Sharp oracle inequalities for high-dimensional matrix prediction

Stéphane Gaïffas[1,3] and Guillaume Lecué[2,3]

August 27, 2010

Abstract

We observe $(X_i, Y_i)_{i=1}^n$ where the Y_i's are real valued outputs and the X_i's are $m \times T$ matrices. We observe a new entry X and we want to predict the output Y associated with it. We focus on the high-dimensional setting, where $mT \gg n$. This includes the matrix completion problem with noise, as well as other problems. We consider linear prediction procedures based on different penalizations, involving a mixture of several norms: the nuclear norm, the Frobenius norm and the ℓ_1-norm. For these procedures, we prove sharp oracle inequalities, using a statistical learning theory point of view. A surprising fact in our results is that the rates of convergence do not depend on m and T directly. The analysis is conducted without the usually considered incoherency condition on the unknown matrix or restricted isometry condition on the sampling operator. Moreover, our results are the first to give for this problem an analysis of penalization (such nuclear norm penalization) as a regularization algorithm: our oracle inequalities prove that these procedures have a prediction accuracy close to the deterministic oracle one, given that the reguralization parameters are well-chosen.

Keywords. High dimensional matrix ; Matrix completion ; Oracle inequalities ; Schatten norms ; Nuclear norm ; Empirical risk minimization ; Empirical process theory ; Sparsity

1 Introduction

1.1 The model and some basic definitions

Let (X, Y) and $D_n = (X_i, Y_i)_{i=1}^n$ be $n + 1$ i.i.d random variables with values in $\mathcal{M}_{m,T} \times \mathbb{R}$, where $\mathcal{M}_{m,T}$ is the set of matrices with m rows and T columns with entries in $\mathbb{R}$. Based on the observations D_n, we have in mind to predict the real-valued ouput

[1]Université Pierre et Marie Curie - Paris 6, Laboratoire de Statistique Théorique et Appliquée. *email*: stephane.gaiffas@upmc.fr

[2]CNRS, Laboratoire d'Analyse et Mathématiques appliquées, Université Paris-Est - Marne-la-vallée *email*: guillaume.lecue@univ-mlv.fr

[3]This work is supported by French Agence Nationale de la Recherce (ANR) ANR Grant "Prognostic" ANR-09-JCJC-0101-01. (http://www.lsta.upmc.fr/prognostic/index.php)

Y by a linear transform of the input variable X. We focus on the high-dimensional setting, where $mT \gg n$. We use a "statistical learning theory point of view": we do not assume that $\mathbb{E}(Y|X)$ has a particular structure, such as $\mathbb{E}(Y|X) = \langle X, A_0 \rangle$ for some $A_0 \in \mathcal{M}_{m,T}$, where $\langle \cdot, \cdot \rangle$ is the standard Euclidean inner product given for any $A, B \in \mathcal{M}_{m,T}$ by

$$\langle A, B \rangle := \mathrm{tr}(A^\top B). \tag{1}$$

The statistical performance of a linear predictor $\langle X, A \rangle$ for some $A \in \mathcal{M}_{m,T}$ is measured by the quadratic risk

$$R(A) := \mathbb{E}[(Y - \langle X, A \rangle)^2]. \tag{2}$$

If $\hat{A}_n \in \mathcal{M}_{m,T}$ is a statistic constructed from the observations D_n, then its risk is given by the conditional expectation

$$R(\hat{A}_n) := \mathbb{E}[(Y - \langle X, \hat{A}_n \rangle)^2 | D_n].$$

A natural candidate for the prediction of Y using D_n is the *empirical risk minimization procedure*, namely any element in $\mathcal{M}_{m,T}$ minimizing the empirical risk $R_n(\cdot)$ defined for all $A \in \mathcal{M}_{m,T}$ by

$$R_n(A) = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \langle X_i, A \rangle)^2.$$

It is well-known that the excess risk of this procedure is of order mT/n. In the high dimensional setting, this rate is not going to zero. So, if $X \mapsto \langle A_0, X \rangle$ is the best linear prediction of Y by X, we need to know more about A_0 in order to construct algorithms with a small risk. In particular, we need to know that A_0 has a "low-dimensional structure". In this setup, this is usually done by assuming that A_0 is low rank. A first idea is then to minimize R_n and to penalize matrices with a large rank. Namely, one can consider

$$\hat{A}_n \in \operatorname*{argmin}_{A \in \mathcal{M}_{m,T}} \big\{ R_n(A) + \lambda \, \mathrm{rank}(A) \big\}, \tag{3}$$

for some regularization parameter $\lambda > 0$. But $A \mapsto \mathrm{rank}(A)$ is far from being a convex function, thus minimizing (3) is very difficult in practice, see [19] for instance on this problem. Convex relaxation of (3) leads to the following convex minimization problem

$$\hat{A}_n \in \operatorname*{argmin}_{A \in \mathcal{M}_{m,T}} \big\{ R_n(A) + \lambda \|A\|_{S_1} \big\}, \tag{4}$$

where $\| \cdot \|_{S_1}$ is the 1-*Schatten* norm, also known as *nuclear norm* or *trace norm*. This comes from the fact that the nuclear norm is the convex envelope of the rank on the unit ball of the spectral norm, see [18]. For any matrix $A \in \mathcal{M}_{m,T}$, we denote

by $s_1(A), \dots, s_{\mathrm{rank}(A)}(A)$ its nonincreasing sequence of singular values. For every $p \in [1, \infty]$, the p-Schatten norm of A is given by

$$\|A\|_{S_p} := \Big(\sum_{j=1}^{\mathrm{rank}(A)} s_j(A)^p \Big)^{1/p}. \tag{5}$$

In particular, the $\|\cdot\|_{S_\infty}$-norm is the *operator norm* or *spectral norm*. The $\|\cdot\|_{S_2}$-norm is the *Frobenius norm*, which satisfies

$$\|A\|_{S_2}^2 = \sum_{i,j} A_{i,j}^2 = \langle A, A \rangle.$$

1.2 Motivations

A particular case of the matrix prediction problem described in Section 1.1 is the problem of *(noisy) matrix completion*, see [38, 39], which became very popular because of the buzz surrounding the Netflix prize[1]. In this problem, it is assumed that X is uniformly distributed over the set $\{e_{p,q} : 1 \le p \le m, 1 \le q \le T\}$, where $e_{p,q} \in \mathcal{M}_{m,T}$ is such that $(e_{p,q})_{i,j} = 0$ when $i \neq q$ or $j \neq p$ and $(e_{p,q})_{p,q} = 1$. If $\mathbb{E}(Y|X) = \langle A_0, X \rangle$ for some $A_0 \in \mathcal{M}_{m,T}$, then the Y_i are n noisy observations of the entries of A_0, and the aim is to denoise the observed entries and to fill the non-observed ones.

First motivation. Quite surprisingly, for matrix completion without noise ($Y_i = \langle X_i, A_0 \rangle$), it is proved in [15] and [16] (see also [21], [32]) that nuclear norm minimization is able, with a large probability (of order $1 - m^{-3}$) to recover *exactly* A_0 when $n > cr(m+T)(\log n)^6$, where r is the rank of A_0. This result is proved under a so-called *incoherency* assumption on A_0. This assumption requires, roughly, that the left and right singular vectors of A_0 are well-spread on the unit sphere. Using this incoherency assumption [14], [23] give results concerning the problem of matrix completion with noise. However, recalling that this assumption was introduced in order to prove *exact* completion, and since in the noisy case it is obvious that exact completion is impossible, a natural goal is then to obtain results for noisy matrix completion without the incoherency assumption. This is a first motivation of this work: we derive very general sharp oracle inequalities without any assumption on A_0, not even that it is low-rank. More than that, we don't need to assume that $\mathbb{E}(Y|X) = \langle X, A_0 \rangle$ for some A_0, since we use a statistical learning point-of-view in the statement of our results. More precisely, we construct procedures $\hat{A}_n$ satisfying *sharp oracle inequalities* of the form

$$R(\hat{A}_n) \le \inf_{A \in \mathcal{M}_{m,T}} \big\{ R(A) + r_n(A) \big\} \tag{6}$$

that hold with a large probability, where $r_n(A)$ is a *residue* related to the penalty used in the definition of $\hat{A}_n$ that we want as small as possible. By "sharp" we mean that in the right hand side of (6), the constant in front of $R(A)$ is equal to one.

[1] `http://www.netflixprize.com/`

A surprising fact in our results is that, for penalization procedures that involve the 1-Schatten norm (and 2-Schatten norm if a mixed penalization is considered), the residue $r_n(\cdot)$ does not depend on m and T directly: it only depends on the 1-Schatten norm of A_0, see Section 2 for details. This was not, as far as we know, previously noticed in literature (all the upper bounds obtained for $\|\hat{A}_n - A_0\|_{S_2}^2$ depend directly on m and T and on $\|A_0\|_{S_1}$ or on its rank and on $\|A_0\|_{S_\infty}$, see the references above and below). This fact can be used to argue that $\|\cdot\|_{S_1}$ is a better measure of sparsity than the rank, and it points out an interesting difference between nuclear-norm penalization (also called "Matrix Lasso") and the Lasso for vectors.

In [34], which is a work close to ours, upper bounds for p-Schatten penalization procedures for $0 < p \leq 1$ are given in the same setting as ours, including in particular the matrix completion problem. The results are stated without the incoherency assumption for matrix completion. But for this problem, the upper bounds are given using the empirical norm $\|\hat{A}_n - A_0\|_n^2 = \sum_{i=1}^n \langle X_i, \hat{A}_n - A_0\rangle^2/n$ only. An upper bound for this measure of accuracy gives information only about the denoising part and not about the filling part of the matrix completion problem. Our results have the form (6), and taking A_0 instead of the minimum in this equation gives an upper bound for $R(\hat{A}_n) - R(A_0)$, which is equal to $\|\hat{A}_n - A_0\|_{S_2}^2/(mT)$ in the matrix completion problem when $\mathbb{E}(Y|X) = \langle X, A_0\rangle$ (see Section 2).

Second motivation. In the setting considered here, an assumption called *Restricted Isometry* (RI) on the sampling operator $\mathcal{L}(A) = (\langle X_1, A\rangle, \ldots, \langle X_n, A\rangle)/\sqrt{n}$ has been introduced in [33] and used in a series of papers, see [34], [13], [29, 30]. This assumption is the matrix version of the restricted isometry assumption for vectors introduced in [12]. Note that in the high-dimensional setting ($mT \gg n$), this assumption is not satisfied in the matrix completion problem, see [34] for instance, which works with and without this assumption. The RI assumption is very restrictive and (up to now) is only satisfied by some special random matrices (cf. [36, 22, 28, 27] and references therein). This is a second motivation for this work: our results do not require any RI assumption. Our assumptions on X are very mild, see Section 2, and are satisfied in the matrix completion problem, as well as other problems, such as the multi-task learning.

Third motivation. Our results are the first to give an analysis of nuclear-norm penalization (and of other penalizations as well, see below) as a regularization algorithm. Indeed, an oracle inequality of the form (6) proves that these penalization procedures have a prediction accuracy close to the deterministic oracle one, given that the reguralization parameters are well-chosen.

Fourth motivation. We give oracle inequalities for penalization procedures involving a mixture of several norms: $\|\cdot\|_{S_1}$, $\|\cdot\|_{S_2}^2$ and the ℓ_1-norm $\|\cdot\|_1$. As far as we know, no result for penalization using several norms was previously given in literature for high-dimensional matrix prediction.

Procedures based on 1-Schatten norm penalization have been considered by many authors recently, with applications to multi-task learning and collaborative filtering. The first studies are probably the ones given in [38, 39], using the hinge loss for binary

classification. In [6], it is proved, under some condition on the X_i, that nuclear norm penalization can consistently recover $\mathrm{rank}(A_0)$ when $n \to +\infty$. Let us recall also the references we mentioned above and close other ones [18, 33], [13, 11, 15, 14, 16], [24, 23], [34], [21], [32, 33], [29, 30], [4, 3, 5], [1].

1.3 The procedures studied in this work

If $\mathbb{E}(Y|X) = \langle X, A_0 \rangle$ where A_0 is low rank, in the sense that $r \ll n$, nuclear norm penalization (4) is likely to enjoy some good prediction performances. But, if we know more about the properties of A_0, then other penalization procedure can be considered. For instance, if we know that the non-zero singular values of A_0 are "well-spread" (that is almost equal) then it may be interesting to use the "regularization effect" of a "S_2 norm" based penalty in the same spirit as "ridge type" penalty for vectors or functions. Moreover, if we know that many entries of A_0 are close or equal to zero, then using also a ℓ_1-penalization

$$A \mapsto \|A\|_1 = \sum_{\substack{1 \le p \le m \\ 1 \le q \le T}} |A_{p,q}| \tag{7}$$

may improve even further the prediction. In this paper, we consider a penalization that uses a mixture of several norms: for $\lambda_1, \lambda_2, \lambda_3 > 0$, we consider

$$\mathrm{pen}_{\lambda_1,\lambda_2,\lambda_3}(A) = \lambda_1 \|A\|_{S_1} + \lambda_2 \|A\|_{S_2}^2 + \lambda_3 \|A\|_1 \tag{8}$$

and we will study the prediction properties of

$$\hat{A}_n(\lambda_1, \lambda_2, \lambda_3) \in \operatorname*{argmin}_{A \in \mathcal{M}_{m,T}} \Big\{ R_n(A) + \mathrm{pen}_{\lambda_1,\lambda_2,\lambda_3}(A) \Big\}. \tag{9}$$

Of course, if more is known on the structure of A_0, other penalty functions can be considered.

We obtain sharp oracle inequalities for the procedure $\hat{A}_n(\lambda_1, \lambda_2, \lambda_3)$ for any values of $\lambda_1, \lambda_2, \lambda_3 \ge 0$ (excepted for $(\lambda_1, \lambda_2, \lambda_3) = (0, 0, 0)$ which provides the well-studied empirical risk minimization procedure). In particular, depending on the "a priori" knowledge that we have on A_0 we will consider different values for the triple $(\lambda_1, \lambda_2, \lambda_3)$. If A_0 is only known to be low-rank, one should choose $\lambda_1 > 0$ and $\lambda_2 = \lambda_3 = 0$. If A_0 is known to be low-rank with many zero entries, one should choose $\lambda_1, \lambda_3 > 0$ and $\lambda_2 = 0$. If A_0 is known to be low-rank with well-spread non-zero singular values, one should choose $\lambda_1, \lambda_2 > 0$ and $\lambda_3 = 0$. Finally, one should choose $\lambda_1, \lambda_2, \lambda_3 > 0$ when a significant part of the entries of A_0 are zero, that A_0 is low rank and that the non-zero singular values of A_0 are well-spread.

2 Results

We will use the following notation: for a matrix $A \in \mathcal{M}_{m,T}$, $\mathrm{vec}(A)$ denotes the vector of $\mathbb{R}^{mT}$ obtained by stacking its columns into a single vector. Note that this is an

isometry between $(\mathcal{M}_{m,T}, \|\cdot\|_{S_2})$ and $(\mathbb{R}^{mT}, |\cdot|_{\ell_2^{mT}})$ since $\langle A, B\rangle = \langle \operatorname{vec} A, \operatorname{vec} B\rangle$. We introduce also the ℓ_∞ norm $\|A\|_\infty = \max_{p,q} |A_{p,q}|$. Let us recall that for $\alpha \geq 1$, the ψ_α-norm of a random variable Z is given by $\|Z\|_{\psi_\alpha} := \inf\{c > 0 : \mathbb{E}[\exp(|Z|^\alpha/c^\alpha))] \leq 2\}$ and a similar norm can be defined for $0 < \alpha < 1$ (cf. [25]).

2.1 Assumptions and examples

The first assumption concers the "covariate" matrix X.

Assumption 1 (Matrix X)**.** *There are positive constants $b_{X,\infty}, b_{X,\ell_\infty}$ and $b_{X,2}$ such that $\|X\|_{S_\infty} \leq b_{X,\infty}$, $\|X\|_\infty \leq b_{X,\ell_\infty}$ and $\|X\|_{S_2} \leq b_{X,2}$ almost surely. Moreover, we assume that the "covariance matrix"*

$$\Sigma := \mathbb{E}[\operatorname{vec} X (\operatorname{vec} X)^\top]$$

is invertible.

This assumption is met in the matrix completion and the multitask-learning problems:

1. In the *matrix completion problem*, the matrix X is uniformly distributed over the set $\{e_{p,q} : 1 \leq p \leq m, 1 \leq q \leq T\}$ (see Section (1.2)), so in this case $\Sigma = (mT)^{-1} I_{m\times T}$ and $b_{X,2} = b_{X,\infty} = b_{X,\ell_\infty} = 1$.

2. In the *multitask-learning problem*, the matrix X is uniformly distributed in $\{A_j(x_{j,s}) : j = 1, \dots, T; s = 1, \dots, k_j\}$, where $(x_{j,s} : j = 1, \dots, T; s = 1, \dots, k_j)$ is a family of vectors in $\mathbb{R}^m$ and for any $j = 1, \dots, T$ and $x \in \mathbb{R}^m$, $A_j(x) \in \mathcal{M}_{m,T}$ is the matrix having the vector x for j-th column and zero everywhere else. So, in this case Σ is equal to T^{-1} times the $mT \times mT$ block matrix with T diagonal blocks of size $m \times m$ made of the T matrices $k_j^{-1} \sum_{i=1}^{k_j} x_{j,s} x_{j,s}^\top$ for $j = 1, \dots, T$.

   If we assume that the smallest singular values of the matrices $k_j^{-1} \sum_{i=1}^{k_j} x_{j,s} x_{j,s}^\top \in \mathcal{M}_{m,m}$ for $j = 1, \dots, T$ are larger than a constant $\sigma_{\min}$ (note that this implies that $k_j \geq m$), then Σ has its smallest singular value larger than $\sigma_{\min} T^{-1}$, so it is invertible. Moreover, if the vectors $x_{j,s}$ are normalized in ℓ_2, then one can take $b_{X,\infty} = b_{X,\ell_\infty} = b_{X,2} = 1$.

The next assumption deals with the regression function of Y given X. It is standard in regression analysis.

Assumption 2 (Noise)**.** *There are positive constants $b_Y, b_{Y,\infty}, b_{Y,\psi_2}, b_{Y,2}$ such that $\|Y - \mathbb{E}(Y|X)\|_{\psi_2} \leq b_{Y,\psi_2}$, $\|\mathbb{E}(Y|X)\|_{L_\infty} \leq b_{Y,\infty}$, $\mathbb{E}[(Y - \mathbb{E}(Y|X))^2|X] \leq b_{Y,2}^2$ almost surely and $\mathbb{E}Y^2 \leq b_Y^2$.*

In particular, any model $Y = \langle A_0, X\rangle + \varepsilon$, where $\|A_0\|_{S_\infty} < +\infty$ and ε is a sub-gaussian noise satisfies Assumption 2. Note that by using the whole strength of Talagrand's concentration inequality on product spaces for ψ_α $(0 < \alpha \leq 1)$ random variables obtained in [2], other type of tail decay of the noise could be considered (yet leading to slower decay of the residual term) depending on this assumption.

2.2 Main results

In this section we state our main results. We give sharp oracle inequalities for the penalized empirical risk minimization procedure

$$\hat{A}_n \in \operatorname*{argmin}_{A \in \mathcal{M}_{m,T}} \Big\{ \frac{1}{n} \sum_{i=1}^{n} (Y_i - \langle X_i, A \rangle)^2 + \mathrm{pen}(A) \Big\}, \tag{10}$$

where $\mathrm{pen}(A)$ is a penalty function which will be either a pure $\|\cdot\|_{S_1}$ penalization, or a "matrix elastic-net" penalization $\|\cdot\|_{S_1} + \|\cdot\|_{S_2}^2$ or other penalty functions involving the $\|\cdot\|_1$ norm.

Theorem 1 (Pure $\|\cdot\|_{S_1}$ penalization)**.** *There is an absolute constants $c > 0$ such that the following holds. Let Assumptions 1 and 2 hold, and let $x > 0$ be the some fixed confidence level. Consider any*

$$\hat{A}_n \in \operatorname*{argmin}_{A \in \mathcal{M}_{m,T}} \big\{ R_n(A) + \lambda_{n,x} \|A\|_{S_1} \big\},$$

for

$$\lambda_{n,x} = c_{X,Y} \frac{(x + \log n) \log n}{\sqrt{n}},$$

where $c_{X,Y} := c(1 + b_{X,2}^2 + b_Y b_X + b_{Y,\psi_1}^2 + b_{Y,\infty}^2 + b_{Y,2}^2 + b_{X,\infty}^2)$. Then one has, with a probability larger than $1 - 5e^{-x}$, that

$$R(\hat{A}_n) \leq \inf_{A \in \mathcal{M}_{m,T}} \big\{ R(A) + \lambda_{n,x}(1 + \|A\|_{S_1}) \big\}.$$

Note that the residue that we obtain is of the form $\|A_0\|_{S_1}/\sqrt{n}$. In particular, this residual term is not deteriorated if A_0 is of full rank but close to a low rank matrix. Classical residue involving the rank of A_0 are useless in this situation. It is also still meaningful when the quantity $m + T$ becomes large compare to n. This is not the case of the residue of the form $r(m + T)/n$ obtained previously for the same procedure (for other risks and under other - stronger - assumptions).

We now state three sharp oracle inequalities for procedures of the form (10) where the penalty function is a mixture of norms.

Theorem 2 (Matrix Elastic-Net)**.** *There is an absolute constant $c > 0$ such that the following holds. Let Assumptions 1 and 2 hold. Fix any $x > 0$, $r_1, r_2 > 0$, and consider*

$$\hat{A}_n \in \operatorname*{argmin}_{A \in \mathcal{M}_{m,T}} \Big\{ R_n(A) + \lambda_{n,x}(r_1 \|A\|_{S_1} + r_2 \|A\|_{S_2}^2) \Big\},$$

where

$$\lambda_{n,x} = c_{X,Y} \frac{\log n}{\sqrt{n}} \Big(\frac{1}{r_1} + \frac{(x + \log n) \log n}{r_2 \sqrt{n}} \Big),$$

where $c_{X,Y} = c(1+b_{X,2}^2+b_{X,2}b_Y+b_{Y,\psi_1}^2+b_{Y,\infty}^2+b_{Y,2}^2)$. *Then one has, with a probability larger than* $1-5e^{-x}$, *that*

$$R(\hat{A}_n) \le \inf_{A\in\mathcal{M}_{m,T}} \big\{R(A) + \lambda_{n,x}(1 + r_1\|A\|_{S_1} + r_2\|A\|_{S_2}^2)\big\}.$$

Theorem 3 ($\|\cdot\|_{S_1} + \|\cdot\|_1$ penalization)**.** *There is an absolute constant* $c > 0$ *such that the following holds. Let Assumptions 1 and 2 hold. Fix any* $x, r_1, r_3 > 0$, *and consider*

$$\hat{A}_n \in \operatorname*{argmin}_{A\in\mathcal{M}_{m,T}} \big\{R_n(A) + \lambda_{n,x}(r_1\|A\|_{S_1} + r_3\|A\|_1)\big\}$$

for

$$\lambda_{n,x} := c_{X,Y}\Big(\frac{1}{r_1} \wedge \frac{\sqrt{\log(mT)}}{r_3}\Big)\frac{(x+\log n)(\log n)^{3/2}}{\sqrt{n}},$$

where $c_{X,Y} = c(1 + b_{X,2}^2 + b_{X,2}b_Y + b_{Y,\psi_1}^2 + b_{Y,\infty}^2 + b_{Y,2}^2 + b_{X,\infty}^2 + b_{X,\ell_\infty}^2)$. *Then one has, with a probability larger than* $1-5e^{-x}$, *that*

$$R(\hat{A}_n) \le \inf_{A\in\mathcal{M}_{m,T}} \big\{R(A) + \lambda_{n,x}(1 + r_1\|A\|_{S_1} + r_3\|A\|_1))\big\}.$$

Theorem 4 ($\|\cdot\|_{S_1} + \|\cdot\|_{S_2}^2 + \|\cdot\|_1$ penalization)**.** *There is an absolute constant* $c > 0$ *such that the following holds. Let Assumptions 1 and 2 hold. Fix any* $x, r_1, r_2, r_3 > 0$, *and consider*

$$\hat{A}_n \in \operatorname*{argmin}_{A\in\mathcal{M}_{m,T}} \Big\{R_n(A) + \lambda_{n,x}(r_1\|A\|_{S_1} + r_2\|A\|_{S_2}^2 + r_3\|A\|_1)\Big\}$$

for

$$\lambda_{n,x} := c_{X,Y}\frac{(\log n)^{3/2}}{\sqrt{n}}\Big(\frac{1}{r_1} \wedge \frac{\sqrt{\log(mT)}}{r_3} + \frac{x+\log n}{r_2\sqrt{n}}\Big),$$

where $c_{X,Y} = c(1+b_{X,2}^2+b_{X,2}b_Y+b_{Y,\psi_1}^2+b_{Y,\infty}^2+b_{Y,2}^2)$. *Then one has, with a probability larger than* $1-5e^{-x}$, *that*

$$R(\hat{A}_n) \le \inf_{A\in\mathcal{M}_{m,T}} \big\{R(A) + \lambda_{n,x}(1 + r_1\|A\|_{S_1} + r_2\|A\|_{S_2}^2 + r_3\|A\|_1))\big\}.$$

The parameters r_1, r_2 and r_3 in the above procedures are completely free and can depend on n, m and T. Intuitively, it is clear that r_2 should be smaller than r_1 since the $\|\cdot\|_{S_2}$ term is used for "regularization" of the non-zero singular values only, while the term $\|\cdot\|_{S_1}$ makes $\hat{A}_n$ of low rank, as for the elastic-net for vectors (see [43]). Indeed, for the $\|\cdot\|_{S_1} + \|\cdot\|_{S_2}^2$ penalization, any choice of r_1 and r_2 such that $r_2 = r_1\log n/\sqrt{n}$ leads to a residual term smaller than

$$c_{X,Y}(1 + x + \log n)\Big(\frac{(\log n)^2}{r_2 n} + \frac{\log n}{\sqrt{n}}\|A\|_{S_1} + \frac{(\log n)^2}{n}\|A\|_{S_2}^2\Big).$$

Note that the rate related to $\|A\|_{S_1}$ is (up to logarithms) $1/\sqrt{n}$ while the rate related to $\|A\|_{S_2}^2$ is $1/n$. The choice of r_3 depends on the number of zeros in the matrix. Note that in the $\|\cdot\|_{S_1} + \|\cdot\|_1$ case, any choice $1 \leq r_3 \leq r_1$ entails a residue smaller than

$$c_{X,Y} \frac{(x + \log n) \log n}{\sqrt{n}} (1 + \|A\|_{S_1} + \|A\|_1),$$

which makes again the residue independent of m and T.

Note that, in the matrix completion case, the term $\sqrt{\log mT}$ can be removed from the regularization (and thus the residual) term thanks to the second statement of Proposition 1 below.

3 Proof of the main results

3.1 Some definitions

For any $r, r_1, r_2, r_3 \geq 0$, we consider the ball

$$B_{r,r_1,r_2,r_3} := \{A \in \mathcal{M}_{m,T} : r_1\|A\|_{S_1} + r_2\|A\|_{S_2}^2 + r_3\|A\|_1 \leq r\}, \tag{11}$$

and we denote by $B_{r,1} = B_{r,1,0,0}$ the *nuclear norm* ball, by $B_{r,r_1,r_2} = B_{r,r_1,r_2,0}$ the *elastic-net* ball. In what follows, B_r will be either $B_{r,1}$, B_{r,r_1,r_2}, B_{r,r_1,r_2,r_3} or $B_{r,r_1,0,r_3}$, depending on the penalization. We consider an oracle matrix in B_r given by:

$$A_r^* \in \operatorname*{argmin}_{A \in B_r} \mathbb{E}(Y - \langle X, A\rangle)^2$$

and the following excess loss function over B_r defined for any $A \in B_r$ by

$$\mathcal{L}_{r,A}(X, Y) := (Y - \langle X, A\rangle)^2 - (Y - \langle X, A_r^*\rangle)^2.$$

Define also the excess loss functions class

$$\mathcal{L}_r := \{\mathcal{L}_{r,A} : A \in B_r\}. \tag{12}$$

The star-shaped-hull at 0 of $\mathcal{L}_r$ is given by

$$V_r := \operatorname{star}(\mathcal{L}_r, 0) = \{\alpha\mathcal{L}_{r,A} : A \in B_r \text{ and } 0 \leq \alpha \leq 1\}$$

and its localized set at level $\lambda > 0$

$$V_{r,\lambda} := \{g \in V_r : \mathbb{E}g \leq \lambda\}. \tag{13}$$

The proof of Theorems 1 to 4 rely on the *isomorphic penalization method*, introduced by P. Bartlett, S. Mendelson and J. Neeman (cf. [8], [26] and [7]). It has improved several results on penalized empirical risk minimization procedures for the Lasso (cf. [7]) and for regularization in reproducing kernel Hilbert spaces (cf. [26]). This

approach relies on a sharp analysis of the complexity of the set $V_{r,\lambda}$. Indeed, an important quantity appearing in learning theory is the maximal deviation of the empirical distribution around its mean uniformly over a class of function. If V is a class of functions, we define the supremum of the deviation of the empirical mean around its actual mean over V by

$$\|P_n - P\|_V = \sup_{h \in V} \Big| \frac{1}{n} \sum_{i=1}^{n} h(X_i, Y_i) - \mathbb{E}h(X, Y) \Big|.$$

3.2 On the importance of convexity

An important parameter driving the quality of concentration of $\|P_n - P\|_V$ to its expectation is the so-called Bernstein's parameter (cf. [9]). We are studying this parameter in our context without introducing a formal definition of this quantity.

For every matrix $A \in \mathcal{M}_{m,T}$, we consider the random variable $f_A := \langle X, A \rangle$ and the following subset of L_2:

$$\mathcal{C}_r := \{f_A : A \in B_r\}, \tag{14}$$

where $B_r = B_{r,r_1,r_2,r_3}$ is given by (11). Because of the convexity of the norms $\|\cdot\|_{S_1}$, $\|\cdot\|_{S_2}$ and $\|\cdot\|_1$, the set $\mathcal{C}_r$ is convex, for any $r, r_1, r_2, r_3 \geq 0$. Now, consider the following minimum

$$f_r^* \in \operatorname*{argmin}_{f \in \mathcal{C}_r} \|Y - f\|_{L_2} \tag{15}$$

and

$$C_r := \min \Big(b_{X,\infty} \frac{r}{r_1}, b_{X,2} \sqrt{\frac{r}{r_2}}, b_{X,\ell_\infty} \frac{r}{r_3} \Big), \tag{16}$$

with the convention $1/0 = +\infty$.

Lemma 5 (Bernstein's parameter)**.** *Let assumptions 1 and 2 hold. There is a unique f_r^* satisfying (15) and a unique $A_r^* \in B_r$ such that $f_r^* = f_{A_r^*}$. Moreover, any $A \in B_r$ satisfies*

$$\mathbb{E}\mathcal{L}_{r,A} \geq \mathbb{E}\langle X, A - A_r^* \rangle^2,$$

and the class $\mathcal{L}_r$ satisfies the following Bernstein's condition: for all $A \in B_r$

$$\mathbb{E}\mathcal{L}_{r,A}^2 \leq 4(b_{Y,2}^2 + (b_{Y,\infty} + C_r)^2) \mathbb{E}\mathcal{L}_{r,A}.$$

Proof. By convexity of $\mathcal{C}_r$ we have $\langle Y - f_r^*, f - f_r^* \rangle_{L^2} \leq 0$ for any $f \in \mathcal{C}_r$. Thus, we have, for any $f \in \mathcal{C}_r$

$$\|Y - f\|_{L_2}^2 - \|Y - f_r^*\|_{L_2}^2 = 2\langle f_r^* - f, Y - f_r^* \rangle + \|f - f_r^*\|_{L_2}^2 \geq \|f - f_r^*\|_{L_2}^2. \tag{17}$$

In particular, the minimum is unique. Moreover, $\mathcal{C}_r$ is a closed set and since Σ is invertible under Assumption 1, there is a unique $A_r^* \in B_r$ such that $f_r^* = f_{A_r^*}$. By

the trace duality formula and Assumption 1, we have, for any $A \in B_{r,r_1,r_2,r_3}$:

$$|f_A| \le \|X\|_{S_\infty}\|A\|_{S_1} \le b_{X,\infty}\frac{r}{r_1}, \quad |f_A| \le \|X\|_{S_2}\|A\|_{S_2} \le b_{X,2}\sqrt{\frac{r}{r_2}},$$
$$\text{and } |f_A| \le \|X\|_\infty \|A\|_1 \le b_{X,\ell_\infty}\frac{r}{r_3}$$

almost surely, so that $|f_A| \le C_r$ for any $A \in B_r$ a.s.. Moreover, for any $A \in B_r$:

$$\mathcal{L}_{r,A} = 2(Y - \mathbb{E}(Y|X))\langle X, A_r^* - A\rangle + (2\mathbb{E}(Y|X) - \langle A + A_r^*, X\rangle)\langle X, A_r^* - A\rangle. \quad (18)$$

Thus, using Assumption 2, we obtain

$$\begin{aligned}\mathbb{E}\mathcal{L}_{r,A}^2 &= \mathbb{E}\big[4(Y - \mathbb{E}(Y|X))^2\langle X, A - A_r^*\rangle^2 + (2\mathbb{E}(Y|X) - \langle X, A + A_r^*\rangle)^2\langle X, A - A_r^*\rangle^2\big] \\ &\le 4\mathbb{E}\big[\langle X, A - A_r^*\rangle^2\mathbb{E}\big[(Y - \mathbb{E}(Y|X))^2|X\big]\big] + 4(b_{Y,\infty} + C_r)^2\mathbb{E}\langle X, A - A_r^*\rangle^2 \\ &\le 4(b_{Y,2}^2 + (b_{Y\infty} + C_r)^2)\mathbb{E}\langle X, A - A_r^*\rangle^2,\end{aligned}$$

which concludes the proof using (17). □

3.3 The isomorphic property of the excess loss functions class

The *isomorphic property* of a functions class has been introduced in [26] and is a consequence of Talagrand's concentration inequality (cf. [40]) applied to a localization of the functions class together with the Bernstein property of this class (here this property was studied in Lemma 5). We recall here the argument in our special case.

Theorem 6 ([10])**.** *There exists an absolute constant $c > 0$ such that the following holds. Let Assumptions 1 and 2 hold. Let $r > 0$ and $\lambda(r) > 0$ be such that*

$$\mathbb{E}\|P_n - P\|_{V_{r,\lambda(r)}} \le \frac{\lambda(r)}{8}.$$

Then, with probability larger than $1 - 4e^{-x}$: for all $A \in B_r$

$$\frac{1}{2}P_n\mathcal{L}_{r,A} - \rho_n(r,x) \le P\mathcal{L}_{r,A} \le 2P_n\mathcal{L}_{r,A} + \rho_n(r,x),$$

where

$$\rho_n(r,x) := c\Big(\lambda(r) + \big[b_{Y,\psi_1} + b_{Y,\infty} + b_{Y,2} + C_r\big]^2\Big(\frac{x\log n}{n}\Big)\Big),$$

and C_r has been introduced in (16).

Proof. We follow the line of [10]. Let $\lambda > 0$ and $x > 0$. Thanks to [2], with probability larger than $1 - 4\exp(-x)$,

$$\|P - P_n\|_{V_{r,\lambda}} \le 2\mathbb{E}\|P - P_n\|_{V_{r,\lambda}} + c_1\sigma(V_{r,\lambda})\sqrt{\frac{x}{n}} + c_2 b_n(V_{r,\lambda})\frac{x}{n} \quad (19)$$

where, by using the Bernstein's properties of $\mathcal{L}_r$ (cf. Lemma 5)

$$\begin{aligned}\sigma^2(V_{r,\lambda}) &:= \sup_{g\in V_{r,\lambda}} \mathrm{Var}(g) \leq \sup\Big(\mathbb{E}(\alpha\mathcal{L}_{r,A})^2 : 0\leq\alpha\leq 1, A\in B_r, \mathbb{E}(\alpha\mathcal{L}_{r,A})\leq\lambda\Big)\\ &\leq \sup\Big(4(b_{Y,2}^2+(b_{Y,\infty}+C_r)^2)\mathbb{E}(\alpha\mathcal{L}_{r,A}) : 0\leq\alpha\leq 1, A\in B_r, \mathbb{E}(\alpha\mathcal{L}_{r,A})\leq\lambda\Big)\\ &\leq 4(b_{Y,2}^2+(b_{Y,\infty}+C_r)^2)\lambda, \end{aligned} \tag{20}$$

and using Pisier's inequality (cf. [42]):

$$\begin{aligned} b_n(V_{r,\lambda}) &:= \Big\| \max_{1\leq i\leq n}\sup_{g\in V_{r,\lambda}} g(X_i,Y_i)\Big\|_{\psi_1} \leq \log n\Big\| \sup_{g\in V_{r,\lambda}} g(X,Y)\Big\|_{\psi_1}\\ &= \log n\Big\| \sup\Big(\alpha(2Y-\langle X, A+A_r^*\rangle)\langle X, A_r^*-A\rangle : 0\leq\alpha\leq 1, A\in B_r\Big)\Big\|_{\psi_1}\\ &\leq 4(\log n)(b_{Y,\psi_1}+b_{Y,\infty}+C_r)C_r, \end{aligned} \tag{21}$$

where we used decomposition (18) and Assumption 2 together with the uniform bound $|\langle A, X\rangle|\leq C_r$ holding for all $A\in B_r$.

Moreover, for any $\lambda>0$, $V_{r,\lambda}$ is star-shaped so $G:\lambda\mapsto\mathbb{E}\|P-P_n\|_{V_{r,\lambda}}/\lambda$ is non-increasing. Since $G(\lambda(r))\leq 1/8$ and $\rho_n(r,x)\geq\lambda(r)$, we have

$$\mathbb{E}\|P-P_n\|_{V_{r,\rho_n(r,x)}}\leq\rho_n(r,x)/8,$$

which yields, in Equation (19) together with the variance control of Equation (20) and the control of Equation (21), that there exists an event Ω_0 of probability measure greater than $1-4\exp(-x)$ such that, on Ω_0,

$$\begin{aligned}\|P-P_n\|_{V_{r,\rho_n(r,x)}} &\leq \frac{\rho_n(r,x)}{4} + c_1(b_{Y,\infty}+b_{Y,2}+C_r)\sqrt{\frac{\rho_n(r,x)x}{n}}\\ &\quad + c_2(b_{Y,\psi_1}+b_{Y,\infty}+C_r)C_r\frac{x\log n}{n}\\ &\leq \frac{\rho_n(r,x)}{2}\end{aligned} \tag{22}$$

in view of the definition of $\rho_n(r,x)$. In particular, on Ω_0, for every $A\in B_r$ such that $P\mathcal{L}_{r,A}\leq\rho_n(r,x)$, we have $|P\mathcal{L}_{r,A}-P_n\mathcal{L}_{r,A}|\leq\rho_n(r,x)/2$. Now, take $A\in B_r$ such that $P\mathcal{L}_{r,A}=\beta>\rho_n(r,x)$ and set $g=\rho_n(r,x)\mathcal{L}_{r,A}/\beta$. Since $g\in V_{r,\rho_n(r,x)}$, Equation (22) yields, on Ω_0, $|Pg-P_ng|\leq\rho_n(r,x)/2<\beta/2$ and so $(1/2)P_n\mathcal{L}_{r,A}\leq P\mathcal{L}_{r,A}\leq(3/2)P_n\mathcal{L}_{r,A}$ which concludes the proof. □

A function $r\mapsto\lambda(r)$ such that $\mathbb{E}\|P_n-P\|_{V_{r,\lambda(r)}}\leq\lambda(r)/8$ is called an *isomorphic function* and is directly connected to the choice of the penalization used in the procedure which was introduced in Section 2. The computation of this function is related to the complexity of Schatten balls, computed in the next section.

3.4 Complexity of Schatten balls

The *generic chaining* technique (see [41]) is a powerful technique for the control of the supremum of empirical processes. For a subgaussian process, such a control is achieved using the γ_2 functional recalled in the next definition.

Definition 7 ([41]). *Let (F, d) be a metric space. We say that $(F_j)_{j\geq 0}$ is an* admissible sequence *of partitions of F if $|F_0| = 1$ and $|F_j| \leq 2^{2^j}$ for all $j \geq 1$. The γ_2 functional is defined by*

$$\gamma_2(F, d) = \inf_{(F_j)_j} \sup_{f\in F} \sum_{j\geq 0} 2^{j/2} d(f, F_j),$$

where the infimum is taken over all admissible sequence $(F_j)_{j\geq 1}$ of F.

A classical upper bound on the γ_2 functional is the Dudley's entropy integral:

$$\gamma_2(F, d) \leq c_0 \int_0^\infty \sqrt{\log N(F, d, \epsilon)} d\epsilon, \tag{23}$$

where $N(B, \|\cdot\|, \varepsilon)$ is the minimal number of balls with respect to the metric d of radius ϵ needed to cover B. When B enjoys some convexity properties, this bound can be improved. Let $(E, \|\cdot\|)$ be a Banach space. We denote by $B(E)$ its unit ball. We say that $(E, \|\cdot\|)$ is 2-convex if there exists some $\rho > 0$ such that for all $x, y \in B(E)$, we have

$$\|x + y\| \leq 2 - 2\rho\|x - y\|^2.$$

In the case of 2-convex bodies, the following theorem gives an upper bound on the γ_2 functional that can improve the one given by Dudley's entropy integral.

Theorem 8 ([41]). *For any $\rho > 0$, there exists $c(\rho) > 0$ such that if $(E, \|\cdot\|)$ is a 2-convex Banach space and $\|\cdot\|_E$ is another norm on E, then*

$$\gamma_2(B(E), \|\cdot\|_E) \leq c(\rho)\Big(\int_0^\infty \epsilon \log N(B(E), \|\cdot\|_E, \epsilon) d\epsilon\Big)^{1/2}.$$

The generic chaining technique provides the following upper bound on Gaussian processes.

Theorem 9 ([41]). *There is an absolute constants $c > 0$ such that the following holds. If $(Z_f)_{f\in F}$ is a subgaussian process for some metric d (i.e. $\|Z_f - Z_g\|_{\psi_2} \leq c_0 d(f, g)$ for all $f, g \in F$) and if $f_0 \in F$, then one has*

$$\mathbb{E} \sup_{f\in F} |Z_f - Z_{f_0}| \leq c\gamma_2(F, d).$$

The metric used to measure the complexity of the excess loss classes we are working on is an empirical one defined for any $A \in \mathcal{M}_{m,T}$ by

$$\|A\|_{\infty,n} := \max_{1\leq i\leq n} |\langle X_i, A\rangle|. \tag{24}$$

This metric comes out of the so-called $L_{\infty,n}$-method of M. Rudelson introduced in [35] and first used in learning theory in [26]. We denote by $B(S_p)$ the unit ball of the Banach space S_p of matrices in $\mathcal{M}_{m,T}$ endowed with the Schatten norm $\|\cdot\|_{S_p}$. We denote also by B_1 the unit ball of $\mathcal{M}_{m,T}$ endowed with the ℓ_1-norm $\|\cdot\|_1$. In the following, we compute the complexity of the balls $B(S_1), B(S_2)$ and B_1 with respect to the empirical metric $\|\cdot\|_{\infty,n}$.

Proposition 1. *There exists an absolute constant $c > 0$ such that the following holds. Assume that $\|X_i\|_{S_2}, \|X_i\|_\infty \leq 1$ for all $i = 1, \ldots, n$. Then, we have*

$$\gamma_2(rB(S_1), \|\cdot\|_{\infty,n}) \leq \gamma_2(rB(S_2), \|\cdot\|_{\infty,n}) \leq cr\log n$$

and

$$\gamma_2(rB_1, \|\cdot\|_{\infty,n}) \leq cr(\log n)^{3/2}\sqrt{\log(mT)}.$$

Moreover, if we assume that $X_1, \ldots, X_n$ have been obtained in the matrix completion model then

$$\gamma_2(rB_1, \|\cdot\|_{\infty,n}) \leq cr(\log n)^{3/2}.$$

Proof. The first inequality is obvious since $B(S_1) \subset B(S_2)$. By using Dual Sudakov's inequality (cf. [31]), we have for all $\epsilon > 0$,

$$\log N(B(S_2), \|\cdot\|_{\infty,n}, \epsilon) \leq c_0\Big(\frac{\mathbb{E}\|G\|_{\infty,n}}{\epsilon}\Big)^2,$$

where G is a $m\times T$ matrix with i.i.d. standard Gaussian random variables for entries. A Gaussian maximal inequality and the fact that $\|X_i\|_{S_2} \leq 1$ for all $i = 1, \ldots, n$ provides $\mathbb{E}\|G\|_{\infty,n} \leq c_1\sqrt{\log n}$, hence

$$\log N(B(S_2), \|\cdot\|_{\infty,n}, \epsilon) \leq \frac{c_2\log n}{\epsilon^2}.$$

Denote by $B_{\infty,n}$ the unit ball of $(\mathcal{M}_{m,T}, \|\cdot\|_{\infty,n})$ in $V_n = \mathrm{span}(X_1, \ldots, X_n)$, the linear subspace of $\mathcal{M}_{m,T}$ spanned by $X_1, \ldots, X_n$. The volumetric argument provides

$$\begin{aligned}\log N(B(S_2), \|\cdot\|_{\infty,n}, \epsilon) &\leq \log N(B(S_2), \|\cdot\|_{\infty,n}, \eta) + \log N(\eta B_{\infty,n}, \epsilon B_{\infty,n})\\ &\leq \frac{c_2\log n}{\eta^2} + n\log\Big(\frac{3\eta}{\epsilon}\Big)\end{aligned}$$

for any $\eta \geq \epsilon > 0$. Thus, for $\eta_n = \sqrt{\log n/n}$, we have, for all $0 < \epsilon \leq \eta_n$

$$\log N(B(S_2), \|\cdot\|_{\infty,n}, \epsilon) \leq c_3 n\log\Big(\frac{3\eta_n}{\epsilon}\Big).$$

Since $B(S_2)$ is the unit ball of a Hilbert space, it is 2-convex. We can thus apply Theorem 8 to obtain the following upper bound

$$\gamma_2(rB(S_2), \|\cdot\|_{\infty,n}) \leq c_4 r \log n.$$

Now, we prove an upper bound on the complexity of B_1 with respect to $\|\cdot\|_{\infty,n}$. Recall that vec : $\mathcal{M}_{m,T} \to \mathbb{R}^{mT}$ concatenates the columns of a matrix into a single vector of size mT. Obviously, vec is an isometry between $(\mathcal{M}_{m,T}, \|\cdot\|_{S_2})$ and $(\mathbb{R}^{mT}, |\cdot|_2)$, since $\langle A, B\rangle = \langle \mathrm{vec}(A), \mathrm{vec}(B)\rangle$. Using this mapping, we see that, for any $\epsilon > 0$,

$$N(B_1, \|\cdot\|_{\infty,n}, \epsilon) = N(b_1^{mT}, |\cdot|_{\infty,n}, \epsilon)$$

where b_1^{mT} is the unit ball of ℓ_1^{mT} and $|\cdot|_{\infty,n}$ is the pseudo norm on $\mathbb{R}^{mT}$ defined for any $x \in \mathbb{R}^{mT}$ by $|x|_{\infty,n} = \max_{1\leq i\leq n} |\langle y_i, x\rangle|$ where $y_i = \mathrm{vec}(X_i)$ for $i = 1, \ldots, n$. Note that $y_1, \ldots, y_n \in b_2^{mT}$, where b_2^{mT} is the unit ball of ℓ_2^{mT}. We use the Carl-Maurey's empirical method to compute the covering number $N(b_1^{mT}, |\cdot|_{\infty,n}, \epsilon)$ for "large scales" of ϵ and the volumetric argument for "small scales". Let us begin with the Carl-Maurey's argument. Let $x \in b_1^{mT}$ and Z be a random variable with values in $\{\pm e_1, \ldots, \pm e_{mT}, 0\}$ - where $(e_1, \ldots, e_{mT})$ is the canonical basis of $\mathbb{R}^{mT}$ - defined by $\mathbb{P}[Z = 0] = 1 - |x|_1$ and for all $i = 1, \ldots, mT$,

$$\mathbb{P}[Z = \mathrm{sign}(x_i) e_i] = |x_i|.$$

Note that $\mathbb{E}Z = x$. Let $s \in \mathbb{N} - \{0\}$ to be defined later and take s i.i.d. copies of Z denoted by $Z_1, \ldots, Z_s$. By the Giné-Zinn symmetrization argument and the fact that Rademacher processes are upper bounded by Gaussian processes, we have

$$\mathbb{E}\Big|\frac{1}{s}\sum_{i=1}^{s} Z_i - \mathbb{E}Z\Big|_{\infty,n} \leq c_0 \mathbb{E}\Big|\frac{1}{s}\sum_{i=1}^{s} g_i Z_i\Big|_{\infty,n} \leq c_1\sqrt{\frac{\log n}{s}} \tag{25}$$

where the last inequality follows by a Gaussian maximal inequality and the fact that $|y_i|_2 \leq 1$. Take $s \in \mathbb{N}$ to be the smallest integer such that $\epsilon \geq c_1\sqrt{(\log n)/s}$. Then, the set

$$\Big\{\frac{1}{s}\sum_{i=1}^{s} z_i : z_1, \ldots, z_s \in \{\pm e_1, \ldots, \pm e_{mT}, 0\}\Big\} \tag{26}$$

is an ϵ-net of b_1^{mT} with respect to $|\cdot|_{\infty,n}$. Indeed, thanks to (25) there exists $\omega \in \Omega$ such that $|s^{-1}\sum_{i=1}^{s} Z_i(\omega) - x|_{\infty,n} \leq \epsilon$. This implies that there exists an element in the set (26) which is ϵ-close to x. Since the cardinality of the set introduced in (26) is, according to [17], at most

$$\binom{2mT + s - 1}{s} \leq \Big(\frac{e(2mT + s - 1)}{s}\Big)^s,$$

we obtain for any $\epsilon \geq \eta_n := \big((\log n)(\log mT)/n\big)^{1/2}$ that

$$\log N(b_1^{mT}, |\cdot|_{\infty,n}, \epsilon) \leq s \log\Big(\frac{e(2mT + s - 1)}{s}\Big) \leq \frac{c_2(\log n)\log(mT)}{\epsilon^2},$$

and a volumetric argument gives

$$\log N(b_1^{mT}, |\cdot|_{\infty,n}, \epsilon) \leq c_3 n \log\left(\frac{3\eta_n}{\epsilon}\right)$$

for any $0 < \epsilon \leq \eta_n$. Now we use the upper bound (23) and compute the Dudley's entropy integral to obtain

$$\gamma_2(rB_1, \|\cdot\|_{\infty,n}) \leq c_4 r (\log n)^{3/2} \sqrt{\log(mT)}.$$

For the "matrix completion case", we have

$$N(b_1^{mT}, |\cdot|_{\infty,n}, \epsilon) \leq N(b_1^n, \epsilon b_\infty^n)$$

where $N(b_1^n, \epsilon b_\infty^n)$ is the minimal number of balls ϵb_∞^n needed to cover b_1^n. We use the following proposition from [37] to compute $N(b_1^n, \epsilon b_\infty^n)$.

Proposition 2 ([37]). *For any $\epsilon > 0$, we have*

$$\log N(b_1^n, \epsilon b_\infty^n) \sim \begin{cases} 0 & \text{if } \epsilon \geq 1 \\ \epsilon^{-1} \log\left(en\epsilon\right) & \text{if } n^{-1} \leq \epsilon \leq 1 \\ n \log\left(1/(\epsilon n)\right) & \text{if } 0 < \epsilon \leq n^{-1}. \end{cases}$$

Then the result follows from (23) and the computation of the Dudley's entropy integral using Proposition 2. □

3.5 Computation of the isomorphic function

Introduce the ellipsoid

$$D := \{A \in \mathcal{M}_{m,T} : \mathbb{E}\langle X, A\rangle^2 \leq 1\}.$$

A consequence of Equation (17) in Lemma 5 is the following inclusion, of importance in what follows. Indeed, since B_r is convex and symmetrical, one has:

$$\{A \in B_r : \mathbb{E}\mathcal{L}_{r,A} \leq \lambda\} \subset A_r^* + K_{r,\lambda}, \tag{27}$$

where

$$K_{r,\lambda} := 2B_r \cap \sqrt{\lambda} D.$$

Hence, the complexity of $\{A \in \mathcal{M}_{m,T} : \mathcal{L}_{r,A} \in \mathcal{L}_{r,\lambda}\}$ will be smaller than the complexity of B_r and $\sqrt{\lambda}D$. This will be of importance in the analysis below. The next result provides an upper bound on the complexity of $V_{r,\lambda}$ where we recall that

$$V_{r,\lambda} := \{\alpha \mathcal{L}_{r,A} : 0 \leq \alpha \leq 1, A \in B_r, \mathbb{E}(\alpha \mathcal{L}_{r,A}) \leq \lambda\}.$$

From this statement we will derive corollaries that provide the shape of the considered penalty functions.

Proposition 3. *There exists two absolute constants c_1 and c_2 such that the following holds. Let Assumptions 1 and 2 hold. For any $r > 0$ and $\lambda > 0$, we have*

$$\mathbb{E}\|P - P_n\|_{V_{r,\lambda}} \leq c_1 \sum_{i \geq 0} 2^{-i} \phi_n(r, 2^{i+1}\lambda),$$

where

$$\phi_n(r,\lambda) := c_2\Big(U_n(K_{r,\lambda})\sqrt{\frac{\lambda}{n}} + U_n(K_{r,\lambda})\sqrt{\frac{R(A_r^*)}{n}} + \frac{U_n(K_{r,\lambda})^2}{n}\Big),$$

for $K_{r,\lambda} = 2B_r \cap \sqrt{\lambda}D$.

Proof. Introduce $\mathcal{L}_{r,\lambda} = \{\mathcal{L}_{r,A} : A \in B_r, \mathbb{E}\mathcal{L}_{r,A} \leq \lambda\}$. Using the Giné-Zinn symmetrization [20] and the inclusion of (27), one has, for any $r > 0$ and $\lambda > 0$,

$$\mathbb{E}\|P - P_n\|_{\mathcal{L}_{r,\lambda}} \leq \mathbb{E}\mathbb{E}_\epsilon \frac{2}{n} \sup_{A \in A_r^* + K_{r,\lambda}} \Big|\sum_{i=1}^n \epsilon_i \mathcal{L}_{r,A}(X_i, Y_i)\Big|,$$

where $\epsilon_1, \ldots, \epsilon_n$ are n i.i.d Rademacher variables. Introduce the Rademacher process $Z_A := \sum_{i=1}^n \epsilon_i \mathcal{L}_{r,A}(X_i, Y_i)$, and note that for any $A, A' \in A_r^* + K_{r,\lambda}$:

$$\begin{aligned}
\mathbb{E}_\epsilon |Z_A - Z_{A'}|^2 &= \sum_{i=1}^n \langle X_i, A - A'\rangle^2 (2Y_i - \langle X_i, A + A'\rangle)^2 \\
&= 4\sum_{i=1}^n \langle X_i, A - A'\rangle^2 (Y_i - \langle X_i, A_r^*\rangle - \langle X_i, \frac{A + A'}{2} - A_r^*\rangle)^2 \\
&\leq 8\|A - A'\|_{n,\infty}^2 \Big(\sum_{i=1}^n (Y_i - \langle X_i, A_r^*\rangle)^2 + \sup_{A \in K_{r,\lambda}} \sum_{i=1}^n \langle X_i, A\rangle^2\Big),
\end{aligned}$$

where we recall that $\|A\|_{n,\infty} = \max_{i=1,\ldots,n} |\langle X_i, A\rangle|$. So, using the generic chaining mechanism (cf. Theorem (9)), we obtain

$$\begin{aligned}
\mathbb{E}\|P - P_n\|_{\mathcal{L}_{r,\lambda}} &\leq \frac{c}{n}\mathbb{E}\Big[\gamma_2(K_{r,\lambda}, \|\cdot\|_{n,\infty})\Big(\sum_{i=1}^n (Y_i - \langle X_i, A_r^*\rangle)^2 + \sup_{A \in K_{r,\lambda}} \sum_{i=1}^n \langle X_i, A\rangle^2\Big)^{1/2}\Big] \\
&\leq \frac{c}{\sqrt{n}} (\mathbb{E}\gamma_2(K_{r,\lambda}, \|\cdot\|_{n,\infty})^2)^{1/2} \Big(R(A_r^*) + \mathbb{E}\sup_{A \in K_{r,\lambda}} \frac{1}{n}\sum_{i=1}^n \langle X_i, A\rangle^2\Big)^{1/2}.
\end{aligned}$$

Now, introduce, for some set $K \subset \mathcal{M}_{m,T}$ the functional

$$U_n(K) := (\mathbb{E}\gamma_2(K, \|\cdot\|_{n,\infty})^2)^{1/2}.$$

Using Theorem 1.2 from [22], we obtain:

$$\mathbb{E}\sup_{A \in K_{r,\lambda}} \frac{1}{n}\sum_{i=1}^n \langle X_i, A\rangle^2 \leq \lambda + c\max\Big(\sqrt{\frac{\lambda}{n}} U_n(K_{r,\lambda}), \frac{U_n(K_{r,\lambda})}{n}^2\Big),$$

and so, we arrive at

$$\mathbb{E}\|P - P_n\|_{\mathcal{L}_{r,\lambda}} \le c\phi_n(r,\lambda),$$

where

$$\begin{aligned}\phi_n(r,\lambda) &:= c\frac{U_n(K_{r,\lambda})}{\sqrt{n}}\Big(\lambda + R(A_r^*) + \frac{\sqrt{\lambda}U_n(K_{r,\lambda})}{\sqrt{n}} + \frac{U_n(K_{r,\lambda})^2}{n}\Big)^{1/2} \\ &\le c\Big(U_n(K_{r,\lambda})\sqrt{\frac{\lambda}{n}} + U_n(K_{r,\lambda})\sqrt{\frac{R(A_r^*)}{n}} + \frac{U_n(K_{r,\lambda})^2}{n}\Big).\end{aligned}$$

We conclude with the peeling argument provided in Lemma 4.6 of [26]:

$$\mathbb{E}\|P - P_n\|_{V_{r,\lambda}} \le c\sum_{i\ge 0} 2^{-i}\mathbb{E}\|P - P_n\|_{\mathcal{L}_{r,2^{i+1}\lambda}}. \qquad \square$$

Now, we can derive the following corollary. It gives several upper bounds for $\mathbb{E}\|P - P_n\|_{V_{r,\lambda}}$, depending on what B_r is (i.e. which penalty function is used).

Corollary 1 ($\|\cdot\|_{S_1}$ penalization)**.** *Let Assumptions 1 and 2 hold and assume that $B_r = B_{r,1,0,0}$ for $r > 0$, see (11). Then, we have*

$$\mathbb{E}\|P - P_n\|_{V_{r,\lambda_1(r)}} \le \frac{\lambda_1(r)}{8}$$

for any $r > 0$, where

$$\lambda_1(r) = c\Big(\frac{b_{X,2}^2 r^2(\log n)^2}{n} + \frac{b_{X,2}b_Y r\log n}{\sqrt{n}}\Big).$$

Proof. If $B_r = rB(S_1)$, we have using the embedding $K_{r,\lambda} \subset 2B_r$ and Proposition 1 that $U_n(K_{r,\lambda}) \le cb_{X,2}r\log n$, so

$$\phi_n(r,\lambda) \le c\Big(b_{X,2}r\log n\sqrt{\frac{\lambda}{n}} + b_{X,2}r\log n\sqrt{\frac{R(A_r^*)}{n}} + \frac{b_{2,X}^2 r^2(\log n)^2}{n}\Big) =: c\phi_{n,1}(r,x).$$

Hence, using Proposition 3 we obtain

$$\mathbb{E}\|P - P_n\|_{V_{r,\lambda}} \le c\sum_{i\ge 0} 2^{-i}\phi_{n,1}(r,2^{i+1}\lambda) \le c\phi_{n,1}(r,\lambda),$$

where we used the fact that the sum is comparable to its first term because of the exponential decay of the summands. Thus, one has $\mathbb{E}\|P - P_n\|_{V_{r,\lambda}} \le \lambda/8$ when $\lambda \ge c\phi_{n,1}(r,\lambda)$. In particular, since $R(A_r^*) \le \mathbb{E}Y^2 \le b_Y^2$ (see Assumption 2), for values of λ such that

$$\lambda \ge c\Big(\frac{b_{X,2}^2 r^2(\log n)^2}{n} + \frac{b_{X,2}b_Y r\log n}{\sqrt{n}}\Big),$$

we have $\mathbb{E}\|P - P_n\|_{V_{r,\lambda}} \le \lambda/8$, which proves the Corollary. $\square$

Corollary 2 ($\|\cdot\|_{S_1}+\|\cdot\|_1$ penalization)**.** *Let Assumptions 1 and 2 hold and assume that* $B_r = B_{r,r_1,0,r_3}$ *for* $r, r_1, r_3 > 0$*, see* (11)*. Then, we have*

$$\mathbb{E}\|P - P_n\|_{V_{r,\lambda_{r_1,0,r_3}(r)}} \leq \frac{\lambda_{r_1,0,r_3}(r)}{8}$$

for any $r > 0$*, where*

$$\lambda_{r_1,0,r_3}(r) = c\Big[\Big(\frac{1}{r_1^2}\wedge\frac{\log(mT)}{r_3^2}\Big)\frac{b_{X,2}^2 r^2(\log n)^2}{n}+\Big(\frac{1}{r_1}\wedge\frac{\sqrt{\log(mT)}}{r_3}\Big)\frac{b_{X,2}b_Y r(\log n)^{3/2}}{\sqrt{n}})\Big].$$

Proof. The proof follows the same steps as the proof of Corollary 1. □

Corollary 3 ($\|\cdot\|_{S_1}+\|\cdot\|_{S_2}^2$ penalization)**.** *Let Assumptions 1 and 2 hold and assume that* $B_r = B_{r,r_1,r_2,0}$ *for* $r, r_1, r_2 > 0$*, see* (11)*. Then, we have*

$$\mathbb{E}\|P - P_n\|_{V_{r,\lambda_{r_1,r_2}(r)}} \leq \frac{\lambda_{r_1,r_2}(r)}{8}$$

for any $r > 0$*, where*

$$\lambda_{r_1,r_2}(r) = c\Big(\frac{b_{X,2}^2 r(\log n)^2}{r_2 n} + \frac{b_{X,2}b_Y r\log n}{r_1\sqrt{n}}\Big).$$

Proof. Use the inclusion

$$B_r \subset \sqrt{\frac{r}{r_2}}B(S_2)\cap\frac{r}{r_1}B(S_1)$$

to obtain using Proposition 1 that

$$\phi_n(r,\lambda) \leq c\Big(b_{X,2}\sqrt{\frac{r}{r_2}}\log n\sqrt{\frac{\lambda}{n}} + b_{X,2}\frac{r}{r_1}\log n\sqrt{\frac{R(A_r^*)}{n}} + \frac{b_{X,2}^2 r(\log n)^2}{r_2 n}\Big).$$

The remaining of the proof is the same as the one of Corollary 1 so it is omitted. □

Corollary 4 ($\|\cdot\|_{S_1}+\|\cdot\|_{S_2}^2+\|\cdot\|_1$ penalization)**.** *Let Assumptions 1 and 2 hold and assume that* $B_r = B_{r,r_1,r_2,r_3}$ *for* $r, r_1, r_2, r_3 > 0$*, see* (11)*. Then, we have*

$$\mathbb{E}\|P - P_n\|_{V_{r,\lambda_{r_1,r_2,r_3}(r)}} \leq \frac{\lambda_{r_1,r_2,r_3}(r)}{8}$$

for any $r > 0$*, where*

$$\lambda_{r_1,r_2,r_3}(r) = c\Big[\frac{b_{X,2}^2 r(\log n)^2}{r_2 n} + \Big(\frac{1}{r_1}\wedge\frac{\sqrt{\log(mT)}}{r_3}\Big)\frac{b_{X,2}b_Y r(\log n)^{3/2}}{\sqrt{n}})\Big].$$

Proof. The proof follows the same steps as the proof of Corollary 3. □

The main difference between $\lambda_1(r), \lambda_{r,r_1,0,r_3}(r)$ and $\lambda_{r_1,r_2}(r), \lambda_{r_1,r_2,r_3}(r)$ is that $\lambda_{r_1,r_2}(r)$ and $\lambda_{r_1,r_2,r_3}(r)$ are linear in r while $\lambda_1(r)$ and $\lambda_{r_1,0,r_3}(r)$ are quadratic. The analysis of the isomorphic functions with quadratic terms will require an extra argument in the proof, in order to remove them from the penality (see below).

Remark 1 (Localization does not work here)**.** *Note that, in Corollaries 1 to 2, we don't use the fact that $K_{r,\lambda} \subset \sqrt{\lambda} D$, that is, we don't use the localization argument which usually allows to derive fast rates in statistical learning theory. Indeed, for the matrix completion problem, one has $\mathbb{E}\langle X, A - A_r^*\rangle^2 = \frac{1}{mT}\|A - A_r^*\|_{S_2}^2$, so when $\mathbb{E}\langle X, A - A_r^*\rangle^2 \leq \lambda$, we only know that $A \in A_r^* + \sqrt{mT\lambda} B(S_2)$, leading to a term of order mT/n (up to logarithms) in the isomorphic function. This term is way too large, since one has typically in matrix completion problems that $mT \gg n$.*

3.6 Isomorphic penalization method

We introduce the *isomorphic penalization method* developed by P. Bartlett, S. Mendelson and J. Neeman in the following general setup. Let $(\mathcal{Z}, \sigma_{\mathcal{Z}}, \nu)$ be a measurable space endowed with the probability measure ν. We consider $Z, Z_1, Z_2, \ldots, Z_n$ i.i.d. random variables having ν for common probability distribution. We are given a class $\mathcal{F}$ of functions on a measurable space $(\mathcal{X}, \sigma_{\mathcal{X}})$, a loss function and a risk function

$$Q : \mathcal{Z} \times \mathcal{F} \to \mathbb{R}; \quad R(f) = \mathbb{E}Q(Z, f).$$

For the problem we have in mind, we will use $Q((X,Y),A) = (Y - \langle X, A\rangle)^2$ for every $A \in \mathcal{M}_{m,T}$.

Now, we go into the core of the isomorphic penalization method. We are given a model $F \subset \mathcal{F}$ and a family $\{F_r : r \geq 0\}$ of subsets of F. We consider the following definition.

Definition 10 (cf. [26])**.** *Let ρ_n be a non-negative function defined on $\mathbb{R}_+ \times \mathbb{R}_+^*$ (which may depend on the sample). We say that the family $\{F_r : r \geq 0\}$ of subsets of F is an ordered, parameterized hierarchy of F with isomorphic function ρ_n when the following conditions are satisfied:*

1. *$\{F_r : r \geq 0\}$ is non-decreasing (that is $s \leq t \Rightarrow F_s \subseteq F_t$);*
2. *for any $r \geq 0$, there exists a unique element $f_r^* \in F_r$ such that $R(f_r^*) = \inf(R(f) : f \in F_r)$; we consider the excess loss function associated with the class F_r*
$$\mathcal{L}_{r,f}(\cdot) = Q(\cdot, f) - Q(\cdot, f_r^*); \tag{28}$$
3. *the map $r \longmapsto R(f_r^*)$ is continuous;*
4. *for every $r_0 \geq 0$, $\cap_{r \geq r_0} F_r = F_{r_0}$;*
5. *$\cup_{r \geq 0} F_r = F$;*

6. *for every* $r \geq 0$ *and* $u > 0$, *with probability at least* $1 - \exp(-u)$

$$(1/2)P_n\mathcal{L}_{r,f} - \rho_n(r,u) \leq P\mathcal{L}_{r,f} \leq 2P_n\mathcal{L}_{r,f} + \rho_n(r,u), \tag{29}$$

for any $f \in F_r$ *and* $P_n\mathcal{L}_{r,f} = (1/n)\sum_{i=1}^n \mathcal{L}_{r,f}(Z_i)$.

In the context of learning theory, ordered, parametrized hierarchy of a set F with isomorphic function ρ_n provides a very general framework for the construction of penalized empirical risk minimization procedure. The following result from [26] proves that the isomorphic function is a "correct penalty function".

Theorem 11 ([26])**.** *There exists absolute positive constants* c_1 *and* c_2 *such that the following holds. Let* $\{F_r : r \geq 0\}$ *be an ordered, parameterized hierarchy of* F *with isomorphic function* ρ_n. *Let* $u > 0$. *With probability at least* $1-\exp(-u)$ *any penalized empirical risk minimization procedure*

$$\hat{f} \in \operatorname*{argmin}_{f \in F} \Big(R_n(f) + c_1\rho_n(2(r(f)+1), \theta(r(f)+1, u)) \Big), \tag{30}$$

where $r(f) = \inf(r \geq 0 : f \in F_r)$ *and* $R_n(f) = (1/n)\sum_{i=1}^n Q(Z_i, f)$ *is the empirical risk of* f, *satisfies*

$$R(\hat{f}) \leq \inf_{f \in F} \Big(R(f) + c_2\rho_n(2(r(f)+1), \theta(r(f)+1, u)) \Big)$$

where for all $r \geq 1$ *and* $x > 0$,

$$\theta(r,x) = x + \ln(\pi^2/6) + 2\ln\Big(1 + \frac{R(f_0^*)}{\rho_n(0, x + \log(\pi^2/6))} + \log r\Big).$$

3.7 End of the proof of Theorems 1 and 2

First, we need to prove that the family of models $\{B_r : r \geq 0\}$ is an ordered, parametrized hierarchy of $\mathcal{M}_{m,T}$. First, fourth and fifth points of Definition 10 are easy to check. Second point follows from Lemma 5. For the third point, we consider $0 \leq q < r < s$, $\beta := q/r$ and $\alpha := r/s$. Since $\alpha A_s^* \in B_r$, we have

$$0 \leq R(A_r^*) - R(A_s^*) \leq R(\alpha A_s^*) - R(A_s^*) \leq (\alpha^2 - 1)\|\langle X, A_s^*\rangle\|_{L^2}^2 + 2(1-\alpha)\|Y\|_2\|\langle X, A_s^*\rangle\|_{L^2}.$$

As $s \to r$, the rights hand side tends to zero (because $\langle X, A_s^*\rangle$ are uniformly bounded in L_2 for $s \in [r, r+1]$). So $r \mapsto R(A_r^*)$ is upper semi-continuous on $(0, \infty)$. The continuity in $r = 0$ follows the same line. In the other direction,

$$0 \leq R(A_q^*) - R(A_r^*) \leq R(\beta A_r^*) - R(A_r^*) \leq (\alpha^2 - 1)\|\langle X, A_r^*\rangle\|_{L^2}^2 + 2(1-\alpha)\|Y\|_2\|\langle X, A_r^*\rangle\|_{L^2}$$

and the right hand side tends to zero for the same reason as before.

Now, we turn to the sixth point of Definition 10. That is the computation of the isomorphic function ρ_n associated with the family $\{B_r : r \geq 0\}$. Using Theorem 6 we obtain that, with a probability larger than $1 - 4e^{-x}$:

$$\frac{1}{2}P_n\mathcal{L}_{r,A} - \rho_n(r,x) \leq P\mathcal{L}_{r,A} \leq 2P_n\mathcal{L}_{r,A} + \rho_n(r,x) \quad \forall A \in B_r,$$

where

$$\rho_n(r,x) := c\Big[\lambda(r) + \big(b'_Y + C_r\big)^2\Big(\frac{x\log n}{n}\Big)\Big],$$

where $b'_Y := b_{Y,\psi_1} + b_{Y,\infty} + b_{Y,2}$, where C_r and $\lambda(r)$ are defined depending on the considered penalization (see (16) and Corollaries 1 to 4). Now, we apply Theorem 11 to the hierarchy $F_r = B_r$ for $r \geq 0$. First of all, note that, for every $x > 0$ and $r \geq 1$

$$\begin{aligned}\theta(r,x) &= x + \ln(\pi^2/6) + 2\ln\Big(1 + \frac{\mathbb{E}Y^2}{\rho_n(0, x + \log(\pi^2/6))} + \log r\Big)\\ &\leq x + c(\log n + \log\log r),\end{aligned}$$

so $\rho_n(2(r+1), \theta(r+1,x)) \leq \rho'_n(r,x)$, with:

$$\rho'_n(r,x) := c\Big[\lambda(2(r+1)) + (b'_Y + C_r)^2\frac{(x + \log n + \log\log r)\log n}{n}\Big].$$

From now on, the analysis depends on the penalization, so we consider them separately.

3.7.1 The $\|\cdot\|_{S_1}$ case

Recall that in this case

$$\lambda(r) = c\Big(\frac{b_{X,2}^2 r^2 (\log n)^2}{n} + \frac{b_{X,2} b_Y r \log n}{\sqrt{n}}\Big)$$

and $C_r = b_{X,\infty} r$, see (16). An easy computation gives $\rho'_n(r,x) \leq \tilde{\rho}_{n,1}(r,x)$ where

$$\tilde{\rho}_{n,1}(r,x) := c_{X,Y}\frac{(r+1)^2(x + \log n \vee \log\log r)\log n}{n} \vee p_{n,1}(r,x),$$

where $c_{X,Y} := c(1 + b_{X,2}^2 + b_Y b_X + b_{Y,\psi_1}^2 + b_{Y,\infty}^2 + b_{Y,2}^2 + b_{X,\infty}^2)$ and where

$$p_{n,1}(r,x) := c_{X,Y}\frac{(r+1)(x + \log n)\log n}{\sqrt{n}}.$$

Note that $p_{n,1}(r,x)$ is the penalty we want (the one considered in Theorem 1). Let us introduce for short $r(A) = \|A\|_{S_1}$ and the following functionals:

$$\begin{aligned}\Lambda_1(A) &= R(A) + \mathrm{pen}_1(A), \quad & \Lambda_{n,1}(A) &= R_n(A) + \mathrm{pen}_1(A),\\ \tilde{\Lambda}_1(A) &= R(A) + \tilde{\mathrm{pen}}_1(A), \quad & \tilde{\Lambda}_{n,1}(A) &= R_n(A) + \tilde{\mathrm{pen}}_1(A),\end{aligned}$$

where $\mathrm{pen}_1(A) := p_{n,1}(r(A),x)$ and where $\tilde{\mathrm{pen}}_1(A) := \tilde{\rho}_{n,1}(r(A),x)$ is a penalization that satisfies that, if $\tilde{A} \in \mathrm{argmin}_A \tilde{\Lambda}_{n,1}(A)$, then we have $R(\tilde{A}) \leq \inf_A \tilde{\Lambda}_1(A)$ with a probability larger than $1-4e^{-x}$. Recall that we want to prove that if $\hat{A} \in \mathrm{argmin}_A \Lambda_{n,1}(A)$, then we have $R(\hat{A}) \leq \inf_A \Lambda_1(A)$ with a probability larger than $1-5e^{-x}$. This will follow if we prove

$$\inf_A \tilde{\Lambda}_1(A) \leq \inf_A \Lambda_1(A) \quad \text{and} \tag{31}$$

$$\operatorname*{argmin}_A \Lambda_{n,1}(A) \subset \operatorname*{argmin}_A \tilde{\Lambda}_{n,1}(A), \tag{32}$$

so we focus on the proof of these two facts. First of all, let us prove that if $\tilde{\rho}_{n,1}(r,x) > p_{n,1}(r,x)$ then both r and $p_{n,1}(r,x)$ cannot be small.

If $\log n < \log\log r$ we have $r > e^n$ and $p_{n,1}(x,r) > c_{X,Y} e^n (\log n)^2/\sqrt{n}$. If $\log n \geq \log\log r$ and $\tilde{\rho}_{n,1}(r,x) > p_{n,1}(r,x)$, then

$$\frac{(r+1)^2(x+\log n)\log n}{n} > \frac{(r+1)(x+\log n)\log n}{\sqrt{n}},$$

so $r > \sqrt{n}-1$ and $p_{n,1}(r,x) > c_{X,Y}(\log n)^2$. Hence, we proved that if $\tilde{\rho}_{n,1}(r,x) > p_{n,1}(r,x)$, then $r > 1$ and $p_{n,1}(r,x) > c_{X,Y}(\log n)^2$. Note also that $p_{n,1}(r,x) > 2(x+\log n)\log n/\sqrt{n}$ since $r > 1$.

Let us turn to the proof of (31). Let A' be such that $\tilde{\Lambda}_1(A') > \Lambda_1(A')$. Then $\tilde{\mathrm{pen}}_1(A') > \mathrm{pen}_1(A')$, ie $\tilde{\rho}_{n,1}(r(A'),x) > p_{n,1}(r(A'),x)$, so that $r(A') > 1$, $p_{n,1}(r(A'),x) > c_{X,Y}(\log n)^2$ and $p_{n,1}(r(A'),x) > 2c_{X,Y}(x+\log n)\log n/\sqrt{n}$. On the other hand, we have $\inf_A \Lambda_1(A) \leq b_Y^2 + \mathrm{pen}_1(0) = b_Y^2 + p_{n,1}(0,x)$. But $p_{n,1}(r(A'),x) > c_{X,Y}(\log n)^2 > 2b_Y^2$ and $p_{n,1}(r(A'),x) > 2p_{n,1}(0,x)$ since $r(A') > 1$, so that $b_Y^2 + p_{n,1}(0,x) < p_{n,1}(r(A'),x)$ and then

$$\inf_A \Lambda_1(A) < p_n(r(A'),x) \leq \Lambda_1(A').$$

Hence, we proved that if A' is such that $\Lambda_1(A') \leq \inf_A \Lambda_1(A)$, we have $\tilde{\Lambda}_1(A') \leq \Lambda_1(A')$, so $\inf_A \tilde{\Lambda}_1(A) \leq \tilde{\Lambda}_1(A') \leq \Lambda_1(A') \leq \inf_A \Lambda_1(A)$, which proves (31).

The proof of (32) is almost the same. Let A' be such that $\tilde{\Lambda}_{n,1}(A') > \Lambda_{n,1}(A')$, so as before we have $r(A') > 1$, $p_{n,1}(r(A'),x) > c_{X,Y}(\log n)^2$ and $p_{n,1}(r(A'),x) > 2c_{X,Y}(x+\log n)\log n/\sqrt{n}$. This time we have $\inf_A \Lambda_{n,1}(A) \leq n^{-1}\sum_{i=1}^n Y_i^2 + p_{n,1}(0,x)$, so we use some concentration for the sum of the Y_i^2's. Indeed, we have, as a consequence of [2], that

$$\frac{1}{n}\sum_{i=1}^n Y_i^2 \leq \mathbb{E}Y^2 + c_1\sqrt{\mathbb{E}(Y^4)\frac{x}{n}} + c_2 \log n \frac{\|Y^2\|_{\psi_1} x}{n} \tag{33}$$

with a probability larger than $1-e^{-x}$. But then, it is easy to infer that for n large enough, the right hand side of (33) is smaller than $p_{n,1}(r(A'),x)/2$, so that we have, on an event of probability larger than $1-e^{-x}$, that

$$\inf_A \Lambda_{n,1}(A) \leq \frac{1}{n}\sum_{i=1}^n Y_i^2 + p_{n,1}(0,x) < p_{n,1}(r(A'),x) < \Lambda_{n,1}(A').$$

So, we proved that if $\Lambda_{n,1}(A') < \tilde{\Lambda}_{n,1}(A')$, then $A' \notin \operatorname{argmin}_A \Lambda_{n,1}(A)$, or equivalently that $\operatorname{argmin}_A \Lambda_{n,1}(A) \subset \{A : \tilde{\Lambda}_{n,1}(A) \leq \Lambda_{n,1}(A)\}$. But $\Lambda_{n,1}(A) \leq \tilde{\Lambda}_{n,1}(A)$ for any A (since $p_{n,1}(r,x) \leq \tilde{\rho}_{n,1}(r,x)$), so (32) follows. This concludes the proof of Theorem 1.

3.7.2 The $\|\cdot\|_{S_1} + \|\cdot\|_1$ case

Recall that in this case

$$\lambda(r) = c\Big[\Big(\frac{1}{r_1} \wedge \frac{\sqrt{\log(mT)}}{r_3}\Big)^2 \frac{b_{X,2}^2 r^2 (\log n)^2}{n} + \Big(\frac{1}{r_1} \wedge \frac{\sqrt{\log(mT)}}{r_3}\Big) \frac{b_{X,2} b_Y r (\log n)^{3/2}}{\sqrt{n}})\Big],$$

and that

$$C_r = \min\Big(b_{X,\infty}\frac{r}{r_1}, b_{X,\ell_\infty}\frac{r}{r_3}\Big),$$

see (16). An easy computation gives that $\rho_n'(r,x) \leq \tilde{\rho}_{n,2}(r,x)$, where

$$\tilde{\rho}_{n,2}(r,x) := c_{X,Y}\Big(\frac{1}{r_1} \wedge \frac{\sqrt{\log(mT)}}{r_3}\Big)^2 \frac{(r+1)^2(x + \log n \vee \log\log r)\log n}{n} \vee p_{n,2}(r,x),$$

where $c_{X,Y} = c(1 + b_{X,2}^2 + b_{X,2}b_Y + b_{Y,\psi_1}^2 + b_{Y,\infty}^2 + b_{Y,2}^2 + b_{X,\infty}^2 + b_{X,\ell_\infty}^2)$ and

$$p_{n,2}(r,x) := c_{X,Y}\Big(\frac{1}{r_1} \wedge \frac{\sqrt{\log(mT)}}{r_3}\Big) \frac{(r+1)(x+\log n)(\log n)^{3/2}}{\sqrt{n}}.$$

Note that $p_{n,2}(r,x)$ is the penalization we want (the one considered in Theorem 3). Introducing $r(A) = r_1\|A\|_{S_1} + r_3\|A\|_1$, the remaining of the proof follows the lines of the pure $\|\cdot\|_{S_1}$ case, so it is omitted.

3.7.3 The $\|\cdot\|_{S_1} + \|\cdot\|_{S_2}^2$ case

This is easier than what we did for the $\|\cdot\|_{S_1}$ case, since we only have a $\log\log r$ term to remove from the penalization. Recall that

$$\lambda(r) = c\Big(\frac{b_{X,2}^2 r (\log n)^2}{r_2 n} + \frac{b_{X,2} b_Y r \log n}{r_1\sqrt{n}}\Big),$$

and

$$C_r = \min\Big(b_{X,\infty}\frac{r}{r_1}, b_{X,2}\sqrt{\frac{r}{r_2}}\Big) \leq b_{X,2}\sqrt{\frac{r}{r_2}},$$

so that $\rho_n'(r,x) \leq \tilde{\rho}_{n,3}(r,x)$ where

$$\tilde{\rho}_{n,3}(r,x) = c_{X,Y}\frac{(r+1)\log n}{\sqrt{n}}\Big(\frac{1}{r_1} + \frac{(x + \log n \vee \log\log r)\log n}{r_2\sqrt{n}}\Big),$$

where $c_{X,Y} = c(1 + b_{X,2}^2 + b_{X,2}b_Y + b_{Y,\psi_1}^2 + b_{Y,\infty}^2 + b_{Y,2}^2)$. This is almost the penalty we want, up to the $\log\log r$ term, so we consider.

$$p_{n,3}(r,x) = c_{X,Y}\frac{(r+1)\log n}{\sqrt{n}}\Big(\frac{1}{r_1} + \frac{(x+\log n)\log n}{r_2\sqrt{n}}\Big),$$

Let us introduce for short

$$r(A) := r_1\|A\|_{S_1} + r_2\|A\|_{S_2}^2 = \inf\big(r \geq 0 : A \in B_r\big)$$

and the following functionals:

$$\begin{aligned}
\Lambda_3(A) &= R(A) + \mathrm{pen}_3(A), \quad & \Lambda_{n,3}(A) &= R_n(A) + \mathrm{pen}_3(A),\\
\tilde{\Lambda}_3(A) &= R(A) + \tilde{\mathrm{pen}}_3(A), \quad & \tilde{\Lambda}_{n,3}(A) &= R_n(A) + \tilde{\mathrm{pen}}_3(A),
\end{aligned}$$

where $\mathrm{pen}_3(A) := p_{n,3}(r(A), x)$ and where $\tilde{\mathrm{pen}}_3(A) := \tilde{\rho}_{n,3}(r(A), x)$. We only need to prove that

$$\inf_A \tilde{\Lambda}_3(A) \leq \inf_A \Lambda_3(A) \quad \text{and} \tag{34}$$

$$\operatorname*{argmin}_A \Lambda_{n,3}(A) \subset \operatorname*{argmin}_A \tilde{\Lambda}_{n,3}(A). \tag{35}$$

Obviously, if $\tilde{\rho}_{n,3}(r, x) > p_{n,3}(r, x)$, then $r > e^n$, so following the arguments we used for the S_1 penalty, it is easy to prove both (34) and (35). This concludes the proof of Theorem 2.

3.7.4 The $\|\cdot\|_{S_1} + \|\cdot\|_{S_2}^2 + \|\cdot\|_1$ case

Recall that in this case

$$\lambda(r) = c\Big[\frac{b_{X,2}^2 r(\log n)^2}{r_2 n} + \Big(\frac{1}{r_1} \wedge \frac{\sqrt{\log(mT)}}{r_3}\Big)\frac{b_{X,2} b_Y r(\log n)^{3/2}}{\sqrt{n}}\Big].$$

and that

$$C_r = \min\Big(b_{X,\infty}\frac{r}{r_1}, b_{X,2}\sqrt{\frac{r}{r_2}}, b_{X,\ell_\infty}\frac{r}{r_3}\Big) \leq b_{X,2}\sqrt{\frac{r}{r_2}}, \tag{36}$$

see (16). An easy computation gives that $\rho_n'(r, x) \leq \tilde{\rho}_{n,4}(r, x)$, where

$$\tilde{\rho}_{n,4}(r, x) := c_{X,Y}\frac{(r+1)(\log n)^{3/2}}{\sqrt{n}}\Big(\frac{1}{r_1} \wedge \frac{\sqrt{\log(mT)}}{r_3} + \frac{x + \log n \vee \log\log r}{r_2\sqrt{n}}\Big)$$

where $c_{X,Y} = c(1 + b_{X,2}^2 + b_{X,2}b_Y + b_{Y,\psi_1}^2 + b_{Y,\infty}^2 + b_{Y,2}^2)$. The penalization we want is

$$p_{n,4}(r, x) := c_{X,Y}\frac{(r+1)(\log n)^{3/2}}{\sqrt{n}}\Big(\frac{1}{r_1} \wedge \frac{\sqrt{\log(mT)}}{r_3} + \frac{x + \log n}{r_2\sqrt{n}}\Big),$$

so introducing $r(A) = r_1\|A\|_{S_1} + r_2\|A\|_{S_2}^2 + r_3\|A\|_1$ and following the lines of the proof of the $S_1 + S_2$ case to remove the $\log\log r$ term, it is easy to conclude the proof of Theorem 4.

References

[1] J. Abernethy, F. Bach, T. Evgeniou, and J.P. Vert. A new approach to collaborative filtering: Operator estimation with spectral regularization. *The Journal of Machine Learning Research*, 10:803–826, 2009.

[2] 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.

[3] A. Argyriou, T. Evgeniou, and M. Pontil. Convex multi-task feature learning. *Machine Learning*, 73(3):243–272, 2008.

[4] A. Argyriou, C.A. Micchelli, and M. Pontil. On spectral learning. *The Journal of Machine Learning Research*, 11:935–953, 2010.

[5] A. Argyriou, C.A. Micchelli, M. Pontil, and Y. Ying. A spectral regularization framework for multi-task structure learning. *Advances in Neural Information Processing Systems*, 20:25–32, 2008.

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

[7] Peter Bartlett, Shahar Mendelson, and Neeman Joseph. ℓ_1-regularized linear regression: Persistence and oracle inequalitie. *To appear in Bernoulli*.

[8] Peter L. Bartlett. Fast rates for estimation error and oracle inequalities for model selection. *Econometric Theory*, 24(2), 2008. (To appear. Was Department of Statistics, U.C. Berkeley Technical Report number 729, 2007).

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

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

[11] J-F Cai, Cands E. J., and Z. Shen. A singular value thresholding algorithm for matrix completion. *SIAM J. on Optimization*, 20(4):1956–1982, 2008.

[12] E.J. Candes and T. Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51(12):4203–4215, 2005.

[13] E. J. Cands and Y. Plan. Tight oracle bounds for low-rank matrix recovery from a minimal number of random measurements. Technical report, Department of Statistics, Stanford Univeristy, 2009.

[14] E. J. Cands and Y. Plan. Matrix completion with noise. *Proceedings of the IEEE*, to appear.

[15] E. J. Cands and B. Recht. Exact matrix completion via convex optimization. *Found. of Comput. Math.*, 9:717–772, 2008.

[16] E. J. Cands and T. Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Trans. Inform. Theory*, to appear.

[17] 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.

[18] M. Fazel, H. Hindi, and S. Boyd. Rank minimization and applications in system theory. In *In American Control Conference*, pages 3273–3278. AACC, 2004.

[19] M. Fazel, H. Hindi, and S.P. Boyd. Log-det heuristic for matrix rank minimization with applications to Hankel and Euclidean distance matrices. In *Proceedings of the American Control Conference*, volume 3, pages 2156–2162, 2003.

[20] Evarist Giné and Joel Zinn. Some limit theorems for empirical processes. *Ann. Probab.*, 12(4):929–998, 1984. With discussion.

[21] D. Gross. Recovering low-rank matrices from few coefficients in any basis. *CoRR, abs/0910.1879*, 2009.

[22] Olivier Guédon, Shahar Mendelson, Alain Pajor, and Nicole Tomczak-Jaegermann. Subspaces and orthogonal decompositions generated by bounded orthogonal systems. *Positivity*, 11(2):269–283, 2007.

[23] Raghunandan H. Keshavan, Andrea Montanari, and Sewoong Oh. Matrix completion from noisy entries. *CoRR*, abs/0906.2027, 2009.

[24] R.H. Keshavan, S. Oh, and A. Montanari. Matrix completion from a few entries. *arxiv*, 901, 2009.

[25] 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.

[26] Shahar Mendelson and Joseph Neeman. Regularization in kernel learning. Technical report, 2009. To appear in Annals of Statistics, availble at http://www.imstat.org/aos/.

[27] Shahar Mendelson, Alain Pajor, and Nicole Tomczak-Jaegermann. Reconstruction and subgaussian operators in asymptotic geometric analysis. *Geom. Funct. Anal.*, 17(4):1248–1282, 2007.

[28] Shahar Mendelson, Alain Pajor, and Nicole Tomczak-Jaegermann. Uniform uncertainty principle for Bernoulli and subgaussian ensembles. *Constr. Approx.*, 28(3):277–289, 2008.

[29] S. Negahban, P. Ravikumar, M.J. Wainwright, and B. Yu. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Advances in Neural Information Processing Systems*, 2009.

[30] S. Negahban and M.J. Wainwright. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *Arxiv preprint arXiv:0912.5100*, 2009.

[31] Alain Pajor and Nicole Tomczak-Jaegermann. Remarques sur les nombres d'entropie d'un opérateur et de son transposé. *C. R. Acad. Sci. Paris Sér. I Math.*, 301(15):743–746, 1985.

[32] B. Recht. A simpler approach to matrix completion. *CoRR, abs/0910.0651*, 2009.

[33] B. Recht, M. Fazel, and P.A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *preprint*, 2007.

[34] Angelika Rohde and Alexandre B. Tsybakov. Estimation of high-dimensional low rank matrices. Technical report, Universität Hamburg and Université Paris 6, 2009.

[35] M. Rudelson. Random vectors in the isotropic position. *J. Funct. Anal.*, 164(1):60–72, 1999.

[36] Mark Rudelson and Roman Vershynin. On sparse reconstruction from Fourier and Gaussian measurements. *Comm. Pure Appl. Math.*, 61(8):1025–1045, 2008.

[37] Carsten Schütt. Entropy numbers of diagonal operators between symmetric Banach spaces. *J. Approx. Theory*, 40(2):121–128, 1984.

[38] N. Srebro, J.D.M. Rennie, and T.S. Jaakkola. Maximum-margin matrix factorization. *Advances in neural information processing systems*, 17:1329–1336, 2005.

[39] N. Srebro and A. Shraibman. Rank, trace-norm and max-norm. *Learning Theory*, pages 545–560, 2005.

[40] Michel Talagrand. New concentration inequalities in product spaces. *Invent. Math.*, 126(3):505–563, 1996.

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

[42] 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.

[43] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 67(2):301–320, 2005.