The Annals of Statistics
2019, Vol. 47, No. 4, 2117–2144
https://doi.org/10.1214/18-AOS1742
© Institute of Mathematical Statistics, 2019

ESTIMATION BOUNDS AND SHARP ORACLE INEQUALITIES OF REGULARIZED PROCEDURES WITH LIPSCHITZ LOSS FUNCTIONS[1]

BY PIERRE ALQUIER[2], VINCENT COTTET AND GUILLAUME LECUÉ

CREST, CNRS, ENSAE, Université Paris Saclay

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 sub-Gaussian 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, and also shape constrained logistic regression; (3) kernel methods, where the loss is the hinge loss, and the regularization function is the RKHS norm.

1. Introduction. Many classification and regression problems are solved in practice by regularized empirical risk minimizers (RERM). The risk is measured via a loss function. The quadratic loss function is the most popular function for regression. It has been extensively studied (cf. [23, 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

Received January 2018; revised June 2018.

[1] Supported by Labex ECODEC (ANR-11-LABEX-0047).

[2] Research programme *New Challenges for New Data* from LCL and GENES, hosted by the *Fondation du Risque*.

MSC2010 subject classifications. Primary 60K35, 62G08; secondary 62C20, 62G05, 62G20.
Key words and phrases. Empirical processes, high-dimensional statistics.

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 [46].

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 [9] under the name "quantile LASSO" while the same penalty with the logistic loss was studied in [45] under the name "logistic LASSO" (cf. [44]). The ERM strategy with Lipschitz proxys of the $0/1$ loss are studied in [25]. The loss functions we will consider in the examples of this paper are:

1. *hinge loss*: $\ell_f(x, y) = (1 - yf(x))_+ = \max(0, 1 - yf(x))$ for every $y \in \{-1, +1\}, x \in \mathcal{X}, f : \mathcal{X} \to \mathbb{R}$,
2. *logistic loss*: $\ell_f(x, y) = \log(1 + \exp(-yf(x)))$ for every $y \in \{-1, +1\}$, $x \in \mathcal{X}, f : \mathcal{X} \to \mathbb{R}$;
3. *quantile regression loss*: for some parameter $\tau \in (0, 1)$, $\ell_f(x, y) = \rho_\tau(y - f(x))$ for every $y \in \mathbb{R}, x \in \mathcal{X}, f : \mathcal{X} \to \mathbb{R}$ 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 sub-Gaussian; 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 (note that the subdifferential of a norm at a nonzero point is a subset of the dual sphere. We informally say that it is large when it covers a large part of the dual sphere without quantifying it).

To be more precise, the method works for any regularization function as long as it is a norm. Following [30], our approach will show a connection between the excess risk bounds and the size of the subdifferential of the regularization norm around the best predictor (the oracle). For example, when the oracle f^* is sparse, the subdifferential of the ℓ_1 norm at f^* is large, and the excess risk bound is small. We refer to these bounds as *sparsity dependent bounds*. In general, good excess risk bounds will be obtained using a regularizer that has some "sparsity inducing power," like the ℓ_1 or nuclear norms, and this will be expressed through the size of the subdifferential of the regularizer around the oracle.

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 ℓ_2 estimation rate achieved by the logistic SLOPE estimator is the classical rate $s\log(p/s)/N$. 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. Note that all those results are obtained for a random design.

A remarkable fact is that no assumption on the output Y is needed (while most results for the quadratic loss rely on—restrictive—assumptions 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.[3]

Mathematical background and notation. 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: $\ell(f(x), y)$. It will actually be convenient to use the notation $\ell_f(x, y) = \ell(f(x), y)$. 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 $y \in \mathcal{Y}$, we have $|\ell(f_1(x), y) - \ell(f_2(x), y)| \le |f_1(x) - f_2(x)|$.

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.

[3] Of course, if Y and X are independent, our results are valid but useless. Our prediction risk bounds (like Theorem 2.1) state that we learn to predict Y by the best possible $f(X)$ for f in a given class F. If there is no $f \in F$ such that $f(X)$ predicts Y well, our results are useless. On the other hand, we point out that it is not necessary to make (restrictive) parametric assumptions on (X, Y) to ensure that there is a function f in a given class F that will lead to acceptable predictions.

REMARK 1.1. Examples were provided above: quantile losses, the hinge loss or the Huber loss. Note that assuming that for any $f \in F$, $\|f\|_\infty \le C_f < \infty$ and that $|Y| \le C_Y$ is bounded a.s, the squared loss $\ell(f(X), Y) = (Y - f(X))^2$ satisfies $|(y - f_1(x))^2 - (y - f_2(x))^2| = |2y - f_1(x) - f_2(x)||f_1(x) - f_2(x)| \le 2(C_F + C_Y)|f_1(x) - f_2(x)|$. It is then possible to use our results in this context. However, we do not recommend this in general: this case excludes classical examples such as Gaussian noise. Our study was partly motivated by [31] that was dedicated to the square loss: in [31], sparse linear regression is covered with a wide set of noises, including Gaussian but also heavy-tailed noise.

We define the oracle predictor as

$$f^* \in \operatorname*{argmin}_{f \in F} P\ell_f \qquad \text{where}^{4}\ P\ell_f = \mathbb{E}\ell_f(X, Y)$$

and (X, Y) is distributed like the (X_i, Y_i)'s. One of the objectives 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 $\mathcal{L}_f = \ell_f - \ell_{f^*}$ 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 an estimator is usually called a Regularized Empirical Risk Minimization procedure (RERM).

For the rest of the paper, we will use the following notation: let rB and rS denote the radius r ball and sphere for the norm $\|\cdot\|$, that is, $rB = \{f \in E : \|f\| \le r\}$ and $rS = \{f \in E : \|f\| = r\}$. For the L_2-norm, we write $rB_{L_2} = \{f \in L_2 : \|f\|_{L_2} \le r\}$ 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, that is, $E = \{\langle t, \cdot\rangle, t \in \mathbb{R}^p\}$. 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 $\langle A, B\rangle = \operatorname{Trace}(A^\top B)$ for

[4] 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 $f_0 \in F$. For any $f \in F$, $\mathbb{E}[\ell_f(X, Y) - \ell_{f_0}(X, Y)]] \le \mathbb{E}|(f - f_0)(X)| < \infty$ 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 definition $f^* \in \operatorname{argmin}_{f \in F} P\ell_f$ when $P\ell_f$ is finite for some $f \in F$.

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 $a \sim b$.

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 [31], and a stochastic assumption on F (sub-Gaussian or bounded). Our two general theorems themselves are eventually presented; note that the proofs of the two main theorems (and extended versions of them) are postponed to Section 9 of Supplement A [3]. 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 7 of Supplement A. For matrix completion, the minimax-optimality of the rates for the logistic and the hinge loss, that were not known, is also stated in Section 4; note that the proof of the optimality is postponed to Section 10 in Supplement A and that an extensive simulation study[5] may also be found in Section 6 of Supplement A. In Section 8 of Supplement A, we discuss the Bernstein condition for the three main loss functions of interest: hinge, logistic and quantile (the corresponding proofs are in Section 11 of Supplement A). Finally, Section 12 of Supplement A contains the study of the (nonpenalized) ERM, that is, the case $\lambda = 0$ under the same assumptions. We also provide a short application to shape-constrained estimation in Section 12.

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 statistician 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 subsection is first to provide the definition of these objects and some hints on their interpretation, through examples. The main theorems are then stated in both settings. Basically, the assumptions for the theorems are of three types:

[5] The code may be downloaded on the page https://sites.google.com/site/vincentcottet/code.

1. The so-called Bernstein condition, which is a quantification of the identifiability condition or a curvature assumption of the objective function $f \to P\ell_f$ at its minimum f^*. Formally, it relates the excess risk $\mathcal{E}(f) = P\mathcal{L}_f = P(\ell_f - \ell_{f^*})$ to the L_2 norm $\|f - f^*\|_{L_2}$ through an inequality of the form $P\mathcal{L}_f \gtrsim \|f - f^*\|_{L_2}^{2\kappa}$.

2. A stochastic assumption on the distribution of the $f(X)$'s for $f \in F$. In this work, we consider both a sub-Gaussian 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 2.5 and Definition 2.7].

3. Finally, we consider the sparsity parameter as introduced in [31]. 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 follows:

1. Find the *Bernstein parameters* $\kappa \geq 1$ and $A > 0$ associated to the loss and the class F;

2. Compute the *Complexity function*

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

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

3. Compute the subdifferential $\partial\|\cdot\|(f^*)$ 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 $\Delta(\rho^*) \geq 4\rho^*/5$", where $\Delta(\cdot)$ is defined in Definition 2.1 below.

4. Apply Theorem 2.1 in the sub-Gaussian framework and Theorem 2.2 in the bounded framework. In each case, with large probability,

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

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 losses.

ASSUMPTION 2.1 (Bernstein condition). There exist $\kappa \geq 1$ 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 8 in Supplement A. In particular, 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 [34, 43]: 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 8 in Supplement A and to the discussions in [28] and Chapter 1.3 in [27] 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 exist $\kappa \geq 1$ 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,

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

Note that the set $\mathcal{C}$ appears to play a central role in the analysis of regularization methods; cf. [31]. 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

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

where A is the constant in Assumption 2.1 and where $\mathrm{comp}(B)$ 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, $\mathrm{comp}(B) = C\,\mathrm{Rad}(B)$ where C is an absolute constant and $\mathrm{Rad}(B)$ is the Rademacher complexity of B (whose definition will be reminded in Section 2.6). In the sub-Gaussian setting, $\mathrm{comp}(B) = CLw(B)$ where C is an absolute constant, L is the sub-Gaussian 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 Section 2.5].

Note that sharper (localized) versions of $r(\cdot)$ are provided in Section 9 in Supplement A. However, as it is the simplest version that is used in most examples, we only introduce this (global) version for now.

2.4. *The sparsity parameter* ρ^*. The size of the subdifferential of the regularization function $\|\cdot\|$ in a neighborhood of the oracle f^* play a central role in our analysis. We recall now its definition: for every $f \in F$

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

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 7 in Supplement A; or when one wants extra *norm dependent* upper bounds (cf. [30] 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 [31]). The *sparsity parameter* is the function $\Delta(\cdot)$ defined for any $\rho > 0$ 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 [31] 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 details in our analysis (a peeling argument that we use below whereas a direct homogeneity argument was enough in [31]).

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 [31].

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

$$\Delta(\rho^*) \geq (4/5)\rho^*. \tag{3}$$

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 $f^* + (\rho/20)B$ [i.e., an element f for which

$\partial\|\cdot\|(f)$ is almost extremal—that is almost as large as the entire dual sphere] yields the existence of a small ρ^* satisfying 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 (3). We will use this observation to obtain *norm dependent* upper bounds, that is, 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 7 in Supplement A.

2.5. *Theorem in the sub-Gaussian setting*. First, we introduce the sub-Gaussian framework (then we will turn to the bounded case in the next section).

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

$$\mathbb{E}\exp\big(\lambda\big|f(X)\big|/\|f\|_{L_2}\big) \leq \exp\big(\lambda^2 L^2\big), \tag{4}$$

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

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

$$F' + f = \{f' + f : f' \in F'\}, \qquad F' - F' = \{f'_1 - f'_2 : f'_1, f'_2 \in F'\}$$

and $d_{L_2}(F') = \sup(\|f'_1 - f'_2\|_{L_2} : f'_1, f'_2 \in F')$.

ASSUMPTION 2.2. The class $F - F$ is L-sub-Gaussian.

Note that there are many equivalent formulations of the sub-Gaussian 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 [15]). The one we will use later is as follows: there exists some absolute constant $\mathbf{C}$ such that $F - F$ is L-sub-Gaussian if and only if for all $f, g \in F$ and $t \geq 1$,

$$\mathbb{P}\big[\big|f(X) - g(X)\big| \geq \mathbf{C}tL\|f - g\|_{L_2}\big] \leq 2\exp\big(-t^2\big). \tag{5}$$

There are several examples of sub-Gaussian 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-sub-Gaussian 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 sub-Gaussian, 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. [5]),
4. $X = (x_j)_{j=1}^p$ is an unconditional vector [meaning that for every signs $(\varepsilon_j)_j \in \{-1, +1\}^p$, $(\varepsilon_j x_j)_{j=1}^p$ 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. [29]).

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

We refer the reader to Section 12 in [18] for the construction of Gaussian processes in L_2. There are many natural situations where Gaussian mean-widths can be derived explicitly; cf. [20] or the examples in Section 3.

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

DEFINITION 2.5. The *complexity parameter* is the nondecreasing function $r(\cdot)$ defined for every $\rho \geq 0$ by

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

where κ (the Bernstein parameter) and A are defined in Assumption 2.1, L is the sub-Gaussian 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 9.2 in Supplement A).

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 (3). Then, with probability larger than

$$1 - \mathbf{C} \exp(-\mathbf{C} N^{1/2\kappa} (\rho^* w(B))^{(2\kappa - 1)/\kappa}) \tag{6}$$

we have

$$\|\hat{f} - f^*\| \le \rho^*, \qquad \|\hat{f} - f^*\|_{L_2} \le r(2\rho^*) = \left[\frac{ACLw(B)2\rho^*}{\sqrt{N}}\right]^{1/2\kappa}$$

and

$$\mathcal{E}(\hat{f}) \le \frac{r(2\rho^*)^{2\kappa}}{A} = \frac{CLw(B)2\rho^*}{\sqrt{N}},$$

where **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

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

we have in particular

$$\|\hat{f} - f^*\|_{L_2} \le r(2\rho^*) = \left[\frac{ACL[w(B) + x]2\rho^*}{\sqrt{N}}\right]^{1/2\kappa}$$

and

$$\mathcal{E}(\hat{f}) \le \frac{r(2\rho^*)^{2\kappa}}{A} = \frac{CL[w(B) + x]2\rho^*}{\sqrt{N}}.$$

REMARK 2.3 (Norm and sparsity dependent error rates). Theorem 2.1 holds for any radius ρ^* satisfying the sparsity equation (3). 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 (3) (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 [23] 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 sub-Gaussian assumption which is in fact a norm equivalence assumption (i.e., Definition 2.3 is equivalent to $\|f\|_{L_2} \le \|f\|_{\psi_2} \le L\|f\|_{L_2}$ for all $f \in \mathcal{F}$ where $\|\cdot\|_{\psi_2}$ is the ψ_2 Orlicz norm; cf. [37]).

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

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* sub-Gaussian, but with a constant L depending on the dimensions m and T, which yields suboptimal 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 that will be studied in depth in Section 4. For this example, the X_i's are chosen randomly in the canonical basis $(E_{1,1},\dots,E_{m,T})$ of $\mathbb{R}^{m\times T}$. Moreover, in that example, the class F is the class of all linear functionals indexed by bB_∞: $F=\{\langle\cdot,M\rangle:\max_{p,q}|M_{pq}|\le b\}$ and, therefore, the study of this problem falls naturally in the bounded framework studied in this section.

Under the boundedness assumption, the "statistical complexity" cannot be anymore characterized by Gaussian mean width. We therefore introduce another complexity parameter known as Rademacher complexity. This complexity measure has been extensively studied in the learning theory literature (cf., for instance, [6, 22, 23]).

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

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

REMARK 2.4. The Rademacher complexity is often defined in the literature as $\operatorname{Rad}'(H)=\mathbb{E}\sup_{f\in H}|\frac{1}{N}\sum_{i=1}^N\varepsilon_i f(X_i)|$, namely, a factor $\sqrt{N}$ smaller than $\operatorname{Rad}(H)$. We chose to use $\operatorname{Rad}(H)=\sqrt{N}\operatorname{Rad}'(H)$ as this allow a unified presentation with the sub-Gaussian case where the complexity is measured with the Gaussian mean width.

Note that when $(f(X))_{f\in H}$ is a version of the isonormal process over L_2 (cf. Chapter 12 in [18]) restricted to H then the Gaussian mean-width and the Rademacher complexity coincide: $w(H)=\operatorname{Rad}(H)$. 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 complexity have been calculated (cf. [36]). Like in the previous *sub-Gaussian* setting the statistical complexity is given by a function $r(\cdot)$. Note that we use the same notation $r(\cdot)$ in the two scenarii, namely the *bounded* and *sub-Gaussian* case. We do this because this $r(\cdot)$ function plays exactly the same role in both cases. However, its definition is not the same in each scenario, as can be seen below.

DEFINITION 2.7. The *complexity parameter* is the nondecreasing function $r(\cdot)$ defined for every $\rho \geq 0$ by

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

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*

$$1 - \mathbf{C}\exp\bigl(-\mathbf{C}N^{1/2\kappa}\bigl(\rho^*\,\mathrm{Rad}(B)\bigr)^{(2\kappa-1)/\kappa}\bigr) \tag{7}$$

we have

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

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 Sections 3, 4 and Section 7 in Supplement A, we compute $r(\rho)$ either in the sub-Gaussian 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. [19, 33, 35, 39, 42]).

We consider the *vector* framework, where $(X_1, Y_1), \dots, (X_N, Y_N)$ are N i.i.d. pairs with values in $\mathbb{R}^p \times \{-1, 1\}$ distributed like (X, Y). Both bounded and sub-Gaussian frameworks can be analyzed in this example. Since an example in the bounded case is provided in the next section, only the sub-Gaussian 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 functionals indexed by RB_{l_2} for some radius $R \geq 1$ and the logistic loss:

$$F = \bigl\{\langle \cdot, t\rangle : t \in RB_{l_2}\bigr\}, \qquad \ell_f(x, y) = \log\bigl(1 + \exp\bigl(-yf(x)\bigr)\bigr). \tag{8}$$

As usual the oracle is denoted by $f^* = \operatorname{argmin}_{f\in F} \mathbb{E}\ell_f(X, Y)$, we also introduce $t^* \in RB_{\ell_2}$ 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 8.2 in Supplement A that Assumption 2.1 is satisfied when the design X is the standard Gaussian variable in $\mathbb{R}^p$ and the class F defined in (8); note that this fact is not obvious, and is new up to our knowledge. In that case, the Bernstein parameter is $\kappa = 1$ and $A = c_0/R^3$ for some absolute constant $c_0 > 0$ which can be deduced from the proof of Proposition 8.2. We consider the l_1 norm $\|\langle\cdot, t\rangle\| = \|t\|_{l_1}$ 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

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

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 (3).

Let us first deal with the complexity parameter of the problem. If one assumes that the design vector X is *isotropic*, that is, $\mathbb{E}\langle X, t\rangle^2 = \|t\|_{l_2}^2$ 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 $w(B_{l_1}) \le c_1\sqrt{\log p}$ for some (known) absolute constant $c_1 > 0$ and so we define, for all $\rho \ge 0$,

$$r(\rho) = \mathbf{C}\left(\rho\sqrt{\frac{\log p}{N}}\right)^{1/2} \tag{9}$$

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 (3). First, note that when the design is isotropic the sparsity parameter is the function

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

where $\Gamma_{t^*}(\rho) = \bigcup_{f \in t^* + (\rho/20)B_{l_1}} \partial\|\cdot\|(f)$.

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 (3) 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 [31].

LEMMA 3.1 (Lemma 4.2 in [31]). *If there exists some $v \in t^* + (\rho/20)B_{l_1}$ such that $\|v\|_0 \le c_0(\rho/r(\rho))^2$ then $\Delta(\rho) \ge 4\rho/5$ 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 $1 \le q \le 2$, the logistic LASSO estimator $\widehat{t_L}$ with regularization parameter*

$$\lambda = \frac{5c_1 CL}{8}\sqrt{\frac{\log p}{N}}$$

satisfies

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

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

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

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

Estimation results for the logistic LASSO estimator in the generalized linear model have been obtained in [45] 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 $(e_k)_{k=1}^p$ is the canonical basis of $\mathbb{R}^p$—are not bounded when the design is $X \sim \mathcal{N}(0, I_{p\times p})$. 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 [45]. 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 classical rates of convergence for the logistic LASSO estimator that have been obtained in the literature so far in the case of the square loss (see, [31]). This rate is the minimax rate obtained over all s-sparse vectors w.r.t. the ℓ_q distance for every $1 \le q \le 2$ as long as $\log(p/s)$ behaves like $\log p$ when the oracle t^* is the one associated with the square loss (see, [8]). 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. [10, 41]): for every $t = (t_j) \in \mathbb{R}^p$,

$$\|t\|_{\mathrm{SLOPE}} = \sum_{j=1}^{p} \sqrt{\log(ep/j)}\, t_j^\sharp, \tag{10}$$

where $t_1^\sharp \geq t_2^\sharp \geq \cdots \geq 0$ is the nonincreasing 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 classical (minimax, for the square loss) rate $\sqrt{(s/N)\log(ep/s))}$ for any $s \in \{1, \dots, p\}$.

Indeed, it follows from Lemma 5.3 in [31] 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 equation* is satisfied by the radius

$$\rho^* \sim \frac{s}{\sqrt{N}} \log\left(\frac{ep}{s}\right) \tag{11}$$

as long as there is a s-sparse vector in $t^* + (\rho^*/20)B_{\mathrm{SLOPE}}$. The *norm dependent* radius is as usual of order $\|t^*\|_{\mathrm{SLOPE}}$. 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 $s \in \{1, \dots, p\}$. Assume that there exists a s-sparse vector in $t^* + (\rho^*/20)B_{\mathrm{SLOPE}}$ for ρ^* as in* (11). *Then, with probability larger than $1 - \mathbf{C}\exp(-\mathbf{C}s\log(p/s))$, the logistic SLOPE estimator*

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

satisfies

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

and

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

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

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

TABLE 1
Key quantities involved in the study of the Logistic LASSO and SLOPE

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

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 rate $s\log(ep/s)/N$ for the ℓ_2-estimation risk (to the square) with absolutely no assumption on the output Y. In the case where a statistical model $Y=\operatorname{sign}(\langle X,t^*\rangle+\xi)$ holds, where ξ is independent of X then Theorem 3.2 shows that the RERM with logistic loss and SLOPE regularization achieves the 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. [29]).

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 subdifferential at sparse vectors and a small Gaussian mean-width $w(B)$ is a good way to construct

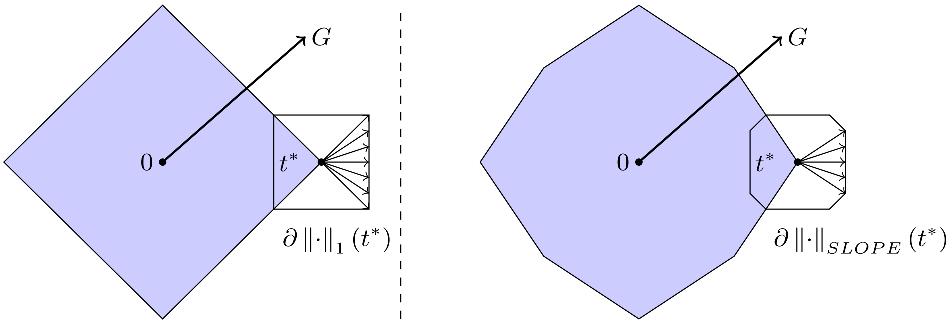

FIG. 1. *Gaussian complexity and size of the subdifferential for the ℓ_1 and SLOPE norms: A "large" subdifferential at sparse vectors and a small Gaussian mean width of the unit ball of the regularization norm is better for sparse recovery. In this figure, G represents a "typical" Gaussian vector used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

"sparsity inducing" regularization norms as it is, for instance, the case of "atomic norms" (cf. [16]).

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 nondifferentiable loss and provide a simulation study that is postponed to Section 6. 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. The class is $F = \{\langle \cdot, M\rangle : M \in bB_\infty\}$, where $bB_\infty = \{M = (M_{pq}) \in \mathbb{R}^{m\times T} : \max_{p,q} |M_{pq}| \leq b\}$ and $b > 0$. In matrix completion, we write the observed location as a mask matrix X: it is an element of the canonical basis $(E_{1,1}, \dots, E_{m,T})$ 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 4.1 (Matrix completion design). The variable X takes value in the canonical basis $(E_{1,1}, \dots, E_{m,T})$ of $\mathbb{R}^{m\times T}$. There are positive constants $\underline{c}$, $\bar{c}$ such that for any $(p,q) \in \{1,\dots,m\} \times \{1,\dots,T\}$, $\underline{c}/(mT) \leq \mathbb{P}(X = E_{p,q}) \leq \bar{c}/(mT)$.

As the design X takes its values in the canonical basis of $\mathbb{R}^{m\times T}$, the boundedness assumption is satisfied. The penalty is taken as the nuclear norm. Thus, the RERM is given by

$$\widehat{M} \in \operatorname*{argmin}_{M\in bB_\infty}\left(\frac{1}{N}\sum_{i=1}^{N} \ell(\langle X_i, M\rangle, Y_i) + \lambda\|M\|_{S_1}\right). \tag{12}$$

Statistical properties of (12) 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 $\langle \overline{M}, \cdot\rangle = \overline{f} = \arg\min_{f \text{ measurable}} \mathbb{E}\ell(f(X), Y)$ (because X takes its values in a finite set). 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, $\bar{M}_{p,q} = \mathbb{P}[Y = 1|X = E_{p,q}] \in [0,1]$ for all p, q and so $\overline{M} = M^*$ 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 $B_{S_1} = \{A \in \mathbb{R}^{m\times T} : \|A\|_{S_1} \le 1\}$:

$$\begin{aligned}\operatorname{Rad}(B_{S_1}) &= \mathbb{E} \sup_{\|A\|_{S_1}\le 1} \left| \frac{1}{\sqrt{N}} \sum_{i=1}^N \varepsilon_i \langle X_i, A\rangle \right| \\ &= \mathbb{E} \left\| \frac{1}{\sqrt{N}} \sum_{i=1}^N \varepsilon_i X_i \right\|_{S_\infty} \\ &\le c_0(\underline{c}, \bar{c}) \sqrt{\frac{\log(m+T)}{\min(m,T)}},\end{aligned} \tag{13}$$

where $\|\cdot\|_{S_\infty}$ is the operator norm (i.e., the largest singular value), the last inequality follows from Lemma 1 in [24] 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 $\rho \ge 0$,

$$r(\rho) = \left[\frac{CA\rho \operatorname{Rad}(B_{S_1})}{\sqrt{N}}\right]^{\frac{1}{2\kappa}} = \mathbf{C}\left[\rho\sqrt{\frac{\log(m+T)}{N\min(m,T)}}\right]^{\frac{1}{2\kappa}}, \tag{14}$$

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 such that, for any $\rho > 0$,

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

where $\Gamma_{M^*}(\rho)$ is the union of all the subdifferential of $\|\cdot\|_{S_1}$ of points 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 "nonnormalized isotropic" property of X: $\underline{c}\|M\|_{S_2}^2/(mT) \le \mathbb{E}\langle X, M\rangle^2 \le \bar{c}\|M\|_{S_2}^2/(mT)$ for all $M \in \mathbb{R}^{m\times T}$. Now, we use a result from [31] to find a solution to the sparsity equation.

LEMMA 4.1 (Lemma 4.4 in [31]). *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* $\operatorname{rank}(V) \leq (c_1\rho/(\sqrt{mT}r(\rho)))^2$ *then* $\Delta(\rho) \geq 4\rho/5$.

It follows from Lemma 4.1 that the sparsity equation (3) is satisfied by ρ^* when it exists $V \in M^* + (\rho^*/20)B_{S_1}$ such that $\operatorname{rank}(V) = c_1(\rho^*/(\sqrt{mT}r(\rho^*)))^2$. Note obviously that V can be M^* itself, in this case, ρ^* can be taken such that $\operatorname{rank}(M^*) = c_1(\rho^*/(\sqrt{mT}r(\rho^*)))^2$. 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

$$\rho_s^* = \mathbf{C}(smT)^{\frac{\kappa}{2\kappa-1}}\left(\frac{\log(m+T)}{N\min(m,T)}\right)^{\frac{1}{2(2\kappa-1)}} \tag{15}$$

then ρ_s^* satisfies the sparsity equation.

Following the remark at the end of Section 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 (12).

THEOREM 4.1. *Assume that Assumption 1.1, 4.1 and 2.1 hold. Consider the estimator in (12) with regularization parameter*

$$\lambda = \frac{c_0(\underline{c}, \bar{c})720}{7}\sqrt{\frac{\log(m+T)}{N\min(m,T)}}, \tag{16}$$

where $c_0(\underline{c}, \bar{c})$ are the constants in Assumption 4.1. 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 $1 - \mathbf{C}\exp(-\mathbf{C}s(m+T)\log(m+T))$, we have

$$\begin{aligned}
&\|\widehat{M} - M^*\|_{S_1} \\
&\quad\leq \mathbf{C}\min\left\{(smT)^{\frac{\kappa}{2\kappa-1}}\left(\frac{\log(m+T)}{N\min(m,T)}\right)^{\frac{1}{2(2\kappa-1)}}, \|M^*\|_{S_1}\right\}, \\
&\frac{\|\widehat{M} - M^*\|_{S_2}}{\sqrt{mT}} \\
&\quad\leq \mathbf{C}\min\left\{\left(\frac{s(m+T)\log(m+T)}{N}\right)^{\frac{1}{2(2\kappa-1)}}, \left(\|M^*\|_{S_1}\sqrt{\frac{\log(m+T)}{N\min(m,T)}}\right)^{\frac{1}{2\kappa}}\right\}, \\
&\mathcal{E}(\widehat{M}) \leq \mathbf{C}\min\left\{\left(\frac{s(m+T)\log(m+T)}{N}\right)^{\frac{\kappa}{2\kappa-1}}, \|M^*\|_{S_1}\sqrt{\frac{\log(m+T)}{N\min(m,T)}}\right\}.
\end{aligned}$$

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

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

$$\|M^*\|_{S_1}^{\frac{p-1+\kappa(2-p)}{p\kappa}} \biggl(\frac{\log(m+T)}{N\min(m,T)}\biggr)^{\frac{p-1}{2\kappa p}} \biggl(\frac{1}{mT}\biggr)^{\frac{2-p}{p}}\biggr\}.$$

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

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

which is the classic excess risk bound under the margin assumption up to a log factor (cf. [4]). 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, for example, take the quadratic loss when Y is bounded and compare to [38]).

4.2. *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 8.2 in Supplement A we know that $\kappa = 1$ for any b [A is also known, $A = 4\exp(2b)$] and, therefore, the 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 4.1 holds. Let $s \in \{1, \dots, \min(m,T)\}$ and assume that there exists a matrix with rank at most s in $M^* + (\rho_s^*/20)B_{S_1}$ where ρ_s^* is defined in (15). With probability at least $1 - \mathbf{C}\exp(-\mathbf{C}s\max(m,T)\log(m+T))$, the estimator*

$$\widehat{M} \in \operatorname*{argmin}_{M \in bB_\infty}\biggl(\frac{1}{N}\sum_{i=1}^{N}\log(1+\exp(-Y_i\langle X_i, M\rangle)) + \lambda\|M\|_{S_1}\biggr) \tag{17}$$

with λ as in equation (16) *satisfies*

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

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

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

Using an interpolation inequality, it is easy to derive estimation bound in S_p for all $1 \le p \le 2$ as in Theorem 4.1 so we do not reproduce it here. Also, note that our bound on $\|\widehat{M}-M^*\|_{S_2}$ is of the same order as the one in [26]. 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 $N \ge s(m + T)\log(2)/(8b^2)$. Then*

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

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 $R_{0/1}(M) = \mathbb{P}[(Y \neq \text{sign}(\langle M, X\rangle)]$ 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 [46] is that

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

But this rate on the excess 0/1-risk may be much better under the margin assumption [34, 43] [cf. equation (36) in Supplement A]. This motivates the use of the hinge loss instead of the logistic loss, for which the results in [46] 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 8.3 in Supplement A 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 4.1 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* (15). *With probability at least* $1 - \mathbf{C}\exp(-\mathbf{C}s \max(m, T)\log(m + T))$, *the estimator*

$$\widehat{M} \in \operatorname*{argmin}_{M \in B_\infty} \left(\frac{1}{N} \sum_{i=1}^{N} (1 - Y_i \langle X_i, M \rangle)_+ + \lambda \|M\|_{S_1} \right) \tag{18}$$

with λ *as in equation* (16) *satisfies*

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

$$\frac{1}{\sqrt{mT}} \|\widehat{M} - \overline{M}\|_{S_2}$$

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

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

In this case, [46] 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 $\operatorname{rank}(\overline{M})(m + T)\log(m + T)/N$.

First, note that [40], obtained a rate in $\sqrt{\operatorname{rank}(\overline{M})(m + T)/N}$ up to log terms without this assumption $\inf_{p,q} |P(Y = 1|X = E_{p,q}) - 1/2| \geq \tau$ for some $\tau > 0$, and proved that this rate is optimal in this case; [11] also obtained a rate in $1/\sqrt{N}$. The rate $\operatorname{rank}(\overline{M})(m + T)\log(m + T)/N$ for the classification excess error was only reached in [17] up to our knowledge (using the PAC-Bayesian technique from [1, 2, 13, 14, 32]), 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 a much general case $\inf_{p,q} |P(Y = 1|X = E_{p,q}) - 1/2| \geq \tau$ even when $\tau > 0$ is very small. Finally, we prove that our rate is actually the minimax rate in this case.

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, for some universal constant $c > 0$,*

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

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.

4.3. *Quantile loss and median matrix completion.* The matrix completion problem with continuous entries has almost always been tackled with a penalized least squares estimator [12, 21, 24, 31, 32], but the use of other loss functions may be very interesting in this case also. 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 aforementioned references provide 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 (e.g., 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 8.4 in Supplement A. Following [44], 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 7 in Supplement A 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 means 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 4.1 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 g with respect to the Lebesgue measure such that $g(u) > 1/c$ for some constant $c > 0$ for any u such that $|u - \overline{M}_{p,q}| \leq 2b$. 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 (15). Then, with probability at least $1 - \mathbf{C}\exp(-\mathbf{C}s\max(m, T)\log(m + T))$, the estimator

$$\widehat{M} \in \operatorname*{argmin}_{M \in bB_\infty}\left(\frac{1}{N}\sum_{i=1}^{N}\rho_\tau(Y_i - \langle X_i, M\rangle) + \lambda\|M\|_{S_1}\right) \tag{19}$$

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

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

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

$$\left.\|\overline{M}\|_{S_1}^{\frac{1}{2}}\left(\frac{\log(m + T)}{N\min(m, T)}\right)^{\frac{1}{4}}\right\},$$

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

We obtain the same rate as for the penalized least squares estimator that is $\sqrt{s(m + T)\log(m + T)/N}$ (cf. [24, 38]).

5. Discussion.

This paper covers several aspects of the regularized empirical risk estimator (RERM) with Lipschitz loss. This Lipschitz property is commonly shared by many loss functions used in practice for robust estimation 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 sub-Gaussian framework and the bounded framework. The first one is illustrated by the classification problem with logistic loss. In particular, the $s\log(p/s)/N$ ℓ_2-estimation rate can be achieved when using the SLOPE regularization norm for estimating an approximately sparse oracle. The second framework is used to derive new results on matrix completion. Finally, Kernel methods are analyzed in Supplement A.

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 subdifferential of the regularization norm and in particular its singularity points which are related to the sparsity equation and to the general sparsity structure we aim at recovering.

SUPPLEMENTARY MATERIAL

Supplementary material to "Estimation bounds and sharp oracle inequalities of regularized procedures with Lipschitz loss functions" (DOI: 10.1214/18-AOS1742SUPP; .pdf). In the supplementary material, we provide a simulation study on the different procedures that have been introduced for matrix completion. The example of kernel estimation is also developed. All the proofs have been gathered in this supplementary material. We finally propose a brief study of the ERM without penalization.

REFERENCES

[1] ALQUIER, P. (2013). Bayesian methods for low-rank matrix estimation: Short survey and theoretical study. In *Algorithmic Learning Theory*. *Lecture Notes in Computer Science* **8139** 309–323. Springer, Heidelberg. MR3133074

[2] ALQUIER, P., RIDGWAY, J. and CHOPIN, N. (2016). On the properties of variational approximations of Gibbs posteriors. *J. Mach. Learn. Res.* **17** 239. MR3595173

[3] ALQUIER, P., COTTET, V. and LECUÉ, G. (2019). Supplement to "Estimation bounds and sharp oracle inequalities of regularized procedures with Lipschitz loss functions." DOI:10.1214/18-AOS1742SUPP.

[4] AUDIBERT, J.-Y. and TSYBAKOV, A. B. (2007). Fast learning rates for plug-in classifiers. *Ann. Statist.* **35** 608–633. MR2336861

[5] BARTHE, F., GUÉDON, O., MENDELSON, S. and NAOR, A. (2005). A probabilistic approach to the geometry of the l_p^n-ball. *Ann. Probab.* **33** 480–513. MR2123199

[6] BARTLETT, P. L., BOUSQUET, O. and MENDELSON, S. (2005). Local Rademacher complexities. *Ann. Statist.* **33** 1497–1537. MR2166554

[7] BARTLETT, P. L. and MENDELSON, S. (2006). Empirical minimization. *Probab. Theory Related Fields* **135** 311–334. MR2240689

[8] BELLEC, P., LECUÉ, G. and TSYBAKOV, A. (2018). Slope meets Lasso: Improved oracle bounds and optimality *Ann. Statist.* **46** 3603–3642. MR3852663

[9] BELLONI, A. and CHERNOZHUKOV, V. (2011). ℓ_1-penalized quantile regression in high-dimensional sparse models. *Ann. Statist.* **39** 82–130. MR2797841

[10] BOGDAN, M., VAN DEN BERG, E., SABATTI, C., SU, W. and CANDÈS, E. J. (2015). SLOPE—adaptive variable selection via convex optimization. *Ann. Appl. Stat.* **9** 1103–1140. MR3418717

[11] CAI, T. and ZHOU, W.-X. (2013). A max-norm constrained minimization approach to 1-bit matrix completion. *J. Mach. Learn. Res.* **14** 3619–3647. MR3159403

[12] CANDÈS, E. J. and PLAN, Y. (2010). Matrix completion with noise. *Proc. IEEE* **98** 925–936.

[13] CATONI, O. (2004). *Statistical Learning Theory and Stochastic Optimization*: *Ecole d'Eté de Probabilités de Saint-Flour, XXXI*-2001. **31**. Springer, Berlin. MR2163920

[14] CATONI, O. (2007). *Pac-Bayesian Supervised Classification*: *The Thermodynamics of Statistical Learning*. *Institute of Mathematical Statistics Lecture Notes—Monograph Series* **56**. IMS, Beachwood, OH. MR2483528

[15] CHAFAÏ, D., GUÉDON, O., LECUÉ, G. and PAJOR, A. (2012). *Interactions Between Compressed Sensing Random Matrices and High Dimensional Geometry*. *Panoramas et Synthèses* [*Panoramas and Syntheses*] **37**. Société Mathématique de France, Paris. MR3113826

[16] CHANDRASEKARAN, V., RECHT, B., PARRILO, P. A. and WILLSKY, A. S. (2012). The convex geometry of linear inverse problems. *Found. Comput. Math.* **12** 805–849. MR2989474

[17] COTTET, V. and ALQUIER, P. (2016). 1-bit Matrix Completion: PAC-Bayesian Analysis of a Variational Approximation. Machine Learning. To appear. Preprint arXiv:1604.04191.

[18] DUDLEY, R. M. (2002). *Real Analysis and Probability*. *Cambridge Studies in Advanced Mathematics* **74**. Cambridge Univ. Press, Cambridge. Revised reprint of the 1989 original. MR1932358

[19] GARCIA-MAGARIÑOS, M., ANTONIADIS, A., CAO, R. and GONZÁLEZ-MANTEIGA, W. (2010). Lasso logistic regression, GSoft and the cyclic coordinate descent algorithm: Application to gene expression data. *Stat*. *Appl*. *Genet*. *Mol*. *Biol*. **9** 30. MR2721710

[20] GORDON, Y., LITVAK, A. E., MENDELSON, S. and PAJOR, A. (2007). Gaussian averages of interpolated bodies and applications to approximate reconstruction. *J*. *Approx*. *Theory* **149** 59–73. MR2371614

[21] KLOPP, O. (2014). Noisy low-rank matrix completion with general sampling distribution. *Bernoulli* **20** 282–303. MR3160583

[22] KOLTCHINSKII, V. (2006). Local Rademacher complexities and oracle inequalities in risk minimization. *Ann*. *Statist*. **34** 2593–2656. MR2329442

[23] KOLTCHINSKII, V. (2011). *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems*. *Lecture Notes in Math*. **2033**. Springer, Heidelberg. MR2829871

[24] KOLTCHINSKII, V., LOUNICI, K. and TSYBAKOV, A. B. (2011). Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *Ann*. *Statist*. **39** 2302–2329. MR2906869

[25] KOLTCHINSKII, V. and PANCHENKO, D. (2002). Empirical margin distributions and bounding the generalization error of combined classifiers. *Ann*. *Statist*. **30** 1–50. MR1892654

[26] LAFOND, J., KLOPP, O., MOULINES, E. and SALMON, J. (2014). Probabilistic low-rank matrix completion on finite alphabets. In *Advances in Neural Information Processing Systems* 1727–1735.

[27] LECUÉ, G. (2011). *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.

[28] LECUÉ, G. and MENDELSON, S. (2012). General nonexact oracle inequalities for classes with a subexponential envelope. *Ann*. *Statist*. **40** 832–860. MR2933668

[29] LECUÉ, G. and MENDELSON, S. (2013). Learning subgaussian classes: Upper and minimax bounds. Technical Report CNRS, Ecole polytechnique and Technion—to appear in Topics in Learning Theory—Societe Mathématique de France (S. Boucheron and N. Vayatis eds.).

[30] LECUÉ, G. and MENDELSON, S. (2017). Regularization and the small-ball method II: Complexity dependent error rates. *J*. *Mach*. *Learn*. *Res*. **18** 146. MR3763780

[31] LECUÉ, G. and MENDELSON, S. (2018). Regularization and the small-ball method I: Sparse recovery. *Ann*. *Statist*. **46** 611–641. MR3782379

[32] MAI, T. T. and ALQUIER, P. (2015). A Bayesian approach for noisy matrix completion: Optimal rate under general sampling distribution. *Electron*. *J*. *Stat*. **9** 823–841. MR3331862

[33] MAK, C. (1999). *Polychotomous Logistic Regression Via the Lasso*. ProQuest LLC, Ann Arbor, MI. Thesis (Ph.D.)—Univ. Toronto. MR2699823

[34] MAMMEN, E. and TSYBAKOV, A. B. (1999). Smooth discrimination analysis. *Ann*. *Statist*. **27** 1808–1829. MR1765618

[35] MEIER, L., VAN DE GEER, S. and BÜHLMANN, P. (2008). The group Lasso for logistic regression. *J*. *R*. *Stat*. *Soc*. *Ser*. *B*. *Stat*. *Methodol*. **70** 53–71. MR2412631

[36] MENDELSON, S. (2004). On the performance of kernel classes. *J*. *Mach*. *Learn*. *Res*. **4** 759–771. MR2075996

[37] RAO, M. M. and REN, Z. D. (1991). *Theory of Orlicz Spaces*. *Monographs and Textbooks in Pure and Applied Mathematics* **146**. Dekker, New York. MR1113700

[38] ROHDE, A. and TSYBAKOV, A. B. (2011). Estimation of high-dimensional low-rank matrices. *Ann. Statist.* **39** 887–930. MR2816342

[39] SABBE, N., THAS, O. and OTTOY, J.-P. (2013). EMLasso: Logistic lasso with missing data. *Stat. Med.* **32** 3143–3157. MR3073790

[40] SREBRO, N., RENNIE, J. and JAAKKOLA, T. S. (2004). Maximum-margin matrix factorization. In *Advances in Neural Information Processing Systems* 1329–1336.

[41] SU, W. and CANDÈS, E. (2016). SLOPE is adaptive to unknown sparsity and asymptotically minimax. *Ann. Statist.* **44** 1038–1068. MR3485953

[42] TIAN, G.-L., TANG, M.-L., FANG, H.-B. and TAN, M. (2008). Efficient methods for estimating constrained parameters with applications to regularized (lasso) logistic regression. *Comput. Statist. Data Anal.* **52** 3528–3542. MR2427362

[43] TSYBAKOV, A. B. (2004). Optimal aggregation of classifiers in statistical learning. *Ann. Statist.* **32** 135–166. MR2051002

[44] VAN DE GEER, S. (2016). *Estimation and Testing Under Sparsity*. *Lecture Notes in Math.* **2159**. Springer, Cham. MR3526202

[45] VAN DE GEER, S. A. (2008). High-dimensional generalized linear models and the lasso. *Ann. Statist.* **36** 614–645. MR2396809

[46] ZHANG, T. (2004). Statistical behavior and consistency of classification methods based on convex risk minimization. *Ann. Statist.* **32** 56–85. MR2051001

ENSAE
3, AVENUE PIERRE LAROUSSE
92245 MALAKOFF
FRANCE
E-MAIL: pierre.alquier@ensae.fr
vincent.cottet@ensae.fr
guillaume.lecue@ensae.fr