General nonexact oracle inequalities for classes with a subexponential envelope

Guillaume Lecué, Shahar Mendelson

Submitted to the Annals of Statistics

GENERAL NON-EXACT ORACLE INEQUALITIES FOR CLASSES WITH A SUBEXPONENTIAL ENVELOPE

By Guillaume Lecué*

CNRS, LAMA, Université Paris-Est Marne-la-vallée, 77454 France

AND

By Shahar Mendelson†

Department of Mathematics, Technion, I.I.T, Haifa 32000, Israel

We show that empirical risk minimization procedures and regularized empirical risk minimization procedures satisfy non-exact oracle inequalities in an unbounded framework, under the assumption that the class has a subexponential envelope function. The main novelty, in addition to the boundedness assumption free setup, is that those inequalities can yield fast rates even in situations in which exact oracle inequalities only hold with slower rates.

We apply these results to show that procedures based on ℓ_1 and nuclear norms regularization functions satisfy oracle inequalities with a residual term that decreases like $1/n$ for every L_q-loss functions ($q \geq 2$), while only assuming that the tail behaviour of the input and output variables are well behaved. In particular, no RIP type of assumption or "incoherence condition" are needed to obtain fast residual terms in those setups. We also apply these results to the problems of Convex aggregation and Model Selection.

1. Introduction and main results. Let $\mathcal{Z}$ be a space endowed with a probability measure P and let Z and $Z_1, \ldots, Z_n$ be $n+1$ independent random variables with values in $\mathcal{Z}$, distributed according to P; from the statistical point of view, $\mathcal{D} = (Z_1, \ldots, Z_n)$ is the set of given data. Let ℓ be a loss function which associates a real number $\ell(f, z)$ to any real-valued measurable function f defined on $\mathcal{Z}$ and any point $z \in \mathcal{Z}$. Denote by ℓ_f the loss function $\ell(f, \cdot)$ associated with f and set $R(f) = \mathbb{E}\ell_f(Z)$ to be the associated risk. The risk of any statistic $\widehat{f}_n(\cdot) = \widehat{f}_n(\cdot, \mathcal{D}) : \mathcal{Z} \longrightarrow \mathbb{R}$ is defined

*This work is supported by the French Agence Nationale de la Recherce (ANR) ANR Grant "Prognostic" ANR-09-JCJC-0101-01.

†partially supported by the Mathematical Sciences Institute – The Australian National University, the European Research Council (under ERC grant agreement n^o [203134]) and the Australian Research Council (under grant DP0986563)

AMS 2000 subject classifications: Primary 62G05; secondary 62H30, 68T10.

Keywords and phrases: statistical learning, fast rates of convergence, oracle inequalities, regularization, classification, aggregation, Model Selection, high-dimensional data.

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

Let F be a class (usually called the *model*) of real-valued measurable functions defined on $\mathcal{Z}$. In learning theory, one wants to assume as little as possible on the class F, or on the measure P. The aim is to use the data to construct learning algorithms whose risk is as close as possible to $\inf_{f\in F} R(f)$ (and when this infimum is attained by a function f_F^* in F, this element is called an *oracle*). Hence, one would like to construct procedures $\widehat{f}_n$ such that, for some $\epsilon \geq 0$, with high probability,

$$R(\widehat{f}_n) \leq (1+\epsilon)\inf_{f\in F} R(f) + r_n(F). \tag{1.1}$$

The role of the *residual term* (or *rate*) $r_n(F)$ is to capture the "complexity" of the problem, and the hope is to make it as small as possible.

When $r_n(F)$ tends to zero as n tends to infinity, Inequality (1.1) is called an *oracle inequality*. When $\epsilon = 0$, we say that $\widehat{f}_n$ satisfies an *exact oracle inequality* (the term *sharp* oracle inequality has been also used) and when $\epsilon > 0$ it satisfies a *non-exact oracle inequality*. Note that the terminology "risk bounds" has been also used for (1.1) in the literature.

A natural algorithm in this setup is the *empirical risk minimization procedure* (ERM) (terminology due to [44]), in which the *empirical risk* functional

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

is minimized and produces $\widehat{f}_n^{ERM} \in \operatorname{Arg\,min}_{f\in F} R_n(f)$. Note that when $R_n(\cdot)$ does not achieve its infimum over F or if the minimizer is not unique, we define $\widehat{f}_n^{ERM}$ to be an element in F for which $R(\widehat{f}_n^{ERM}) \leq \inf_{f\in F} R(f) + 1/n$. This algorithm has been extensively studied and we will compare our first result to the one of [12, 4, 25].

One motivation in obtaining non-exact oracle inequalities (Equation (1.1) for $\epsilon \neq 0$) is the observation that in many situations, one can obtain such an inequality for the ERM procedure with a residual term $r_n(F)$ of the order of $1/n$, while the best residual term achievable by ERM in an exact oracle inequality (Equation (1.1) for $\epsilon = 0$) will only be of the order of $1/\sqrt{n}$ for the same problem. For example, consider the simple case of a finite model F of cardinality M and the bounded regression model with the quadratic loss function (that is $Z = (X,Y) \in \mathcal{X}\times\mathbb{R}$ with $|Y|, \max_{f\in F}|f(X)| \leq C$ for some absolute constant C and $\ell(f,(X,Y)) = (Y-f(X))^2$). It can be verified that for every $x > 0$, with probability greater than $1 - 8\exp(-x)$, $\widehat{f}_n^{ERM}$ satisfies a non-exact oracle inequality with a residual term proportional to $(x + \log M)/(\epsilon n)$. On the other hand, it is known [22, 29, 19] that in the

same setup, there are finite models for which, with probability greater than a positive constant, $\widehat{f}_n^{ERM}$ cannot satisfy an exact oracle inequality with a residual term better than $c_0\sqrt{(\log M)/n}$. Thus, it is possible to establish two optimal oracle inequalities (i.e. oracle inequalities with a non-improvable residual term $r_n(F)$ up to some multiplying constant) for the same procedure with two very different residual terms: one being the square of the other one. We will see below that the same phenomenon occurs in the classification framework for VC classes. Thus our main goal here is to present a general framework for non-exact oracle inequalities for ERM and RERM (regularized ERM), and show that they lead to fast rates in cases when the best known exact oracle inequalities have slow rates.

Although the improved rates are significant, it is clear that exact inequalities are more "valuable" from the statistical point of view. For example, consider the regression model with the quadratic loss. It follows from an exact oracle inequality on the prediction risk (Equation (1.1) for $\epsilon = 0$), an other exact oracle inequality but for the estimation risk:

$$\left\| \widehat{f}_n^{ERM} - f^* \right\|_{L_2}^2 \leq \inf_{f \in F} \|f - f^*\|_{L_2}^2 + r_n(F),$$

where f^* is the regression function of Y given X and $\|\cdot\|_{L_2}$ is the L_2-norm with respect to the marginal distribution of X.

In other words, exact oracle inequalities for the prediction risk $R(\cdot)$ provide both prediction and estimation results (prediction of the output Y and estimation of the regression function f^*) whereas non-exact oracle inequalities provide only prediction results.

Of course, non-exact inequalities are very useful when it suffices to compare the risk $R(\widehat{f}_n)$ with $(1+\epsilon)\inf_{f\in F} R(f)$; and the aim of this note is to show that the residual term can be dramatically improved in such cases.

1.1. *Empirical risk minimization.* The first result of this note is a non-exact oracle inequality for the ERM procedure. To state this result, we need the following notation. Let G be a class of real-valued functions defined on $\mathcal{Z}$. An important part of our analysis relies on the behaviour of the supremum of the empirical process indexed by G

$$\|P - P_n\|_G = \sup_{g \in G} |(P - P_n)(g)| \tag{1.2}$$

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

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

We will control the supremum (1.2) using the quantities

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

Note that for a bounded class G, one has $b_n(G)\leq \sup_{g\in G}\|g\|_\infty$ and in the sub-exponential case, $b_n(G)\lesssim (\log en)\left\|\sup_{g\in G}|g|\right\|_{\psi_1}$ (this follows from Pisier's inequality, cf. Lemma 2.2.2 in [43]). Throughout this note we will also use the notation $b_n(g)=\|\max_{1\leq i\leq n}|g(Z_i)|\|_{\psi_1}$ and for any pseudo-norm $\|\cdot\|$ on $L_2(P)$, we will denote by $\operatorname{diam}(G,\|\cdot\|)=\sup_{g\in G}\|g\|$ the diameter of G with respect to this norm.

Observe that the desired bound depends on the ψ_1 behaviour of the *envelope function* of the class, $\sup_{g\in G}|g(Z)|$, and as noted above, this extends the "classical" framework of a uniformly bounded class in L_∞. Although this extension seems minor at first, the examples we will present show that the assumption is not very restrictive and allows one to deal with LASSO-type situations, in which the indexing class is very small – something which is impossible under the L_∞ assumption. On the other hand, it should be emphasized that this is not a step towards an unbounded learning theory. For such results, the analogous assumption should be that the class has a bounded diameter in ψ_1, which is, of course, a much weaker assumption than a ψ_1 envelope function and requires different methods (see, e.g. [28, 35]).

To obtain the required bound, we will study empirical processes indexed by sets associated with G, namely, the star-shaped hull of G around zero and the localized subsets for different levels $\lambda\geq 0$, defined by

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

Given a model F and a loss function ℓ, consider the loss class and the excess loss class $\ell_F=\{\ell_f: f\in F\}$ and the excess loss class $\mathcal{L}_F=\{\ell_f-\ell_{f_F^*}: f\in F\}$. We will assume that an oracle f_F^* exists in F, and from here on set $\mathcal{L}_f=\ell_f-\ell_{f_F^*}$.

Theorem A. *There exists an absolute constant $c_0>0$ for which the following holds. Let F be a class of functions and assume that there exists $B_n\geq 0$ such that for every $f\in F$, $P\ell_f^2\leq B_nP\ell_f+B_n^2/n$. Let $0<\epsilon<1/2$, set $\lambda_\epsilon^*>0$ for which*

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

and put ρ_n an increasing function satisfying that for every $x>0$,

$$\rho_n(x)\geq\max\Big(\lambda_\epsilon^*, c_0\frac{(b_n(\ell_F)+B_n/\epsilon)x}{n\epsilon}\Big).$$

Then, for every $x > 0$, with probability greater than $1 - 8\exp(-x)$,

$$R(\widehat{f}_n^{ERM}) \leq (1+3\epsilon) \inf_{f\in F} R(f) + \rho_n(x).$$

REMARK 1.1. *Although the formulation of Theorem A requires that for every $\ell \in \ell_F$, $P\ell^2 \leq B_n P\ell + B_n^2/n$, we will show that if ℓ is nonnegative, this condition is trivially satisfied for $B_n \sim \operatorname{diam}(\ell_F, \psi_1)\log(n)$.*

Unfortunately, this type of condition is far from being trivially satisfied for the excess loss class $\mathcal{L}_F = \{\ell_f - \ell_{f_F^*} : f \in F\}$, which is one of the major differences between exact and non-exact oracle inequalities. Indeed, the Bernstein condition, that for every $f \in F$, $\mathbb{E}\mathcal{L}_f^2 \leq B\mathbb{E}\mathcal{L}_f$ (see [4] or Section 6 below), used in [25, 12, 4] to obtain exact oracle inequalities with fast rates (rates of the order of $1/n$), depends on the geometry of the problem [31, 30] and may not be true in general. Theorem A is similar in nature to Corollary 2.9 of [4] and a detailed comparison between the two results can be found in Section 6.

Theorem A is similar in nature to Theorem 2 in [25]:

THEOREM 1.2. *Let $\phi : \mathbb{R} \to \mathbb{R}$ be a non decreasing, continuous function, for which $\phi(1) \geq 1$ and $x \to \phi(x)/x$ is non increasing. Set F to be a class of functions where there is some $0 \leq \beta \leq 1$ such that $\mathbb{E}\mathcal{L}_f^2 \leq B\big(\mathbb{E}\mathcal{L}_f\big)^\beta$, and $\|\ell_f\|_\infty \leq 1$. If $\phi(\lambda) \geq \sqrt{n}\mathbb{E}\sup_{f,g\in F, P(\ell_f-\ell_g)^2\leq\lambda^2}(P-P_n)(\ell_f-\ell_g)$ for any λ satisfying $\phi(\lambda) \leq \sqrt{n}\lambda^2$ and ε_* is the unique solution of the equation $\sqrt{n}\varepsilon_*^2 = \phi(\sqrt{B}\varepsilon_*^\beta)$, then for every $x \geq 1$, with probability greater than $1 - \exp(-x)$,*

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

One of the applications of the above theorem in learning theory is for the loss function $\ell_f(x,y) = \mathbb{1}_{f(x)\neq y}$. It leads to an exact oracle inequality for the ERM procedure, preformed in a class F of VC dimension $V \leq n$ (see [25] for more details), and with a residual term of the order of $\big(V\log(enB^{1/\beta}/V)/n\big)^{1/(2-\beta)}$.

In comparison, in the same situation, for every $f \in F$, $\mathbb{E}\ell_f^2 \leq \mathbb{E}\ell_f$. Therefore, it follows from Theorem A, the argument used to obtain Equation (29) in [25] (or Example 3 in [12]) and the peeling argument which will be presented in (2.5) below, that for every $x \geq 1$, with probability greater than $1 - 8\exp(-x)$,

$$R(\widehat{f}_n^{ERM}) \leq (1+3\epsilon) \inf_{f\in F} R(f) + c_0 \frac{xV\log\big(en/V\big)}{\epsilon^2 n}. \tag{1.3}$$

The residual term ϵ_*^2 obtained in [25] is optimal but since it heavily depends on the parameter β, it ranges between $\sqrt{V/n}$ and V/n (up to a logarithmic factor). In particular, it can be as bad as the *square root* of the residual term of the non-exact oracle inequality (1.3) in the same situation. The main difference between the two results is that the condition $\mathbb{E}\ell_f^2 \leq \mathbb{E}\ell_f$ for every $f \in F$ is always satisfied whereas the condition that for every $f \in F$ $\mathbb{E}\mathcal{L}_f^2 \leq B(\mathbb{E}\mathcal{L}_f)^\beta$ depends on the relative position of Y and F, and thus on geometry of the system (F, Y).

It is interesting to note that the residual term in (1.3) always yields fast rate even for hard classification problem such that $\mathbb{P}[Y = 1|X] = 1/2$. This means that while the prediction problem in classification is completely blind to the geometry of the model, the estimation problem is influenced in a very strong way by the geometry of (F, Y). Thus, estimating the regression function (or the Bayes rule) is in general much harder than predicting the output Y.

Another related result is the one in [12] where (among other results) an exact oracle inequality is proved for the ERM with a residual term $\delta_n(x)$. The residual term is controlled using the empirical oscillation $\phi_n(\delta) = \mathbb{E}\sup_{f,g\in F(\delta)} |(P - P_n)(\ell_f - \ell_g)|$ indexed by $F(\delta) = \{f \in F : P\mathcal{L}_f \leq \delta\}$, and by the L_2 diameter $D(\delta) = \sup_{f,g\in F(\delta)} \sqrt{P(\ell_f - \ell_g)^2}$:

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

Note that all the quantities λ_ϵ^*, ε_*^2 from [25], $\delta_n(x)$ from [12], μ^* from [4] or Theorem 6.1 below, define the residual terms of the oracle inequalities as a fixed point of some equation. Those appear naturally either from *iterative localization* of the excess risk, converging to $\delta_n(x)$ [12, 16], or from an "isomorphic" argument [4] identifying the "level" μ^* at which the actual and the empirical structures are equivalent. We refer the reader to those articles for more details.

Results in [25, 12, 4] were obtained under the boundedness assumption $\sup_{f\in F} \|\ell_f\|_\infty \leq 1$ because the necessary tools from empirical processes theory, like contraction inequalities [21], only hold under such an assumption. In particular, these results do not apply even to the Gaussian regression model. The approach developed in this work provides a slight improvement, since risk bounds hold if the envelope function $\sup_{f\in F} \ell_f$ is sub-exponential (which is the case for the Gaussian regression model with respect to the square loss).

One should also mention the subtle but significant gap between the *margin assumption* and the Bernstein condition which we use. Both state that for

every $f \in F$,

$$\mathbb{E}(\ell_f - \ell_{f^*})^2 \leq B_0(\mathbb{E}(\ell_f - \ell_{f^*}))^{1/\kappa}$$

for some constant $\kappa \geq 1$. However, in the margin condition f^* has the minimal risk *over all measurable functions* (for instance, f^* is the regression function in the regression model with respect to the quadratic loss), while in a Bernstein condition f_F^* is assumed to minimize the risk *over* F.

The two conditions are equivalent only when $f^* \in F$ (and thus $f^* = f_F^*$). But in general, they are very different. As a simple example, in the bounded regression model (i.e. $|Y|, \sup_{f\in F} |f(X)| \leq C$) with respect to the quadratic loss, the margin assumption holds with $\kappa = 1$ whereas the Bernstein condition is not true in general. For more details on the difference between the margin assumption and the Bernstein condition we refer the reader to the discussion in [18].

1.2. *Regularized empirical risk minimization.* The second type of application we will present deals with non-exact regularized oracle inequalities. Usually a model F is chosen or constructed according to the belief that an oracle f_F^* in F is close, in some sense, to some minimizer f^* of the risk function in some larger class of functions $\mathcal{F}$ (for example, in the regression model, f^* can be the regression function and $\mathcal{F} = L^2(P_X)$). Hence, by choosing a particular model $F \subset \mathcal{F}$, it implicitly means that we believe f^* to be close to F in some sense.

It is not always possible to construct a class F that captures properties f^* is believed to have (e.g., a low-dimensional structure or some smoothness properties). In such situation, one is not given a single model F (usually the set $\mathcal{F}$ is too large to be called a model), but a functional crit : $\mathcal{F} \longrightarrow \mathbb{R}^+$, called a *criterion*, that characterizes each function according to its level of compliance with the desired property – and the smaller the criterion, the "closer" one is to the property). For instance, when $\mathcal{F}$ is an RKHS one can take crit$(\cdot)$ to be the norm in the reproducing kernel Hilbert space, or when $\mathcal{F}$ is the set of all linear functionals in $\mathbb{R}^d$, one may chose crit$(\beta) = \|\beta\|_{\ell_p}$ for some $p \in [0, \infty]$. The extreme case here is $p = 0$ and $\|\beta\|_{\ell_0}$ is the cardinality of the support of β; thus a small criterion means that β belongs to a low-dimensional space.

Instead of considering the ERM over the too large class $\mathcal{F}$, the goal is to construct a procedure having both good empirical performances and a small criterion. One idea, that we will not develop here, is to minimize the empirical risk over the set $F_r = \{f \in \mathcal{F} : \text{crit}(f) \leq r\}$ [41, 5], and try to find a data-dependent way of choosing the radius r. Another popular idea is to regularize the empirical risk: consider a non-decreasing function of the

criterion called a *regularizing function* and denoted by $\mathrm{reg} : \mathcal{F} \longrightarrow \mathbb{R}^+$ and construct

$$\widehat{f}_n^{RERM} \in \mathrm{Arg}\min_{f\in\mathcal{F}} \big(R_n(f) + \mathrm{reg}(f)\big), \tag{1.4}$$

with the obvious extension if the infimum is not attained.

The procedure (1.4) is called *regularized empirical risk minimization procedure* (RERM). RERM procedures have been introduced to avoid the "overfitting" effect of large models [3, 24], and later to select functions with additional properties, like smoothness (for instance, SVM estimators in [38]) or an underlying low-dimensional structure (for example, the LASSO estimator).

In this setup, we are interested in constructing estimators $\widehat{f}_n$ realizing the best possible trade-off between the risk and the regularizing function over $\mathcal{F}$: there exists some $\epsilon \geq 0$ such that with high probability

$$R(\widehat{f}_n) + \mathrm{reg}(\widehat{f}_n) \leq (1+\epsilon) \inf_{f\in\mathcal{F}} \big(R(f) + \mathrm{reg}(f)\big). \tag{1.5}$$

Using the same terminology as in (1.1), Inequality (1.5) is called a *regularized oracle inequality*. When $\epsilon = 0$, (1.5) is called an *exact regularized oracle inequality* and when $\epsilon > 0$, (1.5) is called a *non-exact regularized oracle inequality*.

Following our analysis of the ERM algorithm, the next result is a regularized oracle inequality for the RERM. But before stating this result, one has to say a word on the way the regularizing function $\mathrm{reg}(\cdot)$ and the criterion $\mathrm{crit}(\cdot)$ are related.

The choice of $\mathrm{reg}(\cdot)$ is driven by the complexity of the sequence $(F_r)_{r\geq 0}$ of models

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

For any $r \geq 0$, the complexity of F_r is measured by $\lambda_\epsilon^*(r)$ defined as above for some fixed $0 < \epsilon < 1/2$ by

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

Hence, $\lambda_\epsilon^*(r)$ is a "level" in ℓ_{F_r} above which the empirical and the actual structures are equivalent; namely, with high probability, on the set $\{\ell \in \ell_{F_r} : P\ell \geq \lambda_\epsilon^*(r)\}$,

$$(1/2)P_n\ell \leq P\ell \leq (3/2)P_n\ell.$$

Thus, the function $r \to \lambda_\epsilon^*(r)$ captures the "isomorphic profile" of the collection $(\ell_{F_r})_{r\geq 0}$. Up to minor technical adjustments, the regularizing function, defined formally in (1.8), is $\mathrm{reg}(\cdot) = \lambda_\epsilon^*(\mathrm{crit}(\cdot))$.

We will study two separate situations, both motivated by the applications we have in mind. In the first, $\mathrm{crit}(\cdot)$ will be uniformly bounded and may only grow with the sample size n – that is, there is a constant C_n satisfying that for every $f \in \mathcal{F}$, $\mathrm{crit}(f) \leq C_n$. The second case we deal with is when the "isomorphic profile" $r \to \lambda^*_\epsilon(r)$ tends to infinity with r. For technical reasons, we also introduce an auxiliary function α_n, defined in the following assumption.

ASSUMPTION 1.1. *Assume that for every $f \in \mathcal{F}$, $\ell_f(Z) \geq 0$ a.s. and that there are non-decreasing functions ϕ_n and B_n such that for every $r \geq 0$ and every $f \in F_r$,*

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

Let $0 < \epsilon < 1/2$ and consider a function $\rho_n : \mathbb{R}_+ \times \mathbb{R}_+^* \to \mathbb{R}$ non-decreasing in its first argument and such that, for any $r \geq 0$ and $x > 0$,

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

Assume that either:

- *there exists $C_n > 0$ such that for every $f \in \mathcal{F}, \mathrm{crit}(f) \leq C_n$ and in this case define $\alpha_n(\epsilon, x) = C_n$, for all $0 < \epsilon < 1/2$ and $x > 0$, or*
- *the function $r \to \lambda^*_\epsilon(r)$ tends to infinity with r and there exists $K_1 > 0$ such that $2\rho_n(r,x) \leq \rho_n(K_1(r+1), x)$, for all $r \geq 0$ and $x > 0$ and, in this case, let f_0 be any function in $\cup_{r \geq 0} F_r$ and define α_n such that, for every $x > 0$ and $0 < \epsilon < 1/2$,*

$$\alpha_n(\epsilon, x) \geq \max\Big[K_1(\mathrm{crit}(f_0) + 2), \tag{1.6}$$
$$(\lambda^*_\epsilon)^{-1}\big((1+2\epsilon)(3R(f_0) + 2K'(b_n(\ell_{f_0}) + B_n(\mathrm{crit}(f_0)))((x+1)/n))\big)\Big],$$

where $(\lambda^*_\epsilon)^{-1}$ is the generalized inverse function of λ^*_ϵ (i.e. $(\lambda^*_\epsilon)^{-1}(y) = \sup\big(r > 0 : \lambda^*_\epsilon(r) \leq y\big)$, for all $y > 0$) and K' is some absolute constant.

Theorem B. *There exist absolute positive constants c_0, c_1 K and K' for which the following holds. Under Assumption 1.1, for every $x > 0$ and*

$$\widehat{f}_n^{RERM} \in \operatorname{Arg\,min}_{f \in \mathcal{F}} \Big(R_n(f) + \frac{2}{1+2\epsilon}\rho_n(\mathrm{crit}(f) + 1, x + \log \alpha_n(\epsilon, x))\Big), \tag{1.7}$$

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

$$R(\widehat{f}_n^{RERM})+\rho_n(\mathrm{crit}(\widehat{f}_n^{RERM})+1,x+\log\alpha_n(\epsilon,x))\Big)$$
$$\leq \inf_{f\in\mathcal{F}}\Big[(1+3\epsilon)R(f)+2\rho_n(\mathrm{crit}(f)+1,x+\log\alpha_n(\epsilon,x))$$
$$+c_1\frac{(b_n(\ell_f)+B_n(\mathrm{crit}(f))/\epsilon)(x+1)}{n\epsilon}\Big].$$

Fortunately, α_n usually has little impact on the resulting rates. For instance, in the main application we will present here, $\log\alpha_n(\epsilon,x)\lesssim_\epsilon \log(x+n)$.

Like in Theorem A, the Bernstein type condition $P\ell^2\leq B_n(r)P\ell+B_n^2(r)/n$ holds when ℓ is nonnegative and sub-exponential for $B_n(r)\lesssim \mathrm{diam}(\ell_{F_r},\psi_1)\log(n)$. Therefore, and contrary to the situation in exact oracle inequalities, the "geometry" of the family of classes $(F_r)_{r\geq 0}$ does not play a crucial role in the resulting non-exact regularized oracle inequalities.

Observe that now the choice of the regularizing function in terms of the criterion is now made explicit:

$$\mathrm{reg}(f)=\frac{2}{1+2\epsilon}\rho_n(\mathrm{crit}(f)+1,x+\log\alpha_n(\epsilon,x)). \tag{1.8}$$

1.3. ℓ_1*-regularization.* The formulation of Theorem B seems cumbersome, but it is not very difficult to apply it – and here we will present one application dealing with high-dimensional vectors of short support. Other applications on Matrix Completion, Convex aggregation and Model Selection can be found in [20].

Formally, let $(X,Y),(X_i,Y_i)_{1\leq i\leq n}$ be $n+1$ i.i.d. random variables with values in $\mathbb{R}^d\times\mathbb{R}$ and denote by P_X the marginal distribution of X. The dimension d can be much larger than n but we believe that the output Y can be well predicted by a sparse linear combination of covariables of X; in other words, Y can be reasonably approximated by $\langle X,\beta_0\rangle$ for some $\beta_0\in\mathbb{R}^d$ of short support (even though we will not require any assumption of this type to obtain our results).

These kind of problems are called "high-dimensional" because there are more covariables than observations. Nevertheless, one hopes that under the structural assumption that Y "depends" only on a few number of covariables of X, it would still be possible to construct efficient statistical procedures to predict Y.

In this framework, a natural criterion function is the ℓ_0 function measuring the size of the support of a vector. But since this function is far from being convex, using it in practice is hard (see, e.g., [36]). Therefore, it is natural

to consider a convex relaxation of the ℓ_0 function as a criterion: the ℓ_1 norm [41, 8, 10].

In what follows, we will apply Theorem B to establish non-exact regularized oracle inequalities for ℓ_1-based RERM procedures, and with fast error rates – a residual term that tends to 0 like $1/n$ up to logarithmic terms. The regularizing function resulting from Theorem B for the L_q-loss ($q \geq 2$) will be the q-th power of the ℓ_1-norm. In particular, for the quadratic loss, we regularize by $\|\cdot\|_{\ell_1}^2$, the *square of the ℓ_1-norm*:

$$\widehat{\beta}_n \in \operatorname{Arg}\min_{\beta\in\mathbb{R}^d} \Big(\frac{1}{n}\sum_{i=1}^n (Y_i - \langle X_i, \beta\rangle)^2 + \kappa(n,d,x)\frac{\|\beta\|_{\ell_1}^2}{n}\Big), \tag{1.9}$$

while the standard LASSO is regularized by the ℓ_1 norm itself. This choice of the exponent is dictated by the complexity of the underlying models: the sequence of balls $(rB_1^d)_{r\geq 0}$ trough the isomorphic profile function $r \to \lambda_\epsilon^*(r)$. Observe that since $\|\beta\|_{\ell_1}/\sqrt{n} \geq \|\beta\|_{\ell_1}^2/n$ when $\|\beta\|_{\ell_1} \leq \sqrt{n}$, a non-exact oracle inequality for the LASSO estimator itself follows from Theorem B, but with a slow rate of $1/\sqrt{n}$. Using the q-th power of the ℓ_1-norm as a penalty function for the L_q-risk yields a fast $1/n$ rate (see Theorem C).

We will perform this study for the L_q-loss function, and in which case, for every $\beta \in \mathbb{R}^d$,

$$R^{(q)}(\beta) = \mathbb{E}|Y - \langle X, \beta\rangle|^q \text{ and } R_n^{(q)}(\beta) = \frac{1}{n}\sum_{i=1}^n |Y_i - \langle X_i, \beta\rangle|^q.$$

The following result is obtained only under the assumption that Y and $\|X\|_{\ell_\infty^d}$ belong to L_{ψ_q}. Since there are no "statistically reasonable" ψ_q variables for $q > 2$, it sounds more "statistically relevant" to assume that $|Y|, \|X\|_{\ell_\infty^d}$ are almost surely bounded when one wants results for the L_q-risk with $q > 2$, or that the functions are in L_{ψ_2} for $q = 2$ (for example, linear models with sub-gaussian noise and a sub-gaussian design satisfy this condition).

Theorem C. *Let $q \geq 2$. There exist constants c_0 and c_1 that depend only on q for which the following holds. Assume that there exists $K(d) > 0$ such that $\|Y\|_{\psi_q}$, $\Big\|\|X\|_{\ell_\infty^d}\Big\|_{\psi_q} \leq K(d)$. For $x > 0$ and $0 < \epsilon < 1/2$, let*

$$\lambda(n,d,x) = c_0 K(d)^q (\log n)^{(4q-2)/q} (\log d)^2 (x + \log n)$$

and consider the RERM estimator

$$\widehat{\beta}_n \in \operatorname{Arg}\min_{\beta\in\mathbb{R}^d} \Big(R_n^{(q)}(\beta) + \lambda(n,d,x)\frac{\|\beta\|_{\ell_1}^q}{n\epsilon^2}\Big).$$

Then, with probability greater than $1-12\exp(-x)$*, the* L_q*-risk of* $\widehat{\beta}_n$ *satisfies*

$$R^{(q)}(\widehat{\beta}_n) \leq \inf_{\beta\in\mathbb{R}^d}\Big((1+2\epsilon)R^{(q)}(\beta)+\eta(n,d,x)\frac{(1+\|\beta\|_{\ell_1}^q)}{n\epsilon^2}\Big),$$

where $\eta(n,d,x)=c_1K(d)^q(\log n)^{(4q-2)/q}(\log d)^2(x+\log n)$.

Procedures based on the ℓ_1-norm as a regularizing or constraint function have been studied extensively in the last few years. We only mention a small fraction of this very extensive body of work [6, 7, 8, 13, 15, 23, 26, 27, 41, 42, 45, 46]. In fact, it is almost impossible to make a proper comparison even with the results mentioned in this partial list. Some of these results are close enough in nature to Theorem C to allow a comparison. In particular, in [4], the authors prove that with high probability, the LASSO satisfies an exact oracle inequality with a residual term $\sim \|\beta\|_{\ell_1}/\sqrt{n}$ up to logarithm factors, under tail assumptions on Y and X. In [7], upper bounds on the risks $\mathbb{E}[\langle X,\widehat{\beta}_n-\beta_0\rangle^2]$ and $\left\|\widehat{\beta}_n-\beta_0\right\|_{\ell_1}$ were obtained for a weighted LASSO $\widehat{\beta}_n$ when $\mathbb{E}(Y|X)=\langle X,\beta_0\rangle$ for β_0 with short support. Exact oracle inequalities for RERM using an entropy based criterion or on an ℓ_p criterion (with p close to 1) were obtained in [14, 15] for any convex and regular loss function and with fast rates. Similar bounds were obtained in [42] for a RERM using a weighted ℓ_1-criterion. In [6] it is shown that the LASSO and Dantzig estimators [8] satisfy oracle inequalities in the deterministic design setup and under the REC condition. In fact, in most of these results the authors obtained exact oracle inequalities with an optimal residual term of $|\mathrm{Supp}(\beta_0)|(\log d)/n$, which is clearly better than the rate $\|\beta\|_{\ell_1}^2/n$ obtained in Theorem C for the quadratic loss and in the same context.

However, it is important to note that all these exact oracle inequalities were obtained under an assumption that is similar in nature to the Restricted Isometry Property (RIP), whereas in Theorem C one does not need that kind of assumption on the design. Although it seems strange that it is possible to obtain fast rates without RIP there is nothing magical here. In fact, the isomorphic argument used to prove Theorem B (and thus Theorem C) shows that the random operator $\beta\in\mathbb{R}^d\to n^{-1/2}\sum_{i=1}^n(Y_i-\langle X_i,\beta\rangle)e_i\in\mathbb{R}^n$ satisfies some sort of an RIP, which actually coincides with the RIP property in the noise-free case $Y=\langle X,\beta_0\rangle$ for an isotropic design. This indicates that RIP is not the key property in establishing oracle inequalities for the prediction risk, but rather, the "isomorphic profile" of the problem at hand, which takes into account the structure of the class of functions.

Finally, a word about notation. Throughout, we denote absolute constants or constants that depend on other parameters by c, C, c_1, c_2, etc., (and, of

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

2. Preliminaries to the proofs. In this section we obtain a general bound on $\mathbb{E}\,\|P - P_n\|_{(\ell_F)_\lambda}$ for the L_q-loss when $q \geq 2$, and show that a Bernstein type condition is satisfied under weak assumption on the loss function.

2.1. *Isomorphic properties of the loss class.* The *isomorphic property* of a functions class measures the "level" at which empirical means and actual means are equivalent. The notion was introduced in this context in [4]. Although it is not a necessary feature of this method, if one wishes the isomorphic property to hold with exponential probability, one can use a high probability deviation bound on the supremum of the localized process. A standard way (though not the only way, or even the optimal way!) of obtaining such a result is through of Talagrand concentration inequality [39] applied to localizations of the function class, combined with a good control of the variance in terms of the expectation (a Bernstein type condition). When applied to an excess loss class, this argument leads to exact oracle inequalities (see for example, [33, 5]). Here we are interested in non-exact oracle inequality, and thus, we will study the isomorphic properties of the loss class. To make the presentation simpler, we are not dealing with a fully "unbounded theory" like in [28], but rather that the class has an envelope function which is bounded in ψ_1, and to follow the path of [33], in which one obtains the desired high probability bounds using Talagrand's concentration theorem. Since we would like to avoid the assumption that the class consists of uniformly bounded functions, an important part of our analysis is the following ψ_1 version of Talagrand's inequality [1].

THEOREM 2.1. *There exists an absolute constant $K > 0$ for which the following holds. Let $Z_1, \ldots, Z_n$ be n i.i.d. random variables with values in a space $\mathcal{Z}$ and let G be a countable class of real-valued measurable functions defined on $\mathcal{Z}$. For every $x > 0$ and $\alpha > 0$, with probability greater than*

$1-4\exp(-x)$,

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

Using the same truncation argument as in [1], it follows that for every single function $g \in L_2(P)$ and every $\alpha, x > 0$, with probability greater than $1-4\exp(-x)$,

$$P_n g \leq (1+\alpha)Pg + K\sqrt{\frac{xPg^2}{n}} + K(1+\alpha^{-1})\frac{b_n(g)x}{n}$$

and, in particular, if there exists some $B_n \geq 0$ for which $Pg^2 \leq B_n Pg + B_n^2/n$, then for every $0 < \alpha < 1$ and $x > 0$, with probability greater than $1-4\exp(-x)$,

$$P_n g \leq (1+2\alpha)Pg + K'(1+\alpha^{-1})(b_n(g)+B_n)\frac{x+1}{n}. \tag{2.1}$$

Theorem 2.1 can be extended to classes G satisfying some separability property like condition (M) in [25]. We apply Theorem 2.1 in this context and it will be implicitly assumed that every time we use Theorem 2.1, this separability condition holds. In particular, Theorem 2.1 will be applied to the localized sets $V(\ell_F)_\lambda$ to get non-exact oracle inequalities for the ERM algorithm and to the family $(V(\ell_{F_r})_\lambda)_{r\geq 0}$ to get non-exact regularized oracle inequalities for the RERM procedure.

Observe that Theorem 2.1 requires that the envelope function $\sup_{g\in G}|g|$ is sub-exponential, but since $\|\max_{1\leq i\leq n} X_i\|_{\psi_1} \lesssim \|X\|_{\psi_1}\log n$ it follows that $b_n(\ell_F)$ is not much larger than $\|\sup_{g\in G} g(X)\|_{\psi_1}$. However, this condition can be a major drawback. For instance, if the set G consists of linear functions indexed by the Euclidean sphere $\mathcal{S}^{d-1}$, and X is the standard gaussian measure on $\mathbb{R}^d$, the resulting envelope function is bounded in $\psi_1(\mu)$, but its norm is of the order of $\sqrt{d}$. In Theorem C, we bypass this obstacle by assuming that $\|Y\|_{\psi_q}, \left\|\|X\|_{\ell_\infty^d}\right\|_{\psi_q} \leq K(d)$. This assumption is far better suited for situations in which the indexing class is small – like localized subsets of B_1^d that appear naturally in LASSO type results.

THEOREM 2.2. *Let F be a functions class and assume that there exists $B_n \geq 0$ such that for every $f \in F$, $P\ell_f^2 \leq B_n P\ell_f + B_n^2/n$. If $0 < \epsilon < 1/2$ and $\lambda_\epsilon^* > 0$ satisfy that*

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

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

$$P\ell_f \leq (1+2\epsilon)P_n\ell_f + \rho_n(x),$$

where, for K the constant appearing in Theorem 2.1,

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

Proof. The proof follows the ideas from [4]. Fix $\lambda > 0$ and $x > 0$, and note that by Theorem 2.1, with probability larger than $1 - 4\exp(-x)$,

(2.2)
$$\|P - P_n\|_{V(\ell_F)_\lambda} \leq 2\mathbb{E}\|P - P_n\|_{V(\ell_F)_\lambda} + K\sigma(V(\ell_F)_\lambda)\sqrt{\frac{x}{n}} + Kb_n(V(\ell_F)_\lambda)\frac{x}{n}.$$

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

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

Moreover, since $V(\ell_F)$ is star-shaped, $\lambda \geq 0 \to \phi(\lambda) = \mathbb{E}\|P - P_n\|_{V(\ell_F)_\lambda}/\lambda$ is non-increasing, and since $\phi(\lambda^*_\epsilon) \leq \epsilon/8$ and $\rho_n(x) \geq \lambda^*_\epsilon$ then

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

Combined with (2.2), there exists an event $\Omega_0(x)$ of probability greater than $1 - 4\exp(-x)$, and on $\Omega_0(x)$,

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

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

2.2. *The Bernstein condition of loss functions classes.* In Theorem A, the desired concentration properties (and thus the fast rates in Theorem C) rely on a Bernstein type condition, that for every $f \in F$,

$$P\ell_f^2 \leq B_nP\ell_f + B_n^2/n. \tag{2.3}$$

Assumption (2.3) is trivially satisfied when the loss functions are positive and uniformly bounded: if $0 \leq \ell_f \leq B$ then $P\ell_f^2 \leq BP\ell_f$. It also turns out that (2.3) does not require any "global" structural assumption on F and is trivially verified if class members have sub-exponential tails.

LEMMA 2.3. *Let X be a nonnegative subexponential random variable. Then for every $z \geq 1$,*

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

PROOF. Fix $\theta > 0$ and note that

$$\begin{aligned}
\mathbb{E}X^2 \mathbb{I}_{X\geq\theta} &= \int_0^\infty 2t\mathbb{P}[X\mathbb{I}_{X\geq\theta} \geq t]dt = \theta^2\mathbb{P}[X \geq \theta] + 2\int_\theta^\infty t\mathbb{P}[X \geq t]dt \\
&\leq 2\theta^2 \exp\big(-\theta/\|X\|_{\psi_1}\big) + 4\int_\theta^\infty t\exp\big(-t/\|X\|_{\psi_1}\big)dt \\
&\leq (2\theta^2 + 4\theta\|X\|_{\psi_1} + 4)\exp\big(-\theta/\|X\|_{\psi_1}\big). \qquad (2.4)
\end{aligned}$$

Since $X \geq 0$, it follows from (2.4) that, for any $\theta > 0$,

$$\begin{aligned}
\mathbb{E}X^2 &\leq \mathbb{E}X^2\mathbb{I}_{X\leq\theta} + \mathbb{E}X^2\mathbb{I}_{X\geq\theta} \\
&\leq \theta\mathbb{E}X + \big(2\theta^2 + 4\theta\|X\|_{\psi_1} + 4\big)\exp\big(-\theta/\|X\|_{\psi_1}\big)\Big).
\end{aligned}$$

The result follows for $\theta = \|X\|_{\psi_1}\log(ez)$. □

In particular, if $\ell_f \geq 0$ and $\|\ell_f\|_{\psi_1} \leq D$ for some $D \geq 1$, then for every $n \geq 1$,

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

2.3. *Upper bounds on $\mathbb{E}\|P - P_n\|_{V(\ell_F)_\lambda}$.* Let H be the loss class associated with F for the ERM or with a class F_r for some $r \geq 0$ for the RERM. The next step is to obtain bounds on the fixed point of the localized process, that is, for some $c_0 < 1$, to find a small λ^* for which

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

Note that the complexity of the star-shaped hull $V(H)$ is not far from the one of H itself. Actually, a bound on the expectation of the supremum of the empirical process indexed by $V(H)_\lambda$ will follow from one on H_μ for different levels $\mu \in \{2^i\lambda : i \in \mathbb{N}\}$. This follows from the peeling argument of [5]: that $V(H)_\lambda \subset \bigcup_{i=0}^\infty \{\theta h : 0 \leq \theta \leq 2^{-i}, h \in H, \mathbb{E}h \leq 2^{i+1}\lambda\}$. Therefore, setting $H_\mu = \{h \in H : \mathbb{E}h \leq \mu\}$,for all $\mu > 0$ and $R^* = \inf_{h\in H}\mathbb{E}h$,

$$\mathbb{E}\|P - P_n\|_{V(H)_\lambda} \leq \sum_{\{i:2^{i+1}\lambda\geq R^*\}} 2^{-i}\mathbb{E}\|P - P_n\|_{H_{2^{i+1}\lambda}}, \qquad (2.5)$$

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

Let us mention that a naive attempt to control these empirical processes using a contraction argument is likely to fail, and will result in slow rates even in very simple cases (for example, a regression model with a bounded design). We refer to [11, 32, 34] for more details.

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

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

The complexity of these random sets will be measured via a metric invariant, called the γ_2-functional, introduced by Talagrand as a part of the generic chaining mechanism.

DEFINITION 2.4 ([40]). *Let (T, d) be a semi-metric space. An admissible sequence of T is a sequence $(T_s)_{s\in\mathbb{N}}$ of subsets of T such that $|T_0| \leq 1$ and $|T_s| \leq 2^{2^s}$ for any $s \geq 1$. We define*

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

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

We refer the reader to [40] for an extensive survey on chaining methods and on the γ_2-functionals. In particular, one can bound the γ_2-functional using an entropy integral

$$\gamma_2(T, d) \lesssim \int_0^{\mathrm{diam}(T,d)} \sqrt{\log N(T, d, \epsilon)}\, d\epsilon \tag{2.6}$$

where $N(T, d, \epsilon)$ is the minimal number of balls of radius ϵ with respect to the metric d needed to cover T, and $\mathrm{diam}(T, d)$ is the diameter of the metric space (T, d).

We will use the γ_2-functional to state our theoretical bounds because there are examples in which $\gamma_2(T, d)$ is significantly smaller that the corresponding entropy integral. However, in all our concrete applications we will use the bound (2.6) since the computation of those is much simpler, the gap is at most logarithmic, and the purpose of this note is not to obtain the optimal estimates but to show that the residual terms in exact and non-exact oracle inequalities could be very different.

Now, we turn to some concrete examples where H is the loss functions class in the regression model with respect to the L_q-loss.

Let $q \geq 2$ and set the L_q-loss function of f to be $\ell_f^{(q)}(x,y) = |y - f(x)|^q$. In this case, the L_q-loss functions class localized at some level μ is $(\ell_F^{(q)})_\mu = \{\ell_f^{(q)} : f \in F, \mathbb{E}\ell_f^{(q)} \leq \mu\}$.

The following result is a combination of a truncation argument and Rudelson's L_∞^n method. To formulate it, set $M = \left\| \sup_{\ell \in (\ell_F^{(q)})_\mu} |\ell| \right\|_{\psi_1}$, for any $A \subset \mathbb{R}^d$, let $\widetilde{A} = A \cup -A$, and if $F^{(\mu)} = \{f \in F : P\ell_f^{(q)} \leq \mu\}$, put $U_n = \mathbb{E}\gamma_2^2(\widetilde{P_\sigma F^{(\mu)}}, \ell_\infty^n)$.

PROPOSITION 2.5. *For every $q \geq 2$, there exists a constant c_0 depending only on q for which the following holds. If F is a class of functions, then for any $\mu > 0$,*

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

$$c_0 \max\left[\sqrt{\mu\frac{U_n}{n}}\sqrt{(M\log n)^{(q-2)/q}}, \frac{U_n}{n}(M\log n)^{(q-2)/q}, \frac{M\log n}{n}\right].$$

PROOF. Let $\phi(h) = \operatorname{sign}(h)\min(|h|,\theta)$ where $\theta > 0$ is a threshold to be fixed later. For $f \in F$, set $h_f(x,y) = y - f(x)$, let $H_\mu = \{h_f : f \in F,\ \mathbb{E}|h_f|^q \leq \mu\}$, and note that $|h|^q = |\phi(h)|^q + (|h|^q - \theta^q)\mathbb{1}_{|h|\geq\theta}$. Thus,

$$\begin{aligned}
&\mathbb{E}\|P - P_n\|_{(\ell_F^{(q)})_\mu} = \mathbb{E}\sup_{h\in H_\mu}|(P_n - P)(|h|^q)| \\
&\leq \mathbb{E}\sup_{h\in H_\mu}|(P_n - P)(|\phi(h)|^q)| + \mathbb{E}\sup_{h\in H_\mu} P_n|h|^q\mathbb{1}_{|h|\geq\theta} + \sup_{h\in H_\mu} P|h|^q\mathbb{1}_{|h|\geq\theta} \\
&\leq \mathbb{E}\sup_{h\in H_\mu}|(P_n - P)(|\phi(h)|^q)| + 2\mathbb{E}\Big(\sup_{h\in H_\mu}|h|^q\mathbb{1}_{|h|\geq\theta}\Big).
\end{aligned}$$

To upper bound the truncated part of the process, consider the empirical diameter $D_n = \sup_{h\in H_\mu}\left(P_n|\phi(h)|^{2q-2}\right)^{\frac{1}{2q-2}}$. By the Ziné-Ginn symmetrization Theorem [43] and the upper bound on a Rademacher process by a Gaussian one,

$$\mathbb{E}\sup_{h\in H_\mu}|(P_n - P)(|\phi(h)|^q)| \leq \frac{c_0}{\sqrt{n}}\mathbb{E}\mathbb{E}_g\sup_{h\in H_\mu}\Big|\frac{1}{\sqrt{n}}\sum_{i=1}^n g_i|\phi(h)(X_i,Y_i)|^q\Big|$$

where $g_1, \ldots, g_n$ are n independent standard random variables and $\mathbb{E}_g$ denotes the expectation with respect to those variables. For a fixed sample $(X_i, Y_i)_{i=1}^n$, let $(Z(h))_{h\in H_\mu}$ be the gaussian process defined by $Z(h) = n^{-1/2}\sum_{i=1}^n g_i|\phi(h)(X_i,Y_i)|^q$. If $f, g \in F$ then

$$\begin{aligned}\mathbb{E}_g(Z(h_f) - Z(h_g))^2 &= \frac{1}{n}\sum_{i=1}^n \big(|\phi(h_f)(X_i,Y_i)|^q - |\phi(h_g)(X_i,Y_i)|^q\big)^2 \\ &\le \frac{1}{n}\sum_{i=1}^n q^2|f(X_i)-g(X_i)|^2 \max\big(|\phi(h_f)(X_i,Y_i)|, |\phi(h_g)(X_i,Y_i)|\big)^{2q-2} \\ &\le 2q^2 \max_{1\le i\le n}(f(X_i)-g(X_i))^2 D_n^{2q-2},\end{aligned}$$

where we have used that $\big||\phi(u)|^q - |\phi(v)|^q\big| \le q|u-v|\max\big(|\phi(u)|, |\phi(v)|\big)^{q-1}$ for every $u, v \in \mathbb{R}$. By a standard chaining argument it follows that

$$\mathbb{E}_g \sup_{f\in F^{(\mu)}} \Big|\frac{1}{\sqrt{n}}\sum_{i=1}^n g_i|\phi(h_f)(X_i,Y_i)|^q\Big| \le c_1 q\gamma_2(\widetilde{P_\sigma F^{(\mu)}}, \ell_\infty^n) D_n^{q-1}, \tag{2.7}$$

and thus, $\mathbb{E}\sup_{h\in H_\mu}|(P_n - P)(|\phi(h)|^q)| \le c_2 q\sqrt{\frac{\mathbb{E}\gamma_2^2(\widetilde{P_\sigma F^{(\mu)}}, \ell_\infty^n)}{n}}\sqrt{\mathbb{E}D_n^{2q-2}}$.

A bound on the diameter follows from (2.7) and the contraction principle:

$$\begin{aligned}\mathbb{E}D_n^{2q-2} \le& \mathbb{E}\sup_{h\in H_\mu}|(P_n-P)(|\phi(h)|^{2q-2})| + \sup_{h\in H_\mu} P|\phi(h)|^{2q-2} \\ \le& \frac{c_2 q\theta^{q-2}}{\sqrt{n}}\mathbb{E}_g \sup_{h\in H_\mu}\Big|\frac{1}{\sqrt{n}}\sum_{i=1}^n g_i|\phi(h)(X_i,Y_i)|^q\Big| + \theta^{q-2}\mu \\ \le& c_2 q\theta^{q-2}\sqrt{\frac{U_n\mathbb{E}D_n^{2q-2}}{n}} + \theta^{q-2}\mu,\end{aligned}$$

implying that $\mathbb{E}D_n^{2q-2} \le c_3\max\big(q^2\theta^{2q-4}U_n/n, \theta^{q-2}\mu\big)$ and so

$$\mathbb{E}\sup_{h\in H_\mu}|(P_n-P)(|\phi(h)|^q)| \le c_4 q\max\Big(\frac{qU_n\theta^{q-2}}{n}, \sqrt{\frac{U_n\theta^{q-2}\mu}{n}}\Big). \tag{2.8}$$

Next, observe that for $q = 2$, the right hand side in (2.8) does not depend on the truncation level θ, and thus one may take θ arbitrarily large, leading to the desired result.

For $q \ne 2$, consider the unbounded part of the process. Since the envelope

function of H_μ exhibits a subexponential decay, then

$$\begin{aligned}\mathbb{E}\Big(\sup_{h\in H_\mu}|h|^q\mathbb{I}_{|h|\geq\theta}\Big) &= \int_0^\infty \mathbb{P}\Big[\sup_{h\in H_\mu}|h|^q\mathbb{I}_{|h|\geq\theta}\geq t\Big]dt\\ &= \theta^q\mathbb{P}\Big[\sup_{h\in H_\mu}|h|\geq\theta\Big]+\int_{\theta^q}^\infty \mathbb{P}\Big[\sup_{h\in H_\mu}|h|^q\geq t\Big]dt\\ &\leq 2\theta^q\exp\big(-\theta^q/M\big)+2M\exp\big(-\theta^q/M\big).\end{aligned}$$

The result follows by taking $\theta^q = M\log n$. □

3. Proof of Theorem A. In this section, we will present the proof of Theorem A, which follows the same ideas as [5, 4] for the excess loss.

LEMMA 3.1. *There exists an absolute constant $c_0>0$ for which the following holds. Let F be a class of functions and assume that there is some B_n such that for every $f\in F$, $P\ell_f^2\leq B_nP\ell_f+B_n^2/n$. For $x>0$ and $0<\epsilon<1/2$, consider an event $\Omega_0(x)$ on which for every $f\in F$,*

$$R(f)\leq(1+2\epsilon)R_n(f)+\rho_n(x),$$

where $\rho_n(\cdot)$ is some fixed increasing function. Then, with probability greater than $\mathbb{P}(\Omega_0(x))-4\exp(-x)$,

$$R(\widehat{f}_n^{ERM})\leq(1+3\epsilon)\inf_{f\in F}\Big(R(f)+c_0\frac{(b_n(\ell_f)+B_n)(x+1)}{n\epsilon}\Big)+\rho_n(x).$$

PROOF. Fix $x>0$, let K' be the constant introduced in (2.1), consider

$$f^*\in\operatorname{Arg}\min_{f\in F}\Big(R(f)+15K'\frac{(b_n(\ell_f)+B_n)(x+1)}{n\epsilon}\Big),$$

and without loss of generality one assume that the infimum is achieved. By (2.1) (for $\alpha=(\epsilon/2)/(1+2\epsilon)$), the event $\Omega^*(x)$ on which

$$R_n(f^*)\leq\frac{1+3\epsilon}{1+2\epsilon}R(f^*)+5K'\frac{(b_n(\ell_{f^*})+B_n)(x+1)}{n\epsilon},$$

has probability greater than $1-4\exp(-x)$. Hence,

$$-(1+3\epsilon)R(f^*)\leq-(1+2\epsilon)R_n(f^*)+15K'\frac{(b_n(\ell_{f^*})+B_n)(x+1)}{n\epsilon},$$

and on $\Omega_0(x)\cap\Omega^*(x)$, every f in F satisfies that

$$R(f)-(1+3\epsilon)R(f^*)\leq(1+2\epsilon)\big(R_n(f)-R_n(f^*)\big)+\rho_n(x)+15K'\frac{(b_n(\ell_{f^*})+B_n)(x+1)}{n\epsilon}.$$

Since $R_n(\widehat{f}_n^{ERM}) - R_n(f^*) \leq 0$ then

$$R(\widehat{f}_n^{ERM}) \leq (1+3\epsilon)R(f^*) + 15K' \frac{(b_n(\ell_{f^*}) + B_n)(x+1)}{n\epsilon} + \rho_n(x)$$

and the claim now follows from the choice of f^*. □

Proof of Theorem A: Let $x > 0$, $0 < \epsilon < 1/2$ and put

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

By Theorem 2.2, the event $\Omega_0(x)$, on which every $f \in F$ satisfies that

$$R(f) \leq (1+2\epsilon)R_n(f) + \rho_n(x),$$

has probability greater than $1 - 4\exp(-x)$. Now, the result follows from Lemma 3.1.

The remark following Theorem A, that if ℓ is nonnegative, then ℓ_F satisfies a Bernstein type condition with $B_n \sim \operatorname{diam}(\ell_F, \psi_1)\log(en)$ follows from Lemma 2.3.

4. Proof of Theorem B. Although the proof of Theorem B seems rather technical, the idea behind it is rather simple. First, one needs to find a "trivial" bound on $\operatorname{crit}(\widehat{f}_n^{RERM})$, giving preliminary information on where one must look for the RERM function (this is the role played by the function α_n). Then, one combines peeling and fixed point arguments to identify the exact location of the RERM.

Note that for $F = \cup_{r\geq 0} F_r$, we have $\operatorname{crit}(f) = \infty$ for all $f \in \mathcal{F}\backslash F$. Therefore, without loss of generality, we can replace the set $\mathcal{F}$ by F in both the definition of the RERM in (1.4) and in the non-exact regularized oracle inequality of Theorem B.

We begin with the following rough estimate on the criterion of the RERM. In the case where there is a trivial bound $\operatorname{crit}(f) \leq C_n$, for all $f \in F$ then it follows that for any $0 < \epsilon < 1/2$ and $x > 0$, $\operatorname{crit}(\widehat{f}_n^{RERM}) \leq C_n = \alpha_n(\epsilon, x)$. Turning to the second case stated in Assumption 1.1, recall that $r \to \lambda_\epsilon^*(r)$ tends to infinity with r and there exists $K_1 > 0$ such that for every $(r,x) \in \mathbb{R}_+ \times \mathbb{R}_+^*$, $2\rho_n(r,x) \leq \rho_n(K_1(r+1), x)$. Hence, for every $x > 0$ and $0 < \epsilon < 1/2$, we set α_n to satisfy that

$$\alpha_n(\epsilon, x) \geq \max\Big[K_1(\operatorname{crit}(f_0) + 2),$$
$$(\lambda_\epsilon^*)^{-1}\big((1+2\epsilon)(3R(f_0) + 2K'(b_n(\ell_{f_0}) + B_n(\operatorname{crit}(f_0)))((x+1)/n)\big)\Big],$$

where f_0 is any fixed function in F (for instance, when $0 \in F$, one may take $f_0 = 0$), and $(\lambda_\epsilon^*)^{-1}$ is the generalized inverse function of λ_ϵ^*. In this case, we prove the following high probability bound on $\mathrm{crit}(\widehat{f}_n^{RERM})$.

LEMMA 4.1. *Assume that $r \to \lambda_\epsilon^*(r)$ tends to infinity when r tends to infinity and that there exists $K_1 > 0$ such that for every $(r,x) \in \mathbb{R}_+ \times \mathbb{R}_+^*$, $2\rho_n(r,x) \le \rho_n(K_1(r+1),x)$. Then, under the assumptions of Theorem B, for every $x > 0$ and $0 < \epsilon < 1/2$, with probability greater than $1 - 4\exp(-x)$, $\mathrm{crit}(\widehat{f}_n^{RERM}) \le \alpha_n(\epsilon,x)$.*

PROOF. By the definition of $\widehat{f}_n^{RERM}$,

$$R_n(\widehat{f}_n^{RERM}) + \frac{2}{1+2\epsilon}\rho_n\big(\mathrm{crit}(\widehat{f}_n^{RERM})+1, x+\log\alpha_n(\epsilon,x)\big) \le R_n(f_0) + \frac{2}{1+2\epsilon}\rho_n\big(\mathrm{crit}(f_0)+1, x+\log\alpha_n(\epsilon,x)\big).$$

Since ℓ is nonnegative, then $R_n(\widehat{f}_n^{RERM}) \ge 0$, and thus

$$\begin{aligned}&\rho_n\big(\mathrm{crit}(\widehat{f}_n^{RERM})+1, x+\log\alpha_n(\epsilon,x)\big)\\ &\quad\le (1+2\epsilon)R_n(f_0)/2 + \rho_n\big(\mathrm{crit}(f_0)+1, x+\log\alpha_n(\epsilon,x)\big)\\ &\quad\le \max\Big((1+2\epsilon)R_n(f_0), 2\rho_n\big(\mathrm{crit}(f_0)+1, x+\log\alpha_n(\epsilon,x)\big)\Big).\end{aligned}$$

Since $\rho_n(r,x) \ge \lambda_\epsilon^*(r)$, for all $r \ge 0$, one of the following two situations occurs: either

$$\lambda_\epsilon^*(\mathrm{crit}(\widehat{f}_n^{RERM})) \le (1+2\epsilon)R_n(f_0),$$

or, noting that for every $(r,x) \in \mathbb{R}_+ \times \mathbb{R}_+^*$, $2\rho_n(r,x) \le \rho_n(K_1(r+1),x)$, then

$$\begin{aligned}\rho_n\big(\mathrm{crit}(\widehat{f}_n^{RERM})+1, x+\log\alpha_n(\epsilon,x)\big) &\le 2\rho_n\big(\mathrm{crit}(f_0)+1, x+\log\alpha_n(\epsilon,x)\big)\\ &\le \rho_n\big(K_1(\mathrm{crit}(f_0)+2), x+\log\alpha_n(\epsilon,x)\big),\end{aligned}$$

and since ρ_n is monotone in r then $\mathrm{crit}(\widehat{f}_n^{RERM}) \le K_1(\mathrm{crit}(f_0)+2)$.

Hence, in both cases

$$\mathrm{crit}(\widehat{f}_n^{RERM}) \le \max\big((\lambda_\epsilon^*)^{-1}\big((1+2\epsilon)R_n(f_0)\big), K_1(\mathrm{crit}(f_0)+2)\big). \tag{4.1}$$

On the other hand, according to (2.1), with probability greater than $1 - 4\exp(-x)$, $R_n(f_0) \le 3R(f_0) + 2K'(b_n(\ell_{f_0}) + B_n(\mathrm{crit}(f_0)))(x+1)/n$. The result follows by plugging the last inequality in (4.1) and since λ_ϵ is non-decreasing. □

The next step is to find an "isomorphic" result for $\widehat{f}_n^{RERM}$. The idea is to divide the set given by the trivial estimate on $\mathrm{crit}(\widehat{f}_n^{RERM})$ into level sets and analyze each piece separately.

LEMMA 4.2. *Under the assumptions of Theorem B, for every $x > 0$, with probability greater than $1 - 8\exp(-x)$,*

$$R(\widehat{f}_n^{RERM}) \le (1+2\epsilon)R_n(\widehat{f}_n^{RERM}) + \rho_n\big(\mathrm{crit}(\widehat{f}_n^{RERM}) + 1, x + \log\alpha_n(\epsilon,x)\big).$$

PROOF. Let $\Omega_0(x)$ be the event

$$\frac{R(\widehat{f}_n^{RERM}) - R_n(\widehat{f}_n^{RERM})}{2\epsilon R_n(\widehat{f}_n^{RERM}) + \rho_n\big(\mathrm{crit}(\widehat{f}_n^{RERM}) + 1, x + \log\alpha_n(\epsilon,x)\big)} \ge 1,$$

and we will show that this event has the desired small probability.

Clearly,

$$\mathbb{P}[\Omega_0(x)] \le \mathbb{P}\Big[\Omega_0(x) \cap \{\mathrm{crit}(\widehat{f}_n^{RERM}) \le \alpha_n(\epsilon,x)\}\Big] + \mathbb{P}\big[\mathrm{crit}(\widehat{f}_n^{RERM}) > \alpha_n(\epsilon,x)\big],$$

and by Lemma 4.1, $\mathbb{P}\big[\mathrm{crit}(\widehat{f}_n^{RERM}) > \alpha_n(\epsilon,x)\big] \le 4\exp(-x)$ in the second case of Assumption 1.1 or $\mathbb{P}\big[\mathrm{crit}(\widehat{f}_n^{RERM}) > \alpha_n(\epsilon,x)\big] = 0$ when there is a trivial bound on the criterion. Therefore, in any case, we have $\mathbb{P}\big[\mathrm{crit}(\widehat{f}_n^{RERM}) > \alpha_n(\epsilon,x)\big] \le 4\exp(-x)$.

Recall that $F_i = \{f \in F : \mathrm{crit}(f) \le i\}$, for all $i \in \mathbb{N}$, and since ρ_n is monotone in r then

$$\begin{aligned}
&\mathbb{P}\big[\Omega_0(x) \cap \{\mathrm{crit}(\widehat{f}_n^{RERM}) \le \alpha_n(\epsilon,x)\}\big] \\
&\quad\le \sum_{i=0}^{\lfloor\alpha_n(\epsilon,x)\rfloor} \mathbb{P}\big[\Omega_0(x) \cap \{i \le \mathrm{crit}(\widehat{f}_n^{RERM}) \le i+1\}\big] \\
&\quad\le \sum_{i=0}^{\lfloor\alpha_n(\epsilon,x)\rfloor} \mathbb{P}\Big[\exists f \in F_{i+1} : R(f) \ge (1+2\epsilon)R_n(f) + \rho_n(i+1, x+\log\alpha_n(\epsilon,x))\Big].
\end{aligned}$$

By Theorem 2.2, for every $t > 0$ and $i \in \mathbb{N}$, with probability greater than $1 - 4\exp(-t)$, for every $f \in F_{i+1}$, $P\ell_f \le (1+2\epsilon)P_n\ell_f + \rho_n(i+1,t)$. In particular,

$$\begin{aligned}
&\mathbb{P}\Big[\exists f \in F_{i+1} : R(f) \ge (1+2\epsilon)R_n(f) + \rho_n(i+1, x+\log\alpha_n(\epsilon,x))\Big] \\
&\quad\le 4\exp\big(-(x+\log\alpha_n(\epsilon,x))\big).
\end{aligned}$$

Hence, the claim follows, since

$$\begin{aligned}&\mathbb{P}\big[\Omega_0(x)\cap\{\mathrm{crit}(\widehat{f}_n^{RERM})\leq\alpha_n(\epsilon,x)\}\big]\\&\leq\sum_{i=0}^{\lfloor\alpha_n(\epsilon,x)\rfloor}4\exp\big(-(x+\log\alpha_n(\epsilon,x))\big)\leq 4\exp(-x).\end{aligned}$$

□

Proof of Theorem B: Let $x>0$ and $0<\epsilon<1$. Without loss of generality, we assume that, for the constant K' defined in (2.1), there exists $f^*\in F$ minimizing the function

$$\begin{aligned}f\in F\longrightarrow&(1+3\epsilon)R(f)+\rho_n(\mathrm{crit}(f)+1,x+\log\alpha_n(\epsilon,x))\\&+6K'\frac{(b_n(\ell_f)+B_n(\mathrm{crit}(f))(x+1)}{\epsilon n}.\end{aligned}$$

Let $\Omega^*(x)$ be the event on which

$$R_n(f^*)\leq\frac{1+3\epsilon}{1+2\epsilon}R(f^*)+K'\frac{(b_n(\ell_{f^*})+B_n(\mathrm{crit}(f^*)))(x+1)}{n}\Big(\frac{1+3\epsilon}{\epsilon}\Big).$$

Since $f^*\in F_{\mathrm{crit}(f^*)}$ then $P\ell_{f^*}^2\leq B_n(\mathrm{crit}(f^*))P\ell_{f^*}+B_n^2(\mathrm{crit}(f^*))/n$, and by (2.1) (applied with $\alpha=\epsilon/(1+2\epsilon)$), $\mathbb{P}(\Omega^*(x))\geq 1-4\exp(-x)$.

Consider the event $\Omega_0(x)$, on which

$$R(\widehat{f}_n^{RERM})\leq(1+2\epsilon)R_n(\widehat{f}_n^{RERM})+\rho_n\big(\mathrm{crit}(\widehat{f}_n^{RERM})+1,x+\log\alpha_n(\epsilon,x)\big),$$

and observe that by Lemma 4.2, $\mathbb{P}[\Omega_0(x)]\geq 1-8\exp(-x)$. Therefore, on $\Omega_0(x)\cap\Omega^*(x)$, we have

$$\begin{aligned}&R(\widehat{f}_n^{RERM})+\rho_n\big(\mathrm{crit}(\widehat{f}_n^{RERM})+1,x+\log\alpha_n(\epsilon,x)\big)-(1+3\epsilon)R(f^*)\\&\leq(1+2\epsilon)(R_n(\widehat{f}_n^{RERM})-R_n(f^*))\\&\quad+2\rho_n\big(\mathrm{crit}(\widehat{f}_n^{RERM})+1,x+\log\alpha_n(\epsilon,x)\big)+6K'\frac{(b_n(\ell_{f^*})+B_n(\mathrm{crit}(f^*)))(x+1)}{\epsilon n}\\&\leq(1+2\epsilon)\Big(R_n(\widehat{f}_n^{RERM})+\frac{2}{1+2\epsilon}\rho_n\big(\mathrm{crit}(\widehat{f}_n^{RERM})+1,x+\log\alpha_n(\epsilon,x)\big)\\&\quad-R_n(f^*)-\frac{2}{1+2\epsilon}\rho_n\big(\mathrm{crit}(f^*)+1,x+\log\alpha_n(\epsilon,x)\big)\Big)\\&\quad+2\rho_n\big(\mathrm{crit}(f^*)+1,x+\log\alpha_n(\epsilon,x)\big)+6K'\frac{(b_n(\ell_{f^*})+B_n(\mathrm{crit}(f^*)))(x+1)}{\epsilon n}\\&\leq2\rho_n\big(\mathrm{crit}(f^*)+1,x+\log\alpha_n(\epsilon,x)\big)+6K'\frac{(b_n(\ell_{f^*})+B_n(\mathrm{crit}(f^*)))(x+1)}{\epsilon n}\end{aligned}$$

where the last inequality follows from the definition of $\widehat{f}_n^{RERM}$. Hence, by the choice of f^*, it follows that on $\Omega_1(x) \cap \Omega^*(x)$,

$$\begin{aligned}
&R(\widehat{f}_n^{RERM}) + \rho_n\big(\mathrm{crit}(\widehat{f}_n^{RERM}) + 1, x + \log \alpha_n(\epsilon, x)\big) \\
&\leq (1+3\epsilon)R(f^*) + 2\rho_n\big(\mathrm{crit}(f^*) + 1, x + \log \alpha_n(\epsilon, x)\big) \\
&\qquad + 6K' \frac{(b_n(\ell_{f^*}) + B(\mathrm{crit}(f^*)))(x+1)}{\epsilon n} \\
&= \inf_{f \in F} \Big((1+3\epsilon)R(f) + 2\rho_n\big(\mathrm{crit}(f) + 1, x + \log \alpha_n(\epsilon, x)\big) \\
&\qquad + 6K' \frac{(b_n(\ell_f) + B(\mathrm{crit}(f)))(x+1)}{\epsilon n} \Big).
\end{aligned}$$

5. Proofs of Theorem C. Theorem C follows from a direct application of Theorem B, by estimating the specific function ρ_n and the "Bernstein function" $B_n(r)$.

Consider the family of models $(F_r)_{r \geq 0}$ associated with the ℓ_1-criterion $F_r = \{f_\beta : \|\beta\|_1 \leq r\}$, where $f_\beta(x) = \langle x, \beta \rangle$ is a linear functional on $\mathbb{R}^d$.

LEMMA 5.1. *There exists an absolute constant c_0 for which the following holds. For every μ and $r \geq 0$, and every $\sigma = (X_1, \ldots, X_n)$,*

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

Moreover, if $\Big\| \|X\|_{\ell_\infty^d} \Big\|_{\psi_2} \leq K(d)$ then

$$\big(\mathbb{E}\gamma_2^2(\widetilde{P_\sigma F_r}, \ell_\infty^n)\big)^{1/2} \leq c_0 r K(d) (\log n)^{3/2} (\log d).$$

The proof of the first part of the claim is rather standard and has appeared in one form or another in several places (for example, see [5]). It follows from (2.6) and Maurey's empirical method (cf. [9, 37]). The second part is an immediate corollary of the first one.

Proof of Theorem C: Observe that for every $\beta \in rB_1^d$,

$$\begin{aligned}
&\big\| |Y - \langle X, \beta \rangle|^q \big\|_{\psi_1} = \big\| Y - \langle X, \beta \rangle \big\|_{\psi_q}^q \leq \big(\|Y\|_{\psi_q} + \big\| \langle X, \beta \rangle \big\|_{\psi_q} \big)^q \\
&\leq \big(\|Y\|_{\psi_q} + \|\beta\|_1 \, \big\| \|X\|_\infty \big\|_{\psi_q} \big)^q \leq (K(d))^q (1+r)^q.
\end{aligned}$$

Hence, by Lemma 2.3, one may take $B_n(r) = c_0 (2K(d))^q (1+r)^q \log(en)$.

Next, the ψ_1-norm of the envelope of the class F_r satisfies $\big\| \sup_{\beta \in rB_1^d} |Y - \langle X, \beta \rangle|^q \big\|_{\psi_1} \leq (K(d))^q (1+r)^q$, and by (2.5), Proposition 2.5 and Lemma 5.1,

for every $\lambda > 0$,

$$\begin{aligned}
\mathbb{E}\,\|P-P_n\|_{V(\ell_{F_r}^{(q)})_\lambda} &\le \sum_{i=0}^{\infty} 2^{-i}\mathbb{E}\,\|P-P_n\|_{(\ell_{F_r}^{(q)})_{2^{i+1}\lambda}} \\
&\le c_0\sum_{i=0}^{\infty} 2^{-i}\max\left(\sqrt{2^{i+1}\lambda}\sqrt{\frac{r^2(1+r)^{q-2}h(n,d)}{n}}, \frac{r^2(1+r)^{q-2}h(n,d)}{n},\right. \\
&\qquad\qquad \left.\frac{K(d)^q(1+r)^q(\log n)}{n}\right) \\
&\le c_1\max\left(\sqrt{\lambda}\sqrt{\frac{(1+r)^q h(n,d)}{n}}, \frac{(1+r)^q h(n,d)}{n}\right),
\end{aligned}$$

where $h(n,d) = K(d)^q(\log n)^{(4q-2)/q}(\log d)^2$. Set $\lambda_\epsilon^*(r) = c_2(1+r)^q h(n,d)/(n\epsilon^2)$ and observe that $\mathbb{E}\,\|P-P_n\|_{V(\ell_{F_r}^{(q)})_{\lambda_\epsilon^*(r)}} \le (\epsilon/4)\lambda_\epsilon^*(r)$. Since

$$b_n(\ell_{F_r}^{(q)}) = \|\max_{1\le i\le n}\sup_{f\in F_r}\ell_f^{(q)}(X_i,Y_i)\|_{\psi_1} \le c_3(\log en)\|\sup_{f\in F_r}\ell_f^{(q)}(X,Y)\|_{\psi_1},$$

then one can take $\phi_n(r) = c_3K(d)^q(\log n)(1+r)^q$. Thus,

$$\rho_n(r,x) = c_4\frac{h(n,d)(1+r^q)}{n\epsilon^2}(1+x)$$

is a valid isomorphic function for this problem. It is also easy to check that for $f_0 \equiv 0$, $\log\alpha_n(\epsilon,x) \le c_5\log\big(\max(x,n)\,\|Y\|_{\psi_q}^q\big)$. The result now follows by combining these estimates with Theorem B.

6. Remarks on the differences between exact and non-exact oracle inequalities. The goal of this section is to describe the difference between the analysis used in [4] to obtain exact oracle inequalities for the ERM, and the one used in this note to establish non-exact oracle inequalities for the ERM (Theorem A). Our aim is to indicate why one may get faster rates for non-exact inequalities than for exact ones for the same problem.

One should stress that this is not, by any means, a proof that it is impossible to get exact oracle inequalities with fast rates (there are in fact examples in which the ERM satisfies exact oracle inequalities with fast rates: the linear aggregation problem, [12]). It is not even a proof that the localization method presented here is sharp. A detailed study of the isomorphic method and oracle inequalities for a general subgaussian case (i.e., a sub-exponential squared loss), in the sense that the class F has a bounded diameter in L_{ψ_2} rather than an envelope function, will be presented in [28].

However, we believe that this explanation will help to shed some light on the differences between the two types of inequalities, and refer the reader to [28] for a more detailed and accurate analysis.

Our starting point is the following exact oracle inequality for ERM, which is a mild modification of a result from [4]. The only difference is that it uses Adamczak's ψ_1 version of Talagrand's concentration inequality for empirical processes, instead of Massart's version.

THEOREM 6.1. *There exists an absolute constant $c_0 > 0$ for which the following holds. Let F be a class of functions and assume that there exists $B > 0$ such that for every $f \in F$, $P\mathcal{L}_f^2 \leq BP\mathcal{L}_f$. Let $\mu^* > 0$ be such that $\mathbb{E}\|P_n - P\|_{V(\mathcal{L}_F)_{\mu^*}} \leq \mu^*/8$, and consider an increasing function ρ_n which satisfies that, for every $x > 0$, $\rho_n(x) \geq \max\left(\mu^*, c_0(b_n(\mathcal{L}_F) + B)x/n\right)$. Then, for every $x > 0$, with probability greater than $1 - 8\exp(-x)$, the risk of the ERM satisfies $R(\widehat{f}_n^{ERM}) \leq \inf_{f\in F} R(f) + \rho_n(x)$.*

Roughly put, and as indicated by the theorem, localization arguments are based on two main components:

1. A Bernstein type condition, the essence of which is that it allows one to "translate" localization with respect to the loss or the excess loss to a localization with respect to a natural metric. In particular this leads to the necessary control on the ℓ_2^n diameter of a random coordinate projection of the localized class.
2. The fixed point of the empirical process indexed by the localized star-shaped hull of the loss functions class (for non-exact inequalities) or of the excess loss functions class (for exact ones).

Although the two components seem similar for the exact and non-exact cases, they are very different. Indeed, for a non-exact oracle inequality, the Bernstein type condition is almost trivially satisfied and requires no special properties on the model/output couple (F, Y) – as long as the functions involved have well behaved tails. As such, it is an individual property of every class member (see Lemma 2.3).

On the other hand, the Bernstein condition required for the exact oracle inequality is deeply connected to the geometry of the problem (see, for example, [31]). More accurately, when the target Y is far from the set of multiple minimizers of the risk, $N(F, \ell, X) = \{Y : |\{f \in F : R(f) = \inf_{f\in F} R(f)\}| \geq 2\}$, one can show that a Bernstein condition holds for a large variety of loss function ℓ. However, when the target Y gets closer to the set $N(F, \ell, X)$, the Bernstein constant B degenerates, and leads to rates slower than $1/\sqrt{n}$ even

if F is a two functions class. Hence, the geometry of the problem (the relative position of Y and F) is very important when trying to establish exact oracle inequalities, and the Bernstein condition is truly a "global" property of F.

In particular, this explains the gap that we observed in the example preceding the formulation of Theorem A. In that case, the class is a finite set of functions and the set $N(F,\ell,X)$ is nonempty. Thus, one can find a set F and a target Y in a "bad" position, leading to an excess loss class $\mathcal{L}_F$ with a trivial Bernstein constant (i.e. greater than $\sqrt{n}$). On the other hand, regardless of the choice of Y, the Bernstein constant of ℓ_F is well behaved.

Let us mention that when the gap between exact and non-exact oracle inequalities is only due to the Bernstein condition, it is likely that both ERM and RERM will be suboptimal procedures [22, 29, 19]. In particular, when slow rates are due to a lack of convexity of F (which is closely related to a bad Bernstein constant of $\mathcal{L}_F$), one can consider procedures which "improve the geometry" of the model (for instance, the "starification" method of [2] or the "pre-selection-convexification" method in [17]).

The second aspect of the problem is the fixed point of the localized empirical process. Although the complexity of the sets $\mathcal{L}_F$ and ℓ_F seems similar from a metric point of view ($\mathcal{L}_F$ is just a shift of ℓ_F) the localized star-shaped hull $(\mathcal{L}_F)_\lambda$ and $(\ell_F)_\lambda$ are rather different. Since there are many ways of bounding the empirical process indexed by these localized sets, let us show the difference for one of the methods – based on the random geometry of the classes, and for the sake of simplicity, we will only consider the square loss. Using this method of analysis at hand, the dominant term of the bound on $\mathbb{E}\left\|P-P_n\right\|_{V(\ell_F^{(2)})_\mu}$ (for the loss class) which was obtained in Proposition 2.5 is

$$\sqrt{\mu}\sqrt{\frac{\mathbb{E}\gamma_2^2(\widetilde{P_\sigma F^{(\mu)}},\ell_\infty^n)}{n}}. \tag{6.1}$$

A similar bound was obtained for $\mathbb{E}\left\|P-P_n\right\|_{V(\mathcal{L}_F)_\mu}$ in [33] and [5], in which the dominant term is

$$\sqrt{\Big(\inf_{f\in F} R(f)+\mu\Big)}\sqrt{\frac{\mathbb{E}\gamma_2^2(\widetilde{P_\sigma F^{(\mu)}},\ell_\infty^n)}{n}}. \tag{6.2}$$

If this bound is sharp (and it is in many cases), and since $R^*=\inf_{f\in F} R(f)$ is in general a non-zero constant, the fixed point μ^* of Theorem 6.1 is of the order of $\sqrt{\mathbb{E}\gamma_2^2(\widetilde{P_\sigma F^{(\mu^*)}},\ell_\infty^n)/n}$ and thus leads to a rate decaying slower than

$1/\sqrt{n}$. In contrast, in the non-exact case one has $\lambda_\epsilon^* \sim \mathbb{E}\gamma_2^2(\widetilde{P_\sigma F^{(\lambda_\epsilon^*)}}, \ell_\infty^n)/n$ which is of the order of $1/n$ (up to logarithmic factors) when the complexity $\mathbb{E}\gamma_2^2(\widetilde{P_\sigma F}, \ell_\infty^n)$ is "reasonable".

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

In contrast, because there are functions $\mathcal{L}_f$ that can have an arbitrarily small expectation, the complexity of the localized subsets of the star hull of $\mathcal{L}_F$ (normalized properly, of course), can even increase as λ decreases. This happens in very simple situations; for example, even in regression relative to B_1^d, if $R^* \neq 0$, the complexity of the localized sets remains almost stable and starts to decrease only at a very "low" level λ. This is the reason for the phase transition in the error rate ($\sim \max\{\sqrt{(\log d)/n}, d/n\}$) that one encounters in that problem. The first term is due to the fact that the complexity of the localized sets does not change as λ decreases – up to some critical level, while the second captures what happens when the localized sets begin to "shrink". A concrete example of this phenomenon is treated in the Supplementary material [20] in the Convex aggregation context.

SUPPLEMENTARY MATERIAL

Supplement A: Applications to Matrix Completion, Convex aggregation and Model Selection
(http://lib.stat.cmu.edu/aoas/???/???). In the supplementary file, we apply our main results to the problem of Matrix Completion, Convex aggregation and Model Selection. The aim is to expose the fundamental differences between exact and non-exact oracle inequalities on classical problems.

References.

[1] Radosław Adamczak. A tail inequality for suprema of unbounded empirical processes with applications to Markov chains. *Electron. J. Probab.*, 13:no. 34, 1000–1034, 2008.

[2] Jean-Yves Audibert. No fast exponential deviation inequalities for the progressive mixture rule. Technical report, CERTIS, 2007.

[3] Andrew Barron, Lucien Birgé, and Pascal Massart. Risk bounds for model selection via penalization. *Probab. Theory Related Fields*, 113(3):301–413, 1999.

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

[5] Peter L. Bartlett, Shahar Mendelson, and Joseph Neeman. ℓ_1-regularized linear regression: Persistence and oracle inequalities. *To appear in Probability theory and related fields*, 2009.

[6] Peter J. Bickel, Ya'acov Ritov, and Alexandre Tsybakov. Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.*, 37(4):1705–1732, 2009.
[7] Florentina Bunea, Alexandre Tsybakov, and Marten Wegkamp. Sparsity oracle inequalities for the Lasso. *Electron. J. Stat.*, 1:169–194 (electronic), 2007.
[8] Emmanuel J. Candes and Terence Tao. The Dantzig selector: statistical estimation when p is much larger than n. *Ann. Statist.*, 35(6):2313–2351, 2007.
[9] Bernd Carl. Inequalities of Bernstein-Jackson-type and the degree of compactness of operators in Banach spaces. *Ann. Inst. Fourier (Grenoble)*, 35(3):79–118, 1985.
[10] David L. Donoho. Compressed sensing. *IEEE Trans. Inform. Theory*, 52(4):1289–1306, 2006.
[11] Evarist Giné, Rafał Latała, and Joel Zinn. Exponential and moment inequalities for U-statistics. In *High dimensional probability, II (Seattle, WA, 1999)*, volume 47 of *Progr. Probab.*, pages 13–38. Birkhäuser Boston, Boston, MA, 2000.
[12] Vladimir Koltchinskii. Local Rademacher complexities and oracle inequalities in risk minimization. *Ann. Statist.*, 34(6):2593–2656, 2006.
[13] Vladimir Koltchinskii. The Dantzig selector and sparsity oracle inequalities. *Bernoulli*, 15(3):799–828, 2009.
[14] Vladimir Koltchinskii. Sparse recovery in convex hulls via entropy penalization. *Ann. Statist.*, 37(3):1332–1359, 2009.
[15] Vladimir Koltchinskii. Sparsity in penalized empirical risk minimization. *Ann. Inst. Henri Poincaré Probab. Stat.*, 45(1):7–57, 2009.
[16] Vladimir Koltchinskii and Dmitriy Panchenko. Rademacher processes and bounding the risk of function learning. In *High dimensional probability, II (Seattle, WA, 1999)*, volume 47 of *Progr. Probab.*, pages 443–457. Birkhäuser Boston, Boston, MA, 2000.
[17] Guillaume Lecué and Shahar Mendelson. Aggregation via empirical risk minimization. *Probability Theory and Related fields*, 145(3–4):591–613, 2004.
[18] Guillaume Lecué and Shahar Mendelson. On the optimality of the aggregate with exponential weights for low temperature. *submitted*, 2009.
[19] Guillaume Lecué and Shahar Mendelson. Sharper lower bounds on the performance of the empirical risk minimization algorithm. *Bernoulli*, 16(3):605–613, 2010.
[20] Guillaume Lecué and Shahar Mendelson. Supplementary material to "general non-exact oracle inequalities for classes with a subexponential envelope". 2012.
[21] 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.
[22] Wee S. Lee, Peter L. Bartlett, and Robert C. Williamson. The importance of convexity in learning with squared loss. In *Proceedings of the Ninth Annual Conference on Computational Learning Theory*, pages 140–146. ACM Press, 1996.
[23] Karim Lounici. Sup-norm convergence rate and sign concentration property of Lasso and Dantzig estimators. *Electron. J. Stat.*, 2:90–102, 2008.
[24] Pascal Massart. *Concentration inequalities and model selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, Berlin, 2007. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard.
[25] Pascal Massart and Élodie Nédélec. Risk bounds for statistical learning. *Ann. Statist.*, 34(5):2326–2366, 2006.
[26] Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the lasso. *Ann. Statist.*, 34(3):1436–1462, 2006.
[27] Nicolai Meinshausen and Bin Yu. Lasso-type recovery of sparse representations for high-dimensional data. *Ann. Statist.*, 37(1):246–270, 2009.

[28] Shahar Mendelson. Oracle inequalities and the isomorphic method. Submitted.
[29] Shahar Mendelson. Lower bounds for the empirical minimization algorithm. *IEEE Trans. Inform. Theory*, 54(8):3797–3803, 2008.
[30] Shahar Mendelson. Lower bounds for the empirical minimization algorithm. *IEEE Trans. Inform. Theory*, 54(8):3797–3803, 2008.
[31] Shahar Mendelson. Obtaining fast error rates in nonconvex situations. *J. Complexity*, 24(3):380–397, 2008.
[32] Shahar Mendelson. Empirical processes with a bounded ψ_1 diameter. *To appear in GAFA*, 2010.
[33] Shahar Mendelson and Joseph Neeman. Regularization in kernel learning. *Ann. Statist.*, 38(1):526–565, 2010.
[34] Shahar Mendelson, Alain Pajor, and Nicole Tomczak-Jaegermann. Reconstruction and subgaussian operators in asymptotic geometric analysis. *Geom. Funct. Anal.*, 17(4):1248–1282, 2007.
[35] Shahar Mendelson and Grigoris Paouris. On the generic chaining and the smallest singular value of random matrices with heavy tails. Submitted. arXiv:1108.3886.
[36] B. K. Natarajan. Sparse approximate solutions to linear systems. *SIAM J. Comput.*, 24(2):227–234, 1995.
[37] G. Pisier. Remarques sur un résultat non publié de B. Maurey. In *Seminar on Functional Analysis, 1980–1981*, pages Exp. No. V, 13. École Polytech., Palaiseau, 1981.
[38] Ingo Steinwart and Andreas Christmann. *Support vector machines*. Information Science and Statistics. Springer, New York, 2008.
[39] M. Talagrand. Sharper bounds for Gaussian and empirical processes. *Ann. Probab.*, 22(1):28–76, 1994.
[40] Michel Talagrand. *The generic chaining*. Springer Monographs in Mathematics. Springer-Verlag, Berlin, 2005. Upper and lower bounds of stochastic processes.
[41] Robert Tibshirani. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58(1):267–288, 1996.
[42] Sara A. van de Geer. High-dimensional generalized linear models and the lasso. *Ann. Statist.*, 36(2):614–645, 2008.
[43] Aad W. van der Vaart and Jon A. Wellner. *Weak convergence and empirical processes*. Springer Series in Statistics. Springer-Verlag, New York, 1996. With applications to statistics.
[44] Vladimir Vapnik. *Estimation of dependences based on empirical data*. Springer Series in Statistics. Springer-Verlag, New York, 1982. Translated from the Russian by Samuel Kotz.
[45] Tong Zhang. Some sharp performance bounds for least squares regression with L_1 regularization. *Ann. Statist.*, 37(5A):2109–2144, 2009.
[46] Hui Zou. The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.*, 101(476):1418–1429, 2006.

E-MAIL: Guillaume.Lecue@univ-mlv.fr E-MAIL: shahar@tx.technion.ac.il