)\big(R(f)-R(f^*)\big) + 2\rho_n^{\mathcal{E}}(\mathrm{crit}(f)+1, x+\log\gamma_n(\eta,x)) \\ &\quad + c_0\Big(\frac{B_n(\mathrm{crit}(f))x}{\eta n}\Big)^{\frac{1}{2-\beta}} + \frac{c_0 B_n(\mathrm{crit}(f))\sqrt{x}}{n} + \frac{c_0 b_n(\mathcal{E}_f)x}{\eta n}\Big]. \end{aligned}$$

Therefore, in the context of the estimation problem, the choice of the regularizing function in terms of the criterion is given, for every $f \in \mathcal{F}$, by

$$\mathrm{reg}(f) = \frac{2}{1+\eta}\rho_n^{\mathcal{E}}(\mathrm{crit}(f)+1, x+\log\gamma_n(\eta,x)). \tag{3.5.11}$$

As a conclusion, given a criterion function, the way the empirical risk is regularized depends heavily of the problem we want to analyze. If one wants to solve a non-exact prediction problem then one can regularize by (3.5.4), for the exact prediction problem a way of regularizing is given in (3.5.8) and for the estimation problem a regularizing function in terms of the criterion is given in (3.5.11). The three different ways of regularizing depend on the geometry and the complexity of the families $\big(\ell_{F_r}\big)_{r\geq 0}$, $\big(\mathcal{L}_{F_r}\big)_{r\geq 0}$ and $\big(\mathcal{E}_{F_r}\big)_{r\geq 0}$ and therefore can be very different.

3.6 Oracle inequalities for penalized estimators

We consider the Model Selection setup of [76] recalled in Subsection 1.2.4. We are given a collection $\mathcal{M}$ of models. We assume that $\mathcal{M}$ is countable. The aim of this section is to show that one way of penalizing can be derived from the isomorphic profiles of the family of loss functions classes $(\ell_m)_{m\in\mathcal{M}}$ and excess loss functions classes $\big(\mathcal{L}_m\big)_{m\in\mathcal{M}}$ and $\big(\mathcal{E}_m\big)_{m\in\mathcal{M}}$ defined for any $m \in \mathcal{M}$ by

$$\ell_m = \{\ell_f : f \in m\}, \quad \mathcal{L}_m = \{\ell_f - \ell_{f_m^*} : m \in \mathcal{M}\} \text{ and } \mathcal{E}_m = \{\ell_f - \ell_{f^*} : m \in \mathcal{M}\}$$

where we assume that there exists $f_m^* \in \operatorname{argmin}_{f\in m} R(f)$ for any $m \in \mathcal{M}$ (and $f^* \in \operatorname{argmin}_{f\in\mathcal{F}} R(f)$).

We first consider the non-exact prediction problem. Let $0 < \eta < 1/2$. Assume that there exists some function $\rho_n^\ell : \mathcal{M}\times\mathbb{R}^+ \to \mathbb{R}^+$ such that for any $m \in \mathcal{M}$ and $x > 0$, with probability greater than $1 - c_0\exp(-x)$,

$$\big|P_n\ell_f - P\ell_f\big| \leq \eta\max\big(P\ell_f, \rho_n^\ell(m,x)\big), \quad \forall f \in m. \tag{3.6.1}$$

Let $(x_m)_{m\in\mathcal{M}}$ be a family of positive numbers such that $\sum_{m\in\mathcal{M}}\exp(-x_m) \leq c_1$. Let $x > 0$. We consider the penalty function

$$\mathrm{pen}^\ell(m) = \rho_n^\ell(m, x+x_m), \forall m \in \mathcal{M} \tag{3.6.2}$$

and the penalized estimator $\widehat{f}_{\widehat{m}}$ where for any $m \in \mathcal{M}$, $\widehat{f}_m \in \operatorname{argmin}_{f\in m} R_n(f)$ and $\widehat{m} \in \operatorname{argmin}_{m\in\mathcal{M}} \big(R_n(\widehat{f}_m) + \operatorname{pen}^\ell(m)\big)$. Once again, we assume that the ERM procedures $\widehat{f}_m, m \in \mathcal{M}$ and $\widehat{m}$ exist otherwise approximated minimizers can be considered. It follows from a union bound over $\mathcal{M}$ and (3.6.1) that with probability greater than $1 - c_0 c_1 \exp(-x)$,

$$\begin{aligned} P\ell_{\widehat{f}_{\widehat{m}}} &\le (1-\eta)^{-1}\big(P_n\ell_{\widehat{f}_{\widehat{m}}} + \rho_n^\ell(\widehat{m}, x + x_{\widehat{m}})\big) = (1-\eta)^{-1}\big(P_n\ell_{\widehat{f}_{\widehat{m}}} + \operatorname{pen}^\ell(\widehat{m})\big) \\ &= (1-\eta)^{-1} \inf_{m\in\mathcal{M}} \big(P_n\ell_{\widehat{f}_m} + \operatorname{pen}^\ell(m)\big) = (1-\eta)^{-1} \inf_{m\in\mathcal{M}} \big(\inf_{f\in m} P_n\ell_f + \operatorname{pen}^\ell(m)\big) \\ &\le \frac{1+\eta}{1-\eta} \inf_{m\in\mathcal{M}} \big(\inf_{f\in m} P\ell_f + \operatorname{pen}^\ell(m)\big). \end{aligned} \tag{3.6.3}$$

Therefore, one way of obtaining a penalized oracle inequality for the non-exact prediction problem is by penalizing the empirical risk by the penalty function (3.6.2) associated with the isomorphic profile of the family $\big(\ell_m\big)_{m\in\mathcal{M}}$ and the weights $(x_m)_{m\in\mathcal{M}}$.

Any way of controlling the isomorphic profile of a loss functions class will provide a way of penalizing the empirical risk. For instance, we can apply Theorem 3.2.2 to construct a penalty function. We introduce the following function $m \in \mathcal{M} \to \lambda_\eta^*(m)$ defined for some $0 < \eta < 1/2$ by

$$\mathbb{E}\|P_n - P\|_{V(\ell_m)_{\lambda_\eta^*(m)}} \le (\eta/4)\lambda_\eta^*(m). \tag{3.6.4}$$

For any $m \in \mathcal{M}$, $\lambda_\eta^*(m)$ is usually the dominant term in the isomorphic profile of ℓ_m. Thus, $m \to \lambda_\eta^*(m)$ captures the "isomorphic profile" of the family $(\ell_m)_{m\in\mathcal{M}}$. Roughly speaking, the penalty function will be like $\operatorname{pen}(m) = \lambda_\eta^*(m), \forall m \in \mathcal{M}$ (up to some multiplying constants and second order terms; the exact definition of pen is provided in the following result and (3.6.5)). The proof of the following result follows from (3.6.3) and Theorem 3.2.2.

Theorem 3.6.1 *There exists an absolute positive constant c_0 such that the following holds. Assume that there are some functions ϕ_n and B_n such that for every $m \in \mathcal{M}$ and every $f \in m$,*

$$b_n(\ell_m) \le \phi_n(m) \text{ and } P\ell_f^2 \le B_n(m)P\ell_f + B_n^2(m)/n.$$

Let $0 < \eta < 1/2$ and assume that for every $(m, x) \in \mathcal{M} \times \mathbb{R}_+^*$,

$$\rho_n^\ell(m, x) \ge \max\Big(\lambda_\eta^*(m), c_0 \frac{(\phi_n(m) + B_n(m)/\eta)(x+1)}{n\eta}\Big).$$

Let $(x_m)_{m\in\mathcal{M}}$ be a family of positive numbers such that $\sum_{m\in\mathcal{M}} \exp(-x_m) \le c_1$. Let $x > 0$ and consider the penalty function $\operatorname{pen}^\ell(m) = \rho_n^\ell(m, x + x_m)$ and the penalized estimator $\widehat{f}_{\widehat{m}}$ associated with this penalty function. Then, with probability greater than $1 - 12c_1\exp(-x)$,

$$R(\widehat{f}_{\widehat{m}}) \le \frac{1+\eta}{1-\eta} \inf_{m\in\mathcal{M}} \big(\inf_{f\in m} P\ell_f + \operatorname{pen}^\ell(m)\big).$$

Like in Theorem 3.3.1 and Theorem 3.5.1, the condition $P\ell^2 \le B_n(m)P\ell + B_n^2(m)/n, \forall m \in \mathcal{M}, \ell \in \ell_m$ holds when ℓ is non-negative and ψ_1 for some function B_n such that $B_n(m) \lesssim \operatorname{diam}(\ell_m, \psi_1)\log(n)$. The penalty function for the non-exact prediction problem in terms of the fixed point function λ_η^* that comes out of our analysis in Theorem 3.6.1 is

$$\operatorname{pen}^\ell(m) = \max\Big(\lambda_\eta^*(m), c_0 \frac{(\phi_n(m) + B_n(m)/\eta)(x + x_m + 1)}{n\eta}\Big). \tag{3.6.5}$$

As in Section 3.5 for the regularizing function, the choice of the penalty function for the estimation problem may be different.

For the estimation problem, the choice of the penalty function can be derived from the isomorphic profile of the family of excess loss functions classes $\big(\mathcal{E}_m\big)_{m\in\mathcal{M}}$. Let $0<\eta<1/2$. Assume that there exists some function $\rho_n^{\mathcal{E}}:\mathcal{M}\times\mathbb{R}^+\to\mathbb{R}^+$ such that for any $m\in\mathcal{M}$ and $x>0$, with probability greater than $1-c_0\exp(-x)$,

$$\big|P_n\mathcal{E}_f-P\mathcal{E}_f\big|\le\eta\max\big(P\mathcal{E}_f,\rho_n^{\mathcal{E}}(m,x)\big),\quad\forall f\in m. \tag{3.6.6}$$

Let $(x_m)_{m\in\mathcal{M}}$ be a family of positive numbers such that $\sum_{m\in\mathcal{M}}\exp(-x_m)\le c_1$. Let $x>0$. We consider the penalty function

$$\mathrm{pen}^{\mathcal{E}}(m)=\rho_n^{\mathcal{E}}(m,x+x_m),\forall m\in\mathcal{M} \tag{3.6.7}$$

and the penalized estimator $\widehat{f}_{\widehat{m}}$ associated with this penalty function. It follows from a union bound over $\mathcal{M}$ and (3.6.6) that with probability greater than $1-c_0c_1\exp(-x)$,

$$\begin{aligned}P\mathcal{E}_{\widehat{f}_{\widehat{m}}}&\le(1-\eta)^{-1}\big(P_n\mathcal{E}_{\widehat{f}_{\widehat{m}}}+\rho_n^{\mathcal{E}}(\widehat{m},x+x_{\widehat{m}})\big)=(1-\eta)^{-1}\big(P_n\mathcal{E}_{\widehat{f}_{\widehat{m}}}+\mathrm{pen}^{\mathcal{E}}(\widehat{m})\big)\\&=(1-\eta)^{-1}\inf_{m\in\mathcal{M}}\big(P_n\mathcal{E}_{\widehat{f}_m}+\mathrm{pen}^{\mathcal{E}}(m)\big)=(1-\eta)^{-1}\inf_{m\in\mathcal{M}}\big(\inf_{f\in m}P_n\mathcal{E}_f+\mathrm{pen}^{\mathcal{E}}(m)\big)\\&\le\frac{1+\eta}{1-\eta}\inf_{m\in\mathcal{M}}\big(\inf_{f\in m}P\mathcal{E}_f+\mathrm{pen}^{\mathcal{E}}(m)\big).\end{aligned} \tag{3.6.8}$$

Therefore, one way of obtaining a penalized oracle inequality for the estimation problem is by penalizing the empirical risk by the penalty function (3.6.7) associated with the isomorphic profile of the family $\big(\mathcal{E}_m\big)_{m\in\mathcal{M}}$ and the weights $(x_m)_{m\in\mathcal{M}}$.

Following the same strategy as in Theorem 3.6.1, we obtain a penalized oracle inequality for the penalized estimator for the estimation problem. Once again, for some $0<\eta<1/2$, we introduce a fixed point function $m\in\mathcal{M}\to\nu_\eta^*(m)$: for any $m\in\mathcal{M}$,

$$\mathbb{E}\|P_n-P\|_{V(\mathcal{E}_m)_{\nu_\eta^*(m)}}\le(\eta/4)\nu_\eta^*(m). \tag{3.6.9}$$

The fixed point function $m\to\nu_\eta^*(m)$ characterizes the complexity of the family $\big(\mathcal{E}_m\big)_{m\in\mathcal{M}}$. The proof of the following result follows from (3.6.8) and Theorem 3.2.4.

Theorem 3.6.2 *There exist absolute positive constants c_0 and c_1 for which the following holds. Assume that there exists $0<\beta\le1$ and some functions ϕ_n and B_n such that for every $m\in\mathcal{M}$ and every $f\in m$,*

$$b_n(\mathcal{E}_m)\le\phi_n(m)\ \text{and}\ P\mathcal{E}_f^2\le B_n(m)\big(P\mathcal{E}_f\big)^\beta+B_n^2(m)/n.$$

Let $0<\eta<1/2$ and assume that for every $(m,x)\in\mathcal{M}\times\mathbb{R}_+^*$,

$$\rho_n^{\mathcal{E}}(m,x)\ge\max\Big(\nu_\eta^*(m),\frac{16}{\eta}\Big(\frac{xB_n(m)K^2}{n}\Big(\frac{4}{\eta}\Big)^\beta\Big)^{\frac{1}{2-\beta}}+\frac{4KB_n(m)\sqrt{x}}{\eta n}+\frac{4K\phi_n(m)x}{\eta n}\Big).$$

Let $(x_m)_{m\in\mathcal{M}}$ be a family of positive numbers such that $\sum_{m\in\mathcal{M}}\exp(-x_m)\le c_1$. Let $x>0$ and consider the penalty function $\mathrm{pen}^{\mathcal{E}}(m)=\rho_n^{\mathcal{E}}(m,x+x_m)$ and the penalized estimator $\widehat{f}_{\widehat{m}}$ associated with this penalty function. Then, with probability greater than $1-12c_1\exp(-x)$,

$$R(\widehat{f}_{\widehat{m}})-R(f^*)\le\frac{1+\eta}{1-\eta}\inf_{m\in\mathcal{M}}\big(\inf_{f\in m}\big(R(f)-R(f^*)\big)+\mathrm{pen}^{\mathcal{E}}(m)\big).$$

Therefore, in the context of the estimation problem, a possible way of penalizing the empirical risk is by the function

$$\mathrm{pen}^{\mathcal{E}}(m) = \max\Big(\nu_\eta^*(m), c_2(B_n(m)+\phi_n(m))\Big(\frac{x+x_m}{n\eta}\Big)^{\frac{1}{2-\beta}}\Big). \tag{3.6.10}$$

A similar result for the exact prediction problem can be obtained but its proof requires a more subtle argument that can be found in [13]. We introduce a fixed point function $m \in \mathcal{M} \to \mu_{1/2}^*(m)$: for any $m \in \mathcal{M}$,

$$\mathbb{E}\|P_n - P\|_{V(\mathcal{L}_m)_{\mu_{1/2}^*(m)}} \leq (1/8)\mu_{1/2}^*(m) \tag{3.6.11}$$

where $\mathcal{L}_m = \{\ell_f - \ell_{f_m^*} : f \in m\}$ and $f_m^* \in \mathrm{argmin}_{f\in m} R(f)$. Once again for the exact prediction problem we take $\eta = 1/2$. The fixed point function $m \to \mu_{1/2}^*(m)$ characterizes the complexity of the family $\big(\mathcal{L}_m\big)_{m\in\mathcal{M}}$.

Theorem 3.6.3 ([13]) *There exist absolute positive constants c_0 and c_1 for which the following holds. Assume that the models in $\mathcal{M} = \big(m_n\big)_{n\in\mathbb{N}}$ are nested i.e. $m_0 \subset m_1 \subset m_2 \subset \cdots$. Assume that there exists $0 < \beta \leq 1$ and two non-decreasing functions ϕ_n and B_n such that for every $m \in \mathcal{M}$ and every $f \in m$,*

$$b_n(\mathcal{L}_m) \leq \phi_n(m) \text{ and } P\mathcal{L}_{m,f}^2 \leq B_n(m)\big(P\mathcal{L}_{m,f}\big)^\beta + B_n^2(m)/n \text{ where } \mathcal{L}_{m,f} = \ell_f - \ell_{f_m^*}.$$

Let $\rho_n^{\mathcal{L}}$ be an increasing function such that for every $(m,x) \in \mathcal{M}\times\mathbb{R}_+^*$,

$$\rho_n^{\mathcal{L}}(m,x) \geq \max\Big(\mu_{1/2}^*(m), 256\Big(\frac{xB_n(m)K^2}{n}\Big)^{\frac{1}{2-\beta}} + \frac{8KB_n(m)\sqrt{x}}{n} + \frac{8K\phi_n(m)x}{n}\Big).$$

Let $(x_m)_{m\in\mathcal{M}}$ be a family of positive numbers such that $\sum_{m\in\mathcal{M}}\exp(-x_m) \leq c_1$. Let $x > 0$ and consider the penalty function $\mathrm{pen}^{\mathcal{L}}(m) = (7/2)\rho_n^{\mathcal{L}}(m, x+x_m)$ and the penalized estimator $\widehat{f}_{\widehat{m}}$ associated with this penalty function. Then, with probability greater than $1 - 12c_1\exp(-x)$,

$$R(\widehat{f}_{\widehat{m}}) \leq \inf_{m\in\mathcal{M}}\big(\inf_{f\in m} R(f)\big) + (18/7)\,\mathrm{pen}^{\mathcal{L}}(m)\big).$$

Therefore, for the exact prediction problem, a possible way of penalizing the empirical risk is by the function

$$\mathrm{pen}^{\mathcal{L}}(m) = c_2 \max\Big(\mu_{1/2}^*(m), (B_n(m)+\phi_n(m))\Big(\frac{x+x_m}{n}\Big)^{\frac{1}{2-\beta}}\Big) \tag{3.6.12}$$

for some absolute constant c_2.

Similar results hold when the models $m \in \mathcal{M}$ satisfy different Margin/Bernstein condition with different Bernstein parameter $(\beta_m)_{m\in\mathcal{M}}$ (one just have to replace β by β_m in Theorem 3.6.3). Similar results can be found in [76, 77] and a comparison with these results is given in the following section.

As a conclusion, the way the empirical risk is penalized depends heavily of the problem we have in mind. If one wants to solve a non-exact prediction problem then one may penalize by (3.6.2), for the estimation problem one way of penalizing is given in (3.6.10) and for the exact prediction problem, the empirical risk may be penalized by (3.6.12). The three different ways of penalizing depend on the geometry and the complexity of the families $\big(\ell_m\big)_{m\in\mathcal{M}}$, $\big(\mathcal{E}_m\big)_{m\in\mathcal{M}}$ and $\big(\mathcal{L}_m\big)_{m\in\mathcal{M}}$ and therefore can be very different.

3.7 Connections between regularized ERM and penalized estimators

In this section, we show that penalization and regularization of the empirical risk are related. For instance, in the non-exact prediction problem, one can derive Model Selection theorems (like Theorem 3.6.1) from the result of Theorem 3.5.1 originally crafted for regularized ERM. The first step is to show that any regularized ERM procedure is a Model Selection procedure for some particular class $\mathcal{M}$ of models and penalty function. Then, we can use Theorem 3.5.1 and derived non-exact oracle inequalities for the associated penalized estimators. For the sake of completeness, we start to prove that the reverse is also true: any penalized estimator is a regularized ERM procedure for some particular class $\mathcal{F}$ and regularizing function.

We recall the setup of Model Selection as introduced in [76] and recalled in Subsection 1.2.4. We are given a collection $\mathcal{M}$ of models and a penalty function $\text{pen} : \mathcal{M} \to \mathbb{R}^+$. For every model $m \in \mathcal{M}$, an ERM procedure is constructed:

$$\widehat{f}_m \in \operatorname*{argmin}_{f \in m} R_n(f). \tag{3.7.1}$$

Then a model $\widehat{m}$ is empirically selected by

$$\widehat{m} \in \operatorname*{argmin}_{m \in \mathcal{M}} \Big(R_n(\widehat{f}_m) + \text{pen}(m) \Big). \tag{3.7.2}$$

The penalized estimator studied in Model Selection is $\widehat{f}_{\widehat{m}}$. Once again, we assume that the infimum in (3.7.1) and (3.7.2) are achieved. The next result shows that the penalized estimator $\widehat{f}_{\widehat{m}}$ is a regularized ERM.

Lemma 3.7.1 ([P16]) *Define a class $\mathcal{F}$ and a regularizing function by*

$$\mathcal{F} = \bigcup_{m \in \mathcal{M}} m \text{ and } \text{reg}(f) = \inf_{m \in \mathcal{M} : f \in m} \text{pen}(m), \forall f \in \mathcal{F}. \tag{3.7.3}$$

Then the penalized estimator $\widehat{f}_{\widehat{m}}$ satisfies

$$\widehat{f}_{\widehat{m}} \in \operatorname*{argmin}_{f \in \mathcal{F}} \Big(R_n(f) + \text{reg}(f) \Big).$$

It follows from Lemma 3.7.1 that any Model Selection procedure can be studied as a regularized ERM procedure over the function class $\mathcal{F}$ and for the regularizing function defined in (3.7.3). It appears that the reverse is also true.

Lemma 3.7.2 ([P16]) *Let $\mathcal{F}$ be a class of function and $\text{reg} : \mathcal{F} \to \mathbb{R}^+$ be a regularizing function such that for any $f \in \mathcal{F}, \text{reg}(f) < \infty$. Assume that there exists $\widehat{f}_n^{RERM} \in \mathcal{F}$ minimizing $f \longrightarrow R_n(f) + \text{reg}(f)$ over $\mathcal{F}$. Denote by $\text{reg}(\mathcal{F}) \subset \mathbb{R}^+$ the range of reg. For any $r \in \text{reg}(\mathcal{F})$ define the model $m_r = \{f \in \mathcal{F} : \text{reg}(f) \leq r\}$. Define a class $\mathcal{M}$ of models and a penalty function by*

$$\mathcal{M} = \{m_r : r \in \text{reg}(\mathcal{F})\} \text{ and } \text{pen} : m_r \in \mathcal{M} \longrightarrow r \in \mathbb{R}^+. \tag{3.7.4}$$

Then $\widehat{f}_n^{RERM}$ is a penalized estimator for the class of models $\mathcal{M}$ endowed with the penalty function pen.

In particular, we can apply Theorem 3.5.1 to obtain results on regularized ERM procedures, then construct a class $\mathcal{M}$ and a penalty function according to (3.7.4) and finally use Lemma 3.7.2 to obtain non-exact penalized oracle inequalities for the penalized estimator $\widehat{f}_{\widehat{m}}$ constructed in this framework for the non-exact prediction problem.

As an example of application, we consider the Model Selection setup studied in Chapter 8 of [76] on the classification problem over Vapnik-Chervonenkis models. We consider the $0-1$ loss function $\ell_f(x,y) = \mathbb{1}_{f(x)\neq y}$ defined for any $(x,y) \in \mathcal{X} \times \{0,1\}$ and measurable function $f : \mathcal{X} \to \{0,1\}$. We are given a countable set $\mathcal{M}$ of countable models (that is a countable set of measurable functions from $\mathcal{X}$ to $\{0,1\}$) such that any $m \in \mathcal{M}$ has a finite VC dimension denoted by V_m. In this setup, Theorem 3.5.1 can be applied without any extra assumption. In particular, we do not assume any Margin/Bernstein condition. In this case, the penalty function used in p. 285 of [76] is, for any $m \in \mathcal{M}$,

$$\mathrm{pen}(m) = 2\sqrt{\frac{2V_m(1+\log(n/V_m))}{n}} + \sqrt{\frac{\log n}{2n}} \tag{3.7.5}$$

and the penalized estimator $\widehat{f}_{\widehat{m}}$ satisfies the following risk bound (p.285 of [76]):

$$\mathbb{E}R(\widehat{f}_{\widehat{m}}) \leq \inf_{m\in\mathcal{M}} \Big(\inf_{f\in m} R(f) + \mathrm{pen}(m)\Big) + \sqrt{\frac{\pi}{2n}}. \tag{3.7.6}$$

Now we turn to the application of Theorem 3.5.1 for this problem. We first need to consider some "well-adapted" criterion. For that, we define

$$\mathcal{F} = \bigcup_{m\in\mathcal{M}} m \text{ and } \mathrm{crit}(f) = \min_{m\in\mathcal{M}} \big(V_m \wedge n : f \in m\big), \forall f \in \mathcal{F}.$$

The next step is now to calibrate the regularizing function in terms of the criterion. Theorem 3.5.1 provides one way of doing so by computing the isomorphic profile of the family of loss functions classes $(\ell_{F_r})_{r\in\mathbb{N}}$ where for any $r \in \mathbb{N}$ (note that crit takes its values in $\mathbb{N}$) $F_r = \{f \in \mathcal{F} : \mathrm{crit}(f) \leq r\}$. The sets F_r can be very complex if $\mathcal{M}$ is made of many disjoint models of small VC dimension. Our method cannot handle such situations. That is why we assume that the models are embedded in an increasing way:

$$\mathcal{M} = \{m_k : k \in \mathbb{N}\} \text{ such that } m_0 \subset m_1 \subset m_2 \subset \cdots. \tag{3.7.7}$$

This assumption is clearly a weak point compared to the result (3.7.6) which does not require such a structure on $\mathcal{M}$. Nevertheless, assumption (3.7.7) is a classical assumption in Model Selection and allows to get in our situation $F_r = m_{k(r)}$ where $k(r) = \max(k \in \mathbb{N} : V_{m_k} \wedge n \leq r)$. The isomorphic function associated with the family of models $(F_r)_{r\in\mathbb{N}}$ can be obtained in this context following the same strategy used to get (3.3.3): we obtain, for any $r \in \mathbb{N}$,

$$\lambda^*_\epsilon(r) = \frac{c_0(V_{m_{k(r)}} \wedge n)\log\big(en/(V_{m_{k(r)}} \wedge n)\big)}{\epsilon^2 n}. \tag{3.7.8}$$

Moreover, we can check that $b_n(\ell_{F_r}) = 1$, $B_n(r) = 1$ for any $r \geq 0$ and that $\alpha_n \equiv n$ is a valid choice for the auxiliary function α_n since $\mathrm{crit}(f) \leq n, \forall f \in \mathcal{F}$ (cf. (3.5.2)). We can now apply Theorem 3.5.1: let $0 < x \leq \log n$ and $0 < \epsilon < 1/2$ and consider the regularized ERM $\widehat{f}_n^{RERM}$ over $\mathcal{F} = \cup_{m\in\mathcal{M}} m$ associated with the regularizing function

$$\mathrm{reg}(f) = \frac{c_1 V_{m(f)}\log\big(en/V_{m(f)}\big)}{\epsilon^2 n} \tag{3.7.9}$$

where $m(f) = \max\big(m \in \mathcal{M} : V_m \leq \mathrm{crit}(f) + 1\big)$. It follows from Theorem 3.5.1 that with probability greater than $1 - 12\exp(-x)$,

$$R(\widehat{f}_n^{RERM}) + c_0\mathrm{reg}(\widehat{f}_n^{RERM}) \leq (1+2\epsilon)\inf_{f\in\mathcal{F}}\Big(R(f) + c_1\mathrm{reg}(f)\Big) + \frac{c_2(x+1)}{n}. \tag{3.7.10}$$

From this result, we can now derive a non-exact penalized oracle inequality for the penalized estimator associated with the class of models $\mathcal{M}'$ and the penalty function pen' as defined in (3.7.4):

$$\mathcal{M}' = \{m_r : r \in \mathrm{reg}(\mathcal{F})\} \text{ and } \mathrm{pen}'(m_r) = r, \quad \forall r \in \mathrm{reg}(\mathcal{F}) \tag{3.7.11}$$

where $\mathrm{reg}(\mathcal{F}) = \{r_0, \ldots, r_N\}$, $N = \max(k \in \mathbb{N} : V_{m_k} \leq n)$ and

$$r_i = \frac{c_1 V_{m_i'}\log\left(en/V_{m_i'}\right)}{\epsilon^2 n}, \quad \forall 0 \leq i \leq N,$$

for $m_N' = m_N, m_i' = \max\big(m \in \mathcal{M} : V_m < V_{m_{i+1}'}\big), \forall 0 \leq i \leq N-1$. In other words, $\mathcal{M}'$ is the largest subset of $\mathcal{M}$ of models with strictly increasing VC dimension smaller than n and with the largest possible model for each one of these VC dimensions. Each one of these models of VC dimension V is then penalized by $c_1 V\log(en/V)/(\epsilon^2 n)$. We can now state a result for the penalized estimator associated with the class $\mathcal{M}'$ and the penalty function pen'.

Theorem 3.7.3 ([P16]) *There exists some absolute constants c_1, c_2, c_3 and c_4 such that the following holds. Let $\mathcal{M} = \{m_0, \cdots, m_N\}$ be a family of models such that $m_0 \subset \cdots \subset m_N$ and $V_{m_0} < V_{m_1} < \cdots < V_{m_N} \leq n$ where for any $m \in \mathcal{M}$, V_m is the VC dimension of m. Let $0 < \epsilon < 1$. Consider the penalty function $\mathrm{pen} : \mathcal{F} \to \mathbb{R}^+$ defined by $\mathrm{pen}(m) = c_1 V_m \log\big(en/V_m\big)/(\epsilon^2 n)$. Then the penalized estimator $\widehat{f}_{\widehat{m}}$ constructed in this setup is such that for any $0 < x \leq \log n$, with probability greater than $1 - 12\exp(-x)$,*

$$R(\widehat{f}_{\widehat{m}}) + c_2\mathrm{pen}(\widehat{m}) \leq (1+2\epsilon)\min_{m\in\mathcal{M}}\Big(\inf_{f\in m} R(f) + c_3\mathrm{pen}(m)\Big).$$

In particular, it is interesting to note that, up to a logarithmic factor, the penalty function in (3.7.5) is of the order of $\sqrt{V/n}$ whereas, in the same framework (up to the structural assumption (3.7.7) which can be removed if we use a direct approach as in Theorem 3.6.1), the penalty function defined in Theorem 3.7.3 is of the order of V/n up to a logarithmic term. This difference can be explained from a "geometric viewpoint" by the fact that the Bernstein condition on the loss functions class: $\mathbb{E}\ell_f^2 \leq B\mathbb{E}\ell_f, \forall f \in \mathcal{F}$ is trivially satisfied in the setup of Theorem 3.7.3; whereas the Bernstein condition for the excess loss functions class: $\mathbb{E}\mathcal{L}_f^2 \leq B\mathbb{E}\mathcal{L}_f, \forall f \in \mathcal{F}$ or the Margin assumption: $\mathbb{E}(\ell_f - \ell_{f^*})^2 \leq B\mathbb{E}(\ell_f - \ell_{f^*}), \forall f \in \mathcal{F}$ are not true in general and are somehow "required" to obtain oracle inequalities with fast rates for the penalized estimator for the estimation and exact prediction problems. This difference can also be explained from a "statistical viewpoint" since estimation results on the Bayes rules follow from (3.7.6): one just have to subtract the risk of the Bayes rules on both sides of the inequality. But estimation results cannot follow from the non-exact oracle inequality of Theorem 3.7.3. Once again the non-exact oracle inequalities obtained in Theorem 3.3.1, 3.5.1, 3.6.1 and 3.7.3 deal only with the (non-exact) prediction problem and do not provide any estimation result. To summarize, if one wants some estimation results (either on the Bayes rules or the regression function), then one should assume some kind of Margin/Bernstein condition to get fast rates. But if only prediction results are of interest then no geometrical assumption are needed to get fast rates in this context.

Finally, note that a direct approach based on the computation of isomorphic functions for the family of loss functions class $(\ell_m)_{m\in\mathcal{M}}$ like in Theorem 3.6.1 would provide a way of constructing penalty functions and obtaining oracle inequalities for the penalized estimator associated with this penalty function. This approach will not require any structural assumption on $\mathcal{M}$ like (3.7.7). Nevertheless, we did not apply Theorem 3.6.1 in that context. Somehow, we found more interesting to prove that Model Selection methods can be seen as regularized procedures and that Theorem 3.5.1, which was originally designed for regularized estimators, can also be used to prove results for penalized estimators.

3.8 The Restricted Isometry property and the isomorphic profile of the loss functions class

In this section, we assume that the data are n i.i.d. couples $(X_i, Y_i)_{1\leq i\leq n}$ where the output variables $Y_1, \dots, Y_n$ are real numbers and the input variables $X_1, \dots, X_n$ take their values in $\mathbb{R}^d$. The aim of this section is to compare our approach based on the isomorphic profile of some functions classes to the approach in Compressed Sensing based on the restricted isometry property (RIP). To allow this comparison, we consider the Compressed Sensing setup where it is assumed that the output is a free-noise linear combination of the input: there exists some $\beta_0 \in \mathbb{R}^d$ such that $Y_i = \langle X_i, \beta_0\rangle$ for every $1 \leq i \leq n$.

We start with the definition of the RIP and its role in Compressed Sensing. Let $1 \leq s \leq d$ be an integer and $0 < \eta < 1$. The restricted isometry property RIP(s, η) is the following condition on the design $X_1, \dots, X_n$:

$$\forall x \in \Sigma_s, (1-\eta)\|x\|_2^2 \leq \frac{1}{n}\sum_{i=1}^{n}\langle X_i, x\rangle^2 \leq (1+\eta)\|x\|_2^2, \tag{3.8.1}$$

where $\Sigma_s = \{x \in \mathbb{R}^d : |\mathrm{Supp}(x)| \leq s\}$ is the set of all s-sparse vectors of $\mathbb{R}^d$. This condition was first introduced in [35] and since then extensively studied in the Compressed Sensing literature and related fields. The main interest in this property of the sampling process is that under this assumption and as long as $\eta < \sqrt{2}-1$ (cf. [30]), the basis pursuit algorithm

$$\Delta_1(Y_1, \dots, Y_n) \in \mathrm{argmin}\left(\|x\|_1 : \langle X_i, x\rangle = Y_i, i = 1, \dots, n\right) \tag{3.8.2}$$

is such that $\Delta_1(Y_1, \dots, Y_n) = \beta_0$ when $\beta_0 \in \Sigma_{\lceil s/2\rceil}$. This means that any $\lceil s/2\rceil$-sparse vector β_0 can be reconstructed exactly from the n linear measurements $\langle X_1, \beta_0\rangle, \dots, \langle X_n, \beta_0\rangle$ as long as the sampling process (or the design) satisfies RIP(s, η) with $\eta < \sqrt{2}-1$.

Of course, depending on the sparsity parameter, a minimal number of observations is needed for the exact reconstruction problem. It follows from some entropy argument (cf. Chapter 2 in [P2]) that if RIP(s, η) is satisfied with $\eta < \sqrt{2}-1$ then necessarily

$$s\log\left(c_0 d/s\right) \leq c_1 n. \tag{3.8.3}$$

It appears that this bound is sharp up to the constants c_0 and c_1 in the sense that if $X_1, \dots, X_n$ are n i.i.d. standard Gaussian vectors of $\mathbb{R}^d$ and if $s\log\left(ed/s\right) \leq c_2 n$ then with probability greater than $1 - c_3\exp(-c_4 n)$, RIP(s, η) holds with $\eta < \sqrt{2}-1$. This result follows from an ϵ-net argument that can be found for instance in Chapter 2 of [P2].

Now, let us try to understand this problem from the Learning theory point of view. In this setup, a model is a set of linear functions $F = \{f_\beta : \beta \in T\}$ where $T \subset \mathbb{R}^d$ and $f_\beta = \langle \cdot, \beta\rangle, \forall \beta \in$

$\mathbb{R}^d$. We consider the square loss function and thus the regression function is $f^* = f_{\beta_0}$. To simplify notation, we identify the model F to the set T of vectors and write $\ell_\beta(x,y) = (y - \langle x, \beta\rangle)^2$ for all $\beta \in \mathbb{R}^d$. We also denote by $\ell_T = \{\ell_\beta : \beta \in T\}$, the set of all the loss functions indexed by T.

Given $0 < \eta < 1$ and a model $T \subset \mathbb{R}^d$, the isomorphic profile of ℓ_T is defined by

$$r^*(\ell_T)_\eta = \inf\big(r > 0 : \forall \beta \in T, \big|P_n\ell_\beta - P\ell_\beta\big| \le \eta \max(P\ell_\beta, r)\big).$$

In particular, for an isotropic design (i.e. $\mathbb{E}\langle X, \beta\rangle^2 = \|\beta\|_2^2, \forall \beta \in \mathbb{R}^d$), we have

$$r^*(\ell_T)_\eta = \inf\Big(r > 0 : \forall \beta \in T, \Big|\frac{1}{n}\sum_{i=1}^n \langle X_i, \beta_0 - \beta\rangle^2 - \|\beta_0 - \beta\|_2^2\Big| \le \eta \max(\|\beta - \beta_0\|_2^2, r)\Big).$$

If $r^*(\ell_T)_\eta = 0$ then the matrix $A : \mathbb{R}^d \to \mathbb{R}^n$ with row vectors $n^{-1/2}X_i^\top, 1 \le i \le n$ behaves like an η-isometry over the set $\beta_0 - T = \{\beta_0 - \beta : \beta \in T\}$:

$$\forall \beta \in T, (1-\eta)\|\beta_0 - \beta\|_2^2 \le \|A(\beta_0 - \beta)\|_2^2 \le (1+\eta)\|\beta_0 - \beta\|_2^2.$$

On the opposite, if A acts like a η-isometry on $2T = \{t_1 + t_2 : t_1, t_2 \in T\}$ and $\beta_0 \in T$ then $r^*(\ell_T)_\eta = 0$. Therefore, there is an obvious equivalence at saying that the design operator A acts like a η-isometry over the set $\beta_0 - T$ and saying that the isomorphic profile of ℓ_T is null. In the particular case of sparse vectors, for the model $T = \Sigma_s$ and β_0 a $\lceil s/2\rceil$-sparse vector, if $r^*(\ell_{\Sigma_s})_\eta = 0$ then A satisfies RIP$(\lceil s/2\rceil, \eta)$. Reciprocally, if A satisfies RIP(s,η) and β_0 is $\lceil s/2\rceil$-sparse then $r^*(\ell_{\Sigma_{\lceil s/2\rceil}})_\eta = 0$.

Assume now that the design is isotropic. We have seen that if the design operator A satisfies RIP$(2s,\eta)$ then the isomorphic profile $r^*(\ell_{\Sigma_s})_\eta$ is equal to zero. But we proved in Section 3.1 that the isomorphic profile of ℓ_{Σ_s} drives the residual term of the ERM over Σ_s. As a consequence, if A satisfies RIP$(2s,\eta)$ and β_0 is s-sparse then $R(\widehat{f}_n^{ERM}) = 0$ and so $\widehat{\beta}_n^{ERM} = \beta_0$ where $\widehat{f}_n^{ERM} = f_{\widehat{\beta}_n^{ERM}}$ and

$$\widehat{\beta}_n^{ERM} \in \operatorname*{argmin}_{\beta \in \Sigma_s} R_n(f_\beta) = \operatorname*{argmin}_{\beta \in \Sigma_s} \frac{1}{n}\sum_{i=1}^n \big(Y_i - \langle X_i, \beta\rangle\big)^2. \tag{3.8.4}$$

This proves that the ERM can reconstruct exactly the vector β_0 as long as the operator A satisfies RIP$(2s,\eta)$ and β_0 is s-sparse. But the ERM procedure $\widehat{\beta}_n^{ERM}$ of (3.8.4) has two drawbacks: the sparsity parameter s has to be known in advance (which is never the case in practice); and the minimization problem (3.8.4) is combinatorial in nature (since $\beta_0 \in \Sigma_s$ we have to find a s-sparse vector $\widehat{\beta}_n^{ERM}$ such that $A\widehat{\beta}_n^{ERM} = A\beta_0$ - this requires to explore in general an exponential number of support of size s among d coordinates). That is why the ERM procedure is never used in practice. Nevertheless, not all is lost as far as ERM procedures are considered.

The important point in the analysis is that when A acts in a norm preserving way on a set $2T$ and $\beta_0 \in T$ then $r^*(\ell_T)_\eta = 0$ and as a consequence the ERM over T can reconstruct exactly β_0. Therefore, to avoid the two main drawbacks of the procedure (3.8.4) mentioned just before, one has to perform the ERM over a set T that does not depend on the sparsity parameter s and T has to be convex so that convex programming methods may help to construct this ERM estimator. First, we look for a convex set T containing $\Sigma_s \cap \|\beta_0\|_2 B_2^d$ (since there is no need to search β_0 outside the sets $\|\beta_0\|_2 B_2^d$ and Σ_s - the point that we are not supposed to know $\|\beta_0\|_2$ and s in advance will be treated later). Moreover this set T does not have to be too large so that A can still act on T in a norm preserving way (with high probability). A natural candidate is the intersection body $\|\beta_0\|_2 \big(B_2^d \cap \sqrt{s}B_1^d\big)$. Second, as mentioned before, we don't

know in advance $\|\beta_0\|_2$ but we have in mind that A is norm preserving thus there is some hope that with high probability we have $\|\beta_0\|_2^2 \leq 4\|A\beta_0\|_2^2 = 4\|y\|_2^2$ where $y = (Y_1, \dots, Y_n)^\top$ is the vectors of the outputs. Finally, we have to insure that the operator A acts like a η-isometry on $2\|y\|_2 \left(B_2^d \cap \sqrt{s}B_1^d\right)$ for some $0 < \eta < 1$. For that we use a result of [85] (see also Chapter 3 in [P2]) saying that if the Gaussian complexity $\ell_*(T) = \mathbb{E}\sup_{t\in T}\langle G, t\rangle$ - where G is a standard Gaussian vector of $\mathbb{R}^d$ - of some set $T \subset \mathbb{R}^d$ is such that $C\sqrt{n} \leq \ell_*(T) \leq C_2(\eta)\sqrt{n}$ (for some well chosen constant $C_2(\eta)$) and if the row vectors of A are sub-Gaussian, isotropic and independent then with probability larger than $1 - c_0\exp(-c_1 Cn)$, A is a η-isometry on T. In the case we are interested in, we have (cf. Chapter 3 in [P2])

$$c_2\|y\|_2\sqrt{s\log\left(\frac{c_3 d}{s}\right)} \leq \ell_*\big(2\|y\|_2\left(B_2^d \cap \sqrt{s}B_1^d\right)\big) \leq c_4\|y\|_2\sqrt{s\log\left(\frac{c_5 d}{s}\right)}.$$

Therefore, we consider the largest sparsity parameter insuring that A is a η-isometry over $2\|y\|_2\left(B_2^d \cap \sqrt{s}B_1^d\right)$:

$$\widehat{s} = \max\left(s \in \mathbb{N} : c_4\|y\|_2\sqrt{s\log\left(\frac{c_5 d}{s}\right)} \leq C_2(\eta)\sqrt{n}\right).$$

Then the model we consider is $T = 2\|y\|_2\left(B_2^d \cap \sqrt{\widehat{s}}B_1^d\right)$ and the ERM estimator is

$$\tilde{\beta}_n^{ERM} \in \operatorname*{argmin}_{\beta\in 2\|y\|_2\left(B_2^d\cap\sqrt{\widehat{s}}B_1^d\right)} \frac{1}{n}\sum_{i=1}^n \big(Y_i - \langle X_i, \beta\rangle\big)^2. \tag{3.8.5}$$

Note that T is convex and its construction does not require any a priori knowledge on β_0. The ERM procedure $\widehat{\beta}_n^{ERM}$ satisfies the following exact reconstruction property.

Theorem 3.8.1 *There exists two absolute constants c_0 and c_1 such that the following holds. Assume that $X_1, \dots, X_n$ are subgaussian, isotropic and independent. Then with probability greater than $1 - c_0\exp(-c_1 n)$, any vector β_0 such that the size of its support is smaller than $\widehat{s}$ is such that*

$$\operatorname*{argmin}_{\beta\in 2\|y\|_2\left(B_2^d\cap\sqrt{\widehat{s}}B_1^d\right)} \frac{1}{n}\sum_{i=1}^n \big(\langle X_i, \beta_0\rangle - \langle X_i, \beta\rangle\big)^2 = \{\beta_0\}.$$

Roughly speaking, the ERM algorithm introduced in (3.8.5) can reconstruct any vector β_0 as long as $\|\beta_0\|_2\sqrt{\|\beta_0\|_0\log\left(ed/\|\beta_0\|\right)} \lesssim \sqrt{n}$ where $\|\beta_0\|_0 = |\mathrm{Supp}(\beta_0)|$ is the size of the support of β_0. On the other hand, the Basis Pursuit algorithm (3.8.2) can reconstruct any β_0 as long as $\sqrt{\|\beta_0\|_0\log\left(ed/\|\beta_0\|\right)} \lesssim \sqrt{n}$ independently of $\|\beta_0\|_2$. In both cases, an isometry property of the sampling process was used to prove the result: the RIP and a zero-valued isomorphic profile of the loss functions class. Of course, in the noisy setup, one cannot hope to design any exact reconstruction algorithm but still the RIP have been used to prove oracle inequalities for the Basis Pursuit algorithm. Somehow, the RIP should be consider as a special property on the design making the Basis Pursuit algorithm an efficient exact reconstruction algorithm. In the same spirit, isomorphic profile of functions classes should be seen as a special tool designed for the study of ERM procedures. It appears that these two tools coincide in the non-noisy case but they are not by any means necessary properties of the sampling process or the design in both setups (exact reconstruction of sparse vectors and Learning theory respectively). In fact, these two properties are sometimes even too strong. One example for the Compressed Sensing problem can be found in [P2] and an example in Learning theory is given in the following section.

3.9 A counter-example in Convex aggregation

We have seen in Section 3.1 that the isomorphic profile of $\mathcal{L}_F$ bounds the residual term of exact oracle inequalities of the ERM procedure over F. It appears that this bound is in many cases of the correct order. That is some lower bounds matching the upper bound obtained from the isomorphic profile can be stated. But there are some examples for which this method does not provide the correct residual term.

A counter-example was constructed in [15] showing that for every integer n, there exists a class G star-shaped in zero and a probability measure P for which for any sample $Z_1, \dots, Z_n$ there is $g \in G$ such that $Pg = 1/4$ and $P_n g = 0$ and therefore the isomorphic profile of G is at least $1/4$ leading to a trivial bound for the ERM $\widehat{g}^{ERM} \in \operatorname{argmin}_{g \in G} P_n g$. But still, by a direct approach, one has $\mathbb{E}\widehat{g}^{ERM} \leq 1/n$ with constant probability. Therefore, in some situation, the isomorphic profile does not provide the correct residual term.

The counter-example of [15] is somehow a bit artificial. The aim of this section is to construct a "not too much artificial" example for which the approach based on the isomorphic profile is sub-optimal. For that we consider the bounded regression framework with respect to the square loss. We consider $\phi_1, \dots, \phi_{M+1}$ real-valued functions defined on $\mathcal{X}$ and a random variable X such that $\phi_1, \dots, \phi_{M+1}$ are orthogonal in $L^2(P_X)$, $\phi_1(X), \dots, \phi_M(X)$ are uniformly distributed over $[-\sqrt{3}, \sqrt{3}]$ and $Y = \phi_{M+1}(X)$ is a Rademacher variable independent of $\phi_1(X), \dots, \phi_{M+1}(X)$. We consider the problem of convex aggregation over the class $F = \{0, \pm\phi_1, \dots, \pm\phi_M\}$ and for the orthogonal target $Y = \phi_{M+1}(X)$. A natural candidate is the ERM over the convex hull

$$\tilde{f}^{ERM-C} \in \operatorname*{argmin}_{f \in \operatorname{conv}(F)} R_n(f).$$

It follows from a direct approach the following exact oracle inequality for $\tilde{f}^{ERM-C}$.

Theorem 3.9.1 ([P18]) *There exist absolute positive constants c_0, c_1 for which the following holds. Consider the dictionary F and the couple (X, Y) introduced previously. When $M \geq \sqrt{n}$, with probability larger than $7/12$,*

$$R(\tilde{f}^{ERM-C}) \leq \min_{f \in \operatorname{conv}(F)} R(f) + c_3 \frac{c_3}{\sqrt{n \log(eM/\sqrt{n})}}.$$

The counter example used in Theorem 3.9.1 is the same as the one used in Theorem 2.3.1. Since both residual term are of the same order up to multiplying absolute constants and, in this case $\min_{f \in \operatorname{conv}(F)} R(f) = \min_{f \in F} R(f)$, Theorem 3.9.1 proves that the residual term obtained in Theorem 2.3.1 is optimal in the case $M \geq \sqrt{n}$. Now, it remains to see that the residual term

$$\frac{1}{\sqrt{n \log\left(eM/\sqrt{n}\right)}} \tag{3.9.1}$$

is not the one obtained by using the "isomorphic profile approach" and also to understand why there is indeed a gap in this example.

The counter example we use for this result is a class $F_M = \{0, \pm\phi_1, \dots, \pm\phi_M\}$ such that $(\phi)_{i=1}^M$ is a bounded orthonormal family of $L_2(P^X)$ and $Y = \phi_{M+1}(X)$ is orthogonal to this family in 0. We also assume that $\Phi(X) = (\phi_1(X), \dots, \phi_M(X))$ is isotropic (i.e. $\mathbb{E}\langle \Phi(X), \lambda \rangle^2 = \|\lambda\|_2^2, \forall \lambda \in \mathbb{R}^M$) so that the complexity is measured with respect to the ℓ_2^M-norm. An element

in $\mathrm{conv}(F_M)$ is of the form $f_\lambda = \langle \Phi, \lambda \rangle$ for some $\lambda \in B_1^M$. Its excess risk is $\mathcal{L}_{f_\lambda} = \langle \Phi, \lambda \rangle^2 - 2\langle \Phi, \lambda \rangle \phi_{M+1}$ and the empirical process indexed by B_1^M that we study is for any $\lambda \in B_1^M$

$$P_n \mathcal{L}_{f_\lambda} = \frac{1}{n}\sum_{i=1}^n \langle \Phi(X_i), \lambda \rangle^2 - \frac{2}{n}\sum_{i=1}^n \langle \Phi(X_i), \lambda \rangle \phi_{M+1}(X_i). \tag{3.9.2}$$

It follows from [85] that the oscillations of the quadratic term $\lambda \in B_1^M \to |(P_n - P)(\langle \Phi, \lambda \rangle^2)|$ are second order terms and that the empirical process (3.9.2) behaves like $\lambda \in B_1^M \to \|\lambda\|_2^2 - 2n^{-1/2}\langle V, \lambda \rangle$ where $V = n^{-1/2}\sum_{i=1}^n \phi_{M+1}(X_i)\Phi(X_i)$. Then, we use a Gaussian approximation result from [90] saying that V essentially behaves like G a standard normal vector of $\mathbb{R}^M$. It follows that the risk $P\mathcal{L}_{\widehat{f}} = \|\widehat{\lambda}\|_2^2$ of the empirical risk minimization procedure $\widehat{f} = f_{\widehat{\lambda}}$ will be essentially located around

$$\operatorname*{argmin}_{0 \le r \le 1} \min_{\lambda \in B_1^M \cap \sqrt{r}\mathcal{S}^{M-1}} \left(r - 2n^{-1/2}\langle G, \lambda \rangle \right) = \operatorname*{argmin}_{0 \le r \le 1} \left(r - 2n^{-1/2} \|G\|_{A_r^\circ} \right),$$

where $A_r = B_1^M \cap \sqrt{r}\mathcal{S}^{M-1}$ and $\|G\|_{A_r^\circ} = \sup_{\lambda \in A_r} \langle G, \lambda \rangle$. For every radius $1 \le r \le 1$, we end up with computing the interpolation norm $\|G\|_{A_r^\circ}$. It appears that, for the range $1/M \le r \le 1$ we are interested in, a slight modification of the radius r results in a very small logarithmic change in the value of $\|G\|_{A_r^\circ}$. That is the reason why we have to compute a second order term approximation of $\|G\|_{A_r^\circ}$ for every r. This finally yields that $\|\widehat{\lambda}\|_2^2$ is essentially located around $\operatorname{argmin}_{0 \le r \le 1} \left(r - 2n^{-1/2}\ell_*(B_1^M \cap \sqrt{r}\mathcal{S}^{M-1}) \right)$ where $\ell_*(T) = \mathbb{E}\sup_{t \in T}\langle G, t \rangle$ is the Gaussian complexity for some $T \subset \mathbb{R}^M$ and since $\ell_*(B_1^M \cap \sqrt{r}\mathcal{S}^{M-1}) \sim \sqrt{\log(eMr)}$, $\|\widehat{\lambda}\|_2^2$ is indeed of the order of the rate (3.9.1). This is essentially the strategy we used to prove Theorem 2.3.1 and Theorem 3.9.1 in the case $M \ge \sqrt{n}$.

An approach based on the isomorphic profile of $\mathcal{L}_{\mathrm{conv}(F)}$ will provide a residual term of the order of

$$\mu_{1/2}^* = \inf\left(\mu > 0 : \frac{\ell_*(B_1^M \cap \sqrt{\mu}\mathcal{S}^{M-1})}{\sqrt{n}} \le c_0 \mu \right) \sim \sqrt{\frac{\log\left(eM/\sqrt{n}\right)}{n}}. \tag{3.9.3}$$

Therefore, there is a logarithmic gap between the direct approach used to prove Theorem 3.9.1 and the approach based on the isomorphic profile of $\mathcal{L}_{\mathrm{conv}(F)}$. Therefore, the Learning setup of a model F which is B_1^M in $L_2(P_X)$ with an output Y orthogonal to B_1^M in zero and constant far away from B_1^M is a setup for which the isomorphic profile approach is suboptimal.

The surprising fact is that the rate (3.9.1) — which cannot be improved in this situation — tends to zero as the dimension M grows. Since the model is B_1^M and the target Y is orthogonal to B_1^M in 0, somehow this means that the complexity with respect to ℓ_2^M of B_1^M increases around 0 as M increases. At a first analysis, we did not expect such a behaviour for B_1^M in high dimensions and that is the reason why we chose Y to be orthogonal to B_1^M in 0 and not in some other point of B_1^M. We believe that this unexpected high dimensional behaviour of B_1^M can find some geometrical explanation in the following paragraph.

We first try to find some structure inside B_1^M which is the source of complexity of B_1^M with respect to ℓ_2^M. Denote by $(e_1, \ldots, e_M)$ the canonical basis of $\mathbb{R}^M$ and for any $I \subset \{1, \ldots, M\}$ define $x_I = |I|^{-1}\sum_{i \in I} e_i$. It is known that for any $k \in \{1, \ldots, M\}$ there exists $\Lambda_k \subset \{I \subset \{1, \ldots, M\} : |I| = k\}$ such that $\log|\Lambda_k| \ge c_0 k \log\left(eM/k\right)$ and the symmetrical difference for any $I \ne J \in \Lambda_k$ is such that $|I \Delta J| \ge k/8$ (cf. for instance [76] or [P2]). It is easy to check that

$$\cup_{1 \le k \le M}\{x_I : I \in \Lambda_k\} \subset \cup_{1 \le k \le M} \cup_{I \in \Lambda_k} k^{-1/2} B_2^I \subset B_1^M$$

where B_2^I is the set of vectors in B_2^M supported in I.

According to [95], the logarithm of the minimal number of translated of $\sqrt{r}B_2^M$ needed to cover B_1^M is of the order of $r^{-1}\log\big(eMr\big)$ when $r \geq 1/M$. When $1/r$ is an integer, two different points x_I and x_J in $\big\{x_I : I \in \Lambda_{1/r}\big\}$ are such that $\|x_I - x_J\|_2 = r\sqrt{|I\Delta J|} \geq \sqrt{r}/8$. Therefore, $\big\{x_I : I \in \Lambda_{1/r}\big\}$ is a set of points $\sqrt{r}/8$-separated of log-cardinality at least $c_0 r^{-1}\log\big(eMr\big)$. Therefore, the entropy numbers $\log N(B_1^M, \sqrt{r}B_2^N)$ and $\log N\big(\big\{x_I : I \in \Lambda_{1/r}\big\}, \sqrt{r}/8B_2^M\big)$ are proportional and thus, the set $\big\{x_I : I \in \Lambda_{1/r}\big\}$ is one of the source of complexity of B_1^M with respect to ℓ_2^M at the resolution level $\sqrt{r}$. It is also interesting to note that, from Sudakov inequality,

$$\ell_*(\big\{x_I : I \in \Lambda_{1/r}\big\}) \geq c_1 \min_{I\neq J\in\Lambda_{1/r}} \|x_I - x_J\|_2 \sqrt{\log|\Lambda_{1/r}|} \geq c_2\sqrt{\log\big(eMr\big)}$$

which is of the same order as the Gaussian complexity of $B_1^M \cap \sqrt{r}\mathcal{S}^{M-1}$ and $\big\{x_I : I \in \Lambda_{1/r}\big\} \subset \sqrt{r}\mathcal{S}^{M-1}$, meaning that all the complexity of B_1^M (w.r.t. ℓ_2^M) at the scale $\sqrt{r}$ lies in $B_1^M \cap \sqrt{r}\mathcal{S}^{M-1}$. In other words, one source of complexity of B_1^M (w.r.t. ℓ_2^M) is an exponential number of points with almost disjoint supports at any level $\sqrt{r}$ when $1/r$ is an integer and, in general, of an exponential number of euclidean balls of dimension $\lceil 1/r\rceil$ with radius $\sqrt{1/\lceil 1/r\rceil}$ having almost disjoint support.

Roughly speaking, the ERM will be such that $\|\widehat{\lambda}\|_2^2$ is close to the argmin of $r \in [0,1] \to r - n^{-1/2}\ell_*\big(\cup_{I\in\Lambda_{\lceil 1/r\rceil}}\sqrt{r}B_2^I\big)$. This argmin is proportional to (3.9.1). In particular, it decreases with the dimension M increasing because for any two radii $1/M \leq s < r \leq 1$ we have

$$\begin{aligned}
&\ell_*\big(\cup_{I\in\Lambda_{\lceil 1/r\rceil}}\sqrt{r}B_2^I\big) - \ell_*\big(\cup_{I\in\Lambda_{\lceil 1/s\rceil}}\sqrt{s}B_2^I\big)\\
&\quad\sim \sqrt{\log\big(c_4 Mr\big)} - \sqrt{\log\big(c_5 Ms\big)} \sim \frac{\log(c_6 r/s)}{\sqrt{\log\big(c_7 Mr\big)}}
\end{aligned}$$

thus the complexity of $\cup_{I\in\Lambda_{\lceil 1/r\rceil}}\sqrt{r}B_2^I$ remains the same (up to some absolute multiplying constant) for any radius r in the bandwidth

$$\left[\frac{c_8}{\sqrt{n\log\big(eM/\sqrt{n}\big)}}, c_9\sqrt{\frac{\log\big(eM/\sqrt{n}\big)}{n}}\right]. \tag{3.9.4}$$

It appears that the residual term derived from the isomorphic profile of $\mathcal{L}_{\mathrm{conv}(F)}$ chooses the largest radius in this interval. This is the worst case: largest radius of a set of complexity $c_{10}\sqrt{\log\big(eM/\sqrt{n}\big)}$. Whereas a direct approach picks up the smallest radius in this interval. Somehow, the isomorphic approach starts the search of the localization of the ERM from "outside" the model (in (3.9.3), the point $\mu_{1/2}^*$ is obtained by decreasing μ up to a point where $8\ell_*\big(B_1^M \cap \sqrt{\mu}\mathcal{S}^{M-1}\big) \leq \sqrt{n}\mu$). Whereas in the direct approach, we start from the center of B_1^M — where the target Y is projected on the model — and we increase the radius r up to the first point where $\ell_*\big(B_1^M \cap \sqrt{r}\mathcal{S}^{M-1}\big) = 2\sqrt{n}r$ (cf. Figure 3.1). Since there is a whole interval (3.9.4) for which $r \to \ell_*\big(B_1^M \cap \sqrt{r}\mathcal{S}^{M-1}\big)$ is constant equal to $\sqrt{\log\big(eM/\sqrt{n}\big)}$ up to some multiplying constants, $\mu_{1/2}^*$ picks the upper bound of this interval whereas the direct approach picks its lower bound. This is the source of the logarithmic gap between the two approaches and the reason of the sub-optimality of the approach based on the isomorphic profile in this context.

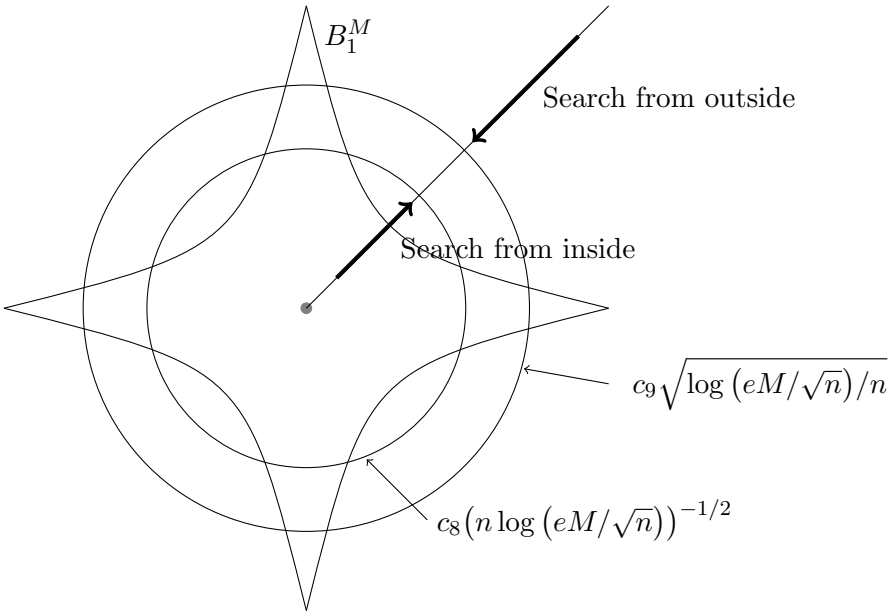

Figure 3.1: The direct approach is searching the localization of the ERM from the center of B_1^M whereas the isomorphic profile approach is searching from outside the model B_1^M. Since there are localized sets $B_1^M \cap \sqrt{r}\mathcal{S}^{M-1}$ with the same complexity — up to some multiplying constants — but different radii the isomorphic profile approach fails to find the correct localization of the ERM in this context.

To finish on the complexity of B_1^M around 0 in high-dimension. The source of complexity of B_1^M with respect to ℓ_2^M at scale $\sqrt{r}$ can be described by an exponential number of ℓ_2^I-balls with short support I of size $\lceil 1/r \rceil$ and radius $\sqrt{r}$ with almost disjoint supports. The complexity of B_1^M results from a trade-off between the distance between the center of these balls and the number of such balls. In B_1^M, there is a whole bandwidth of radii r for which this trade-off results in the same complexity (up to multiplying constants) which is of the order of $\sqrt{\log\left(eM/\sqrt{n}\right)}$. In particular the lower bound of this bandwidth (3.9.4) decreases with the dimension M increasing. This means that, as the dimension M increases, we can find at smaller and smaller radii the same complexity than at larger radii in smaller dimensions. Therefore the complexity of B_1^M around zero increases with the dimension M increasing resulting in a decreasing rate of convergence for the ERM when the oracle in B_1^M is chosen in zero.

At the time we consider this counter-example we did not expect such a residual term (3.9.1). The residual term coming out of the isomorphic approach $\sqrt{\log\left(eM/\sqrt{n}\right)/n}$ was our first guess. But since there are matching upper and lower bounds with the rate (3.9.1) this is the correct rate for the ERM in that context. This means that as the dimension M grows the rate of the ERM get smaller and this due to the fact that the complexity of B_1^M around 0 increases as the dimension grows. This is not the usual way of representing B_1^M in high dimension (cf. the picture of B_1^M in Figure 3.1) since the complexity of B_1^M around zero is usually represented by a ball $M^{-1/2}B_2^M$ which is of constant complexity ($\ell_*(M^{-1/2}B_2^M) = 1$) whereas "outside the

ball $M^{-1/2}B_2^M$" B_1^M is usually represented by spikes. Each spike corresponds to a unit vector $\pm e_i$ on the canonical axis. In particular, the number of spikes increases with the dimension and thus at a first glance, one may think that the complexity outside of $M^{-1/2}B_2^M$ in B_1^M may grow faster than the complexity of B_1^M around zero as the dimension grows. This counter-example in Learning theory shows that this is not the case.

Nevertheless, we still believe that it is possible to construct a counter-example for the ERM-C with a residual term of the order of $\sqrt{\log\big(eM/\sqrt{n}\big)/n}$ when $M > \sqrt{n}$. A possible counter-example can be found in Chapter 5.

3.10 The shifted empirical process and non-exact oracle inequalities

In [P19], we used another approach to prove non-exact oracle inequalities. The idea relies on the fact that comparing an empirical mean $\bar{\zeta}_n = n^{-1}\sum_{i=1}^n \zeta_i$ of i.i.d. real-valued random variables $\zeta_1, \dots, \zeta_n$ to the actual mean $\mathbb{E}\zeta$ is sometimes harder than comparing it to $(1+\epsilon)\mathbb{E}\zeta$ or $(1-\epsilon)\mathbb{E}\zeta$ for some $\epsilon > 0$ when a Margin/Bernstein condition $\mathbb{E}\zeta^2 \le B\big(\mathbb{E}\zeta\big)^\beta$ holds.

Indeed, it follows from Bernstein inequality that for any $0 < \epsilon < 1$ and $0 < x < n$, with probability greater than $1 - 2\exp(-x)$,

$$\begin{aligned}\big|\bar{\zeta}_n - \mathbb{E}\zeta\big| &\le K\sigma(\zeta)\sqrt{\frac{x}{n}} + K\,\|\zeta\|_\infty \frac{x}{n} \le K\sqrt{\frac{xB\big(\mathbb{E}\zeta\big)^\beta}{n}} + K\,\|\zeta\|_\infty \frac{x}{n}\\ &\le \epsilon\mathbb{E}\zeta + \Big[\Big(\frac{K\sqrt{B}}{\epsilon}\Big)^{\frac{2}{2-\beta}} + K\,\|\zeta\|_\infty\Big]\Big(\frac{x}{n}\Big)^{\frac{1}{2-\beta}}.\end{aligned}$$

In other words, for $c(\zeta) = \big[\big(K\sqrt{B}/\epsilon\big)^{\frac{2}{2-\beta}} + K\,\|\zeta\|_\infty\big]$, with probability greater than $1-2\exp(-x)$,

$$-c(\zeta)\Big(\frac{x}{n}\Big)^{\frac{1}{2-\beta}} + (1-\epsilon)\mathbb{E}\zeta \le \bar{\zeta}_n \le (1+\epsilon)\mathbb{E}\zeta + c(\zeta)\Big(\frac{x}{n}\Big)^{\frac{1}{2-\beta}}. \tag{3.10.1}$$

The residual term $(x/n)^{1/(2-\beta)}$ in (3.10.1) is always better than the residual term $\sqrt{x/n}$ in the Bernstein inequality.

We want to use this remark to obtain non-exact oracle inequalities for the ERM with fast residual terms. Let $F \subset \mathcal{F}$ be a model and consider the ERM over F:

$$\widehat{f}_n^{ERM} \in \operatorname*{argmin}_{f\in F} R_n(f).$$

First consider the non-exact prediction problem. We have

$$\begin{aligned}&R(\widehat{f}_n^{ERM}) - (1+4\epsilon)\inf_{f\in F} R(f) = P\ell_{\widehat{f}_n^{ERM}} - (1+2\epsilon)P_n\ell_{\widehat{f}_n^{ERM}}\\ &+ (1+2\epsilon)P_n\ell_{\widehat{f}_n^{ERM}} - (1+2\epsilon)P_n\ell_{f_F^*} + (1+2\epsilon)P_n\ell_{f_F^*} - (1+4\epsilon)P\ell_{f_F^*}\\ &\le \sup_{f\in F}\Big(\big(P-(1+2\epsilon)P_n\big)(\ell_f)\Big) + (1+2\epsilon)\sup_{f\in F}\Big(P_n - \frac{1+4\epsilon}{1+2\epsilon}P\Big)(\ell_f).\end{aligned}$$

Thus, it is enough to bound the supremum of the shifted empirical processes $\big((P-(1+\eta)P_n)\ell\big)_{\ell\in\ell_F}$ and $\big((P_n-(1+\eta)P)\ell\big)_{\ell\in\ell_F}$, for some $0 < \eta < 1$ and $\ell_F = \{\ell_f : f \in F\}$, to

obtain oracle inequalities for the ERM for the non-exact prediction problem. In an identical manner, we obtain the following bound for the estimation problem:

$$R(\widehat{f}_n^{ERM}) - R(f^*) - (1+4\epsilon)\inf_{f\in F}\big(R(f) - R(f^*)\big)$$
$$\leq \sup_{f\in F}\Big(\big(P-(1+2\epsilon)P_n\big)(\ell_f - \ell_{f^*})\Big) + (1+2\epsilon)\sup_{f\in F}\Big(P_n - \frac{1+4\epsilon}{1+2\epsilon}P\Big)(\ell_f - \ell_{f^*}).$$

Therefore, oracle inequalities for the ERM in the estimation problem follow from upper bounds on the supremum of the shifted empirical processes $\big((P-(1+\eta)P_n)\mathcal{E}\big)_{\mathcal{E}\in\mathcal{E}_F}$ and $\big((P_n-(1+\eta)P)\mathcal{E}\big)_{\mathcal{E}\in\mathcal{E}_F}$ for some $0<\eta<1$ and $\mathcal{E}_F = \{\ell_f - \ell_{f^*} : f\in F\}$.

In the following result, we control the deviation and expectation of the supremum of the "shifted" empirical process.

Theorem 3.10.1 ([P19]) *Let $\eta>0$ and G be a set of real-valued measurable functions defined on $\mathcal{Z}$. Let $Z, Z_1, \ldots, Z_n$ be i.i.d. random variables with values in $\mathcal{Z}$ such that $\forall g\in G, Pg = \mathbb{E}g(Z)\geq 0$. Suppose that there exists some constants $c, L, \lambda_{min}>0$ such that for all $\lambda\geq\lambda_{min}$ and all $u\geq 1$, with probability greater than $1-L\exp(-cu)$*

$$\sup_{g\in G: Pg\leq\lambda}((P-P_n)g)_+ \leq \frac{uJ(\lambda)}{\sqrt{n}}, \tag{3.10.2}$$

where J is a strictly increasing function such that J^{-1} is strictly convex. Let ψ be the convex conjugate of J^{-1} defined by $\psi(u) = \sup_{v>0}(uv - J^{-1}(v)), \forall u>0$. Assume that for some $r\geq 1$, $x>0\longmapsto\psi(x)/x^r$ decreases and define for $q>1$ and $u\geq 1$,

$$\lambda_q(u) = \psi\Big(\frac{2q^{r+1}(1+\eta)u}{\eta\sqrt{n}}\Big)\vee\lambda_{min}.$$

Then, there exists a constant L_1 (depending only on L) such that for every $u\geq 1$, with probability greater than $1-L_1\exp(-cu)$

$$\sup_{g\in G}\Big((P-(1+\eta)P_n)g\Big)_+ \leq \frac{\eta\lambda_q(u/q)}{q}.$$

Moreover, assume that ψ increases such that $\psi(\infty)=\infty$, then there exists a constant c_1 depending only on L and c such that

$$\mathbb{E}\sup_{g\in G}\Big((P-(1+\eta)P_n)g\Big)_+ \leq \frac{\eta c_1\lambda_q(1/q)}{q}$$

The function $\lambda>0\longmapsto\sup_{g\in G:Pg\leq\lambda}(P-P_n)g$, appearing in Equation (3.10.4), is a classical measure of the complexity of the set of functions G (cf. for instance [109], [15], [57] and references therein). A common way to upper bound this function is to use some exponential bounds on the increments of the empirical process (where some Margin/Bernstein condition may improve the bounds) together with a chaining argument. This results in an exponential bounds depending on some metric complexity measure like the Dudley entropy integral (cf. for instance [113] or Section 3.2.3) or the gamma functional (cf. [101] or Section 3.2.3). In particular, this way of bounding empirical processes does not require Talagrand concentration inequality and has been extensively used in Statistics before [100] and also after to handle the unbounded case. Thanks to [1] (cf. Theorem 3.2.1), the unbounded case can now also be handled by Talagrand inequality.

It follows from Theorem 3.10.1, oracle inequalities for the ERM over F in the non-exact prediction problem.

Theorem 3.10.2 *Let $F \subset \mathcal{F}$ be a model, ℓ be a loss function and $0 < \epsilon < 1$. Assume that for any $f \in F, P\ell_f \geq 0$. Suppose that there exists some constants $c, L, \lambda_{min} > 0$ such that for all $\lambda \geq \lambda_{min}$ and all $u \geq 1$, with probability greater than $1 - L\exp(-cu)$*

$$\sup_{\ell \in \ell_F : P\ell \leq \lambda} ((P - P_n)\ell)_+, \sup_{\ell \in \ell_F : P\ell \leq \lambda} ((P_n - P)\ell)_+ \leq \frac{uJ(\lambda)}{\sqrt{n}}, \tag{3.10.3}$$

where J is a strictly increasing function such that J^{-1} is strictly convex. Let ψ be the convex conjugate of J^{-1} defined by $\psi(u) = \sup_{v>0}(uv - J^{-1}(v)), \forall u > 0$. Assume that for some $r \geq 1$, $x > 0 \longmapsto \psi(x)/x^r$ decreases and define for $q > 1$ and $u \geq 1$,

$$\lambda_q(u) = \psi\Big(\frac{2q^{r+1}(1+4\epsilon)u}{\epsilon\sqrt{n}}\Big) \vee \lambda_{min}.$$

Then, there exists a constant L_1 (depending only on L) such that for every $u \geq 1$, with probability greater than $1 - L_1\exp(-cu)$

$$R(\widehat{f}_n^{ERM}) \leq (1+4\epsilon) \inf_{f \in F} R(f) + \frac{8\epsilon\lambda_q(u/q)}{q}.$$

It follows from Theorem 3.10.1, oracle inequalities for the ERM in the estimation problem.

Theorem 3.10.3 *Let $F \subset \mathcal{F}$ be a model, ℓ be a loss function and $0 < \epsilon < 1$. Suppose that there exists some constants $c, L, \lambda_{min} > 0$ such that for all $\lambda \geq \lambda_{min}$ and all $u \geq 1$, with probability greater than $1 - L\exp(-cu)$*

$$\sup_{\mathcal{E} \in \mathcal{E}_F : P\mathcal{E} \leq \lambda} ((P - P_n)\mathcal{E})_+, \sup_{\mathcal{E} \in \mathcal{E}_F : P\mathcal{E} \leq \lambda} ((P_n - P)\mathcal{E})_+ \leq \frac{uJ(\lambda)}{\sqrt{n}}, \tag{3.10.4}$$

where J is a strictly increasing function such that J^{-1} is strictly convex. Let ψ be the convex conjugate of J^{-1} defined by $\psi(u) = \sup_{v>0}(uv - J^{-1}(v)), \forall u > 0$. Assume that for some $r \geq 1$, $x > 0 \longmapsto \psi(x)/x^r$ decreases and define for $q > 1$ and $u \geq 1$,

$$\lambda_q(u) = \psi\Big(\frac{2q^{r+1}(1+4\epsilon)u}{\epsilon\sqrt{n}}\Big) \vee \lambda_{min}.$$

Then, there exists a constant L_1 (depending only on L) such that for every $u \geq 1$, with probability greater than $1 - L_1\exp(-cu)$

$$R(\widehat{f}_n^{ERM}) - R(f^*) \leq (1+4\epsilon) \inf_{f \in F} \big(R(f) - R(f^*)\big) + \frac{8\epsilon\lambda_q(u/q)}{q}.$$

Theorem 3.10.2 and Theorem 3.10.3 provide an alternative way of proving oracle inequalities for the ERM to the results in Section 3.3. In particular, the two last results do not require any tail assumption on the loss functions $\ell_f(Z), f \in F$ (nevertheless, explicit computation of the function J requires such deviation bounds). Whereas in Section 3.3, the envelop $\sup_{f \in F} \ell_f$ needs to be sub-exponential. We use these results in [P19] to obtain oracle inequalities for Cross-Validation type procedures. These results are recalled in the following chapter devoted to applications.

Chapter 4

Applications to High-Dimensional data analysis

In this chapter, we obtain oracle inequalities for different ERM and regularized ERM procedures. These results are derived from the general oracle inequalities obtained in Chapter 3.

4.1 ℓ_1-regularization

The formulation of Theorem 3.5.1 seems cumbersome, but it is not very difficult to apply it — and here we will present one application dealing with high-dimensional vectors of short support.

Formally, let $(X,Y),(X_i,Y_i)_{1\leq i\leq n}$ be $n+1$ i.i.d. random variables with values in $\mathbb{R}^d\times\mathbb{R}$ and denote by P_X the marginal distribution of X. The dimension d can be much larger than n but we believe that the output Y can be well predicted by a sparse linear combination of covariables of X: Y can be reasonably approximated by $\langle X,\beta_0\rangle$ for some $\beta_0\in\mathbb{R}^d$ of short support (even though, we do not need any assumption of this type to obtain our results). These kind of problems are called high-dimensional problems because there are more covariables than observations. Nevertheless, one hopes that under the structural assumption that Y "depends" only on a few number of covariables of X, it would still be possible to construct efficient statistical procedures to predict Y.

In this framework, a natural criterion function is the ℓ_0 function measuring the size of the support of a vector. But since this function is far from being convex, using it in practice is hard (cf. [87]). Therefore, it is natural to consider a convex relaxation of the ℓ_0 function as a criterion: the ℓ_1 norm (see e.g. [102, 36]). In what follows, we apply Theorem 3.5.1 to show non-exact regularized oracle inequalities for ℓ_1-based regularized ERM procedures, and with fast error rates — a residual term that tends to 0 like $1/n$ up to logarithmic terms. The regularizing function resulting from Theorem 3.5.1 for the L_q-loss ($q\geq 2$) will be the q-th power of the ℓ_1-norm. In particular, for the quadratic loss, we regularize by $\|\cdot\|_{\ell_1}^2$, the *square of the ℓ_1-norm*:

$$\widehat{\beta}_n\in\operatorname*{argmin}_{\beta\in\mathbb{R}^d}\Big(\frac{1}{n}\sum_{i=1}^n(Y_i-\langle X_i,\beta\rangle)^2+\kappa(n,d,x)\frac{\|\beta\|_{\ell_1}^2}{n}\Big) \tag{4.1.1}$$

while the standard LASSO regularizes by the ℓ_1 norm itself. This choice of the exponent is dictated by the complexity of the underlying models: the sequence of ℓ_1-balls $(rB_1^d)_{r\geq 0}$ in the spirit of [12, 76], trough the isomorphic profile function $r\to\lambda_\eta^*(r)$ of the family of loss functions class $\big(\ell_{F_r}\big)_{r\geq 0}$ as defined in Section 3.5. Observe that since $\|\beta\|_{\ell_1}/\sqrt{n}\geq\|\beta\|_{\ell_1}^2/n$

when $\|\beta\|_{\ell_1} \leq \sqrt{n}$, a non-exact oracle inequality for the LASSO estimator itself follows from Theorem 3.5.1, but with a slow rate of $1/\sqrt{n}$. Using the q-th power of the ℓ_1-norm as a penalty function for the L_q-risk yields a fast $1/n$ rate (see Theorem 4.1.1 below). Finally, note that for the case $q = 2$, it follows from the theory of proximal operator (cf. [93]) that the estimator $\widehat{\beta}_n$ defined in (4.1.1) is the solution of the fixed point equation $\widehat{\beta}_n = \mathrm{prox}(\widehat{\beta}_n + 2\mathbb{X}^\top y - 2\mathbb{X}^\top \mathbb{X}\widehat{\beta}_n)$ where prox is some multidimensional threshold operator. Indeed, denote $y = (Y_1, \ldots, Y_n)^\top$ the vector of outputs and $\mathbb{X} \in \mathcal{M}_{n,d}$ the design matrix with rows vectors $X_i^\top, 1 \leq i \leq n$. Define the functions $f_1(\beta) = \kappa \|\beta\|_{\ell_1}^2$ where $\kappa = \kappa(n, d, x)$ and $f_2(\beta) = \|y - \mathbb{X}\beta\|_{\ell_2}^2$ for any $\beta \in \mathbb{R}^d$ where $\|z\|_{\ell_2}^2 = \sum_{i=1}^n z_i^2, \forall z \in \mathbb{R}^n$. We have $\widehat{\beta}_n \in \mathrm{argmin}_{\beta \in \mathbb{R}^d} \big(f_2(\beta) + f_1(\beta)\big)$ in particular, $0 \in \partial^-(f_2 + f_1)(\widehat{\beta}_n) = \{\nabla f_2(\widehat{\beta}_n)\} + \partial^- f_1(\widehat{\beta}_n)$ where ∂^- denotes the sub-differential multidimensional mapping and ∇ denotes the gradient operator. In particular, $\widehat{\beta}_n$ is such that

$$-\nabla f_2(\widehat{\beta}_n) \in \partial^- f_1(\widehat{\beta}_n). \tag{4.1.2}$$

On the other hand, the proximal operator of the convex function f_1 is defined for any $\alpha \in \mathbb{R}^d$ by

$$\mathrm{prox}_{f_1}(\alpha) = \underset{\beta \in \mathbb{R}^d}{\mathrm{argmin}} \Big(\frac{1}{2} \|\alpha - \beta\|_{\ell_2}^2 + f_1(\beta) \Big),$$

the minimizer being unique because of the strict convexity of $\beta \to (1/2)\|\alpha - \beta\|_{\ell_2}^2 + f_1(\beta)$. But since $\partial^-\big((1/2)\|\alpha - \cdot\|_{\ell_2}^2 + f_1(\cdot)\big)(\beta) = \{\beta - \alpha\} + \partial^- f_1(\beta)$ the point $\mathrm{prox}_{f_1}(\alpha)$ is the unique solution of

$$\alpha - \mathrm{prox}_{f_1}(\alpha) \in \partial^-(\mathrm{prox}_{f_1}(\alpha)). \tag{4.1.3}$$

This holds for any $\alpha \in \mathbb{R}^d$. In particular, for $\alpha = \widehat{\beta}_n - \nabla f_2(\widehat{\beta}_n)$, we have $\widehat{\beta}_n - \nabla f_2(\widehat{\beta}_n) - \mathrm{prox}_{f_1}(\widehat{\beta}_n - \nabla f_2(\widehat{\beta}_n)) \in \partial^- f_1(\mathrm{prox}_{f_1}(\widehat{\beta}_n - \nabla f_2(\widehat{\beta}_n)))$. But it follows from (4.1.2) that $\widehat{\beta}_n - \nabla f_2(\widehat{\beta}_n) - \widehat{\beta}_n \in \partial^- f_1(\widehat{\beta}_n)$. Thus $\widehat{\beta}_n$ and $\mathrm{prox}_{f_1}(\widehat{\beta}_n - \nabla f_2(\widehat{\beta}_n))$ satisfy the same equation having only one solution therefore $\widehat{\beta}_n = \mathrm{prox}_{f_1}(\widehat{\beta}_n - \nabla f_2(\widehat{\beta}_n))$. To compute the proximal function of f_1, we use (4.1.3) and $\partial^- f_1(\beta) = 2\|\beta\|_{\ell_1} \big(\partial^- \|\cdot\|_{\ell_1}\big)(\beta)$ to check that for any $\alpha \in \mathbb{R}^d$, we can take $\mathrm{prox}_{f_1}(\alpha)$ such that, for any $1 \leq i \leq d$,

$$\big(\mathrm{prox}_{f_1}(\alpha)\big)_i = \begin{cases} 0 & \text{if } |\alpha_i| \leq 2\kappa t_0(\alpha) \\ \alpha_i - 2\kappa t_0(\alpha) & \text{if } \alpha_i > 2\kappa t_0(\alpha) \\ \alpha_i + 2\kappa t_0(\alpha) & \text{if } \alpha_i < -2\kappa t_0(\alpha), \end{cases}$$

where $t_0(\alpha)$ is the unique solution of $t_0(\alpha) = \sum_{i:|\alpha_i| > 2\kappa t_0(\alpha)} \big(|\alpha_i| - 2\kappa t_0(\alpha)\big)$. In particular, in addition to its prediction properties, $\widehat{\beta}_n$ may also enjoy some support recovery or estimation properties.

Now, we turn to oracle inequalities for the regularized ERM introduced here. We will perform this study for the L_q-loss function, and in which case, for every $\beta \in \mathbb{R}^d$,

$$R^{(q)}(\beta) = \mathbb{E}|Y - \langle X, \beta \rangle|^q \text{ and } R_n^{(q)}(\beta) = \frac{1}{n} \sum_{i=1}^n |Y_i - \langle X_i, \beta \rangle|^q.$$

The following result is obtained only under the assumption that Y and $\|X\|_{\ell_\infty^d}$ belong to L_{ψ_q}. Since there are no "statistically reasonable" ψ_q variables for $q > 2$, it sounds more "statistically relevant" to assume that $|Y|, \|X\|_{\ell_\infty^d}$ are almost surely bounded when one wants results for the L_q-risk with $q > 2$, or that the functions are in L_{ψ_2} for $q = 2$ (for example, linear models with sub-gaussian noise and a sub-gaussian design satisfy this condition).

Theorem 4.1.1 ([P16]) *Let $q \geq 2$. There exist constants c_0 and c_1 that depend only on q for which the following holds. Assume that there exists some constant $c_d > 0$ (which may depend only on d) such that $\|Y\|_{\psi_q}$, $\left\| \|X\|_{\ell_\infty^d} \right\|_{\psi_q} \leq c_d$. For $x > 0$ and $0 < \epsilon < 1/2$, let*

$$\lambda(n,d,x) = c_0 c_d^q (\log n)^{(4q-2)/q} (\log d)^2 (x + \log n)$$

and consider the regularized ERM estimator

$$\widehat{\beta}_n \in \operatorname*{argmin}_{\beta \in \mathbb{R}^d} \Big(R_n^{(q)}(\beta) + \lambda(n,d,x) \frac{\|\beta\|_{\ell_1}^q}{n\epsilon^2} \Big).$$

Then, with probability greater than $1 - 12\exp(-x)$, the L_q-risk of $\widehat{\beta}_n$ satisfies

$$R^{(q)}(\widehat{\beta}_n) \leq \inf_{\beta \in \mathbb{R}^d} \Big((1+2\epsilon) R^{(q)}(\beta) + \eta(n,d,x) \frac{(1 + \|\beta\|_{\ell_1}^q)}{n\epsilon^2} \Big),$$

where $\eta(n,d,x) = c_1 c_d^q (\log n)^{(4q-2)/q} (\log d)^2 (x + \log n)$.

Procedures based on the ℓ_1-norm as a regularizing or constraint function have been studied extensively in the last few years. We only mention a small fraction of this very extensive body of work [19, 27, 36, 58, 60, 73, 78, 79, 102, 110, 122, 123]. In fact, it is almost impossible to make a proper comparison even with the results mentioned in this partial list. Some of these results are close enough in nature to Theorem 4.1.1 to allow a comparison. In particular, in [16], the authors prove that with high probability, the LASSO satisfies an exact oracle inequality with a residual term $\sim \|\beta\|_{\ell_1} / \sqrt{n}$ up to logarithm factors, under tail assumptions on Y and X. In [27], upper bounds on the risks $\mathbb{E}[\langle X, \widehat{\beta}_n - \beta_0 \rangle^2]$ and $\left\| \widehat{\beta}_n - \beta_0 \right\|_{\ell_1}$ were obtained for a weighted LASSO $\widehat{\beta}_n$ when $\mathbb{E}(Y|X) = \langle X, \beta_0 \rangle$ for β_0 with short support. Exact oracle inequalities for regularized ERM, based on an entropy or on an ℓ_p with p close to 1 criterion were obtained in [59, 60] for any convex and regular loss function and with fast rates. Similar bounds were obtained in [110] for a regularized ERM using a weighted ℓ_1-criterion. In [19] it is shown that the LASSO and Dantzig estimators [36] satisfy oracle inequalities in the deterministic design setup and under the REC condition. In fact, in most of these results the authors obtained exact oracle inequalities with an optimal residual term of $|\mathrm{Supp}(\beta_0)|(\log d)/n$, which is clearly better than the rate $\|\beta\|_{\ell_1}^2 / n$ obtained in Theorem 4.1.1 for the quadratic loss and in the same context.

However, it is important to note that all these exact oracle inequalities were obtained under an assumption that is similar in nature to the Restricted Isometry Property (RIP), whereas in Theorem 4.1.1 one does not need that kind of assumption on the design. Although it seems strange that it is possible to obtain fast rates without RIP there is nothing magical. In fact, the isomorphic argument used to prove Theorem 3.5.1 (and thus Theorem 4.1.1) shows that the random operator $\beta \in \mathbb{R}^d \to n^{-1/2} \sum_{i=1}^n (Y_i - \langle X_i, \beta \rangle) e_i \in \mathbb{R}^n$ satisfies some sort of an RIP. And that this isomorphic property coincides with the RIP property in the non noisy case $Y = \langle X, \beta_0 \rangle$ with isotropic design (cf. Section 3.8 for more details on the comparisons between the RIP and the isomorphic properties of the loss functions class). This indicates that RIP is not the key property in establishing oracle inequalities for the prediction risk, but rather, the "isomorphic profile" of the problem at hand, which takes into account the structure of the class of functions.

4.2 S_1-regularization

For the second application of Theorem 3.5.1, we observe n i.i.d. couples input/output $(X_i, Y_i)_{1\leq i\leq n}$ where the input variables $X_1, \dots, X_n$ take their values in the space $\mathcal{X} = \mathcal{M}_{m\times T}$ of all $m \times T$ matrices with entries in $\mathbb{R}$ and the output variables $Y_1, \dots, Y_n$ are real-valued. Being given a new input X, the goal is to predict the output Y using a linear function of X when (X, Y) is assumed to have the same probability distribution as the (X_i, Y_i)'s. In this setup, it is now common to assume that there are more covariables than observations ($mT \gg n$) and thus more information on the best linear prediction of Y by X are required. A common assumption is that Y can be well predicted by a function of the form $\langle X, A_0\rangle = \mathrm{Tr}(X^\top A_0)$ where A_0 is a $m \times T$ matrix of low rank. Once again this will not constitute any assumption in our framework since our techniques are taken from Learning theory where one try to assume as little as possible. Nevertheless, it can be useful to treat this problem having this low-dimensional structure in mind.

Indeed, having this structure in mind, it is natural to penalize linear estimators $\langle X, A\rangle$ where A is of large rank. Unfortunately, the rank$(\cdot)$ function is not convex and thus cannot be used in practice as a criterion. A more popular choice is to use a convex relaxation of the rank$(\cdot)$ function: the S_1 norm ("Schatten one" norm) (see [11, 31, 32, 29, 34, 55, P6, 50, 88, 94, 63] and references therein), which is the ℓ_1-norm of the singular values of a matrix. Formally, for every $A \in \mathcal{M}_{m\times T}$, $\|A\|_{S_1} = \sum_{i=1}^{m\wedge T} s_i(A)$, where $s_1(A), \dots, s_{m\wedge T}(A)$ are the singular values of A and, in general for $p \geq 1$, $\|A\|_{S_p} = \left(\sum_{i=1}^{m\wedge T} s_i(A)^p\right)^{1/p}$. The S_1-norm was originally used in this type of problems to study exact reconstruction (see, for example, [34, 92, 33]), but other regularizing functions have been used in this context (e.g. [56, 49, P6]) for the prediction and estimation problems.

In the following result, we apply Theorem 3.5.1 to obtain non-exact oracle inequalities for a S_1-based regularized ERM procedure for the L_q-loss function, for some $q \geq 2$. For every $A \in \mathcal{M}_{m\times T}$ let

$$R^{(q)}(A) = \mathbb{E}|Y - \langle X, A\rangle|^q \text{ and } R_n^{(q)}(A) = \frac{1}{n}\sum_{i=1}^{n} |Y_i - \langle X_i, A\rangle|^q.$$

Again, it seems more "statistically relevant" to assume that $|Y|$ and $\|X\|_{S_2}$ are almost surely bounded rather than bounded in ψ_q for $q > 2$, and the two most interesting cases are the uniformly bounded one and the subgaussian case for $q = 2$. We have stated the results under the more general ψ_q assumption to point out the places in which the decay properties of the functions involved are really needed.

Theorem 4.2.1 ([P16]) *For every $q \geq 2$ there are constants c_0 and c_1 depending only on q for which the following holds. Let m and T be as above and assume that there exists some constant c_{mT} which may depend only on the product mT such that $\|Y\|_{\psi_q}, \big\|\|X\|_{S_2}\big\|_{\psi_q} \leq c_{mT}$. Let $x > 0$ and $0 < \epsilon < 1/2$, and put $\lambda(n, mT, x) = c_0 c_{mT}^q (\log n)^{(4q-2)/q}(x + \log n)$. Consider the regularized ERM procedure*

$$\widehat{A}_n \in \operatorname*{argmin}_{A\in\mathcal{M}_{m\times T}} \left(R_n^{(q)}(A) + \lambda(n, mT, x)\frac{\|A\|_{S_1}^q}{n\epsilon^2}\right)$$

Then, with probability greater than $1 - 10\exp(-x)$, the L_q-risk of $\widehat{A}_n$ satisfies for $\eta_\epsilon(n, mT, x) = c_1 c_{mT}^q (\log n)^{(4q-2)/q}(x + \log n)$

$$R^{(q)}(\widehat{A}_n) \leq \inf_{A\in\mathcal{M}_{m\times T}} \left((1 + 2\epsilon)R^{(q)}(A) + \eta(n, mT, x)\frac{(1 + \|A\|_{S_1}^q)}{n\epsilon^2}\right).$$

Once again, in the same spirit as in Theorem 4.1.1, it can be interesting to note that for the quadratic loss ($q = 2$), the estimator which comes out of our analysis is

$$\widehat{A}_n \in \operatorname*{argmin}_{A \in \mathcal{M}_{m\times T}} \Big(\frac{1}{n}\sum_{i=1}^{n} (Y_i - \langle X_i, A\rangle)^2 + \lambda(n, mT, x)\frac{\|A\|_{S_1}^2}{n\epsilon^2} \Big)$$

where the regularizing function uses the *square* of the S_1-norm unlike the classical estimator for this problem which uses the S_1 norm itself as a regularizing function.

The first results in the direction of matrix completion have focused on the exact reconstruction of a low-rank matrix A_0 where $Y_i = \langle X_i, A_0\rangle, i = 1, \ldots, n$ [31, 32, 34, 92, 50]. The best results [92, 50] to date are that if the number of measurements n is larger than $\operatorname{rank}(A_0)(m+T)\log(m+T)$ and if the "incoherence condition" holds (see [34] for more details), then with high probability, a constrained nuclear norm minimization algorithm can reconstruct A_0 exactly.

Prediction results and statistical estimation involving low-rank matrices has become a very active field. The most popular methods are regularized ERM based on S_1-norm penalty functions (cf. for instance [11, 26, 31, 32, 88, 94, P6, 56, 63, 94]). To specify some results, fast rates for the noisy matrix completion problem are derived in [94] — in the context of empirical prediction and under an RIP-type assumption. In [63] the authors prove exact oracle inequalities for the prediction error $\mathbb{E}\langle X, \widehat{A}_n - A_0\rangle^2$ when $\mathbb{E}(Y|X) = \langle X, A_0\rangle$ and A_0 is either of low-rank or of a small S_1-norm, and when the probability distribution of the design of X is known. In [88], optimal rates for the quadratic risk were obtained under a "spikiness assumption" on the SVD of A_0, and in [56], fast convergence rates were derived for regularized ERM based on the von Neuman entropy penalization and for a known design. However, so far only few results have been obtained for the prediction risk as considered here. Probably the closest result in this setup is an exact oracle inequality with slow rates satisfied by a regularized ERM using a mixture of several norms in [P6] which is developed in the next section.

Note that for the two applications in Theorem 4.1.1 and Theorem 4.2.1, we obtain fast convergence rates under only tail assumptions on the design X and the output Y for every L_q-loss (for $q \geq 2$). In particular, one does not need to assume that $\mathbb{E}(Y|X)$ is a linear combination of the covariables of X, nor that Y has any low-dimensional structure. If one happens to be in a "low-dimensional" situation, the residual terms of Theorem 4.1.1 and Theorem 4.2.1 will be small. Hence, the ℓ_1 and S_1 based regularized ERM procedures used there automatically adapt to this low-dimensional structure.

4.3 Exact oracle inequalities for high-Dimensional Matrix prediction

In this third application, we consider the matrix completion setup as introduced in the previous Section 4.2. We prove an exact oracle inequality for a regularized ERM with a regularizing function being a mixture of several norms. In particular, instead of applying Theorem 3.5.1 which led to a non-exact oracle inequality in Theorem 4.2.1, we apply Theorem 3.5.3.

If $\mathbb{E}(Y|X) = \langle X, A_0\rangle$ where A_0 is low rank, in the sense that $\operatorname{rank}(A_0) \ll n$, nuclear norm regularization is likely to enjoy some good prediction performances. But, if we know more about the properties of A_0, then some additional regularization can be considered. For instance, if we know that the non-zero singular values of A_0 are "well-spread" (that is almost equal) then it may be interesting to use the "regularization effect" of the S_2 norm in the same spirit as a "ridge"

regularization for vectors or functions. Moreover, if we know that many entries of A_0 are close or equal to zero, then using also a ℓ_1-regularization on the entries

$$A \mapsto \|A\|_1 = \sum_{\substack{1\le p\le m\\ 1\le q\le T}} |A_{p,q}| \tag{4.3.1}$$

may improve even further the prediction. As a consequence, we consider in this section, a regularizing function that uses a mixture of several norms: for $\lambda_1, \lambda_2, \lambda_3 > 0$, we consider

$$\mathrm{reg}_{\lambda_1,\lambda_2,\lambda_3}(A) = \lambda_1 \|A\|_{S_1} + \lambda_2 \|A\|_{S_2}^2 + \lambda_3 \|A\|_1 \tag{4.3.2}$$

and we study the prediction properties of the regularized ERM

$$\widehat{A}_n(\lambda_1,\lambda_2,\lambda_3) \in \operatorname*{argmin}_{A\in\mathcal{M}_{m,T}} \Big\{ R_n(A) + \mathrm{reg}_{\lambda_1,\lambda_2,\lambda_3}(A) \Big\} \tag{4.3.3}$$

where $R_n(A) = n^{-1}\sum_{i=1}^n \big(Y_i - \langle X_i, A\rangle\big)^2, \forall A \in \mathcal{M}_{m,T}$. Of course, if more is known on the structure of A_0, other regularizing functions can be considered.

We obtain sharp oracle inequalities for the procedure $\widehat{A}_n(\lambda_1,\lambda_2,\lambda_3)$ for any values of $\lambda_1, \lambda_2, \lambda_3 \ge 0$ (excepted for $(\lambda_1,\lambda_2,\lambda_3) = (0,0,0)$ which provides the well-studied empirical risk minimization procedure). In particular, depending on the "a priori" knowledge that we have on A_0 we will consider different values for the triple $(\lambda_1,\lambda_2,\lambda_3)$. If A_0 is only known to be low-rank, one should choose $\lambda_1 > 0$ and $\lambda_2 = \lambda_3 = 0$. If A_0 is known to be low-rank with many zero entries, one should choose $\lambda_1, \lambda_3 > 0$ and $\lambda_2 = 0$. If A_0 is known to be low-rank with well-spread non-zero singular values, one should choose $\lambda_1, \lambda_2 > 0$ and $\lambda_3 = 0$. Finally, one should choose $\lambda_1, \lambda_2, \lambda_3 > 0$ when a significant part of the entries of A_0 are zero, that A_0 is low rank and that the non-zero singular values of A_0 are well-spread.

4.3.1 Assumptions and examples

We will use the following notation: for a matrix $A \in \mathcal{M}_{m,T}$, $\mathrm{vec}(A)$ denotes the vector of $\mathbb{R}^{mT}$ obtained by stacking its columns into a single vector. Note that this is an isometry between $(\mathcal{M}_{m,T}, \|\cdot\|_{S_2})$ and $(\mathbb{R}^{mT}, |\cdot|_{\ell_2^{mT}})$ since $\langle A, B\rangle = \langle \mathrm{vec}\, A, \mathrm{vec}\, B\rangle$. We introduce also the ℓ_∞ norm $\|A\|_\infty = \max_{p,q}|A_{p,q}|$. Let us recall that for $\alpha \ge 1$, the ψ_α-norm of a random variable Z is given by $\|Z\|_{\psi_\alpha} := \inf\{c > 0 : \mathbb{E}[\exp(|Z|^\alpha/c^\alpha))] \le 2\}$ (cf. [66], p. 10) and a similar norm can be defined for $0 < \alpha < 1$.

The first assumption concerns the "covariate" matrix X.

Assumption 4.3.1 (Matrix X) *There are positive constants $b_{X,\infty}, b_{X,\ell_\infty}$ and $b_{X,2}$ such that $\|X\|_{S_\infty} \le b_{X,\infty}$, $\|X\|_\infty \le b_{X,\ell_\infty}$ and $\|X\|_{S_2} \le b_{X,2}$ almost surely. Moreover, we assume that the "covariance matrix"*

$$\Sigma := \mathbb{E}[\mathrm{vec}\, X (\mathrm{vec}\, X)^\top]$$

is invertible.

This assumption is very mild. It is met in the matrix completion and the multitask-learning problems, defined below.

Example. [*Uniform matrix completion*] The matrix X is uniformly distributed over the set $\{e_{p,q} : 1 \le p \le m, 1 \le q \le T\}$ where $e_{p,q}$ is the $m \times T$ matrix with zero entries everywhere

except $(e_{p,q})_{p,q} = 1$. In this case $\Sigma = (mT)^{-1} I_{mT}$ (where I_{mT} stands for the identity matrix on $\mathbb{R}^{mT}$) and $b_{X,2} = b_{X,\infty} = b_{X,\ell_\infty} = 1$.

Example. [*Weighted matrix completion*] The distribution of X is such that $\mathbb{P}(X = e_{p,q}) = \pi_{p,q}$ where $(\pi_{p,q})_{1\le p\le m, 1\le q\le T}$ is a family of positive numbers summing to 1. In this situation Σ is invertible and again $b_{X,2} = b_{X,\infty} = b_{X,\ell_\infty} = 1$.

Example. [Multitask-learning, or "column-masks"] The distribution of X is uniform over a set of matrices with only one non-zero column (all the columns have the same probability to be non-zero). The distribution is such that the j-th column takes values in a set $\{x_{j,s} : s = 1, \dots, k_j\}$, each vector having the same probability. So, in this case Σ is equal to T^{-1} times the $mT \times mT$ block matrix with T diagonal blocks of size $m \times m$ made of the T matrices $k_j^{-1} \sum_{i=1}^{k_j} x_{j,s} x_{j,s}^\top$ for $j = 1, \dots, T$.

If we assume that the smallest singular values of the matrices $k_j^{-1} \sum_{i=1}^{k_j} x_{j,s} x_{j,s}^\top \in \mathcal{M}_{m,m}$ for $j = 1, \dots, T$ are larger than a constant $\sigma_{\min}$ (note that this implies $k_j \ge m$), then Σ has its smallest singular value larger than $\sigma_{\min} T^{-1}$, so it is invertible. Moreover, if the vectors $x_{j,s}$ are normalized in ℓ_2, then one can take $b_{X,\infty} = b_{X,\ell_\infty} = b_{X,2} = 1$.

The next assumption deals with the regression function of Y given X. It is standard in regression analysis.

Assumption 4.3.2 (Noise) *There are positive constants $b_Y, b_{Y,\infty}, b_{Y,\psi_2}, b_{Y,2}$ such that $\mathbb{E}Y^2 \le b_Y^2$, $\|\mathbb{E}(Y|X)\|_{L_\infty} \le b_{Y,\infty}$, $\mathbb{E}[(Y - \mathbb{E}(Y|X))^2|X] \le b_{Y,2}^2$ almost surely and $\|Y - \mathbb{E}(Y|X)\|_{\psi_2} \le b_{Y,\psi_2}$.*

In particular, any model $Y = \langle A_0, X\rangle + \varepsilon$ where $\|A_0\|_{S_\infty} < +\infty$ and ε is a centered sub-gaussian noise satisfies Assumption 4.3.2. Note that in the matrix completion problem, if $\sigma^2 = \mathbb{E}(\varepsilon^2)$, the signal-to-noise ratio is given by $\mathbb{E}(\langle X, A_0\rangle^2)/\sigma^2 = \|A_0\|_{S_2}^2 / (\sigma^2 mT)$, so that σ^2 has to scale like $1/(mT)$ for the signal-to-noise ratio to have a reasonable value.

4.3.2 Main results

In this section we state our main results. We give sharp oracle inequalities for the regularized ERM procedure

$$\widehat{A}_n \in \operatorname*{argmin}_{A \in \mathcal{M}_{m,T}} \Big\{ \frac{1}{n} \sum_{i=1}^{n} (Y_i - \langle X_i, A\rangle)^2 + \mathrm{reg}(A) \Big\}, \tag{4.3.4}$$

where $\mathrm{reg}(A)$ is a regularizing function which will be either a pure $\|\cdot\|_{S_1}$ regularization, or a "matrix elastic-net" regularization $\|\cdot\|_{S_1} + \|\cdot\|_{S_2}^2$ or other regularizing functions involving the $\|\cdot\|_1$ norm.

Theorem 4.3.1 (Pure $\|\cdot\|_{S_1}$ regularization, [P6]) *There is an absolute constant $c > 0$ for which the following holds. Let Assumptions 4.3.1 and 4.3.2 hold, and let $x > 0$ be some fixed confidence level. Consider any*

$$\widehat{A}_n \in \operatorname*{argmin}_{A \in \mathcal{M}_{m,T}} \big\{ R_n(A) + \lambda_{n,x} \|A\|_{S_1} \big\},$$

for

$$\lambda_{n,x} = c_{X,Y} \frac{(x + \log n) \log n}{\sqrt{n}},$$

where $c_{X,Y} := c(1 + b_{X,2}^2 + b_Y b_X + b_{Y,\psi_1}^2 + b_{Y,\infty}^2 + b_{Y,2}^2 + b_{X,\infty}^2)$. *Then one has, with a probability larger than* $1 - 5e^{-x}$, *that*

$$R(\widehat{A}_n) \leq \inf_{A \in \mathcal{M}_{m,T}} \big\{ R(A) + \lambda_{n,x}(1 + \|A\|_{S_1}) \big\}.$$

When there is an underlying model, namely if $\mathbb{E}(Y|X) = \langle X, A_0 \rangle$ for some matrix A_0, an immediate corollary of Theorem 4.3.1 is that for any $x > 0$, we have

$$\mathbb{E}\langle X, \widehat{A}_n - A_0 \rangle^2 \leq c_{X,Y} \frac{(x + \log n) \log n}{\sqrt{n}} (1 + \|A_0\|_{S_1})$$

with a probability larger than $1 - 5e^{-x}$. The rate obtained here involves the nuclear norm of A_0 and not the rank. In particular, this rate is not deteriorated if A_0 is of full rank but close to a low rank matrix, and it is also still meaningful when $m + T$ is large compared to n. This is not the case for rates of the form $\mathrm{rank}(A_0)(m + T)/n$, obtained for instance in [55] and [94], which are obtained under stronger assumptions.

Concerning the optimality of Theorem 4.3.1, the following lower bound can be proved by using the classical tools of [106]. Consider the model

$$Y = \big\langle A_0, X \big\rangle + \sigma\zeta \tag{4.3.5}$$

where ζ is a standard Gaussian variable and X is distributed like the $m \times T$ diagonal matrix $\mathrm{diag}[\epsilon_1, \ldots, \epsilon_{m \wedge T}]$ where $\epsilon_1, \ldots, \epsilon_{m \wedge T}$ are i.i.d. Rademacher variables. Then, there exists absolute constants $c_0, c_1 > 0$ such that the following holds. Let $n, m, T \in \mathbb{N} - \{0\}$ and $R > 0$. Assume that $m \wedge T \geq \sqrt{n}$. For any procedure $\widehat{A}$ constructed from n observations in the model (4.3.5) (and denote by $\mathbb{P}_{A_0}^{\otimes n}$ the probability distribution of such a sample), there exists $A_0 \in RB(S_1)$ such that with $\mathbb{P}_{A_0}^{\otimes n}$-probability greater than c_1,

$$R(\widehat{A}) - R(A_0) \geq c_0 \sigma R \sqrt{\frac{1}{n} \log\Big(\frac{c_0 \sigma m \wedge T}{R\sqrt{n}}\Big)}.$$

This shows that, up to some logarithmic factor, the residual term obtained in Theorem 4.3.1 is optimal. The only point is that the S_2 norm of the design in (4.3.5) is not nicely upper bounded ($\|X\|_{S_2} = m \wedge T$ a.s.) as it is required in Assumption 4.3.1. Nevertheless, the assumption $\|X\|_{S_2} \leq b_{X,2}$ a.s. is mostly technical: it comes from the fact that we use the weak inclusion $B(S_1) \subset B(S_2)$ for the computation of the complexity of $B(S_1)$. This inclusion is clearly a source of looseness and we believe that Theorem 4.3.1 is also valid if we only assume that $\|X\|_{S_\infty} \leq b_{X,\infty}$ a.s. in place of $\|X\|_{S_2} \leq b_{X,2}$ a.s..

We now state three sharp oracle inequalities for procedures of the form (4.3.4) where the regularizing function is a mixture of norms.

Theorem 4.3.2 (Matrix Elastic-Net, [P6]) *There is an absolute constant* $c > 0$ *for which the following holds. Let Assumptions 4.3.1 and 4.3.2 hold. Fix any* $x > 0$, $r_1, r_2 > 0$, *and consider*

$$\widehat{A}_n \in \operatorname*{argmin}_{A \in \mathcal{M}_{m,T}} \Big\{ R_n(A) + \lambda_{n,x}(r_1 \|A\|_{S_1} + r_2 \|A\|_{S_2}^2) \Big\},$$

where

$$\lambda_{n,x} = c_{X,Y} \frac{\log n}{\sqrt{n}} \Big(\frac{1}{r_1} + \frac{(x + \log n) \log n}{r_2 \sqrt{n}} \Big),$$

where $c_{X,Y} = c(1 + b_{X,2}^2 + b_{X,2}b_Y + b_{Y,\psi_1}^2 + b_{Y,\infty}^2 + b_{Y,2}^2)$. *Then one has, with a probability larger than* $1 - 5e^{-x}$, *that*

$$R(\widehat{A}_n) \leq \inf_{A \in \mathcal{M}_{m,T}} \left\{ R(A) + \lambda_{n,x}(1 + r_1 \|A\|_{S_1} + r_2 \|A\|_{S_2}^2) \right\}.$$

Theorem 4.3.2 guarantees the performances of the Matrix Elastic-net estimator (mixture of the S_1-norm and the S_2-norm to the square). The use of this algorithm is particularly relevant for matrices with a spectra spread out on the few first singular values, namely for matrices with a singular value decomposition of the form

$$U \operatorname{diag}[a_1, \ldots, a_r, \epsilon_{r+1}, \ldots, \epsilon_{m \wedge T}] V^\top, \tag{4.3.6}$$

where U and V are orthonormal matrices, where $(a_1, \ldots, a_r)$ is well-spread (roughly speaking, the a_i's are of the same order) and where the ε_i are small compared to the a_i.

Theorem 4.3.3 ($\|\cdot\|_{S_1} + \|\cdot\|_1$ regularization, [P6]) *There is an absolute constant* $c > 0$ *for which the following holds. Let Assumptions 4.3.1 and 4.3.2 hold. Fix any* $x, r_1, r_3 > 0$, *and consider*

$$\widehat{A}_n \in \operatorname*{argmin}_{A \in \mathcal{M}_{m,T}} \left\{ R_n(A) + \lambda_{n,x}(r_1 \|A\|_{S_1} + r_3 \|A\|_1) \right\}$$

for

$$\lambda_{n,x} := c_{X,Y} \Big(\frac{1}{r_1} \wedge \frac{\sqrt{\log(mT)}}{r_3} \Big) \frac{(x + \log n)(\log n)^{3/2}}{\sqrt{n}},$$

where $c_{X,Y} = c(1 + b_{X,2}^2 + b_{X,2}b_Y + b_{Y,\psi_1}^2 + b_{Y,\infty}^2 + b_{Y,2}^2 + b_{X,\infty}^2 + b_{X,\ell_\infty}^2)$. *Then one has, with a probability larger than* $1 - 5e^{-x}$, *that*

$$R(\widehat{A}_n) \leq \inf_{A \in \mathcal{M}_{m,T}} \left\{ R(A) + \lambda_{n,x}(1 + r_1 \|A\|_{S_1} + r_3 \|A\|_1)) \right\}.$$

Theorem 4.3.3 guarantees the statistical performances of a mixture of the S_1-norm and the ℓ_1-norm. This mixed regularization shall improve upon the pure S_1 regularization when the underlying matrix contains many zeros. Note that, in the matrix completion case, the term $\sqrt{\log mT}$ can be removed from the regularization (and thus the residual) term thanks to Theorem 1 in [95].

Theorem 4.3.4 ($\|\cdot\|_{S_1} + \|\cdot\|_{S_2}^2 + \|\cdot\|_1$ regularization, [P6]) *There is an absolute constant* $c > 0$ *for which the following holds. Let Assumptions 4.3.1 and 4.3.2 hold. Fix any* $x, r_1, r_2, r_3 > 0$, *and consider*

$$\widehat{A}_n \in \operatorname*{argmin}_{A \in \mathcal{M}_{m,T}} \left\{ R_n(A) + \lambda_{n,x}(r_1 \|A\|_{S_1} + r_2 \|A\|_{S_2}^2 + r_3 \|A\|_1) \right\}$$

for

$$\lambda_{n,x} := c_{X,Y} \frac{(\log n)^{3/2}}{\sqrt{n}} \Big(\frac{1}{r_1} \wedge \frac{\sqrt{\log(mT)}}{r_3} + \frac{x + \log n}{r_2 \sqrt{n}} \Big),$$

where $c_{X,Y} = c(1 + b_{X,2}^2 + b_{X,2}b_Y + b_{Y,\psi_1}^2 + b_{Y,\infty}^2 + b_{Y,2}^2)$. *Then one has, with a probability larger than* $1 - 5e^{-x}$, *that*

$$R(\widehat{A}_n) \leq \inf_{A \in \mathcal{M}_{m,T}} \left\{ R(A) + \lambda_{n,x}(1 + r_1 \|A\|_{S_1} + r_2 \|A\|_{S_2}^2 + r_3 \|A\|_1)) \right\}.$$

All these results follow from Theorem 3.5.3 or [84, 16]. In particular, we can point out that Theorem 3.5.3 can be used to handle very general criterion functions. The parameters r_1, r_2 and r_3 in the above procedures are completely free and can depend on n, m and T. Intuitively, it is clear that r_2 should be smaller than r_1 since the $\|\cdot\|_{S_2}$ term is used for "regularization" of the non-zero singular values only, while the term $\|\cdot\|_{S_1}$ makes $\widehat{A}_n$ of low rank, as for the elastic-net for vectors (see [124]). Indeed, for the $\|\cdot\|_{S_1} + \|\cdot\|_{S_2}^2$ regularization, any choice of r_1 and r_2 such that $r_2 = r_1 \log n/\sqrt{n}$ leads to a residual term smaller than

$$c_{X,Y}(1 + x + \log n)\Big(\frac{(\log n)^2}{r_2 n} + \frac{\log n}{\sqrt{n}}\|A\|_{S_1} + \frac{(\log n)^2}{n}\|A\|_{S_2}^2\Big).$$

Note that the rate related to $\|A\|_{S_1}$ is (up to logarithms) $1/\sqrt{n}$ while the rate related to $\|A\|_{S_2}^2$ is $1/n$. The choice of r_3 depends on the number of zeros in the matrix. Note that in the $\|\cdot\|_{S_1} + \|\cdot\|_1$ case, any choice $1 \leq r_3 \leq r_1$ entails a residue smaller than

$$c_{X,Y}\frac{(x + \log n)\log n}{\sqrt{n}}(1 + \|A\|_{S_1} + \|A\|_1),$$

which makes again the residue independent of m and T.

It is interesting to point that regularized procedures that involve the 1-Schatten norm (and also for regularizations involving other norms), the residue does not depend on m and T directly: it only depends on the 1-Schatten norm of A_0. This fact points out an interesting difference between nuclear-norm regularization (also called "Matrix Lasso") and the Lasso for vectors. In [94], upper bounds for p-Schatten regularization procedures for $0 < p \leq 1$ are given in the same setting as ours, including in particular the matrix completion problem. The results are stated without the incoherency assumption for matrix completion. But for this problem, the upper bounds are given using the empirical norm $\big\|\widehat{A}_n - A_0\big\|_n^2 = \sum_{i=1}^n \big\langle X_i, \widehat{A}_n - A_0\big\rangle^2/n$ only. An upper bound using this empirical norm gives information only about the denoising part and not about the generalizing/filling part of the matrix completion problem. Estimation results follow from our exact oracle inequalities: when $\mathbb{E}(Y|X) = \langle X, A_0\rangle$ for some $A_0 \in \mathcal{M}_{m,T}$ our estimators $\widehat{A}_n$ satisfy

$$\mathbb{E}\big\langle X, \widehat{A}_n - A_0\big\rangle^2 \leq \inf_{A \in \mathcal{M}_{m,T}} \big\{\mathbb{E}\big\langle X, A - A_0\big\rangle^2 + r_n(A)\big\},$$

and taking A_0 in the infimum leads to the upper bound

$$\mathbb{E}\big\langle X, \widehat{A}_n - A_0\big\rangle^2 \leq r_n(A_0).$$

Note that $\mathbb{E}\big\langle X, \widehat{A}_n - A_0\big\rangle^2 = \big\|\widehat{A}_n - A_0\big\|_{S_2}^2/(mT)$ in the uniform matrix completion problem (see Example 4.3.1 below).

4.4 Selection of variables and dimension reduction in high-dimensional non-parametric regression

We consider the non-parametric Gaussian regression model

$$Y_i = f(X_i) + e_i, \quad i = 1, \ldots, n,$$

where the design variables (or input variables) $X_1, \ldots, X_n$ are n i.i.d. random variables with values in $\mathbb{R}^d$, the noise $e_1, \ldots, e_n$ are n i.i.d. Gaussian random variables with variance σ^2 independent of the X_i's and f is the unknown regression function. In this section, we are interested in the pointwise estimation of f at a fixed point $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$. We want to construct some estimation procedures $\widehat{f}_n$ having the smallest pointwise integrated quadratic risk

$$\mathbb{E}(\widehat{f}_n(x) - f(x))^2 \tag{4.4.1}$$

using only the set of data $D_n = (Y_i, X_i)_{1 \leq i \leq n}$.

We assume f to be β-Hölderian around x. We recall that a function $f : \mathbb{R}^d \longmapsto \mathbb{R}$ is *β-Hölderian at the point x* with $\beta > 0$, denoted by $f \in \Sigma(\beta, x)$, when the two following points hold:

- f is l-times differentiable in x (where $l = \lfloor \beta \rfloor$ is the largest integer which is strictly smaller than β),
- there exists $L > 0$ such that for any $t = (t_1, \ldots, t_n) \in B_\infty^d(x, 1)$,

$$|f(t) - P_l(f)(t, x)| \leq L \|t - x\|_1^\beta,$$

where $P_l(f)(\cdot, x)$ is the Taylor polynomial of order l associated with f at the point x, $\|\cdot\|_1$ is the l_1 norm and $B_\infty^d(x, 1)$ is the unit l_∞-ball of center x and radius 1.

When f is only assumed to be in $\Sigma(\beta, x)$, no estimator can converge to f (for the pointwise risk given in equation (4.4.1)) faster than

$$n^{-2\beta/(2\beta+d)}. \tag{4.4.2}$$

This rate can be very slow when the dimension d of the input variable X is large compared to the regularity β of the regression function f. In many practical problems, the dimension d can depend on the number n of observations in such a way that the rate (4.4.2) does not even tend to zero when n tends to infinity. This phenomenon was called the "curse of dimensionality" by R.Bellman (cf. [115] for some discussion on this phenomenon). Fortunately, in some of these problems the regression function really depends only on a few number of coordinates of the input variables. We formulate this heuristic by the following assumption:

Assumption 4.4.1 *There exist an integer $d^* \leq d$, a function $g : \mathbb{R}^{d^*} \to \mathbb{R}$ and a subset $J = \{i_1, \ldots, i_{d^*}\} \subset \{1, \ldots, d\}$ of cardinality d^* such that for any $(x_1, \ldots, x_d) \in \mathbb{R}^d$*

$$f(x_1, \ldots, x_d) = g(x_{i_1}, \ldots, x_{i_{d^*}}).$$

Under Assumption 4.4.1, the "real" dimension of the problem is not anymore d but d^*. Then, we hope that if $f \in \Sigma(\beta, x)$ (which is equivalent to say that g is β-Hölderian at the point x), it would be possible to estimate $f(x)$ at the rate given in equation (4.4.2) where d is replaced by d^*, leading to a real improvement of the convergence rate when $d^* \ll d$. Nevertheless, starting from the data D_n, it is not clear that detecting the set J of interesting coordinates is an easy task. To select this set, we use a l_1 regularization technique. This technique has been mostly used in the parametric setup. We adapt it to the non-parametric setup and we obtain the following informal selection result which is a short version of Theorem 4.4.2 below.

Theorem A (selection of the subset J) [P1] *Under Assumption 4.4.1 it is possible to construct, only from the data D_n, a subset $\widehat{J} \subset \{1, \ldots, d\}$ such that, with probability greater than $1 - c_0 \exp(c_0 d - c_1 n h^{d+2})$ (for a free parameter $0 < h < 1$), $\widehat{J} = J$.*

Once the set J is empirically determined with high probability, we then run a classical local polynomial estimation procedure on the set of indices $\widehat{J}$ to obtain the following informal estimation result which is a short version of Theorem 4.4.3 below.

Theorem B (estimation of f) [P1] *For any $f \in \Sigma(\beta, x)$, with $\beta > 1$, satisfying Assumption 4.4.1, it is possible to construct, only from the data D_n, an estimation procedure $\widehat{f}_n$ such that $\mathbb{P}[|\widehat{f}_n(x) - f(x)| \geq \delta] \leq c \exp(-c\delta^2 n^{2\beta/(2\beta+d^*)}), \forall \delta > 0$ where c does not depend on n.*

The last theorem proves that it is possible, only from the set of data, to reduce and to detect the "real" dimension of the problem under Assumption 4.4.1.

The problem we consider in the section is called a high-dimensional problem. In the last years, many papers have studied these kinds of problems and summarizing here the state of the art is not possible (we refer the reader to the bibliography of [65]). We just mention some papers. In [18, 72, 17, 44], it is assumed that the design variable X belongs to a low dimensional smooth manifold of dimension $d^* < d$. All of these work are based on heuristics techniques. In [65], the same problem as the one considered here is handled. Their strategy is a greedy method that incrementally searches through bandwidth in small steps. If the regression f is in a Sobolev ball of order 2, their procedure is nearly optimal for the pointwise estimation of f in x. It achieves the convergence rate $n^{-4/(4+d^*+\epsilon)}$ for every $\epsilon > 0$, when $d = \mathcal{O}(\log n / \log\log n)$ and $d^* = \mathcal{O}(1)$. Our procedure improves this result. First, the optimal rate of convergence is achieved. Second, the regression function does not have to be twice differentiable (actually Theorem B holds for any smoothness parameter $\beta > 1$). Third, the dimension d can be taken of the order of $\log n$.

Our goal is twofold. First, we want to determine the set of indices $J = \{i_1, \ldots, i_{d^*}\}$. Second, we want to construct an estimator of the value $f(x)$ that converges to the rate $n^{-2\beta/(2\beta+d^*)}$ when $f \in \Sigma(\beta, x)$ for $\beta > 1$. To achieve the first goal, we use a l_1 regularization of local polynomial estimators.

4.4.1 Selection Procedure

We consider the following set of vectors

$$\bar{\Theta}(\lambda) = \underset{\theta \in \mathbb{R}^{d+1}}{\operatorname{argmin}} \left[\frac{1}{nh^d} \sum_{i=1}^{n} \left(Y_i - U\left(\frac{X_i - x}{h}\right)\theta \right)^2 K\left(\frac{X_i - x}{h}\right) + 2\lambda \|\theta\|_1 \right], \tag{4.4.3}$$

where $U(v) = (1, v_1, \ldots, v_d)$ for any $v = (v_1, \ldots, v_d)^\top \in \mathbb{R}^d$, $\|\theta\|_1 = \sum_{j=0}^{d} |\theta_j|$ for any $\theta = (\theta_0, \ldots, \theta_d)^\top \in \mathbb{R}^{d+1}$, $h > 0$ is called the *bandwidth*, $\lambda > 0$ is called the *regularization parameter* and $K : \mathbb{R}^d \longrightarrow \mathbb{R}$ is called the *kernel*. We will explain how to choose the parameters h and λ in what follows. In the following, we denote $U_0(v) = 1$ and $U_i(v) = v_i$, for $i = 1, \ldots, d$ for any $v = (v_1, \ldots, v_d) \in \mathbb{R}^d$. The kernel K is taken such that the following set of assumptions holds:

Assumption 4.4.2 *The kernel $K : \mathbb{R}^d \longrightarrow \mathbb{R}$ is symmetric, supported in $B_\infty^d(0,1)$, the matrix $(\int_{\mathbb{R}^d} K(y) U_i(y) U_j(y) dy)_{i,j \in \{0,\ldots,d\}}$ is diagonal with positive coefficients independent of d in the diagonal and there exists a constant $M_K \geq 1$ independent of d which upper bounds the quantities $\max_{u \in \mathbb{R}^d} |K(u)|$, $\max_{u \in \mathbb{R}^d} K(u)^2$, $\max_{u \in \mathbb{R}^d} |K(u)| \|u\|_1^2$, $\max_{u \in \mathbb{R}^d} |K(u)| \|u\|_2^2$, $\int_{\mathbb{R}^d} K(y)^2 (1 + \|y\|_2^2) dy$, $\int_{\mathbb{R}^d} |K(u)|^2 \|u\|_1^4 du$ and $\int_{\mathbb{R}^d} K(y)^2 (U_i(y) U_j(y))^2 dy$.*

Note that for example the uniform kernel $K(u) = \frac{1}{2^d} \mathbb{1}_{B_\infty^d(0,1)}(u)$ satisfies Assumption 4.4.2.

Any statistic $\bar{\theta} \in \bar{\Theta}(\lambda)$ is a l_1 regularized version of the classical local polynomial estimator. Usually, for the estimation problem of $f(x)$, only the first coordinate of $\bar{\theta}$ is used. Here, for the

selection problem, we use all the coordinates except the first one. We denote by $\widehat{\theta}$ the vector of $\mathbb{R}^d$ made of the d last coordinates of $\bar{\theta}$.

We expect the vector $\widehat{\theta}$ to be sparse (that is with many zero coordinates) such that the set of all the non-zero coordinates of $\widehat{\theta}$, denoted by $\widehat{J}$, will be the same as the set J of all the non-zero coordinates of $(\theta_1^*, \dots, \theta_d^*)^\top$ where $\theta_i^* = h\partial_i f(x)$, for $i \in \{1, \dots, d\}$, and $\partial_i f(x)$ stands for the $i-$th derivative of f at point x. We remark that, under Assumption 4.4.1, the vector $(\theta_1^*, \dots, \theta_d^*)^\top$ is sparse.

Note that, the estimator $\bar{\theta} \in \bar{\Theta}(\lambda)$ may not be unique (depending on d and n). Hence, the subset selection method may provide different subsets $\widehat{J}$ depending on the choice of $\bar{\theta}$. Nevertheless, Theorem 4.4.3 holds for any subset $\widehat{J}$, whatever is the vector $\bar{\theta}$ chosen in $\bar{\Theta}(\lambda)$.

We also consider another selection procedure close to the previous one which requires less assumption on the regression function. We just need to assume that there exists $f_{\max} > 0$ such that $|f(x)| \leq f_{\max}$. With the same notation, we consider the following set of vectors

$$\bar{\Theta}_2(\lambda) = \underset{\theta \in \mathbb{R}^{d+1}}{\operatorname{argmin}} \left[\frac{1}{nh^d} \sum_{i=1}^{n} \left(Y_i + f_{max} + Ch - U\left(\frac{X_i - x}{h}\right)\theta \right)^2 K\left(\frac{X_i - x}{h}\right) + 2\lambda \|\theta\|_1 \right], \tag{4.4.4}$$

where C and h are defined later. We just translate the outputs Y_i's by $f_{max} + Ch$. This translation affects the estimator since the LASSO method is not a linear procedure. We denote by $\widehat{J}_2$, this subset selection procedure.

Remark 4.4.1 *The l_1 penalization technique can be related to the problem of linear aggregation (cf. [105] and [89]) in a sparse setup. Indeed, l_1 penalization is known to provide sparse estimators if the underlying object to estimate is sparse with respect to a given dictionary. Assumption 4.4.1 can be interpreted in terms of sparsity of f w.r.t. to a certain dictionary. For that, we consider the set $\mathcal{F} = \{f_0, f_1, \dots, f_d\}$ of functions from $\mathbb{R}^d$ to $\mathbb{R}$ where $f_0 = \mathbb{1}$ is the constant function equals to 1 and $f_j(t) = (t_j - x_j)/h$ for any $j \in \{1, \dots, d\}$ and $t = (t_1, \dots, t_d) \in \mathbb{R}^d$. The set $\mathcal{F}$ is the dictionary. That is the set within we are looking for the best sparse linear combination of elements in $\mathcal{F}$ approaching f in a neighborhood of x. In this setup, the Taylor polynomial of order 1 at point x, denoted by $P_1(f)(\cdot, x)$, is a linear combination of the elements in the dictionary $\mathcal{F}$. When f is assumed to belong to $\Sigma(\beta, x)$, the polynomial $P_1(f)(\cdot, x)$ is a good approximation of f in a neighborhood of x. Moreover, under Assumption 4.4.1, this linear combination is sparse w.r.t. the dictionary $\mathcal{F}$. Thus, we hope that, with high probability, minimizing a localized version of the empirical L_2-risk penalized by the l_1 norm over the set of all the linear combinations of elements in $\mathcal{F}$ will detect the right locations of the interesting indices $i_1, \dots, i_{d^*}$ (which correspond to the non-zero coefficients of $P_1(f)(\cdot, x)$ in the dictionary $\mathcal{F}$). That is the main idea behind the procedures introduced in this section since we have:*

$$\bar{\Theta}(\lambda) = \underset{\theta \in \mathbb{R}^{d+1}}{\operatorname{argmin}} \left[\frac{1}{nh^d} \sum_{i=1}^{n} \left(Y_i - \sum_{j=0}^{d} f_j(X_i)\theta_j \right)^2 K\left(\frac{X_i - x}{h}\right) + 2\lambda \|\theta\|_1 \right],$$

Of course, we can generalize this approach to other dictionaries (this will lead to other sparsity and regularity properties of f) provided that some kind of "orthogonality properties" of $\mathcal{F}$ still holds.

4.4.2 Estimation Procedure

We now construct a classical local polynomial estimator (LPE) (cf. [64, 104]) on the set of coordinates $\widehat{J}_2$ previously constructed.

We assume that the selection step is now done. We have at hand a subset $\widehat{J}_2 = \{\widehat{\imath}_1, \dots, \widehat{\imath}_{\widehat{d^*}}\} \subset \{1, \dots, d\}$ of cardinality $\widehat{d^*}$. For the second step, we consider γ_x a polynomial on $\mathbb{R}^{\widehat{d^*}}$ of degree $l = \lfloor \beta \rfloor$ which minimizes

$$\sum_{i=1}^{n} (Y_i - \gamma_x(p(X_i - x)))^2 K^\star\left(p\Big(\frac{X_i - x}{h^\star}\Big)\right)$$

where $h^\star = n^{-1/(2\beta + \widehat{d^*})}$, $p(v) = (v_{\widehat{\imath}_1}, \dots, v_{\widehat{\imath}_{\widehat{d^*}}})^\top$ for any $v = (v_1, \dots, v_d)^\top \in \mathbb{R}^d$ and $K^\star : \mathbb{R}^{\widehat{d^*}} \longrightarrow \mathbb{R}$ is a kernel function. The local polynomial estimator of f at the point x is $\widehat{\gamma}_x(0)$ if $\widehat{\gamma}_x$ is unique and 0 otherwise. We denote by $\widehat{f}(x)$ the projection onto $[-f_{max}, f_{max}]$ of the LPE of $f(x)$. Here, we don't use the other coefficients of $\widehat{\gamma}_x(0)$ like we did in the selection step.

For the estimation step, we use a result on the convergence of multivariate LPE from [10]. We recall here the properties of the kernel required in [10] to obtain this result.

Assumption 4.4.3 *The kernel $K^\star : \mathbb{R}^{\widehat{d^*}} \longrightarrow \mathbb{R}$ is such that: there exists $c > 0$ satisfying*

$$K^\star(u) \geq c \mathbb{1}_{\|x\|_2 \leq c}, \forall u \in \mathbb{R}^{\widehat{d^*}}; \int_{\mathbb{R}^{\widehat{d^*}}} K^\star(u) du = 1;$$

$$\int_{\mathbb{R}^{\widehat{d^*}}} (1 + \|u\|_2^{4\beta})(K^\star(u))^2 du < \infty; \sup_{u \in \mathbb{R}^{\widehat{d^*}}} (1 + \|u\|_2^{2\beta}) K^\star(u) < \infty.$$

4.4.3 A selection and estimation theorem

In this subsection, we provide the main results of this section. To avoid any technical complexity we will assume that the density function μ of the design X satisfies the following assumption:

Assumption 4.4.4 *There exists some constants η, $\mu_m > 0$, $\mu_M \geq 1$ and $L_\mu > 0$ such that*

- *$B_\infty^d(x, \eta) \subset \mathrm{supp}(\mu)$ and $\mu_m \leq \mu(y) \leq \mu_M$ for almost every $y \in B_\infty^d(x, \eta)$,*
- *μ is L_μ-Lipschitz around x, that is for any $t \in B_\infty^d(x, 1)$, $|\mu(x) - \mu(t)| \leq L_\mu \|x - t\|_\infty$ (remark that the value $\mu(x)$ is the value of the continuous version of μ around x).*

The first result deals with the statistical properties of the selection procedure. For this step, we require a weaker regularity assumption for the regression function f. This assumption is satisfied for any β-Hölderian function in x with $\beta > 1$.

Assumption 4.4.5 *There exists an absolute constant $L > 0$ such that the following holds. The regression function f is differentiable and*

$$|f(t) - P_1(f)(t, x)| \leq L \|t - x\|_1^\beta, \quad \forall t \in B_\infty^d(x, 1),$$

where $P_1(f)(\cdot, x)$ is the Taylor polynomial of degree 1 of f at the point x.

To achieve an efficient selection of the interesting coordinates, we have to be able to distinguish the non-zero partial derivatives of f from the null partial derivatives. For that, we consider the following assumption:

Assumption 4.4.6 *There exists a constant $C \geq 72(\mu_M/\mu_m) L M_K \sqrt{d_0}$ such that $|\partial_j f(x)| \geq C$ for any $j \in J$, where the set J is given in Assumption 4.4.1 and d_0 is an integer such that $d^* \leq d_0$.*

Theorem 4.4.2 ([P1]) *There exists some constants $c_0 > 0$ and $c_1 > 0$ depending only on L_μ, μ_m, μ_M, M_K, L, C and σ for which the following holds. We assume that the regression function f satisfies the regularity Assumption 4.4.5, the sparsity Assumption 4.4.1 such that the integer d^* is smaller than a known integer d_0 and the distinguishable Assumption 4.4.6. We assume that a density function μ of the input variable X satisfies Assumption 4.4.4.*

We consider $\bar{\theta} = (\bar{\theta}_0, \dots, \bar{\theta}_d) \in \bar{\Theta}(\lambda) \subset \mathbb{R}^{d+1}$ and $\bar{\theta}_2 = ((\bar{\theta}_2)_0, \dots, (\bar{\theta}_2)_d) \in \bar{\Theta}_2(\lambda) \subset \mathbb{R}^{d+1}$ where $\bar{\Theta}(\lambda)$ and $\bar{\Theta}_2(\lambda)$ are defined in equations (4.4.3) and (4.4.4) with a kernel satisfying Assumption 4.4.2, a bandwidth and a regularization parameter such that

$$0 < h < \frac{\mu_m}{32(d_0+1)L_\mu M_K} \wedge \eta \text{ and } \lambda = 8\sqrt{3M_K}\mu_M L h \tag{4.4.5}$$

We denote by $\widehat{J}$ the set $\{j \in \{1, \dots, d\} : \bar{\theta}_j \neq 0\}$ and by $\widehat{J}_2$ the set $\{j \in \{1, \dots, d\} : (\bar{\theta}_2)_j \neq 0\}$.

- *If $|f(x)| > Ch$, where C is defined in Assumption 4.4.6 or $f(x) = 0$, then with probability greater than $1 - c_1 \exp(c_1 d - c_0 n h^{d+2})$, $\widehat{J} = J$.*
- *If $|f(x)| \leq f_{max}$, then with probability greater than $1 - c_1 \exp(c_1 d - c_0 n h^{d+2})$, $\widehat{J}_2 = J$.*

We remark that Theorem 4.4.2 still holds when we only assume that there exists a subset $J \subset \{1, \dots, d\}$ such that $\partial_j f(x) = 0$ for any $j \notin J$ instead of the more global Assumption 4.4.1.

Theorem 4.4.3 ([P1]) *We assume that the regression function f belongs to the Hölder class $\Sigma(\beta, x)$ with $\beta > 1$ and satisfies the sparsity Assumption 4.4.1 such that the integer d^* is smaller than a known integer d_0 and the distinguishable Assumption 4.4.6. We assume that a density function μ of the input variable X satisfies Assumption 4.4.4 and $|f(x)| \leq f_{max}$. We assume that the dimension d is such that $d + 2 \leq (\log n)/(-2 \log h)$ (h satisfies (4.4.5)).*

We construct the set $\widehat{J}_2$ of selected coordinates with a kernel, a bandwidth and a regularization parameter as in Theorem 4.4.2. The LPE estimator $\widehat{f}(x)$ constructed in subsection 4.4.2 on the subset $\widehat{J}_2$ and a kernel $K^\star$ satisfying Assumption 4.4.3, satisfies

$$\forall \delta > 0, \mathbb{P}[|\widehat{f}(x) - f(x)| \geq \delta] \leq c_1 \exp\Big(- c_2 n^{\frac{2\beta}{2\beta + d^*}} \delta^2 \Big),$$

where $c_1, c_2 > 0$ are constants independent of n, d, d^*.

Note that, by taking the expectation, we obtain $\mathbb{E}[(\widehat{f}(x) - f(x))^2] \leq c n^{\frac{-2\beta}{2\beta+d^*}}$. The selection procedure is efficient provided that $c_1 n h^{d+2} - c_0 d$ tends to infinity when n tends to infinity. Namely, we need (with $0 < h < 1$)

$$d + 2 < \frac{\log n}{-\log h}. \tag{4.4.6}$$

It is interesting to note that, for d of the order of $\log n$ (like in (4.4.6)), the rate of convergence in (4.4.2) does not tend to zero. Therefore, in this case and without any previous selection step, a classical LPE can fail to estimate $f(x)$.

A remarkable point of Theorem 4.4.2 is that the bandwidth h does not have to tend to 0 when n tends to infinity. This particular behavior does not appear when LPE are used for estimation and not for selection. This can be explained because, we do not need to control any bias term in the selection step. In the selection context, the restriction on h comes only from the fact that we need the dictionary $\mathcal{F}$ to be approximatively orthogonal.

Finally, once the set of interesting coordinates is selected, we can use it to run other non-parametric methods to estimate the function f with other pointwise risks or integrated risks

and under other smoothness assumptions on f. Note that, by considering other order of the l_1-penalized LPE in the selection step, it is easy to find other properties of the function f. For instance, inflection points or convexity of f can be detected with a second order method in the selection step.

4.5 Non-exact oracle inequalities for the Convex aggregation problem

The problem of Convex aggregation is the following: take a finite model $F = \{f_1, \dots, f_M\}$ for some $M \in \mathbb{N}$ and try to find a procedure that is "as good as" the best convex combination of elements in F. To define what is meant by "as good as", we introduce some notation.

For any $\lambda = (\lambda_1, \dots, \lambda_M)^\top \in \mathbb{R}^M$, we define $f_\lambda = \sum_{j=1}^M \lambda_j f_j$ and the convex hull of F is the set $\mathrm{conv}(F) = \left\{ f_\lambda : \sum_{j=1}^M \lambda_j = 1 \text{ and } \lambda_j \geq 0 \right\}$. There are many different ways of defining the Convex aggregation problem. The one that we will be interested in is the following: for some $0 \leq \epsilon \leq 1$ construct a procedure $\tilde{f}_n$ such that, for any $x > 0$ with probability larger than $1 - \exp(-x)$,

$$R(\tilde{f}_n) \leq (1+\epsilon) \inf_{f \in \mathrm{conv}(F)} R(f) + c_0 \max\left(r_n(M), \frac{x}{n} \right) \tag{4.5.1}$$

where the residual term $r_n(M)$ should be as small as possible. From both mathematical and statistical point of view, the most interesting case to study is for $\epsilon = 0$. In this case, it follows from classical minimax result that no algorithm can do better than the rate

$$\psi_n^C(M) = \begin{cases} M/n & \text{if } M \leq \sqrt{n}, \\ \sqrt{\frac{\log\left(eM/\sqrt{n}\right)}{n}} & \text{otherwise.} \end{cases} \tag{4.5.2}$$

It is shown in [103] that there is indeed a procedure $\tilde{f}_n$ achieving this rate in expectation: $\mathbb{E}R(\tilde{f}_n) \leq \inf_{f \in \mathrm{conv}(F)} R(f) + \psi_n^C(M)$ and in Theorem 2.8.1, this rate is achieved in deviation.

In this setup, we apply Theorem 3.3.1 to obtain inequalities like (4.5.1) with $0 < \epsilon < 1$ for the ERM over $\mathrm{conv}(F)$:

$$\tilde{f}^{ERM-C} \in \operatorname*{argmin}_{f \in \mathrm{conv}(F)} R_n(f). \tag{4.5.3}$$

To make the argument simple, we consider the bounded regression framework with respect to the quadratic loss: $|Y|, \sup_{f \in F} |f(X)| \leq 1$ a.s. and $\ell_f(x,y) = (y - f(x))^2, \forall f \in F, \forall (x,y) \in \mathcal{X} \times \mathbb{R}$.

Theorem 4.5.1 ([P16]) *There exists an absolute constant c_0 such that the following holds. For any $0 < x < \log n$, with probability greater than $1 - 8\exp(-x)$,*

$$R(\tilde{f}^{ERM-C}) \leq (1+2\epsilon) \inf_{f \in \mathrm{conv}(F)} R(f) + \frac{c_0 (\log M)(\log n)}{n}.$$

It is interesting to note that the proofs of Theorem 4.5.1 and Theorem 4.1.1 are closely related. In the case of Theorem 4.1.1, the result follows from the analysis of the loss functions classes indexed for the family of models $(rB_1^d)_{r \geq 0}$. In the case of Theorem 4.5.1, the result follows from the analysis of the loss function class indexed by the model $\Lambda = \{\lambda \in \mathbb{R}^M : \lambda_j \geq 0, \|\lambda\|_{\ell_1} = 1\}$ (we identify every function $f_\lambda \in \mathrm{conv}(F)$ to its parameter $\lambda \in \Lambda$) which is included in B_1^M. Therefore, the proof of Theorem 4.5.1 follows the same path as the proof of

Theorem 4.1.1 but since we assume boundedness some extra logarithms appearing in Theorem 4.1.1 can be saved here. Indeed, for the loss functions class $\ell_{\mathrm{conv}(F)} = \{\ell_f : f \in \mathrm{conv}(F)\}$, we can apply Theorem 3.3.1 with $b_n(\ell_{\mathrm{conv}(F)}) = 1$, $B_n = 1$ and take for isomorphic function $\rho_n(x) \equiv c_0(\log M)(\log n)/n$.

Note that the residual term of the non-exact oracle inequality satisfied by $\tilde{f}^{ERM-C}$ in Theorem 4.5.1 is of the order of $(\log M)(\log n)/n$ and is thus uniformly better than the optimal rate $\psi_n^C(M)$ for exact oracle inequalities in this setup. Up to logarithms, this residual term can even be the *square* of $\psi_n^C(M)$ when $M \geq \sqrt{n}$.

For this example, the gap between the rates obtained for non-exact oracle inequalities and exact oracle inequalities is not anymore due to the Bernstein condition since in both situations this condition holds: for any $f \in \mathrm{conv}(F)$,

$$\mathbb{E}\ell_f^2 \leq \mathbb{E}\ell_f \text{ and } \mathbb{E}\mathcal{L}_f^2 \leq B\mathbb{E}\mathcal{L}_f.$$

The Bernstein condition for the excess loss functions class $\mathcal{L}_{\mathrm{conv}(F)}$ holds in this context because of the convexity of the set $\mathrm{conv}(F)$ (cf. Proposition 1.3.3). In other words, the geometry of the Convex aggregation problem is good. Thus to explain this gap we have to look somewhere else. It appears that, in this case, the gap comes from complexity. Indeed, it follows from Proposition 3.2.7 that (up to some logarithmic factors), for any $\lambda > 0$,

$$\mathbb{E}\,\|P - P_n\|_{V(\ell_{\mathrm{conv}(F)})_\lambda} \lesssim \sqrt{\frac{\lambda}{n}}.$$

Therefore, the fixed point λ_η^* defined in Theorem 3.3.1 and associated with the loss functions class $\ell_{\mathrm{conv}(F)}$ will be of the order of $1/n$ (up to some logarithmic factors). Whereas, for the particular case $\mathrm{conv}\big(\{f(X) : f \in F\}\big) = \{\sum_{j=1}^M \lambda_j \epsilon_j : \lambda \in B_1^M\}$ and $Y = \epsilon_{M+1}$, where $\epsilon_1, \ldots, \epsilon_{M+1}$ are i.i.d. Rademacher variables, we can prove that for any $\mu \geq 1/M$,

$$\mathbb{E}\,\|P - P_n\|_{V(\mathcal{L}_{\mathrm{conv}(F)})_\mu} \gtrsim \frac{1}{\sqrt{n}}.$$

Hence, the fixed point μ_η^* of Theorem 3.3.2 associated with the excess loss class $\mathcal{L}_{\mathrm{conv}(F)}$ will be of the order of $1/\sqrt{n}$ when $M \geq \sqrt{n}$. To conclude, the problem of Convex aggregation is an instructive case showing that the complexity aspect of the problem plays also a key role in understanding the difference between exact and non-exact oracle inequalities. Unlike the example of (MS) aggregation for which the geometrical aspect was more underlined through the Bernstein condition.

The result of Theorem 4.5.1 rises two questions in Convex aggregation theory: What is the optimal rate of aggregation for the Convex aggregation problem for non-exact oracle inequalities? Is the ERM an optimal procedure for this problem? Furthermore, for the Model Selection aggregation problem (where one wants to be as good as the best in F), it is known that non-exact oracle inequalities and exact oracle inequalities share the same optimal rate of aggregation $(\log M)/n$. Therefore, a natural question would be to characterize aggregation problems (we can extend the definition of aggregation problem to any subsets $A \subset \mathbb{R}^M$: $A = \{e_1, \ldots, e_M\}$ for Model Selection aggregation, $A = \Lambda$ for the Convex aggregation, $A = \mathbb{R}^M$ for Linear aggregation, etc.) for which there is indeed a gap between the residual terms for non-exact and exact oracle inequalities (up to some absolute multiplying constants). This question will be analyzed in Chapter 5.

4.6 Oracle inequalities for cross-validation type procedures

In this section, we construct adaptation procedures inspired by cross-validation. Adaptation procedures are of particular interest when one wants to adapt to an unknown parameter. Such a parameter can appear in statistical procedures for two reasons: either it is an unknown parameter of the model (complexity parameter, "concentration" parameter, geometric parameter, variance of the noise, etc.), or the construction of the procedure requires fitting a parameter that no theory is able to determine (regularization parameter, smoothing parameter, threshold, etc.). Thus it is very useful to have at hand some statistical procedures which can choose these unknown parameters in a data-dependent way. The construction of adaptation procedures has been one of the main topics in non-parametric statistics for the two last decades. Retracing the entire bibliography here is not possible. Nevertheless, we would like to refer the reader to some classical — and now pioneering — steps in this field like the Model Selection approach (cf. i.e. [12] and [76]), aggregation methods (cf. i.e. [89], [39] and [40]), empirical risk minimization (cf. i.e. [116], [57] and [15]), Lepskii's adaptation method in [70, 69] or wavelet thresholding methods in [45]. Of course many other approaches in some particular setups have been developed. But one of the most popular and universal strategy used for fitting unknown parameters or more generally to select algorithms is the Cross-Validation (CV). Cross-validation is a very important and widely applied family of model/ estimator/ parameter selection methods. Among other, the CV procedure was studied for the selection of the bandwidth in kernel density estimation in [52] and [98], for the regression model in [99], in classification in [43]. Many other authors have been studying or using this method and we refer the reader to the survey of CV methods in Model Selection [4], the PhD thesis [41], [96] or [112, 111] for more bibliographical references on this topic. The aim of this section is to present and to study three procedures inspired by the CV procedure in the general framework introduced in Section 1.1.1.

We say that a statistic is a sequence of functions $\widehat{f} = (\widehat{f}^{(n)})_{n\in\mathbb{N}}$ such that each $\widehat{f}^{(n)}$ is a map associating a function $\widehat{f}^{(n)}(\cdot) = \widehat{f}^{(n)}(D^{(n)})(\cdot)$ in $\mathcal{F}$ to each data set $D^{(n)} = \{Z_1, \dots, Z_n\}$. If $\widehat{f}$ is a statistic, the risk of $\widehat{f}$ is given for each n by

$$R(\widehat{f}^{(n)}(D^{(n)})) = \mathbb{E}[\ell(\widehat{f}(D^{(n)}), Z)|D^{(n)}].$$

We assume that we know how to construct some statistics $\widehat{f}_\lambda$ for λ in a set of indexes Λ. We want to construct procedures $\bar{f} = (\bar{f}^{(n)})_{n\in\mathbb{N}}$ satisfying oracle inequalities that is, for any sample size n,

$$\mathbb{E}[R(\bar{f}^{(n)}(D^{(n)})) - R^*] \leq C \inf_{\lambda\in\Lambda} \mathbb{E}[R(\widehat{f}_\lambda^{(n)}(D^{(n)})) - R^*] + r(n, \Lambda) \tag{4.6.1}$$

where $C \geq 1$ is a constant and $r(n, \Lambda)$ is a residue term which we would like to keep as small as possible. Controlling this residue will depend on some complexity parameter of the excess loss functions class $\{\ell(\widehat{f}_\lambda^{(n)}(D^{(n)}), \cdot) - \ell(f^*, \cdot) : \lambda \in \Lambda\}$, as well as on a Margin parameter that limits the behaviour of the contrast function around the risk minimizer (cf. Assumptions (A) below).

We introduce now different adaptation procedures: two modified versions of the cross-validation procedure and then the cross-validation procedure itself.

4.6.1 Classical Cross-validation procedures

The key feature of the CV procedure, the use of multiple splits to train and test the candidate estimator, renders it somewhat more difficult to handle in a theoretial way. Nevertheless, we shall show that a carefully crafted risk inequality opens the door to oracle inequalities for cross-validation too. In this section, we have to pay careful attention to the exact choice of the splits of our data, especially when retraining the selected model to obtain our final estimator(s).

First we shall introduce some notation. Let n be an integer, and V a divisor of n. We split the data set $D^{(n)}$ into V disjoint subsets of equal size $n_C = n/V$, namely, for every $k = 1, \ldots, V$,

$$B_k = \{Z_{(k-1)n_C+1}, \ldots, Z_{kn_C}\}, \tag{4.6.2}$$

which shall be test sets, and their complements

$$D_k = \cup_{j=1:j\neq k}^{V} B_j, \tag{4.6.3}$$

the corresponding training sets. Note that D_k is a data set of size $n_V = n - n_C$.

Let $\ell(f, Z)$ be a contrast function whose arguments are a data point Z and a parameter $f \in \mathcal{F}$. For a statistic $\widehat{f} = (\widehat{f}^{(n)})_n$, we define the *V-fold CV empirical risk* by

$$R_{n,V}(\widehat{f}) = \frac{1}{V}\sum_{k=1}^{V} \frac{1}{n_C} \sum_{i=(k-1)n_C+1}^{kn_C} \ell(\widehat{f}^{(n_V)}(D_k), Z_i). \tag{4.6.4}$$

Let p statistics $\widehat{f}_1, \ldots, \widehat{f}_p$ be given. The *V-fold CV procedure* is the procedure $\bar{f}_{VCV} = (\bar{f}^{(n)}_{VCV})_n$ defined, for any n, by

$$\bar{f}^{(n)}_{VCV}(D^{(n)}) = \widehat{f}^{(n)}_{\widehat{\jmath}(D^{(n)})}(D^{(n)}) \text{ s.t. } \widehat{\jmath}(D^{(n)}) \in \operatorname*{argmin}_{j\in\{1,\ldots,p\}} R_{n,V}(\widehat{f}_j). \tag{4.6.5}$$

Perhaps the oldest, and certainly the most frequently studied, cross-validation scheme is n-fold or *leave-one-out* cross-validation. It forms the intersection between the class of V-fold cross-validation schemes and the class of *leave-m-out CV* schemes, defined by

$$\bar{f}^{(n)}_{lmo}(D^{(n)}) = \widehat{f}^{(n)}_{\widehat{\jmath}(D^{(n)})}(D^{(n)}) \text{ s.t. } \widehat{\jmath}(D^{(n)}) \in \operatorname*{argmin}_{j\in\{1,\ldots,p\}} R_{n,-m}(\widehat{f}_j), \tag{4.6.6}$$

where $R_{n,-m}$ is defined as

$$R_{n,-m}(\widehat{f}) = \binom{n}{m}^{-1} \sum_{C\subset\{1,\ldots,n\}:|C|=m} \frac{1}{m}\sum_{i\in C} \ell(\widehat{f}^{(n-m)}((Z_k)_{k\in\{1,\ldots,n\}\setminus C}), Z_i).$$

This method does however become very computationally inadequate as soon as m is no longer 1, as there are far too many subsets of $\{1, \ldots, n\}$ to average over. One possible solution for this is *balanced incomplete cross-validation*, where cross-validation is treated as a block design and the available pieces of data are all used equally often for training, and equally often for testing. Alternatively, we could use *Monte Carlo cross-validation*, where the training and testing subsets are drawn randomly — without replacement — from the available data. See [96] for a discussion of all these methods.

We can place all of these cross-validation schemes into one general framework as follows. For any subset $C \subset \{1, \ldots, n\}$ of indices, write $D_{(C)}$ for $\{Z_i : i \in C\}$ and $D_{(C)'}$ for $\{Z_i : i \notin C\}$. Assume that a fixed value n_C be given (the size of test sets), and define $n_V = n - n_C$. Let $C_1, \ldots, C_{N_C}$ be N_C subsets of $\{1, \ldots, n\}$, each of size n_V. Now for any statistic $\widehat{f}$ define the *CV risk*

$$R_{n_C}(\widehat{f}) = \frac{1}{N_C}\sum_{k=1}^{N_C} \frac{1}{n_C} \sum_{i\notin C_k} \ell(\widehat{f}^{(n_V)}(D_{(C_k)}), Z_i), \tag{4.6.7}$$

and its minimizer by

$$\widehat{f}^{(n)}_{CV}(D^{(n)}) = \widehat{f}^{(n)}_{\widehat{\jmath}(D^{(n)})}(D^{(n)}) \text{ s.t. } \widehat{\jmath}(D^{(n)}) \in \operatorname*{argmin}_{j\in\{1,\ldots,p\}} R_{n_C}(\widehat{f}_j). \tag{4.6.8}$$

4.6.2 The modified CV procedure and its average version

In this subsection, we introduce the selecting procedures that we will be studying later. We use the notations introduced in the previous subsection.

To introduce the modified CV procedure, we consider some integer V and we assume that V divides n. We consider the splits $(B_1, D_1), \ldots, (B_V, D_V)$ of the data introduced in (4.6.2) and (4.6.3). We define the **modified CV procedure (mCV)** by

$$\bar{f}^{(n)}_{mCV}(D^{(n)}) = \widehat{f}^{(n_V)}_{\widehat{\jmath}(D^{(n)})}(D^{(n_V)}) \tag{4.6.9}$$

where $D^{(n_V)} = \{Z_1, \ldots, Z_{n_V}\}$ and, for the V-fold CV empirical risk $R_{n,V}$ introduced in (4.6.4),

$$\widehat{\jmath}(D^{(n)}) \in \operatorname*{argmin}_{j\in\{1,\ldots,p\}} R_{n,V}(\widehat{f}_j).$$

For the average version of the mCV procedure, we don't have to split the data in the same "organized" way as in (4.6.2) and (4.6.3). We can consider the more general second partition scheme introduced in the second part of the previous subsection that we recall now for the reader convenience: Let N_C and $1 \leq n_C < n$ be two integers and set $n_V = n - n_C$. Let $C_1, \ldots, C_{N_C}$ be subsets of $\{1, \ldots, n\}$ each of size n_V. We define the **averaged version of the modified CV procedure (amCV)** by:

$$\widehat{f}^{(n)}_{amCV}(D^{(n)}) = \frac{1}{N_C}\sum_{k=1}^{N_C} \widehat{f}^{(n_V)}_{\widehat{\jmath}(D^{(n)})}(D_{(C_k)}) \tag{4.6.10}$$

where, for the CV-risk R_{n_C} introduced in (4.6.7),

$$\widehat{\jmath}(D^{(n)}) \in \operatorname*{argmin}_{j\in\{1,\ldots,p\}} R_{n_C}(\widehat{f}_j).$$

We did not consider the same partition scheme of the data for the two procedures. The one considered for the amCV is more general but to obtain oracle inequalities for the amCV we will need the convexity of the risk. Whereas for the mCV, the partition scheme is the one used for the VCV method and will only require a weak assumption on the basis statistics $\widehat{f}_1, \ldots, \widehat{f}_p$. For each one of our results, we will consider two different setups depending on the procedure that we want to study and the assumptions of the problem.

Note that the difference between the classical VCV procedure defined in (4.6.5) and our mCV procedure is that $\bar{f}^{(n)}_{mCV}$ takes its values in $\{\widehat{f}^{(n_V)}_1, \ldots, \widehat{f}^{(n_V)}_p\}$ whereas $\bar{f}^{(n)}_{VCV}$ takes its values in $\{\widehat{f}^{(n)}_1, \ldots, \widehat{f}^{(n)}_p\}$. Therefore, under some extra "regularity" assumptions on the basis statistics $\widehat{f}_1, \ldots, \widehat{f}_p$ saying that for every j, $\widehat{f}^{(n)}_j$ is somehow more efficient as n increases (cf., for instance, the "stability" assumption in [23]) the VCV procedure should outperform our mCV procedure. Nevertheless, we will not explore this kind of regularity assumption and will require only weak assumptions on the basis estimators. Under these weak assumptions, the mCV (as well as the amCV) will, in fact, outperform the classical VCV and CV procedures (cf. Theorem 4.6.2 and Example 4.6.5 below).

4.6.3 Oracle inequalities for the modified CV procedures (mCV) and its average version (amCV)

In this subsection, we shall not yet introduce any conditions on how a candidate statistic $\widehat{f}$ behaves when its training sample size changes, i.e. about the relationship of $\widehat{f}_m$ and $\widehat{f}_n$ for

$m \neq n$. As the usual application of cross-validation involves retraining the selected model using *all* the available data to obtain a final estimator, such assumptions are crucial for avoiding such pathological "counter-examples" as that found in Example 4.6.5 below. As we shall only introduce such conditions in Subsection 4.6.4, we will first prove a simpler case — the case where even after selection involving estimation with training size n_V, we still only use training samples of size n_V to build the final estimator. The case where we retrain on all available data will then be handled in Subsection 4.6.4.

We will require some simple (fixed sample size) properties on the estimators $\widehat{f}_1, \ldots, \widehat{f}_p$ to obtain an oracle inequality for the modified CV procedure.

Definition 4.6.1 We say that a statistic $\widehat{f} = (\widehat{f}^{(n)})_n$ is *exchangeable* when for any integer n, for any permutation $\phi : \{1, \ldots, n\} \longmapsto \{1, \ldots, n\}$ for any $\pi^{\otimes n}$-almost vector $(z_1, \ldots, z_n) \in \mathcal{Z}^n$, we have $\widehat{f}^{(n)}(z_1, \ldots, z_n) = \widehat{f}^{(n)}(z_{\phi(1)}, \ldots, z_{\phi(n)})$.

Remark that most of the statistics in the batch setup (the setup of this section) satisfy this property. On the other side, statistics coming from the on-line setup are likely to be un-exchangeable.

We shall also use the following assumptions on the tail behavior and the "Margin" (cf. [75] and [107]) of the excess loss function of an estimator $\widehat{f}$.

(A) *There exist $\kappa \geq 1$ and $K_0, K_1 > 0$ such that the following holds. For any $m \in \mathbb{N}$ and any data set $D^{(m)} = \{Z_1, \ldots, Z_m\}$*

1. $\left\| \ell(\widehat{f}^{(m)}(D^{(m)}), \cdot) - \ell(f^*, \cdot) \right\|_{L_{\psi_1}(\pi)} \leq K_0$
2. $\left\| \ell(\widehat{f}^{(m)}(D^{(m)}), \cdot) - \ell(f^*, \cdot) \right\|_{L_2(\pi)} \leq K_1 \big(R(\widehat{f}^{(m)}(D^{(m)})) - R(f^*) \big)^{1/2\kappa}$.

The next oracle inequality for the amCV and the mCV procedures follow from a result similar in nature to the result on the shifted empirical process in Theorem 3.10.1.

Theorem 4.6.2 ([P19]) *Let $\widehat{f}_1, \ldots, \widehat{f}_p$ be p statistics satisfying Assumption (A). We have two different setups depending on the procedure that we want to study. Assume that one of the two conditions holds:*

1. *The risk function $f \longmapsto R(f)$ is convex and our estimator is the amCV procedure $\bar{f}^{(n)} = \widehat{f}^{(n)}_{amCV}$ introduced in (4.6.10).*
2. *The statistics $\widehat{f}_1, \ldots, \widehat{f}_p$ are exchangeable and our procedure is the modified CV procedure $\bar{f}^{(n)} = \bar{f}^{(n)}_{mCV}$ introduced in (4.6.9).*

Then for any $a > 0$, there exists a constant $c = c(a, \kappa)$ such that

$$\mathbb{E}_{D^{(n)}} \Big(R(\bar{f}^{(n)}(D^{(n)})) - R(f^*) \Big) \leq (1+a) \min_{j=1,\ldots,p} \Big[\mathbb{E}_{D^{(n_V)}} R(\widehat{f}_j^{(n_V)}(D^{(n_V)})) - R(f^*) \Big] + c \Big(\frac{\log p}{n_C} \Big)^{\frac{\kappa}{2\kappa - 1}} \vee \Big(\frac{\log n_C \log p}{n_C} \Big).$$

4.6.4 Oracle inequalities for cross-validation itself

In Part 1 of Theorem 4.6.2, we make the assumption that the risk $R(\cdot)$ is convex — for which e.g. the conditional convexity of the contrast function $\ell(f, z)$, for all z, would suffice, and thereafter in Part 2 we assume that our candidate statistics are exchangeable. To derive a result for a

CV estimator retrained on the full data $D^{(n)}$ (instead of the only data $D^{(n_V)}$ like in (4.6.9) and (4.6.10)), we shall combine and strengthen these two assumptions.

Regard the mCV procedure, whose final estimator $\bar{f}^{(n)}_{mCV}(D^{(n)}) = \widehat{f}^{(n_V)}_{\widehat{\jmath}(D^{(n)})}(D^{(n_V)})$ is retrained on the first n_V pieces of data. For symmetry reasons, Part 2 of Theorem 4.6.2 remains true for any $k = 1, \dots, V$, if we replace $\bar{f}^{(n)}_{mCV}(D^{(n)})$ by $\bar{f}^{(n)}_{mCV,k}(D^{(n)}) = \widehat{f}^{(n_V)}_{\widehat{\jmath}(D^{(n)})}(D_k)$ using the training set D_k from the $k-$th split.

Now assume that $\mathcal{Z} = \mathbb{R}$ and the statistics $\widehat{f}_1, \dots, \widehat{f}_p$ can all be written as functionals on the cumulative distribution function of the data, i.e. that there exist functionals $G_1, \dots, G_p$ such that

$$\widehat{f}^{(m)}_j(D^{(m)}) = G_j(F_{D^{(m)}}),\ j = 1, \dots, p, m \in \mathbb{N}, \tag{4.6.11}$$

where $F_{D^{(m)}}(z) = \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}\{Z_i \le z\}, \forall z \in \mathbb{R}$. (This assumption automatically implies the exchangeability of the statistics. In particular, all M-estimators, such as the mean or median, have such a functional form). Obviously $F_{D^{(n)}} = V^{-1}\sum_{k=1}^{V} F_{D_k}$. Thus if the risk $R(\cdot)$ is convex, and all the compositions $R \circ G_j$ too, then we can combine the upper bounds for the estimators $\bar{f}^{(n)}_{mCV,k}(D^{(n)})$ obtained in Part 2 of Theorem 4.6.2 to derive a bound for the VCV procedure (4.6.5) as follows:

$$\begin{aligned} R\Big(\bar{f}^{(n)}_{VCV}(D^{(n)})\Big) &= R\left(G_{\widehat{\jmath}(D^{(n)})}\left(F_{D^{(n)}}\right)\right) = R\left(G_{\widehat{\jmath}(D^{(n)})}\left(\frac{1}{V}\sum_{k=1}^{V} F_{D_k}\right)\right) \\ &\le \frac{1}{V}\sum_{k=1}^{V} R\left(G_{\widehat{\jmath}(D^{(n)})}\left(F_{D_k}\right)\right) = \frac{1}{V}\sum_{k=1}^{V} R\left(\bar{f}^{(n)}_{mCV,k}(D^{(n)})\right), \end{aligned}$$

and thus it easily follows from Part 2 of Theorem 4.6.2 the result:

Theorem 4.6.3 *Let $\widehat{f}_1, \dots, \widehat{f}_p$ be p statistics that can be written as functionals $G_1, \dots, G_p$ as in (4.6.11) and which satisfy Assumption (A), and assume that all the compositions $R \circ G_1, \dots, R \circ G_p$ are convex, as also is the risk function $R(\cdot)$. Then for the V-fold cross-validation procedure, we have the oracle inequality*

$$\begin{aligned} &\mathbb{E}_{D^{(n)}}\Big(R(\bar{f}^{(n)}_{VCV}(D^{(n)})) - R(f^*)\Big) \\ &\le (1+a)\min_{j=1,\dots,p}\Big[\mathbb{E}_{D^{(n_V)}} R(\widehat{f}^{(n_V)}_j(D^{(n_V)})) - R(f^*)\Big] + c\Big(\frac{\log p}{n_C}\Big)^{\frac{\kappa}{2\kappa-1}} \vee \Big(\frac{\log n_C \log p}{n_C}\Big). \end{aligned}$$

Remark 4.6.4 *The "functional convexity condition" on the $R \circ G_j$ is a strong one, but need not be exactly fulfilled — it suffices for it to hold up to a summand that converges to zero no slower than the residual term in Theorem 4.6.2, and versions of it averaged over the training data may also suffice. In most practical cases, the only straightforward way of showing the convexity of the $R \circ G_j$ (with high certainty) is by simulation. In the standard example of least-squares regression with underlying Gaussian linear model, for instance, $R \circ G_j$ is convex for the fixed-design setup, regardless of other parameters, but for the random-design setup we need additional conditions such as a reasonable signal-to-noise ratio or large enough sample size (indicating that such a convexity condition does in fact hold* up to *a quickly-decaying extra summand). Simulations of a straightforward sparse Lasso example with 100-dimensional Gaussian covariates and Gaussian noise have shown that the neccessary functional convexity condition for 10-fold cross-validation holds from a sample size of 40 and a signal-to-noise ratio of 2.0 upwards, for a range of penalty tuning parameters. However, discussing this issue at length is beyond the scope of this section.*

The reason why we need extra assumptions such as the functional form of the candidate statistics is that the computation of the index $\widehat{\jmath}(D^{(n)})$ only involves the performances of the estimators for n_V observations ($R_{n_C}(\widehat{f})$ depends only on $\widehat{f}^{(n_V)}$). Without extra assumptions, it is thus easy to contrive counter-examples for which $\widehat{f}^{(n_V)}$ performs well and $\widehat{f}^{(n)}$ performs badly:

Example 4.6.5 *Fix an integer V and a sample size $n > 1$ that is a multiple of V. We will construct a set $\mathcal{F}_n = \{\widehat{f}_1, \widehat{f}_2\}$ of two estimators (which are functionals of the training data) for which V-fold cross-validation does* not *satisfy the oracle inequality from Theorem 4.6.3.*

We consider the classification problem with $0-1$ loss $\ell(f, Z) = \ell(f, (X, Y)) = \mathbb{1}_{f(X)\neq Y}$. Assume that $Y \equiv 1$ a.s. and X is uniformly distributed on $[0, 1]$. The Bayes rule is thus given by $f^*(x) = \mathbb{P}(Y = 1|X = x) = 1, \forall x \in [0, 1]$. We define statistics $\widehat{f}_1 = (\widehat{f}_1^{(n)})_n$ and $\widehat{f}_2 = (\widehat{f}_2^{(n)})_n$ by

$$\widehat{f}_1^{(p)} \equiv \begin{cases} 0 & \text{if} \quad 1 \leq p \leq n-1 \\ 1 & \text{if} \quad p \geq n \end{cases} \quad \text{and } \widehat{f}_2^{(p)} \equiv \begin{cases} 1 & \text{if} \quad 1 \leq p \leq n-1 \\ 0 & \text{if} \quad p \geq n \end{cases} .$$

It is easy to see that $\widehat{\jmath}(D^{(n)}) = \arg\min_{j\in\{1,2\}} R_{n,V}(\widehat{f}_j)$ is always equal to 2. Thus the V-fold CV procedure is $\bar{f}_{VCV}^{(n)}(D^{(n)}) = \widehat{f}_{\widehat{\jmath}(D^{(n)})}^{(n)}(D^{(n)}) = \widehat{f}_2^{(n)}(D^{(n)})$. Set $\mathcal{F}_n = \{\widehat{f}_1, \widehat{f}_2\}$. For any $1 \leq p \leq n$, it is easy to check that

$$\min_{\widehat{f}\in\mathcal{F}_n} \mathbb{E}_{D^{(n)}}[R(\widehat{f}^{(p)}(D^{(p)})) - R^*] = 0 \text{ and } \mathbb{E}_{D^{(n)}}[R(f_{VCV}^{(n)}(D^{(n)})) - R^*] = 1.$$

As we can do this for arbitrarily high sample sizes n, V-fold cross-validation is not even risk-consistent at this level of generality — and certainly does not satisfy any meaningful oracle inequalities.

4.6.5 The continuous case

We consider Θ a set of indexes and $F = \{\widehat{f}_\theta : \theta \in \Theta\}$ a set of statistics indexed by Θ. In the previous part of this section, we have explored the case $\Theta = \{1, \ldots, p\}$. In this section, we need not assume Θ to be finite.

We consider the notation introduced in Subsection 4.6.1, and define the continuous version of the modified CV procedure by

$$\bar{f}_{mCV}^{(n)}(D^{(n)}) = \widehat{f}_{\widehat{\theta}(D^{(n)})}^{(n_V)}(D^{(n_V)}) \text{ where } \widehat{\theta}(D^{(n)}) \in \operatorname*{argmin}_{\theta\in\Theta} R_{n,V}(\widehat{f}_\theta) \tag{4.6.12}$$

and the continuous version of the averaged version of the modified CV procedure by

$$\widehat{f}_{amCV}^{(n)}(D^{(n)}) = \frac{1}{N_C}\sum_{k=1}^{N_C} \widehat{f}_{\widehat{\theta}(D^{(n)})}^{(n_V)}(D_{(C_k)}) \text{ where } \widehat{\theta}(D^{(n)}) \in \operatorname*{argmin}_{\theta\in\Theta} R_{n_C}(\widehat{f}_\theta). \tag{4.6.13}$$

Remark that we assume that the infimum of $\theta \longmapsto R_{n,V}(\widehat{f}_\theta)$ and $\theta \longmapsto R_{n_C}(\widehat{f}_\theta)$ are achieved. We also called these two infima by the same name but there will be no ambiguity since we will use them in two clearly separated setups. It follows from similar results on the shifted empirical process as in Theorem 3.10.1, a continuous version of Theorem 4.6.2.

Theorem 4.6.6 ([P19]) *Let Θ be a set of indexes and $F = \{\widehat{f}_\theta : \theta \in \Theta\}$ be a set of statistics indexed by Θ. Fix $n_V \leq n$ the size of the validation sample and define the set of excess loss*

functions associated with F by $\mathcal{E}_F = \{\ell(\widehat{f}_\theta^{(n_V)}(D^{(n_V)}), \cdot) - \ell(f^*, \cdot) : \theta \in \Theta\}$. We assume that the tail behavior of the statistics in F and the complexity of F satisfy the following assumptions:

Any statistic $\widehat{f}$ in F satisfies (A) and there exist λ_{min} and a strictly increasing function J such that J^{-1} is strictly convex, the convex conjugate ψ of J^{-1} increases, $\psi(\infty) = \infty$ and there exists $r \geq 1$ such that $x \mapsto \psi(x)/x^r$ decreases and

$$J(\lambda) \geq \gamma_2(\mathcal{E}_F^\lambda, \|\cdot\|_{L_2}) + \frac{(\log n_C)\gamma_1(\mathcal{E}_F^\lambda, \|\cdot\|_{\psi_1})}{\sqrt{n_C}}, \forall \lambda > \lambda_{min}$$

where $\mathcal{E}_F^\lambda = \{\mathcal{E} \in \mathcal{E}_F : \|\mathcal{E}(Z)\|_{L_2} \leq \lambda^{1/2\kappa}\}$.

We consider two different setups depending on the procedure we want to study. Assume that one of the two conditions holds:

1. *The risk function $f \longmapsto R(f)$ is convex and our procedure is the amCV procedure $\bar{f}^{(n)} = \widehat{f}_{amCV}^{(n)}$ defined in (4.6.13).*

2. *The statistics $\widehat{f}_1, \ldots, \widehat{f}_p$ are exchangeable and our procedure is the mCV procedure $\bar{f}^{(n)} = \bar{f}_{mCV}^{(n)}$ introduced in (4.6.12).*

Then, for every $a > 0$ and $q > 1$, the following inequality holds

$$\mathbb{E}_{D^{(n)}}\left(R(\bar{f}^{(n)}(D^{(n)})) - R(f)\right) \leq (1+a)\inf_{\theta\in\Theta}\left[\mathbb{E}_{D^{(n_V)}}R(\widehat{f}_\theta^{(n_V)}(D^{(n_V)})) - R(f^*)\right] + \frac{ac\lambda_q(1/q)}{q},$$

where we set $\lambda_q(u) = \psi\left(\frac{2q^{r+1}(1+a)u}{a\sqrt{n_C}}\right) \vee \lambda_{min}, \forall u > 0$.

Note that Theorem 4.6.6 generalizes Theorem 4.6.2 to a continuous family of estimators. Indeed, it is easy to verify that, in the finite case $|\Theta| = p$, we obtain the same result as in Theorem 4.6.2. For instance, under the assumptions of Theorem 4.6.2 by using Equation (3.2.10), we have, for any $\lambda > 0$,

$$\frac{(\log n_C)\gamma_1(\mathcal{Q}_\lambda^{L_2}, \|\cdot\|_{\psi_1})}{\sqrt{n_C}} + \gamma_2(\mathcal{Q}_\lambda^{L_2}, \|\cdot\|_{L_2}) \leq \frac{K_0(\log n_C)\log p}{\sqrt{n_C}} + \lambda^{1/2\kappa}\sqrt{\log p} = J(\lambda);$$

thus, the convex conjugate of J^{-1} is

$$\psi(v) = \frac{K_0(\log n_C)\log p}{\sqrt{n_C}}v + c_\kappa\left(v\sqrt{\log p}\right)^{\frac{2\kappa}{2\kappa-1}}, \forall v > 0.$$

Thus, $\lambda_q(1/q)$ is, up to some constant depending only on K_0 and κ, of the same order as the residue of the oracle inequality of Theorem 4.6.2. Furthermore, the same reasoning used for Theorem 4.6.3 can also be applied here in sufficiently convex setups where the full data set is used for retraining. Nevertheless, from a technical point of view, there is a major difference between the finite and the continuous cases. In the finite case, it is only a side effect of the Margin assumption (cf. second point of Assumption (A)) that is actually used, namely a better concentration of the empirical risk to the actual risk. Whereas in the continuous case, all the strength of the Margin assumption is used: a reduction of the complexity of the localized sets.

4.6.6 Adaptive choice of the regularization parameter for the Lasso

We consider the linear regression model $Y = \langle X, \beta^* \rangle + \sigma\epsilon$, where $Y \in \mathbb{R}$ is a random variable, $X \in \mathbb{R}^p$ is a random vector and $\epsilon \in \mathbb{R}$ is a random variable (the noise) independent of X such that $\mathbb{E}\epsilon = 0$ and $\mathbb{E}\epsilon^2 = 1$. We have n i.i.d. observations in this model, and the total dataset consists of $\mathbf{Y} = (Y_1, \ldots, Y_n)^\top$ and $\mathbf{X} = (X_1, \ldots, X_n)^\top$. We consider the function $\Phi : \mathbb{R}^p \times \mathbb{R}^+ \longmapsto \mathbb{R}$ defined by

$$\Phi(\beta, \lambda) = |\mathbf{Y} - \mathbf{X}\beta|_n^2 + \lambda|\beta|_1.$$

Given a regularization parameter λ, the Lasso estimator $\widehat{f}_\lambda$ is defined by

$$\widehat{f}_\lambda^{(n)}(\cdot, D^{(n)}) = \langle \cdot, \widehat{\beta}(\lambda)^{(n)}(D^{(n)}) \rangle \text{ where } \widehat{\beta}(\lambda)^{(n)}(D^{(n)}) \in \operatorname{Arg\,min}_{\beta \in \mathbb{R}^p} \Phi(\beta, \lambda).$$

We consider the regularization parameter λ to be normalized so as to lie in $[0, 1]$. Such a normalization is possible, since for $\lambda_{max} = 2 \max_i |\langle X_i, Y \rangle|$, the zero vector is a minimizer of $\Phi(\beta, \lambda_{max})$; that is, the Lasso penalty is always able to shrink the coefficient estimate for β down to zero. Thus the dictionary of estimators that we consider is a finite set $\{\widehat{f}_\lambda : \lambda \in \mathcal{G}\}$ where $\mathcal{G}$ is a finite grid of $[0, 1]$.

Now, we construct the mCV procedure (cf. (4.6.9)) in this setup. Let $(B_1, D_1), \ldots, (B_V, D_V)$ be the family of splits of $D^{(n)}$ defined in (4.6.2) and (4.6.3) for some $1 \leq V \leq n$ dividing n. For any Lasso estimator $\widehat{f}_\lambda$ the r-V-fold CV empirical risk, for $r > 0$, is defined by

$$R_{n,V}^{(r)}(\widehat{f}_\lambda) = \frac{1}{V} \sum_{k=1}^{V} \frac{1}{n_C} \sum_{i=(k-1)n_C+1}^{kn_C} |Y_i - \langle X_i, \widehat{\beta}(\lambda)^{(n_V)}(D_k) \rangle|^r.$$

The mCV procedure is defined in this context by

$$\bar{f}_{mCV}^{(n)}(\cdot, D^{(n)}) = \widehat{f}^{(n_V)}_{\widehat{\lambda}_r(D^{(n)})}(\cdot, D^{(n_V)}) = \langle \cdot, \widehat{\beta}(\widehat{\lambda}_r(D^{(n)}))^{(n_V)}(D^{(n_V)}) \rangle = \langle \cdot, \bar{\beta}_{mCV}^{(n)}(D^{(n)}) \rangle$$

where

$$\widehat{\lambda}_r(D^{(n)}) \in \operatorname*{argmin}_{\lambda \in \mathcal{G}} R_{n,V}^{(r)}(\widehat{f}_\lambda).$$

Now, we construct the amCV (cf. (4.6.10)) procedure using the subsets $C_1, \ldots, C_{N_C}$ of $\{1, \ldots, n\}$ each of size n_V: the mCV is defined, in this context, by

$$\begin{aligned} \widehat{f}_{amCV}(D^{(n)})(\cdot, D^{(n)}) &= \frac{1}{N_C} \sum_{k=1}^{N_C} \widehat{f}^{(n_V)}_{\widehat{\lambda}_r(D^{(n)})}(\cdot, D_{(C_k)}) \\ &= \langle \cdot, \frac{1}{N_C} \sum_{k=1}^{N_C} \widehat{\beta}(\widehat{\lambda}_r(D^{(n)}))^{(n_V)}(D_{(C_k)}) \rangle = \langle \cdot, \widehat{\beta}_{amCV}^{(n)}(D^{(n)}) \rangle, \end{aligned}$$

where $\widehat{\lambda}_r(D^{(n)}) \in \operatorname{argmin}_{\lambda \in \mathcal{G}} R_{n_C}^{(r)}(\widehat{f}_\lambda)$ and $R_{n_C}^{(r)}$ is the r-CV risk.

From a theoretical point of view, of course, we should have minimized the r-CV risk over $\lambda \in [0, 1]$ (for the mCV and the amCV). But we have in mind to perform the Lasso procedure by means of the LARS algorithm. This algorithm provides a family of regularization parameters $0 = \lambda^{(0)} < \lambda^{(1)} < \ldots < \lambda^{(N)}$, where N may differ from n, and the corresponding Lasso estimators $\widehat{f}_{\lambda^{(j)}}, j = 1, \ldots, N$. Thus we believe that using the LARS algorithm combined with the mCV or amCV procedures with a grid $\mathcal{G} \subset \{\lambda^{(0)}, \ldots, \lambda^{(N)}\}$ will prove to be efficient.

Note that for values of r close to 0, the Lasso vector $\bar{\beta}^{(n)}_{mCV}$ constructed with a data-driven choice of the regularization parameter $\widehat{\lambda}_r(D^{(n)})$ is likely to enjoy some model selection (or sign consistency) properties. Nevertheless, from a theoretical point of view, we will obtain results only for the prediction problem with respect to the L_2-risk.

We would like to apply Theorem 4.6.2 to the two procedures that we have introduced here. To this end, we have to check assumption (A) for the elements of the dictionary $F = \{\widehat{f}_\lambda : \lambda \in \mathcal{G}\}$ and so the design vector X has to enjoy some properties.

Definition 4.6.7 Let X be a random vector of $\mathbb{R}^p$ and denote by μ its probability distribution. We say that X is *log-concave* when for all nonempty measurable sets $A, B \subset \mathbb{R}^p$ and every $\alpha \in [0,1]$, $\mu(\alpha A + (1-\alpha)B) \geq \mu(A)^\alpha \mu(B)^{1-\alpha}$. We say that X is a ψ_2 *vector* when $\|X\|_{\psi_2} = \sup_{x \in \mathcal{S}_2^{p-1}} \left\|\langle X, x\rangle\right\|_{\psi_2} < \infty$.

Many natural measures are log-concave. Among the examples are measures that have a log-concave density, the volume measure of a convex body, and many others. A well known fact on a log-concave random vector X of $\mathbb{R}^p$ follows from Borell's inequality (cf. [86]): for every $x \in \mathbb{R}^p$, $\left\|\langle X, x\rangle\right\|_{\psi_1} \leq c \left\|\langle X, x\rangle\right\|_{L_1}$ where c is an absolute constant. In particular, the moments of linear functionals satisfy, for all $p \geq 1$, $\left\|\langle X, x\rangle\right\|_{L_p} \leq cp \left\|\langle X, x\rangle\right\|_{L_1}$.

In the following we assume that X is a ψ_2, log-concave vector and the noise ϵ is ψ_2.

Let $m \in \mathbb{N}$, $\beta = \widehat{\beta}(\lambda)^{(m)}(D^{(m)})$ be fixed for the moment, and let $\mathcal{L}_\beta(X,Y) = (Y - \langle X, \beta\rangle)^2 - (Y - \langle X, \beta^*\rangle)^2$ be the corresponding excess loss function. We need to bound the ψ_1-norm of $\mathcal{L}_\beta$ and to check the Margin assumption. For the second task, we use the log-concavity of X to obtain

$$\begin{aligned}\mathbb{E}\mathcal{L}_\beta(X,Y)^2 \leq& 2\mathbb{E}\langle X, \beta - \beta^*\rangle^4 + 8\sigma^2 \mathbb{E}\langle X, \beta - \beta^*\rangle^2 \\ \leq& (c + 8\sigma^2)\mathbb{E}\langle X, \beta - \beta^*\rangle^2 = (c + 8\sigma^2)\mathbb{E}\mathcal{L}_\beta.\end{aligned}$$

This proves that the dictionary F satisfies the Margin assumption with $\kappa = 1$. For the first task, we use the fact that X is ψ_2 to get

$$\begin{aligned}\|\mathcal{L}_\beta(X,Y)\|_{\psi_1} &= \left\|\langle X, \beta - \beta^*\rangle^2 + 2\sigma\epsilon\langle X, \beta^* - \beta\rangle\right\|_{\psi_1} \\ &\leq (1 + 2\sigma)\left\|\langle X, \beta - \beta^*\rangle\right\|_{\psi_2}^2 + 2\sigma\|\epsilon\|_{\psi_2}^2 \\ &\leq (1 + 2\sigma)\|X\|_{\psi_2}^2 \|\beta - \beta^*\|_2^2 + 2\sigma\|\epsilon\|_{\psi_2}^2.\end{aligned}$$

Now for the construction of the dictionary, we threshold all the Lasso vectors $\widehat{\beta}(\lambda^{(j)})$ provided by the LARS algorithm, in such a way that the ℓ_2-norm of these vectors is smaller than a constant K_0'. Then the dictionary F satisfies Assumption (A) (with $K_0 = K_0' + \|\beta^*\|_2$). Thus, we are now in position to apply Theorem 4.6.2.

Let $\widehat{\beta}$ be either $\bar{\beta}^{(n)}_{mCV}(D^{(n)})$ or $\widehat{\beta}^{(n)}_{amCV}(D^{(n)})$, we have

$$\begin{aligned}\mathbb{E}[(Y - \langle X, \widehat{\beta}\rangle)^2] \leq& (1+a)\min_{\lambda \in \mathcal{G}} \mathbb{E}[(Y - \langle X, \tau(\widehat{\beta}(\lambda)^{(n_V)}(D^{(n_V)}))\rangle)^2] \\ &+ c\frac{\log|\mathcal{G}|\log(n_C)}{n_C},\end{aligned}$$

where τ is a thresholded function such that $\forall \beta \in \mathbb{R}^p, \|\tau(\beta)\|_2 \leq K_0'$.

This proves that the adaptation procedures constructed in this section optimize the prediction task of the Lasso thanks to a data-driven choice of the regularization parameter.

Chapter 5

Open problems

5.1 Optimality of the AEW in the regression model with random design

The suboptimality in expectation of the AEW, obtained in Theorem 2.6.1, is rather surprising for two reasons. First of all, it is known that the progressive mixture rule is optimal in expectation for T larger than some parameters of the model (see [39], [117], [119], [54] or [8]). This procedure is defined by

$$\bar{f} = \frac{1}{n}\sum_{k=1}^{n} \tilde{f}_k^{AEW},$$

where $\tilde{f}_k^{AEW}$ is the function generated by AEW (with common temperature parameter T) associated with the dictionary F and constructed using the first k observations $Z_1, \dots, Z_k$. Thus, this aggregate is the mean of $\tilde{f}_k^{AEW}$ for $1 \leq k \leq n$, where, for every $k < n$, $\tilde{f}_k^{AEW}$ is constructed using only the first k observations. In particular, $\bar{f}$ is the mean of aggregates that are (or should be) less "efficient" than $\tilde{f}_n^{AEW}$, since the latter is constructed using all the observations $Z_1, \dots, Z_n$, rather than a subset of the given observations. That is why one expects the AEW to be an optimal aggregation procedure in expectation — at least in the high temperature regime. Theorem 2.6.1 shows that, even for temperature of the order of a constant, $\tilde{f}_n^{AEW}$ might have a very bad behavior, of the order of $(1/\sqrt{n})$.

Second, the optimality in expectation of AEW was obtained in [42] for the regression model $Y_i = f(x_i) + \epsilon_i$ with a deterministic design $x_1, \dots, x_n \in \mathcal{X}$ with respect to the risk $\|g - f\|_n^2 = n^{-1}\sum_{i=1}^n (g(x_i) - f(x_i))^2$ (with its empirical version being $R_n(g) = n^{-1}\sum_{i=1}^n (Y_i - g(x_i))^2$); that is, it was shown that for $T \geq c\max(b, \sigma^2)$ (where σ^2 is the variance of the noise ϵ),

$$\mathbb{E}\left\|\tilde{f}_n^{AEW} - f\right\|_n^2 \leq \min_{g\in F} \|g - f\|_n^2 + \frac{T\log M}{n+1}. \tag{5.1.1}$$

Theorem 2.6.1 shows that the behavior of the AEW is very different, at least in the low temperature regime. The fact that the same procedure (although in different models) can exhibit such two extreme behaviors - and for roughly the same temperature parameter is rather striking. The $1/\sqrt{n}$ lower bound of Theorem 2.6.1 vs. the $1/n$ upper bound derived from the oracle inequality (5.1.1) can have one of the two following explanations. Either that the two seemingly similar scenarios are, in fact very different, or that AEW exhibits a sharp phase transition at $T \sim c$. And, if the latter is true, then an important outcome of Theorem 2.6.1 is that the temperature parameter is of the highest importance with regard to the optimality of the AEW in expecta-

tion and that a slightly modified choice of this parameter can result in a huge difference in the residual term.

All the optimal upper bounds on AEW or on the progressive mixture need T to be larger than some unknown parameters of the model (the variance of the noise in particular). This means that in practice, AEW is likely to be a very "risky" aggregation procedure because of its sensitivity to the temperature parameter. Moreover, and to make things even worse, even for large values of T, AEW is suboptimal with a constant probability for small dictionaries (Part 2 of Theorem 2.6.1) and with probability that tends to 1 for larger dictionaries (Theorem 2.6.2). Hence, given a set of data and a dictionary, AEW *is likely* to behave very poorly regardless of what T is. In contrast, Theorem 2.7.1 shows that the choice of the temperature parameter has no significant effect on the performance of the AEW (residual term of the order of $T(\log M)/n$) under the Bernstein condition.

From a practical point of view, we believe that exponential aggregating schemes simply should not be used in the regression setup with random design because (cf. also the comments in [7]):

1. for any temperature $T \leq c_0\sqrt{n}/\log n$, there is an event of constant probability such that AEW performs poorly (second point of Theorem 2.6.1);

2. if the temperature parameter is chosen too small (like a constant) then even in expectation the AEW can perform badly (first part of Theorem 2.6.1).

Nevertheless, from a theoretical point of view it remains to be seen whether AEW is an optimal aggregation procedure in expectation and for high temperatures (larger than some constant) in the regression model with random design. This question is mainly interesting from a theoretical point of view.

5.2 Optimality of ERM in Convex aggregation

In Section 2.8, we studied the problem of Convex aggregation. We constructed an optimal aggregation procedure in deviation for this problem. We also proved that the ERM over the convex hull of the dictionary (called ERM-C) achieves the rate M/n when $M \leq \sqrt{n}$ and $\sqrt{\log M/n}$ when $M > \sqrt{n}$. Moreover, when the dictionary is orthogonal then the optimal rate $\psi_n^{(C)}(M)$ is achieved by ERM-C. It still remains to prove that this procedure is indeed an optimal aggregation procedure for the Convex aggregation problem in deviation in its full generality: for any integers n and M, any dictionary F of size M any couple of random variables (X, Y) such that $|Y|, \max_{f\in F}|f(X)| \leq 1$ a.s. and any $x > 0$, with probability greater than $1 - c_0\exp(-x)$, the quadratic risk of ERM-C is such that

$$R(\tilde{f}_n^{ERM-C}) \leq \min_{f\in \mathrm{conv}(F)} R(f) + c_1 \max\left(\psi_n^{(C)}(M), \frac{x}{n}\right).$$

I solved this problem one month after the submission of this work in [P20].

5.3 An optimal lower bound for the ERM-C in the context of (MS) aggregation

In Section 2.3, we improve the geometry of the model F by considering the ERM over the convex hull of F and try to use this procedure $\tilde{f}_n^{ERM-C}$ for the (MS) aggregation problem. It appears

that this procedure is suboptimal for this problem. We prove in Theorem 2.3.1 that for any n and M such that $\log M \leq c_0 n^{1/3}$, there exists a dictionary F of cardinality M and a probability measure P for (X,Y) such that with $P^{\otimes n}$-probability greater than $3/4$,

$$R(\tilde{f}_n^{ERM-C}) \geq \min_{f\in F} R(f) + c_2\psi_n(M) \tag{5.3.1}$$

where $\psi_n(M) = M/n$ when $M \leq \sqrt{n}$ and $\left(n\log\left(eM/\sqrt{n}\right)\right)^{-1/2}$ when $M > \sqrt{n}$. We proved in Theorem 3.9.1 that the rate $\psi_n(M)$ is optimal for the counter-example used in Theorem 2.3.1. Roughly speaking the counter-example used here is B_1^M for the model and the target Y is orthogonal to the model in 0. It appears that the complexity around 0 in B_1^M increases as the dimension M grows that is the reason why the residual term $\psi_n(M)$ decreases as M increases.

We believe that a better counter-example would provide an optimal lower bound for the procedure $\tilde{f}_n^{ERM-C}$ for the (MS) aggregation problem of the order of $\psi_n^{(C)}(M)$ which is M/n when $M \leq \sqrt{n}$ and $\sqrt{\log\left(eM/\sqrt{n}\right)/n}$ when $M > \sqrt{n}$. It may be possible that the following counter-example would provide the desired lower bound: Consider the bounded regression framework with respect to the square loss. Let $\phi_1,\dots,\phi_{M+1}$ be real-valued functions defined on $\mathcal{X}$ and X be a random variable such that $\phi_1,\dots,\phi_{M+1}$ are orthogonal in $L^2(P_X)$, $\phi_1(X),\dots,\phi_M(X)$ are uniformly distributed over $[-\sqrt{3},\sqrt{3}]$ and $\phi_{M+1}(X)$ is a Rademacher variable independent of $\phi_1(X),\dots,\phi_M(X)$. Take $m = \left\lceil \left(n/\log\left(eM/\sqrt{n}\right)\right)^{1/2}\right\rceil$ and defined $\phi = m^{-1}\sum_{j=1}^m \phi_j$. We consider the model $F = \{\phi, \pm\phi_1,\dots,\pm\phi_M\}$ and the target $Y = \phi_{M+1}(X) + \phi(X)$. Since $\min_{f\in F} R(f) = \min_{f\in \mathrm{conv}(F)} R(f)$, the model is roughly speaking B_1^M and the target is orthogonal to the model at the point $x_0 = m^{-1}\sum_{j=1}^m e_j$ where $e_1,\dots,e_M$ is the canonical basis of $\mathbb{R}^M$. Somehow, since the complexity of B_1^M increases around 0, we believe that the complexity of B_1^M around x_0 may slightly decrease when the dimension M grows. This may then provide this extra logarithmic term in the lower bound of the ERM-C that we were not able to achieve in (5.3.1) when $M > \sqrt{n}$.

5.4 Convex model and the ERM

In this section, we sum up some results on the problem of aggregation and we expose a problem which follows.

Given a functions class F, the natural candidate to achieve $\inf_{f\in F} R(f)$ is the ERM over F: $\widehat{f} \in \mathrm{argmin}_{f\in F} R_n(f)$. This procedure has been intensively studied in the Learning theory literature. Nevertheless, finding properties on F for which the ERM algorithm is an "optimal" algorithm remains still an open problem.

For the (MS) aggregation problem, according to Theorem 2.2.1, for any n and M, there exists a set F of M functions and a couple (X,Y) for which the ERM will be a suboptimal procedure (a $\sqrt{(\log M)/n}$ lower bound can be obtained for the ERM whereas, in the same setting, a procedure achieving the fast rate $(\log M)/n$ can be constructed). On the other side, for the (L) aggregation problem, when $F = \mathrm{span}(f_1,\dots,f_M)$ then the ERM is an optimal procedure. And for the (C) aggregation problem, when $F = \mathrm{conv}(f_1,\dots,f_M)$, we conjecture that the ERM is also an optimal aggregation procedure (note that in Theorem 2.8.3 we were able to prove that the ERM-C achieves the optimal rate of Convex aggregation when all the elements in F are orthogonal). Therefore, we wonder why in the linear and convex situation, the ERM should be optimal whereas for the (MS) aggregation problem, this is not the case. Thus, one can ask the question: what is the fundamental difference between, on one side, a set of M functions and,

on the other side, the convex hull or the linear span of M functions which is at the heart of the sub-optimality of the ERM in one case and its optimality in the other case?

Considering the complexity point of view to this question leads to the important concept of "over-fitting": for classes of functions F which are too rich the ERM is not optimal (cf. [21]) because it is likely that in the set F one function will perform very nicely on the data but will have very bad generalization capability. Understanding this concept was the leading idea to the introduction of penalized empirical risk minimization procedures (cf. penalized estimators).

Now, fix the complexity of F to be "small", for instance F is contained in a linear span of M functions where M is small compared to n, so that the reason of the suboptimality of the ERM cannot lie anymore in the concept of "over-fitting". The question on the fundamental structure of F at the heart of the difference between "the ERM over F is on optimal procedure to mimic the oracle in F" and "the ERM over F fails to mimic the oracle in F at the optimal rate" becomes now a problem on the geometrical structure of F. Understanding what is the fundamental concept behind this question may provide algorithms which can encounter this structural problem (like the concept of "over-fitting" was the reason of the introduction of penalty functions). This reasoning is behind the procedures introduced in Chapter 2.

To be more precise, we now introduce the problem in a formal way. Let F be a functions class. We say that $(r_n(F))_{n\in\mathbb{N}}$ is the *optimal rate of aggregation of* F when, there exists some absolute positive constants c_1, c_2 and c_3, such that for any $n \in \mathbb{N}$:

1. there exists an estimators $\widehat{f}_n$ such that for any random couple (X, Y) of probability measure P such that $|Y| \leq 1$ and $\max_{f\in F} |f(X)| \leq 1$ and for any $x > 0$, with $P^{\otimes n}$-probability greater than $1 - c_0 \exp(-x)$,
$$R(\widehat{f}_n) \leq \inf_{f\in F} R(f) + c_1 \max\left(r_n(F), \frac{x}{n}\right), \tag{5.4.1}$$

2. for any statistic $\widehat{f}_n$, there exists a couple (X, Y) of probability measure P such that $|Y|, |f(X)| \leq 1$ almost surely for all $f \in F$ and, with $P^{\otimes n}$-probability greater than c_2,
$$R(\widehat{f}_n) \geq \inf_{f\in F} R(f) + c_3 r_n(F). \tag{5.4.2}$$

The question now boils down to finding a property (P) such that the ERM over F can achieve the optimal rate of aggregation of F if and only if F satisfies (P).

This question can also be asked in the aggregation framework. Let $M \in \mathbb{N}^*$ and $\Lambda \subset \mathbb{R}^M$. We define the *optimal rate of aggregation of* Λ by a sequence $(r_n(\Lambda))_{n\in\mathbb{N}}$ such that there exists some absolute positive constants c_1, c_2 and c_3 for which for any $n \in \mathbb{N}$,

1. there exists an aggregation procedure $\widehat{f}_n(\cdot)$ such that for any set $F = \{f_1, \ldots, f_M\}$ and any couple (X, Y) of random variables such that $|Y|, \max_{f\in F} |f(X)| \leq 1$ and for any $x > 0$, with $P^{\otimes n}$-probability greater than $1 - c_0 \exp(-x)$,
$$R(\widehat{f}_n) \leq \inf_{f\in\Lambda(F)} R(f) + c_1 \max\left(r_n(\Lambda), \frac{x}{n}\right), \tag{5.4.3}$$
where $\Lambda(F) = \{\sum_{j=1}^M \lambda_j f_j : \lambda \in \Lambda\}$.

2. for any statistic $\widehat{f}_n$, there exist a set $F = \{f_1, \ldots, f_M\}$ and a couple (X, Y) of probability measure P satisfying $|Y|, \max_{f\in F} |f(X)| \leq 1$ such that with $P^{\otimes n}$-probability greater than c_2,
$$R(\widehat{f}_n) \geq \inf_{f\in\Lambda(F)} R(f) + c_3 r_n(\Lambda). \tag{5.4.4}$$

The problem is here to find a property (P') on Λ such that for any set $F = \{f_1, \dots, f_M\}$ of bounded functions the ERM over $\Lambda(F)$ can achieve the optimal rate of aggregation of Λ if and only if Λ satisfies (P').

Good progresses on these questions were made by the introduction of the Bernstein's condition (cf. [15]). Nevertheless, it appears that this condition does not seem to be necessary and sufficient for the optimality of the ERM over F or $\Lambda(F)$. It may be true that the properties (P) and (P') are geometrical properties like convexity. This would mean that ERM is optimal only over convex model and that for non-convex models one has to find some surrogate procedures like the one introduced in Chapter 2.

5.5 Optimal rate of aggregation for exact and non-exact oracle inequalities

In the previous section, we define the optimal rate of aggregation of a function class. This definition was given with respect to exact oracle inequalities. Here, we consider non-exact oracle inequalities and define an optimal rate of aggregation for the non-exact prediction problem.

Let $M \in \mathbb{N}^*$, $\Lambda \subset \mathbb{R}^M$ and $0 < \epsilon < 1$. We define the *optimal rate of aggregation of* Λ *for the non-exact prediction problem* by a sequence $(r_{n,\epsilon}(\Lambda))_{n\in\mathbb{N}}$ such that there exists some absolute positive constants c_0, c_1, c_2 and c_3 for which for any $n \in \mathbb{N}$,

1. there exists an aggregation procedure $\widehat{f}_n$ such that for any set $F = \{f_1, \dots, f_M\}$ and any couple (X, Y) of random variables of probability measure P satisfying $|Y|, \max_{f\in F} |f(X)| \leq 1$ a.s. and for any $x > 0$, with $P^{\otimes n}$-probability greater than $1 - c_0 \exp(-x)$,

$$R(\widehat{f}_n) \leq (1+\epsilon) \inf_{f\in\Lambda(F)} R(f) + c_1 \max\left(r_{n,\epsilon}(\Lambda), \frac{x}{n}\right), \tag{5.5.1}$$

   where $\Lambda(F) = \{\sum_{j=1}^M \lambda_j f_j : \lambda \in \Lambda\}$.

2. for any statistic $\widehat{f}_n$, there exist a set $F = \{f_1, \dots, f_M\}$ and a couple (X, Y) of random variables of probability measure P satisfying $|Y|, \max_{f\in F} |f(X)| \leq 1$ such that with $P^{\otimes n}$-probability greater than c_2,

$$R(\widehat{f}_n) \geq (1+\epsilon) \inf_{f\in\Lambda(F)} R(f) + c_3 r_{n,\epsilon}(\Lambda). \tag{5.5.2}$$

The problem is to find a property (P'') such that the optimal rates of aggregation $r_n(\Lambda)$ and $r_{n,\epsilon}(\Lambda)$ are proportional up to some constant depending only on ϵ if and only if Λ satisfies (P''). We already know that this is the case for $\Lambda = \{e_1, \dots, e_M\}$ and $\Lambda = \mathbb{R}^M$ and that this is not the case for $\Lambda = B_1^M$. In particular property (P'') is not a convexity property.

Chapter 6

Proofs

Proof of Theorem 3.2.3 and Theorem 3.2.4

We prove Theorem 3.2.3. The proof of Theorem 3.2.4 follows exactly from the same argument (just replace $\mathcal{L}_F$ by $\mathcal{E}_F$, $\mathcal{L}_f$ by $\mathcal{E}_f$ and μ_η^* by ν_η^* every time these terms appear).

The proof follows the ideas from [15]. Fix $\lambda > 0$ and $x > 0$, and note that by Theorem 3.2.1, with probability larger than $1 - 4\exp(-x)$,

$$\|P - P_n\|_{V(\mathcal{L}_F)_\lambda} \leq 2\mathbb{E}\,\|P - P_n\|_{V(\mathcal{L}_F)_\lambda} + K\sigma(V(\mathcal{L}_F)_\lambda)\sqrt{\frac{x}{n}} + Kb_n(V(\mathcal{L}_F)_\lambda)\frac{x}{n}. \tag{6.0.1}$$

Clearly, we have $b_n(V(\mathcal{L}_F)_\lambda) \leq b_n(\mathcal{L}_F)$ and

$$\sigma^2(V(\mathcal{L}_F)_\lambda) = \sup\Big(P(\alpha\mathcal{L}_f)^2 : 0 \leq \alpha \leq 1, f \in F, P(\alpha\mathcal{L}_f) \leq \lambda\Big) \leq B_n\lambda^\beta + B_n^2/n.$$

Since $V(\mathcal{L}_F)$ is star-shaped, $\lambda > 0 \to \phi(\lambda) = \mathbb{E}\,\|P - P_n\|_{V(\mathcal{L}_F)_\lambda}/\lambda$ is non-increasing, and since $\phi(\mu_\eta^*) \leq \eta/4$ and $\rho_n(x) \geq \mu_\eta^*$ where

$$\rho_n(x) = \max\Big(\mu_\eta^*, \frac{16}{\eta}\Big(\frac{xB_nK^2}{n}\Big(\frac{4}{\eta}\Big)^\beta\Big)^{\frac{1}{2-\beta}} + \frac{4KB_n\sqrt{x}}{\eta n} + \frac{4Kb_n(\mathcal{L}_F)x}{\eta n}\Big),$$

we have

$$\mathbb{E}\,\|P - P_n\|_{V(\mathcal{L}_F)_{\rho_n(x)}} \leq (\eta/4)\rho_n(x).$$

Combined with (6.0.1), there exists an event $\Omega_0(x)$ of probability greater than $1 - 4\exp(-x)$, and on $\Omega_0(x)$,

$$\begin{aligned}\|P - P_n\|_{V(\mathcal{L}_F)_{\rho_n(x)}} &\leq (\eta/2)\rho_n(x) + K\sqrt{\frac{(B_n\rho_n(x)^\beta + B_n^2/n)x}{n}} + K\frac{b_n(\mathcal{L}_F)x}{n} \\ &\leq (3\eta/4)\rho_n(x) + 4\Big(\frac{xB_nK^2}{n}\Big(\frac{4}{\eta}\Big)^\beta\Big)^{\frac{1}{2-\beta}} + \frac{KB_n\sqrt{x}}{n} + \frac{Kb_n(\mathcal{L}_F)x}{n} \leq \eta\rho_n(x).\end{aligned}$$

Hence, on $\Omega_0(x)$, if $g \in V(\mathcal{L}_F)$ satisfies that $Pg \leq \rho_n(x)$, then $|Pg - P_ng| \leq \eta\rho_n(x)$. Moreover, if $P\mathcal{L}_f = \beta > \rho_n(x)$, then $g = \rho_n(x)\mathcal{L}_f/\beta \in V(\mathcal{L}_F)_{\rho_n(x)}$; hence $|Pg - P_ng| \leq \eta\rho_n(x)$, and so $(1-\eta)P\mathcal{L}_f \leq P_n\mathcal{L}_f \leq (1+\eta)P\mathcal{L}_f$.

Sketch of the proof of Theorem 3.5.3

First, like in the proof of Theorem 3.5.1 (cf. Lemma 6.0.1 below as well), we prove that with probability greater than $1-4\exp(-x)$,

$$\widehat{r} = \mathrm{crit}(\widehat{f}_n^{RERM}) \le \beta_n(x).$$

Second, recall that for any $r \ge 0$, $f_r^* \in \mathrm{argmin}_{f\in F_r} R(f)$. Set

$$r^* \in \operatorname*{argmin}_{r\ge 0} \big(R(f_r^*) + c_2\rho_n^{\mathcal{L}}(2(r+1),\theta(x))\big).$$

In particular, we have $\inf_{f\in F}\big(R(f)+c_2\rho_n^{\mathcal{L}}(2(\mathrm{crit}(f)+1),\theta(x))\big) = \big(R(f_{r^*}^*)+c_2\rho_n^{\mathcal{L}}(2(r^*+1),\theta(x))$. Let $\Omega_0(x)$ be the event such that $\widehat{r} = \mathrm{crit}(\widehat{f}_n^{RERM}) \le \beta_n(x)$ and $\rho_n^{\mathcal{L}}$ is an upper bound on the isomorphic profile of $\mathcal{L}_{F_{r^*}}$ and of $\mathcal{L}_{F_r}$ for any $r^* \le r \le \beta_n(x)$ (when $r^* \le \beta_n(x)$), that is

$$\frac{1}{2}P\mathcal{L}_{r^*,f} - 4\rho_n^{\mathcal{L}}(2(r^*+1),\beta_n(x)) \le P_n\mathcal{L}_{r^*,f} \le 2P\mathcal{L}_{r^*,f} + 8\rho_n^{\mathcal{L}}(2(r^*+1),\beta_n(x)), \forall f \in F_{r^*} \quad (6.0.2)$$

and for any $r^* \le r \le \beta_n(x)$ (when $r^* \le \beta_n(x)$),

$$\frac{1}{2}P\mathcal{L}_{r,f} - 4\rho_n^{\mathcal{L}}(2(r+1),\beta_n(x)) \le P_n\mathcal{L}_{r,f} \le 2P\mathcal{L}_{r,f} + 8\rho_n^{\mathcal{L}}(2(r+1),\beta_n(x)), \forall f \in F_r. \quad (6.0.3)$$

On the event $\Omega_0(x)$, we adapt the argument of [13] to get the following. If $r^* \ge \widehat{r}$ then $\widehat{f}_n^{RERM} \in F_{r^*}$ and thus by (6.0.2), we have

$$R(\widehat{f}_n^{RERM}) \le R(f_{r^*}^*) + 2P_n\mathcal{L}_{r^*,\widehat{f}_n^{RERM}} + 8\rho_n^{\mathcal{L}}(2(r^*+1),\beta_n(x))$$

and by definition of $\widehat{f}_n^{RERM}$, we have $P_n\mathcal{L}_{r^*,\widehat{f}_n^{RERM}} \le c_2\rho_n^{\mathcal{L}}(2(r^*+1),\beta_n(x))$. It follows that $R(\widehat{f}_n^{RERM}) \le R(f_{r^*}^*) + (8+c_2)\rho_n^{\mathcal{L}}(2(r^*+1),\beta_n(x))$. Otherwise, when $r^* < \widehat{r}$, we use the isomorphic properties of $\mathcal{L}_{F_{\widehat{r}}}$ (since $r^* \le \widehat{r} \le \beta_n(x)$):

$$\begin{aligned}
&R(f_{r^*}^*) - R(f_{\widehat{r}}^*) \ge \frac{1}{2}\big(R_n(f_{r^*}^*) - R_n(f_{\widehat{r}}^*)\big) - 4\rho_n^{\mathcal{L}}(2(\widehat{r}+1),\beta_n(x))\\
&= \frac{1}{2}\big(R_n(f_{r^*}^*) + c_2\rho_n^{\mathcal{L}}(2(r^*+1),\beta_n(x)) - c_2\rho_n^{\mathcal{L}}(2(r^*+1),\beta_n(x)) - R_n(f_{\widehat{r}}^*)\big) - 4\rho_n^{\mathcal{L}}(2(\widehat{r}+1),\beta_n(x))\\
&\ge \frac{1}{2}\big(R_n(\widehat{f}_n^{RERM}) + c_2\rho_n^{\mathcal{L}}(2(\widehat{r}+1),\beta_n(x)) - c_2\rho_n^{\mathcal{L}}(2(r^*+1),\beta_n(x)) - R_n(f_{\widehat{r}}^*)\big) - 4\rho_n^{\mathcal{L}}(2(\widehat{r}+1),\beta_n(x))\\
&\ge \frac{1}{2}\big(c_2\rho_n^{\mathcal{L}}(2(\widehat{r}+1),\beta_n(x)) - c_2\rho_n^{\mathcal{L}}(2(r^*+1),\beta_n(x)) - 4\rho_n^{\mathcal{L}}(2(\widehat{r}+1),\beta_n(x))\big) - 4\rho_n^{\mathcal{L}}(2(\widehat{r}+1),\beta_n(x))\\
&= \rho_n^{\mathcal{L}}(2(\widehat{r}+1),\beta_n(x))\Big(\frac{c_2}{2}-6\Big) - \frac{c_2}{2}\rho_n^{\mathcal{L}}(2(r^*+1),\beta_n(x)) \qquad (6.0.4)
\end{aligned}$$

where we use in the last but one inequality the isomorphic properties of $\mathcal{L}_{F_{\widehat{r}}}$ for $\widehat{f}_n^{RERM}$: $P_n\mathcal{L}_{\widehat{r},\widehat{f}_n^{RERM}} \ge (1/2)P\mathcal{L}_{\widehat{r},\widehat{f}_n^{RERM}} - 4\rho_n^{\mathcal{L}}(2(\widehat{r}+1),\beta_n(x)) \ge -4\rho_n^{\mathcal{L}}(2(\widehat{r}+1),\beta_n(x))$. On the other side, we have by definition $R_n(\widehat{f}_n^{RERM}) + c_2\rho_n^{\mathcal{L}}(2(\widehat{r}+1),\beta_n(x)) \le R_n(f_{\widehat{r}}^*) + c_2\rho_n^{\mathcal{L}}(2(\widehat{r}+1),\beta_n(x))$ therefore $R_n(\widehat{f}_n^{RERM}) \le R_n(f_{\widehat{r}}^*)$ and thus thanks to (6.0.3) for the level $\widehat{r}$,

$$R(\widehat{f}_n^{RERM}) - R(f_{\widehat{r}}^*) \le 2\big(R_n(\widehat{f}_n^{RERM}) - R_n(f_{\widehat{r}}^*)\big) + 8\rho_n^{\mathcal{L}}(2(\widehat{r}+1),\beta_n(x)) \le 8\rho_n^{\mathcal{L}}(2(\widehat{r}+1),\beta_n(x)).$$

It follows from (6.0.4) and the last inequality that

$$R(\widehat{f}_n^{RERM}) \le R(f^*_{\widehat{r}}) + 8\rho_n^{\mathcal{L}}(2(\widehat{r}+1), \beta_n(x)) \\ \le R(f^*_{r^*}) + \rho_n^{\mathcal{L}}(2(\widehat{r}+1), \beta_n(x))\Big(14 - \frac{c_2}{2}\Big) + \frac{c_2}{2}\rho_n^{\mathcal{L}}(2(r^*+1), \beta_n(x)).$$

For $c_2 = 28$, we obtain $R(\widehat{f}_n^{RERM}) \le R(f^*_{r^*}) + 14\rho_n^{\mathcal{L}}(2(r^*+1), \beta_n(x))$. Therefore, in any case, we obtain $R(\widehat{f}_n^{RERM}) \le R(f^*_{r^*}) + 36\rho_n^{\mathcal{L}}(2(r^*+1), \beta_n(x))$.

We conclude with the same peeling argument of [84, 16] together with the fact that $\widehat{r} \le \beta_n(x)$ on $\Omega_0(x)$ to prove that $\mathbb{P}[\Omega_0(x)] \ge 1 - 5\exp(-x)$.

Proof of Theorem 3.5.4

The proof of Theorem 3.5.4 follows the same lines as the proof of Theorem 3.5.1: First, one needs to find a "trivial" bound on $\mathrm{crit}(\widehat{f}_n^{RERM})$, giving preliminary information on where one must look for that function (this is the role played by the function γ_n). Then, one combines peeling and fixed point arguments to identify the exact location of the RERM.

We begin with the following rough estimate on the criterion of the regularized ERM in the case where no trivial bound $\mathrm{crit}(f) \le C_n, \forall f \in F$ holds but when $r \to \nu^*_\eta(r)$ tends to infinity when r tends to infinity and there exists $K_1 > 0$ such that for every $(r,x) \in \mathbb{R}_+ \times \mathbb{R}_+^*$, $2\rho_n^{\mathcal{E}}(r,x) \le \rho_n^{\mathcal{E}}(K_1(r+1), x)$. Recall that, in this case, for every $x > 0$ and $0 < \eta < 1/2$, we set γ_n to satisfy that

$$\gamma_n(\eta, x) \ge \max\Big[K_1(\mathrm{crit}(f_0)+2), (\nu^*_\eta)^{-1}\Big(2(1+\eta)\Big(3R(f_0) \\ + 4\Big(\frac{B_n(\mathrm{crit}(f_0))K^2x}{n}\Big)^{\frac{1}{2-\beta}} + \frac{KB_n(\mathrm{crit}(f_0))\sqrt{x}}{n} + \frac{2Kb_n(\mathcal{E}_{f_0})x}{n}\Big)\Big)\Big],$$

where f_0 is any fixed function in F (for instance, when $0 \in F$, one may take $f_0 \equiv 0$), and $(\nu^*_\eta)^{-1}$ is the generalized inverse function of ν^*_η.

We also need the following concentration result: for every single function $g \in L_2(P)$ and every $\alpha, x > 0$, with probability greater than $1 - 4\exp(-x)$,

$$P_n g \le (1+\alpha)Pg + K\sqrt{\frac{xPg^2}{n}} + K(1+\alpha^{-1})\frac{b_n(g)x}{n},$$

where $b_n(g) = \|\max_{1\le i\le n} g(Z_i)\|_{\psi_1}$ and, in particular, if there exists some $B_n \ge 0$ and $0 < \beta \le 1$ for which $Pg^2 \le B_n\big(Pg\big)^\beta + B_n^2/n$, then for every $0 < \alpha < 1$ and $x > 0$, with probability greater than $1 - 4\exp(-x)$,

$$P_n g \le (1+2\alpha)Pg + 4\Big(\frac{B_n K^2 x}{n\alpha^\beta}\Big)^{\frac{1}{2-\beta}} + \frac{KB_n\sqrt{x}}{n} + \frac{2Kb_n(g)x}{\alpha n}. \tag{6.0.5}$$

This result follows from the truncation argument of [1] (cf. for instance [P19]).

Lemma 6.0.1 *Assume that $r \to \nu^*_\eta(r)$ tends to infinity when r tends to infinity and that there exists $K_1 > 0$ such that for every $(r,x) \in \mathbb{R}_+ \times \mathbb{R}_+^*$, $2\rho_n^{\mathcal{E}}(r,x) \le \rho_n^{\mathcal{E}}(K_1(r+1), x)$. Then, under the assumptions of Theorem 3.5.4 , for every $x > 0$ and $0 < \eta < 1/2$, with probability greater than $1 - 4\exp(-x)$, $\mathrm{crit}(\widehat{f}_n^{RERM}) \le \gamma_n(\eta, x)$.*

Proof. By the definition of $\widehat{f}_n^{RERM}$,

$$R_n(\widehat{f}_n^{RERM}) + \frac{2}{1+\eta}\rho_n^{\mathcal{E}}\big(\mathrm{crit}(\widehat{f}_n^{RERM}) + 1, x + \log\gamma_n(\eta, x)\big) \\ \leq R_n(f_0) + \frac{2}{1+\eta}\rho_n^{\mathcal{E}}\big(\mathrm{crit}(f_0) + 1, x + \log\gamma_n(\eta, x)\big).$$

Since for every $f \in \mathcal{F}$, $\ell_f(Z) \geq 0$ a.s., then $R_n(\widehat{f}_n^{RERM}) \geq 0$, and thus

$$\rho_n^{\mathcal{E}}\big(\mathrm{crit}(\widehat{f}_n^{RERM}) + 1, x + \log\gamma_n(\eta, x)\big) \\ \leq (1+\eta)R_n(f_0) + \rho_n^{\mathcal{E}}\big(\mathrm{crit}(f_0) + 1, x + \log\gamma_n(\eta, x)\big) \\ \leq \max\Big(2(1+\eta)R_n(f_0), 2\rho_n^{\mathcal{E}}\big(\mathrm{crit}(f_0) + 1, x + \log\gamma_n(\eta, x)\big)\Big).$$

Since $\rho_n^{\mathcal{E}}(r, x) \geq \nu_\eta^*(r), \forall r \geq 0$, one of the following two situations occurs: either

$$\nu_\eta^*(\mathrm{crit}(\widehat{f}_n^{RERM})) \leq 2(1+\eta)R_n(f_0),$$

or, noting that for every $(r, x) \in \mathbb{R}_+ \times \mathbb{R}_+^*$, $2\rho_n^{\mathcal{E}}(r, x) \leq \rho_n^{\mathcal{E}}(K_1(r+1), x)$, then

$$\rho_n^{\mathcal{E}}\big(\mathrm{crit}(\widehat{f}_n^{RERM}) + 1, x + \log\gamma_n(\eta, x)\big) \leq 2\rho_n^{\mathcal{E}}\big(\mathrm{crit}(f_0) + 1, x + \log\gamma_n(\eta, x)\big) \\ \leq \rho_n^{\mathcal{E}}\big(K_1(\mathrm{crit}(f_0) + 2), x + \log\gamma_n(\eta, x)\big),$$

and since $\rho_n^{\mathcal{E}}$ is monotone in r then $\mathrm{crit}(\widehat{f}_n^{RERM}) \leq K_1(\mathrm{crit}(f_0) + 2)$.

Hence, in both cases

$$\mathrm{crit}(\widehat{f}_n^{RERM}) \leq \max\big((\nu_\eta^*)^{-1}\big(2(1+\eta)R_n(f_0)\big), K_1(\mathrm{crit}(f_0) + 2)\big). \tag{6.0.6}$$

On the other hand, according to (6.0.5), with probability greater than $1 - 4\exp(-x)$,

$$R_n(f_0) \leq 3R(f_0) + 4\Big(\frac{B_n(\mathrm{crit}(f_0))K^2 x}{n}\Big)^{\frac{1}{2-\beta}} + \frac{KB_n(\mathrm{crit}(f_0))\sqrt{x}}{n} + \frac{2Kb_n(\mathcal{E}_{f_0})x}{n}.$$

The result follows by plugging the last inequality in (6.0.6). ∎

The next step is to find an "isomorphic" result for $\widehat{f}_n^{RERM}$. The idea is to divide the set given by the trivial estimate on $\mathrm{crit}(\widehat{f}_n^{RERM})$ into level sets and analyze each piece separately.

Lemma 6.0.2 *Under the assumptions of Theorem 3.5.4, for every $x > 0$, with probability greater than $1 - 8\exp(-x)$,*

$$P\mathcal{E}_{\widehat{f}_n^{RERM}} \leq (1+\eta)P_n\mathcal{E}_{\widehat{f}_n^{RERM}} + \rho_n^{\mathcal{E}}\big(\mathrm{crit}(\widehat{f}_n^{RERM}) + 1, x + \log\gamma_n(\eta, x)\big).$$

Proof. Let $\Omega_0(x)$ be the event

$$P\mathcal{E}_{\widehat{f}_n^{RERM}} \geq (1+\eta)P_n\mathcal{E}_{\widehat{f}_n^{RERM}} + \rho_n^{\mathcal{E}}\big(\mathrm{crit}(\widehat{f}_n^{RERM}) + 1, x + \log\gamma_n(\eta, x)\big),$$

and we will show that this event has the desired small probability.

Clearly,

$$\mathbb{P}[\Omega_0(x)] \leq \mathbb{P}\Big[\Omega_0(x) \cap \{\mathrm{crit}(\widehat{f}_n^{RERM}) \leq \gamma_n(\eta, x)\}\Big] + \mathbb{P}\big[\mathrm{crit}(\widehat{f}_n^{RERM}) \geq \gamma_n(\eta, x)\big], \tag{6.0.7}$$

and by Lemma 6.0.1,

$$\mathbb{P}\big[\mathrm{crit}(\widehat{f}_n^{RERM}) \geq \gamma_n(\eta, x)\big] \leq 4\exp(-x). \tag{6.0.8}$$

Recall that $F_i = \{f \in F : \mathrm{crit}(f) \leq i\}, \forall i \in \mathbb{N}$, and since $\rho_n^{\mathcal{E}}$ is monotone in r then

$$\begin{aligned}
&\mathbb{P}\big[\Omega_0(x) \cap \{\mathrm{crit}(\widehat{f}_n^{RERM}) \leq \gamma_n(\eta, x)\}\big] \\
&\leq \sum_{i=0}^{\lfloor \gamma_n(\eta,x)\rfloor} \mathbb{P}\big[\Omega_0(x) \cap \{i \leq \mathrm{crit}(\widehat{f}_n^{RERM}) \leq i+1\}\big] \\
&\leq \sum_{i=0}^{\lfloor \gamma_n(\eta,x)\rfloor} \mathbb{P}\Big[\exists f \in F_{i+1} : P\mathcal{E}_f \geq (1+\eta)P_n\mathcal{E}_f + \rho_n^{\mathcal{E}}(i+1, x + \log\gamma_n(\eta,x))\Big].
\end{aligned} \tag{6.0.9}$$

By Theorem 3.2.3, for every $t > 0$ and $i \in \mathbb{N}$, with probability greater than $1 - 4\exp(-t)$, for every $f \in F_{i+1}$, $P\mathcal{E}_f \leq (1+\eta)P_n\mathcal{E}_f + \rho_n^{\mathcal{E}}(i+1, t)$. In particular,

$$\begin{aligned}
&\mathbb{P}\Big[\exists f \in F_{i+1} : P\mathcal{E}_f \geq (1+\eta)P_n\mathcal{E}_f + \rho_n^{\mathcal{E}}(i+1, x + \log\gamma_n(\eta,x))\Big] \\
&\quad \leq 4\exp\big(-(x + \log\gamma_n(\eta,x))\big).
\end{aligned}$$

Hence, the claim follows, since

$$\begin{aligned}
&\mathbb{P}\big[\Omega_0(x) \cap \{\mathrm{crit}(\widehat{f}_n^{RERM}) \leq \gamma_n(\eta, x)\}\big] \\
&\leq \sum_{i=0}^{\lfloor \gamma_n(\eta,x)\rfloor} 4\exp\big(-(x + \log\gamma_n(\eta,x))\big) \leq 4\exp(-x).
\end{aligned}$$

■

End of the proof of Theorem 3.5.4: Let $x > 0$ and $0 < \eta < 1$. Without loss of generality, we assume that, for the constant K' defined in (6.0.5), there exists $f^{**} \in F$ minimizing the function

$$\begin{aligned}
f \in F \longrightarrow\ & (1+2\eta)P\mathcal{E}_f + \rho_n^{\mathcal{E}}(\mathrm{crit}(f)+1, x + \log\gamma_n(\eta,x)) \\
& + 2K'\Big(\frac{B_n(\mathrm{crit}(f^{**}))x}{\eta n}\Big)^{\frac{1}{2-\beta}} + \frac{2K'B_n(\mathrm{crit}(f^{**}))\sqrt{x}}{n} + \frac{2K'b_n(\mathcal{E}_{f^{**}})x}{\eta n}.
\end{aligned}$$

Let $\Omega^*(x)$ be the event on which

$$P_n\mathcal{E}_{f^{**}} \leq \frac{1+2\eta}{1+\eta}P\mathcal{E}_{f^{**}} + K'\Big(\frac{B_n(\mathrm{crit}(f^{**}))x}{\eta n}\Big)^{\frac{1}{2-\beta}} + \frac{K'B_n(\mathrm{crit}(f^{**}))\sqrt{x}}{n} + \frac{K'b_n(\mathcal{E}_{f^{**}})x}{\eta n}.$$

Since $f^{**} \in F_{\mathrm{crit}(f^{**})}$ then $P\mathcal{E}_{f^{**}}^2 \leq B_n(\mathrm{crit}(f^{**}))\big(P\mathcal{E}_{f^{**}}\big)^{\beta} + B_n^2(\mathrm{crit}(f^{**}))/n$, and by (6.0.5) (applied with $\alpha = \eta/(1+\eta)$), $\mathbb{P}(\Omega^*(x)) \geq 1 - 4\exp(-x)$.

Consider the event $\Omega_0(x)$, on which

$$P\mathcal{E}_{\widehat{f}_n^{RERM}} \leq (1+\eta)P_n\mathcal{E}_{\widehat{f}_n^{RERM}} + \rho_n^{\mathcal{E}}\big(\mathrm{crit}(\widehat{f}_n^{RERM})+1, x + \log\gamma_n(\eta,x)\big),$$

and observe that by Lemma 6.0.2, $\mathbb{P}[\Omega_0(x)] \geq 1 - 8\exp(-x)$. Therefore, on $\Omega_0(x) \cap \Omega^*(x)$, we

have

$$
\begin{aligned}
&P\mathcal{E}_{\widehat{f}_n^{RERM}} + \rho_n^{\mathcal{E}}\big(\mathrm{crit}(\widehat{f}_n^{RERM}) + 1, x + \log\gamma_n(\eta, x)\big) - (1+2\eta)P\mathcal{E}_{f^{**}} \\
\leq &(1+\eta)\big(P_n\mathcal{E}_{\widehat{f}_n^{RERM}} - P_n\mathcal{E}_{f^{**}}\big) + 2\rho_n^{\mathcal{E}}\big(\mathrm{crit}(\widehat{f}_n^{RERM}) + 1, x + \log\gamma_n(\eta, x)\big) \\
&+ 2K'\Big(\frac{B_n(\mathrm{crit}(f^{**}))x}{\eta n}\Big)^{\frac{1}{2-\beta}} + \frac{2K'B_n(\mathrm{crit}(f^{**}))\sqrt{x}}{n} + \frac{2K'b_n(\mathcal{E}_{f^{**}})x}{\eta n} \\
\leq &(1+\eta)\Big(P_n\mathcal{E}_{\widehat{f}_n^{RERM}} + \frac{2}{1+\eta}\rho_n^{\mathcal{E}}\big(\mathrm{crit}(\widehat{f}_n^{RERM}) + 1, x + \log\gamma_n(\eta, x)\big) - P_n\mathcal{E}_{f^{**}} \\
&- \frac{2}{1+\eta}\rho_n^{\mathcal{E}}\big(\mathrm{crit}(f^{**}) + 1, x + \log\gamma_n(\eta, x)\big)\Big) + 2\rho_n^{\mathcal{E}}\big(\mathrm{crit}(f^{**}) + 1, x + \log\gamma_n(\eta, x)\big) \\
&+ 2K'\Big(\frac{B_n(\mathrm{crit}(f^{**}))x}{\eta n}\Big)^{\frac{1}{2-\beta}} + \frac{2K'B_n(\mathrm{crit}(f^{**}))\sqrt{x}}{n} + \frac{2K'b_n(\mathcal{E}_{f^{**}})x}{\eta n} \\
\leq &2\rho_n^{\mathcal{E}}\big(\mathrm{crit}(f^{**}) + 1, x + \log\gamma_n(\eta, x)\big) \\
&+ 2K'\Big(\frac{B_n(\mathrm{crit}(f^{**}))x}{\eta n}\Big)^{\frac{1}{2-\beta}} + \frac{2K'B_n(\mathrm{crit}(f^{**}))\sqrt{x}}{n} + \frac{2K'b_n(\mathcal{E}_{f^{**}})x}{\eta n}
\end{aligned}
$$

where the last inequality follows from the definition of $\widehat{f}_n^{RERM}$. Hence, by the choice of f^{**}, it follows that on $\Omega_1(x) \cap \Omega^*(x)$,

$$
\begin{aligned}
&P\mathcal{E}_{\widehat{f}_n^{RERM}} + \rho_n^{\mathcal{E}}\big(\mathrm{crit}(\widehat{f}_n^{RERM}) + 1, x + \log\gamma_n(\eta, x)\big) \\
&\leq (1+2\eta)P\mathcal{E}_{f^{**}} + 2\rho_n^{\mathcal{E}}\big(\mathrm{crit}(f^{**}) + 1, x + \log\gamma_n(\eta, x)\big) \\
&\qquad + 2K'\Big(\frac{B_n(\mathrm{crit}(f^{**}))x}{\eta n}\Big)^{\frac{1}{2-\beta}} + \frac{2K'B_n(\mathrm{crit}(f^{**}))\sqrt{x}}{n} + \frac{2K'b_n(\mathcal{E}_{f^{**}})x}{\eta n} \\
&= \inf_{f\in F}\Big((1+2\eta)P\mathcal{E}_f + 2\rho_n^{\mathcal{E}}\big(\mathrm{crit}(f) + 1, x + \log\gamma_n(\eta, x)\big) \\
&+ 2K'\Big(\frac{B_n(\mathrm{crit}(f))x}{\eta n}\Big)^{\frac{1}{2-\beta}} + \frac{2K'B_n(\mathrm{crit}(f))\sqrt{x}}{n} + \frac{2K'b_n(\mathcal{E}_f)x}{\eta n}\Big).
\end{aligned}
$$

■

Proof of Theorem 2.8

When $M \leq \sqrt{n}$ then $\bar{f}_n = \widehat{f}_n^{ERM-L}$ and it is proved in [57] that for any $x > 0$, with probability greater than $1 - 2\exp(-x)$,

$$
\begin{aligned}
R(\bar{f}_n) = R(\widehat{f}_n^{ERM-L}) &\leq \min_{f\in\mathrm{span}(F)} R(f) + c_1\max\Big(\frac{M}{n}, \frac{x}{n}\Big) \\
&\leq \min_{f\in\mathrm{conv}(F)} R(f) + c_1\max\Big(\psi_n^{(C)}(M), \frac{x}{n}\Big).
\end{aligned}
$$

We now turn to the case $M > \sqrt{n}$. We recall that $m = \lceil \big(n/\log\big(eM/\sqrt{n}\big)\big)^{1/2}\rceil$ and

$$
\mathcal{C}' = \Big\{\frac{1}{m}\sum_{i=1}^{m}\theta_i : \theta_i \in F\Big\}.
$$

We know by Carl-Maurey empirical method that $\mathcal{C}'$ is a $\big(b/\sqrt{m}\big)$-net of $\mathcal{C} = \mathrm{conv}(F)$ with respect to the $L_2(P_X)$-norm (cf. [38] and [91]). But here we don't use this global approximation result and instead we follow [89] and [103] to prove that

$$\min_{f\in\mathcal{C}'} R(f) \leq \min_{f\in \mathrm{conv}(F)} R(f) + \frac{4b^2}{m}. \tag{6.0.10}$$

Indeed, take $f_{\mathcal{C}}^* \in \mathcal{C}$ such that $R(f_{\mathcal{C}}^*) = \min_{f\in\mathcal{C}} R(f)$ and denote $f_{\mathcal{C}}^* = \sum_{j=1}^M \lambda_j f_j$ where $\sum_{j=1}^M \lambda_j = 1$ and $\lambda_j \geq 0, \forall j = 1,\dots,M$. Consider a random variable $\Theta : \Omega' \to F$ such that $\mathbb{P}'[\Theta = f_j] = \lambda_j, \forall j = 1,\dots,M$. In particular, note that $\mathbb{E}'\theta = f_{\mathcal{C}}^*$. Let $\Theta_1,\dots,\Theta_m$ be m i.i.d. realizations of Θ independent of X and Y. We denote by $\mathbb{E}'_\Theta$ the expectation with respect to $\Theta_1,\dots,\Theta_m$ and by $\mathbb{E}_{X,Y}$ the expectation with respect to X, Y. We have

$$\begin{aligned}\mathbb{E}'_\Theta\mathbb{E}_{X,Y}\left\|\frac{1}{m}\sum_{i=1}^m \Theta_i(X) - Y\right\|_2^2 &= \mathbb{E}_{X,Y}\mathbb{E}'_\Theta \frac{1}{m^2}\sum_{i,j=1}^m (Y-\Theta_i(X))(Y-\Theta_j(X)) \\ &= \frac{m^2-m}{m^2}\mathbb{E}\big(Y - f_{\mathcal{C}}^*(X)\big)^2 + \frac{\mathbb{E}(Y-f_{\mathcal{C}}^*(X))^2}{m} \leq \min_{f\in\mathcal{C}} R(f) + \frac{4b^2}{m}.\end{aligned}$$

In particular, Equation (6.0.10) holds. Moreover, it is proved in [37], that

$$|\mathcal{C}'| \leq \left(\begin{array}{c} M+m-1 \\ m \end{array}\right) \leq \Big(\frac{2eM}{m}\Big)^m. \tag{6.0.11}$$

We consider an optimal aggregation procedure for the (MS) aggregation problem and we run this procedure over the dictionary $\mathcal{C}'$. We denote this procedure by $\tilde{f}_n$ and we set $\bar{f}_n = \tilde{f}_n$. Let $x > 0$, we have (cf. for instance Section 2.4 or 2.5), with probability greater than $1 - 2\exp(-x)$,

$$R(\tilde{f}_n) \leq \min_{f\in\mathcal{C}'} R(f) + c_0 \max\Big(\frac{\log|\mathcal{C}'|}{n}, \frac{x}{n}\Big).$$

Thus, it follows from (6.0.10) and (6.0.11) that for any $x > 0$, with probability greater than $1 - 2\exp(-x)$,

$$R(\bar{f}_n) \leq \min_{f\in\mathcal{C}} R(f) + c_0 \max\Big(\frac{\psi_n^{(C)}(M)}{n}, \frac{x}{n}\Big).$$

Publications

[P1] Karine Bertin and Guillaume Lecué. Selection of variables and dimension reduction in high-dimensional non-parametric regression. *Electron. J. Stat.*, 2:1224–1241, 2008.

[P2] Djalil Chafaï, Olivier Guédon, Guillaume Lecué, and Alain Pajor. *Interaction between Compressed Sensing, Random matrices and High dimensional geometry*. En révision pour le collection "Panoramas et synthéses". SMF, 2011.

[P3] Christophe Chesneau and Guillaume Lecué. Adapting to unknown smoothness by aggregation of thresholded wavelet estimators. *Statist. Sinica*, 19(4):1407–1417, 2009.

[P4] Stéphane Gaïffas and Guillaume Lecué. Optimal rates and adaptation in the single-index model using aggregation. *Electron. J. Stat.*, 1:538–573, 2007.

[P5] Stéphane Gaïffas and Guillaume Lecué. Hyper-sparse optimal aggregation. *Journal of machine learning research*, 12:1813–1833, June 2011.

[P6] Stéphane Gaïffas and Guillaume Lecué. Sharp oracle inequalities for high-dimensional matrix prediction. To appear in IEEE transaction on information theory, 2011.

[P7] Stéphane Gaïffas and Guillaume Lecué. Weighted algorithms for compressed sensing and matrix completion. Submitted, 2011.

[P8] Guillaume Lecué. Lower bounds and aggregation in density estimation. *J. Mach. Learn. Res.*, 7:971–981, 2006.

[P9] Guillaume Lecué. Optimal oracle inequality for aggregation of classifiers under low noise condition. In *Learning theory*, volume 4005 of *Lecture Notes in Comput. Sci.*, pages 364–378. Springer, Berlin, 2006.

[P10] Guillaume Lecué. Optimal rates of aggregation in classification under low noise assumption. *Bernoulli*, 13(4):1000–1022, 2007.

[P11] Guillaume Lecué. Simultaneous adaptation to the margin and to complexity in classification. *Ann. Statist.*, 35(4):1698–1721, 2007.

[P12] Guillaume Lecué. Suboptimality of penalized empirical risk minimization in classification. In Gentile (Eds.) Proceedings. Bshouty, editor, *20th Annual Conference On Learning Theory, COLT07*, volume LNAI 4539, pages 142–156. Springer, 2007.

[P13] Guillaume Lecué. Classification with minimax fast rates for classes of Bayes rules with sparse representation. *Electron. J. Stat.*, 2:741–773, 2008.

[P14] Guillaume Lecué and Shahar Mendelson. Aggregation via empirical risk minimization. *Probability Theory and Related fields*, 145(3–4):591–613, 2004.

[P15] Guillaume Lecué and Shahar Mendelson. Sharper lower bounds on the performance of the empirical risk minimization algorithm. *Bernoulli*, 16(3):605–613, 2010.

[P16] Guillaume Lecué and Shahar Mendelson. General non-exact oracle inequalities in the unbounded case. Under revision in Annals of Statistics, 2011.

[P17] Guillaume Lecué and Shahar Mendelson. On the optimality of the aggregate with exponential weights for low temperatures. To appear in Bernoulli Journal, 2011.

[P18] Guillaume Lecué and Shahar Mendelson. On the optimality of the empirical risk minimization procedure for the Convex aggregation problem. To appear in Annales de l'IHP, 2011.

[P19] Guillaume Lecué and Charles Mitchell. Oracle inequalities for cross-validation type procedures. Submitted, 2010.

[P20] Guillaume Lecué. Empirical risk minimization is optimal for the Convex aggregation problem. Submitted, 2011.

Bibliography

[1] Radosław Adamczak. A tail inequality for suprema of unbounded empirical processes with applications to Markov chains. *Electron. J. Probab.*, 13:no. 34, 1000–1034, 2008.

[2] Pierre Alquier. *Transductive and Inductive Adaptative Inference for Density and Regression Estimation*. PhD thesis, Paris 6, December 2006.

[3] Sylvain Arlot and Peter L. Bartlett. Margin-adaptive model selection in statistical learning. *Bernoulli*, 17(2):687–713, 2011.

[4] Sylvain Arlot and Alain Celisse. A survey of cross-validation procedures for model selection. *Stat. Surv.*, 4:40–79, 2010.

[5] Jean-Yves Audibert. Aggregated estimators and empirical complexity for least square regression. *Ann. Inst. H. Poincaré Probab. Statist.*, 40(6):685–736, 2004.

[6] Jean-Yves Audibert. *PAC-Bayesian Statistical Learning Theory*. PhD thesis, Paris 6, July 2004.

[7] Jean-Yves Audibert. Progressive mixture rules are deviation suboptimal. *Advances in Neural Information Processing Systems*, 20:2, 2007.

[8] Jean-Yves Audibert. Fast learning rates in statistical inference through aggregation. *Ann. Statist.*, 37(4):1591–1646, 2009.

[9] Jean-Yves Audibert. *PAC-Bayesian aggregation and multi-armed bandits*. Habilitation à Diriger des Recherches Université. Paris-Est Marne-la-vallée, December 2010.

[10] Jean-Yves Audibert and Alexandre Tsybakov. Fast learning rates for plug-in classifiers under the margin condition. *The Annals of Statistics*, 35(2):608–633, April 2007.

[11] Francis R. Bach. Consistency of trace norm minimization. *J. Mach. Learn. Res.*, 9:1019–1048, 2008.

[12] Andrew Barron, Lucien Birgé, and Pascal Massart. Risk bounds for model selection via penalization. *Probab. Theory Related Fields*, 113(3):301–413, 1999.

[13] Peter L. Bartlett. Fast rates for estimation error and oracle inequalities for model selection. *Econometric Theory*, 24(2), 2008. (To appear. Was Department of Statistics, U.C. Berkeley Technical Report number 729, 2007).

[14] Peter L. Bartlett, Michael I. Jordan, and Jon D. McAuliffe. Convexity, classification, and risk bounds. *J. Amer. Statist. Assoc.*, 101(473):138–156, 2006.

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

[16] Peter L. Bartlett, Shahar Mendelson, and Joseph Neeman. ℓ_1-regularized linear regression: Persistence and oracle inequalities. *To appear in Probability theory and related fields*, 2009.

[17] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15:1373–1396, 2003.

[18] Peter J. Bickel and Bo Li. *Local polynomial regression on unknown manifolds*, volume 54 of *IMS Lecture Notes-Monograph Series. Complex Datasets and Inverse Problems: Tomography, Networks and Beyond*, pages 177–186. 2007.

[19] Peter J. Bickel, Ya'acov Ritov, and Alexandre B. Tsybakov. Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.*, 37(4):1705–1732, 2009.

[20] Lucien Birgé. Model selection via testing: an alternative to (penalized) maximum likelihood estimators. *Ann. Inst. H. Poincaré Probab. Statist.*, 42(3):273–325, 2006.

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

[22] Stéphane Boucheron, Olivier Bousquet, and Gábor Lugosi. Theory of classification: a survey of some recent advances. *ESAIM Probab. Stat.*, 9:323–375, 2005.

[23] Olivier Bousquet and André Elisseeff. Stability and generalization. *J. Mach. Learn. Res.*, 2(3):499–526, 2002.

[24] Olivier Bousquet, Vladimir Koltchinskii, and Dmitriy Panchenko. Some local measures of complexity of convex hulls and generalization bounds. In *Computational learning theory (Sydney, 2002)*, volume 2375 of *Lecture Notes in Comput. Sci.*, pages 59–73. Springer, Berlin, 2002.

[25] Florentina Bunea and Andrew Nobel. Sequential procedures for aggregating arbitrary estimators of a conditional mean. *IEEE Trans. Inform. Theory*, 54(4):1725–1735, 2008.

[26] Florentina Bunea, Yiyuan She, and Marten Wegkamp. Optimal selection of reduced rank estimators of high-dimensional matrices. *arXiv:1004.2995v2*, 2010.

[27] Florentina Bunea, Alexandre Tsybakov, and Marten Wegkamp. Sparsity oracle inequalities for the Lasso. *Electron. J. Stat.*, 1:169–194 (electronic), 2007.

[28] Florentina Bunea, Alexandre B. Tsybakov, and Marten H. Wegkamp. Aggregation for Gaussian regression. *Ann. Statist.*, 35(4):1674–1697, 2007.

[29] E.J. Candes and Benjamin Recht. Exact matrix completion via convex optimization. *Found. Comput. Math.*, 9(6):717–772, 2009.

[30] Emmanuel J. Candes. The restricted isometry property and its implications for compressed sensing. *C. R. Math. Acad. Sci. Paris*, 346(9-10):589–592, 2008.

[31] Emmanuel J. Candes and Yaniv Plan. Matrix completion with noise. Proceedings of IEEE, 2009.

[32] Emmanuel J. Candes and Yaniv Plan. Tight oracle bounds for low-rank matrix recovery from a minimal number of random measurements. arXiv:1001.0339.

[33] Emmanuel J. Candes and Benjamin Recht. Exact matrix completion via convex optimization. *Found. Comput. Math.*, 9(6):717–772, 2009.

[34] Emmanuel J. Candes and Terence Tao. The power of convex relaxation: Near-optimal matrix completion. arXiv:0903.1476.

[35] Emmanuel J. Candes and Terence Tao. Decoding by linear programming. *IEEE Trans. Inform. Theory*, 51(12):4203–4215, 2005.

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

[37] Bernd Carl. Inequalities of Bernstein-Jackson-type and the degree of compactness of operators in Banach spaces. *Ann. Inst. Fourier (Grenoble)*, 35(3):79–118, 1985.

[38] Bernd Carl and Irmtraud Stephani. *Entropy, compactness and the approximation of operators*, volume 98 of *Cambridge Tracts in Mathematics*. Cambridge University Press, Cambridge, 1990.

[39] Olivier Catoni. *Statistical learning theory and stochastic optimization*, volume 1851 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 2004. Lecture notes from the 31st Summer School on Probability Theory held in Saint-Flour, July 8–25, 2001.

[40] Olivier Catoni. *Pac-Bayesian supervised classification: the thermodynamics of statistical learning*. Institute of Mathematical Statistics Lecture Notes—Monograph Series, 56. Institute of Mathematical Statistics, Beachwood, OH, 2007.

[41] Matthieu Cornec. *Probability bounds for the cross-validation estimate in the context of the statistical learning theory and statistical models applied to economics and finance*. PhD thesis, CREST - Centre de Recherche en économie et statistique, 2009.

[42] Arnak S. Dalalyan and Alexandre B. Tsybakov. Aggregation by exponential weighting and sharp oracle inequalities. In *Learning theory*, volume 4539 of *Lecture Notes in Comput. Sci.*, pages 97–111. Springer, Berlin, 2007.

[43] Luc P. Devroye and T. J. Wagner. Distribution-free performance bounds for potential function rules. *IEEE Trans. Inform. Theory*, 25(5):601–604, 1979.

[44] David L. Donoho and Carrie Grimes. Hessian locally linear embeddings techniques for high-dimensional data. *Proc. Natl. Acad. Sci. USA*, 100:5591–5596, 2003.

[45] David L. Donoho, Iain M. Johnstone, Gérard Kerkyacharian, and Dominique Picard. Wavelet shrinkage: asymptopia? *Journal of the Royal Statistical Society. Series B. Methodological*, 57(2):301–369, 1995.

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

[47] M. Emery, A. Nemirovski, and D. Voiculescu. *Lectures on probability theory and statistics*, volume 1738 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 2000. Lectures from the 28th Summer School on Probability Theory held in Saint-Flour, August 17–September 3, 1998, Edited by Pierre Bernard.

[48] Evarist Giné, Rafał Latała, and Joel Zinn. Exponential and moment inequalities for U-statistics. In *High dimensional probability, II (Seattle, WA, 1999)*, volume 47 of *Progr. Probab.*, pages 13–38. Birkhäuser Boston, Boston, MA, 2000.

[49] Christophe Giraud. Low rank multivariate regression. arXiv:1009.5165.

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

[51] László Györfi, Michael Kohler, Adam Krzyżak, and Harro Walk. *A distribution-free theory of nonparametric regression*. Springer Series in Statistics. Springer-Verlag, New York, 2002.

[52] Peter Hall. Large sample optimality of least squares cross-validation in density estimation. *Ann. Statist.*, 11(4):1156–1174, 1983.

[53] Anatoli Juditsky and Arkadii Nemirovski. Functional aggregation for nonparametric regression. *Ann. Statist.*, 28(3):681–712, 2000.

[54] Anatoli Juditsky, Philippe Rigollet, and Alexandre B. Tsybakov. Learning by mirror averaging. *Ann. Statist.*, 36(5):2183–2206, 2008.

[55] Raghunandan H. Keshavan, Andrea Montanari, and Sewoong Oh. Matrix completion from noisy entries. *J. Mach. Learn. Res.*, 11:2057–2078, 2010.

[56] Vladimir Koltchinskii. Von neumann entropy penalization and low rank matrix estimation. arXiv:1009.2439.

[57] Vladimir Koltchinskii. Local Rademacher complexities and oracle inequalities in risk minimization. *Ann. Statist.*, 34(6):2593–2656, 2006.

[58] Vladimir Koltchinskii. The Dantzig selector and sparsity oracle inequalities. *Bernoulli*, 15(3):799–828, 2009.

[59] Vladimir Koltchinskii. Sparse recovery in convex hulls via entropy penalization. *Ann. Statist.*, 37(3):1332–1359, 2009.

[60] Vladimir Koltchinskii. Sparsity in penalized empirical risk minimization. *Ann. Inst. Henri Poincaré Probab. Stat.*, 45(1):7–57, 2009.

[61] Vladimir Koltchinskii. Local Rademacher Complexities and Oracle Inequalities in Risk Minimization. *Ann. Statist.*, 34(6):1–50, December 2006. 2004 IMS Medallion Lecture.

[62] Vladimir Koltchinskii and Dmitriy Panchenko. Rademacher processes and bounding the risk of function learning. In *High dimensional probability, II (Seattle, WA, 1999)*, volume 47 of *Progr. Probab.*, pages 443–457. Birkhäuser Boston, Boston, MA, 2000.

[63] Vladimir Koltchinskii, Alexandre B. Tsybakov, and Karim Lounici. Nuclear norm penalization and optimal rates for noisy low rank matrix completion. arXiv:1011.6256.

[64] A. P. Korostelëv and A. B. Tsybakov. *Minimax theory of image reconstruction*, volume 82 of *Lecture Notes in Statistics*. Springer-Verlag, New York, 1993.

[65] John Lafferty and Larry Wasserman. Rodeo: sparse, greedy nonparametric regression. *Ann. Statist.*, 36(1):28–63, 2008.

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

[67] W. S. Lee, P. L. Bartlett, and R. C. Williamson. The importance of convexity in learning with squared loss. In *Proceedings of the Ninth Annual Conference on Computational Learning Theory*, pages 140–146. ACM Press, 1996.

[68] Wee Sun Lee, Peter L. Bartlett, and Robert C. Williamson. The importance of convexity in learning with squared loss. *IEEE Trans. Inform. Theory*, 44(5):1974–1980, 1998.

[69] Oleg V. Lepski. On a problem of adaptive estimation in Gaussian white noise. *Theory of Probability and its Applications*, 35(3):454–466, 1990.

[70] Oleg V. Lepski. On problems of adaptive estimation in white gaussian noise. *Advances in Soviet Mathematics*, 12:87–106, 1992.

[71] Gilbert Leung and Andrew R. Barron. Information theory and mixing least-squares regressions. *IEEE Trans. Inform. Theory*, 52(8):3396–3410, 2006.

[72] Elizaveta Levina and Peter J. Bickel. *Maximum Likelihood Estimation of Intrinsec Dimension*, volume 17 of *Advances in NIPS*. 2005.

[73] Karim Lounici. Sup-norm convergence rate and sign concentration property of Lasso and Dantzig estimators. *Electron. J. Stat.*, 2:90–102, 2008.

[74] Gábor Lugosi and Marten Wegkamp. Complexity regularization via localized random penalties. *Ann. Statist.*, 32(4):1679–1697, 2004.

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

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

[77] Pascal Massart and Élodie Nédélec. Risk bounds for statistical learning. *Ann. Statist.*, 34(5):2326–2366, 2006.

[78] Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the lasso. *Ann. Statist.*, 34(3):1436–1462, 2006.

[79] Nicolai Meinshausen and Bin Yu. Lasso-type recovery of sparse representations for high-dimensional data. *Ann. Statist.*, 37(1):246–270, 2009.

[80] Shahar Mendelson. Lower bounds for the empirical minimization algorithm. *IEEE Trans. Inform. Theory*, 54(8):3797–3803, 2008.

[81] Shahar Mendelson. Lower bounds for the empirical minimization algorithm. *IEEE Trans. Inform. Theory*, 54(8):3797–3803, 2008.

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

[83] Shahar Mendelson. Empirical processes with a bounded ψ_1 diameter. *Geom. Funct. Anal.*, 20(4):988–1027, 2010.

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

[85] Shahar Mendelson, Alain Pajor, and Nicole Tomczak-Jaegermann. Reconstruction and subgaussian operators in asymptotic geometric analysis. *Geom. Funct. Anal.*, 17(4):1248–1282, 2007.

[86] Vitali D. Milman and Gideon Schechtman. *Asymptotic theory of finite-dimensional normed spaces*, volume 1200 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 1986. With an appendix by M. Gromov.

[87] B. K. Natarajan. Sparse approximate solutions to linear systems. *SIAM J. Comput.*, 24(2):227–234, 1995.

[88] Sahand Negahban and Martin J. Wainwright. Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. arXiv:1009.2118.

[89] Arkadi Nemirovski. Topics in non-parametric statistics. In *Lectures on probability theory and statistics (Saint-Flour, 1998)*, volume 1738 of *Lecture Notes in Math.*, pages 85–277. Springer, Berlin, 2000.

[90] Valentin V. Petrov. *Limit theorems of probability theory*, volume 4 of *Oxford Studies in Probability*. The Clarendon Press Oxford University Press, New York, 1995.

[91] Gilles Pisier. Remarques sur un résultat non publié de B. Maurey. In *Seminar on Functional Analysis, 1980–1981*, pages Exp. No. V, 13. École Polytech., Palaiseau, 1981.

[92] Benjamin Recht. A simpler approach to matrix completion. arXiv:0910.0651.

[93] R. Tyrrell Rockafellar. *Convex analysis*. Princeton Landmarks in Mathematics. Princeton University Press, Princeton, NJ, 1997. Reprint of the 1970 original, Princeton Paperbacks.

[94] Angelika Rohde and Alexandre Tsybakov. Estimation of high-dimensional low-rank matrice. To appear in Ann. Statist.. arXiv:0912.5338.

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

[96] Jun Shao. Linear model selection by cross-validation. *J. Amer. Statist. Assoc.*, 88(422):486–494, 1993.

[97] Ingo Steinwart and Andreas Christmann. *Support vector machines*. Information Science and Statistics. Springer, New York, 2008.

[98] Charles J. Stone. An asymptotically optimal window selection rule for kernel density estimates. *Ann. Statist.*, 12(4):1285–1297, 1984.

[99] Mervyn Stone. Cross-validatory choice and assessment of statistical predictions. *J. Roy. Statist. Soc. Ser. B*, 36:111–147, 1974. With discussion by G. A. Barnard, A. C. Atkinson, L. K. Chan, A. P. Dawid, F. Downton, J. Dickey, A. G. Baker, O. Barndorff-Nielsen, D. R. Cox, S. Giesser, D. Hinkley, R. R. Hocking, and A. S. Young, and with a reply by the authors.

[100] Michel Talagrand. New concentration inequalities in product spaces. *Invent. Math.*, 126(3):505–563, 1996.

[101] Michel Talagrand. *The generic chaining*. Springer Monographs in Mathematics. Springer-Verlag, Berlin, 2005. Upper and lower bounds of stochastic processes.

[102] Robert Tibshirani. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58(1):267–288, 1996.

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

[104] Alexandre B. Tsybakov. Robust reconstruction of functions by the local-approximation method. *Problems of Information Transmission*, 22:133–146, 1986.

[105] Alexandre B. Tsybakov. Optimal rates of aggregation. *Computational Learning Theory and Kernel Machines. B.Schölkopf and M.Warmuth, eds. Lecture Notes in Artificial Intelligence*, 2777:303–313, 2003. Springer, Heidelberg.

[106] Alexandre B. Tsybakov. *Introduction à l'estimation non-paramétrique*, volume 41 of *Mathématiques & Applications (Berlin) [Mathematics & Applications]*. Springer-Verlag, Berlin, 2004.

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

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

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

[110] Sara A. van de Geer. High-dimensional generalized linear models and the lasso. *Ann. Statist.*, 36(2):614–645, 2008.

[111] Mark J. van der Laan, Sandrine Dudoit, and Aad W. van der Vaart. The cross-validated adaptive epsilon-net estimator. *Statist. Decisions*, 24(3):373–395, 2006.

[112] Aad W. van der Vaart, Sandrine Dudoit, and Mark J. van der Laan. Oracle inequalities for multi-fold cross validation. *Statist. Decisions*, 24(3):351–371, 2006.

[113] Aad W. van der Vaart and Jon A. Wellner. *Weak convergence and empirical processes*. Springer Series in Statistics. Springer-Verlag, New York, 1996. With applications to statistics.

[114] Vladimir Vapnik. *Estimation of dependences based on empirical data*. Springer Series in Statistics. Springer-Verlag, New York, 1982. Translated from the Russian by Samuel Kotz.

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

[116] Vladimir N. Vapnik. *The nature of statistical learning theory*. Statistics for Engineering and Information Science. Springer-Verlag, New York, second edition, 2000.

[117] Yuhong Yang. Combining different procedures for adaptive regression. *J. Multivariate Anal.*, 74(1):135–161, 2000.

[118] Yuhong Yang. Mixing strategies for density estimation. *Ann. Statist.*, 28(1):75–87, 2000.

[119] Yuhong Yang. Adaptive regression by mixing. *J. Amer. Statist. Assoc.*, 96(454):574–588, 2001.

[120] Yuhong Yang. Aggregating regression procedures to improve performance. *Bernoulli*, 10(1):25–47, 2004.

[121] Tong Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *Ann. Statist.*, 32(1):56–85, 2004.

[122] Tong Zhang. Some sharp performance bounds for least squares regression with L_1 regularization. *Ann. Statist.*, 37(5A):2109–2144, 2009.

[123] Hui Zou. The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.*, 101(476):1418–1429, 2006.

[124] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 67(2):301–320, 2005.