Estimation bounds and sharp oracle inequalities of regularized procedures with Lipschitz loss functions

Pierre Alquier[1,3,4] , Vincent Cottet[1,3,4] , Guillaume Lecué[2,3,4]
(1) CREST, ENSAE, Université Paris Saclay
(2) CREST, CNRS, Université Paris Saclay

February 6, 2017

Abstract

We obtain estimation error rates and sharp oracle inequalities for regularization procedures of the form

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

when $\|\cdot\|$ is any norm, F is a convex class of functions and ℓ is a Lipschitz loss function satisfying a Bernstein condition over F. We explore both the bounded and subgaussian stochastic frameworks for the distribution of the $f(X_i)$'s, with no assumption on the distribution of the Y_i's. The general results rely on two main objects: a complexity function, and a sparsity equation, that depend on the specific setting in hand (loss ℓ and norm $\|\cdot\|$).

As a proof of concept, we obtain minimax rates of convergence in the following problems: 1) matrix completion with any Lipschitz loss function, including the hinge and logistic loss for the so-called 1-bit matrix completion instance of the problem, and quantile losses for the general case, which enables to estimate any quantile on the entries of the matrix; 2) logistic LASSO and variants such as the logistic SLOPE; 3) kernel methods, where the loss is the hinge loss, and the regularization function is the RKHS norm.

1 Introduction

Many classification and prediction problems are solved in practice by regularized empirical risk minimizers (RERM). The risk is measured by a loss function and the quadratic loss function is the most popular function for regression. It has been extensively studied (cf. [39, 31] among others). Still many other loss functions are popular among practitioners and are indeed extremely useful in specific situations.

First, let us mention the quantile loss in regression problems. The 0.5-quantile loss (also known as absolute or L_1 loss) is known to provide an indicator of conditional central tendency more robust to outliers than the quadratic loss. An alternative to the absolute loss for robustification is provided by the Huber loss. On the other hand, general quantile losses are used to estimate conditional quantile functions and are extremely useful to build confidence intervals and measures of risk, like *Values at Risk* (VaR) in finance.

Let us now turn to classification problems. The natural loss in this context, the so called 0/1 loss, leads very often to computationally intractable estimators. Thus, it is usually replaced by a convex loss function, such as the hinge loss or the logistic loss. A thorough study of convex loss functions in classification can be found in [66].

[3]Email: {pierre.alquier}, {vincent.cottet}, {guillaume.lecue}@ensae.fr

[4]The authors gratefully acknowledge financial support from Labex ECODEC (ANR - 11-LABEX-0047). Author n. 1 also acknowledge financial support from the research programme *New Challenges for New Data* from LCL and GENES, hosted by the *Fondation du Risque*. Author n. 3 acknowledge financial support from the "Chaire Economie et Gestion des Nouvelles Données", under the auspices of Institut Louis Bachelier, Havas-Media and Paris-Dauphine.

All the aforementioned loss functions (quantile, Huber, hinge and logistic) share a common property: they are Lipschitz functions. This motivates a general study of RERM with any Lipschitz loss. Note that some examples were already studied in the literature: the $\|\cdot\|_1$-penalty with a quantile loss was studied in [8] under the name "quantile LASSO" while the same penalty with the logistic loss was studied in [64] under the name "logistic LASSO" (cf. [62]). The ERM strategy with Lipschitz proxys of the 0/1 loss are studied in [29]. The loss functions we will consider in the examples of this paper are reminded below:

1. **hinge loss:** $\ell(y', y) = (1 - yy')_+ = \max(0, 1 - yy')$ for every $y \in \{-1, +1\}, y' \in \mathbb{R}$,
2. **logistic loss:** $\ell(y', y) = \log(1 + \exp(-yy'))$ for every $y \in \{-1, +1\}, y' \in \mathbb{R}$;
3. **quantile regression loss:** for some parameter $\tau \in (0, 1)$, $\ell(y', y) = \rho_\tau(y - y')$ for every $y \in \mathbb{R}, y' \in \mathbb{R}$ where $\rho_\tau(z) = z(\tau - I(z \leq 0))$ for all $z \in \mathbb{R}$.

The two main theoretical results of the paper, stated in Section 2, are general in the sense that they do not rely on a specific loss function or a specific regularization norm. We develop two different settings that handle different assumptions on the design. In the first one, we assume that the family of predictors is subgaussian; in the second setting we assume that the predictors are uniformly bounded, this setting is well suited for classification tasks, including the 1-bit matrix completion problem. The rates of convergence rely on quantities that measure the complexity of the model and the size of the subdifferential of the norm.

To be more precise, the method works for any regularization function as long as it is a norm. If this norm has some sparsity inducing power, like the ℓ_1 or nuclear norms, thus the statistical bounds depend on the underlying sparsity around the oracle because the subdifferential is large. We refer these bounds as *sparsity dependent bounds*. If the norm does not induce sparsity, it is still possible to derive bounds that are now depending on the norm of the oracle because the subdifferential of the norm is very large in 0. We call it *norm dependent bounds* (aka "complexity dependent bounds" in [40]).

We study many applications that give new insights on diverse problems: the first one is a classification problem with logistic loss and LASSO or SLOPE regularizations. We prove that the rate of the SLOPE estimator is minimax in this framework. The second one is about matrix completion. We derive new excess risk bounds for the 1-bit matrix completion issue with both logistic and hinge loss. We also study the quantile loss for matrix completion and prove it reaches sharp bounds. We show several examples in order to assess the general methods as well as simulation studies. The last example involves the SVM and proves that "classic" regularization method with no special sparsity inducing power can be analyzed in the same way as sparsity inducing regularization methods.

A remarkable fact is that no assumption on the output Y is needed (while most results for the quadratic loss rely on an assumption of the tails of the distribution of Y). Neither do we assume any statistical model relating the "output variable" Y to the "input variable" X.

Mathematical background and notations. The observations are N i.i.d pairs $(X_i, Y_i)_{i=1}^N$ where $(X_i, Y_i) \in \mathcal{X} \times \mathcal{Y}$ are distributed according to P. We consider the case where $\mathcal{Y}$ is a subset of $\mathbb{R}$ and let μ denote the marginal distribution of X_i. Let L_2 be the set of real valued functions f defined on $\mathcal{X}$ such that $\mathbb{E}f(X)^2 < +\infty$ where the distribution of X is μ. In this space, we define the L_2-norm as $\|f\|_{L_2} = (\mathbb{E}f(X)^2)^{1/2}$ and the L_∞ norm such that $\|f\|_{L_\infty} = \text{esssup}(|f(X)|)$. We consider a set of predictors $F \subseteq E$, where E is a subspace of L_2 and $\|\cdot\|$ is a norm over E (actually, in some situations we will simply have $F = E$, but in some natural examples we will consider bounded set of predictors, in the sense that $\sup_{f \in F} \|f\|_{L_\infty} < \infty$, which implies that F cannot be a subspace of L_2).

For every $f \in F$, the loss incurred when we predict $f(x)$, while the true output / label is actually y, is measured using a loss function ℓ: $\ell(f(x), y)$. For short, we will also use the notation $\ell_f(x, y) = \ell(f(x), y)$ the loss function associated with f. In this work, we focus on loss functions that are nonnegative, and Lipschitz, in the following sense.

Assumption 1.1 (Lipschitz loss function)**.** *For every $f_1, f_2 \in F$, $x \in \mathcal{X}$ and $y \in \mathbb{R}$, we have*

$$\big|\ell(f_1(x), y) - \ell(f_2(x), y)\big| \leq |f_1(x) - f_2(x)|.$$

Note that we chose a Lipschitz constant equal to one in Assumption 1.1. This can always be achieved by a proper normalization of the loss function. We define the oracle predictor as

$$f^* \in \operatorname*{argmin}_{f \in F} P\ell_f \text{ where}^{1} \; P\ell_f = \mathbb{E}\ell_f(X,Y)$$

and (X,Y) is distributed like the (X_i,Y_i)'s. The objective of machine learning is to provide an estimator $\hat{f}$ that predicts almost as well as f^*. We usually formalize this notion by introducing the excess risk $\mathcal{E}(f)$ of $f \in F$ by

$$\mathcal{L}_f = \ell_f - \ell_{f^*} \text{ and } \mathcal{E}(f) = P\mathcal{L}_f.$$

Thus we consider the estimator of the form

$$\hat{f} \in \operatorname*{argmin}_{f \in F} \{P_N \ell_f + \lambda \|f\|\} \tag{1}$$

where $P_N \ell_f = (1/N)\sum_{i=1}^N \ell_f(X_i,Y_i)$ and λ is a regularization parameter to be chosen. Such estimators are usually called Regularized Empirical Risk Minimization procedure (RERM).

For the rest of the paper, we will use the following notations: let rB and rS denote the radius r ball and sphere for the norm $\|\cdot\|$, i.e. $rB = \{f \in E : \|f\| \leq r\}$ and $rS = \{f \in E : \|f\| = r\}$. For the L_2-norm, we write $rB_{L_2} = \{f \in L_2 : \|f\|_{L_2} \leq r\}$ and $rS_{L_2} = \{f \in L_2 : \|f\|_{L_2} = r\}$ and so on for the other norms.

Even though our results are valid in the general setting introduced above, we will develop the examples mainly in two directions that we will refer to *vector* and *matrix*. The *vector* case involves $\mathcal{X}$ as a subset of $\mathbb{R}^p$; we then consider the class of linear predictors, i.e. $E = \{\langle t, \cdot \rangle, t \in \mathbb{R}^p\}$. In this case, we denote for $q \in [1, +\infty]$, the l_q-norm in $\mathbb{R}^p$ as $\|\cdot\|_{l_q}$. The *matrix* case is also referred as the trace regression model: X is a random matrix in $\mathbb{R}^{m \times T}$ and we consider the class of linear predictors $E = \{\langle M, \cdot \rangle, M \in \mathbb{R}^{m \times T}\}$ where $\langle A, B \rangle = \mathrm{Trace}(A^\top B)$ for any matrices A, B in $\mathbb{R}^{m \times T}$. The norms we consider are then, for $q \in [1, +\infty[$, the Schatten-q-norm for a matrix: $\forall M \in \mathbb{R}^{m \times T}, \|M\|_{S_q} = (\sum \sigma_i(M)^q)^{1/q}$ where $\sigma_1(M) \geq \sigma_2(M) \geq \cdots$ is the family of the singular values of M. The Schatten-1 norm is also called trace norm or nuclear norm. The Schatten-2 norm is also known as the Frobenius norm. The S_∞ norm, defined as $\|M\|_{S_\infty} = \sigma_1(M)$ is known as the operator norm.

The notation $\mathbf{C}$ will be used to denote positive constants, that might change from one instance to the other. For any real numbers a, b, we write $a \lesssim b$ when there exists a positive constant $\mathbf{C}$ such that $a \leq \mathbf{C}b$. When $a \lesssim b$ and $b \lesssim a$, we write $a \sim b$.

Proof of Concept. We now present briefly one of the outputs of our global approach: an oracle inequality for the 1-bit matrix completion problem with hinge loss (we refer the reader to Section 4 for a detailed exposition of this example). While the general matrix completion problem has been extensively studied in the case of a quadratic loss, see [32, 39] and the references therein, we believe that there is no satisfying solution to the so-called 1-bit matrix completion problem, that is for binary observations $\mathcal{Y} = \{-1, +1\}$. Indeed, the attempts in [55, 18] to use the hinge loss did not lead to rank dependent learning rates. On the other hand, [34] studied RERM procedure using a statistical modeling approach and the logistic loss. While these authors prove optimal rates of convergence of their estimator with respect to the Frobenius norm, the excess classification risk, is not studied in their paper. However we believe that the essence of machine learning is to focus on this quantity – it is directly related to the average number of errors in prediction.

From now on we assume that $\mathcal{Y} = \{-1, +1\}$ and we consider the *matrix* framework. In matrix completion, we write the observed location as a mask matrix X: it is an element of the canonical basis

[1] Note that without any assumption on Y it might be that $P\ell_f = \mathbb{E}\ell_f(X,Y) = \infty$ for any $f \in F$. Our results remain valid in this case, but it is no longer possible to use the definition $f^* \in \operatorname{argmin}_{f \in F} P\ell_f$. A general definition is as follows: fix any $f_0 \in F$. Note that for any $f \in F$, $\mathbb{E}[\ell_f(X,Y) - \ell_{f_0}(X,Y)]] \leq \mathbb{E}|(f - f_0)(X)| < \infty$ under the assumptions on F that will be stated in Section 2. It is then possible to define f^* as any minimizer of $\mathbb{E}[\ell_f(X,Y) - \ell_{f_0}(X,Y)]]$. This definition obviously coincides with the defintion $f^* \in \operatorname{argmin}_{f \in F} P\ell_f$ when $P\ell_f$ is finite for some $f \in F$.

$(E_{1,1}, \cdots, E_{m,T})$ of $\mathbb{R}^{m\times T}$ where for any $(p,q) \in \{1,\dots,m\} \times \{1,\dots,T\}$ the entry of $E_{p,q}$ is 0 everywhere except for the (p,q)-th entry where it equals to 1. We assume that there are constants $0 < \underline{c} \leq \bar{c} < \infty$ such that, for any (p,q), $\underline{c}/(mT) \leq \mathbb{P}(X = E_{p,q}) \leq \bar{c}/(mT)$ (this extends the uniform sampling distribution for which $\underline{c} = \bar{c} = 1$). These assumptions are encompassed in the following definition.

Assumption 1.2 (Matrix completion design)**.** *The sample size N is in $\{\min(m,T), \dots, \max(m,T)^2\}$ and X takes value in the canonical basis $(E_{1,1}, \cdots, E_{m,T})$ of $\mathbb{R}^{m\times T}$. There are positive constants $\underline{c}, \bar{c}$ such that for any $(p,q) \in \{1,\dots,m\} \times \{1,\dots,T\}$,*

$$\underline{c}/(mT) \leq \mathbb{P}(X = E_{p,q}) \leq \bar{c}/(mT).$$

A predictor can be seen, for this problem, as the natural inner product with a real $m \times T$ matrix: $f(X) = \langle M, X\rangle = \mathrm{Tr}(X^\top M)$. The class F that we consider in Section 4 is the set of linear predictors where every entry of the matrix is bounded: $F = \{\langle \cdot, M\rangle : M \in bB_\infty\}$ where $bB_\infty = \{M = (M_{pq}) : \max_{p,q} |M_{pq}| \leq b\}$ for a specific b. This set is very common in matrix completion studies. But it is especially natural in this setting: indeed, the Bayes classifier, defined by $\overline{M} = \mathrm{argmin}_{M\in\mathbb{R}^{m\times T}} \mathbb{E}\left(1 - Y\langle X, M\rangle\right)_+$, has entries in $[-1,1]$. So, by taking $b = 1$ in the definition of F, we ensure that the oracle $M^* = \mathrm{argmin}_{M\in\mathbb{R}^{m\times T}} \mathbb{E}\left(1 - Y\langle X, M\rangle\right)_+$ satisfies $M^* = \overline{M}$, so there would be no point in taking $b > 1$. We will therefore consider the following RERM (using the hinge loss)

$$\widehat{M} \in \operatorname*{argmin}_{M\in B_\infty} \left(\frac{1}{N} \sum_{i=1}^{N} \left(1 - Y_i\langle X_i, M\rangle\right)_+ + \lambda \|M\|_{S_1} \right) \tag{2}$$

where $\lambda > 0$ is some parameter to be chosen. We prove in Section 4 the following result.

Theorem 1.1. *Assume that Assumption 1.2 holds and there is $\tau > 0$ such that, for any $(p,q) \in \{1,\dots,m\}\times\{1,\dots,T\}$,*

$$\left|\overline{M}_{p,q} - \frac{1}{2}\right| \geq \tau. \tag{3}$$

There is a $c_0(\underline{c}, \bar{c}) > 0$, that depends only on $\underline{c}$ and $\bar{c}$, and that is formally introduced in Section 4 below, such that if one chooses the regularization parameter

$$\lambda = c_0(\underline{c}, \bar{c}) \sqrt{\frac{\log(m+T)}{N \min(m,T)}}$$

then, with probability at least

$$1 - \mathbf{C}\exp\left(-\mathbf{C}\mathrm{rank}(\overline{M}) \max(m,T) \log(m+T)\right), \tag{4}$$

the RERM estimator $\widehat{M}$ defined in (2) satisfies for every $1 \leq p \leq 2$,

$$\frac{1}{(mT)^{\frac{1}{p}}} \left\|\widehat{M} - \overline{M}\right\|_{S_p} \leq \mathbf{C}\mathrm{rank}(\overline{M})^{\frac{1}{p}} \sqrt{\frac{\log(m+T)}{N}} \frac{\max(m,T)^{1-\frac{1}{p}}}{\min(m,T)^{\frac{1}{p}-\frac{1}{2}}}$$

and as a special case for $p = 2$,

$$\frac{1}{\sqrt{mT}} \left\|\widehat{M} - \overline{M}\right\|_{S_2} \leq \mathbf{C}\sqrt{\frac{\mathrm{rank}(\overline{M}) \max(m,T) \log(m+T)}{N}} \tag{5}$$

and its excess hinge risk is such that

$$\mathcal{E}_{hinge}(\widehat{M}) = \mathbb{E}(1 - Y\langle X, \widehat{M}\rangle)_+ - \mathbb{E}(1 - Y\langle X, \overline{M}\rangle)_+ \leq \mathbf{C}\frac{\mathrm{rank}(\overline{M}) \max(m,T) \log(m+T)}{N}$$

where the notation $\mathbf{C}$ is used for constants that might change from one instance to the other but depend only on $\underline{c}$, $\bar{c}$ and τ.

The excess hinge risk bound from Theorem 1.1 is of special interest as it can be related to the classic excess 0/1 risk. The excess 0/1 risk of a procedure is really the quantity we want to control since it measures the difference between the average number of mistakes of a procedure with the best possible theoretical classification rule. Indeed, let us define the 0/1 risk of M by $R_{0/1}(M) = \mathbb{P}[Y \neq \text{sign}(\langle M, X\rangle)]$. It is clear that $\overline{M} \in \text{argmin}_{M\in\mathbb{R}^{m\times T}} R_{0/1}(M)$. Then, it follows from Theorem 2.1 in [66] that for some universal constant $c > 0$, for every $M \in \mathbb{R}^{m\times T}$,

$$R_{0/1}(M) - \inf_{M\in B_\infty} R_{0/1}(M) \leq c\mathcal{E}_{hinge}(M).$$

Therefore, the RERM from (2) for the choice of regularization parameter λ as in Theorem 1.1 satisfies with probability larger than in (4),

$$\mathcal{E}_{0/1}(\widehat{M}) = R_{0/1}(\widehat{M}) - \inf R_{0/1}(M) \leq \mathbf{C}\frac{\text{rank}(\overline{M})\max(m,T)\log(m+T)}{N} \tag{6}$$

where $\mathbf{C}$ depends on c, $\underline{c}$, $\bar{c}$ and τ. This yields a bound on the average of excess number of mistakes of $\widehat{M}$. To our knowledge such a prediction bound was not available in the literature on the 1-bit matrix completion problem. Let us compare Theorem 1.1 to the main result in [34]. In [34], the authors focus on the estimation error $\|\widehat{M} - M^*\|_{S_2}$, which seems less relevant for practical applications. In order to connect such a result to the excess classification risk, one can use the results in [66] and in this case, the best bound that can be derived is of the order of $\sqrt{\text{rank}(M^*)\max(m,T)/N}$. Note that other authors focused on the classification error: [55] proved an excess error bound, but the bound does not depend on the rank of the oracle. The rate $\text{rank}(M^*)\max(m,T)/N$ derived from Theorem 1.1 for the 0/1-classification excess risk was only reached in [18], but in the very restrictive noiseless setting, which is equivalent to $\inf_M R_{0/1}(M) = 0$.

We hope that this example convinced the reader of the practical interest of the general study of $\hat{f}$ in (1). The rest of the paper is organized as follows. In Section 2 we introduce the concepts necessary to the general study of (1): namely, a complexity parameter, and a sparsity parameter. Thanks to these parameters, we define the assumptions necessary to our general results: the Bernstein condition, which is classic in learning theory to obtain fast rates [39], and a stochastic assumption on F (subgaussian, or bounded). The general results themselves are eventually presented. The remaining sections are devoted to applications of our results to different estimation methods: the logistic LASSO and logistic SLOPE in Section 3, matrix completion in Section 4 and Support Vector Machines (SVM) in Section 5. For matrix completion, the optimality of the rates for the logistic and the hinge loss, that were not known, is also derived. In Section 6 we discuss the Bernstein condition for the three main loss functions of interest: hinge, logistic and quantile.

2 Theoretical Results

2.1 Applications of the main results: the strategy

The two main theorems in Sections 2.5 and 2.6 below are general in the sense that they allow the user to deal with any (nonnegative) Lipschitz loss function and any norm for regularization, but they involve quantities that depend on the loss and the norm. The aim of this Section is first to provide the definition of these objects and some hints on their interpretation, through examples. The theorems are then stated in both settings. Basically, the assumptions for the theorems are of three types:

1. the so-called Bernstein condition, which is a quantification of the identifiability condition. It basically tells how the excess risk $\mathcal{E}(f) = P\mathcal{L}_f = P(\ell_f - \ell_{f^*})$ is related to the L_2 norm $\|f - f^*\|_{L_2}$.

2. a stochastic assumption on the distribution of the $f(X)$'s for $f \in F$. In this work, we consider both a subgaussian assumption and a uniform boundedness assumption. Analysis of the two setups

differ only on the way the "statistical complexity of F" is measured (cf. below the functions $r(\cdot)$ in Definition 8.1 and Definition 8.2).

3. finally, we introduce a sparsity parameter as in [39]. It reflects how the norm $\|\cdot\|$ used as a regularizer can induce sparsity - for example, think of the "sparsity inducing power" of the l_1-norm used to construct the LASSO estimator.

Given a scenario, that is a loss function ℓ, a random design X, a convex class F and a regularization norm, statistical results (exact oracle inequalities and estimation bounds w.r.t. the L_2 and regularization norms) for the associated regularized estimator together with the choice of the regularization parameter follow from the derivation of the three parameters (κ, r, ρ^*) as explained in the next box together with Theorem 2.1 and Theorem 2.2.

Application of the main results

1. find the **Bernstein parameter** $\kappa \geq 1$ and $A > 0$ associated to the loss and the class F;
2. compute the **Complexity function**

$$r(\rho) = \left[\frac{A\rho \text{comp}(B)}{\sqrt{N}}\right]^{1/2\kappa}$$

where $\text{comp}(B)$ is defined either through the Gaussian mean width $w(B)$, in the subgaussian case, or the Rademacher complexity $\text{Rad}(B)$, in the bounded case;

3. Compute the sub-differential $\partial\|\cdot\|(f^*)$ of $\|\cdot\|$ at the oracle f^* (or in the neighborhood $f^* + (\rho/20)B$ for approximately sparse oracles) and solve the **sparsity equation** "find ρ^* such that $\Delta(\rho^*) \geq 4\rho^*/5$".
4. Apply Theorem 2.1 in the subgaussian framework and Theorem 2.2 in the bounded framework. In each case, with large probability,

$$\left\|\hat{f} - f^*\right\| \leq \rho^*, \ \left\|\hat{f} - f^*\right\|_{L_2} \leq r(2\rho^*) \text{ and } \mathcal{E}(\hat{f}) \leq \mathbf{C}\left[r(2\rho^*)\right]^{2\kappa}.$$

For the sake of simplicity, we present the two settings in different subsections with both the exact definition of the complexity function and the theorem. As the sparsity equation is the same in both settings, we define it before even though it involves the complexity function.

2.2 The Bernstein condition

The first assumption needed is called *Bernstein* assumption and is very classic in order to deal with Lipschitz loss.

Assumption 2.1 (Bernstein condition)**.** *There exists $\kappa \geq 1$ and $A > 0$ such that for every $f \in F$, $\|f - f^*\|_{L_2}^{2\kappa} \leq AP\mathcal{L}_f$.*

The most important parameter is κ and will be involved in the rate of convergence. As usual fast rates will be derived when $\kappa = 1$. In many situations, this assumption is satisfied and we present various cases in Section 6. In particular, we prove that it is satisfied with $\kappa = 1$ for the logistic loss in both bounded and Gaussian framework, and we exhibit explicit conditions to ensure that Assumption 2.1 holds for the hinge and the quantile loss functions.

We call Assumption 2.1 a *Bernstein condition* following [7] and that it is different from the margin assumption from [44, 60]: in the so-called margin assumption, the oracle f^* in F is replaced by the

minimizer $\overline{f}$ of the risk function $f \to P\ell_f$ over all measurable functions f, sometimes called the Bayes rules. We refer the reader to Section 6 and to the discussions in [37] and Chapter 1.3 in [36] for more details on the difference between the margin assumption and the Bernstein condition.

Remark 2.1. *The careful reader will actually realize that the proof of Theorem 2.1 and Theorem 2.2 requires only a weaker version of this assumption, that is: there exists $\kappa \geq 1$ and $A > 0$ such that for every $f \in \mathcal{C}$, $\|f - f^*\|_{L_2}^{2\kappa} \leq AP\mathcal{L}_f$, where $\mathcal{C}$ is defined in terms of the complexity function $r(\cdot)$ and the sparsity parameter ρ^* to be defined in the next subsections,*

$$\mathcal{C} := \left\{ f \in F : \|f - f^*\|_{L_2} \geq r(2\|f - f^*\|) \text{ and } \|f - f^*\| \geq \rho^* \right\}. \tag{7}$$

Note that the set $\mathcal{C}$ appears to play a central role in the analysis of regularization methods, cf. [39]. However, in all the examples presented in this paper, we prove that the Bernstein condition holds on the entire set F.

2.3 The complexity function $r(\cdot)$

The complexity function $r(\cdot)$ is defined by

$$\forall \rho > 0, \quad r(\rho) = \left[\frac{A\rho \text{comp}(B)}{\sqrt{N}} \right]^{1/2\kappa}$$

where A is the constant in Assumption 2.1 and where $\text{comp}(B)$ is a measure of the complexity of the unit ball B associated to the regularization norm. Note that this complexity measure will depend on the stochastic assumption of F. In the bounded setting, $\text{comp}(B) = C\text{Rad}(B)$ where C is an absolute constant and $\text{Rad}(B)$ is the Rademacher complexity of B (whose definition will be reminded in Subsection 2.6). In the subgaussian setting, $\text{comp}(B) = CLw(B)$ where C is an absolute constant, L is the subgaussian parameter of the class $F - F$ and $w(B)$ is the Gaussian mean-width of B (here again, exact definitions of L and $w(B)$ will be reminded in Subsection 2.5).

Note that sharper (localized) versions of $r(\cdot)$ are provided in Section 8. However, as it is the simplest version that is used in most examples, we only introduce this version for now.

2.4 The sparsity parameter ρ^*

The size of the sub-differential of the regularization function $\|\cdot\|$ in a neighborhood of the oracle f^* will play as well a central role in our analysis. We recall now its definition: for every $f \in F$

$$\partial \|\cdot\| (f) = \left\{ g \in E : \|f + h\| - \|f\| \geq \langle g, h \rangle \text{ for all } h \in E \right\}.$$

It is well-known that $\partial \|\cdot\| (f)$ is a subset of the unit sphere of the dual norm of $\|\cdot\|$ when $f \neq 0$. Note also that when $f = 0$, $\partial \|\cdot\| (f)$ is the entire unit dual ball, a fact we will also use in two situations, either when the regularization norm has no "sparsity inducing power" – in particular, when it is a smooth function as in the RKHS case treated in Section 5; or when one wants extra *norm dependent* upper bounds (cf. [40] for more details where these bounds are called *complexity dependent*) in addition to *sparsity dependent* upper bounds. In the latter, the statistical bounds that we get are the minimum between an error rate that depends on the notion of sparsity naturally associated to the regularization norm (when it exists) and an error rate that depends on $\|f^*\|$.

Definition 2.1 (From [39])**.** *The **sparsity parameter** is the function $\Delta(\cdot)$ defined by*

$$\Delta(\rho) = \inf_{h \in \rho S \cap r(2\rho) B_{L_2}} \sup_{g \in \Gamma_{f^*}(\rho)} \langle h, g \rangle$$

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

Note that there is a slight difference with the definition of the *sparsity parameter* from [39] where there $\Delta(\rho)$ is defined taking the infimum over the sphere ρS intersected with a L_2-ball of radius $r(\rho)$ whereas in Definition 2.1, ρS is intersected with a L_2-ball of radius $r(2\rho)$. Up to absolute constants this has no effect on the behavior of $\Delta(\rho)$ and the difference comes from technical detains in our analysis (a peeling argument that we use below whereas a direct homogeneity argument was enough in [39]).

In the following, estimation rates with respect to the regularization norm $\|\cdot\|$, the norm $\|\cdot\|_{L_2}$ as well as sharp oracle inequalities are given. All the convergence rates depend on a single radius ρ^* that satisfies the *sparsity equation* as introduced in [39].

Definition 2.2. *The radius ρ^* is any solution of the sparsity equation:*

$$\Delta(\rho^*) \geq (4/5)\rho^*. \tag{8}$$

Since ρ^* is central in the results and drives the convergence rates, finding a solution to the sparsity equation will play an important role in all the examples that we worked out in the following. Roughly speaking, if the regularization norm induces sparsity, a sparse element in $f^* + (\rho/20)B$ (that is an element f for which $\partial\|\cdot\|(f)$ is almost extremal – that is almost as large as the dual sphere) yields the existence of a small ρ^*. In this case, ρ^* satisfies the sparsity equation.

In addition, if one takes $\rho = 20\|f^*\|$ then $0 \in \Gamma_{f^*}(\rho)$ and since $\partial\|\cdot\|(0)$ is the entire dual ball associate to $\|\cdot\|$, one has directly that $\Delta(\rho) = \rho$ and so ρ satisfies the sparsity Equation (8). We will use this observation to obtain *norm dependent* upper bounds, i.e. rates of convergence depending on $\|f^*\|$ and that do not depend on any sparsity parameter. Such a bound holds for any norm; in particular, for norms with no sparsity inducing power as in Section 5.

2.5 Theorem in the subgaussian setting

First, we introduce the subgaussian framework (then we will turn to the bounded case in the next section).

Definition 2.3 (Subgaussian class)**.** *We say that a class of functions $\mathcal{F}$ is L-subgaussian (w.r.t. X) for some constant $L \geq 1$ when for all $f \in \mathcal{F}$ and all $\lambda \geq 1$,*

$$\mathbb{E}\exp\left(\lambda|f(X)|/\|f\|_{L_2}^2\right) \leq \exp\left(\lambda^2 L^2\right) \tag{9}$$

where $\|f\|_{L_2} = \left(\mathbb{E}f(X)^2\right)^{1/2}$.

We will use the following operations on sets: for any $F' \subset E$ and $f \in E$,

$$F'+f = \{f'+f : f' \in F'\}, \quad F'-F' = \{f_1'-f_2' : f_1', f_2' \in F'\} \text{ and } d_{L_2}(F') = \sup\left(\|f_1' - f_2'\|_{L_2} : f_1', f_2' \in F'\right).$$

Assumption 2.2. *The class $F - F$ is L-subgaussian.*

Note that there are many equivalent formulations of the subgaussian property of a random variable based on ψ_2-Orlicz norms, deviations inequalities, exponential moments, moments growth characterization, etc. (cf., for instance Theorem 1.1.5 in [16]). The one we should use later is as follows: there exists some absolute constant $\mathbf{C}$ such that $F - F$ is L-subgaussian if and only if for all $f, g \in F$ and $t \geq 1$,

$$\mathbb{P}[|f(X) - g(X)| \geq \mathbf{C}tL\|f - g\|_{L_2}] \leq 2\exp(-t^2). \tag{10}$$

There are several examples of subgaussian classes. For instance, when F is a class of linear functionals $F = \{\langle\cdot, t\rangle : t \in T\}$ for $T \subset \mathbb{R}^p$ and X is a random variable in $\mathbb{R}^p$ then $F - F$ is L-subgaussian in the following cases:

1. X is a Gaussian vector in $\mathbb{R}^p$,

2. $X = (x_j)_{j=1}^p$ has independent coordinates that are subgaussian, that is, there are constants $c_0 > 0$ and $c_1 > 0$ such that $\forall j$, $\forall t > c_0, \mathbb{P}[|x_j| \geq t(\mathbb{E}x_j^2)^{1/2}] \leq 2\exp(-c_1 t^2)$,

3. for $2 \leq q < \infty$, X is uniformly distributed over $p^{1/q} B_{l_q}$ (cf. [3]),

4. $X = (x_j)_{j=1}^p$ is an unconditional vector (meaning that for every signs $(\epsilon_j)_j \in \{-1, +1\}^p$, $(\epsilon_j x_j)_{j=1}^p$ has the same distribution as $(x_j)_{j=1}^p$), $\mathbb{E}x_j^2 \geq c^2$ for some $c > 0$ and $\|X\|_{l_\infty} \leq R$ almost surely then one can choose $L \leq \mathbf{C}R/c$ (cf. [38]).

In the *subgaussian framework*, a natural way to measure the *statistical complexity* of the problem is via Gaussian mean-width that we introduce now.

Definition 2.4. *Let H be a subset of L_2 and denote by d the natural metric in L_2. Let $(G_h)_{h\in H}$ be the canonical centered Gaussian process indexed by H (in particular, the covariance structure of $(G_h)_{h\in H}$ is given by d: $\left(\mathbb{E}(G_{h_1} - G_{h_2})^2\right)^{1/2} = \left(\mathbb{E}(h_1(X) - h_2(X))^2\right)^{1/2}$ for all $h_1, h_2 \in H$). The **Gaussian mean-width** of H (as a subset of L_2) is*

$$w(H) = \mathbb{E}\sup_{h\in H} G_h.$$

We refer the reader to Section 12 in [21] for the construction of Gaussian processes in L_2. There are many natural situations where Gaussian mean-widths can be computed. To familiarize with this quantity let us consider an example in the *matrix* framework. Let $H = \{\langle M, \cdot\rangle : \|M\|_{S_1} \leq 1\}$ be the class of linear functionals indexed by the unit ball of the S_1-norm and d be the distance associated with the Frobenius norm (i.e. $d(\langle \cdot, M_1\rangle, \langle \cdot, M_2\rangle) = d(M_1, M_2) = \|M_1 - M_2\|_{S_2}$) then

$$w(H) = w(B_{S_1}) = \mathbb{E}\sup_{\|M\|_{S_1}\leq 1} \langle \mathbb{G}, M\rangle = \mathbb{E}\,\|\mathbb{G}\|_{S_1}^* = \mathbb{E}\,\|\mathbb{G}\|_{S_\infty} \sim \sqrt{m+T}$$

where $\mathbb{G}$ is a standard Gaussian matrix in $\mathbb{R}^{m\times T}$, $\|\cdot\|_{S_1}^*$ is the dual norm of the nuclear norm which is the operator norm $\|\cdot\|_{S_\infty}$.

We are now in position to define the complexity parameter as announced previously.

Definition 2.5. *The **complexity parameter** is the non-decreasing function $r(\cdot)$ defined for every $\rho \geq 0$,*

$$r(\rho) = \left(\frac{ACLw(B)\rho}{\sqrt{N}}\right)^{\frac{1}{2\kappa}}$$

where κ, A are the Bernstein parameters from Assumption 2.1, L is the subgaussian parameter from Assumption 2.2 and $C > 0$ is an absolute constant (the exact value of C can be deduced from the proof of Proposition 8.2). The Gaussian mean-width $w(B)$ of B is computed with respect to the the metric associated with the covariance structure of X, i.e. $d(f_1, f_2) = \|f_1 - f_2\|_{L_2}$ for every $f_1, f_2 \in F$.

After the computation of the Bernstein parameter κ, the complexity function $r(\cdot)$ and the radius ρ^*, it is now possible to explicit our main result in the sub-Gaussian framework.

Theorem 2.1. *Assume that Assumption 1.1, Assumption 2.1 and Assumption 2.2 hold and let $C > 0$ from the definition of $r(\cdot)$ in Definition 2.5. Let the regularization parameter λ be*

$$\lambda = \frac{5}{8}\frac{CLw(B)}{\sqrt{N}}$$

and ρ^* satisfying (8). Then, with probability larger than

$$1 - \mathbf{C}\exp\left(-\mathbf{C}N^{1/2\kappa}(\rho^* w(B))^{(2\kappa-1)/\kappa}\right) \tag{11}$$

we have

$$\left\|\hat{f}-f^*\right\| \leq \rho^*, \quad \left\|\hat{f}-f^*\right\|_{L_2} \leq r(2\rho^*) = \left[\frac{ACLw(B)2\rho^*}{\sqrt{N}}\right]^{1/2\kappa} \text{ and } \mathcal{E}(\hat{f}) \leq \frac{r(2\rho^*)^{2\kappa}}{A} = \frac{CLw(B)2\rho^*}{\sqrt{N}}$$

where $\mathbf{C}$ *denotes positive constants that might change from one instance to the other and depend only on* A, κ, L *and* C.

Remark 2.2 (Deviation parameter)**.** *Replacing* $w(B)$ *by any upper bound does not affect the validity of the result. As a special case, it is possible to increase the confidence level of the bound by replacing* $w(B)$ *by* $w(B)+x$*: then, with probability at least*

$$1-\mathbf{C}\exp\left(-\mathbf{C}N^{1/2\kappa}(\rho^*[w(B)+x])^{(2\kappa-1)/\kappa}\right)$$

we have in particular

$$\left\|\hat{f}-f^*\right\|_{L_2} \leq r(2\rho^*) = \left[\frac{ACL[w(B)+x]2\rho^*}{\sqrt{N}}\right]^{1/2\kappa} \text{ and } \mathcal{E}(\hat{f}) \leq \frac{r(2\rho^*)^{2\kappa}}{A} = \frac{CL[w(B)+x]2\rho^*}{\sqrt{N}}.$$

Remark 2.3 (Norm and sparsity dependent error rates)**.** *Theorem 2.1 holds for any radius* ρ^* *satisfying the sparsity equation (8). We have noticed in Section 2.4 that* $\rho^* = 20\|f^*\|$ *satisfies the sparsity equation since in that case* $0 \in \Gamma_{f^*}(\rho^*)$ *and so* $\Delta(\rho^*) = \rho^*$*. Therefore, one can apply Theorem 2.1 to both* $\rho^* = 20\|f^*\|$ *(this leads to* norm dependent *upper bounds) and to the smallest* ρ^* *satisfying the sparsity equation (8) (this leads to* sparsity dependent *upper bounds) at the same time. Both will lead to meaningful results (a typical example of such a combined result is Theorem 9.2 from [31] or Theorem 3.1 below).*

2.6 Theorem in the bounded setting

We now turn to the *bounded framework*; that is we assume that all the functions in F are uniformly bounded in L_∞. This assumption is very different in nature than the subgaussian assumption which is in fact a norm equivalence assumption (i.e. Definition 2.3 is equivalent to $\|f\|_{L_2} \leq \|f\|_{\psi_2} \leq L\|f\|_{L_2}$ for all $f \in \mathcal{F}$ where $\|\cdot\|_{\psi_2}$ is the ψ_2 Orlicz norm, cf. [51]).

Assumption 2.3 (Boundedness assumption)**.** *There exist a constant* $b > 0$ *such that for all* $f \in F$, $\|f\|_{L_\infty} \leq b$.

The main motivation to consider the *bounded setup* is for sampling over the canonical basis of a finite dimensional space like $\mathbb{R}^{m\times T}$ or $\mathbb{R}^p$. Note that this type of sampling is *stricto sensu* subgaussian, but with a constant L depending on the dimensions m and T, which yields sub-optimal rates. This is the reason why the results in the bounded setting are more relevant in this situation. This is especially true for the 1-bit matrix completion problem as introduced in Section 1. For this example, the X_i's are chosen randomly in the canonical basis $(E_{1,1},\cdots,E_{m,T})$ of $\mathbb{R}^{m\times T}$. Moreover, in that example, the class F is the class of all linear functionals indexed by bB_∞: $F = \{\langle\cdot, M\rangle : \max_{p,q}|M_{pq}| \leq b\}$ and therefore the study of this problem falls naturally in the bounded framework studied in this section.

Under the boundedness assumption, the natural way to measure the "statistical complexity" cannot be anymore characterized by Gaussian mean width. We therefore introduce another complexity parameter known as Rademacher complexities. This complexity measure has been extensively studied in the learning theory literature (cf., for instance, [30, 31, 4]).

Definition 2.6. *Let* H *be a subset of* L_2*. Let* $(\epsilon_i)_{i=1}^N$ *be* N *i.i.d. Rademacher variables (i.e.* $\mathbb{P}[\epsilon_i = 1] = \mathbb{P}[\epsilon_i = -1] = 1/2$*) independent of the* X_i*'s. The* ***Rademacher complexity*** *of* H *is*

$$\mathrm{Rad}(H) = \mathbb{E}\sup_{f\in H}\left|\frac{1}{\sqrt{N}}\sum_{i=1}^N \epsilon_i f(X_i)\right|.$$

Note that when $(f(X))_{f\in H}$ is a version of the isonormal process over L_2 (cf. Chapter 12 in [21]) restricted to H then the Gaussian mean-width and the Rademacher complexity coincide: $w(H) = \mathrm{Rad}(H)$. But, in that case, H is not bounded in L_∞ and, in general, the two complexity measures are different.

There are many examples where Rademacher complexities have been computed (cf. [48]). Like in the previous *subgaussian* setting the statistical complexity is given by a function $r(\cdot)$ (we use the same name in the two *bounded* and *subgaussian* setups because this $r(\cdot)$ function plays exactly the same role in both scenarii even though it uses different notion of complexity).

Definition 2.7. *The **complexity parameter** is the non-decreasing function $r(\cdot)$ defined for every $\rho \geq 0$ by*

$$r(\rho) = \left(\frac{CA\mathrm{Rad}(B)\rho}{\sqrt{N}}\right)^{\frac{1}{2\kappa}}, \text{ where } C = \frac{1920}{7}.$$

Theorem 2.2. *Assume that Assumption 1.1, Assumption 2.1 and Assumption 2.3 hold. Let the regularization parameter λ be chosen as $\lambda = 720\mathrm{Rad}(B)/7\sqrt{N}$. Then, with probability larger than*

$$1 - \mathbf{C}\exp\left(-\mathbf{C}N^{1/2\kappa}(\rho^*\mathrm{Rad}(B))^{(2\kappa-1)/\kappa}\right) \tag{12}$$

we have

$$\left\|\hat{f} - f^*\right\| \leq \rho^*, \quad \left\|\hat{f} - f^*\right\|_{L_2} \leq r(2\rho^*) = \left[\frac{CA\mathrm{Rad}(B)2\rho^*}{\sqrt{N}}\right]^{1/2\kappa} \text{ and } \mathcal{E}(\hat{f}) \leq \frac{r(2\rho^*)^{2\kappa}}{A} = \frac{C\mathrm{Rad}(B)2\rho^*}{\sqrt{N}},$$

where $\mathbf{C}$ denotes positive constants that might change from one instance to the other and depend only on A, b, κ and $r(\cdot)$ is the function introduced in Definition 2.7.

In the next Sections 3, 4 and 5 we compute $r(\rho)$ either in the subgaussian setup or in the bounded setup and solve the sparsity equation in various examples, showing the versatility of the main strategy.

3 Application to logistic LASSO and logistic SLOPE

The first example of application of the main results in Section 2 involves one very popular method developed during the last two decades in binary classification which is the Logistic LASSO procedure (cf. [43, 46, 59, 24, 54]).

We consider the *vector* framework, where $(X_1, Y_1), \ldots, (X_N, Y_N)$ are N i.i.d. pairs with values in $\mathbb{R}^p \times \{-1, 1\}$ distributed like (X, Y). Both bounded and subgaussian framework can be analyzed in this example. For the sake of shortness and since an example in the bounded case is provided in the next section, only the subgaussian case is considered here and we leave the bounded case to the interested reader. We therefore shall apply Theorem 2.1 to get estimation and prediction bounds for the well known logistic LASSO and the new logistic SLOPE.

In this section, we consider the class of linear functional indexed by RB_{l_2} for some radius $R \geq 1$ and the logistic loss:

$$F = \left\{\langle \cdot, t\rangle : t \in RB_{l_2}\right\}, \ell_f(x, y) = \log(1 + \exp(-yf(x))).$$

As usual the oracle is denoted by $f^* = \mathrm{argmin}_{f\in F}\mathbb{E}\ell_f(X, Y)$, we also introduce t^* such that $f^* = \langle \cdot, t^*\rangle$.

3.1 Logistic LASSO

The logistic loss function is Lipschitz with constant 1, so Assumption 1.1 is satisfied. It follows from Proposition 6.2 in Section 6.1 that Assumption 2.1 is satisfied when the design X is the standard Gaussian variable in $\mathbb{R}^p$ and the considered class F. In that case, the Bernstein parameter is $\kappa = 1$, and we have $A = c_0/R^3$ for some absolute constant $c_0 > 0$ which can be deduced from the proof of Proposition 6.2. We

consider the l_1 norm $\left\|\langle \cdot, t\rangle\right\| = \|t\|_{l_1}$ for regularization. We will therefore obtain statistical results for the RERM estimator $\widehat{f_L} = \langle \widehat{t_L}, \cdot \rangle$ that is defined by

$$\widehat{t_L} \in \operatorname*{argmin}_{t \in RB_{l_2}} \left(\frac{1}{N} \sum_{i=1}^{N} \log\left(1 + \exp(-Y_i \langle X_i, t\rangle\right) + \lambda \|t\|_{l_1} \right)$$

where λ is a regularization parameter to be chosen according to Theorem 2.1.

The two final ingredients needed to apply Theorem 2.1 are 1) the computation of the Gaussian mean width of the unit ball B_{l_1} of the regularization function $\|\cdot\|_{l_1}$ 2) find a solution ρ^* to the sparsity equation (8).

Let us first deal with the complexity parameter of the problem. If one assumes that the design vector X is **isotropic**, i.e. $\mathbb{E}\langle X, t\rangle^2 = \|t\|_{l_2}^2$ for every $t \in \mathbb{R}^p$ then the metric naturally associated with X is the canonical l_2-distance in $\mathbb{R}^p$. In that case, it is straightforward to check that $w(B_{l_1}) \leq c_1\sqrt{\log p}$ for some (known) absolute constant $c_1 > 0$ and so we define, for all $\rho \geq 0$,

$$r(\rho) = \mathbf{C}\left(\rho\sqrt{\frac{\log p}{N}}\right)^{1/2} \tag{13}$$

for the complexity parameter of the problem (from now and until the end of Section 3, the constants $\mathbf{C}$ depends only on L, C, c_0 and c_1).

Now let us turn to a solution ρ^* of the sparsity equation (8). First note that when the design is isotropic the sparsity parameter is the function

$$\Delta(\rho) = \inf\left\{ \sup_{g \in \Gamma_{t^*}(\rho)} \langle h, g\rangle : h \in \rho S_{l_1} \cap r(2\rho)B_{l_2} \right\}$$

where $\Gamma_{t^*}(\rho) = t^* + (\rho/20)B_{l_1}$.

A first solution to the sparsity equation is $\rho^* = 20\|t^*\|_{l_1}$ because it leads to $0 \in \Gamma_{t^*}(\rho^*)$. This solution is called *norm dependent*.

Another radius ρ^* solution to the sparsity equation (8) is obtained when t^* is close to a sparse-vector, that is a vector with a small support. We denote by $\|v\|_0 := |\mathrm{supp}(v)|$ the size of the support of $v \in \mathbb{R}^p$. Now, we recall a result from [39].

Lemma 3.1 (Lemma 4.2 in [39]). *If there exists some $v \in t^* + (\rho/20)B_{l_1}$ such that $\|v\|_0 \leq c_0(\rho/r(\rho))^2$ then $\Delta(\rho) \geq 4\rho/5$ where c_0 is an absolute constant.*

In particular, we get that $\rho^* \sim s\sqrt{(\log p)/N}$ is a solution to the sparsity equation if there is a s-sparse vector which is $(\rho^*/20)$-close to t^* in l_1. This radius leads to the so-called *sparsity dependent* bounds.

After the derivation of the Bernstein parameter $\kappa = 1$, the complexity $w(B)$ and a solution ρ^* to the sparsity equation, we are now in a position to apply Theorem 2.1 to get statistical bounds for the Logistic LASSO.

Theorem 3.1. *Assume that X is a standard Gaussian vector in $\mathbb{R}^p$. Let $s \in \{1, \dots, p\}$. Assume that there exists a s-sparse vector in $t^* + \mathbf{C}s\sqrt{(\log p)/N}B_{l_1}$. Then, with probability larger than $1 - \mathbf{C}\exp(-\mathbf{C}s\log p)$, for every $1 \leq q \leq 2$, the logistic LASSO estimator $\widehat{t_L}$ with regularization parameter*

$$\lambda = \frac{5c_1CL}{8}\sqrt{\frac{\log p}{N}}$$

satisfies

$$\left\|\widehat{t_L} - t^*\right\|_{l_q} \leq \mathbf{C}\min\left(s^{1/q}\sqrt{\frac{\log p}{N}}, \|t^*\|_{l_1}^{1/q}\left(\frac{\log p}{N}\right)^{\frac{1}{2}-\frac{1}{2q}} \right)$$

and the excess logistic risk of $\widehat{t_L}$ *is such that*

$$\mathcal{E}_{logistic} = R(\widehat{t_L}) - R(t^*) \leq \mathbf{C} \min \left(\frac{s \log(p)}{N}, \|t^*\|_{l_1} \sqrt{\frac{\log(p)}{N}} \right).$$

Note that an estimation result for any l_q-norm for $1 \leq q \leq 2$ follows from results in l_1 and l_2 and the interpolation inequality $\|v\|_{l_q} \leq \|v\|_{l_1}^{-1+2/q} \|v\|_{l_2}^{2-2/q}$.

Estimation results for the logistic LASSO estimator in the generalized linear model have been obtained in [64] under the assumption that the basis functions and the oracle are bounded. This assumption does not hold here since the *basis functions* – defined here by $\psi_k(\cdot) = \langle e_k, \cdot \rangle$ where $(e_k)_{k=1}^d$ is the canonical basis of $\mathbb{R}^p$ – are not bounded when the design is $X \sim \mathcal{N}(0, I_{d \times p})$. Moreover, we do not make the assumption that f^* is bounded in L_∞. Nevertheless, we recover the same estimation result for the l_2-loss and l_1-loss as in [64]. But we also provide a prediction result since an excess risk bound is also given in Theorem 3.1.

Note that Theorem 3.1 recovers the classic rates of convergence for the logistic LASSO estimator that have been obtained in the literature so far. This rates is the minimax rate as long as $\log(p/s)$ behaves like $\log p$. This is indeed the case when $s \ll p$ which is the classic setup in high-dimensional statistics. But when s is proportional to p this rate is not minimax since there is a logarithmic loss. To overcome this issue we introduce a new estimator: the logistic SLOPE.

3.2 Logistic Slope

The construction of the logistic Slope is similar to the one of the logistic LASSO except that the regularization norm used in this case is the SLOPE norm (cf. [57, 9]): for every $t = (t_j) \in \mathbb{R}^p$,

$$\|t\|_{SLOPE} = \sum_{j=1}^{p} \sqrt{\log(ep/j)} t_j^\sharp \tag{14}$$

where $t_1^\sharp \geq t_2^\sharp \geq \cdots \geq 0$ is the non-increasing rearrangement of the absolute values of the coordinates of t and e is the base of the natural logarithm. Using this estimator with a regularization parameter $\lambda \sim 1/\sqrt{N}$ we recover the same result as for the Logistic LASSO case except that one can get, in that case, the optimal minimax rate for any $s \in \{1, \ldots, p\}$:

$$\sqrt{\frac{s}{N} \log\left(\frac{ep}{s}\right)}.$$

Indeed, it follows from Lemma 5.3 in [39] that the Gaussian mean width of the unit ball B_{SLOPE} associated with the SLOPE norm is of the order of a constant. The *sparsity dependent* radius satisfies

$$\rho^* \sim \frac{s}{\sqrt{N}} \log\left(\frac{ep}{s}\right) \tag{15}$$

as long as there is a s-sparse vector in $t^* + (\rho^*/20)B_{SLOPE}$. The *norm dependent* radius is as usual of order $\|t^*\|_{SLOPE}$. Then, the next result follows from Theorem 2.1. It improves the best known bounds on the logistic LASSO.

Theorem 3.2. *Assume that X is a standard Gaussian vector in $\mathbb{R}^p$. Let $s \in \{1, \ldots, p\}$. Assume that there exists a s-sparse vector in $t^* + (\rho^*/20)B_{SLOPE}$ for ρ^* as in (15). Then, with probability larger than $1 - \mathbf{C}\exp(-\mathbf{C}s\log(p/s))$, the logistic SLOPE estimator*

$$\widehat{t_S} \in \operatorname*{argmin}_{t \in RB_{l_2}} \left(\frac{1}{N} \sum_{i=1}^{N} \log\left(1 + \exp(-Y_i \langle X_i, t \rangle)\right) + \frac{\mathbf{C}}{\sqrt{N}} \|t\|_{SLOPE} \right)$$

satisfies

$$\left\|\widehat{t_S} - t^*\right\|_{SLOPE} \leq \mathbf{C} \min \left(\frac{s}{\sqrt{N}} \log\left(\frac{ep}{s}\right), \|t^*\|_{SLOPE} \right)$$

and

$$\left\|\widehat{t_S} - t^*\right\|_{l_2} \leq \mathbf{C} \min\left(\sqrt{\frac{s}{N}\log\left(\frac{ep}{s}\right)}, \sqrt{\frac{\|t^*\|_{SLOPE}}{\sqrt{N}}}\right)$$

and the excess logistic risk of $\widehat{t_S}$ *is such that*

$$\mathcal{E}_{logistic}(\widehat{t_S}) = R(\widehat{t_S}) - R(t^*) \leq \mathbf{C} \min\left(\frac{s\log ep/s}{N}, \|t^*\|_{l_1}\sqrt{\frac{\log ep/s}{N}}\right).$$

Let us comment on Theorem 3.2 together with the fact that we do not make any assumption on the output Y all along this work. Theorem 3.2 proves that there exists an estimator achieving the minimax rate $s\log(ep/s)/N$ for the ℓ_2-estimation risk (to the square) with absolutely no assumption on the output Y; neither a moment assumption nor a "connection" between Y and X. In the case where a statistical model $Y = \text{sign}(\langle X, t^*\rangle + \xi)$ holds, where ξ is independent of X then Theorem 3.2 shows that the RERM with logistic loss and SLOPE regularization achieves the minimax rate $s\log(ep/s)/N$ under no assumption on the noise ξ. In particular, ξ does not need to have any moment and, for instance, the mimimax rate $s\log(ep/s)/N$ can still be achieved when the noise has a Cauchy distribution. Moreover, this estimation rate holds with exponentially large probability as if the noise had a Gaussian distribution (cf. [38]). This is a remarkable feature of Lipschitz loss functions genuinely understood in Huber's seminal paper [27].

	LASSO	SLOPE
$w(B)$	$\sqrt{\log p}$	1
ρ^*	$\frac{s}{\sqrt{N}}\sqrt{\log p}$	$\frac{s}{\sqrt{N}}\log\frac{ep}{s}$
$r(\rho^*)$	$\frac{s}{N}\log p$	$\frac{s}{N}\log\frac{ep}{s}$

Table 1: Comparison of the key quantities involved in our study for the ℓ_1 (LASSO) and SLOPE norms

In Table 1, the different quantities playing an important role in our analysis have been collected for the ℓ_1 and SLOPE norms: the Gaussian mean width $w(B)$ of the unit ball B of the regularization norm, a radius ρ^* satisfying the sparsity equation and finally the L_2 estimation rate of convergence $r(\rho^*)$ summarizing the two quantities. As mentioned in Figure 1, having a large sub-differential at sparse vectors and a small Gaussian mean-width $w(B)$ is a good way to construct "sparsity inducing"regularization norms as it is, for instance the case of "atomic norms" (cf. [17]).

4 Application to matrix completion via S_1-regularization

The second example involves matrix completion and uses the bounded setting from Section 2.6. The goal is to derive new results on two ways: the 1-bit matrix completion problem where entries are binary, and the quantile completion problem. The main theorems in this section yield upper bounds on completion in S_p norms ($1 \leq p \leq 2$) and on various excess risks. We also propose algorithms in order to compute efficiently the RERM in the matrix completion issue but with non differentiable loss and provide a simulation study. We first present a general theorem and then turn to specific loss functions because they induce a discussion about the Bernstein assumption and the κ parameter and lead to more particular theorems.

4.1 General result

In this section, we consider the matrix completion problem. Contrary to the introduction, we do not immediately focus on the case $Y \in \{-1, +1\}$. So for the moment, Y is a general real random variable and ℓ is any Lipschitz loss. The class is $F = \{\langle \cdot, M\rangle : M \in bB_\infty\}$, where $bB_\infty = \{M = (M_{pq}) : \max_{p,q}|M_{pq}| \leq b\}$ and $b > 0$. As the design X takes its values in the canonical basis of $\mathbb{R}^{m\times T}$, the boundedness assumption

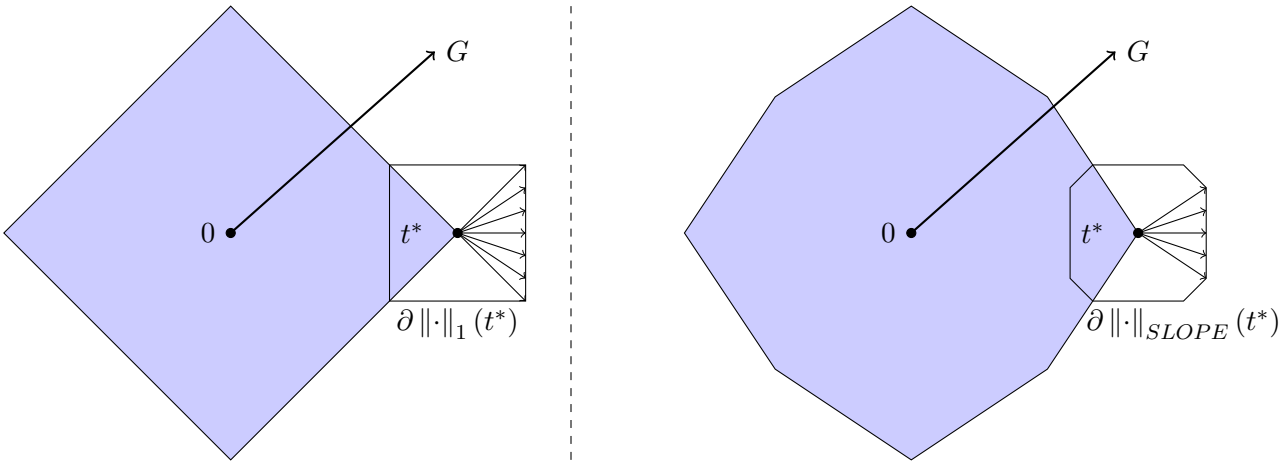

Figure 1: **Gaussian complexity and size of the sub-differential for the ℓ_1 ans SLOPE norms:** A "large" sub-differential at sparse vectors and a small Gaussian mean width of the unit ball of the regularization norm is better for sparse recovery. In this figure, G represents a "typical" Gaussian vector used to compute the Gaussian mean width of the unit regularization norm ball.

is satisfied. Apart from that, the notations and assumptions are as in the introduction, that is, we assume that X satisfies Assumption 1.2, with parameters $(\underline{c}, \bar{c})$, and the penalty is the nuclear norm. Thus, the RERM is given by

$$\widehat{M} \in \operatorname*{argmin}_{M \in bB_\infty} \left(\frac{1}{N} \sum_{i=1}^{N} \ell(\langle X_i, M \rangle, Y_i) + \lambda \|M\|_{S_1} \right). \tag{16}$$

Statistical properties of (16) will follow from Theorem 2.2 since one can recast this problem in the setup of Section 2.6. The oracle matrix M^* is defined by $f^* = \langle \cdot, M^* \rangle$, that is, $M^* = \operatorname{argmin}_{M \in bB_\infty} \mathbb{E}\ell(\langle M, X \rangle, Y)$.

Let us also introduce the matrix $\overline{M} = \operatorname{argmin}_{M \in \mathbb{R}^{m \times T}} \mathbb{E}\ell(\langle M, X \rangle, Y)$. Note that $\langle \overline{M}, \cdot \rangle = \overline{f} = \arg\min_{f \text{ measurable}} \mathbb{E}\ell(f(X), Y)$. Our general results usually are on f^* rather than on $\overline{f}$ as it is usually impossible to provide rates on the estimation of $\overline{f}$ without stringent assumptions on Y and F. However, as noted in the introduction, in 1-bit matrix completion with the hinge loss, we have $\overline{M} = M^*$ without any extra assumption when $b = 1$ (this is a favorable case). On the other hand, to get fast rates in matrix completion with quantile loss requires that $\overline{M} = M^*$ (which is a stringent assumption in this setting).

Complexity function We first compute the complexity parameter $r(\cdot)$ as introduced in Definition 2.7. To that end one just needs to compute the global Rademacher complexity of the unit ball of the regularization function which is $B_{S_1} = \{A \in \mathbb{R}^{m \times T} : \|A\|_{S_1} \leq 1\}$:

$$\mathrm{Rad}(B_{S_1}) = \mathbb{E} \sup_{\|A\|_{S_1} \leq 1} \left| \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \epsilon_i \langle X_i, A \rangle \right| = \mathbb{E} \left\| \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \epsilon_i X_i \right\|_{S_\infty} \leq c_0(\underline{c}, \bar{c}) \sqrt{\frac{\log(m+T)}{\min(m,T)}} \tag{17}$$

where $\|\cdot\|_{S_\infty}$ is the operator norm (i.e. the largest singular value), the last inequality follows from Lemma 1 in [32] and $c_0(\underline{c}, \bar{c}) > 0$ is some constant that depends only on $\underline{c}$ and $\bar{c}$.

The complexity parameter $r(\cdot)$ is derived from Definition 2.7: for any $\rho \geq 0$,

$$r(\rho) = \left[\frac{CA\rho \mathrm{Rad}(B_{S_1})}{\sqrt{N}} \right]^{\frac{1}{2\kappa}} = \mathbf{C} \left[\rho \sqrt{\frac{\log(m+T)}{N \min(m,T)}} \right]^{\frac{1}{2\kappa}} \tag{18}$$

where from now the constants $\mathbf{C}$ depend only on $\underline{c}$, $\bar{c}$, b, A and κ.

Sparsity parameter The next important quantity is the sparsity parameter. Its expression in this particular case is, for any $\rho > 0$,

$$\Delta(\rho) = \inf \left\{ \sup_{G \in \Gamma_{M^*}(\rho)} \langle H, G \rangle : H \in \rho S_{S_1} \cap ((\sqrt{mT}/\underline{c}) r(2\rho)) B_{S_2} \right\}$$

where $\Gamma_{M^*}(\rho)$ is the union of all the sub-differential of $\|\cdot\|_{S_1}$ in a S_1-ball of radius $\rho/20$ centered in M^*. Note that the normalization factor $\sqrt{mT}$ in the localization $(\sqrt{mT} r(2\rho)) B_{S_2}$ comes from the "non normalized isotropic" property of X: $\underline{c} \|M\|_{S_2}^2 / (mT) \leq \mathbb{E}\langle X, M \rangle^2 \leq \bar{c} \|M\|_{S_2}^2 / (mT)$ for all $M \in \mathbb{R}^{m \times T}$. Now, we use a result from [39] to find a solution to the sparsity equation.

Lemma 4.1 (Lemma 4.4 in [39]). *There exists an absolute constant $c_1 > 0$ for which the following holds. If there exists $V \in M^* + (\rho/20) B_{S_1}$ such that $\operatorname{rank}(V) \leq \left(c_1 \rho / (\sqrt{mT} r(\rho)) \right)^2$ then $\Delta(\rho) \geq 4\rho/5$.*

It follows from Lemma 4.1 that the sparsity equation (8) is satisfied by ρ^* when it exists $V \in M^* + (\rho^*/20) B_{S_1}$ such that $\operatorname{rank}(V) = c_1 \left(\rho^* / (\sqrt{mT} r(\rho^*)) \right)^2$. Note obviously that V can be M^* itself, in this case, ρ^* can be taken such that $\operatorname{rank}(M^*) = c_1 \left(\rho^* / (\sqrt{mT} r(\rho^*)) \right)^2$. However, when M^* is not low-rank, it might still be that a low-rank approximation V of M^* is close enough to M^* w.r.t. the S_1-norm. As a consequence, if for some $s \in \{1, \dots, \min(m, T)\}$ there exists a matrix V with rank at most s in $M^* + (\rho_s^*/20) B_{S_1}$ where

$$\rho_s^* = \mathbf{C} \, (smT)^{\frac{\kappa}{2\kappa - 1}} \left(\frac{\log(m+T)}{N \min(m, T)} \right)^{\frac{1}{2(2\kappa-1)}} . \tag{19}$$

then ρ_s^* satisfies the sparsity equation.

Following the remark at the end of Subsection 2.4, another possible choice is $\rho^* = 20 \|M^*\|_{S_1}$ in order to get *norm dependent* rates. In the end, we choose $\rho^* = \mathbf{C} \min [\rho_s^*, \|M^*\|_{S_1}]$. We are now in a position to apply Theorem 2.2 to derive statistical properties for the RERM $\widehat{M}$ defined in (16).

Theorem 4.1. *Assume that Assumption 1.1, 1.2 and 2.1 hold. Consider the estimator in (16) with regularization parameter*

$$\lambda = \frac{c_0(\underline{c}, \bar{c}) 720}{7} \sqrt{\frac{\log(m+T)}{N \min(m, T)}} \tag{20}$$

where $c_0(\underline{c}, \bar{c})$ are the constants in Assumption 1.2. Let $s \in \{1, \dots, \min(m, T)\}$ and assume that there exists a matrix with rank at most s in $M^* + (\rho_s^*/20) B_{S_1}$. Then, with probability at least

$$1 - \mathbf{C} \exp\left(-\mathbf{C} s (m+T) \log(m+T) \right)$$

we have

$$\left\| \widehat{M} - M^* \right\|_{S_1} \leq \mathbf{C} \min \left\{ (smT)^{\frac{\kappa}{2\kappa-1}} \left(\frac{\log(m+T)}{N \min(m,T)} \right)^{\frac{1}{2(2\kappa-1)}} , \|M^*\|_{S_1} \right\},$$

$$\frac{1}{\sqrt{mT}} \left\| \widehat{M} - M^* \right\|_{S_2} \leq \mathbf{C} \min \left\{ \left(\frac{s(m+T)\log(m+T)}{N} \right)^{\frac{1}{2(2\kappa-1)}} , \left(\|M^*\|_{S_1} \sqrt{\frac{\log(m+T)}{N \min(m,T)}} \right)^{\frac{1}{2\kappa}} \right\}$$

$$\mathcal{E}(\widehat{M}) \leq \mathbf{C} \min \left\{ \left(\frac{s(m+T)\log(m+T)}{N} \right)^{\frac{\kappa}{2\kappa-1}} , \|M^*\|_{S_1} \sqrt{\frac{\log(m+T)}{N \min(m,T)}} \right\}.$$

Note that the interpolation inequality also allows to get a bound for the S_p norm, when $1 \leq p \leq 2$:

$$\frac{1}{(mT)^{\frac{1}{p}}} \left\| \widehat{M} - M^* \right\|_{S_p} \leq \mathbf{C} \min \left\{ \left[\left(\frac{s^{2(p-1)+\kappa(2-p)}(m+T)^{p-1}}{\min(m,T)^{\frac{2-p}{2}}} \right)^{\frac{1}{p}} \sqrt{\frac{\log(m+T)}{N}} \right]^{\frac{1}{2\kappa-1}}, \right.$$
$$\left. \|M^*\|_{S_1}^{\frac{p-1+\kappa(2-p)}{p\kappa}} \left(\frac{\log(m+T)}{N \min(m,T)} \right)^{\frac{p-1}{2\kappa p}} \left(\frac{1}{mT} \right)^{\frac{2-p}{p}} \right\}.$$

Theorem 4.1 shows that the sparsity dependent error rate in the excess risk bound is (for $s = \operatorname{rank}(M^*)$)

$$\left(\frac{\operatorname{rank}(M^*)(m+T)\log(m+T)}{N} \right)^{\frac{\kappa}{2\kappa-1}}$$

which is the classic excess risk bound under the margin assumption up to a log factor (cf. [2]). As for the S_2-estimation error, when $\kappa = 1$, we recover the classic S_2-estimation rate

$$\sqrt{\frac{\operatorname{rank}(M^*)(m+T)\log(m+T)}{N}}$$

which is minimax in general (up to log terms, e.g. take the quadratic loss when Y is bounded and compare to [53]).

4.2 Algorithm and Simulation Outlines

Since this part provides new methods and results on matrix completion, we propose an algorithm in order to compute efficiently the RERM using the hinge loss and the quantile loss. This section explains the structure of the algorithm that is used with specific loss functions in next sections. Although many algorithms exist for the least squares matrix completion, at our knowledge many of them treat only the exact recovery such as in [12] and [45], or at least they all deal with differentiable loss functions, see [26]. On the other hand, the two losses that we mainly consider here are non differentiable because they are piecewise linear (in the case of hinge and $0,5$-quantile loss functions): new algorithms are hence needed. It has been often noted that the RERM with respect to the hinge loss or 0.5-quantile loss can been solved by a semidefinite programming but the cost is prohibitive for large matrices, say dimensions larger than 100. It actually works for small matrices as we ran SDP solver in Python in very small examples.

We propose here an *alternating direction method of multiplier* (ADMM) algorithm. For a clear and self-contained introduction to this class of algorithms, the reader is referred to [11] and we do not explain all the details here and we keep the same vocabulary. When the optimization problem is a sum of two parts, the core idea is to split the problem by introducing an extra variable. In our case, the two following problems are equivalent:

$$\underset{M}{\text{minimize}} \left\{ \frac{1}{N} \sum_{i=1}^{N} \ell(\langle X_i, M \rangle, Y_i) + \lambda \|M\|_{S_1} \right\}, \quad \underset{M,L}{\text{minimize}} \left\{ \frac{1}{N} \sum_{i=1}^{N} \ell(\langle X_i, M \rangle, Y_i) + \lambda \|L\|_{S_1} \right\}$$
$$\text{subject to } M = L$$

Below, we use the scaled form and the $m \times T$ matrix U is then called the *scaled dual variable*. Note that the S_2 norm is also the Froebenius norm and is thus elementwise. We can now exhibit the *augmented Lagrangian*:

$$L_\alpha(M, L, U) = \frac{1}{N} \sum_{i=1}^{N} \ell(\langle X_i, M \rangle, Y_i) + \lambda \|L\|_{S_1} + \frac{\alpha}{2} \|M - L + U\|_{S_2}^2 - \frac{\alpha}{2} \|U\|_{S_2}^2,$$

where α is a positive constant, called the *augmented Lagrange parameter*. The ADMM algorithm [11] is then:

$$M^{k+1} = \underset{M}{\operatorname{argmin}} \left(\frac{1}{N} \sum_{i=1}^{N} \ell(\langle X_i, M \rangle, Y_i) + \frac{\alpha}{2} \left\| M - L^k + U^k \right\|_{S_2}^2 \right) \tag{21}$$

$$L^{k+1} = \underset{L}{\operatorname{argmin}} \left(\lambda \|L\|_{S_1} + \frac{\alpha}{2} \left\| M^{k+1} - L + U^k \right\|_{S_2}^2 \right) \tag{22}$$

$$U^{k+1} = U^k + M^{k+1} - L^{k+1}$$

The starting point (M^0, L^0, U^0) uses one random matrix with independent Gaussian entries for M^0 and two zero matrices for L^0 and U^0. Another choice of starting point is to use a previous estimator with a larger λ. The stopping criterion is, as explained in [11], $\left\|M^{k+1} - M^k\right\|_{S_2}^2 + \left\|U^{k+1} - U^k\right\|_{S_2}^2 \leq \varepsilon$ for a fixed threshold ε. It means that it stops when both (U_k) and (M_k) start converging.

General considerations The second step (22) is independent of the loss function. It is well-known that the solution of this problem is $S_{\lambda/\alpha}(M^{k+1}+U^k)$ when $S_a(M)$ is the soft-thresholding operator with magnitude a applied to the singular values of the matrix M. It is defined for a rank r matrix M with SVD $M = U\Sigma V^\top$ where $\Sigma = \operatorname{diag}\left((d_i)_{1\leq i\leq r}\right)$ by $S_a(M) = US_a(\Sigma)V^\top$ where $S_a(\Sigma) = \operatorname{diag}\left((\max(0, d_i - a))_{1\leq i\leq r}\right)$.

It requires the SVD of a $m \times T$ matrix at each iteration and is the main bottleneck of this algorithm (the other main step (21) can be performed elementwise since the X_i's take their values in the canonical basis of $\mathbb{R}^{m\times T}$; so it needs only at most N operations). Two methods may be used in order to speed up the algorithm: efficient algorithms for computing the n largest singular values and the associate subspaces, such as the well-known PROPACK routine in Fortran. It can be plugged in order to solve (22) by computing the n largest and stop at this stage if the lowest computed singular values is lower than the threshold. It is obviously more relevant when the target is expected to have a very small rank. This method has been implemented in Python and works well in practice even though the parameter n has to be tuned carefully. An alternative method is to use approximate SVD such as in [25].

Moreover, the first step (21) (which may be performed elementwise) has a closed form solution for hinge and quantile loss: it is a soft-thresholding applied to a specified quantity.

Simulated observations as well as real-world data (cf. the MovieLens dataset[2]) are considered in the examples below. Finally note that parameter λ is tuned by cross-validation.

4.3 1-bit matrix completion

In this subsection we assume that $Y \in \{-1, +1\}$, and we challenge two loss functions: the logistic loss, and the hinge loss. It is worth noting that the minimizer $\overline{M} = \operatorname{argmin}_{M\in\mathbb{R}^{m\times T}} \mathbb{E}\ell(\langle M, X\rangle, Y)$ is not the same for both losses. For the hinge loss, it is known that it is the matrix formed by the Bayes classifier. This matrix has entries bounded by 1 so $M^* = \overline{M}$ as soon as $b = 1$. In opposite to this case, the logistic loss leads to a matrix $\overline{M}$ with entries formed by the odds ratio. It may even be infinite when there is no noise.

Logistic loss. Let us start by assuming that ℓ is the logistic loss. Thanks to Proposition 6.1 we know that $\kappa = 1$ for any b (A is also known, $A = 4\exp(2b)$) and therefore next result follows from Theorem 4.1. Note that we do not assume that $\overline{M}$ is in F and therefore our results provides estimation and prediction bounds for the oracle M^*.

Theorem 4.2 (1-bit Matrix Completion with logistic loss)**.** *Assume that Assumption 1.2 holds. Let $s \in \{1, \ldots, \min(m, T)\}$ and assume that there exists a matrix with rank at most s in $M^* + (\rho_s^*/20)B_{S_1}$ where ρ_s^* is defined in (19). With probability at least*

$$1 - \mathbf{C}\exp\left(-\mathbf{C}s\max(m, T)\log(m + T)\right)$$

[2] available in http://grouplens.org/datasets/movielens/

the estimator

$$\widehat{M} \in \operatorname*{argmin}_{M \in bB_\infty} \left(\frac{1}{N} \sum_{i=1}^{N} \log\left(1 + \exp\left(-Y_i \langle X_i, M \rangle\right)\right) + \lambda \|M\|_{S_1} \right) \tag{23}$$

with λ as in Equation (20) satisfies

$$\frac{1}{mT} \left\| \widehat{M} - M^* \right\|_{S_1} \leq \mathbf{C} \min \left\{ s \sqrt{\frac{\log(m+T)}{N \min(m,T)}}, \frac{\|M^*\|_{S_1}}{mT} \right\},$$

$$\frac{1}{\sqrt{mT}} \left\| \widehat{M} - M^* \right\|_{S_2} \leq \mathbf{C} \min \left\{ \sqrt{\frac{s \max(m,T) \log(m+T)}{N}}, \|M^*\|_{S_1}^{\frac{1}{2}} \left(\frac{\log(m+T)}{N \min(m,T)} \right)^{\frac{1}{4}} \right\}$$

$$\mathcal{E}_{logistic}(\widehat{M}) \leq \mathbf{C} \min \left\{ \frac{s \max(m,T) \log(m+T)}{N}, \|M^*\|_{S_1} \sqrt{\frac{\log(m+T)}{N \min(m,T)}} \right\}.$$

Using an interpolation inequality, it is easy to derive estimation bound in S_p for all $1 \leq p \leq 2$ as in Theorem 4.1 so we do not reproduce it here. Also, note that our bound on $\left\| \widehat{M} - M^* \right\|_{S_2}$ is of the same order as the one in [34]. We actually now prove that this rate is minimax-optimal (up to log terms).

Theorem 4.3 (Lower bound with logistic loss)**.** *For a given matrix $M \in B_\infty$, define $\mathbb{P}_M^{\otimes N}$ as the probability distribution of the N-uplet $(X_i, Y_i)_{i=1}^N$ of i.i.d. pairs distributed like (X, Y) such that X is uniformly distributed on the canonical basis $(E_{p,q})$ of $\mathbb{R}^{m \times T}$ and $\mathbb{P}_M(Y = 1 | X = E_{p,q}) = \exp(M_{pq})/[1 + \exp(M_{pq})]$ for every $(p,q) \in \{1, \dots, m\} \times \{1, \dots, T\}$. Fix $s \in \{1, \dots, \min(m,T)\}$ and assume that $N \geq s(m+T)\log(2)/(8b^2)$. Then*

$$\inf_{\widehat{M}} \sup_{\substack{M^* \in bB_\infty \\ \operatorname{rank}(M^*) \leq s}} \mathbb{P}_{M^*}^{\otimes N} \left(\frac{1}{\sqrt{mT}} \left\| \widehat{M} - M^* \right\|_{S_2} \geq c \sqrt{\frac{(m+T)s}{N}} \right) \geq \beta$$

for some universal constants $\beta, c > 0$.

Also, as pointed out in the introduction, the quantity of interest is not the logistic excess risk, but the classification excess risk: let us remind that $R_{0/1}(M) = \mathbb{P}[(Y \neq \operatorname{sign}(\langle M, X \rangle)]$ for all $M \in \mathbb{R}^{m \times T}$. Even if we assume that $M^* = \overline{M}$, all that can be deduced from Theorem 2.1 in [66] is that

$$\mathcal{E}_{0/1}(\widehat{M}) = R_{0/1}(\widehat{M}) - \inf_{M \in \mathbb{R}^{m \times T}} R_{0/1}(M) \leq \mathbf{C} \sqrt{\mathcal{E}_{logistic}(\widehat{M})} \leq \mathbf{C} \sqrt{\frac{\operatorname{rank}(\overline{M})(m+T)\log(m+T)}{N}}.$$

But this rate on the excess 0/1-risk may be much better under the margin assumption [44, 60] (cf. Equation (36) below). This motivates the use of the hinge loss instead of the logistic loss, for which the results in [66] do not lead to a loss of a square root in the rate.

Hinge loss. As explained above, the choice $b = 1$ ensures $\overline{M} = M^*$ without additional assumption. Thanks to Proposition 6.3 we know that as soon as $\inf_{p,q} |\overline{M}_{p,q} - 1/2| \geq \tau$ for some $\tau > 0$, the Bernstein assumption is satisfied by the hinge loss with $\kappa = 1$ and $A = 1/(2\tau)$. This assumption seems very mild in many situations and we derive the results with it.

Theorem 4.4 (1-bit Matrix Completion with hinge loss)**.** *Assume that Assumption 1.2 holds. Assume that $\inf_{p,q} |P(Y = 1 | X = E_{p,q}) - 1/2| \geq \tau$ for some $\tau > 0$. Let $s \in \{1, \dots, \min(m,T)\}$ and assume that there exists a matrix with rank at most s in $\overline{M} + (\rho_s^*/20) B_{S_1}$ where ρ_s^* is defined in (19). With probability at least*

$$1 - \mathbf{C} \exp\left(-\mathbf{C} s \max(m,T) \log(m+T)\right)$$

the estimator

$$\widehat{M} \in \operatorname*{argmin}_{M \in B_\infty} \left(\frac{1}{N} \sum_{i=1}^{N} \left(1 - Y_i \langle X_i, M \rangle\right)_+ + \lambda \|M\|_{S_1} \right) \tag{24}$$

with λ as in Equation (20) satisfies

$$\frac{1}{mT} \left\| \widehat{M} - \overline{M} \right\|_{S_1} \leq \mathbf{C} \min \left\{ s \sqrt{\frac{\log(m+T)}{N \min(m,T)}}, \frac{\|\overline{M}\|_{S_1}}{mT} \right\},$$

$$\frac{1}{\sqrt{mT}} \left\| \widehat{M} - \overline{M} \right\|_{S_2} \leq \mathbf{C} \min \left\{ \sqrt{\frac{s(m+T)\log(m+T)}{N}}, \|\overline{M}\|_{S_1}^{\frac{1}{2}} \left(\frac{\log(m+T)}{N \min(m,T)} \right)^{\frac{1}{4}} \right\}$$

$$\mathcal{E}_{hinge}(\widehat{M}) \leq \mathbf{C} \min \left\{ \frac{s(m+T)\log(m+T)}{N}, \|\overline{M}\|_{S_1} \sqrt{\frac{\log(m+T)}{N \min(m,T)}} \right\}.$$

In this case, [66] implies that the excess risk bound for the classification error (using the 0/1-loss) is the same as the one for the hinge loss: it is therefore of the order of $\operatorname{rank}(\overline{M}) \max(m,T)/N$.

Note that the rate $\operatorname{rank}(\overline{M}) \max(m,T)/N$ for the classification excess error was only reached in [18] up to our knowledge (using the PAC-Bayesian technique from [14, 15, 42, 1]), in the very restrictive noiseless setting - that is, $P(Y = 1|X = E_{p,q}) \in \{0,1\}$ which is equivalent to $P(Y = \operatorname{sign}(\langle \overline{M}, X \rangle)) = 1$. Here this rate is proved to hold in the general case. Other works, including [55], obtained only rates in $1/\sqrt{N}$. Finally, we prove that this rate is the minimax rate in the next result.

Theorem 4.5 (Lower bound with hinge loss)**.** *For a given matrix $M \in B_\infty$, let $\mathbb{E}_M^{\otimes N}$ be the expectation w.r.t. the N-uplet $(X_i, Y_i)_{i=1}^N$ of i.i.d. pairs distributed like (X,Y) such that X is uniformly distributed on the canonical basis $(E_{p,q})$ of $\mathbb{R}^{m \times T}$ and $\mathbb{P}_M(Y = 1|X = E_{p,q}) = M_{pq}$ for every $(p,q) \in \{1,\dots,m\} \times \{1,\dots,T\}$. Fix $s \in \{1,\dots,\min(m,T)\}$ and assume that $N \geq s \max(m,T)\log(2)/8$. Then*

$$\inf_{\widehat{M}} \sup_{\substack{M^* \in B_\infty \\ \operatorname{rank}(M^*) \leq r}} \mathbb{E}_{M^*}^{\otimes N} \left(\mathcal{E}_{hinge}(\widehat{M}) \right) \geq c \frac{s \max(m,T)}{N}$$

for some universal constants $c > 0$.

Theorem 4.5 provides a minimax lower bound in expectation whereas Theorem 4.4 provides an excess risk bound with large deviation. The two residual terms of the excess hinge risk from Theorem 4.5 and Theorem 4.4 match up to the $\log(m+T)$ factor.

Simulation Study. As the hinge loss has not been often studied in the matrix context, we provide many simulations in order to show the robustness of our method and the opportunity of using the hinge loss rather than the logistic loss. We follow the simulations ran in [18] and compare several methods. An estimator based on the logistic model, studied in [20], is also challenged[3].

A first set of simulations. The simulations are all based on a low-rank 200×200 matrix $M^\star$ from which the data are generated and which is the target for the predictions. $M^\star$ is also a minimizer of $R_{0/1}$ so the error criterion that we will report for a matrix M is the difference of the predictions between $M^\star$ and M, which is $\mathbb{P}[\operatorname{sign}(\langle M^\star, X \rangle) \neq \operatorname{sign}(\langle M, X \rangle)]$. The X_i's correspond to 20% of the entries randomly picked so the misclassification rate is also $1/mT \sum_{p,q} I\{\operatorname{sign}(M_{p,q}) \neq \operatorname{sign}(M^\star_{p,q})\}$.

Two different scenarios are tested: the first one (called A), involves a matrix $M^\star$ with only entries in $\{-1,+1\}$ so the Bayes classifier is low rank and favors the hinge loss. The second test (called B) involves

[3] In the followings, the four estimators will be referred to *Hinge* for estimator given in (24), *Hinge Bayes* and *Logit Bayes* for the two Bayesian estimators from [18] with respectively hinge and logistic loss functions, and *Logit* for the estimator from [20]. The Bayesian estimators use the Gammma prior distribution.

a matrix $M^\star = LR^\top$ where L, R have i.i.d. Gaussian entries and the rank is the number of columns. In this case, the Bayes matrix contains the signs of a low-rank matrix, but it is not itself low rank in general. We also test the impact of the noise structure on the results:

1. (noiseless) $Y_i = \text{sign}(\langle M^\star, X_i \rangle)$
2. (logistic) $Y_i = \text{sign}(\langle M^\star, X_i \rangle + Z_i)$, where Z_i follows a logistic distribution
3. (switch) $Y_i = \epsilon_i \text{sign}(\langle M^\star, X_i \rangle)$ where $\epsilon_i = (1-p)\delta_1 + p\delta_{-1}$

Finally, we run all the simulations on rank 3 and rank 5 matrices. λ is tuned by cross validation. All the simulations are run one time.

	Model	A1	A2 ($p = .1$)	A3	B1	B2 ($p = .1$)	B3
Rank 3	Hinge	0	0	14.5	6.7	10.9	21.0
	Logit	0	0.5	17.3	5.1	10.7	19.8
	Hinge Bayes	0	0.1	8.5	5.3	10.8	22.1
	Logit Bayes	0	0.5	16.0	4.1	10.1	16.0
Rank 5	Hinge	0	0.8	29.0	11.7	19.3	23.3
	Logit	0	3.1	30.1	9.0	18.3	22.1
	Hinge Bayes	0	0.5	27	9.4	17.9	24.4
	Logit Bayes	0	4.4	32.5	7.8	17.3	21.5

Table 2: Misclassification error rates on simulated matrices in various cases. Model $\in \{A, B\}\{1, 2, 3\}$ refers to scenario $\in \{A, B\}$ and noise structure $\in \{1, 2, 3\}$. For the noise-free Model $= A0$, the 0 column shows the exact reconstruction property of all procedures.

The results are very similar among the methods, see Table 2. The logistic loss performs better for matrices of type B and especially for high level of noise in the logistic data generation as expected. For type A matrices, the hinge loss performs slightly better. The Bayesian models performs as good as the frequentist estimators even though the program solved is not convex.

Impact of the noise level. The second experiment is a focus on the switch noise and matrices that are well separated (as A2 in the previous example). The noise lies between $p = 0$ and almost full noise ($p = .4$). The performance of the RERM with the hinge loss is slightly worse than the Bayesian estimator with hinge loss but always better than the RERM with the logistic loss, see Figure 4.3.

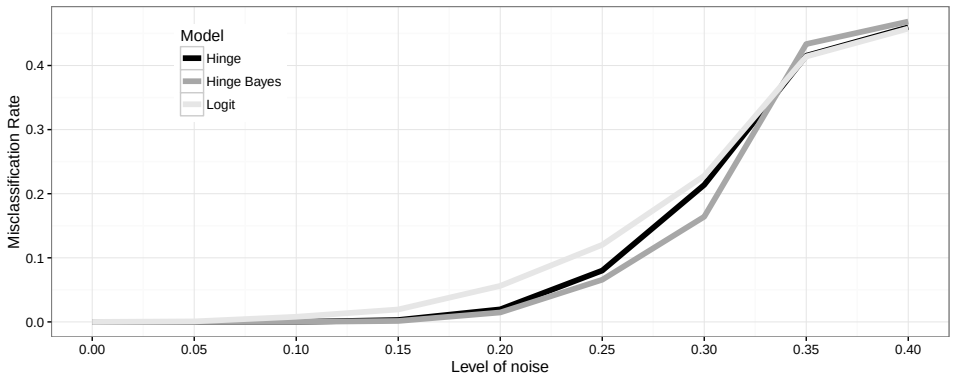

Figure 2: Misclassification error rates for a large range of switch noise (noise structure number 3).

Real dataset. We finally run the hinge loss estimator on the MovieLens dataset. The ratings, that lie in $\{1,2,3,4,5\}$, are split between good ratings $(4,5)$ and bad ratings (others). The goal is therefore to predict whether the user will like a movie or not. On a test set that contains 20% of the data, the misclassification rate in prediction are almost the same for all the methods (Table 3).

Model	Hinge Bayes	Logit	Hinge
misclassification rate	.28	.27	.28

Table 3: Misclassification Rate on MovieLens 100K dataset

4.4 Quantile loss and median matrix completion

The matrix completion problem with continuous entries has almost always been tackled with a penalized least squares estimator [13, 32, 28, 39, 42], but the use of other loss functions may be very interesting in this case too. Our last result on matrix completion is a result for the quantile loss ρ_τ for $\tau \in (0,1)$. Let us recall that $\rho_\tau(u) = u(\tau - I(u \leq 0))$ for all $u \in \mathbb{R}$ and $\ell_M(x,y) = \rho_\tau(y - \langle M, x \rangle)$. While the aforementionned references provided ways to estimate the conditional mean of $Y|X = E_{p,q}$, here, we thus provide a way to estimate conditional quantiles of order τ. When $\tau = 0.5$, it actually estimates the conditional median, which is known to be an indicator of central tendency that is more robust than the mean in the presence of outliers. On the other hand, for large and small τ's (for example the 0.05 and 0.95 quantiles), this allows to build confidence intervals for $Y|X = E_{p,q}$. Confidence bounds for the entries of matrices in matrix completion problems are something new up to our knowledge.

The following result studies a particular case in which the Bernstein Assumption is proved in Proposition 6.4. Following [62], it assumes that the conditional distribution of Y given X is continuous and that the density is not too small on the domain of interest – this ensures that Bernstein's condition is satisfied with $\kappa = 1$ and A depending on the lower bound on the density, see Section 6 for more details. It can easily be derived for a specific distribution such as Gaussian, Student and even Cauchy. But we also have to assume that $\overline{M} \in bB_\infty$, or in other words $\overline{M} = M^*$, which is a more stringent assumption: in practice, it meands that we should know *a priori* an upper bound b on the quantiles to be estimated.

Theorem 4.6 (Quantile matrix completion)**.** *Assume that Assumption 1.2 holds. Let $b > 0$ and assume that $\overline{M} \in bB_\infty$. Assume that for any (p,q), $Y|X = E_{p,q}$ has a density with respect to the Lebesgue measure, g, and that $g(u) > 1/c$ for some constant $c > 0$ for any u such that $|u - \overline{M}_{i,j}| \leq 2b$. Let $s \in \{1, \dots, \min(m,T)\}$ and assume that there exists a matrix with rank at most s in $\overline{M} + (\rho_s^*/20)B_{S_1}$ where ρ_s^* is defined in (19). Then, with probability at least*

$$1 - \mathbf{C}\exp\left(-\mathbf{C}s\max(m,T)\log(m+T)\right)$$

the estimator

$$\widehat{M} \in \operatorname*{argmin}_{M \in bB_\infty} \left(\frac{1}{N}\sum_{i=1}^{N} \rho_\tau(Y_i - \langle X_i, M \rangle) + \lambda \|M\|_{S_1} \right) \tag{25}$$

with $\lambda = c_0(\underline{c}, \bar{c})\sqrt{\log(m+T)/(N\min(m,T))}$ satisfies

$$\frac{1}{mT}\left\|\widehat{M} - \overline{M}\right\|_{S_1} \leq \mathbf{C}\min\left\{ s\sqrt{\frac{\log(m+T)}{N\min(m,T)}}, \frac{\|\overline{M}\|_{S_1}}{mT} \right\},$$

$$\frac{1}{\sqrt{mT}}\left\|\widehat{M} - \overline{M}\right\|_{S_2} \leq \mathbf{C}\min\left\{ \sqrt{\frac{s(m+T)\log(m+T)}{N}}, \|\overline{M}\|_{S_1}^{\frac{1}{2}} \left(\frac{\log(m+T)}{N\min(m,T)}\right)^{\frac{1}{4}} \right\}$$

$$\mathcal{E}_{quantile}(\widehat{M}) \leq \mathbf{C}\min\left\{ \frac{s(m+T)\log(m+T)}{N}, \|\overline{M}\|_{S_1}\sqrt{\frac{\log(m+T)}{N\min(m,T)}} \right\}.$$

We obtain the same rate as for the penalized least squares estimator that is $\sqrt{s(m+T)\log(m+T)/N}$ (cf. [53, 32]).

Simulation study.

The goal of this part is to challenge the regularized least squares estimator by the RERM with quantile loss. The quantile used here is therefore the median. The main conclusion of our study is that median based estimators are more robust to outliers and noise than mean based estimators. We first test them on simulated datasets and then turn to use a real dataset.

Simulated matrices. The observations come from a base matrix $M^\star$ which is a 200×200 low rank matrix. It is built by $M^\star = LR^\top$ where the entries of L, R are i.i.d. gaussian and L, R have 3 columns (and therefore, the rank of $M^\star$ is 3). The X_i's correspond to 20% randomly picked entries. The criterion that we retain is the l_1 reconstruction of $M^\star$ that is: $1/mT \sum_{p,q} |M^\star_{p,q} - M_{p,q}|$.

The observations are made according to this flexible model:

$$Y_i = \langle M^\star, X \rangle + z_i + o\zeta_i.$$

z_i is the noise, o is the magnitude of outliers and ζ_i is the outlier indicator parametrized by the share p such that $\zeta_i = p/2\delta_{-1} + (1-p)\delta_0 + p/2\delta_1$. The different parameters for the different scenarios are summarized in Table 4.

On the first experiment, p is fixed to 10% and the magnitude o increases. As expected for least squares, the results are better for low magnitude of outliers (it corresponds to the penalized maximum likelihood estimator), see Figure 3. Quickly, the performance of the least squares estimator is getting worse and when the outliers are large enough, the best least squares predictor is a matrix with null entries. In opposite to this estimator, the median of the distribution is almost not affected by outliers and it is completely in line with the results: the performances are strictly the same for mid-range to high-range magnitude of outliers. The robustness of the quantile reconstruction is totally independent to the magnitude of the outliers.

	z_i	o	ζ_i
Figure 3	$\mathcal{N}(0, 1/4)$	$o = 0..30$	$p = 0.1$
Figure 4	$\mathcal{N}(0, 1/4)$	10	$p = 0..0.25$
Figure 5	$t_\alpha, \alpha = 1..10$	0	$p = 0$

t_α: t-distribution with α degrees of freedom.

Table 4: Parameters and distributions of the simulations

A second experiment involves fixed magnitude of outliers but the share of them increases, see Figure 4. The median completion is, as expected, more robust and the results deteriorate less than the ones from least squares. When the outliers ratio is greater than 20%, the least squares estimator completely fails while the median completion still works.

The third simulation involves non gaussian noise without outliers: we use the t-distribution, that has heavy tails. In this challenge, a lower degree of freedom involves heavier tails and the worst case is for Student distribution with degree 1. We can see that the least squares is inadequate for small degrees of freedom (1 to 2) and behaves better than the median completion for larger degrees of freedom, see Figure 5.

Real dataset. The last experiment involves the MovieLens dataset. We keep one fifth of the sample for test set to check the prediction accuracy. Even though the least squares estimator remains very efficient in the standard case, see Table 5, the results are quite similar for the MAE criterion. In a second step, we add artificial outliers. In order to do that, we change 20% of 5 ratings to 1 ratings. It can be seen as malicious users that change ratings in order to distort the perception of some movies. As expected, it depreciates

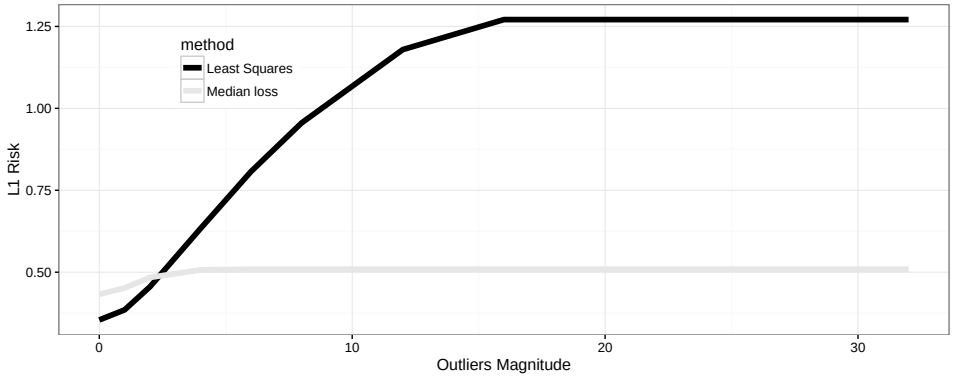

Figure 3: l_1 reconstruction for different magnitude of outliers

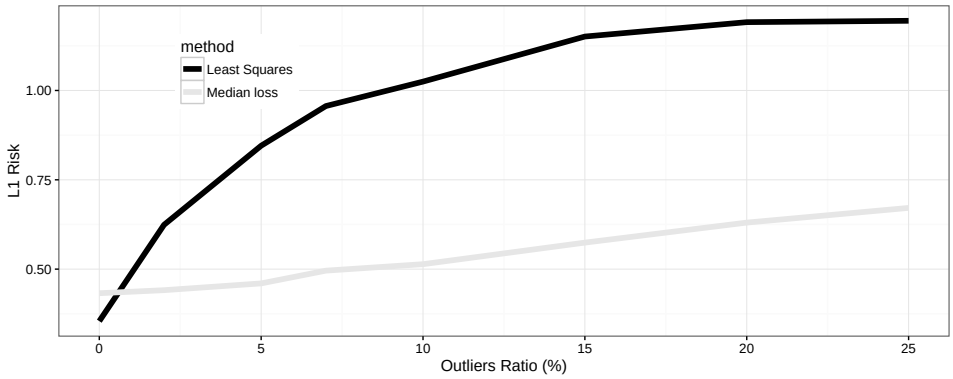

Figure 4: l_1 reconstruction for different percentage of outliers

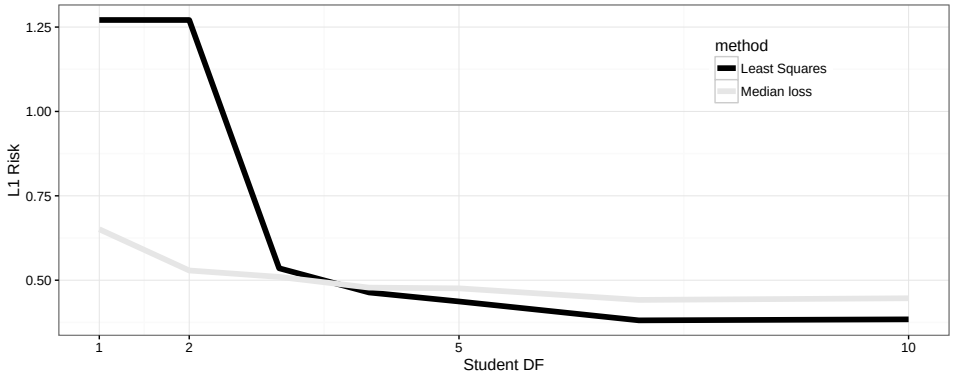

Figure 5: l_1 reconstruction for student noise with various magnitude degrees of freedom

the least squares estimator performance but the median estimator returns almost as good performances as in the standard case.

5 Kernel methods via the hinge loss and a RKHS-norm regularization

In this section, we consider regularization methods in some general Reproducing Kernel Hilbert Space (RKHS) (cf. [19], Chapter 4 in [56] or Chapter 3 of [65] for general references on RKHS).

Unlike the previous examples, the regularization norm here, which is the norm $\|\cdot\|_{\mathcal{H}_K}$ of a RKHS $\mathcal{H}_K$, is not associated with some "hidden" concept of sparsity. In particular, RKHS norms have no singularity since they are differentiable at any point except in 0. As a consequence the sparsity parameter $\Delta(\rho)$ cannot be larger than $4\rho/5$, i.e. ρ does not satisfy the sparsity equation, unless the set $\Gamma_{f^*}(\rho)$ contains 0 that is

	MSE	MAE
Raw Data, LS	0.89	0.75
Raw Data, Median	0.93	0.75
Outliers, LS	1.04	0.84
Outliers, Median	0.96	0.78

Table 5: Prediction power of Least Squares and Median Loss on MovieLens 100K dataset

for $\rho \geq 20 \|f^*\|_{\mathcal{H}_K}$. Indeed, one key observation is that any norm is non differentiable at 0 and that its subdifferential at 0 is somehow extremal:

$$\partial \|\cdot\| (0) = B_* := \{f : \|f\|_* \leq 1\}, \tag{26}$$

where $\|\cdot\|_*$ is the dual norm.

As a consequence, the rates obtained in this section do not depend on some *hidden sparsity parameter* associated with the oracle f^* but on the RKHS norm at f^*, that is $\|f^*\|_{\mathcal{H}_K}$. The aim of this section is therefore to show that our main results apply beyond "sparsity inducing regularization methods" by showing that "classic" regularization method, inducing smoothness for instance, may also be analyzed the same way and fall into the scope of Theorem 2.1 and Theorem 2.2. This section also shows an explicit expression for the Gaussian mean-width with localization as used in Definition 8.1 (a sharper way to measure statistical complexity via a local $r(\cdot)$ function provided below).

Mathematical background In this setup, the data are still N i.i.d. pairs $(X_i, Y_i)_{i=1}^N$ where the X_i's take their values in some set $\mathcal{X}$ and $Y_i \in \{-1, +1\}$. A "similarity measure" is provided over the set $\mathcal{X}$ by means of a kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ so that $x_1, x_2 \in \mathcal{X}$ are "similar" when $K(x_1, x_2)$ is small. One can think for instance of $\mathcal{X}$ the set of all DNA sequences (that is finite words over the alphabet $\{A, T, C, G\}$) and $K(w_1, w_2)$ is the minimal number of changes like insertion, deletion and mutation needed to transform word $w_1 \in \mathcal{X}$ into word $w_2 \in \mathcal{X}$.

The core idea behind kernel methods is to transport the design data X_i's from $\mathcal{X}$ to a Hilbert space via the application $x \to K(x, \cdot)$ and then construct statistical procedures based on the "transported" dataset $(K(X_i, \cdot), Y_i)_{i=1}^N$. The advantage of doing so is that the space where the $K(X_i, \cdot)$'s belong have much structure than the initial set $\mathcal{X}$ which may have no algebraic structure at all. The first thing to set is to define somehow the "smallest" Hilbert space containing all the functions $x \to K(x, \cdot)$. We recall now one classic way of doing so that will be used later to define the objects that need to be considered in order to construct RERM in this setup and to obtain estimation rates for them via Theorem 2.1 and Theorem 2.2.

Recall that if $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a positive definite kernel such that $\|K\|_{L_2} < \infty$, then by Mercer's theorem, there is an orthogonal basis $(\phi_i)_{i \in \mathbb{N}}$ of L_2 such that $\mu \otimes \mu$-almost surely, $K(x, x') = \sum_{i=1}^{\infty} \lambda_i \phi_i(x) \phi_i(x')$ where $(\lambda_i)_{i \in \mathbb{N}}$ is the sequence of eigenvalues of the positive self-adjoint integral operator T_K (arranged in a non-increasing order) defined for every $f \in L_2$ and μ-almost every $x \in \mathcal{X}$ by

$$(T_K f)(x) = \int K(x, x') f(x') d\mu(x').$$

In particular, for all $i \in \mathbb{N}$, ϕ_i is an eigenvector of T_K corresponding to the eigenvalue λ_i; and $(\phi_i)_i$ is an orthonormal system in L_2.

The reproducing kernel Hilbert space $\mathcal{H}_K$ is the set of all function series $\sum_{i=1}^{\infty} a_i K(x_i, \cdot)$ converging in L_2 endowed with the inner product

$$\Big\langle \sum a_i K(x_i, \cdot), \sum b_j K(x'_j, \cdot) \Big\rangle = \sum_{i,j} a_i b_j K(x_i, x'_j)$$

where a_i, b_j's are any real numbers and the x_i's and x'_j's are any points in $\mathcal{X}$.

Estimator. The RKHS $\mathcal{H}_K$ is therefore a class of functions from $\mathcal{X}$ to $\mathbb{R}$ that can be used as a learning model and the norm naturally associated to its Hilbert structure can be used as a regularization function. Given a Lipschitz loss function ℓ, the oracle is defined as

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

and it is believed that $\|f^*\|_{\mathcal{H}_K}$ is small which justified the use of the RERM with regularization function given by the RKHS norm $\|\cdot\|_{\mathcal{H}_K}$:

$$\hat{f} \in \operatorname*{argmin}_{f\in\mathcal{H}_K} \left(\frac{1}{N}\sum_{i=1}^N \ell_f(X_i,Y_i) + \lambda \|f\|_{\mathcal{H}_K} \right)$$

Statistical properties of this RERM may be obtained from Theorem 2.1 in the subgaussian case and from Theorem 2.2 in the bounded case. To that end, we only have to compute the Gaussian mean width and/or the Rademacher complexities of $B_{\mathcal{H}_K}$. In this example, we rather compute the localized version of those quantities because it is possible to derive explicit formula. They are obtained by intersecting the ball with $r\mathcal{E}$. In order not to induce any confusion, we still use the global ones in estimation bounds.

Localized complexity parameter. The goal is to compute $w(\rho B_{\mathcal{H}_K} \cap r\mathcal{E})$ and $\mathrm{Rad}(\rho B_{\mathcal{H}_K} \cap r\mathcal{E})$ for all $\rho, r > 0$ where $B_{\mathcal{H}_K} = \{f \in \mathcal{H}_K : \|f\|_{\mathcal{H}_K} \leq 1\}$ is the unit ball of the RKHS and $\mathcal{E} = \{f \in \mathcal{H}_K : \mathbb{E}f(X)^2 \leq 1\}$ is the ellipsoid associated with X. In the following, we embed the two sets $B_{\mathcal{H}_K}$ and $\mathcal{E}$ in $l_2 = l_2(\mathbb{N})$ so that we simply have to compute the Gaussian mean width and the Rademacher complexities of the intersection of two ellipsoids sharing the same coordinates structure.

The unit ball of $\mathcal{H}_K$ can be constructed from the eigenvalue decomposition of T_K by considering the feature map $\Phi : \mathcal{X} \to l_2$ defined by $\Phi(x) = \big(\sqrt{\lambda_i}\phi_i(x)\big)_{i\in\mathbb{N}}$ and then the unit ball of $\mathcal{H}_K$ is just

$$B_{\mathcal{H}_K} = \big\{ f_\beta(\cdot) = \big\langle \beta, \Phi(\cdot) \big\rangle : \|\beta\|_{l_2} \leq 1 \big\}.$$

One can use the feature map Φ to show that there is an isometry between the two Hilbert spaces $\mathcal{H}_K$ and l_2 endowed with the norm $\|\beta\|_K = \big(\sum \beta_i^2/\lambda_i\big)^{1/2}$. The unit ball of l_2 endowed with the norm $\|\cdot\|_K$ is an ellipsoid denoted by $\mathcal{E}_K$.

Let us now determine the ellipsoid in l_2 associated with the design X obtained via this natural isomorphism $\beta \in l_2 \to f_\beta(\cdot) = \big\langle \beta, \Phi(\cdot) \big\rangle \in \mathcal{H}_K$ between l_2 and $\mathcal{H}_K$. Since $(\phi_i)_i$ is an orthonormal system in L_2, the covariance operator of $\Phi(X)$ in l_2 is simply the diagonal operator with diagonal elements $(\lambda_i)_i$. As a consequence the ellipsoid associated with X is isomorphic to $\tilde{\mathcal{E}} = \{\beta \in l_2 : \mathbb{E}\big\langle \beta, \Phi(X) \big\rangle^2 \leq 1\}$; it has the same coordinate structure as the canonical one in l_2 endowed with $\|\cdot\|_K$: $\tilde{\mathcal{E}} = \{\beta \in l_2 : \sum \lambda_i \beta_i^2 \leq 1\}$. So that, we obtain

$$w(K_\rho(f^*) \cap r\mathcal{E}_{f^*}) = w(\rho\mathcal{E}_K \cap r\tilde{\mathcal{E}}) \sim \left(\sum_j (\rho^2\lambda_j) \wedge r^2 \right)^{1/2} \tag{27}$$

where the last inequality follows from Proposition 2.2.1 in [58] (note that we defined the Gaussian mean widths in Definition (2.4) depending on the covariance of X). We also get from Theorem 2.1 in [48] that

$$\mathrm{Rad}(K_\rho(f^*) \cap r\mathcal{E}_{f^*}) \sim \left(\sum_j (\rho^2\lambda_j) \wedge r^2 \right)^{1/2}. \tag{28}$$

Note that unlike the previous examples, we do not have to assume isotropicity of the design. Indeed, in the RKHS case, the unit ball of the regularization function is isomorphic to the ellipsoid $\mathcal{E}_K$. Since $\mathcal{E}$ is also an ellipsoid having the same coordinates structure as $\mathcal{E}_K$ (cf. paragraph above), for all $\rho, r > 0$, the

intersection $\rho B_{\mathcal{H}_K} \cap r\mathcal{E}$ is equivalent to an ellipsoid, meaning that, it contains an ellipsoid and is contained in a multiple of this ellipsoid. Therefore, the Gaussian mean width and the Rademacher complexity of $\rho B_{\mathcal{H}_K} \cap r\mathcal{E}$ has been computed without assuming isotropicity (thanks to general results on the complexity of Ellipsoids from Proposition 2.2.1 in [58] and Theorem 2.1 in [48]).

It follows from (27) and (28) that the Gaussian mean width and the Rademacher complexities are equal. Therefore, up to constant (L in the subgaussian case and b in the bounded case), the two subgaussian and bounded setups may be analyzed at the same time. Nevertheless, since we will only consider in this setting the hinge loss and that the Bernstein condition (cf. Assumption 2.1) with respect to the hinge loss has been studied in Proposition 6.3 only in the bounded case. We therefore continue the analysis only for the bounded framework.

We are now able to identify the complexity parameter of the problem. We actually do not use the localization in this and rather use only the global complexity parameter as defined in Definition 2.7: for all $\rho > 0$:

$$r(\rho) = \left[\frac{\mathbf{C}\rho \left(\sum_j \lambda_j \right)^{1/2}}{\sqrt{N}} \right]^{\frac{1}{2\kappa}} \tag{29}$$

where $\kappa \geq 1$ is the Bernstein parameter.

Results in the bounded setting Finally, let us discuss about the boundedness assumption. It is known (cf., for instance, Lemma 4.23 in [56]) that if the kernel K is bounded then the functions in the RKHS $\mathcal{H}_K$ are bounded: for any $f \in \mathcal{H}_K$, $\|f\|_{L_\infty} \leq \|K\|_\infty \|f\|_{\mathcal{H}_K}$ where $\|K\|_\infty := \sup_{x \in \mathcal{X}} \sqrt{K(x,x)}$. As a consequence, if one restricts the search space of the RERM to a RKHS ball of radius R, one has $F := RB_{\mathcal{H}_K} \subset \|K\|_\infty B_{L_\infty}$ and therefore the boundedness assumption is satisfied by F. However, note that a refinement of the proof of Theorem 8.2 using a boundedness parameter b depending on the radius of the RKHS balls used while performing the peeling device yields statistical properties for the RERM with no search space constraint. For the sake of shortness, we do not provide this analysis here.

We are now in a position to provide estimation and prediction results for the RERM

$$\hat{f} \in \operatorname*{argmin}_{f \in RB_{\mathcal{H}_K}} \left(\frac{1}{N} \sum_{i=1}^{N} (1 - Y_i f(X_i))_+ + \frac{\mathbf{C} \left(\sum_j \lambda_j \right)^{1/2}}{\sqrt{N}} \|f\|_{\mathcal{H}_K} \right) \tag{30}$$

where the choice of the regularization parameter λ follows from Theorem (2.2) and (28) (for $r = +\infty$). Note that unlike the examples in the previous sections, we do not have to find some radius ρ^* satisfying the sparsity equation (8) to apply Theorem 2.2 since we simply take $\rho^* = 20 \|f^*\|_{\mathcal{H}_K}$ to insure that $0 \in \Gamma_{f^*}(\rho^*)$.

Theorem 5.1. *Let $\mathcal{X}$ be some space, $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a bounded kernel and denote by $\mathcal{H}_K$ the associated RKHS. Denote by $(\lambda_i)_i$ the sequence of eigenvalues associated to $\mathcal{H}_K$ in L_2. Assume that the Bayes rule $\overline{f}$ from (35) belongs to $RB_{\mathcal{H}_K}$ and that the margin assumption (36) is satisfied for some $\kappa \geq 1$.*

Then the RERM defined in (30) satisfies with probability larger than

$$1 - \mathbf{C} \exp\left(-\mathbf{C} N^{1/2\kappa} \left(\|\overline{f}\|_{\mathcal{H}_K} \left(\sum_j \lambda_j \right)^{1/2} \right)^{(2\kappa-1)/\kappa} \right),$$

that

$$\left\| \hat{f} - \overline{f} \right\|_{L_2} \leq \mathbf{C} \left[\frac{\|\overline{f}\|_{\mathcal{H}_K} \left(\sum_j \lambda_j \right)^{1/2}}{\sqrt{N}} \right]^{1/2\kappa} \quad \text{and } \mathcal{E}_{hinge}(\hat{f}) \leq \mathbf{C} \frac{\|\overline{f}\|_{\mathcal{H}_K} \left(\sum_j \lambda_j \right)^{1/2}}{\sqrt{N}}$$

where $\mathcal{E}(\hat{f})$ is the excess hinge risk of $\hat{f}$.

Note that classic procedures in the literature on RKHS are mostly developed in the classification framework. They are usually based on the hinge loss and the regularization function is the square of the RKHS norm. For such procedures, oracle inequalities have been obtained in Chapter 7 from [56] under the margin assumption (cf. [60]). A result that is close to the one obtained in Theorem 5.1 is Corollary 4.12 in [50]. Assuming that $\|Y\|_\infty \leq \mathbf{C}$, $\mathcal{X} \subset \mathbb{R}^d$, $\|K\|_\infty \leq 1$, that the eigenvalues of the integral operator satisfies

$$\lambda_i \leq ci^{-1/p} \tag{31}$$

for some $0 < p < 1$ and that the eigenvectors (ϕ_i) are such that $\|\phi_i\|_\infty \leq A$ for any i and some constant A then the RERM $\tilde{f}$ over the entire RKHS space, w.r.t. the quadratic loss and for a regularization function of the order of (up to logarithmic terms)

$$f \mapsto \rho(\|f\|_{\mathcal{H}}) := \max\left(\frac{\|f\|_{\mathcal{H}}^{2p/(1+p)}}{N^{1/(1+p)}}, \frac{\|f\|_{\mathcal{H}}^2}{N}\right) \tag{32}$$

satisfies with large probability an oracle inequality like

$$\mathbb{E}(Y - \tilde{f}(X))^2 \leq \inf_{r \geq 1}\left(\inf_{\|f\|_{\mathcal{H}} \leq r} \mathbb{E}(Y - f(X))^2 + \mathbf{C}\rho(r)\right).$$

In particular, an error bound (up to log factors) follows from this result: with high probability,

$$\left\|\tilde{f} - f^*\right\|_{L_2}^2 \leq \mathbf{C}\rho(\|f^*\|_{\mathcal{H}}) = \mathbf{C}\max\left(\frac{\|f^*\|_{\mathcal{H}}^{2p/(1+p)}}{N^{1/(1+p)}}, \frac{\|f^*\|_{\mathcal{H}}^2}{N}\right). \tag{33}$$

One may compare this result to the one from Theorem 5.1 under assumption (31) even though the two procedures $\tilde{f}$ and $\hat{f}$ use different loss functions, regularization function and different search space. If assumption (31) holds then $\left(\sum_j \lambda_j\right)^{1/2} \leq c$ and so, one can take $r(\rho) = \left(\mathbf{C}c\rho/(\theta\sqrt{N})\right)^{1/(2\kappa)}$ and $\lambda = \mathbf{C}\sqrt{C/N}$. For such a choice of regularization parameter, Theorem 5.1 provides an error bound of the order of

$$\left\|\tilde{f} - \overline{f}\right\|_{L_2(\mu)}^2 \leq \mathbf{C}\left[\frac{\|\overline{f}\|_{\mathcal{H}_K} C}{\sqrt{N}}\right]^{1/\kappa} \tag{34}$$

which is almost the same as the one obtained in (33) when $\kappa = 1$ and p is close to 1. But our result is worse when $\kappa > 1$ and p is far from 1. This is the price that we pay by using the hinge loss – note that the quadratic loss satisfies the Bernstein condition with $\kappa = 1$ – and by fixing a regularization function which is the norm $\|\cdot\|_{\mathcal{H}_K}$ instead of fitting the regularization function in a "complexity dependent way" as in (32). In the last case, our procedure $\hat{f}$ does not benefit from the "real complexity" of the problem which is localized Rademacher complexities – note that we used global Rademacher complexities to fit λ and construct the complexity function $r(\cdot)$.

6 A review of the Bernstein and margin conditions

In order to apply the main results from Theorem 2.1 and Theorem 2.2, one has to check the Bernstein condition. This section is devoted to the study of this condition for three loss functions: the hinge loss, the quantile loss and the logistic loss. This condition has been extensively studied in Learning theory (cf. [5, 66, 49, 7, 64, 23]). We can identify mainly two approaches to study this condition: when the class F is convex and the loss function ℓ is "strongly convex", then the risk function inherits this property and automatically satisfies the Bernstein condition (cf. [5]). On the other hand, for loss functions like the hinge or quantile loss, that are affine by parts, one has to use a different path. In such cases, one may go back to a statistical framework and try to check the margin assumption. As a consequence, in the latter case, the Bernstein condition is usually more restrictive and requires strong assumptions on the distribution of the observations.

6.1 Logistic loss

In this section, we study the Bernstein condition of the logistic loss function which is defined for every $f : \mathcal{X} \to \mathbb{R}$, $x \in \mathcal{X}$, $y \in \{-1, 1\}$ and $u \in \mathbb{R}$ by

$$\ell_f(x, y) = \tilde{\ell}(yf(x)) \text{ where } \tilde{\ell}(u) = \log(1 + \exp(-u)).$$

Function $\tilde{\ell}$ is strongly convex on every compact interval in $\mathbb{R}$. As it was first observed in [5, 6], one may use this property to check the Bernstein condition for the loss function ℓ. This approach was extended to the bounded regression problem with respect to L_p loss functions ($1 < p < \infty$) in [47] and to non convex classes in [49].

In the bounded scenario, [6] proved that the logistic loss function satisfies the Bernstein condition for $\kappa = 1$. One may therefore use that result to apply Theorem 2.2. The analysis is pretty straightforward in the bounded case. It becomes more delicate in the subgaussian scenario as considered in Theorem 2.1.

Proposition 6.1 ([5])**.** *Let F be a convex class of functions from $\mathcal{X}$ to $\mathbb{R}$. Assume that for every $f \in F$, $\|f\|_{L_\infty} \leq b$. Then the class F satisfies the Bernstein condition with Bernstein parameter $\kappa = 1$ and constant $A = 4\exp(2b)$.*

This result solves the problem of the Bernstein condition with respect to the logistic loss function over a convex class F of functions as long as all functions in F are uniformly bounded by some constant b. We will therefore use this result only in the bounded framework, for instance, when F is a class of linear functional indexed by a bounded set of vectors and when the design takes its values in the canonical basis.

In the subgaussian framework, one may proceed as in [64] and assume that a statistical model holds. In that case, the Bernstein condition is reduced to the study of the Margin assumption since, in that case, the "Bayes rule" $\overline{f}$ (which is called the log-odds ratio in the case of the logistic loss function) is assumed to belong to the class F and so $f^* = \overline{f}$. The margin assumption with respect to the logistic loss function has been studied in Example 1 from [64] but for a slightly different definition of the Margin assumption. Indeed, in [64] only functions f in a L_∞ neighborhood of $\overline{f}$ needs to satisfy the Margin assumption whereas in Assumption 2.1 it has to be satisfied in the non-bounded set $\mathcal{C}$.

From our perspective, we do not want to make no "statistical modeling assumption". In particular, we do not want to assume that $\overline{f}$ belongs to F. We therefore have to prove the Bernstein condition when $\overline{f}$ may not belong to F. We used this result in Section 3 in order to obtain statistical bounds for the Logistic LASSO and Logistic Slope procedures. In those cases, F is a class of linear functionals. We now state that the Bernstein condition is satisfied for a class of linear functional when X is a standard Gaussian vector.

Proposition 6.2. *Let $F = \{\langle \cdot, t \rangle : t \in RB_{l_2}\}$ be a class of linear functionals indexed by RB_{l_2} for some radius $R \geq 1$. Let X be a standard Gaussian vector in $\mathbb{R}^d$ and let Y be a $\{-1, 1\}$ random variable. For every $f \in F$, the excess logistic risk of f, denoted by $P\mathcal{L}_f$, satisfies*

$$\mathcal{E}_{logistic}(f) = P\mathcal{L}_f \geq \frac{c_0}{R^3} \|f - f^*\|_{L_2}^2$$

where c_0 is some absolute constant.

6.2 Hinge loss

Unlike the logistic loss function, both the hinge loss and the quantile losses does not enjoy a strong convexity property. Therefore, one has to turn to a different approach as the one used in the previous section to check the Bernstein condition for those two loss functions.

For the hinge loss function, Bernstein condition is more stringent and is connected to the margin condition in classification. So, let us first introduce some notations specific to classification. In this setup, one is given N labeled pairs $(X_i, Y_i), i = 1, \ldots, N$ where X_i takes its values in $\mathcal{X}$ and Y_i is a label taking values in $\{-1, +1\}$. The aim is to predict the label Y associated with X from the data when (X, Y) is

distributed like the (X_i, Y_i)'s. The classic loss function considered in this setup is the $0-1$ loss function $\ell_f(x,y) = I(y \neq f(x))$ defined for any $f : \mathcal{X} \to \{-1, +1\}$. The $0-1$ loss function is not convex, this may result in some computational issues when dealing with it. A classic approach is to use a " convex relaxation function" as a surrogate to the $0-1$ loss function: note that this is a way to motivate the introduction of the hinge loss $\ell_f(x,y) = \max(1 - yf(x), 0)$. It is well known that the Bayes rules minimizes both the standard $0-1$ risk as well as the hinge risk: put $\eta(x) := \mathbb{E}[Y|X = x]$ for all $x \in \mathcal{X}$ and define the Bayes rule as

$$\overline{f}(x) = \operatorname{sgn}(\eta(x)), \tag{35}$$

then $\overline{f}$ minimizes $f \to P\ell_f$ over all measurable functions from $\mathcal{X}$ to $\mathbb{R}$ when ℓ_f is the hinge loss of f.

Let F be a class of functions from $\mathcal{X}$ to $[-1, 1]$. Assume that $\overline{f} \in F$ so that $\overline{f}$ is an oracle in F and thus (using the notations from Section 2) $f^* = \overline{f}$. In this situation, Margin assumption with respect to the hinge loss (cf. [60, 35]) restricted to the class F and Bernstein condition (cf. Assumption 2.1) coincide. Therefore, Assumption 2.1 holds when the Margin assumption w.r.t. the hinge loss holds. According to Proposition 1 in [35], the Margin assumption with respect to the hinge loss is equivalent the Margin assumption with respect to the $0-1$ loss for a class F of functions with values in $[-1, 1]$. Then, according to Proposition 1 in [60] and [10] the margin assumption with respect to the $0-1$ loss with parameter κ is equivalent to

$$\begin{cases} \mathbb{P}(|\eta(X)| \leq t) \leq ct^{\frac{1}{\kappa-1}}, \forall 0 \leq t \leq 1 & \text{when } \kappa > 1 \\ |\eta(X)| \geq \tau \text{ a.s. for some } \tau > 0 & \text{when } \kappa = 1. \end{cases} \tag{36}$$

As a consequence, one can state the following result on the Bernstein condition for the hinge loss in the bounded case scenario.

Proposition 6.3 (Proposition 1, [35]). *Let F be a class of functions from $\mathcal{X}$ to $[-1, 1]$. Define $\eta(x) = \mathbb{E}[Y|X = x]$ for all $x \in \mathcal{X}$ and assume that the Bayes rule (35) belongs to F. If (36) is satisfied for some $\kappa \geq 1$ then Assumption 2.1 holds with parameter κ for the hinge loss, and A depending on c, κ and τ (which is explicitly given in the mentioned references). In the special case when $\kappa = 1$ then $A = 1/(2\tau)$.*

Note that up to a modification of the constant A, the same result holds for functions with values in $[-b, b]$ for $b > 0$, a fact we used in Section 5.

6.3 Quantile loss

In this section, we study the Bernstein parameter of the **quantile loss** in the bounded regression model, that is when for all $f \in F, \|f\|_{L_\infty} \leq b$ a.s.. Let $\tau \in (0, 1)$ and, for all $x \in \mathcal{X}$, define $\overline{f}(x)$ as the quantile of order τ of $Y|X = x$ and assume that $\overline{f}$ belongs to F, in that case, $\overline{f} = f^*$ and Bernstein condition and margin assumption are the same. Therefore one may follow the study of the margin assumption for the quantile loss in [23] to obtain the following result.

Proposition 6.4 ([23]). *Assume that for any $x \in \mathcal{X}$, it is possible to define a density f_x w.r.t the Lebesgue measure for $Y|X = x$ such that $f_x(u) \geq 1/C$ for some $C > 0$ for all $u \in \mathbb{R}$ with $|u - f^*(x)| \leq 2b$. Then the quantile loss satisfies the Bernstein's assumption with $\kappa = 1$ and $A = 2C$ over F.*

7 Discussion

This paper covers many aspects of the regularized empirical risk estimator (RERM) with Lipschitz loss. This property is commonly shared by many loss functions used in practice such as the hinge loss, the logistic loss or the quantile regression loss. This work offers a general method to derive estimation bounds as well as excess risk upper bounds. Two main settings are covered: the subgaussian framework and the bounded framework. The first one is illustrated by the classification problem with logistic loss. In particular, minimax rates are achieved when using the SLOPE regularization norm. The second framework is used to derive new results on matrix completion and in kernel methods.

A possible extension of this work is to study other regularization norms. In order to do that, one has to compute the complexity parameter in one of the settings and a solution of the sparsity equation. The latter usually involves to understand the sub-differential of the regularization norm and in particular its singularity points which are related to the sparsity equation.

8 Proof of Theorem 2.1 and Theorem 2.2

8.1 More general statements: Theorems 8.2 and 8.1

First, we state two theorems: Theorem 8.1 in the subgaussian setting, and Theorem 8.2 in the bounded setting. These two theorems rely on localized versions of the complexity function $r(\cdot)$ that will be defined first. Note that the localized version of $r(\cdot)$ can always be upper bounded by the simpler version used in the core of the paper. Thus, Theorem 2.1 is a direct corollary of Theorem 8.1, and Theorem 2.2 is a direct corollary of Theorem 8.2.

So let us start with a localized complexity parameters. The "statistical size" of the family of "sub-models" $(\rho B)_{\rho>0}$ is now measured by local Gaussian mean-widths in the subgaussian framework.

Definition 8.1. *Let $\theta > 0$. The **complexity parameter** is a non-decreasing function $r(\cdot)$ such that for every $\rho \geq 0$,*

$$CLw\left(\rho B \cap r(\rho) B_{L_2}\right) \leq \theta r(\rho)^{2\kappa}\sqrt{N}$$

In the boundedness case, it is written as follows.

Definition 8.2. *Let $\theta > 0$. The **complexity parameter** is a non-decreasing function $r(\cdot)$ such that for every $\rho \geq 0$,*

$$48\mathrm{Rad}\left(\rho B \cap r(\rho) B_{L_2}\right) \leq \theta r(\rho)^{2\kappa}\sqrt{N}$$

where κ is the Bernstein parameter from Assumption 2.1.

To obtain the complexity functions from Definition 2.5 and 2.7, we use the fact that $w\left(\rho B \cap r(\rho) B_{L_2}\right) \leq w(\rho B)$ and $\mathrm{Rad}(\rho B \cap r(\rho) B_{L_2}) \leq \mathrm{Rad}(\rho B)$: it indeed does not use the localization. We also set $\theta = 7/40A$ in those definitions because it is the largest value allowed in the following theorems.

Theorem 8.1. *Assume that Assumption 1.1, Assumption 2.1 and Assumption 2.2 hold where $r(\cdot)$ is a function as in Definition 8.1 for some θ such that $40A\theta \leq 7$ and assume that $\rho \to r(2\rho)/\rho$ is non-increasing. Let the regularization parameter λ be chosen such that*

$$\frac{10\theta r(2\rho)^{2\kappa}}{7\rho} < \lambda < \frac{r(2\rho)^{2\kappa}}{2A\rho}, \quad \forall \rho \geq \rho^* \tag{37}$$

where ρ^* satisfies (8). Then, with probability larger than

$$1 - \sum_{j=0}^{\infty} \sum_{i \in I_j} \exp\left(-\frac{\theta^2 N (2^{(i-1)\vee 0} r(2^j \rho^*))^{4\kappa-2}}{4C^2 L^2} \right) \tag{38}$$

where for all $j \in \mathbb{N}$, $I_j = \{1\} \cup \{i \in \mathbb{N}^* : 2^{i-1} r(2^j \rho^*) \leq 2^j \rho^* d_{L_2}(B)\}$, we have

$$\left\| \hat{f} - f^* \right\| \leq \rho^*, \quad \left\| \hat{f} - f^* \right\|_{L_2} \leq r(2\rho^*) \text{ and } \mathcal{E}(\hat{f}) \leq r(2\rho^*)^{2\kappa}/A.$$

Proof of Theorem 2.1: Let $r(\cdot)$ be chosen as in (2.5). For this choice, one can check that the regularization parameter used for the construction of the RERM satisfies (37) with an adequate constant choice. Moreover, for this choice of function $r(\cdot)$ it is straightforward to lower bound the sum in the probability estimate in (38). The parameter λ is chosen in the middle of the range. ∎

The bounded case is in the same spirit.

Theorem 8.2. *Assume that Assumption 1.1, Assumption 2.1 and Assumption 2.3 hold where $r(\cdot)$ is a function as in Definition 8.2 for some θ such that $40A\theta \leq 7$ and assume that $\rho \to r(2\rho)/\rho$ is non-increasing. Let the regularization parameter λ be chosen such that*

$$\frac{10\theta r(2\rho)^{2\kappa}}{7\rho} < \lambda < \frac{r(2\rho)^{2\kappa}}{2A\rho}, \quad \forall \rho \geq \rho^* \tag{39}$$

where ρ^* satisfies (8). Then, with probability larger than

$$1 - 2\sum_{j=0}^{\infty}\sum_{i\in I_j} \exp\left(-c_0\theta^2 N(2^i r(2^{j+1}\rho^*))^{4\kappa-2}\right) \tag{40}$$

where $c_0 = 1/\max\left(48, 207\theta b^{2\kappa-1}\right)$ and for all $j \in \mathbb{N}$, $I_j := \{1\} \cup \{i \in \mathbb{N}^* : 2^{i-1} r(2^j\rho^*) \leq \min(2^j\rho^* d_{L_2}(B), b)\}$, we have

$$\left\|\hat{f} - f^*\right\| \leq \rho^*, \quad \left\|\hat{f} - f^*\right\|_{L_2} \leq r(2\rho^*) \text{ and } \mathcal{E}(\hat{f}) \leq r(2\rho^*)^{2\kappa}/A.$$

The proof of Theorem 2.2 is identical to the one of Theorem 2.1 and we do not reproduce it here.

8.2 Proofs of Theorems 8.2 and 8.1

Proof of Theorem 8.1 and and Theorem 8.2 follow the same strategy. They are split into two parts. First, we identify an event onto which the statistical behavior of the regularized estimator $\hat{f}$ can be controlled using only deterministic arguments. Then, we prove that this event holds with a probability at least as large as the one in (38) in the case of Theorem 8.1 and as in (40) in the case of Theorem 8.2. We first introduce this event which is common to the subgaussian and the bounded setups:

$$\Omega_0 := \left\{ \left|(P - P_N)\mathcal{L}_f\right| \leq \theta \max\left(r(2\max(\|f - f^*\|, \rho^*))^{2\kappa}, \|f - f^*\|_{L_2}^{2\kappa}\right) : \text{ for all } f \in F \right\}$$

where θ is a parameter appearing in the definition of $r(\cdot)$ in Definition 8.1 and Definition 8.2, $\kappa \geq 1$ is the Bernstein parameter from Definition 2.1 and ρ^* is a radius satisfying the sparsity Equation (8).

Proposition 8.1. *Let λ be as in (37) (or equivalently as in (39)) and let ρ^* satisfy (8), on the event Ω_0, one has*

$$\left\|\hat{f} - f^*\right\| \leq \rho^*, \quad \left\|\hat{f} - f^*\right\|_{L_2} \leq r(2\rho^*) \text{ and } \mathcal{E}(\hat{f}) \leq \theta r(2\rho^*)^{2\kappa}.$$

Proof. Denote $\hat{\rho} = \left\|\hat{f} - f^*\right\|$. We first prove that $\hat{\rho} < \rho^*$. To that end, we assume that the reverse inequality holds and show some contradiction. Assume that $\hat{\rho} \geq \rho^*$. Since $\rho \to r(2\rho)/\rho$ is non-increasing then by Lemma A.1, $\rho \to \Delta(\rho)/\rho$ is non-decreasing and so we have

$$\frac{\Delta(\hat{\rho})}{\hat{\rho}} \geq \frac{\Delta(\rho^*)}{\rho^*} \geq \frac{4}{5}.$$

Now, we consider two cases: either $\left\|\hat{f} - f^*\right\|_{L_2} \leq r(2\hat{\rho})$ or $\left\|\hat{f} - f^*\right\|_{L_2} > r(2\hat{\rho})$.

First assume that $\left\|\hat{f} - f^*\right\|_{L_2} \leq r(2\hat{\rho})$. Since $\Delta(\hat{\rho}) \geq 4\hat{\rho}/5$ and $h = \hat{f} - f^* \in \hat{\rho}S \cap r(2\hat{\rho})B_{L_2}$, it follows from the definition of the sparsity parameter $\Delta(\hat{\rho})$ that there exists some $f \in F$ such that $\|f - f^*\| \leq \hat{\rho}/20$ and for which

$$\|f + h\| - \|f\| \geq \frac{4\hat{\rho}}{5}.$$

It follows that

$$\left\|\hat{f}\right\| - \|f^*\| = \|f^* + h\| - \|f^*\| \geq \|f + h\| - \|f\| - 2\|f - f^*\| \geq \frac{4\hat{\rho}}{5} - \frac{\hat{\rho}}{10} = \frac{7\hat{\rho}}{10}.$$

Let us now introduce the excess regularized loss: for all $f \in F$,

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

On the event Ω_0, we have

$$\begin{aligned} P_N\mathcal{L}_{\hat{f}}^\lambda &= P_N\mathcal{L}_{\hat{f}} + \lambda\left(\left\|\hat{f}\right\| - \|f^*\|\right) \geq (P_N - P)\mathcal{L}_{\hat{f}} + \lambda\left(\left\|\hat{f}\right\| - \|f^*\|\right) \\ &\geq -\theta\max\left(r(2\hat{\rho})^{2\kappa}, \left\|\hat{f} - f^*\right\|_{L_2}^{2\kappa}\right) + \frac{7\lambda\hat{\rho}}{10} = -\theta r(2\hat{\rho})^{2\kappa} + \frac{7\lambda\hat{\rho}}{10} > 0 \end{aligned}$$

because by definition of λ, $7\lambda\hat{\rho} > 10\theta r(2\hat{\rho})^{2\kappa}$. Therefore, $P_N\mathcal{L}_{\hat{f}}^\lambda > 0$. But, by construction, one has $P_N\mathcal{L}_{\hat{f}}^\lambda \leq 0$.

Then, assume that $\left\|\hat{f} - f^*\right\|_{L_2} > r(2\hat{\rho})$. In particular, $f \in \mathcal{C}$ where $\mathcal{C}$ is the set introduced in 7 below Assumption 2.1. By definition of $\hat{f}$ we have $P_N\mathcal{L}_{\hat{f}}^\lambda \leq 0$ so it follows from Assumption 2.1 that

$$\begin{aligned} &\left\|\hat{f} - f^*\right\|_{L_2}^{2\kappa} \leq AP\mathcal{L}_{\hat{f}} = A\left[(P - P_N)\mathcal{L}_{\hat{f}} + P_N\mathcal{L}_{\hat{f}}^\lambda + \lambda\left(\|f^*\| - \left\|\hat{f}\right\|\right)\right] \\ &\leq A\theta\max\left(r(2\hat{\rho})^{2\kappa}, \left\|\hat{f} - f^*\right\|_{L_2}^{2\kappa}\right) + A\lambda\left\|\hat{f} - f^*\right\| = A\theta\left\|\hat{f} - f^*\right\|_{L_2}^{2\kappa} + A\lambda\hat{\rho}. \end{aligned} \tag{41}$$

Hence, if $A\theta \leq 1/2$ then

$$r(2\hat{\rho})^{2\kappa} \leq \left\|\hat{f} - f^*\right\|_{L_2}^{2\kappa} \leq 2A\lambda\hat{\rho}.$$

But, by definition of λ one has $r(2\hat{\rho})^{2\kappa} > 2A\lambda\hat{\rho}$.

Therefore, none of the two cases is possible when one assumes that $\hat{\rho} \geq \rho^*$ and so we necessarily have $\hat{\rho} < \rho^*$.

Now, assuming that $\left\|\hat{f} - f^*\right\|_{L_2} > r(2\rho^*)$ and following (41) step by step also leads to a contradiction, so $\left\|\hat{f} - f^*\right\|_{L_2} \leq r(2\rho^*)$.

Next, we prove the result for the excess risk. One has

$$\begin{aligned} &P_N\mathcal{L}_{\hat{f}}^\lambda = P_N\mathcal{L}_{\hat{f}} + \lambda\left(\left\|\hat{f}\right\| - \|f^*\|\right) = (P_N - P)\mathcal{L}_{\hat{f}} + P\mathcal{L}_{\hat{f}} + \lambda\left(\left\|\hat{f}\right\| - \|f^*\|\right) \\ &\geq -\theta\max\left(r(2\rho^*)^{2\kappa}, \left\|\hat{f} - f^*\right\|_{L_2}^{2\kappa}\right) + P\mathcal{L}_{\hat{f}} - \lambda\hat{\rho} \geq -\theta r(2\rho^*)^{2\kappa} - \lambda\rho^* + P\mathcal{L}_{\hat{f}} \\ &\geq -\left(\theta + \frac{1}{2A}\right)r(2\rho^*)^{2\kappa} + P\mathcal{L}_{\hat{f}} \geq \frac{-r(2\rho^*)^{2\kappa}}{A} + P\mathcal{L}_{\hat{f}}. \end{aligned}$$

In particular, if $P\mathcal{L}_{\hat{f}} > r(2\rho^*)^{2\kappa}/A$ then $P_N\mathcal{L}_{\hat{f}}^\lambda > 0$ which is not possible by construction of $\hat{f}$ so we necessarily have $P\mathcal{L}_{\hat{f}} \leq r(2\rho^*)^{2\kappa}/A$. ∎

Proposition 8.1 shows that $\hat{f}$ satisfies some estimation and prediction properties on the event Ω_0. Next, we prove that Ω_0 holds with large probability in both subgaussian and bounded frameworks. We start with the subgaussian framework. To that end, we introduce several tools.

Recall that the ψ_2-norm of a real valued random variable Z is defined by

$$\|Z\|_{\psi_2} = \inf\{c > 0 : \mathbb{E}\psi_2(|Z|/c) \leq \psi_2(1)\}$$

where $\psi_2(u) = \exp(u^2) - 1$ for all $u \geq 0$. The space L_{ψ_2} of all real valued random variables with finite ψ_2-norm is called the Orlicz space of subgaussian variables. We refer the reader to [51, 52] for more details on Orlicz spaces.

We recall several facts on the ψ_2-norm and subgaussian processes. First, it follows from Theorem 1.1.5 from [16] that $\|Z\|_{\psi_2} \leq \max(K_0, K_1)$ if

$$\mathbb{E}\exp(\lambda|Z|) \leq \exp\left(\lambda^2 K_1^2\right), \quad \forall \lambda \geq 1/K_0. \tag{42}$$

It follows from Lemma 1.2.2 from [16] that, if Z is a centered ψ_2 random variable then, for all $\lambda > 0$,

$$\mathbb{E}\exp\left(\lambda Z\right) \leq \exp\left(e\lambda^2 \|Z\|_{\psi_2}^2\right). \tag{43}$$

Then, it follows from Theorem 1.2.1 from [16] that if $Z_1, \ldots, Z_N$ are independent centered real valued random variables then

$$\left\|\sum_{i=1}^N Z_i\right\|_{\psi_2} \leq 16\left(\sum_{i=1}^N \|Z_i\|_{\psi_2}^2\right)^{1/2}. \tag{44}$$

Finally, let us turn to some properties of subgaussian processes. Let (T, d) be a pseudo-metric space. Let $(X_t)_{t\in T}$ be a random process in L_{ψ_2} such that for all $s, t \in T$, $\|X_t - X_s\|_{\psi_2} \leq d(s,t)$. It follows from the comment below Theorem 11.2 p.300 in [41] that for all measurable set A and all $s, t \in T$,

$$\int_A |X_s - X_t| d\mathbb{P} \leq d(s,t)\mathbb{P}(A)\psi_2^{-1}\left(\frac{1}{\mathbb{P}(A)}\right).$$

Therefore, it follows from equation (11.14) in [41] that for every $u > 0$,

$$\mathbb{P}\left(\sup_{s,t\in T} |X_s - X_t| > c_0(\gamma_2 + Du)\right) \leq \psi_2(u)^{-1} \tag{45}$$

where D is the diameter of (T, d), c_0 is an absolute constant and γ_2 is the majorizing measure integral $\gamma(T, d; \psi_2)$ (cf. Chapter 11 in [41]). When T is a subset of L_2 and d is the natural metric of L_2 it follows from the majorizing measure theorem that $\gamma_2 \leq c_1 w(T)$ (cf. Chapter 1 in [58]).

Lemma 8.1. *Assume that Assumption 1.1 and Assumption 2.2 hold. Let $F' \subset F$ then for every $u > 0$, with probability at least $1 - 2\exp(-u^2)$*

$$\sup_{f,g\in F'} |(P - P_N)(\mathcal{L}_f - \mathcal{L}_g)| \leq \frac{c_0 L}{\sqrt{N}}\left(w(F') + u d_{L_2}(F')\right)$$

where d is the L_2 metric and $d_{L_2}(F')$ is the diameter of (F', d).

Proof. To prove Lemma 8.1, it is enough to show that $((P - P_N)\mathcal{L}_f)_{f\in F'}$ has $(L/\sqrt{N})$-subgaussian increments and then to apply (45) where $\gamma_2 \sim w(F')$ in this case.

Let us prove that for some absolute constant c_0: for all $f, g \in F'$,

$$\|(P - P_N)(\mathcal{L}_f - \mathcal{L}_g)\|_{\psi_2} \leq c_0(L/\sqrt{N})\|f - g\|_{L_2}$$

It follows from (44) that

$$\|(P - P_N)(\mathcal{L}_f - \mathcal{L}_g)\|_{\psi_2} \leq 16\left(\sum_{i=1}^N \frac{\|(\mathcal{L}_f - \mathcal{L}_g)(X_i, Y_i) - \mathbb{E}(\mathcal{L}_f - \mathcal{L}_g)\|_{\psi_2}^2}{N^2}\right)^{1/2} = \frac{16}{\sqrt{N}}\|\zeta_{f,g}\|_{\psi_2}.$$

where $\zeta_{f,g} = (\mathcal{L}_f - \mathcal{L}_g)(X, Y) - \mathbb{E}(\mathcal{L}_f - \mathcal{L}_g)$.Therefore, it only remains to show that $\|\zeta_{f,g}\|_{\psi_2} \leq c_1 L\|f - g\|_{L_2}$.

It follows from (42), that the last inequality holds if one proves that for all $\lambda \geq c_1/(L\|f-g\|_{L_2})$,

$$\mathbb{E}\exp\left(\lambda|\zeta_{f,g}|\right) \leq \exp(c_2\lambda^2 L^2\|f-g\|_{L_2}^2) \tag{46}$$

for some absolute constants c_1 and c_2. To that end, it is enough to prove that, for some absolute constant c_3 – depending only on c_1 and c_2 – and all $\lambda > 0$,

$$\mathbb{E}\exp\left(\lambda|\zeta_{f,g}|\right) \leq 2\exp(c_3\lambda^2 L^2\|f-g\|_{L_2}^2).$$

Note that if Z is a real valued random variable and ϵ is a Rademacher variable independent of Z then $\mathbb{E}\exp(|Z|) \leq 2\exp(\epsilon Z)$. Hence, it follows from a symmetrization argument (cf. Lemma 6.3 in [41]), (a simple version of) the contraction principle (cf. Theorem 4.4 in [41]) and (43) that, for all $\lambda > 0$,

$$\begin{aligned}\mathbb{E}\exp\left(\lambda|\zeta_{f,g}|\right) &\leq 2\mathbb{E}\exp(\lambda\epsilon\zeta_{f,g}) \leq 2\mathbb{E}\exp\left(2\lambda\epsilon(\mathcal{L}_f-\mathcal{L}_g)(X,Y)\right)\\ &\leq 2\mathbb{E}\exp\left(2\lambda\epsilon(f-g)(X)\right) \leq 2\mathbb{E}\exp\left(c_4\lambda^2L^2\|f-g\|_{\psi_2}^2\right)\end{aligned}$$

where ϵ is a Rademacher variable independent of (X,Y) and where we used in the last but one inequality that $|\mathcal{L}_f(X,Y)-\mathcal{L}_g(X,Y)| \leq |f(X)-g(X)|$ a.s.. ∎

Proposition 8.2. *We assume that Assumption 1.1, 2.2 and 2.1 hold. Then the probability measure of Ω_0 is at least as large as the one in (38).*

Proof. The proof is based on a peeling argument (cf. [63]) with respect to the two distances naturally associated with this problem: the regularization norm $\|\cdot\|$ and the L_2-norm $\|\cdot\|_{L_2}$ associated with the design X. The peeling according to $\|\cdot\|$ is performed along the radii $\rho_j = 2^j\rho^*$ for $j \in \mathbb{N}$ and the peeling according to $\|\cdot\|_{L_2}$ is performed within the class $\{f \in F : \|f-f^*\| \leq \rho_j\} := f^* + \rho_j B$ along the radii $2^i r(\rho_j)$ for all $i = 0,1,2,\cdots$ up to a radius such that $2^i r(\rho_j)$ becomes larger than the radius of $f^*+\rho_j B$ in L_2, that is for all $i \in I_j$.

We introduce the following partition of the class F. We first introduce the "true model", i.e. the subset of F where we want to show that $\hat{f}$ belongs to with high probability:

$$F_{0,0} = \left\{f \in F : \|f-f^*\| \leq \rho_0 \text{ and } \|f-f^*\|_{L_2} \leq r(\rho_0)\right\}$$

(note that $\rho_0 = \rho^*$). Then we peel the remaining set $F\backslash F_{0,0}$ according to the two norms: for every $i \in I_0$,

$$F_{0,i} = \left\{f \in F : \|f-f^*\| \leq \rho_0 \text{ and } 2^{i-1}r(\rho_0) < \|f-f^*\|_{L_2} \leq 2^i r(\rho_0)\right\},$$

for all $j \geq 1$,

$$F_{j,0} = \left\{f \in F : \rho_{j-1} < \|f-f^*\| \leq \rho_j \text{ and } \|f-f^*\|_{L_2} \leq r(\rho_j)\right\}$$

and for every integer $i \in I_j$,

$$F_{j,i} = \left\{f \in F : \rho_{j-1} < \|f-f^*\| \leq \rho_j \text{ and } 2^{i-1}r(\rho_j) < \|f-f^*\|_{L_2} \leq 2^i r(\rho_j)\right\}.$$

We also consider the sets $F_{j,i}^* = \rho_j B \cap (2^i r(\rho_j))B_{L_2}$ for all integers i and j.

Let j and $i \in I_j$ be two integers. It follows from Lemma 8.1 that for any $u > 0$, with probability larger than $1-2\exp(-u^2)$,

$$\sup_{f\in F_{j,i}} |(P-P_N)\mathcal{L}_f| \leq \sup_{f,g\in F_{j,i}^*+f^*} |(P-P_N)(\mathcal{L}_f-\mathcal{L}_g)| \leq \frac{c_0 L}{\sqrt{N}}\left(w(F_{j,i}^*) + u d_{L_2}(F_{j,i}^*)\right) \tag{47}$$

where $d_{L_2}(F_{j,i}^*) \leq 2^{i+1}r(\rho_j)$.

Note that for any $\rho > 0$, $h : r \to w(\rho B \cap rB_{L_2})/r$ is non-increasing (cf. Lemma A.2 in the Appendix) and note that, by definition of $r(\rho)$ (cf. Definition 8.1), $h(r(\rho)) \leq \theta r(\rho)^{2\kappa-1}\sqrt{N}/(CL)$. Since $h(\cdot)$ is

non-increasing, we have $w(F^*_{j,i})/(2^i r(\rho_j)) \le h(2^i r(\rho_j)) \le h(r(\rho_j)) \le \theta r(\rho_j)^{2\kappa-1}\sqrt{N}/(CL)$ and so $w(F^*_{j,i}) \le \theta 2^i r(\rho_j)^{2\kappa}\sqrt{N}/(CL)$. Therefore, it follows from (47) for $u = \theta\sqrt{N}(2^{(i-1)\vee 0} r(\rho_j))^{2\kappa-1}/(2CL)$, if $C \ge 4c_0$ then, with probability at least

$$1 - 2\exp\left(-\theta^2 N(2^{(i-1)\vee 0} r(\rho_j))^{4\kappa-2}/(4C^2L^2)\right), \tag{48}$$

for every $f \in F_{j,i}$,

$$|(P-P_N)\mathcal{L}_f| \le \theta(2^{(i-1)\vee 0} r(\rho_j))^{2\kappa} \le \theta \max\left(r(2\max(\|f-f^*\|, \rho^*))^{2\kappa}, \|f-f^*\|_{L_2}^{2\kappa}\right).$$

The result follows from a union bound. ∎

Now we turn to the proof of Theorem 8.1 under the boundedness assumption. The proof follows the same strategy as in the "subgaussian case": we first use Proposition 8.1 and then show (under the boundedness assumption) that event Ω_0 holds with probability at least as large as the one in (40).

Similar to Proposition 8.2, we prove the following result under the boundedness assumption.

Proposition 8.3. *We assume that Assumption 1.1, 2.3 and 2.1 hold. Then the probability measure of Ω_0 is at least as large as the one in (40).*

Proof. Using the same notation as in the proof of Proposition 8.2, we have for any integer j and i such that $2^i r(\rho_j) \le b$ that by Talagrand's concentration inequality: for any $x > 0$, with probability larger than $1 - 2e^{-x}$,

$$Z_{j,i} \le 2\mathbb{E}Z_{j,i} + \sigma(\mathcal{L}_{F_{j,i}})\sqrt{\frac{8x}{N}} + \frac{69\left\|\mathcal{L}_{F_{j,i}}\right\|_\infty x}{2N} \tag{49}$$

where

$$Z_{j,i} = \sup_{f\in F_{j,i}} |(P-P_N)\mathcal{L}_f|, \quad \sigma(\mathcal{L}_{F_{j,i}}) = \sup_{f\in F_{j,i}} \sqrt{\mathbb{E}\mathcal{L}_f^2} \text{ and } \left\|\mathcal{L}_{F_{j,i}}\right\|_\infty = \sup_{f\in F_{j,i}} \|\mathcal{L}_f\|_\infty.$$

By the Lipschitz assumption, one has

$$\sigma(\mathcal{L}_{F_{j,i}}) \le 2^{i+1} r(\rho_j) \text{ and } \left\|\mathcal{L}_{F_{j,i}}\right\|_\infty \le 2b.$$

Therefore, it only remains to upper bound the expectation $\mathbb{E}Z_{j,i}$. Let $\epsilon_1, \dots, \epsilon_N$ be a N i.i.d. Rademacher variables independent of the (X_i, Y_i)'s. For all function f, we set

$$P_{N,\epsilon} f = \frac{1}{N}\sum_{i=1}^N \epsilon_i f(X_i)$$

It follows from a symmetrization and a contraction argument (cf. Chapter 4 in [41]) that

$$\mathbb{E}Z_{j,i} \le 4\mathbb{E}\sup_{f\in F_{j,i}} |P_{N,\epsilon}(f-f^*)| \le \frac{4\mathrm{Rad}(\rho_j B \cap (2^i r(\rho_j))B_{L_2})}{\sqrt{N}} \le (\theta/12)2^i r(\rho_j)^{2\kappa}.$$

Now, we take $x = c_2\theta^2 N(2^{i-1} r(\rho_j))^{4\kappa-2}$ in (49) and note that $2^i r(\rho_j) \le b$ and $\kappa \ge 1$: with probability larger than

$$1 - 2\exp(-c_2\theta N(2^i r(\rho_j))^{4\kappa-2}), \tag{50}$$

for any $f \in F_{j,i}$,

$$\begin{aligned}
|(P-P_N)\mathcal{L}_f| &\le \theta 2^{i-1} r(\rho_j)^{2\kappa}/3 + 2\sqrt{8c_2}\theta\left(2^{i-1} r(\rho_j)\right)^{2\kappa} + 69c_2\theta^2 b(2^{i-1} r(\rho_j))^{4\kappa-2} \\
&\le \theta\left(2^{(i-1)\vee 0} r(\rho_j)\right)^{2\kappa}\left[\frac{1}{3} + 2\sqrt{8c_2} + 69c_2\theta b(2^i r(\rho_j))^{2\kappa-2}\right] \\
&\le \theta\left(2^{(i-1)\vee 0} r(\rho_j)\right)^{2\kappa}\left[\frac{1}{3} + 2\sqrt{8c_2} + 69c_2\theta b^{2\kappa-1}\right] \\
&\le \theta\max\left(r(2\max(\|f-f^*\|, \rho^*))^{2\kappa}, \|f-f^*\|_{L_2}^{2\kappa}\right)
\end{aligned}$$

if c_2 is defined by

$$c_2 = \min\left(\frac{1}{48}, \frac{1}{207\theta b^{2\kappa-1}}\right). \tag{51}$$

We conclude with a union bound. ■

9 Proof of Theorem 4.3

For the sake of simplicity, assume that $m \geq T$ so $\max(m,T) = m$. Fix $r \in \{1,\dots,T\}$. Fix $x > 0$ such that $\exp(x)/[1+\exp(x)] \leq b$, we define the set of matrices

$$\mathcal{C}_x = \left\{A \in \mathbb{R}^{m\times r} : \forall (p,q), A_{p,q} \in \{0,x\}\right\}$$

and

$$\mathcal{M}_x = \{A \in \mathbb{R} : A = (B|\dots|B|O), B \in \mathcal{C}_x\}$$

where the block B is repeated $\lfloor T/r \rfloor$ times (this construction is taken from [33]). Varshamov-Gilbert bound (Lemma 2.9 in [61]) implies that there is a finite subset $\mathcal{M}_x^0 \subset \mathcal{M}_x$ with $\mathrm{card}(\mathcal{M}_x^0) \geq 2^{rm/8}+1$ with $0 \in \mathcal{M}_x^0$, and for any distinct $A, B \in \mathcal{M}_x^0$,

$$\|A-B\|_{S_2}^2 \geq \frac{mr\lfloor T/r \rfloor}{8}x^2 \geq \frac{mT}{16}x^2$$

and so

$$\frac{1}{mT}\|A-B\|_{S_2}^2 \geq \frac{x^2}{16}.$$

Then, for $A \in \mathcal{M}_x^0 \setminus \{0\}$,

$$\begin{aligned}
\mathcal{K}(\mathbb{P}_0, \mathbb{P}_A) &= \frac{n}{mT}\sum_{i=1}^{m}\sum_{j=1}^{T}\left[\frac{1}{2}\log\left(\frac{1+\exp(M_{i,j})}{2\exp(M_{i,j})}\right) + \frac{1}{2}\log\left(\frac{1+\exp(M_{i,j})}{2}\right)\right] \\
&= \frac{n}{mT}\sum_{i=1}^{m}\sum_{j=1}^{T}\left[\log\left(\frac{1+\exp(M_{i,j})}{2}\right) - \frac{1}{2}M_{i,j}\right] \\
&\leq n\left[\log\left(\frac{1+\exp(x)}{2}\right) - \frac{1}{2}x\right] \\
&\leq c(b)nx^2
\end{aligned}$$

where $c(b) > 0$ is a constant that depends only on b. So:

$$\frac{1}{\mathrm{card}(\mathcal{M}_x^0)-1}\sum_{A\in\mathcal{M}_x^0}\mathcal{K}(\mathbb{P}_0,\mathbb{P}_A) \leq c(b)nx^2 \leq c(b)\log(\mathrm{card}(\mathcal{M}_x^0)-1)$$

as soon as we choose

$$x \leq \sqrt{\frac{\log(\mathrm{card}(\mathcal{M}_x^0)-1)}{n}} \leq \sqrt{\frac{rm\log(2)}{8n}}$$

(note that the condition $n \geq rm\log(2)/(8b^2)$ implies that $\exp(x)/[1+\exp(x)] \leq b$). Then, Theorem 2.5 in [61] leads to the existence of $\beta, c > 0$ such that

$$\inf_{\widehat{M}}\sup_{A\in\mathcal{M}_x^0}\mathbb{P}_A\left(\frac{1}{mT}\|\widehat{M}-A\|_{S_2}^2 \geq c\frac{mr}{N}\right) \geq \beta.$$

■

10 Proof of Theorem 4.5

For the sake of simplicity, assume that $m \geq T$ so $\max(m,T) = m$. Fix $r \in \{2,\dots,T\}$ and assume that $rT \leq N \leq mT$.

We recall that $\{E_{p,q} : 1 \leq p \leq m, 1 \leq p \leq T\}$ is the canonical basis of $\mathbb{R}^{m\times T}$. We consider the following "blocks of coordinates": for every $1 \leq k \leq r-1$ and $1 \leq l \leq T$,

$$B_{kl} = \left\{ E_{p,l} : \frac{(k-1)mT}{N} + 1 \leq p < \frac{kmT}{N} + 1 \right\}$$

(note that $(r-1)mT/N + 1 \leq m$ when $rT \leq N \leq mT$). We also introduce the "blocks" of "remaining" coordinates:

$$B_0 = \left\{ E_{p,q} : \frac{(r-1)mT}{N} + 1 \leq p, 1 \leq q \leq T \right\}$$

For every $\sigma = (\sigma_{kl}) \in \{0,1\}^{(r-1)\times T}$, we denote by $\mathbb{P}_\sigma$ the probability distribution of a pair (X,Y) taking its values in $\mathbb{R}^{m\times T} \times \{-1,1\}$ where X is uniformly distributed over the basis $\{E_{p,q} : 1 \leq p \leq m, 1 \leq p \leq T\}$ and for every $(p,q) \in \{1,\dots,m\} \times \{1,\dots,T\}$,

$$\mathbb{P}_\sigma[Y = 1 | X = E_{p,q}] = \left\{ \begin{array}{cc} \sigma_{kl} & \text{if } E_{p,q} \in B_{kl} \\ 1 & \text{otherwise.} \end{array} \right.$$

We also introduce $\eta_\sigma(E_{p,q}) = \mathbb{E}[Y = 1 | X = E_{p,q}] = 2\mathbb{P}_\sigma[Y = 1 | X = E_{p,q}] - 1$. It follows from [66] that the Bayes rules minimizes the Hinge risk, that is $f^*_\sigma \in \operatorname{argmin}_f \mathbb{E}_\sigma(Y - f(X))_+$, where the minimum runs over all measurable functions and $\mathbb{E}_\sigma$ denotes the expectation w.r.t. (X,Y) when $(X,Y) \sim \mathbb{P}_\sigma$, is achieved by $f^*_\sigma = \operatorname{sgn}(\eta_\sigma(\cdot))$. Therefore, $f^*_\sigma(\cdot) = \langle M^*_\sigma, \cdot \rangle$ where for every $(p,q) \in \{1,\dots,m\} \times \{1,\dots,T\}$,

$$(M^*_\sigma)_{pq} = \left\{ \begin{array}{cc} 2\sigma_{kl} - 1 & \text{if } E_{p,q} \in B_{kl} \\ 1 & \text{otherwise.} \end{array} \right. = \eta_\sigma(E_{p,q}).$$

In particular, M^*_σ has a rank at most equal to r.

Let $\sigma = (\sigma_{p,q}), \sigma' = (\sigma'_{pq})$ be in $\{0,1\}^{(r-1)T}$. We denote by $\rho(\sigma,\sigma')$ the Hamming distance between σ and σ' (i.e. the number of times the coordinates of σ and σ' are different). We denote by $H(\mathbb{P}_\sigma, \mathbb{P}_{\sigma'})$ the Hellinger distance between the probability measures $\mathbb{P}_\sigma$ and $\mathbb{P}_{\sigma'}$. We have

$$H(\mathbb{P}_\sigma, \mathbb{P}_{\sigma'}) = \int \left(\sqrt{d\mathbb{P}_\sigma} - \sqrt{d\mathbb{P}_{\sigma'}} \right)^2 = \frac{2\rho(\sigma,\sigma')}{N}.$$

Then, if $\rho(\sigma,\sigma') = 1$, it follows that (cf. Section 2.4 in [61]),

$$H^2(\mathbb{P}_\sigma^{\otimes N}, \mathbb{P}_{\sigma'}^{\otimes N}) = 2\left(1 - \left(1 - \frac{H^2(\mathbb{P}_\sigma, \mathbb{P}_{\sigma'})}{2}\right)^N\right) = 2\left(1 - \left(1 - \frac{1}{N}\right)^N\right) \leq 2(1 - e^{-2}) := \alpha.$$

Now, it follows from Theorem 2.12 in [61], that

$$\inf_{\hat\sigma} \max_{\sigma \in \{0,1\}^{(r-1)T}} \mathbb{E}_\sigma^{\otimes N} \left\| \hat\sigma - \sigma \right\|_{l_1} \geq \frac{(r-1)T}{8} \left(1 - \sqrt{\alpha(1 - \alpha/4)}\right) \tag{52}$$

where the infimum $\inf_{\hat\sigma}$ runs over all measurable functions $\hat\sigma$ of the data $(X_i, Y_i)_{i=1}^N$ with values in $\mathbb{R}$ (note that Theorem 2.12 in [61] is stated for functions $\hat\sigma$ taking values in $\{0,1\}^{(r-1)T}$ but its is straightforward to extend this result to any $\hat\sigma$ valued in $\mathbb{R}$) and $\mathbb{E}_\sigma^{\otimes N}$ denotes the expectation w.r.t. those data distributed according to $\mathbb{P}_\sigma^{\otimes N}$.

Now, we lower bound the excess risk of any estimator. Let $\hat{f}$ be an estimator with values in $\mathbb{R}$. Using a truncation argument it is not hard to see that one can restrict the values of $\hat{f}$ to $[-1,1]$. In that case, We have

$$\mathcal{E}_{hinge}(\hat{f}) = \mathbb{E}\left[|2\eta_\sigma(X)-1||\hat{f}(X)-f^*_\sigma(X)|\right] = \mathbb{E}|\hat{f}(X)-f^*_\sigma(X)|$$
$$= \sum_{p,q}|\hat{f}(E_{p,q})-f^*_\sigma(E_{p,q})|\mathbb{P}[X=E_{p,q}] \geq \sum_{kl}\frac{1}{mT}\sum_{E_{p,q}\in B_{kl}}|\hat{f}(E_{p,q})-(2\sigma_{pq}-1)| \geq \frac{2}{N}\sum_{kl}|\hat{\sigma}_{kl}-\sigma_{pq}|$$

where $\hat{\sigma}_{kl}$ is the mean of $\{(\hat{f}(E_{p,q})+1)/2 : E_{p,q}\in B_{kl}\}$. Then we obtain,

$$\inf_{\hat{f}}\sup_{\sigma\in\{0,1\}^{(r-1)T}}\mathbb{E}_\sigma^{\otimes N}\mathcal{E}_{hinge}(\hat{f}) \geq \frac{2}{N}\inf_{\hat{\sigma}}\max_{\sigma\in\{0,1\}^{(r-1)T}}\mathbb{E}_\sigma^{\otimes N}\|\hat{\sigma}-\sigma\|_{l_1}$$

and, using (52), we get

$$\inf_{\hat{f}}\sup_{\sigma\in\{0,1\}^{(r-1)T}}\mathbb{E}_\sigma^{\otimes N}\mathcal{E}_{hinge}(\hat{f}) \geq c_0\frac{rT}{N}$$

for $c_0 = \left(1-\sqrt{\alpha(1-\alpha/4)}\right)/4$. ∎

11 Proofs of Section 6

11.1 Proof of Section 6.1

The proof of Proposition 6.1 may be found in several papers (cf., for instance, [5]). Let us recall this argument since we will be using it at a starting point to prove the Bernstein condition in the subgaussian case.

Proof of Proposition 6.1: The logistic risk of a function $f : \mathcal{X} \to \mathbb{R}$ can be written as $P\ell_f = \mathbb{E}[g(X,f(X))]$ where for all $x,a\in\mathbb{R}$, $g(x,a) := ((1+\eta(x))/2)\log(1+e^{-a}) + ((1-\eta(x))/2)\log(1+e^a)$ and $\eta(x) = \mathbb{E}[Y|X=x]$ is the conditional expectation of Y given $X=x$.

Since f^* minimizes $f \to P\ell_f$ over the convex class F, one has by the first order condition that for every $f\in F$, $\mathbb{E}\partial_2 g(X,f^*(X))(f-f^*)(X) \geq 0$. Therefore, it follows from a second order Taylor expansion that the excess logistic loss of every $f\in F$ is such that

$$\mathcal{E}_{logistic}(f) = P\mathcal{L}_f \geq \mathbb{E}\left[(f(X)-f^*(X))^2\int_0^1(1-u)\delta(f^*(X)+u(f-f^*)(X))du\right] \tag{53}$$

where $\delta(u) = \partial_2^2 g(x,u) = e^u/(1+e^u)^2$ for every $u\in\mathbb{R}$.

Since $|f^*(X)|, |f(X)| \leq b$ a.s. then for every $u\in[0,1]$, $|f^*(X)+u(f-f^*)(X)| \leq 2b$, a.s. and since $\delta(v)\geq\delta(2b)\geq\exp(-2b)/4$ for every $|v|\leq 2b$, it follows from (53) that $P\mathcal{L}_f \geq \delta(2b)\|f-f^*\|_{L_2}^2$. ∎

Proof of Proposition 6.2: Let $t^*\in RB_{l_2}$ be such that $f^* = \langle\cdot,t^*\rangle$, where f^* is an oracle in $F = \{\langle\cdot,t\rangle : t\in RB_{l_2}\}$ w.r.t. the logistic loss risk. Let $f = \langle\cdot,t\rangle\in F$ for some $t\in RB_{l_2}$. It follows from (53) that the excess logistic risk of f satisfies

$$P\mathcal{L}_f \geq \int_0^1 \mathbb{E}\left[\langle X,t^*-t\rangle^2\delta\left(\langle X,t^*+u(t-t^*)\rangle\right)\right]du.$$

The result will follow if one proves that for every $t_0,t\in\mathbb{R}^d$,

$$\mathbb{E}\left[\langle X,t\rangle^2\delta\left(\langle X,t_0\rangle\right)\right] \geq \frac{\min\left(\pi,\pi^2\left(\|t_0\|_2\sqrt{2\pi+\|t_0\|_2^2}\right)^{-1}\right)}{\sqrt{2\pi+\|t_0\|_2^2}+(\pi-1)\|t_0\|_2}\frac{\|t\|_2^2}{8\sqrt{2\pi}}. \tag{54}$$

Let us now prove (54). We write $t = t_0^\perp + \lambda t_0$ where $t_0^\perp$ is a vector orthogonal to t_0 and $\lambda \in \mathbb{R}$. Since $\langle X, t_0^\perp \rangle$ and $\langle X, t_0 \rangle$ are independent random variables, we have

$$\begin{aligned}\mathbb{E}\left[\langle X,t\rangle^2 \delta\left(\langle X,t_0\rangle\right)\right] &= \mathbb{E}\left[\langle X,t_0^\perp\rangle^2\right]\mathbb{E}\left[\delta\left(\langle X,t_0\rangle\right)\right] + \lambda^2 \mathbb{E}\left[\langle X,t_0\rangle^2 \delta\left(\langle X,t_0\rangle\right)\right], \\ &= \left\|t_0^\perp\right\|_2^2 \mathbb{E}\delta(\|t_0\|_2\, g) + \lambda^2 \|t_0\|_2^2\, \mathbb{E} g^2 \delta(\|t_0\|_2\, g)\end{aligned}$$

where $g \sim \mathcal{N}(0,1)$ is standard Gaussian variable and we recall that $\delta(v) = e^v/(1+e^v)^2$ for all $v \in \mathbb{R}$. Now, it remains to lower bound $\mathbb{E}\delta(\sigma g)$ and $\mathbb{E} g^2\delta(\sigma g)$ for every $\sigma > 0$.

Since $\delta(v) \geq \exp(-|v|)/4$ for all $v \in \mathbb{R}$, one has for all $\sigma > 0$,

$$\mathbb{E}\delta(\sigma g) \geq \mathbb{E}\exp(-\sigma|g|)/4 = \exp(\sigma^2/2)\mathbb{P}[g \geq \sigma]/2$$

and

$$\mathbb{E} g^2\delta(\sigma g) \geq \mathbb{E} g^2 \exp(-\sigma|g|)/4 = (1/2)\exp(\sigma^2/2)\left[(1+\sigma^2)\mathbb{P}[g \geq \sigma] - \frac{\sigma \exp(-\sigma^2/2)}{\sqrt{2\pi}}\right].$$

Therefore, for $\sigma = \|t_0\|_2$,

$$\begin{aligned}\mathbb{E}\left[\langle X,t\rangle^2 \delta\left(\langle X,t_0\rangle\right)\right] \geq{}& \exp(\sigma^2/2)\mathbb{P}[g \geq \sigma]\left\|t_0^\perp\right\|_2^2 \\ &+ 2\lambda^2 \|t_0\|_2^2 \exp(\sigma^2/2)\left[\left(1+\sigma^2\right)\mathbb{P}[g \geq \sigma] - \frac{\sigma\exp(-\sigma^2/2)}{\sqrt{2\pi}}\right]\end{aligned}$$

and since $\|t\|_2^2 = \left\|t_0^\perp\right\|_2^2 + \lambda^2\|t_0\|_2^2$, one has,

$$\mathbb{E}\left[\langle X,t\rangle^2 \delta\left(\langle X,t_0\rangle\right)\right] \geq \frac{\|t\|_2^2}{\sqrt{2\pi}} \min\left\{\left(\frac{1-\Phi(\sigma)}{\phi(\sigma)}\right), (1+\sigma^2)\left(\frac{1-\Phi(\sigma)}{\phi(\sigma)}\right) - \sigma\right\} \tag{55}$$

where ϕ and Φ denote the standard Gaussian density and distribution functions, respectively.

We lower bound the right-hand side of (55) using estimates on the Mills ratio $(1-\Phi)/\phi$ that follows from Equation (10) in [22]: for every $\sigma > 0$,

$$\frac{1-\Phi(\sigma)}{\phi(\sigma)} > \frac{\pi}{\sqrt{2\pi+\sigma^2} + (\pi-1)\sigma}.$$

■

11.2 Proof of Section 6.3

Proof of Proposition 6.4: We globally follow a proof of [23]. We have

$$P\mathcal{L}_f = \mathbb{E}[\rho_\tau(Y - f(X)) - \rho_\tau(Y - f^*(X))] = \mathbb{E}\Big\{\mathbb{E}[\rho_\tau(Y - f(X)) - \rho_\tau(Y - f^*(X))|X]\Big\}.$$

For all $x \in \mathcal{X}$, denote by F_x the c.d.f. associated with f_x. We have

$$\begin{aligned}&\mathbb{E}[\rho_\tau(Y - f(X))|X = x] = (\tau-1)\int_{y<f(x)} (y - f(x))F_x(\mathrm{d}y) + \tau\int_{y\geq f(x)} (y-f(x))F_x(\mathrm{d}y) \\ &= \int_{y\geq f(x)} (y-f(x))F_x(\mathrm{d}y) + (\tau-1)\int_{\mathbb{R}} (y-f(x))F_x(\mathrm{d}y) \\ &= \int_{y\geq f(x)} (1-F_x(y))\mathrm{d}y + (\tau-1)\left(\int_{\mathbb{R}} yF_x(\mathrm{d}y) - f(x)\right) = g(x,f(x)) + (\tau-1)\int_{\mathbb{R}} yF_x(\mathrm{d}y)\end{aligned}$$

where $g(x,a) = \int_{y \geq a}(1 - F_x(y))\mathrm{d}y + (1-\tau)a$. Note that $\partial_2 g(x, f^*(x)) = 0$ (can be checked by calculations but also obvious from the definition). So

$$\mathbb{E}[\rho_\tau(Y - f(X)) - \rho_\tau(Y - f^*(X))|X = x] = g(x, f(x)) - g(x, f^*(x)) = \int_{f^*(x)}^{f(x)} (f(x) - u)\partial_2^2 g(x,u)\mathrm{d}u$$
$$= \int_{f^*(x)}^{f(x)} (f(x) - u) f_x(u)\mathrm{d}u \geq \frac{1}{C}\int_{f^*(x)}^{f(x)} (f(x) - u)\mathrm{d}u = \frac{(f(x) - f^*(x))^2}{2C^2}.$$

It follows that

$$\mathcal{E}_{quantile}(f) = P\mathcal{L}_f \geq \mathbb{E}\left\{\frac{(f(X) - f^*(X))^2}{2C}\right\} = \frac{1}{2C}\|f - f^*\|_{L_2}^2.$$

■

A Technical lemmas

Lemma A.1. *If $\rho \to r(2\rho)/\rho$ is non-increasing then $\rho \to \Delta(\rho)/\rho$ is non-decreasing.*

Proof. We have for all $\rho > 0$

$$\frac{\Delta(\rho)}{\rho} = \inf_{H \in S \cap (r(2\rho)/\rho)B_{L_2}} \sup_{G \in \partial\|\cdot\|(M^*)} \big\langle H, G \big\rangle.$$

The result follows since $\rho \to S \cap (r(2\rho)/\rho)B_{L_2}$ is non-increasing. ■

Lemma A.2. *Let $\rho > 0$. The function $h : r > 0 \to w(\rho B \cap rB_{L_2})/r$ is non-increasing.*

Proof. Let $r_1 \geq r_2$. By convexity of B and B_{L_2}, we have

$$(\rho B \cap r_1 B_{L_2})/r_1 = (\rho/r_1)B \cap B_{L_2} \subset (\rho/r_2)B \cap B_{L_2} = (\rho B \cap r_2 B_{L_2})/r_2. \tag{56}$$

■

References

[1] Pierre Alquier, James Ridgway, and Nicolas Chopin. On the properties of variational approximations of gibbs posteriors. *Journal Of Machine Learning Research*, 17(239):1–41, 2016.

[2] Jean-Yves Audibert and Alexandre B. Tsybakov. Fast learning rates for plug-in classifiers. *Ann. Statist.*, 35(2):608–633, 2007.

[3] Franck Barthe, Olivier Guédon, Shahar Mendelson, and Assaf Naor. A probabilistic approach to the geometry of the l_p^n-ball. *Ann. Probab.*, 33(2):480–513, 2005.

[4] Peter L. Bartlett, Olivier Bousquet, and Shahar Mendelson. Local Rademacher complexities. *Ann. Statist.*, 33(4):1497–1537, 2005.

[5] Peter L Bartlett, Michael I Jordan, and Jon D McAuliffe. Large margin classifiers: Convex loss, low noise, and convergence rates. In *NIPS*, pages 1173–1180, 2003.

[6] Peter L Bartlett, Michael I Jordan, and Jon D McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

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

[8] A. Belloni and V. Chernozhukov. ℓ-1-penalized quantile regression in high-dimensional sparse models. *The Annals of Statistics*, 39(1):82–130, 2011.

[9] Małgorzata Bogdan, Ewout van den Berg, Chiara Sabatti, Weijie Su, and Emmanuel J. Candès. SLOPE—adaptive variable selection via convex optimization. *Ann. Appl. Stat.*, 9(3):1103–1140, 2015.

[10] Stéphane Boucheron, Olivier Bousquet, and Gábor Lugosi. Theory of classification: a survey of some recent advances. *ESAIM Probab. Stat.*, 9:323–375, 2005.

[11] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122, 2011.

[12] Jian-Feng Cai, Emmanuel J Candès, and Zuowei Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4):1956–1982, 2010.

[13] E. J. Candès and Y. Plan. Matrix Completion With Noise. *Proceedings of the IEEE*, 98(6):925–936, 2010.

[14] Olivier Catoni. *Statistical learning theory and stochastic optimization: Ecole d'Eté de Probabilités de Saint-Flour, XXXI-2001*, volume 31. Springer, 2004.

[15] Olivier Catoni. Pac-bayesian supervised classification: the thermodynamics of statistical learning. *arXiv preprint arXiv:0712.0248*, 2007.

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

[17] Venkat Chandrasekaran, Benjamin Recht, Pablo A. Parrilo, and Alan S. Willsky. The convex geometry of linear inverse problems. *Found. Comput. Math.*, 12(6):805–849, 2012.

[18] V. Cottet and P. Alquier. 1-bit Matrix Completion: PAC-Bayesian Analysis of a Variational Approximation. *ArXiv e-prints*, April 2016.

[19] Felipe Cucker and Steve Smale. On the mathematical foundations of learning. *Bull. Amer. Math. Soc. (N.S.)*, 39(1):1–49 (electronic), 2002.

[20] Mark A Davenport, Yaniv Plan, Ewout van den Berg, and Mary Wootters. 1-bit matrix completion. *Information and Inference*, 3(3):189–223, 2014.

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

[22] Lutz Dümbgen. Bounding standard gaussian tail probabilities. Technical report, University of Bern, 2010.

[23] Andreas Elsener and Sara van de Geer. Robust low-rank matrix estimation. *arXiv preprint arXiv:1603.09071*, 2016.

[24] Manuel Garcia-Magariños, Anestis Antoniadis, Ricardo Cao, and Wenceslao González-Manteiga. Lasso logistic regression, GSoft and the cyclic coordinate descent algorithm: application to gene expression data. *Stat. Appl. Genet. Mol. Biol.*, 9:Art. 30, 30, 2010.

[25] Nathan Halko, Per-Gunnar Martinsson, and Joel A Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, 53(2):217–288, 2011.

[26] Cho-Jui Hsieh and Peder A Olsen. Nuclear norm minimization via active subspace selection. In *ICML*, pages 575–583, 2014.

[27] Peter J. Huber. Robust estimation of a location parameter. *Ann. Math. Statist.*, 35:73–101, 1964.

[28] O. Klopp. Noisy low-rank matrix completion with general sampling distribution. *Bernoulli*, 20(1):282–303, 2014.

[29] V. Koltchinskii and D. Panchenko. Empirical margin distributions and bounding the generalization error of combined classifiers. *Ann. Statist.*, 30(1):1–50, 2002.

[30] Vladimir Koltchinskii. Local Rademacher complexities and oracle inequalities in risk minimization. *Ann. Statist.*, 34(6):2593–2656, 2006.

[31] Vladimir Koltchinskii. *Oracle inequalities in empirical risk minimization and sparse recovery problems*, volume 2033 of *Lecture Notes in Mathematics*. Springer, Heidelberg, 2011. Lectures from the 38th Probability Summer School held in Saint-Flour, 2008, École d'Été de Probabilités de Saint-Flour. [Saint-Flour Probability Summer School].

[32] Vladimir Koltchinskii, Karim Lounici, and Alexandre B. Tsybakov. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *Ann. Statist.*, 39(5):2302–2329, 2011.

[33] Vladimir Koltchinskii, Karim Lounici, Alexandre B Tsybakov, et al. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics*, 39(5):2302–2329, 2011.

[34] Jean Lafond, Olga Klopp, Eric Moulines, and Joseph Salmon. Probabilistic low-rank matrix completion on finite alphabets. In *Advances in Neural Information Processing Systems*, pages 1727–1735, 2014.

[35] Guillaume Lecué. Optimal rates of aggregation in classification under low noise assumption. *Bernoulli*, 13(4):1000–1022, 2007.

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

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

[38] Guillaume Lecué and Shahar Mendelson. Learning subgaussian classes: Upper and minimax bounds. Technical report, CNRS, Ecole polytechnique and Technion, 2013.

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

[40] Guillaume Lecué and Shahar Mendelson. Regularization and the small-ball method II: complexity dependent error rates. Technical report, CNRS, Ecole Polytechnique and Technion, 2015.

[41] 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.

[42] T. T. Mai and P. Alquier. A bayesian approach for noisy matrix completion: Optimal rate under general sampling distribution. *Electronic Journal of Statistics*, 9:823–841, 2015.

[43] Carmen Mak. *Polychotomous logistic regression via the Lasso*. ProQuest LLC, Ann Arbor, MI, 1999. Thesis (Ph.D.)–University of Toronto (Canada).

[44] E. Mammen and A. Tsybakov. Smooth discrimination analysis. *The Annals of Statistics*, 27(6):1808–1829, 1999.

[45] Rahul Mazumder, Trevor Hastie, and Robert Tibshirani. Spectral regularization algorithms for learning large incomplete matrices. *Journal of machine learning research*, 11(Aug):2287–2322, 2010.

[46] Lukas Meier, Sara van de Geer, and Peter Bühlmann. The group Lasso for logistic regression. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 70(1):53–71, 2008.

[47] Shahar Mendelson. Improving the sample complexity using global data. *IEEE transactions on Information Theory*, 48(7):1977–1991, 2002.

[48] Shahar Mendelson. On the performance of kernel classes. *J. Mach. Learn. Res.*, 4(5):759–771, 2004.

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

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

[51] M. M. Rao and Z. D. Ren. *Theory of Orlicz spaces*, volume 146 of *Monographs and Textbooks in Pure and Applied Mathematics*. Marcel Dekker, Inc., New York, 1991.

[52] M. M. Rao and Z. D. Ren. *Applications of Orlicz spaces*, volume 250 of *Monographs and Textbooks in Pure and Applied Mathematics*. Marcel Dekker, Inc., New York, 2002.

[53] Angelika Rohde and Alexandre B Tsybakov. Estimation of high-dimensional low-rank matrices. *The Annals of Statistics*, 39(2):887–930, 2011.

[54] N. Sabbe, O. Thas, and J.-P. Ottoy. EMLasso: logistic lasso with missing data. *Stat. Med.*, 32(18):3143–3157, 2013.

[55] Nathan Srebro, Jason Rennie, and Tommi S Jaakkola. Maximum-margin matrix factorization. In *Advances in neural information processing systems*, pages 1329–1336, 2004.

[56] Ingo Steinwart and Andreas Christmann. *Support vector machines*. Information Science and Statistics. Springer, New York, 2008.

[57] Weijie Su and Emmanuel Candès. SLOPE is adaptive to unknown sparsity and asymptotically minimax. *Ann. Statist.*, 44(3):1038–1068, 2016.

[58] Michel Talagrand. *The generic chaining*. Springer Monographs in Mathematics. Springer-Verlag, Berlin, 2005. Upper and lower bounds of stochastic processes.

[59] Guo-Liang Tian, Man-Lai Tang, Hong-Bin Fang, and Ming Tan. Efficient methods for estimating constrained parameters with applications to regularized (lasso) logistic regression. *Comput. Statist. Data Anal.*, 52(7):3528–3542, 2008.

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

[61] Alexandre B Tsybakov. *Introduction to nonparametric estimation*. Springer Series in Statistics, 2009.

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

[63] 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.

[64] Sara A Van de Geer. High-dimensional generalized linear models and the lasso. *The Annals of Statistics*, pages 614–645, 2008.

[65] 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.

[66] Tong Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *Annals of Statistics*, pages 56–85, 2004.