SUPPLEMENTARY MATERIAL TO "GENERAL NON-EXACT ORACLE INEQUALITIES FOR CLASSES WITH A SUBEXPONENTIAL ENVELOPE"

By Guillaume Lecué[*]

CNRS, LAMA, Marne-la-vallée, 77454 France

and

By Shahar Mendelson[†]

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

We apply Theorem A to the problem of Convex aggregation and show that the optimal rate of Convex aggregation for non-exact oracle inequalities is much faster than the optimal rate for exact oracle inequalities.

We apply Theorem B to show that regularized procedures based on a nuclear norm criterion 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, an RIP type of assumption or an "incoherence condition" are not needed to obtain fast residual terms in this setup.

Finally, we relate the problem of Model Selection to the problem of regularization and apply Theorem B to obtain non-exact oracle inequalities in the Model Selection setup.

1. Application to the prediction of low-rank matrices. For this application, we observe n i.i.d. couples input/output $(X_i, Y_i)_{1 \leq i \leq n}$ where the input variables $X_1, \ldots, X_n$ take their values in the space $\mathcal{X} = \mathcal{M}_{m \times T}$ of all $m \times T$ matrices with entries in $\mathbb{R}$ and the output variables $Y_1, \ldots, Y_n$ are real-valued. Being given a new input X, the goal is to predict the output Y using a linear function of X when (X, Y) is assumed to have the same probability distribution as the (X_i, Y_i)'s. In this setup, it is now common to assume that there are more covariables than observations ($mT >> n$) and

[*] 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: classification, statistical learning, fast rates of convergence, excess risk, aggregation, margin, complexity of classes of sets, SVM.

thus more information on the best linear prediction of Y by X is required. A common assumption is that Y can be well predicted by a function of the form $\langle X, A_0\rangle = \mathrm{Tr}(X^\top A_0)$ where A_0 is an $m \times T$ matrix of low rank. Once again we will not have to make such an assumption, but it helps to keep this low-dimensional structure in mind.

Indeed, with a "small rank" intuition, it is natural to penalize linear estimators $\langle X, A\rangle$ by $\mathrm{rank}(A)$. Unfortunately, since the $\mathrm{rank}(\cdot)$ function is not convex it cannot be used in practice as a criterion. A more popular choice is to use a convex relaxation of the $\mathrm{rank}(\cdot)$ function: the S_1 norm ("Schatten one" norm) (see [1, 3, 4, 5, 7, 11, 8, 10, 18, 20, 13] and references therein), which is the ℓ_1-norm of the singular values of a matrix. Formally, for every $A \in \mathcal{M}_{m\times T}$, $\|A\|_{S_1} = \sum_{i=1}^{m\wedge T} s_i(A)$, where $s_1(A), \ldots, s_{m\wedge T}(A)$ are the singular values of A and, in general for $p \geq 1$, $\|A\|_{S_p} = \Big(\sum_{i=1}^{m\wedge T} s_i(A)^p\Big)^{1/p}$. The S_1-norm was originally used in this type of problems to study exact reconstruction (see, for example, [7, 19, 6]), but other regularizing functions have been used in this context (e.g. [12, 9, 8]) for the prediction and estimation problems.

In the following result, we apply Theorem B to obtain non-exact oracle inequalities for an S_1-based criterion RERM procedure, under an L_q-loss function for some $q \geq 2$. For every $A \in \mathcal{M}_{m\times T}$ let

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

Again, it seems more "statistically relevant" to assume that $|Y|$ and $\|X\|_{S_2}$ are almost surely bounded rather than bounded in ψ_q for $q > 2$, and the two most interesting cases are the uniformly bounded one and $q = 2$. We have stated the results under the more general ψ_q assumption to point out the places in which the decay properties of the functions involved are really needed – in the hope that it would be possible to improve and extend the results at a later date, by relaxing the ψ_q boundedness assumption.

Theorem D *For every $q \geq 2$ there are constants c_0 and c_1 depending only on q for which the following holds. Let m and T as above and assume that $\|Y\|_{\psi_q}, \Big\|\|X\|_{S_2}\Big\|_{\psi_q} \leq K(mT)$ for some constant $K(mT)$ which depends only on the product mT. Let $x > 0$ and $0 < \epsilon < 1/2$, and put $\lambda(n, mT, x) = c_0 K(mT)^q (\log n)^{(4q-2)/q}(x + \log n)$. Consider the RERM procedure*

$$\widehat{A}_n \in \mathrm{Arg}\min_{A\in\mathcal{M}_{m\times T}} \Big(R_n^{(q)}(A) + \lambda(n, mT, x)\frac{\|A\|_{S_1}^q}{n\epsilon^2}\Big)$$

Then, with probability greater than $1-10\exp(-x)$*, the* L_q*-risk of* $\widehat{A}_n$ *satisfies for* $\eta_\epsilon(n, mT, x) = c_1 K(mT)^q (\log n)^{(4q-2)/q}(x+\log n)$

$$R^{(q)}(\widehat{A}_n) \leq \inf_{A\in\mathcal{M}_{m\times T}} \Big((1+2\epsilon)R^{(q)}(A) + \eta(n, mT, x)\frac{(1+\|A\|_{S_1}^q)}{n\epsilon^2}\Big).$$

Sketch of the proof of Theorem D. The proof of Theorem D follows the same line as the one of Theorem C. The only different ingredient is an entropy estimate that can be found in [8] on the complexity of the Schatten S_1-ball.

Proposition 1.1 ([8]). *There exists an absolute constant* $c_0 > 0$ *such that the following holds. Assume that* $\Big\| \|X\|_{S_2} \Big\|_{\psi_2} \leq K(mT)$*. Then,*

$$\left(\mathbb{E}\gamma_2^2(rB(S_1), \|\cdot\|_{\infty,n})\right)^{1/2} \leq c_0 K(mT) r \log n.$$

■

Once again, in the same spirit as in Theorem C, it can be interesting to note that for the quadratic loss ($q = 2$), the resulting estimator is

$$\widehat{A}_n \in \text{Arg}\min_{A\in\mathcal{M}_{m\times T}} \Big(\frac{1}{n}\sum_{i=1}^n (Y_i - \langle X_i, A\rangle)^2 + \lambda(n, mT, x)\frac{\|A\|_{S_1}^2}{n\epsilon^2}\Big)$$

where the regularizing function uses the *square* of the S_1-norm unlike the classical estimator for this problem which uses the S_1 norm itself as a regularizing function.

The first results in the direction of matrix completion have focused on the exact reconstruction of a low-rank matrix A_0 where $Y = \langle X, A_0\rangle$ [3, 4, 7, 19, 10]. The best results [19, 10] to date are that if the number of measurements n is larger than $\text{rank}(A_0)(m+T)\log(m+T)$ and if the "incoherence condition" holds (see [7] for more details), then with high probability, a constraint nuclear norm minimization algorithm can reconstruct A_0 exactly.

Prediction results and statistical estimation involving low-rank matrices has become a very active field. The most popular methods are RERM based on S_1-norm penalty functions (see for instance [1, 2, 3, 4, 18, 20, 8, 12, 13, 20]). To specify some results, fast rates for the noisy matrix completion problem are derived in [20] – in the context of empirical prediction and under an RIP-type assumption. In [13] the authors prove exact oracle inequalities for the prediction error $\mathbb{E}\langle X, \widehat{A}_n - A_0\rangle^2$ when $\mathbb{E}(Y|X) = \langle X, A_0\rangle$ and A_0

is either of low-rank or of a small S_1-norm, and when the design of X is known. In [18], optimal rates for the quadratic risk were obtained under a "spikiness assumption" on the SVD of A_0, and in [12], fast convergence rates were derived for RERM based on the von Neuman entropy penalization and for a known design. However, so far only a few results have been obtained for the prediction risk as considered here. Probably the closest result in this setup is an exact oracle inequality with slow rates satisfied by a RERM using a mixture of several norms in [8].

Note that for the two applications in Theorem C and D, we obtain fast convergence rates under only tail assumptions on the design X and the output Y for every L_q-loss (for $q \geq 2$). In particular, one does not need to assume that $\mathbb{E}(Y|X)$ is a linear combination of the covariables of X, nor that Y has any low-dimensional structure. If one happens to be in a low-dimensional situation, the residual terms of Theorem C and D will be small. Hence, the ℓ_1 and S_1 based RERM procedures used there automatically adapt to this low-dimensional structure.

2. Non-exact oracle inequalities for the Convex aggregation problem.

The problem of Convex aggregation is the following: consider a finite model $F = \{f_1, \ldots, f_M\}$ for some $M \in \mathbb{N}$ and try to find a procedure that is "as good as" the best convex combination of elements in F. To define what is meant by "as good as", we introduce some notation.

For any $\lambda = (\lambda_1, \ldots, \lambda_M) \in \mathbb{R}^M$, let $f_\lambda = \sum_{j=1}^M \lambda_j f_j$ and the convex hull of F is the set $\mathrm{conv}(F) = \left\{ f_\lambda : \sum_{j=1}^M \lambda_j = 1 \text{ and } \lambda_j \geq 0 \right\}$. There are many different ways of defining the convex aggregation problem. The one that we will be interested in is the following: for some $0 \leq \epsilon \leq 1/2$ construct a procedure $\tilde{f}_n$ such that, for any $x > 0$ with probability larger than $1 - \exp(-x)$,

$$R(\tilde{f}_n) \leq (1+\epsilon) \inf_{f \in \mathrm{conv}(F)} R(f) + r_n(M) \tag{2.1}$$

where the residual term $r_n(M)$ should be as small as possible. From both mathematical and statistical point of view, the most interesting case to study is for $\epsilon = 0$. In this case, it follows from classical minimax results (cf. [21]) that no algorithm can do better than the rate

$$\psi_n^C(M) = \begin{cases} M/n & \text{if } M \leq \sqrt{n}, \\ \sqrt{\frac{\log\left(eM/\sqrt{n}\right)}{n}} & \text{otherwise.} \end{cases} \tag{2.2}$$

It is shown in [21] that there is a procedure $\tilde{f}_n$ achieving this rate in expectation: $\mathbb{E}R(\tilde{f}_n) \leq \inf_{f \in \mathrm{conv}(F)} R(f) + \psi_n^C(M)$ and in [14], the ERM is proved to achieve this optimal rate in deviation (we refer to [21] and [14] for more details).

In this setup, we apply Theorem A to obtain inequalities like (2.1) with $0 < \epsilon < 1/2$ for the ERM over $\mathrm{conv}(F)$:

$$\tilde{f}_n^{ERM-C} \in \mathrm{Arg} \min_{f \in \mathrm{conv}(F)} R_n(f). \tag{2.3}$$

To make the argument simple, we consider the bounded regression framework with respect to the square loss: $|Y|, \sup_{f \in F} |f(X)| \leq 1$ a.s. and $\ell_f(x,y) = (y - f(x))^2, \forall f \in F, \forall (x,y) \in \mathcal{X} \times \mathbb{R}$.

Theorem E. There exists an absolute constant c_0 such that the following holds. For any $0 < x < \log n$ and $0 < \epsilon < 1/2$, with probability greater than $1 - 8\exp(-x)$,

$$R(\tilde{f}_n^{ERM-C}) \leq (1 + 3\epsilon) \inf_{f \in \mathrm{conv}(F)} R(f) + \frac{c_0 (\log M)(\log n)}{n\epsilon}.$$

Sketch of the proof of Theorem E. The proof of Theorem E and Theorem C are closely related. In the case of Theorem C, the result follows from the analysis of the loss functions classes indexed for the family of models $(rB_1^d)_{r \geq 0}$. In the case of Theorem E, the result follows from the analysis of the loss function class indexed by the model $\Lambda = \{\lambda \in \mathbb{R}^M : \lambda_j \geq 0, \|\lambda\|_{\ell_1} = 1\}$ (we identify every function $f_\lambda \in \mathrm{conv}(F)$ with its parameter $\lambda \in \Lambda$) which is included in B_1^M. Therefore, the proof of Theorem E follows the same path as the proof of Theorem C but since we assume boundedness some extra logarithms appearing in Theorem C can be saved here. Indeed, for the loss functions class $\ell_{\mathrm{conv}(F)} = \{\ell_f : f \in \mathrm{conv}(F)\}$, we can apply Theorem A with $b_n(\ell_{\mathrm{conv}(F)}) = 1$, $B_n = 1$ and the isomorphic function $\rho_n(x) \equiv c_0 (\log M)(\log n)/n$. ■

Note that the residual term of the non-exact oracle inequality satisfied by $\tilde{f}_n^{ERM-C}$ in Theorem E is of the order of $(\log M)(\log n)/n$ and is thus uniformly better than the optimal rate $\psi_n^C(M)$ for exact oracle inequalities in this setup. Up to logarithms, this residual term can even be the *square* of $\psi_n^C(M)$ when $M \geq \sqrt{n}$.

In this example, the gap between the rates obtained in the non-exact and exact cases is not due to a difference in the Bernstein condition, since it holds in both: for any $f \in \mathrm{conv}(F)$,

$$\mathbb{E}\ell_f^2 \leq \mathbb{E}\ell_f \text{ and } \mathbb{E}\mathcal{L}_f^2 \leq B\mathbb{E}\mathcal{L}_f,$$

where the latter follows from the convexity of $\mathrm{conv}(F)$. Therefore, to explain this gap one has to look elsewhere, and it appears that the reason is complexity. Indeed, it follows from Proposition 2.5 that (up to some logarithmic factors), for any $\lambda > 0$,

$$\mathbb{E}\left\|P - P_n\right\|_{V(\ell_{\mathrm{conv}(F)})_\lambda} \lesssim \sqrt{\frac{\lambda}{n}}.$$

Therefore, the fixed point λ_ϵ^* defined in Theorem A and associated with the loss functions class $\ell_{\mathrm{conv}(F)}$ will be of the order of $1/n$ (up to some logarithmic factors). On the other hand, if $g_1, ..., g_M$ are independent, standard Gaussian random variables, $\mathrm{conv}\big(\{f(X) : f \in F\}\big) = \{\sum_{j=1}^M \lambda_j g_j : \lambda \in B_1^M\}$ and $Y = g_{M+1}$, one can show that for any $\mu \geq 1/M$,

$$\mathbb{E}\left\|P - P_n\right\|_{V(\mathcal{L}_{\mathrm{conv}(F)})_\mu} \gtrsim \frac{1}{\sqrt{n}}.$$

Hence, the fixed point μ^* of Theorem 6.1 associated with the excess loss class $\mathcal{L}_{\mathrm{conv}(F)}$ is of the order of $1/\sqrt{n}$ when $M \geq \sqrt{n}$.

To conclude, in Convex aggregation, the complexity of the localized classes $V(\ell_{\mathrm{conv}(F)})_\lambda$ and $V(\mathcal{L}_{\mathrm{conv}(F)})_\mu$ for all $\lambda > 0$ plays a key role in understanding the difference between exact and non-exact oracle inequalities. The results of [17] show that this is a generic situation, and that one should expect a gap between the two inequalities even if the loss and excess loss classes satisfy a Bernstein condition.

3. Model Selection and regularization.

In this section, we obtain results on Model Selection by applying Theorem B.

The first step is to show that any RERM procedure is a Model Selection procedure for a particular class of models $\mathcal{M}$ and some penalty function. Then, one may apply Theorem B and derived non-exact oracle inequalities for the *penalized estimators* (see [16] for the terminology of this section) associated with the class $\mathcal{M}$ and the penalty.

For the sake of completeness, we will show that the converse is also true, and any penalized estimator is a RERM procedure for some particular class $\mathcal{F}$ and regularizing function.

Recall the setup of Model Selection from [16]. One is given a collection of models, denoted by $\mathcal{M}$, and a penalty function pen : $\mathcal{M} \to \mathbb{R}^+$. For every model $m \in \mathcal{M}$, an ERM procedure is constructed:

$$\widehat{f}_m \in \mathop{\mathrm{Arg\,min}}_{f \in m} R_n(f). \tag{3.1}$$

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

$$\widehat{m} \in \operatorname{Arg}\min_{m\in\mathcal{M}}\Big(R_n(\widehat{f}_m) + \operatorname{pen}(m)\Big). \tag{3.2}$$

The penalized estimator studied in Model Selection is $\widehat{f}_{\widehat{m}}$, where, as before, we assume that the infimum in (3.1) and in (3.2) are achieved. The next result shows that the penalized estimator $\widehat{f}_{\widehat{m}}$ is an RERM.

Lemma 3.1. *Define a class $\mathcal{F}$ and a regularizing function by*

$$\mathcal{F} = \bigcup_{m\in\mathcal{M}} m \text{ and } \operatorname{reg}(f) = \inf_{\{m\in\mathcal{M}: f\in m\}} \operatorname{pen}(m). \tag{3.3}$$

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

$$\widehat{f}_{\widehat{m}} \in \operatorname{Arg}\min_{f\in\mathcal{F}}\Big(R_n(f) + \operatorname{reg}(f)\Big).$$

Proof. By definition of $\widehat{f}_{\widehat{m}}$, for any $m \in \mathcal{M}$ and $f \in m$,

$$R_n(\widehat{f}_{\widehat{m}}) + \operatorname{pen}(\widehat{m}) \le R_n(f) + \operatorname{pen}(m). \tag{3.4}$$

Therefore, given $f \in \mathcal{F}$, (3.4) is true for any $m \in \mathcal{M}$ such that $f \in m$. Taking the infimum over all $m \in \mathcal{M}$ for which $f \in m$ in the right hand side of (3.4), we obtain

$$R_n(\widehat{f}_{\widehat{m}}) + \operatorname{pen}(\widehat{m}) \le R_n(f) + \operatorname{reg}(f). \tag{3.5}$$

Since (3.5) holds for any $f \in \mathcal{F}$, thus the claim follows since $\widehat{f}_{\widehat{m}} \in \widehat{m}$ and thus $\operatorname{reg}(\widehat{f}_{\widehat{m}}) \le \operatorname{pen}(\widehat{m})$. □

It follows from Lemma 3.1 that any Model Selection procedure is an RERM procedure over the function class $\mathcal{F}$ and for the regularizing function defined in (3.3). The next lemma proves the converse.

Lemma 3.2. *Let $\mathcal{F}$ be a class of function and $\operatorname{reg} : \mathcal{F} \to \mathbb{R}^+$ be a regularizing function such that for any $f \in \mathcal{F}, \operatorname{reg}(f) < \infty$. Assume that there exists $\widehat{f}_n^{RERM} \in \mathcal{F}$ minimizing $f \longrightarrow R_n(f) + \operatorname{reg}(f)$ over $\mathcal{F}$. Denote by $\operatorname{reg}(\mathcal{F}) \subset \mathbb{R}^+$ the range of reg. For any $r \in \operatorname{reg}(\mathcal{F})$ define the model $m_r = \{f \in \mathcal{F} : \operatorname{reg}(f) \le r\}$. If*

$$\mathcal{M} = \{m_r : r \in \operatorname{reg}(\mathcal{F})\} \text{ and } \operatorname{pen} : m_r \in \mathcal{M} \longrightarrow r \in \mathbb{R}^+, \tag{3.6}$$

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

Proof. Let $\widehat{r} = \mathrm{reg}(\widehat{f}_n^{RERM})$, set $\widehat{m} = m_{\widehat{r}}$ and for any $m \in \mathcal{M}$, let $\widehat{f}_m$ be the output of an ERM performed in m. Thus, for any $f \in \widehat{m}$,

$$R_n(\widehat{f}_n^{RERM}) + \mathrm{reg}(\widehat{f}_n^{RERM}) \leq R_n(f) + \mathrm{reg}(f) \leq R_n(f) + \widehat{r}.$$

Therefore, $R_n(\widehat{f}_n^{RERM}) \leq \inf_{f \in \widehat{m}} R_n(f)$ and $\widehat{f}_n^{RERM}$ is an ERM over $\widehat{m}$, that is, $\widehat{f}_n^{RERM} = \widehat{f}_{\widehat{m}}$.

It remains to show that $\widehat{m} \in \mathrm{Arg\,min}_{m \in \mathcal{M}} \left(R_n(\widehat{f}_m) + \mathrm{pen}(m)\right)$, which is evident because

$$\begin{aligned}
R_n(\widehat{f}_{\widehat{m}}) + \mathrm{pen}(\widehat{m}) &= R_n(\widehat{f}_n^{RERM}) + \mathrm{reg}(\widehat{f}_n^{RERM}) = \min_{f \in \mathcal{F}} \left(R_n(f) + \mathrm{reg}(f)\right) \\
&= \min_{r \in \mathrm{reg}(\mathcal{F})} \min_{f \in \mathcal{F}: \mathrm{reg}(f) \leq r} \left(R_n(f) + \mathrm{reg}(f)\right) = \min_{r \in \mathrm{reg}(\mathcal{F})} \min_{f \in m_r} \left(R_n(f) + \mathrm{reg}(f)\right) \\
&\leq \min_{r \in \mathrm{reg}(\mathcal{F})} \min_{f \in m_r} \left(R_n(f) + r\right) = \min_{r \in \mathrm{reg}(\mathcal{F})} \Big(\min_{f \in m_r} R_n(f) + \mathrm{pen}(m_r)\Big) \\
&= \min_{m \in \mathcal{M}} \Big(\min_{f \in m} R_n(f) + \mathrm{pen}(m)\Big) = \min_{m \in \mathcal{M}} \Big(R_n(\widehat{f}_m) + \mathrm{pen}(m)\Big).
\end{aligned}$$

□

With this equivalence in mind, one can apply Theorem B to obtain results on RERM procedures, then construct a class $\mathcal{M}$ and a penalty function according to (3.6) and finally use Lemma 3.2 to obtain oracle inequalities for the penalized estimator $\widehat{f}_{\widehat{m}}$ constructed in this framework.

As an example of application, we will use the Model Selection problem studied in Chapter 8 of [16] for Vapnik-Chervonenkis models.

Consider the loss function $\ell_f(x,y) = \mathbb{1}_{f(x) \neq y}$ defined for any $(x,y) \in \mathcal{X} \times \{0,1\}$ and measurable function $f : \mathcal{X} \to \{0,1\}$. One is given a countable set $\mathcal{M}$ of countable models (that is, a countable set of measurable functions from $\mathcal{X}$ to $\{0,1\}$) and any $m \in \mathcal{M}$ has a finite VC dimension denoted by V_m. In this setup, Theorem B can be applied without any extra assumption. In particular, we will not require any Margin assumption or Bernstein condition. Let the penalty function be the one used in pg. 285 of [16], that is, for any $m \in \mathcal{M}$,

$$\mathrm{pen}(m) = 2\sqrt{\frac{2V_m(1 + \log(n/V_m))}{n}} + \sqrt{\frac{\log n}{2n}} \tag{3.7}$$

and recall that the penalized estimator $\widehat{f}_{\widehat{m}}$ satisfies the following risk bound ([16], pg. 285):

$$\mathbb{E}R(\widehat{f}_{\widehat{m}}) \leq \inf_{m \in \mathcal{M}} \Big(\inf_{f \in m} R(f) + \mathrm{pen}(m)\Big) + \sqrt{\frac{\pi}{2n}}. \tag{3.8}$$

To apply Theorem B, we consider the following regularization setup defined by

$$\mathcal{F} = \bigcup_{m\in\mathcal{M}} m \text{ and } \mathrm{crit}(f) = \min_{m\in\mathcal{M}}\left(V_m \wedge n : f \in m\right)$$

and to make things simpler assume that

$$\mathcal{M} = \{m_k : k \in \mathbb{N}\} \text{ such that } m_0 \subset m_1 \subset m_2 \subset \cdots . \tag{3.9}$$

Theorem B allows one to calibrate the regularization term using the isomorphic profile of the family of loss functions classes $(\ell_{F_r})_{r\in\mathbb{N}}$ where for any $r \in \mathbb{N}$ (note that crit takes its values in $\mathbb{N}$) $F_r = \{f \in \mathcal{F} : \mathrm{crit}(f) \leq r\}$.

It follows from assumption (3.9) that $F_r = m_{k(r)}$ where $k(r) = \max(k \in \mathbb{N} : V_{m_k} \wedge n \leq r)$. The isomorphic function associated with the family of models $(F_r)_{r\in\mathbb{N}}$ can be obtained in this context following the same strategy used to get Equation (1.3) in the main paper of this supplementary material: we obtain, for any $r \in \mathbb{N}$,

$$\lambda_\epsilon^*(r) = \frac{c_0(V_{m_{k(r)}} \wedge n)\log\left(en/(V_{m_{k(r)}} \wedge n)\right)}{\epsilon^2 n}. \tag{3.10}$$

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

$$\mathrm{reg}(f) = \frac{c_1 V_{m(f)}\log\left(en/V_{m(f)}\right)}{\epsilon^2 n} \tag{3.11}$$

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

$$R(\widehat{f}_n^{RERM}) + c_0\mathrm{reg}(\widehat{f}_n^{RERM}) \leq (1+3\epsilon)\inf_{f\in\mathcal{F}}\Big(R(f) + c_1\mathrm{reg}(f)\Big) + \frac{c_2(x+1)}{n}. \tag{3.12}$$

From this result, we can now derive a non-exact regularized oracle inequality for the penalized estimator associated with the class of models $\mathcal{M}'$ and the penalty function pen$'$ as defined in (3.6):

$$\mathcal{M}' = \{m_r : r \in \mathrm{reg}(\mathcal{F})\} \text{ and } \mathrm{pen}'(m_r) = r, \quad \forall r \in \mathrm{reg}(\mathcal{F}) \tag{3.13}$$

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

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

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

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

$$R(\widehat{f}_{\widehat{m}}) + c_2 \mathrm{pen}(\widehat{m}) \leq (1+3\epsilon) \min_{m \in \mathcal{M}} \Big(\inf_{f \in m} R(f) + c_3 \mathrm{pen}(m) \Big).$$

In particular, it is interesting to note that, up to a logarithmic factor, the penalty function in (3.7) is of the order of $\sqrt{V/n}$ whereas, in the same framework (up to the structural assumption (3.9)), the penalty function defined in Theorem F is of the order of V/n. This difference comes from the Bernstein conditions since for the loss functions class: $\mathbb{E}\ell_f^2 \leq B\mathbb{E}\ell_f, \forall f \in \mathcal{F}$ is trivially satisfied in the setup of Theorem F; whereas the Bernstein condition for the excess loss functions classes: $\mathbb{E}\mathcal{L}_{f,m}^2 \leq B\mathbb{E}\mathcal{L}_{f,m}, \forall f \in m, \forall m \in \mathcal{M}$ (where $\mathcal{L}_{f,m} = \ell_f - \ell_{f_m^*}$ and $f_m^* \in \mathrm{argmin}_{f \in m} R(f)$) is not true in general and is somehow "required" to obtain fast rates for exact oracle inequalities.

Finally, note that a direct approach based on the computation of isomorphic functions for the family of loss functions class $(\ell_m)_{m \in \mathcal{M}}$ would provide a way of constructing penalty functions and obtaining oracle inequalities for the penalized estimator associated with this penalty function. This approach do not require the structural assumption (3.9). Nevertheless, we did not prove such a result which would mainly follow the same line as the proof of Theorem B combined with the approach in [16] (cf. Section 3.6 in [15] for such a result). Somehow, we found more interesting to prove that Model Selection methods can be seen as regularized procedures and that Theorem B, which was originally designed for regularized estimators, can also be used to prove results for penalized estimators.

References.

[1] Francis R. Bach. Consistency of trace norm minimization. *J. Mach. Learn. Res.*, 9:1019–1048, 2008.

[2] Florentina Bunea, Yiyuan She, and Marten Wegkamp. Optimal selection of reduced rank estimators of high-dimensional matrices. *arXiv:1004.2995v2*, 2010.

[3] Emmanuel J. Candes and Yaniv Plan. Matrix completion with noise. Proceedings of IEEE, 2009.

[4] Emmanuel J. Candes and Yaniv Plan. Tight oracle bounds for low-rank matrix recovery from a minimal number of random measurements. arXiv:1001.0339.

[5] Emmanuel J. Candes and Benjamin Recht. Exact matrix completion via convex optimization. *Found. of Comput. Math.*, 9:717–772, 2008.

[6] Emmanuel J. Candes and Benjamin Recht. Exact matrix completion via convex optimization. *Found. Comput. Math.*, 9(6):717–772, 2009.

[7] Emmanuel J. Candes and Terence Tao. The power of convex relaxation: Near-optimal matrix completion. arXiv:0903.1476.

[8] Stéphane Gaïffas and Guillaume Lecué. Sharp oracle inequalities for high-dimensional matrix prediction. *To appear in IEEE transaction on information theory. arXiv:1008.4886.*

[9] Christophe Giraud. Low rank multivariate regression. arXiv:1009.5165.

[10] David Gross. Recovering low-rank matrices from few coefficients in any basis. *ArXiv 0910.1879*, 2009.

[11] Raghunandan Keshavan, Andrea H., Montanari, and Sewoong Oh. Matrix completion from noisy entries. *arXiv:0906.2027*, 2009.

[12] Vladimir Koltchinskii. Von neumann entropy penalization and low rank matrix estimation. arXiv:1009.2439.

[13] Vladimir Koltchinskii, Alexandre B. Tsybakov, and Karim Lounici. Nuclear norm penalization and optimal rates for noisy low rank matrix completion. arXiv:1011.6256.

[14] Guillaume Lecué. Empirical risk minimization is optimal for the convex aggregation problem. 2011.

[15] Guillaume Lecué. Interplay between concentration, complexity and geometry in learning theory with applications to high dimensional data analysis. Habilitation à diriger des recherches. 2011.

[16] 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.

[17] Shahar Mendelson. Oracle inequalities and the isomorphic method. Under preparation.

[18] Sahand Negahban and Martin J. Wainwright. Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. arXiv:1009.2118.

[19] Benjamin Recht. A simpler approach to matrix completion. arXiv:0910.0651.

[20] Angelika Rohde and Alexandre Tsybakov. Estimation of high-dimensional low-rank matrice. To appear in Ann. Statist.. arXiv:0912.5338.

[21] Alexandre Tsybakov. Optimal rate of aggregation. In *Computational Learning Theory and Kernel Machines (COLT-2003)*, volume 2777 of *Lecture Notes in Artificial Intelligence*, pages 303–313. Springer, Heidelberg, 2003.

E-mail: Guillaume.Lecue@univ-mlv.fr

E-mail: shahar@tx.technion.ac.il