Low Rank Estimation of Smooth Kernels on Graphs

Vladimir Koltchinskii * **and Pedro Rangel** †
School of Mathematics
Georgia Institute of Technology
Atlanta, GA 30332-0160
vlad@math.gatech.edu, prangel@math.gatech.edu

July 23, 2012

Abstract

Let (V, A) be a weighted graph with a finite vertex set V, with a symmetric matrix of nonnegative weights A and with Laplacian Δ. Let $S_* : V \times V \mapsto \mathbb{R}$ be a symmetric kernel defined on the vertex set V. Consider n i.i.d. observations $(X_j, X_j', Y_j), j = 1, \dots, n$, where X_j, X_j' are independent random vertices sampled from the uniform distribution in V and $Y_j \in \mathbb{R}$ is a real valued response variable such that $\mathbb{E}(Y_j | X_j, X_j') = S_*(X_j, X_j'), j = 1, \dots, n$. The goal is to estimate the kernel S_* based on the data $(X_1, X_1', Y_1), \dots, (X_n, X_n', Y_n)$ and under the assumption that S_* is low rank and, at the same time, smooth on the graph (the smoothness being characterized by discrete Sobolev norms defined in terms of the graph Laplacian). We obtain several results for such problems including minimax lower bounds on the L_2-error and upper bounds for penalized least squares estimators both with nonconvex and with convex penalties.

1 Introduction

We study a problem of estimation of a symmetric kernel $S_* : V \times V \mapsto \mathbb{R}$ defined on a large weighted graph with a vertex set V and $m := \mathrm{card}(V)$, based on a finite number of noisy linear measurements of S_*. For simplicity, assume that these are the measurements of randomly picked entries of $m \times m$ matrix $(S_*(u, v))_{u,v \in V}$, which is a standard sampling model in matrix completion. More precisely, let $(X_j, X_j', Y_j), j = 1, \dots, n$ be n

*Partially supported by NSF Grants DMS-1207808, DMS-0906880 and CCF-0808863
†Supported by NSF Grant CCF-0808863

independent copies of a random triple (X, X', Y), where X, X' are independent random vertices sampled from the uniform distribution Π in V and $Y \in \mathbb{R}$ is a "measurement" of the kernel S_* at a random location (X, X') in the sense that $\mathbb{E}(Y|X, X') = S_*(X, X')$. In what follows, we assume that, for some constant $a > 0$, $|Y| \le a$ a.s., which implies that $|S_*(u, v)| \le a, u, v \in V$. The target kernel S_* is to be estimated based on its i.i.d. measurements $(X_j, X_j', Y_j), j = 1, \dots, n$. We would like to study this problem in the case when the target kernel S_* is, on the one hand, "low rank" (that is, $\mathrm{rank}(S_*)$ is relatively small comparing with m) and, on the other hand, it is "smooth" in the sense that its "Sobolev type norm" is not too large. Discrete versions of Sobolev norms can be defined for functions and kernels on weighted graphs in terms of their graph Laplacians. As a typical example, one can consider a problem of learning a binary relationship (say, "similarity") between vertices of the graph (see Koltchinskii and Rangel (2012)). In this case, $(X, X', Y) \in V \times V \times \{-1, 1\}$, $Y = +1$ meaning that the vertices X, X' are "similar" and $Y = -1$ meaning that they are not. The goal is to predict Y for a given couple of vertices (X, X') based on the training data $(X_j, X_j', Y_j), j = 1, \dots, n$. Clearly, the optimal classifier is $\mathrm{sign}(S_*(X, X'))$, where $S_*(X, X') = \mathbb{E}(Y|X, X')$ is the regression function. In the learning theory literature, there has been a number of attempts to develop classification methods based on similarities between the objects (see, e.g., Balcan et al (2008), Maurer (2008), Chen et al (2009)). In problems of this kind, it is of importance to learn kernels suitable for representing and predicting such similarity relationships. This is also important in various classification problems in large complex networks (see, e.g., Leskovec et al (2010)). Our main motivation, however, is mostly theoretical: we would like to explore to which extent taking into account smoothness of the target kernel could improve the existing methods of low rank recovery.

We introduce some notations used throughout the paper. Let $\mathcal{S}_V$ be the linear space of *symmetric kernels* $S : V \times V \mapsto \mathbb{R}$, $S(u, v) = S(v, u), u, v \in V$ (or, equivalently, symmetric $m \times m$ matrices with real entries). Given $S \in \mathcal{S}_V$, we use the notation $\mathrm{rank}(S)$ for the rank of S and $\mathrm{tr}(S)$ for its trace. For two functions $f, g : V \mapsto \mathbb{R}$, $(f \otimes g)(u, v) := f(u)g(v)$. Suppose that $S = \sum_{j=1}^r \mu_j(\psi_j \otimes \psi_j)$ is the spectral representation of S with $r = \mathrm{rank}(S)$, $\mu_1, \dots, \mu_r$ being non-zero eigenvalues of S repeated with their multiplicities and $\psi_1, \dots, \psi_r$ being the corresponding orthonormal eigenfunctions (obviously, there are multiple choices of ψ_js in the case of repeated eigenvalues). We will define $\mathrm{sign}(S)$ as $\mathrm{sign}(S) := \sum_{j=1}^r \mathrm{sign}(\mu_j)(\psi_j \otimes \psi_j)$ and the support of S as $\mathrm{supp}(S) := \mathrm{l.s.}\{\psi_1, \dots, \psi_r\}$.[1]

[1] "l.s." means "the linear span".

For $1 \le p < \infty$, define the Schatten p-norm of S as

$$\|S\|_p := (\mathrm{tr}(|S|^p))^{1/p} = \left(\sum_{j=1}^{r} |\mu_j|^p \right)^{1/p},$$

where $|S| := \sqrt{S^2}$. For $p = 1$, $\|\cdot\|_1$ is also called the nuclear norm and, for $p = 2$, $\|\cdot\|_2$ is called the Hilbert–Schmidt or Frobenius norm. This norm is induced by the Hilbert–Schmidt inner product which will be denoted by $\langle \cdot, \cdot \rangle$. The operator norm of S is defined as $\|S\| := \max_j |\mu_j|$.[2]

Let $\Pi^2 := \Pi \otimes \Pi$ be the distribution of random couple (X, X'). The $L_2(\Pi^2)$-norm of kernel S,

$$\|S\|^2_{L_2(\Pi^2)} = \int_{V \times V} |S(u,v)|^2 \Pi^2(du, dv) = \mathbb{E}|S(X, X')|^2,$$

is naturally related to the sampling model studied in the paper and it will be used to measure the estimation error. Denote by $\langle \cdot, \cdot \rangle_{L_2(\Pi^2)}$ the corresponding inner product. Since Π is the uniform distribution in V, $\|S\|_{L_2(\Pi^2)} = m^{-2}\|S\|_2^2$ and $\langle S_1, S_2 \rangle_{L_2(\Pi^2)} = m^{-2}\langle S_1, S_2 \rangle$. In what follows, it will be often more convenient to use these rescaled versions rather than the actual Hilbert–Schmidt norm or inner product.

We will also denote by $\{e_v : v \in V\}$ the canonical orthonormal basis of the space $\mathbb{R}^V$. Based on this basis, one can construct matrices $E_{u,v} = E_{v,u} = \frac{1}{2}(e_u \otimes e_v + e_v \otimes e_u)$. If $v_1, \ldots, v_m$ is an arbitrary ordering of the vertices in V, then $\{E_{v_j, v_j} : j = 1, \ldots, m\} \cup \{\sqrt{2} E_{v_i, v_j} : 1 \le i < j \le m\}$ is an orthonormal basis of the space $\mathcal{S}_V$ of symmetric matrices with Hilbert–Schmidt inner product.

In standard matrix completion problems, V is a finite set with no further structure (that is, the set of edges of the graph or the weight matrix are not specified). In the noiseless matrix completion problems, the target matrix S_* is to be recovered from the measurements $(X_j, X'_j, Y_j), j = 1, \ldots, n$, where $Y_j = S_*(X_j, X'_j)$. The following method is based on nuclear norm minimization over the space of all matrices that "agree" with the data:

$$\hat{S} := \mathrm{argmin}\{\|S\|_1 : S \in \mathcal{S}_V, S(X_j, X'_j) = Y_j, j = 1, \ldots, n\}, \tag{1.1}$$

It has been studied in detail in the recent literature, see Candes and Recht (2009), Recht, Fazel and Parrilo (2010), Candes and Tao (2010), Gross (2011) and references therein. Clearly, there are low rank matrices S_* that can not be recovered based on a random

[2] With some abuse of notation, we also denote occasionally the canonical Euclidean inner product in $\mathbb{R}^V$ by $\langle \cdot, \cdot \rangle$ and the corresponding Euclidean norm by $\|\cdot\|$.

sample of n entries unless n is comparable with the total number of the entries of the matrix. For instance, for given $u, v \in V$, let $S_* = E_{u,v}$. Then, $\mathrm{rank}(S_*) \le 2$. However, the probability that the only two non-zero entries of S_* are not present in the sample is $(1 - \frac{2}{m^2})^n$, and it is close to 1 when $n = o(m^2)$. In this case, the matrix S_* can not be recovered. So called *low coherence* assumptions have been developed to define classes of "generic" matrices that are not "low rank" and "sparse" at the same time and for which noiseless low rank recovery is possible with a relatively small number of measurements. For a linear subspace $L \subset \mathbb{R}^V$, let $L^\perp$ be the orthogonal complement of L and let P_L be the orthogonal projector onto the subspace L. Denote $L := \mathrm{supp}(S_*)$, $r = \mathrm{rank}(S_*)$. A *coherence coefficient* is a constant $\nu \ge 1$ such that

$$\|P_L e_v\|^2 \le \frac{\nu r}{m}, \ v \in V \ \text{ and } \ |\langle \mathrm{sign}(S_*) e_u, e_v \rangle|^2 \le \frac{\nu r}{m^2}, u, v \in V. \tag{1.2}$$

(it is easy to see that ν can not be smaller than 1).

The following highly nontrivial result is essentially due to Candes and Tao (2010) (a version stated here is due to Gross (2011) and it is an improvement of the initial result of Candes and Tao). It shows that target matrices of "low coherence" (for which ν is a relatively small constant) can be recovered exactly using the nuclear norm minimization algorithm (1.1) provided that the number of observed entries is of the order mr (up to a log factor).

Theorem 1 *Suppose conditions (1.2) hold for some $\nu \ge 1$. Then, there exists a numerical constant $C > 0$ such that, for all $n \ge C\nu r m \log^2 m$, $\hat{S} = S_*$ with probability at least $1 - m^{-2}$.*

In the case of noisy matrix completion, a matrix version of LASSO is based on a trade-off between fitting the target matrix to the data using least squares and minimizing the nuclear norm:

$$\hat{S} := \mathrm{argmin}_{S \in \mathcal{S}_V} \left[n^{-1} \sum_{j=1}^n (Y_j - S(X_j, X_j'))^2 + \varepsilon \|S\|_1 \right]. \tag{1.3}$$

This method has been studied by a number of authors, including Candes and Plan (2011), Rohde and Tsybakov (2011), Negahban and Wainwright (2010), Koltchinskii, Lounici and Tsybakov (2011), Koltchinskii (2011b). In the case of known design distribution Π (in particular, in the case of uniform design) one can use instead of (1.3) the following modification of nuclear norm penalized least squares method:

$$\hat{S} := \mathrm{argmin}_{S \in \mathcal{S}_V} \left[\|S\|^2_{L_2(\Pi^2)} - \frac{2}{n} \sum_{j=1}^n Y_j S(X_j, X_j') + \varepsilon \|S\|_1 \right]. \tag{1.4}$$

Note that, if the norm $\|S\|_{L_2(\Pi^2)}$ in (1.4) is replaced by the $L_2(\Pi_n)$-norm, where Π_n is the empirical distribution based on $(X_1, X_1'), \ldots, (X_n, X_n')$, then the resulting estimator coincides with (1.3).

The next result was proved by Koltchinskii, Lounici and Tsybakov (2011) (see their Theorem 4).

Theorem 2 *Assume that, for some constant $a > 0$, $|Y| \le a$ a.s. Let $t > 0$ and suppose that*

$$\varepsilon \ge 4a\left(\sqrt{\frac{t+\log(2m)}{nm}} \bigvee \frac{2(t+\log(2m))}{n}\right).$$

Then, there exists a constant $C > 0$ such that with probability at least $1 - e^{-t}$

$$\|\hat{S} - S_*\|^2_{L_2(\Pi^2)} \le \inf_{S \in \mathcal{S}_V}\left[\|S - S_*\|^2_{L_2(\Pi^2)} + Cm^2\varepsilon^2 \mathrm{rank}(S)\right].$$

In particular, Theorem 2 implies that, with probability at least $1 - e^{-t}$,

$$\|\hat{S} - S_*\|^2_{L_2(\Pi^2)} \le Cm^2\varepsilon^2 \mathrm{rank}(S_*).$$

Very recently, Klopp (2012) proved a similar bound for the matrix LASSO estimator (1.3) in the case when the domain of optimization problem is $\{S : \|S\|_{L_\infty} \le a\}$, where $\|S\|_{L_\infty} := \max_{u,v \in V} |S(u,v)|$.[3]

In the current paper, we are more interested in the case when the target kernel S_* is defined on the set V of vertices of a weighted graph with a weight matrix A. This allows one to define the notion of graph Laplacian and to introduce Sobolev type norms characterizing smoothness of functions on V as well as symmetric kernels on $V \times V$.

Let $G = (V, A)$ be a weighted graph with vertex set V and weight matrix A. It is assumed that $A := (a(u,v))_{u,v \in V}$ is a symmetric $m \times m$ matrix with nonnegative entries (or, equivalently, a symmetric kernel on V). Denote

$$\deg(u) := \sum_{v \in V} a(u,v), u \in V.$$

It is common in graph theory to call $\deg(u)$ the degree of vertex u. Let D be the diagonal $m \times m$ matrix (kernel) with the degrees of vertices on the diagonal (it is assumed that the vertices of the graph have been ordered in an arbitrary, but fixed way). The Laplacian

[3] In fact, Koltchinskii, Lounici and Tsybakov (2011) and Klopp (2012) studied low rank recovery problems for rectangular matrices. However, modification of their results to the case of symmetric matrices is straightforward.

of the weighted graph G is defined as $\Delta := D - A$. Denote $\langle \cdot, \cdot \rangle$ the canonical Euclidean inner product in the m-dimensional space $\mathbb{R}^V$ of functions $f : V \mapsto \mathbb{R}$ and let $\| \cdot \|$ be the corresponding norm. It is easy to see that

$$\langle \Delta f, f \rangle = \frac{1}{2} \sum_{u,v \in V} a(u,v)(f(u) - f(v))^2,$$

implying that $\Delta : \mathbb{R}^V \mapsto \mathbb{R}^V$ is a symmetric nonnegatively definite linear transformation. In a special case of a usual graph (V, E) with vertex set V and edge set E, one defines $A(u,v) = 1$ iff $u \sim v$ (that is, vertices u and v are connected with an edge) and $A(u,v) = 0$ otherwise. In this case, $\deg(u)$ is the number of edges incident to the vertex u and

$$\langle \Delta f, f \rangle = \sum_{u \sim v} (f(u) - f(v))^2.$$

The notion of graph Laplacian allows one to define Sobolev type norms $\|\Delta^{p/2} f\|, p > 0$ for functions on the vertex set of the graph and, thus, to describe their smoothness on the graph. Given a symmetric kernel $S : V \times V \mapsto \mathbb{R}$, one can also describe its smoothness in terms of the norms $\|\Delta^{p/2} S\|_2$. Suppose S has the following spectral representation: $S = \sum_{j=1}^m \mu_j (\psi_j \otimes \psi_j)$, where $\mu_j, j = 1, \dots, m$ are the eigenvalues of S (repeated with their multiplicities) and $\psi_j, j = 1, \dots, m$ are the corresponding orthonormal eigenfunctions in $\mathbb{R}^V$, then

$$\|\Delta^{p/2} S\|_2^2 = \operatorname{tr}(\Delta^{p/2} S^2 \Delta^{p/2}) = \operatorname{tr}(\Delta^p S^2) = \sum_{j=1}^m \mu_j^2 \langle \Delta^p \psi_j, \psi_j \rangle = \sum_{j=1}^m \mu_j^2 \|\Delta^{p/2} \psi_j\|^2.$$

Basically, it means that the smoothness of the kernel S depends on the smoothness of its eigenfunctions. In what follows, we will often use rescaled versions of Sobolev norms:

$$\|\Delta^{p/2} f\|_{L_2(\Pi)} = m^{-1/2} \|\Delta^{p/2} f\|^2, \quad \|\Delta^{p/2} S\|_{L_2(\Pi^2)} = m^{-1} \|\Delta^{p/2} S\|_2.$$

It will be convenient for our purposes to fix $p > 0$ and to define a nonnegatively definite symmetric kernel $W := d\Delta^p$, where d is a fixed constant. We will characterize smoothness of a kernel $S \in \mathcal{S}_V$ by the squared Sobolev type norm $\|W^{1/2} S\|_{L_2(\Pi^2)}^2$. The kernel W will be fixed throughout the paper and its spectral properties are crucial in our analysis.[4] Assume that W has the following spectral representation $W = \sum_{k=1}^m \lambda_k (\phi_k \otimes \phi_k)$, where $0 \le \lambda_1 \le \cdots \le \lambda_m$ are the eigenvalues repeated with their multiplicities

[4] In fact, the relationship of W to the graph and its Laplacian will be of little importance allowing, possibly, other interpretations of the problem.

and $\phi_1, \dots, \phi_m$ are the corresponding orthonormal eigenfunctions (of course, there is a multiple choice of ϕ_k in the case of repeated eigenvalues). Let $k_0 := \min\{k \le m : \lambda_k > 0\}$. We will assume in what follows that, for some constant $c \ge 1$, $\lambda_{k+1} \le c\lambda_k$ for all $k \ge k_0$. It will be also convenient to set $\lambda_k := +\infty, k > m$.

Let $\rho := \|W^{1/2} S_*\|_{L_2(\Pi^2)}$ and $r := \mathrm{rank}(S_*)$. It is easy to show (see the proof of Theorem 5 below) that kernel S_* can be approximated by the following kernel $S_{*,l} := \sum_{i,j}^{l} \langle S_* \phi_i, \phi_j \rangle (\phi_i \otimes \phi_j)$ with the approximation error

$$\|S_* - S_{*,l}\|^2_{L_2(\Pi^2)} \le \frac{2\rho^2}{\lambda_{l+1}}. \tag{1.5}$$

Note that the kernel $S_{*,l}$ can be viewed as an $l \times l$ matrix (represented in the basis of eigenfunctions $\{\phi_j\}$) and $\mathrm{rank}(\mathrm{S}_*) \le r \wedge l$, so, one needs $\sim (r \wedge l)l$ parameters to characterize such matrices. Thus, one can expect, that such a kernel can be estimated, based on n linear measurements, with the squared $L_2(\Pi^2)$-error of the order $\frac{a^2(r \wedge l)l}{n}$. Taking into account the bound on the approximation error (1.5) and optimizing with respect to $l = 1, \dots, m$, it would be also natural to expect the following error rate in the problem of estimation of the target kernel S_* : [5]

$$\min_{1 \le l \le m} \left[\frac{a^2(r \wedge l)l}{n} \bigvee \frac{\rho^2}{\lambda_{l+1}} \right]. \tag{1.6}$$

We will show that such a rate is attained (up to constants and log factors) for a version of least squares method with a nonconvex complexity penalty (see Section 3). This method is not computationally tractable, so, we also study another method, based on convex penalization with a combination of nuclear norm and squared Sobolev type norm, and show that the rates are attained for such a method, too, provided that the target matrix satisfies a version low coherence assumption with respect to the basis of eigenfunctions of W (see Section 4). Finally, we prove minimax lower bounds on the error rate that are roughly of the order

$$\max_{1 \le l \le m} \left[\frac{a^2(r \wedge l)l}{n} \bigwedge \frac{\rho^2}{\lambda_l} \right]$$

(subject to some extra conditions and with additional terms; see Section 2). In typical situations, this expression is of the same order as the upper bound (1.6). For instance, if $\lambda_l \asymp l^{2\beta}$ for some $\beta > 1/2$, then the minimax error rate of estimation of the target

[5] It is easy to modify the results of the paper and to control the error in terms of variance of the noise of observations Y_j rather than the "range" a of these observations.

kernel S_* is of the order

$$\left(\left(\frac{a^2\rho^{1/\beta}r}{n}\right)^{2\beta/(2\beta+1)} \bigwedge \left(\frac{a^2\rho^{2/\beta}}{n}\right)^{\beta/(\beta+1)} \bigwedge \frac{a^2rm}{n}\right) \bigvee \frac{a^2}{n}$$

(up to log factors). When m is sufficiently large, the term $\frac{a^2rm}{n}$ will be dropped from the minimum and we end up with a nonparametric convergence rate controlled by the smoothness parameter β and the rank r of the target matrix S_* (the dependence on m in the first two terms of the minimum is only in the log factors).

The focus of the paper is on the matrix completion problems with uniform random design, but it is very straightforward to extend the results of the following sections to sampling models with more general design distributions discussed in the literature on low rank recovery (such as, for instance, the models of random linear measurements studied in Koltchinskii, Lounici and Tsybakov (2011), Koltchinskii (2011b)). It is also not hard to replace the range a of the response variable Y by the standard deviation of the noise in the upper and lower bounds obtained below. This is often done in the literature on low rank recovery and it can be easily extended to the framework discussed in the paper by modifying our proofs. We have not discussed this in the paper due to the lack of space.

2 Minimax Lower Bounds

In this section, we derive minimax lower bounds on the $L_2(\Pi^2)$-error of an arbitrary estimator $\hat{S}$ of the target kernel S_* under the assumptions that the response variable Y is bounded by a constant $a > 0$, the rank of S_* is bounded by $r \leq m$ and its Sobolev norm $\|W^{1/2}S_*\|_{L_2(\Pi^2)}$ is bounded by $\rho > 0$. More precisely, given $r = 1, \dots, m$ and $\rho > 0$, denote by $\mathcal{S}_{r,\rho}$ the set of all symmetric kernels $S : V \times V \mapsto \mathbb{R}$ such that

(i) $\mathrm{rank}(S) \leq r$;

(ii) $\|W^{1/2}S\|_{L_2(\Pi^2)} \leq \rho$.

Given r, ρ and $a > 0$, let $\mathcal{P}_{r,\rho,a}$ be the set of all probability distributions of (X, X', Y) such that

(i) (X, X') is uniformly distributed in $V \times V$;

(ii) $|Y| \leq a$ a.s.;

(iii) $\mathbb{E}(Y|X, X') = S_*(X, X')$, where $S_* \in \mathcal{S}_{r,\rho}$.

For $P \in \mathcal{P}_{r,\rho,a}$, denote $S_P(u, v) := \mathbb{E}_P(Y|X = u, X' = v), u, v \in V$.

Recall that $\{\phi_j, j = 1, \dots, m\}$ are the eigenfunctions of W orthonormal in the space

$(\mathbb{R}^V, \langle\cdot,\cdot\rangle)$. Then $\bar{\phi}_j := \sqrt{m}\phi_j, j = 1, \ldots, m$ are orthonormal in $L_2(\Pi)$.

We will obtain minimax lower bounds for classes of distributions $\mathcal{P}_{r,\rho,a}$ in two different cases. In the first case, we assume that for some (relatively large) value of $p \geq 2$ and some (not too large) constant $Q_p > 0$

$$\max_{1\leq j\leq m} \|\bar{\phi}_j\|^2_{L_p(\Pi)} \leq Q_p. \tag{2.1}$$

Roughly, condition (2.1) means that most of the components of vectors $\phi_j \in \mathbb{R}^V$ are uniformly small, say, $\phi_j(v) \asymp m^{-1/2}, v \in V, j = 1, \ldots, m$. In other words, the $m \times m$ matrix $(\phi_j(v))_{j=1,\ldots,m;v\in V}$ is "dense", so, we refer to this case as a *"dense case"*. The opposite case is when this matrix is *"sparse"*. Suppose, for instance, that for some (relatively small) $d \geq 1$

$$\mathrm{card}\{j : \phi_j(v) \neq 0\} \leq d, v \in V. \tag{2.2}$$

A typical example is the case when basis of eigenfunctions $\{\phi_j, j = 1, \ldots, m\}$ coincides with the canonical basis $\{e_v : v \in V\}$ of $\mathbb{R}^V$ (then, $d = 1$).

Denote $l_0 := k_0 \wedge 32$. In the *dense case*, the following theorem holds.

Theorem 3 *Suppose condition (2.1) holds. Define*

$$\delta_n^{(1)}(r,\rho,a) := \max_{l_0\leq l\leq m}\left[\frac{a^2(r\wedge l)l}{n}\bigwedge\frac{\rho^2}{\lambda_l}\bigwedge\frac{1}{p-1}\frac{1}{Q_p^2}\frac{a^2(r\wedge l)}{l}\frac{1}{m^{4/p}}\right].$$

There exist constants $c_1, c_2 > 0$ such that

$$\inf_{\hat{S}_n}\sup_{P\in\mathcal{P}_{r,\rho,a}} \mathbb{P}_P\Big\{\|\hat{S}_n - S_P\|^2_{L_2(\Pi^2)} \geq c_1\delta_n^{(1)}(r,\rho,a)\Big\} \geq c_2,$$

where the infimum is taken over all the estimators $\hat{S}_n$ based on n i.i.d. copies of (X, X', Y).

In fact, it will follow from the proof that, if $\lambda_{k_0} \leq \frac{n\rho^2}{a^2(r\wedge k_0)k_0}$ (that is, the smallest nonzero eigenvalue of W is not too large), then the maximum in the definition of $\delta_n^{(1)}(r,\rho,a)$ can be extended to all $l = 1, \ldots, m$.

Corollary 1 *Suppose condition (2.1) holds with $p = \log m$. Let*

$$\delta_n^{(2)}(r,\rho,a) := \max_{l_0\leq l\leq m}\left[\frac{a^2(r\wedge l)l}{n}\bigwedge\frac{\rho^2}{\lambda_l}\bigwedge\frac{1}{Q^2_{\log m}}\frac{a^2(r\wedge l)}{l}\frac{1}{\log m}\right].$$

There exist constants $c_1, c_2 > 0$ such that

$$\inf_{\hat{S}_n}\sup_{P\in\mathcal{P}_{r,\rho,a}} \mathbb{P}_P\Big\{\|\hat{S}_n - S_P\|^2_{L_2(\Pi^2)} \geq c_1\delta_n^{(2)}(r,\rho,a)\Big\} \geq c_2.$$

Proof. Take $p = \log m$ in the statement of Theorem 3 and observe that $m^{4/p} = e^4$ and $\frac{1}{p-1} \geq \frac{1}{\log m}$.

□

It is obvious that one can replace the quantity $\delta_n^{(1)}(r, \rho, a)$ in Theorem 3 (or the quantity $\delta_n^{(2)}(r, \rho, a)$ in Corollary 1) by the following smaller quantity:

$$\delta_n^{(3)}(r, \rho, a) := \max_{l_0 \leq l \leq L} \left[\frac{a^2(r \wedge l)l}{n} \bigwedge \frac{\rho^2}{\lambda_l} \right],$$

where $L := \left[\frac{1}{Q_p m^{2/p}} \sqrt{\frac{n}{p-1}} \right] \wedge m$. Moreover, denote

$$\bar{l} := \max \left\{ l = l_0, \ldots, m : (r \vee l) l \lambda_l \leq \frac{\rho^2 n}{a^2} \right\}.$$

It is straightforward to check that

$$\max_{l_0 \leq l \leq m} \left[\frac{a^2(r \wedge l)l}{n} \bigwedge \frac{\rho^2}{\lambda_l} \right] = \frac{a^2(r \wedge \bar{l})\bar{l}}{n} \bigvee \frac{\rho^2}{\lambda_{\bar{l}+1}}$$

and, if $\bar{l} \leq L$, then $\delta_n^{(3)}(r, \rho, a) = \frac{a^2(r \wedge \bar{l})\bar{l}}{n} \bigvee \frac{\rho^2}{\lambda_{\bar{l}+1}}$.

Example. Suppose that, for some $\beta > 1/2$, $\lambda_l \asymp l^{2\beta}, l = 1, \ldots, m$ (in particular, it means that $\lambda_l \neq 0$ and $l_0 = k_0 = 1$). Then, an easy computation shows that

$$\bar{l} = (\check{l} \wedge m) \vee 1, \quad \check{l} \asymp \left(\frac{\rho^2}{a^2} \frac{n}{r} \right)^{1/(2\beta+1)} \bigwedge \left(\frac{\rho^2 n}{a^2} \right)^{1/(2\beta+2)}.$$

Let $p = \log m$ and suppose that $\max_{1 \leq j \leq m} \|\bar{\phi}_j\|^2_{L_p(\Pi)} \leq Q_p$. Take $L := \left[\frac{1}{e^2 Q_p} \sqrt{\frac{n}{\log(m/e)}} \right] \bigwedge m$. The condition $\bar{l} \leq L$ is satisfied, for instance, when either

$$e^2 Q_p \sqrt{\log(m/e)} \left(\frac{\rho^2}{a^2 r} \right)^{1/(2\beta+1)} \leq c' n^{\frac{1}{2} - \frac{1}{2\beta+1}}, \text{ or } e^2 Q_p \sqrt{\log(m/e)} \left(\frac{\rho}{a} \right)^{1/(\beta+1)} \leq c' n^{\frac{1}{2} - \frac{1}{2\beta+2}}, \tag{2.3}$$

where $c' > 0$ is a small enough constant (this, essentially, means that n is sufficiently large). Under this condition, we get the following expression for a minimax lower bound:

$$\left(\left(\frac{a^2 \rho^{1/\beta} r}{n} \right)^{2\beta/(2\beta+1)} \bigwedge \left(\frac{a^2 \rho^{2/\beta}}{n} \right)^{\beta/(\beta+1)} \bigwedge \frac{a^2 r m}{n} \right) \bigvee \frac{a^2}{n}. \tag{2.4}$$

We now turn to the *sparse case*.

Theorem 4 *Suppose condition (2.2) holds and let*

$$\delta_n^{(4)}(r,\rho,a) := \max_{l_0\le l\le m}\left[\frac{a^2(r\wedge l)l}{n}\bigwedge\frac{\rho^2}{\lambda_l}\bigwedge\frac{a^2}{d\log m}\frac{l^2}{m^2}\right].$$

There exist constants $c_1, c_2 > 0$ such that

$$\inf_{\hat S_n}\sup_{P\in\mathcal{P}_{r,\rho,a}}\mathbb{P}_P\Big\{\|\hat S_n - S_P\|^2_{L_2(\Pi^2)} \ge c_1\delta_n^{(4)}(r,\rho,a)\Big\} \ge c_2.$$

It will be clear from the upper bounds of Section 3 (see the remark after Theorem 5) that, at least in a special case when $\{\phi_j\}$ coincides with the canonical basis of $\mathbb{R}^V$, the additional term $\frac{a^2}{d\log m}\frac{l^2}{m^2}$ is correct (up to a log factor). At the same time, most likely, the "third terms" of the bounds of Theorem 3 (in the dense case) and Theorem 4 (in the sparse case) have not reached their final form yet. A more sophisticated construction of "well separated" subsets of $\mathcal{P}_{r,\rho,a}$ might be needed to achieve this goal. The main difficulty in the proof given below is related to the fact that we have to impose constraints, on the one hand, on the entries of the target matrix represented in the canonical basis and, on the other hand, on the Soblolev type norm $\|W^{1/2}S\|_{L_2(\Pi^2)}$ (for which it is convenient to use the representation in the basis of eigenfunctions of W). Due to this fact, we are using the last representation in our construction and we have to use an argument based on the properties of Rademacher sums to ensure that the entries of the matrix represented in the canonical basis are uniformly bounded by a. This is the reason why the "third terms" occur in the bounds of theorems 3 and 4. In the case, when the constraints are only on the norm $\|W^{1/2}S\|_{L_2(\Pi^2)}$ and on the variance of the noise and there are no constraints on $\|S\|_{L_\infty}$, it is much easier to prove the lower bound of the order

$$\max_{l_0\le l\le m}\left[\frac{\sigma^2(r\wedge l)l}{n}\bigwedge\frac{\rho^2}{\lambda_l}\right]$$

without any additional terms. Note, however, that the condition $\|S_*\|_{L_\infty}\le a$ is of importance in the following sections to obtain the upper bounds for penalized least squares estimators that match the lower bounds up to log factors.

Proof of Theorem 3. The proof relies on several well known facts stated below. In what follows, $K(\mu\|\nu) := -\mathbb{E}_\mu\log\frac{d\nu}{d\mu}$ denotes Kullback–Leibler divergence between two probability measures μ,ν defined on the same space and such that $\nu\ll\mu$ (that is, ν is absolutely continuous with respect to μ). We will denote by $P^{\otimes n}$ the n-fold product measure $P^{\otimes n} := P\otimes P\cdots\otimes P$. The following proposition is a version of Theorem 2.5 in Tsybakov (2009).

Proposition 1 *Let $\mathcal{P}$ be a finite set of distributions of (X, X', Y) such that the following assumptions hold:*

1. *There exists $P_0 \in \mathcal{P}$ such that for all $P \in \mathcal{P}$, $P \ll P_0$*
2. *There exists $\alpha \in (0, 1/8)$ such that*

$$\sum_{P \in \mathcal{P}} K(P_0^{\otimes n} \| P^{\otimes n}) \leq \alpha(\mathrm{card}(\mathcal{P}) - 1) \log(\mathrm{card}(\mathcal{P}) - 1)$$

3. *For all $P_1, P_2 \in \mathcal{P}$, $\|S_{P_1} - S_{P_2}\|^2_{L_2(\Pi^2)} \geq 4s^2 > 0$.*

Then, there exists a constant $\beta > 0$ such that

$$\inf_{\hat{S}_n} \max_{P \in \mathcal{P}} \mathbb{P}_P\{\|\hat{S}_n - S_P\|^2_{L_2(\Pi^2)} \geq s^2\} \geq \beta > 0. \tag{2.5}$$

We will also use Varshamov–Gilbert bound stated below.

Lemma 1 (Varshamov–Gilbert bound) *Let $d \geq 8$. There exists a subset $E \subset \{-1, 1\}^d$ such that* $\mathrm{card}(E) \geq 2^{d/8} + 1$ *and*

$$\sum_{i=1}^{d} I(\sigma'_i \neq \sigma''_i) \geq d/8, \ \sigma', \sigma'' \in E, \sigma' \neq \sigma''. \tag{2.6}$$

Another well known fact we need is Sauer's lemma.

Lemma 2 (Sauer's Lemma) *Let $N \geq 1$ and let $\Lambda \subset \{-1, 1\}^N$. If*

$$\mathrm{card}(\Lambda) \geq \binom{N}{\leq k-1} := \sum_{j=0}^{k-1} \binom{N}{j},$$

then there exists $J \subset \{1, \ldots, N\}$ such that $\mathrm{card}(J) = k$ *and $\pi_J \Lambda = \{-1, 1\}^J$, where $\pi_J : \{-1, 1\}^N \mapsto \{-1, 1\}^J$, $\pi_J(t_1, \ldots, t_N) = (t_j : j \in J)$.*

Finally, we use the following elementary bound for Rademacher sums (see de la Pena and Giné (1998), p. 21).

Lemma 3 *Let $\varepsilon_1, \ldots, \varepsilon_N$ be i.i.d. Rademacher random variables (that is, $\varepsilon_j = +1$ with probability $1/2$ and $\varepsilon_j = -1$ with the same probability). Then, for all $p \geq 2$,*

$$\mathbb{E}^{1/p} \left| \sum_{j=1}^{N} \varepsilon_j t_j \right|^p \leq \sqrt{p-1} \left(\sum_{j=1}^{N} t_j^2 \right)^{1/2}, (t_1, \ldots, t_N) \in \mathbb{R}^N.$$

We will start the proof with constructing a "well separated" subset $\mathcal{P}$ of the class of distributions $\mathcal{P}_{r,\rho,a}$ that will allow us to use Proposition 1. Fix $l \leq m$, $l \geq 32$ and $\kappa > 0$. Denote $l' = [l/2], l'' = l - l'$. First assume that $r \leq l''$. Denote $R_\sigma := \kappa\Big((\sigma_{ij}) : i = 1, \ldots, l', j = 1, \ldots, r\Big)$, where $\sigma_{ij} = +1$ or $\sigma_{ij} = -1$. Let $\mathcal{R}_{l',r} = \{R_\sigma : \sigma \in \{-1,1\}^{l' \times r}\}$ (so, $\mathcal{R}_{l',r}$ is the class of all $l' \times r$ matrices with entries $+\kappa$ or $-\kappa$). Given $R \in \mathcal{R}_{l',r}$, let

$$\tilde{R} := \left(\begin{array}{cccc} R & R & \ldots & R & O_{l',l^*} \end{array} \right)$$

be the $l' \times l''$ matrix that consists of $[l''/r]$ blocks R and the last block O_{l',l^*}, where $l^* := l'' - [l''/r]r$ and O_{k_1,k_2} is the $k_1 \times k_2$ zero matrix. Finally, define the following symmetric $m \times m$ matrix:

$$R^\diamondsuit := \left(\begin{array}{ccc} O_{l',l'} & \tilde{R} & O_{l',m-l} \\ \tilde{R}^T & O_{l'',l''} & O_{l'',m-l} \\ O_{m-l,l'} & O_{m-l,l''} & O_{m-l,m-l} \end{array} \right).$$

Now, given $\sigma \in \{-1,1\}^{l' \times r}$, define a symmetric kernel $K_\sigma : V \times V \mapsto \mathbb{R}$:

$$K_\sigma := \sum_{i,j=1}^{m} (R_\sigma^\diamondsuit)_{ij} (\phi_i \otimes \phi_j).$$

It is easy to see that

$$K_\sigma(u,v) = \kappa \sum_{i=1}^{l'} \sum_{j=1}^{r} \sigma_{ij} \phi_i(u) \sum_{k=0}^{[l''/r]-1} \phi_{l'+rk+j}(v) + \kappa \sum_{i=1}^{r} \sum_{j=1}^{l'} \sigma_{ji} \sum_{k=0}^{[l''/r]-1} \phi_{l'+rk+i}(u) \phi_j(v). \tag{2.7}$$

Consider the following set: $\Lambda := \{\sigma \in \{-1,1\}^{l' \times r} : \max_{u,v \in V} |K_\sigma(u,v)| \leq a\}$. We will show that, if κ is sufficiently small (its precise value to be specified later), then the set Λ contains at least three quarters of the points of the combinatorial cube $\{-1,1\}^{l' \times r}$. To this end, define $\xi := \max_{u,v \in V} |K_\varepsilon(u,v)|$, where $\varepsilon \in \{-1,1\}^{l' \times r}$ is a random vector with i.i.d. Rademacher components. Assume, in addition, that ε and (X, X') are independent. It is enough to show that $\xi \leq a$ with probability at least $3/4$. We have

$$\mathbb{P}\{\xi \geq a\} \leq \sum_{u,v \in V} \mathbb{P}\{|K_\varepsilon(u,v)| \geq a\} = m^2 \mathbb{E}\mathbb{P}\{|K_\varepsilon(X,X')| \geq a | X, X'\} =$$

$$m^2 \mathbb{P}\{|K_\varepsilon(X,X')| \geq a\} \leq \frac{m^2 \mathbb{E}|K_\varepsilon(X,X')|^p}{a^p}. \tag{2.8}$$

We will use Lemma 3 to control $\mathbb{E}(|K_\varepsilon(X,X')|^p|X,X')$ (recall that $K_\varepsilon(u,v), u,v \in V$ is a Rademacher sum). By representation (2.7), $K_\varepsilon(u,v) = K_\varepsilon^{'}(u,v) + K_\varepsilon^{'}(v,u)$, where

$$K_\varepsilon^{'}(u,v) = \kappa \sum_{i=1}^{l'} \sum_{j=1}^{r} \varepsilon_{ij} \phi_i(u) \sum_{k=0}^{[l''/r]-1} \phi_{l'+rk+j}(v).$$

Denote

$$\tau^2(u,v) := \sum_{i=1}^{l'} \sum_{j=1}^{r} \phi_i^2(u) \Big(\sum_{k=0}^{[l''/r]-1} \phi_{l'+rk+j}(v) \Big)^2.$$

Observe that

$$\tau^2(u,v) \le \frac{l''}{r} q(l',u) q(l'',v) \le q(l,u) q(l,v) \frac{l}{r},$$

where $q(l,u) := \sum_{j=1}^{l} \phi_j^2(u), u \in V$, and we used the bound

$$\Big(\sum_{k=0}^{[l''/r]-1} \phi_{l'+rk+j}(v) \Big)^2 \le \frac{l''}{r} \sum_{k=0}^{[l''/r]-1} \phi_{l'+rk+j}^2(v). \tag{2.9}$$

Thus, applying Lemma 3 to the Rademacher sum $K_\varepsilon^{'}$, we get

$$\mathbb{E}|K_\varepsilon(u,v)|^p \le 2^{p-1} \Big(\mathbb{E}|K_\varepsilon^{'}(u,v)|^p + \mathbb{E}|K_\varepsilon^{'}(v,u)|^p \Big) \le$$

$$2^p (p-1)^{p/2} \kappa^p (\tau^2(u,v) \vee \tau^2(v,u))^{p/2} \le 2^p (p-1)^{p/2} \kappa^p q^{p/2}(l,u) q^{p/2}(l,v) \Big(\frac{l}{r} \Big)^{p/2}.$$

Given $p \in [2, +\infty]$, denote

$$Q_p(l) := \Big\| l^{-1} q(l,\cdot) \Big\|_{L_{p/2}(\Pi)} = \Big\| \frac{1}{l} \sum_{j=1}^{l} \bar{\phi}_j^2 \Big\|_{L_{p/2}(\Pi)}, \ l = 1, \dots, m.$$

This yields

$$\begin{aligned}
&\mathbb{E}|K_\varepsilon(X,X')|^p = \mathbb{E}\mathbb{E}(|K_\varepsilon(X,X')|^p|X,X') \le \\
&2^p (p-1)^{p/2} \kappa^p \Big(\frac{l}{r} \Big)^{p/2} \mathbb{E}(q^{p/2}(l,X) q^{p/2}(l,X')) = \\
&2^p (p-1)^{p/2} \kappa^p \Big(\frac{l}{r} \Big)^{p/2} (\mathbb{E} q^{p/2}(l,X))^2 \\
&2^p (p-1)^{p/2} \kappa^p \Big(\frac{l}{r} \Big)^{p/2} \Big(\frac{l}{m} \Big)^p Q_p^p(l).
\end{aligned}$$

Substituting the last bound into (2.8), we get

$$\mathbb{P}\{\xi \ge a\} \le \frac{m^2 \mathbb{E}|K_\varepsilon(X,X')|^p}{a^p} \le m^2 2^p (p-1)^{p/2} \frac{\kappa^p}{a^p} \Big(\frac{l}{r} \Big)^{p/2} \Big(\frac{l}{m} \Big)^p Q_p^p(l).$$

Now, to get $\mathbb{P}\{\xi \geq a\} \leq 1/4$, it is enough to take

$$\kappa \leq 2^{-(1+2/p)}(p-1)^{-1/2}\frac{1}{Q_p(l)}\frac{m}{l}\frac{a\sqrt{r}}{\sqrt{l}}\frac{1}{m^{2/p}}. \tag{2.10}$$

Next observe that

$$\mathrm{card}(\Lambda) \geq \frac{3}{4}2^{l'r} > \binom{l'r}{\leq [l'r/2]} = \sum_{k=0}^{[l'r/2]}\binom{l'r}{k}.$$

It follows from Sauer's Lemma (see Lemma 2) that there exists a subset $J \subset \{(i,j) : 1 \leq i \leq l', 1 \leq j \leq r\}$ with $\mathrm{card}(J) = [l'r/2]+1$ and such that $\pi_J(\Lambda) = \{-1,1\}^J$, where $\pi_J : \{-1,1\}^{l'\times r} \mapsto \{-1,1\}^J$, $\pi_J\Big(\sigma_{ij} : i = 1,\dots,l', j = 1,\dots,r\Big) = \Big(\sigma_{ij} : (i,j) \in J\Big)$. Since $l \geq 32$, we have $l'r \geq 16$ and $\mathrm{card}(J) \geq 8$. We can now apply Varshamov-Gilbert bound (see Lemma 1) to the combinatorial cube $\{-1,1\}^J$ to prove that there exists a subset $E \subset \{-1,1\}^J$ such that $\mathrm{card}(E) \geq 2^{l'r/16}+1$ and, for all $\sigma', \sigma'' \in E, \sigma' \neq \sigma''$, $\sum_{(i,j)\in J} I(\sigma_{ij}^{'} \neq \sigma_{ij}^{''}) \geq \frac{l'r}{16}$. It is now possible to choose a subset Λ' of Λ such that $\mathrm{card}(\Lambda') = \mathrm{card}(E)$ and $\pi_J(\Lambda') = E$. Then, we have $\mathrm{card}(\Lambda') \geq 2^{l'r/16}+1$ and

$$\sum_{i=1}^{l'}\sum_{j=1}^{r} I(\sigma_{ij}^{'} \neq \sigma_{ij}^{''}) \geq \frac{l'r}{16}, \tag{2.11}$$

for all $\sigma', \sigma'' \in \Lambda', \sigma' \neq \sigma''$.

We are now in a position to define the set of distributions $\mathcal{P}$. For $\sigma \in \Lambda'$, denote by P_σ the distribution of (X, X', Y) such that

(i) (X, X') is uniform in $V \times V$;

(ii) conditional distribution of Y given (X, X') is defined as follows:

$$\mathbb{P}_{P_\sigma}\{Y = +a|X, X'\} = p_\sigma(X, X') = 1/2 + K_\sigma(X, X')/8a,$$

$$\mathbb{P}_{P_\sigma}\{Y = -a|X, X'\} = 1 - p_\sigma(X, X') = 1/2 - K_\sigma(X, X')/8a.$$

Since $|K_\sigma(X, X')| \leq a$ for all $\sigma \in \Lambda'$, we have $p_\sigma(X, X') \in [3/8, 5/8], \sigma \in \Lambda$. Denote $\mathcal{P} := \{P_\sigma : \sigma \in \Lambda'\}$. For $P = P_\sigma \in \mathcal{P}$, we have

$$S_P(u, v) = \mathbb{E}(Y|X = u, X' = v) = \frac{1}{4}K_\sigma(u, v).$$

Note that $\mathrm{rank}(S_P) = \mathrm{rank}(K_\sigma) = \mathrm{rank}(R_\sigma^\diamondsuit) \leq r$ (see the definitions of K_σ and $R_\sigma^\diamondsuit$). Moreover, we have

$$\|W^{1/2}K_\sigma\|_2^2 = \left\|W^{1/2}\sum_{i,j=1}^{m}(R_\sigma^\diamondsuit)_{ij}(\phi_i \otimes \phi_j)\right\|_2^2 = \sum_{i,j=1}^{l}\lambda_i(R_\sigma^\diamondsuit)_{ij}^2 \leq \lambda_l\|K_\sigma\|_2^2$$

and

$$\|K_\sigma\|_2^2 = \left\| \kappa \sum_{i=1}^{l'} \sum_{j=1}^{r} \sigma_{ij} \sum_{k=0}^{[l''/r]-1} \phi_i \otimes \phi_{l'+rk+j} + \kappa \sum_{i=1}^{r} \sum_{j=1}^{l'} \sigma_{ji} \sum_{k=0}^{[l''/r]-1} \phi_{l'+rk+i} \otimes \phi_j \right\|_2^2$$

$$\leq 2\kappa^2 l' r [l''/r] \leq \kappa^2 l^2.$$

Therefore, $\|W^{1/2} K_\sigma\|^2_{L_2(\Pi^2)} \leq \lambda_l \kappa^2 \frac{l^2}{m^2}$, so, we have

$$\|W^{1/2} S_{P_\sigma}\| = \frac{1}{16} \|W^{1/2} K_\sigma\|^2_{L_2(\Pi^2)} \leq \rho^2, \tag{2.12}$$

provided that

$$\kappa \leq \frac{m}{l} \frac{4\rho}{\sqrt{\lambda_l}}. \tag{2.13}$$

We can conclude that, for all $P \in \mathcal{P}$, $S_P \in \mathcal{S}_{r,\rho}$ provided that κ satisfies conditions (2.10) and (2.13). Since also $|Y| \leq a$, we have that $\mathcal{P} \subset \mathcal{P}_{r,\rho,a}$.

Next we check that $\mathcal{P}$ satisfies the conditions of Proposition 1. It is easy to see that, for all $\sigma, \sigma' \in \Lambda'$ $P_{\sigma'} \ll P_\sigma$ and

$$K(P_\sigma \| P_{\sigma'}) = \mathbb{E}\left(p_\sigma(X, X') \log \frac{p_\sigma(X, X')}{p_{\sigma'}(X, X')} + (1 - p_\sigma(X, X')) \log \frac{1 - p_\sigma(X, X')}{1 - p_{\sigma'}(X, X')} \right).$$

Using the following elementary inequality $-\log(1+u) \leq -u + u^2, |u| \leq 1/2$ and the fact that $p_\sigma(X, X') \in [3/8, 5/8], \sigma \in \Lambda$, we get that

$$K(P_\sigma \| P_{\sigma'}) \leq \frac{6}{8^2 a^2} \|K_\sigma - K_{\sigma'}\|_{L_2(\Pi^2)} \leq \frac{1}{10 a^2 m^2} \|K_\sigma - K_{\sigma'}\|_2^2, \sigma, \sigma' \in \Lambda'.$$

A simple computation based on the definition of $K_\sigma, K_{\sigma'}$ easily yields that

$$\|K_\sigma - K_{\sigma'}\|_2^2 \leq 8\kappa^2 l' r [l''/r] \leq 8\kappa^2 l' l'' \leq 4\kappa^2 l^2.$$

Thus, for the n-fold product-measures $P_\sigma^{\otimes n}, P_{\sigma'}^{\otimes n}$, we get

$$K(P_\sigma^{\otimes n} \| P_{\sigma'}^{\otimes n}) = n K(P_\sigma \| P_{\sigma'}) \leq \frac{4n\kappa^2}{10a^2} \frac{l^2}{m^2}.$$

For a fixed $\sigma \in \Lambda'$, this yields

$$\frac{1}{\mathrm{card}(\Lambda') - 1} \sum_{\sigma' \in \Lambda'} K(P_\sigma^{\otimes n} \| P_{\sigma'}^{\otimes n}) \leq \frac{4n\kappa^2}{10a^2} \frac{l^2}{m^2} \leq \frac{1}{10} \frac{l' r}{16} \leq \frac{1}{10} \log(\mathrm{card}(\Lambda') - 1), \tag{2.14}$$

provided that

$$\kappa \leq \frac{1}{16} a \frac{m}{l} \sqrt{\frac{rl}{n}}. \tag{2.15}$$

It remains to use (2.11) and the definition of kernels K_σ to bound from below the squared distance $\|K_\sigma - K_{\sigma'}\|^2_{L_2(\Pi^2)}$ for $\sigma, \sigma' \in \Lambda', \sigma \neq \sigma'$:

$$\|K_\sigma - K_{\sigma'}\|^2_{L_2(\Pi^2)} = m^{-2}\|K_\sigma - K_{\sigma'}\|_2^2 \geq 4m^{-2}\kappa^2 \frac{l'r}{16}[l''/r] \geq \frac{1}{64}\kappa^2 \frac{l^2}{m^2}.$$

Since $S_{P_\sigma} = \frac{1}{4}K_\sigma$, this implies that

$$\|S_P - S_{P'}\|^2_{L_2(\Pi^2)} \geq 2^{-10}\kappa^2 \frac{l^2}{m^2}. \tag{2.16}$$

In view of (2.10), (2.15) and (2.13), we now take

$$\kappa := \frac{1}{16} a \frac{m}{l} \sqrt{\frac{rl}{n}} \bigwedge \frac{m}{l} \frac{4\rho}{\sqrt{\lambda_l}} \bigwedge 2^{-(1+2/p)}(p-1)^{-1/2} \frac{1}{Q_p(l)} \frac{m}{l} \frac{a\sqrt{r}}{\sqrt{l}} \frac{1}{m^{2/p}}.$$

With this choice of κ, $\mathcal{P} := \{P_\sigma : \sigma \in \Lambda'\} \subset \mathcal{P}_{r,a,\rho}$. In view of (2.16) and (2.14), we can use Proposition 1 to get

$$\inf_{\hat{S}} \sup_{P \in \mathcal{P}_{r,a,\rho}} \mathbb{P}_P\Big\{\|\hat{S} - S_P\|^2_{L_2(\Pi^2)} \geq c_1\delta_n\Big\} \geq \inf_{\hat{S}} \sup_{P \in \mathcal{P}} \mathbb{P}_P\Big\{\|\hat{S} - S_P\|^2_{L_2(\Pi^2)} \geq c_1\delta_n\Big\} \geq c_2, \tag{2.17}$$

where

$$\delta_n := \frac{a^2 rl}{n} \bigwedge \frac{\rho^2}{\lambda_l} \bigwedge \frac{1}{p-1} \frac{1}{Q_p^2(l)} \frac{a^2 r}{l} \frac{1}{m^{4/p}}$$

and $c_1, c_2 > 0$ are constants.

In the case when $r > l''$, bound (2.17) still holds with

$$\delta_n := \frac{a^2 l^2}{n} \bigwedge \frac{\rho^2}{\lambda_l} \bigwedge \frac{1}{p-1} \frac{a^2}{Q_p^2(l)} \frac{1}{m^{4/p}}.$$

The proof is an easy modification of the argument in the case when $r \leq l''$. For $r > l''$, the construction becomes simpler: namely, we define

$$R^\flat := \begin{pmatrix} O_{l',l'} & R & O_{l',m-l} \\ R^T & O_{l'',l''} & O_{l'',m-l} \\ O_{m-l,l'} & O_{m-l,l''} & O_{m-l,m-l} \end{pmatrix},$$

where $R \in \mathcal{R}_{l',l''}$, and, based on this, redefine kernels $K_\sigma, \sigma \in \{-1,1\}^{l' \times l''}$. The proof then goes through with minor simplifications.

Thus, in both cases $r > l''$ and $r \leq l''$, (2.17) holds with

$$\delta_n = \delta_n(l) := \frac{a^2(r \wedge l)l}{n} \bigwedge \frac{\rho^2}{\lambda_l} \bigwedge \frac{1}{p-1} \frac{1}{Q_p^2(l)} \frac{a^2(r \wedge l)}{l} \frac{1}{m^{4/p}}.$$

This is true under the assumption that $l \geq 32$. Note also that $Q_p(l) \leq \max_{1\leq j\leq m} \|\bar{\phi}_j\|^2_{L_p(\Pi)} \leq Q_p$. Thus, we can replace $Q_p^2(l)$ by the upper bound Q_p^2 in the definition of $\delta_n(l)$.

We can now choose $l \in \{32, \dots, m\}$ that maximizes $\delta_n(l)$ to get bound (2.17) with $\delta_n := \min_{32\leq l\leq m} \delta_n(l)$. This completes the proof in the case when $k_0 \geq 32$ and $l_0 = 32$. If $k_0 < 32$, it is easy to use the condition $\lambda_{l+1} \leq c\lambda_l, l \geq k_0$ and to show that

$$\min_{32\leq l\leq m} \delta_n(l) \leq c' \min_{k_0\leq l\leq m} \delta_n(l),$$

where c' is a constant depending only on c. This completes the proof in the remaining case.

□

Proof of Theorem 4. The only modification of the previous proof is to replace bound (2.9) by

$$\Big(\sum_{k=0}^{[l''/r]-1} \phi_{l'+rk+j}(v)\Big)^2 \leq d \sum_{k=0}^{[l''/r]-1} \phi^2_{l'+rk+j}(v).$$

Then, the outcome of the next several lines of the proof is that $\mathbb{P}\{\xi \geq a\} \leq 1/4$ provided that (instead of (2.10))

$$\kappa \leq 2^{-(1+2/p)}(p-1)^{-1/2} \frac{1}{Q_p(l)} \frac{m}{l} \frac{a}{\sqrt{d}} \frac{1}{m^{2/p}}.$$

As a result, at the end of the proof, we get that (2.17) holds with

$$\delta_n = \delta_n(l) := \frac{a^2(r \wedge l)l}{n} \bigwedge \frac{\rho^2}{\lambda_l} \bigwedge \frac{1}{p-1} \frac{1}{Q_p^2(l)} \frac{a^2}{d} \frac{1}{m^{4/p}}.$$

It remains to observe that $Q_p(l) \leq \frac{m}{l}$, which follows from the fact that

$$\sum_{j=1}^{l} \phi_j^2(v) = \sum_{j=1}^{l} \langle \phi_j, e_v\rangle^2 \leq \sum_{j=1}^{m} \langle \phi_j, e_v\rangle^2 = 1, v \in V,$$

and to take $p = \log m$ to complete the proof.

□

3 Least Squares Estimators with Nonconvex Penalties

In this section, we derive upper bounds on the squared $L_2(\Pi^2)$-error of the following least squares estimator of the target matrix S_* :

$$\hat{S}_l := \hat{S}_{r,l,a} := \operatorname{argmin}_{S\in\bar{\mathcal{S}}_r(l;a)} \frac{1}{n} \sum_{j=1}^{n} (Y_j - S(X_j, X'_j))^2, \tag{3.1}$$

where $\bar{\mathcal{S}}_r(l;a) := \{S^a : S \in \mathcal{S}_r(l;a)\}, l = 1, \dots, m$,

$$\mathcal{S}_r(l;a) := \left\{ S : S \in \mathcal{S}_V, \operatorname{rank}(S) \le r, \|S\|_{L_2(\Pi^2)} \le a, S = \sum_{i,j=1}^{l} s_{ij}(\phi_i \otimes \phi_j) \right\}. \quad (3.2)$$

Here S^a denotes a truncation of kernel S : $S^a(u,v) = S(u,v)$ if $|S(u,v)| \le a$, $S^a(u,v) = a$ if $S(u,v) > a$ and $S^a(u,v) = -a$ if $S(u,v) < -a$. Note that the kernels in the class $\mathcal{S}_r(l;a)$ are symmetric and $\operatorname{rank}(S) \le r \wedge l, S \in \mathcal{S}_r(l;a)$. Note also that the sets $\mathcal{S}_r(l;a)$, $\bar{\mathcal{S}}_r(l;a)$ and optimization problem (3.1) are not convex. We will prove the following result under the assumption that $|Y| \le a$ a.s. Recall the definition of the class of kernels $\mathcal{S}_{r,\rho}$ in Section 2.

Theorem 5 *There exist constants $C > 0, A > 0$ such that, for all $t > 0$, with probability at least $1 - e^{-t}$*

$$\|\hat{S}_l - S_*\|^2_{L_2(\Pi^2)} \le 2 \inf_{S \in \bar{\mathcal{S}}_r(l;a)} \|S - S_*\|^2_{L_2(\Pi^2)} + C\left(\frac{a^2(r \wedge l)l}{n} \log\left(\frac{Anm}{(r \wedge l)l} \right) + \frac{a^2 t}{n} \right). \quad (3.3)$$

In particular, for some constants $C, A > 0$, for $S_* \in \mathcal{S}_{r,\rho}$ and for all $t > 0$, with probability at least $1 - e^{-t}$,

$$\|\hat{S}_l - S_*\|^2_{L_2(\Pi^2)} \le C\left[\frac{a^2(r \wedge l)l}{n} \log\left(\frac{Anm}{(r \wedge l)l} \right) \bigvee \frac{\rho^2}{\lambda_{l+1}} \bigvee \frac{a^2 t}{n} \right]. \quad (3.4)$$

Proof. Without loss of generality, assume that $a = 1$; this would imply the general case by a simple rescaling of the problem. We will use a version of well known bounds for least squares estimators over uniformly bounded function classes in terms of Rademacher complexities. Specifically, consider the following least squares estimator:

$$\hat{g} := \operatorname{argmin}_{g \in \mathcal{G}} n^{-1} \sum_{j=1}^{n} (Y_j - g(X_j))^2,$$

where $(X_1, Y_1), \dots, (X_n, Y_n)$ are i.i.d. copies of a random couple (X, Y) in $T \times \mathbb{R}$, $(T, \mathcal{T})$ being a measurable space, $|Y| \le 1$ a.s., $\mathcal{G}$ being a class of measurable functions on T uniformly bounded by 1. The goal is to estimate the regression function $g_*(x) := \mathbb{E}(Y|X = x)$. Define localized Rademacher complexity

$$\psi_n(\delta) := \mathbb{E} \sup_{g_1, g_2 \in \mathcal{G}, \|g_1 - g_2\|^2_{L_2(\Pi)} \le \delta} |R_n(g_1 - g_2)|,$$

where Π is the distribution of X and $R_n(g) := n^{-1} \sum_{j=1}^{n} \varepsilon_j g(X_j)$ is the Rademacher process, $\{\varepsilon_j\}$ being a sequence of i.i.d. random variables independent of $\{X_j\}$. Denote

$\psi_n^\flat(\delta) := \sup_{\sigma \geq \delta} \frac{\psi_n(\sigma)}{\sigma}$ and $\psi_n^\sharp(\varepsilon) := \inf\{\delta > 0 : \psi_n(\delta) \leq \varepsilon\}$. The next result easily follows from Theorem 5.2 in Koltchinskii (2011b):

Proposition 2 *There exist constants $c_1, c_2 > 0$ such that, for all $t > 0$, with probability at least $1 - e^{-t}$*

$$\|\hat{g} - g_*\|^2_{L_2(\Pi)} \leq 2 \inf_{g \in \mathcal{G}} \|g - g_*\|^2_{L_2(\Pi)} + c_1 \left(\psi_n^\sharp(c_2) + \frac{t}{n} \right).$$

We will apply this proposition to prove Theorem 5. In what follows in the proof, denote $\hat{S} := \hat{S}_l$. In our case, $T = V \times V$, (X, X') plays the role of X and Π^2 plays the role of Π. Let $\mathcal{G} := \bar{\mathcal{S}}_r(l; 1)$, $g_* = S_*$ and $\hat{g} = \hat{S}$. First, we need to upper bound the Rademacher complexity $\psi_n(\delta)$ for the class $\mathcal{G}$. Let $\mathbb{S}_{r,m}(R)$ be the set of all symmetric $m \times m$ matrices S with $\mathrm{rank}(S) \leq r$ and $\|S\|_2 \leq R$. The ε-covering number $N(\mathbb{S}_{r,m}(R); \|\cdot\|_2; \varepsilon)$ of the set $\mathbb{S}_{r,m}(R)$ with respect to the Hilbert–Schmidt distance (that is, the minimal number of balls of radius ε needed to cover this set) can be bounded as follows:

$$N(\mathbb{S}_{r,m}(R); \|\cdot\|_2; \varepsilon) \leq \left(\frac{18R}{\varepsilon} \right)^{(m+1)r}. \tag{3.5}$$

Such bounds are well known (see, e.g., Koltchinskii (2011b), Lemma 9.3 and references therein; the proof of this lemma can be easily modified to obtain (3.5)). Bound (3.5) will be used to control the covering numbers of the set of kernels $\mathcal{S}_r(l; 1)$. This set can be easily identified with a subset of the set $\mathbb{S}_{r \wedge l, l}(m)$ (since kernels $S \in \mathcal{S}_r(l; 1)$ can be viewed as symmetric $l \times l$ matrices of rank at most $r \wedge l$ with $\|S\|_{L_2(\Pi^2)} \leq 1$ and $\|S\|_2 = m\|S\|_{L_2(\Pi^2)} \leq m$). Therefore, we get the following bound:

$$N(\mathcal{S}_r(l; 1); \|\cdot\|_2; \varepsilon) \leq \left(\frac{18m}{\varepsilon} \right)^{(l+1)(r \wedge l)}.$$

Since $\|S_1^1 - S_2^1\|_2^2 \leq \|S_1 - S_2\|_2^2$ (truncation of the entries reduces the Hilbert–Schmidt distance), we also have

$$N(\bar{\mathcal{S}}_r(l; 1); \|\cdot\|_2; \varepsilon) \leq \left(\frac{18m}{\varepsilon} \right)^{(l+1)(r \wedge l)}.$$

Note that

$$\|S_1 - S_2\|^2_{L_2(\Pi_n)} = n^{-1} \sum_{j=1}^n \langle S_1 - S_2, E_{X_j, X_j'} \rangle^2 \leq n^{-1} \sum_{j=1}^n \|E_{X_j, X_j'}\|_2^2 \|S_1 - S_2\|_2^2 \leq \|S_1 - S_2\|_2^2.$$

Therefore, we get the following bound on the $L_2(\Pi_n)$-covering numbers of the set $\bar{\mathcal{S}}_r(l;1)$:

$$N(\bar{\mathcal{S}}_r(l;1); L_2(\Pi_n); \varepsilon) \le \left(\frac{18m}{\varepsilon}\right)^{(l+1)(r\wedge l)}.$$

Here Π_n denotes the empirical distribution based on observations $(X_1, X_1'), \dots, (X_n, X_n')$. The last bound allows us to use inequality (3.17) in Koltchinskii (2011b) to control the localized Rademacher complexity $\psi_n(\delta)$ of the class $\mathcal{G}$ as follows:

$$\psi_n(\delta) = \mathbb{E} \sup_{S_1, S_2 \in \bar{\mathcal{S}}_r(l;1), \|S_1 - S_2\|^2_{L_2(\Pi^2)} \le \delta} \left| n^{-1} \sum_{j=1}^n \varepsilon_j (S_1(X_j, X_j') - S_2(X_j, X_j')) \right|$$

$$\le C_1 \left[\sqrt{\frac{\delta l(r\wedge l)}{n}} \sqrt{\log\left(\frac{Am}{\sqrt{\delta}}\right)} \bigvee \frac{l(r\wedge l)}{n} \log\left(\frac{Am}{\sqrt{\delta}}\right) \right] \tag{3.6}$$

with some constant $A, C_1 > 0$. This easily yields

$$\psi_n^\sharp(c_2) \le C_2 \frac{(r\wedge l)l}{n} \log\left(\frac{Anm}{(r\wedge l)l}\right)$$

with some constants $A, C_2 > 0$. Proposition 2 now implies bound (3.3).

To prove bound (3.4), it is enough to observe that, for $S_* \in \mathcal{S}_{r,\rho}$,

$$\inf_{S \in \bar{\mathcal{S}}_r(l;1)} \|S - S_*\|^2_{L_2(\Pi^2)} \le \frac{2\rho^2}{\lambda_{l+1}}. \tag{3.7}$$

Indeed, since $S_* \in \mathcal{S}_{r,\rho}$, we can approximate this kernel by $S_l := \sum_{i,j=1}^l \langle S_* \phi_i, \phi_j\rangle (\phi_i \otimes \phi_j)$. For the error of this approximation, we have

$$\|S_l - S_*\|^2_{L_2(\Pi^2)} = m^{-2}\|S_l - S_*\|_2^2 = m^{-2} \sum_{i\vee j > l} \langle S_* \phi_i, \phi_j\rangle^2 \le$$

$$m^{-2} \frac{1}{\lambda_{l+1}} \sum_{i>l} \sum_{j=1}^m \lambda_i \langle S_* \phi_i, \phi_j\rangle^2 + m^{-2} \frac{1}{\lambda_{l+1}} \sum_{i=1}^m \sum_{j>l} \lambda_j \langle S_* \phi_i, \phi_j\rangle^2 \le \frac{2\rho^2}{\lambda_{l+1}},$$

which implies $\|S_l^1 - S_*\|^2_{L_2(\Pi)} \le \|S_l - S_*\|^2_{L_2(\Pi^2)} \le \frac{2\rho^2}{\lambda_{l+1}}$ (since the entries of matrix S_* are bounded by 1 and truncation of the entries reduces the Hilbert–Schmidt distance). We also have $\mathrm{rank}(S_l) \le \mathrm{rank}(S_*) \le r$ and

$$\|S_l\|_{L_2(\Pi^2)} = m^{-1}\|S_l\|_2 \le m^{-1}\|S_*\|_2 = \|S_*\|_{L_2(\Pi^2)} \le \|S_*\|_{L_\infty} \le 1.$$

Therefore, $S_l^1 \in \bar{\mathcal{S}}_r(l;1)$ and bound (3.7) follows. Bound (3.4) is a consequence of (3.3) and (3.7).

□

Remark. Note that, in the case when the basis of eigenfunctions $\{\phi_j\}$ coincides with the canonical basis of space $\mathbb{R}^V$, the following bound holds trivially:

$$\|\hat{S}_l - S_*\|^2_{L_2(\Pi^2)} \le \frac{4a^2l^2}{m^2} + \frac{2\rho^2}{\lambda_{l+1}} \tag{3.8}$$

This follows from the fact that the entries of both matrices $\hat{S}_l$ and S_l are bounded by a and their nonzero entries are only in the first l rows and the first l columns, so, $\|\hat{S}_l - S_l\|^2_{L_2(\Pi^2)} \le \frac{4a^2l^2}{m^2}$. Combining this with (3.4) and minimizing the resulting bound with respect to l yields the following upper bound (up to a constant) that holds for the optimal choice of l :

$$\min_{1\le l\le m}\left[\left(\frac{a^2(r\wedge l)l}{n}\log\left(\frac{Anm}{(r\wedge l)l}\right)\bigwedge\frac{a^2l^2}{m^2}\right)\bigvee\frac{\rho^2}{\lambda_{l+1}}\right]\bigvee\frac{a^2t}{n}.$$

It is not hard to check that, typically, this expression is of the same order (up to log factors) as the lower bound of Theorem 4 for $d = 1$.

Next we consider a penalized version of least squares estimator which is adaptive to unknown parameters of the problem (such as the rank of the target matrix and the optimal value of parameter l which minimizes the error bound of Theorem 5). We still assume that $|Y| \le a$ a.s. for some known constant $a > 0$. Define

$$(\hat{r},\hat{l}) := \mathrm{argmin}_{r,l=1,\dots,m}\left\{n^{-1}\sum_{j=1}^n(Y_j - \hat{S}_{r,l,a}(X_j,X_j'))^2 + K\frac{a^2(r\wedge l)l}{n}\log\left(\frac{Anm}{(r\wedge l)l}\right)\right\}$$

and let $\hat{S} := \hat{S}_{\hat{r},\hat{l},a}$. Here $K > 0$ and $A > 0$ are fixed constants.

The following theorem provides an oracle inequality for the estimator $\hat{S}$.

Theorem 6 *There exists a choice of constants $K > 0$, $A > 0$ in (3.9) and $C > 0$ in the inequality below such that for all $t > 0$ with probability at least $1 - e^{-t}$*

$$\|\hat{S} - S_*\|^2_{L_2(\Pi^2)} \le 2\min_{1\le r\le m,1\le l\le m}\left[\inf_{S\in\mathcal{S}_r(l;a)}\|S - S_*\|^2_{L_2(\Pi^2)} + C\left(\frac{a^2(r\wedge l)l}{n}\log\left(\frac{Anm}{(r\wedge l)l}\right) + \frac{a^2(t+\log m)}{n}\right)\right]. \tag{3.9}$$

Proof. As in the proof of the previous theorem, we can assume that $a = 1$; the general case follows by rescaling. We will use oracle inequalities in abstract penalized empirical risk minimization problems (see Koltchinskii (2011b), Theorem 6.5). We only

sketch the proof here skipping the details that are standard. As in the proof of Theorem 5, first consider i.i.d. copies $(X_1, Y_1), \dots, (X_n, Y_n)$ of a random couple (X, Y) in $T \times \mathbb{R}$, where $(T, \mathcal{T})$ is a measurable space and $|Y| \leq 1$ a.s.. Let $\{\mathcal{G}_k : k \in I\}$ be a finite family of classes of measurable functions from T into $[-1, 1]$. Consider the corresponding family of least squares estimators

$$\hat{g}_k := \mathrm{argmin}_{g \in \mathcal{G}_k} n^{-1} \sum_{j=1}^{n} (Y_j - g(X_j))^2, k \in I.$$

Suppose the following upper bounds on localized Rademacher complexities for classes $\mathcal{G}_k, k \in I$ hold:

$$\mathbb{E} \sup_{g_1, g_2 \in \mathcal{G}_k, \|g_1 - g_2\|^2_{L_2(\Pi)} \leq \delta} |R_n(g_1 - g_2)| \leq \psi_{n,k}(\delta), \delta > 0,$$

where $\psi_{n,k}$ are nondecreasing functions of δ that do not depend on the distribution of (X, Y). Let

$$\hat{k} := \mathrm{argmin}_{k \in I} \left[n^{-1} \sum_{j=1}^{n} (Y_j - \hat{g}_k(X_j))^2 + K \left(\psi^{\sharp}_{n,k}(c_1) + \frac{t_k}{n} \right) \right], \tag{3.10}$$

and K, c_1 are constants and $\{t_k, k \in I\}$ are positive numbers. Define the following penalized least squares estimator of the regression function $g_* : \hat{g} := \hat{g}_{\hat{k}}$.

The next result is essentially due to Massart (2000). It can be also deduced from Theorem 6.5 in Koltchinskii (2011b).

Proposition 3 *There exists constants $K, c_1 > 0$ in the definition (3.10) of $\hat{k}$ and a constant $K_1 > 0$ such that, for all $t_k > 0$, with probability at least $1 - \sum_{k \in I} e^{-t_k}$*

$$\|\hat{g} - g_*\|^2_{L_2(\Pi)} \leq 2 \inf_{k \in I} \left[\inf_{g \in \mathcal{G}_k} \|g - g_*\|^2_{L_2(\Pi)} + K_1 \left(\psi^{\sharp}_{n,k}(c) + \frac{t_k}{n} \right) \right].$$

We apply this result to the estimator $\hat{S} = \hat{S}_{\hat{r}, \hat{l}, 1}$, where $(\hat{r}, \hat{l})$ is defined by (3.9) (with $a = 1$). In this case, $T = V \times V$, (X, X') plays the role of X, $g_* = S_*$, $I = \{(r, l) : 1 \leq r, l \leq m\}$, $\mathcal{G}_{r,l} = \bar{\mathcal{S}}_r(l; 1)$. In view of (3.6), we can use the following bounds on localized Rademacher complexities for these function classes

$$\psi_{n,r,l}(\delta) := C_1 \left[\sqrt{\frac{\delta l(r \wedge l)}{n}} \sqrt{\log \left(\frac{Am}{\sqrt{\delta}} \right)} \bigvee \frac{l(r \wedge l)}{n} \log \left(\frac{Am}{\sqrt{\delta}} \right) \right]$$

with some constant C_1, and we have

$$\psi^{\sharp}_{n,r,l}(c_1) \le C_2 \frac{(r \wedge l)l}{n} \log\left(\frac{Anm}{(r \wedge l)l}\right)$$

with some constant $C_2 > 0$. Define $t_{r,l} := t + 2\log m, (r,l) \in I$. This yields the bound $\sum_{(r,l)\in I} e^{-t_{r,l}} \le e^{-t}$. These considerations and Proposition 3 imply the claim of the theorem.

□

It follows from Theorem 6 that, for some constant $C > 0$ and for all $t > 0$,

$$\sup_{P\in\mathcal{P}_{r,\rho,a}} \mathbb{P}_P\left\{\|\hat{S} - S_P\|^2_{L_2(\Pi^2)} \ge C\left(\Delta_n(r,\rho,a) \bigvee \frac{a^2 t}{n}\right)\right\} \le e^{-t}, \tag{3.11}$$

where

$$\Delta_n(r,\rho,a) := \min_{1\le l\le m}\left[\frac{a^2(r\wedge l)l}{n}\log\left(\frac{Anm}{(r\wedge l)l}\right) \bigvee \frac{\rho^2}{\lambda_{l+1}}\right].$$

Denoting

$$\tilde{l} := \min\left\{l = 1,\dots,m : (r\vee l)l\lambda_{l+1}\log\left(\frac{Anm}{(r\wedge l)l}\right) \ge \frac{\rho^2 n}{a^2}\right\},$$

it is easy to see that

$$\Delta_n(r,\rho,a) = \frac{a^2(r\wedge\tilde{l})\tilde{l}}{n}\log\left(\frac{Anm}{(r\wedge\tilde{l})\tilde{l}}\right) \bigvee \frac{\rho^2}{\lambda_{\tilde{l}}}.$$

Example. Suppose that, for some $\beta > 1/2$, $\lambda_l \asymp l^{2\beta}, l = 1,\dots,m$. Under this assumption, it is easy to show that the upper bound on the squared $L_2(\Pi^2)$-error of the estimator $\hat{S}$ is of the order

$$\left(\left(\frac{a^2\rho^{1/\beta} r}{n}\log\frac{Anm}{r}\right)^{2\beta/(2\beta+1)} \bigwedge \left(\frac{a^2\rho^{2/\beta}\log(Anm)}{n}\right)^{\beta/\beta+1} \bigwedge \frac{a^2 rm\log(Anm)}{n}\right) \bigvee \frac{a^2 t}{n}$$

(in fact, the log factors can be written in a slightly better, but more complicated way). Up to the log factors, this is the same error rate as in the lower bounds of Section 2 (see (2.4)).

4 Least Squares with Convex Penalization: Combining Nuclear Norm and Squared Sobolev Norm

Our main goal in this section is to study the following penalized least squares estimator with a combination of two convex penalties:

$$\hat{S}_{\varepsilon,\bar{\varepsilon}} := \mathrm{argmin}_{S\in\mathbb{D}}\left[\frac{1}{n}\sum_{j=1}^{n}(Y_j - S(X_j,X'_j))^2 + \varepsilon\|S\|_1 + \bar{\varepsilon}\|W^{1/2}S\|^2_{L_2(\Pi^2)}\right], \tag{4.1}$$

where $\mathbb{D} \subset \mathcal{S}_V$ is a closed convex set of symmetric kernels such that

$$\|S\|_{L_\infty} := \max_{u,v \in V} |S(u,v)| \leq a, S \in \mathbb{D},$$

and $\varepsilon, \bar{\varepsilon} > 0$ are regularization parameters. The first penalty involved in (4.1) is based on the nuclear norm $\|S\|_1$ and it is used to "promote" low rank solutions. The second penalty is based on a "Sobolev type norm" $\|W^{1/2}S\|^2_{L_2(\Pi^2)}$. It is used to "promote" the smoothness of the solution on the graph.

We will derive an upper bound on the error $\|\hat{S}_{\varepsilon,\bar{\varepsilon}} - S_*\|^2_{L_2(\Pi^2)} = m^{-2}\|\hat{S}_{\varepsilon,\bar{\varepsilon}} - S_*\|_2^2$ of estimator $\hat{S}_{\varepsilon,\bar{\varepsilon}}$ in terms of spectral characteristics of the target kernel S_* and matrix W.

As before, W is a nonnegatively definite symmetric kernel with spectral representation $W = \sum_{k=1}^m \lambda_k(\phi_k \otimes \phi_k)$, where $0 \leq \lambda_1 \leq \cdots \leq \lambda_m$ are the eigenvalues of W repeated with their multiplicities and $\phi_1, \ldots, \phi_m$ are the corresponding orthonormal eigenfunctions. We will also use the decomposition of identity associated with W :

$$E(\lambda) := \sum_{\lambda_j \leq \lambda} (\phi_j \otimes \phi_j), \ \lambda \geq 0.$$

Clearly, $\{E(\lambda), \lambda \geq 0\}$ is a nondecreasing projector-valued function of λ. Despite the fact that the eigenfunctions $\{\phi_k\}$ are not uniquely defined in the case when W has multiple eigenvalues, the decomposition of identity $\{E(\lambda), \lambda \geq 0\}$ is uniquely defined (in fact, it can be rewritten in terms of spectral projectors of W). The distribution of the eigenvalues of W is characterized by the following spectral function:

$$F(\lambda) := \mathrm{tr}(E(\lambda)) = \|E(\lambda)\|_2^2 = \sum_{j=1}^m I(\lambda_j \leq \lambda), \lambda \geq 0.$$

Denote $k_0 := F(0) + 1$ (in other words, k_0 is the smallest k such that $\lambda_k > 0$). It was assumed in the Introduction that there exists a constant $c \geq 1$ such that $\lambda_{k+1} \leq c\lambda_k$ for all $k \geq k_0$.

Let $\bar{F} : \mathbb{R}_+ \mapsto \mathbb{R}_+$ be a nondecreasing function such that $F(\lambda) \leq \bar{F}(\lambda), \lambda \geq 0$, the function $\lambda \mapsto \frac{\bar{F}(\lambda)}{\lambda}$ is nonincreasing and, for some $\gamma \in (0,1)$,

$$\int_\lambda^\infty \frac{\bar{F}(s)}{s^2} ds \leq \frac{1}{\gamma}\frac{\bar{F}(\lambda)}{\lambda}, \lambda > 0.$$

Without loss of generality, we assume in what follows that $\bar{F}(\lambda) = m, \lambda \geq \lambda_m$ (otherwise, one can take the function $\bar{F}(\lambda) \wedge m$ instead). The conditions on $\bar{F}$ are satisfied if for some $\gamma \in (0,1)$, the function $\frac{\bar{F}(\lambda)}{\lambda^{1-\gamma}}$ is nonincreasing: in this case, $\frac{\bar{F}(\lambda)}{\lambda}$ is also nonincreasing and

$$\int_\lambda^\infty \frac{\bar{F}(s)}{s^2} ds = \int_\lambda^\infty \frac{\bar{F}(s)}{s^{1-\gamma}} \frac{ds}{s^{1+\gamma}} \leq \frac{\bar{F}(\lambda)}{\lambda^{1-\gamma}} \int_\lambda^\infty \frac{ds}{s^{1+\gamma}} = \frac{1}{\gamma}\frac{\bar{F}(\lambda)}{\lambda}.$$

Consider a kernel $S \in \mathcal{S}_V$ (an oracle) with spectral representation: $S = \sum_{k=1}^{r} \mu_k(\psi_k \otimes \psi_k)$, where $r = \text{rank}(S) \geq 1$, μ_k are non-zero eigenvalues of S (possibly repeated) and ψ_k are the corresponding orthonormal eigenfunctions. Denote $L = \text{supp}(S) = \text{l.s.}(\psi_1, \ldots, \psi_r)$. The following function will be used to characterize the relationship between the kernels S and W :

$$\varphi(S;\lambda) := \langle P_L, E(\lambda)\rangle := \sum_{\lambda_j \leq \lambda} \|P_L \phi_j\|^2, \lambda \geq 0. \tag{4.2}$$

It is immediate from this definition that $\varphi(S,\lambda) \leq F(\lambda) \leq \bar{F}(\lambda), \lambda \geq 0$. Note also that $\varphi(S,\lambda) = \sum_{j=1}^{m} \|P_L \phi_j\|^2 = r,\ \lambda \geq \lambda_m$. Denote by $\Psi = \Psi_{S,W}$ the set of all nondecreasing functions $\varphi : \mathbb{R}_+ \mapsto \mathbb{R}_+$ such that $\lambda \mapsto \frac{\varphi(\lambda)}{F(\lambda)}$ is nonincreasing and $\varphi(S;\lambda) \leq \varphi(\lambda), \lambda \geq 0$. It is easy to see that the class of functions $\Psi_{S,W}$ contains the smallest function (uniformly in $\lambda \geq 0$) that will be denoted by $\bar{\varphi}(S;\lambda)$ and it is given by the following expression:

$$\bar{\varphi}(S;\lambda) := \sup_{\sigma \leq \lambda} \bar{F}(\sigma) \sup_{\sigma' \geq \sigma} \frac{\varphi(S;\sigma')}{\bar{F}(\sigma')}.$$

It easily follows from this definition that $\bar{\varphi}(S,\lambda) = r, \lambda \geq \lambda_m$. Note that since the function $\frac{\bar{\varphi}(S,\lambda)}{F(\lambda)}$ is nonincreasing and it is equal to $\frac{r}{m}$ for $\lambda \geq \lambda_m$, we have

$$\bar{\varphi}(S;\lambda) \geq \frac{r}{m}\bar{F}(\lambda) \geq \frac{r}{m} F(\lambda), \lambda \geq 0. \tag{4.3}$$

Given $t > 0$ and $\tilde{\lambda} \in (0, \lambda_{k_0}]$, let $t_{n,m} := t + 3\log\left(2\log_2 n + \frac{1}{2}\log_2 \frac{\lambda_m}{\tilde{\lambda}} + 2\right)$. Suppose that, for some constant $D > 0$,

$$\varepsilon \geq Da\left(\sqrt{\frac{\log(2m)}{nm}} \bigvee \frac{\log(2m)}{n}\right). \tag{4.4}$$

Theorem 7 *There exists constants C, D depending only on c, γ such that, for all $\bar{\varepsilon} \in [0, \tilde{\lambda}^{-1}]$ with probability at least $1 - e^{-t}$,*

$$\|\hat{S}_{\varepsilon,\bar{\varepsilon}} - S_*\|^2_{L_2(\Pi^2)} \leq \inf_{S \in \mathbb{D}} \Big[\|S - S_*\|^2_{L_2(\Pi^2)} + Cm^2\varepsilon^2 \bar{\varphi}(S;\bar{\varepsilon}^{-1}) \tag{4.5}$$
$$+ \bar{\varepsilon}\|W^{1/2}S\|^2_{L_2(\Pi^2)} \Big] + C\frac{a^2 t_{n,m}}{n}.$$

Remarks. 1. Under the additional assumption that $m\log(2m) \leq n$, one can take $\varepsilon = Da\sqrt{\frac{\log(2m)}{nm}}$. In this case, the main part of the random error term in the right hand side of bound (4.5) becomes

$$Cm^2\varepsilon^2\bar{\varphi}(S;\bar{\varepsilon}^{-1}) + \bar{\varepsilon}\|W^{1/2}S\|^2_{L_2(\Pi^2)} = C'\frac{a^2\bar{\varphi}(S;\bar{\varepsilon}^{-1})m\log(2m)}{n} + \bar{\varepsilon}\|W^{1/2}S\|^2_{L_2(\Pi^2)}.$$

2. Note also that Theorem 7 holds in the case when $\bar{\varepsilon} = 0$. In this case, our method coincides with nuclear norm penalized least squares (matrix LASSO) and $\bar{\varphi}(S; \bar{\varepsilon}^{-1}) = \mathrm{rank}(S)$, so the bound of Theorem 7 becomes

$$\|\hat{S}_{\varepsilon,0} - S_*\|^2_{L_2(\Pi^2)} \le \inf_{S \in \mathbb{D}} \left[\|S - S_*\|^2_{L_2(\Pi^2)} + Cm^2\varepsilon^2 \mathrm{rank}(S) \right] + C\frac{a^2 t_{n,m}}{n}. \tag{4.6}$$

Similar oracle inequalities were proved by Koltchinskii, Lounici and Tsybakov (2011) (see Theorem 2 in the Introduction) for a modified least squares method with nuclear norm penalty (1.4). Clearly, (4.6) implies that

$$\|\hat{S}_{\varepsilon,0} - S_*\|^2_{L_2(\Pi^2)} \le Cm^2\varepsilon^2 \mathrm{rank}(S_*) + C\frac{a^2 t_{n,m}}{n}, \tag{4.7}$$

which is a version of a bound proved very recently by Klopp (2012) for the matrix LASSO with constrained L_∞-norm.

3. Suppose that

$$\mathbb{D} = \mathcal{S}(l; a) := \left\{ S : S \in \mathcal{S}_V, S = \sum_{i,j=1}^{l} s_{ij}(\phi_i \otimes \phi_j), \|S\|_{L_\infty} \le a \right\}$$

and consider the following estimator

$$\hat{S}_l := \mathrm{argmin}_{S \in \mathcal{S}(l;a)} \left[\frac{1}{n} \sum_{j=1}^{n} (Y_j - S(X_j, X_j'))^2 + \varepsilon \|S\|_1 \right].$$

Let now W be the orthogonal projection onto l.s.$\{\phi_{l+1}, \dots, \phi_m\}$ (for an orthonormal system $\{\phi_1, \dots, \phi_m\}$.) Since, for all $S \in \mathcal{S}(l; a)$, $\|W^{1/2} S\|_{L_2(\Pi^2)} = 0$, the estimator $\hat{S}_l$ coincides with $\hat{S}_{\varepsilon,\bar{\varepsilon}}$ for an arbitrary $\bar{\varepsilon} \ge 0$. It is easy to check that, for this choice of W, one can take $\bar{F}(\lambda) := (m\sqrt{\lambda} \vee l) \wedge m$, and we have $\bar{\varphi}(S; \lambda) = \sum_{j=1}^{l} \|P_L \phi_j\|^2 \bigvee \frac{r\bar{F}(\lambda)}{m}$, where $r = \mathrm{rank}(S)$. We will choose $\bar{\varepsilon} := m^2$ and $\tilde{\lambda} = m^{-2}$. Then, $\bar{\varphi}(S; \bar{\varepsilon}^{-1}) = \sum_{j=1}^{l} \|P_L \phi_j\|^2 \bigvee \frac{rl}{m}$, and the bound of Theorem 7 becomes

$$\|\hat{S}_l - S_*\|^2_{L_2(\Pi^2)} \le$$
$$\inf_{S \in \mathcal{S}(l;a)} \left[\|S - S_*\|^2_{L_2(\Pi^2)} + Cm^2\varepsilon^2 \left(\sum_{j=1}^{l} \|P_{\mathrm{supp}(S)} \phi_j\|^2 \bigvee \frac{\mathrm{rank}(S) l}{m} \right) \right] + C\frac{a^2 t_{n,m}}{n}.$$

In fact, an inspection of the proof shows that the term $\frac{\mathrm{rank}(S) l}{m}$ is not needed in this special case.

Using simple aggregation techniques, it is easy to construct an adaptive estimator for which the oracle inequality of Theorem 7 holds with the optimal value of $\bar{\varepsilon}$ that minimizes

the right hand side of the bound. To this end, divide the sample $(X_1, X_1', Y_1), \dots, (X_n, X_n', Y_n)$ into two parts,

$$(X_j, X_j', Y_j), j = 1, \dots, n' \text{ and } (X_{n'+j}, X_{n'+j}', Y_{n'+j}), j = 1, \dots, n - n',$$

where $n' := [n/2]+1$. The first part of the sample will be used to compute the estimators $\hat{S}_l := \hat{S}_{\varepsilon,\bar{\varepsilon}_l}$, $\varepsilon_l := \lambda_l^{-1}$, $l = k_0, \dots, m+1$ (they are defined by (4.1), but they are based only on the first n' observations). The second part of the sample is used for model selection:

$$\hat{l} := \mathrm{argmin}_{l=k_0,\dots,m+1} \frac{1}{n-n'} \sum_{j=1}^{n-n'} \Big(Y_{n'+j} - \hat{S}_l(X_{n'+j}, X_{n'+j}')\Big)^2.$$

Finally, let $\hat{S} := \hat{S}_{\hat{l}}$.

Theorem 8 *Under the assumptions and notations of Theorem 7, with probability at least* $1 - e^{-t}$,

$$\|\hat{S} - S_*\|^2_{L_2(\Pi^2)} \le \inf_{S\in\mathbb{D}} \Big[2\|S - S_*\|^2_{L_2(\Pi^2)} + C \inf_{\bar{\varepsilon}\in[0,\lambda_{k_0}^{-1}]} \Big(m^2\varepsilon^2\bar{\varphi}(S;\bar{\varepsilon}^{-1}) + \bar{\varepsilon}\|W^{1/2}S\|^2_{L_2(\Pi^2)}\Big)\Big]$$
$$+C\frac{a^2(\log(m+1) + t_{n,m})}{n}. \tag{4.8}$$

Proof. The idea of aggregation result behind this theorem is rather well known (see Massart (2007), Chapter 8). The proof can be deduced, for instance, from Proposition 2 used in Section 3. Specifically, this proposition has to be applied in the case when $\mathcal{G}$ is a finite class of functions bounded by 1. Let $N := \mathrm{card}(\mathcal{G})$. Then, for some numerical constant $C_1 > 0$

$$\psi_n(\delta) \le C_1\Big[\delta\sqrt{\frac{\log N}{n}} \bigvee \frac{\log N}{n}\Big]$$

(see, e.g., Koltchinskii (2011b), Theorem 3.5) and Proposition 2 easily implies that, for all $t > 0$, with probability at least $1 - e^{-t}$

$$\|\hat{g} - g_*\|^2_{L_2(\Pi)} \le 2 \inf_{g\in\mathcal{G}} \|g - g_*\|^2_{L_2(\Pi)} + C_2\frac{\log N + t}{n}, \tag{4.9}$$

where $C_2 > 0$ is a constant. We will assume that $a = 1$ (in the general case, the result would follow by rescaling) and use bound (4.9), conditionally on the first part of the sample, in the case when $\mathcal{G} := \{\hat{g}_l : l = k_0, \dots, m+1\}$. Then, given $(X_j, X_j', Y_j), j = 1, \dots, n'$, with probability at least $1 - e^{-t}$

$$\|\hat{S} - S_*\|^2_{L_2(\Pi^2)} \le 2 \min_{k_0\le l\le m+1} \|\hat{S}_l - S_*\|^2_{L_2(\Pi)} + C_2\frac{\log(m+1) + t}{n}. \tag{4.10}$$

By Theorem 7 (with t replaced by $t + \log(m+1)$) and the union bound, we get that, with probability at least $1 - e^{-t}$, for all $l = k_0, \dots, m+1$,

$$\|\hat{S}_l - S_*\|^2_{L_2(\Pi^2)} \le \inf_{S \in \mathbb{D}} \Bigg[\|S - S_*\|^2_{L_2(\Pi^2)} + C_3 m^2 \varepsilon^2 \bar{\varphi}(S; \bar{\varepsilon}_l^{-1})$$
$$+ \bar{\varepsilon}_l \|W^{1/2} S\|^2_{L_2(\Pi^2)} \Bigg] + C_3 \frac{\log(m+1) + t_{n,m}}{n}$$

with some constant $C_3 > 0$. This yields the following upper bound on the minimal error $\min_{k_0 \le l \le m+1} \|\hat{S}_l - S_*\|^2_{L_2(\Pi)}$:

$$\min_{k_0 \le l \le m+1} \|\hat{S}_l - S_*\|^2_{L_2(\Pi^2)} \le \inf_{S \in \mathbb{D}} \Bigg[\|S - S_*\|^2_{L_2(\Pi^2)} + \tag{4.11}$$
$$C_3 \min_{k_0 \le l \le m+1} \Big(m^2 \varepsilon^2 \bar{\varphi}(S; \bar{\varepsilon}_l^{-1}) + \bar{\varepsilon}_l \|W^{1/2} S\|^2_{L_2(\Pi^2)} \Big) \Bigg] + C_3 \frac{\log(m+1) + t_{n,m}}{n}.$$

Using monotonicity of the function $\lambda \mapsto \varphi(S; \lambda)$ and the condition that $\lambda_{l+1} \le c\lambda_l, l = k_0, \dots, m-1$, it is easy to replace the minimum over l in (4.11) by the infimum over $\bar{\varepsilon}$. Combining (4.11) and (4.10) and adjusting the constants yields the result.

□

Using more sophisticated aggregation methods (for instance, such as the methods studied by Gaifas and Lecué (2011)), it is possible to construct an estimator $\hat{S}$ for which the oracle inequality similar to (4.8) holds with constant 1 in front of the approximation error term $\|S - S_*\|^2_{L_2(\Pi^2)}$.

To understand better the meaning of function $\bar{\varphi}$ involved in the statements of theorems 7 and 8 it makes sense to relate it to the low coherence assumptions discussed in the Introduction. Indeed, suppose that, for some $\nu \ge 1$,

$$\|P_L \phi_k\|^2 \le \frac{\nu r}{m}, k = 1, \dots, m. \tag{4.12}$$

This is a part of standard low coherence assumptions on matrix S_* with respect to the orthonormal basis $\{\phi_k\}$ (see (1.2)). Clearly, it implies that[6]

$$\bar{\varphi}(S; \lambda) \le \frac{\nu r \bar{F}(\lambda)}{m}, \lambda \ge 0. \tag{4.13}$$

Suppose that $n \ge m \log(2m)$ and $\varepsilon = Da\sqrt{\frac{\log(2m)}{nm}}$. If condition (4.13) holds for the target kernel S_* with $r = \operatorname{rank}(S_*)$ and some $\nu \ge 1$, then Theorem 7 implies that with

[6]Compare (4.13) with (4.3)

probability at least $1-e^{-t}$,

$$\|\hat{S}_{\varepsilon,\bar{\varepsilon}}-S_*\|^2_{L_2(\Pi^2)} \le C\frac{a^2\nu r\bar{F}(\bar{\varepsilon}^{-1})\log(2m)}{n}+\bar{\varepsilon}\|W^{1/2}S_*\|^2_{L_2(\Pi^2)}+C\frac{a^2t_{n,m}}{n}$$

and Theorem 8 implies that with the same probability

$$\|\hat{S}-S_*\|^2_{L_2(\Pi^2)} \le C\inf_{\bar{\varepsilon}\in[0,\lambda_{k_0}^{-1}]}\left(\frac{a^2\nu r\bar{F}(\bar{\varepsilon}^{-1})\log(2m)}{n}+\bar{\varepsilon}\|W^{1/2}S_*\|^2_{L_2(\Pi^2)}\right)+C\frac{a^2(\log(m+1)+t_{n,m})}{n}.$$

Example. If $\lambda_k \asymp k^{2\beta}$ for some $\beta > 1/2$, then it is easy to check that $\bar{F}(\lambda) \asymp \lambda^{1/2\beta}$. Under the assumption that $\|W^{1/2}S_*\|^2_{L_2(\Pi^2)} \le \rho^2$, we get the bound

$$\|\hat{S}-S_*\|^2_{L_2(\Pi^2)} \le C\left(\left(\left(\frac{a^2\rho^{1/\beta}\nu r\log(2m)}{n}\right)^{2\beta/(2\beta+1)}\bigwedge\frac{a^2rm}{n}\right)\right. \tag{4.14}$$
$$\left.\bigvee\frac{a^2(\log(m+1)+t_{n,m})}{n}\right).$$

Under the following slightly modified version of low coherence assumption (4.13),

$$\bar{\varphi}(S;\lambda) \le \frac{\nu(r\wedge\bar{F}(\lambda))\bar{F}(\lambda)}{m}, \lambda \ge 0, \tag{4.15}$$

one can almost recover upper bounds of Section 3:

$$\|\hat{S}-S_*\|^2_{L_2(\Pi^2)} \le C\left(\left(\left(\frac{\nu a^2\rho^{1/\beta}r\log(2m)}{n}\right)^{2\beta/(2\beta+1)}\bigwedge\left(\frac{\nu a^2\rho^{2/\beta}\log(2m)}{n}\right)^{\beta/(\beta+1)}\right.\right.$$
$$\left.\left.\bigwedge\frac{a^2rm}{n}\right)\bigvee\frac{a^2(\log(m+1)+t_{n,m})}{n}\right).$$

The main difference with what was proved in Section 3 is that now the low coherence constant ν is involved in the bounds, so, the methods discussed in this section yield correct (up to log factors) error rates provided that the target kernel S_* has "low coherence" with respect to the basis of eigenfunctions of W.

Proof of Theorem 7. Bound (4.5) will be proved for a fixed oracle $S \in \mathbb{D}$ and an arbitrary function $\varphi \in \Psi_{S,W}$ with $\varphi(\lambda) = r, \lambda \ge \lambda_m$ instead of $\bar{\varphi}$. It then can be applied to the function $\bar{\varphi}$ (which is the smallest function in $\Psi_{S,W}$). Without loss of generality, we assume that $a = 1$; the general case then follows by a simple rescaling. Finally, we will denote $\hat{S} := \hat{S}_{\varepsilon,\bar{\varepsilon}}$ throughout the proof.

Define the following orthogonal projectors $\mathcal{P}_L, \mathcal{P}_L^{\perp}$ in the space $\mathcal{S}_V$ with Hilbert–Schmidt inner product: $\mathcal{P}_L(A) := A - P_{L^{\perp}}AP_{L^{\perp}}$, $\mathcal{P}_L^{\perp}(A) = P_{L^{\perp}}AP_{L^{\perp}}, A \in \mathcal{S}_V$. We will use a well known representation of subdifferential of convex function $S \mapsto \|S\|_1$:

$$\partial\|S\|_1 = \left\{\mathrm{sign}(S) + \mathcal{P}_L^{\perp}(M) : M \in \mathcal{S}_V, \|M\| \le 1\right\},$$

where $L = \mathrm{supp}(S)$ (see Koltchinskii (2011b), Appendix A.4 and references therein). Denote

$$L_n(S) := \frac{1}{n}\sum_{j=1}^{n}(Y_j - S(X_j, X_j'))^2 + \varepsilon\|S\|_1 + \bar{\varepsilon}\|W^{1/2}S\|^2_{L_2(\Pi^2)},$$

so that $\hat{S} := \mathrm{argmin}_{S\in\mathbb{D}} L_n(S)$. An arbitrary matrix $A \in \partial L_n(\hat{S})$ can be represented as

$$A = \frac{2}{n}\sum_{i=1}^{n}\hat{S}(X_i, X_i')E_{X_i,X_i'} - \frac{2}{n}\sum_{i=1}^{n}Y_iE_{X_i,X_i'} + \varepsilon\hat{V} + 2\frac{\bar{\varepsilon}}{m^2}W\hat{S}, \tag{4.16}$$

where $\hat{V} \in \partial\|\hat{S}\|_1$. Since $\hat{S}$ is a minimizer of $L_n(S)$, there exists a matrix $A \in \partial L_n(\hat{S})$ such that $-A$ belongs to the normal cone of $\mathbb{D}$ at the point $\hat{S}$ (see Aubin and Ekeland (1984), Chap. 2, Corollary 6). This implies that $\langle A, \hat{S} - S\rangle \le 0$ and, in view of (4.16),

$$\begin{aligned} &2P_n(\hat{S}(\hat{S} - S)) - \left\langle \frac{2}{n}\sum_{i=1}^{n}Y_iE_{X_i,X_i'}, \hat{S} - S\right\rangle + \\ &\varepsilon\langle\hat{V}, \hat{S} - S\rangle + 2\frac{\bar{\varepsilon}}{m^2}\langle W\hat{S}, \hat{S} - S\rangle \le 0. \end{aligned} \tag{4.17}$$

Here and in what follows P_n denotes the empirical distribution based on the sample $(X_1, X_1', Y_1), \dots, (X_n, X_n', Y_n)$. The corresponding true distribution of (X, X', Y) will be denoted by P. It easily follows from (4.17) that

$$\begin{aligned} &2\langle\hat{S} - S_*, \hat{S} - S\rangle_{L_2(P_n)} - 2\langle\Xi, \hat{S} - S\rangle + \\ &\varepsilon\langle\hat{V}, \hat{S} - S\rangle + 2\bar{\varepsilon}\langle W^{1/2}\hat{S}, W^{1/2}(\hat{S} - S)\rangle_{L_2(\Pi^2)} \le 0, \end{aligned}$$

where

$$\Xi := \frac{1}{n}\sum_{j=1}^{n}\xi_jE_{X_j,X_j'}, \quad \xi_j := Y_j - S_*(X_j, X_j').$$

We can now rewrite the last bound as

$$\begin{aligned} &2\langle\hat{S} - S_*, \hat{S} - S\rangle_{L_2(P)} + \varepsilon\langle\hat{V}, \hat{S} - S\rangle + 2\bar{\varepsilon}\langle W^{1/2}(\hat{S} - S), W^{1/2}(\hat{S} - S)\rangle_{L_2(\Pi^2)} \\ &\le -2\bar{\varepsilon}\langle W^{1/2}S, W^{1/2}(\hat{S} - S)\rangle_{L_2(\Pi^2)} + 2\langle\Xi, \hat{S} - S\rangle + 2(P - P_n)((\hat{S} - S_*)(\hat{S} - S)) \end{aligned}$$

and use a simple identity

$$2\langle\hat{S}-S_*, \hat{S}-S\rangle_{L_2(P)} = 2\langle\hat{S}-S_*, \hat{S}-S\rangle_{L_2(\Pi^2)} = \|\hat{S}-S_*\|^2_{L_2(\Pi^2)} + \|\hat{S}-S\|^2_{L_2(\Pi^2)} - \|S-S_*\|^2_{L_2(\Pi^2)}$$

to get the following bound:

$$\begin{aligned} &\|\hat{S} - S_*\|^2_{L_2(\Pi^2)} + \|\hat{S} - S\|^2_{L_2(\Pi^2)} + 2\bar{\varepsilon}\|W^{1/2}(\hat{S} - S)\|^2_{L_2(\Pi^2)} + \varepsilon\langle\hat{V}, \hat{S} - S\rangle \\ &\le \|S - S_*\|^2_{L_2(\Pi^2)} - 2\bar{\varepsilon}\langle W^{1/2}S, W^{1/2}(\hat{S} - S)\rangle_{L_2(\Pi^2)} + \\ &2\langle\Xi, \hat{S} - S\rangle + 2(P - P_n)(S - S_*)(\hat{S} - S) + 2(P - P_n)(\hat{S} - S)^2. \end{aligned} \tag{4.18}$$

For an arbitrary $V \in \partial\|S\|_1$, $V = \mathrm{sign}(S) + \mathcal{P}_L^{\perp}(M)$, where M is a matrix with $\|M\| \leq 1$. It follows from the trace duality property that there exists an M with $\|M\| \leq 1$ such that

$$\langle \mathcal{P}_L^{\perp}(M), \hat{S} - S\rangle = \langle M, \mathcal{P}_L^{\perp}(\hat{S} - S)\rangle = \langle M, \mathcal{P}_L^{\perp}(\hat{S})\rangle = \|\mathcal{P}_L^{\perp}(\hat{S})\|_1,$$

where the first equality is based on the fact that $\mathcal{P}_L^{\perp}$ is a self-adjoint operator and the second equality is based on the fact that S has support L. Using this equation and monotonicity of subdifferentials of convex functions, we get

$$\langle \mathrm{sign}(S), \hat{S} - S\rangle + \|\mathcal{P}_L^{\perp}(\hat{S})\|_1 = \langle V, \hat{S} - S\rangle \leq \langle \hat{V}, \hat{S} - S\rangle.$$

Substituting this into the left hand side of (4.18), it is easy to get

$$\begin{aligned} &\|\hat{S} - S_*\|^2_{L_2(\Pi^2)} + \|\hat{S} - S\|^2_{L_2(\Pi^2)} + \varepsilon\|\mathcal{P}_L^{\perp}(\hat{S})\|_1 + 2\bar{\varepsilon}\|W^{1/2}(\hat{S} - S)\|^2_{L_2(\Pi^2)} \quad (4.19)\\ &\leq \|S - S_*\|^2_{L_2(\Pi^2)} - \varepsilon\langle \mathrm{sign}(S), \hat{S} - S\rangle - 2\bar{\varepsilon}\langle W^{1/2}S, W^{1/2}(\hat{S} - S)\rangle_{L_2(\Pi^2)}\\ &+2\langle \Xi, \hat{S} - S\rangle + 2(P - P_n)(S - S_*)(\hat{S} - S) + 2(P - P_n)(\hat{S} - S)^2. \end{aligned}$$

We need to bound the right hand side of (4.19). We start with deriving a bound on $\langle \mathrm{sign}(S), \hat{S} - S\rangle$, expressed in terms of function φ. Note that, for all $\lambda > 0$,

$$\begin{aligned} &\langle \mathrm{sign}(S), \hat{S} - S\rangle = \sum_{k=1}^{m} \langle \mathrm{sign}(S)\phi_k, (\hat{S} - S)\phi_k\rangle = \\ &\sum_{\lambda_k \leq \lambda} \langle \mathrm{sign}(S)\phi_k, (\hat{S} - S)\phi_k\rangle + \sum_{\lambda_k > \lambda} \left\langle \frac{\mathrm{sign}(S)\phi_k}{\sqrt{\lambda_k}}, \sqrt{\lambda_k}(\hat{S} - S)\phi_k \right\rangle, \end{aligned}$$

which easily implies

$$\begin{aligned} &|\langle \mathrm{sign}(S), \hat{S} - S\rangle| \leq \left(\sum_{\lambda_k \leq \lambda} \|\mathrm{sign}(S)\phi_k\|^2\right)^{1/2} \left(\sum_{\lambda_k \leq \lambda} \|(\hat{S} - S)\phi_k\|^2\right)^{1/2} + \quad (4.20)\\ &\left(\sum_{\lambda_k > \lambda} \frac{\|\mathrm{sign}(S)\phi_k\|^2}{\lambda_k}\right)^{1/2} \left(\sum_{\lambda_k > \lambda} \lambda_k \|(\hat{S} - S)\phi_k\|^2\right)^{1/2} \leq \\ &\left(\sum_{\lambda_k \leq \lambda} \|P_L\phi_k\|^2\right)^{1/2} \|\hat{S} - S\|_2 + \left(\sum_{\lambda_k > \lambda} \frac{\|P_L\phi_k\|^2}{\lambda_k}\right)^{1/2} \|W^{1/2}(\hat{S} - S)\|_2. \end{aligned}$$

We will now use the following elementary lemma.

Lemma 4 *For all* $\lambda > 0$,

$$\sum_{\lambda_k > \lambda} \frac{\|P_L \phi_k\|^2}{\lambda_k} \le c_\gamma \frac{\varphi(\lambda)}{\lambda} \quad \text{and} \quad \sum_{\lambda_k > \lambda} \frac{1}{\lambda_k} \le c_\gamma \frac{\bar{F}(\lambda)}{\lambda},$$

where $c_\gamma := \frac{c+\gamma}{\gamma}$.

Proof. Denote $H_k := \sum_{j=1}^{l} \|P_L \phi_j\|^2, k = 1, \dots, m$. Suppose that $\lambda \in [\lambda_l, \lambda_{l+1}]$ for some $l = k_0 - 1, \dots, m-1$. We will use the properties of functions $\varphi \in \Psi_{S,W}$ and $\bar{F}$. In particular, recall that the functions $\frac{\varphi(\lambda)}{F(\lambda)}$ and $\frac{\bar{F}(\lambda)}{\lambda}$ are nonincreasing. Using these properties and the condition that $\lambda_{k+1} \le c\lambda_k, k \ge k_0$ we get

$$\sum_{\lambda_k > \lambda} \frac{\|P_L \phi_k\|^2}{\lambda_k} = \sum_{k=l+1}^{m-1} H_k \left(\frac{1}{\lambda_k} - \frac{1}{\lambda_{k+1}} \right) + \frac{H_m}{\lambda_m} - \frac{H_l}{\lambda_{l+1}} \le$$
$$\sum_{k=l+1}^{m-1} \varphi(\lambda_k) \left(\frac{1}{\lambda_k} - \frac{1}{\lambda_{k+1}} \right) + \frac{\varphi(\lambda_m)}{\lambda_m} \le c \sum_{k=l+1}^{m-1} \frac{\varphi(\lambda_{k+1})}{\lambda_{k+1}^2} (\lambda_{k+1} - \lambda_k) + \frac{\varphi(\lambda_m)}{\lambda_m} \le$$
$$c \int_\lambda^\infty \frac{\varphi(s)}{s^2} ds + \frac{\varphi(\lambda)}{\lambda} \le c \int_\lambda^\infty \frac{\varphi(s)}{\bar{F}(s)} \frac{\bar{F}(s)}{s^2} ds + \frac{\varphi(\lambda)}{\lambda} \le$$
$$c \frac{\varphi(\lambda)}{\bar{F}(\lambda)} \int_\lambda^\infty \frac{\bar{F}(s)}{s^2} ds + \frac{\varphi(\lambda)}{\lambda} \le \frac{c}{\gamma} \frac{\varphi(\lambda)}{\bar{F}(\lambda)} \frac{\bar{F}(\lambda)}{\lambda} + \frac{\varphi(\lambda)}{\lambda} = \frac{c+\gamma}{\gamma} \frac{\varphi(\lambda)}{\lambda},$$

which proofs the first bound. To prove the second bound, replace in the inequalities above $\|P_L \phi_k\|^2$ by 1 and $\varphi(\lambda)$ by $\bar{F}(\lambda)$. In the case when $\lambda \ge \lambda_m$, both bounds are trivial since their left hand sides are equal to zero.

□

It follows from from (4.20) and the first bound of Lemma 4 that

$$|\langle \operatorname{sign}(S), \hat{S} - S \rangle| \le \sqrt{\varphi(\lambda)} \|\hat{S} - S\|_2 + \sqrt{c_\gamma \frac{\varphi(\lambda)}{\lambda}} \|W^{1/2}(\hat{S} - S)\|_2 =$$
$$m \sqrt{\varphi(\lambda)} \|\hat{S} - S\|_{L_2(\Pi^2)} + m \sqrt{c_\gamma \frac{\varphi(\lambda)}{\lambda}} \|W^{1/2}(\hat{S} - S)\|_{L_2(\Pi^2)}. \tag{4.21}$$

This implies the following bound:

$$\varepsilon |\langle \operatorname{sign}(S), \hat{S} - S \rangle| \le \tag{4.22}$$
$$\varphi(\lambda) m^2 \varepsilon^2 + \frac{1}{4} \|\hat{S} - S\|_{L_2(\Pi^2)}^2 + c_\gamma \frac{\varphi(\lambda)}{\lambda} \frac{m^2 \varepsilon^2}{\bar{\varepsilon}} + \frac{\bar{\varepsilon}}{4} \|W^{1/2}(\hat{S} - S)\|_{L_2(\Pi^2)}^2,$$

where we used twice an elementary inequality $ab \le a^2 + \frac{1}{4} b^2, a, b > 0$. We will apply this

bound for $\lambda = \bar{\varepsilon}^{-1}$ to get the following inequality:

$$\varepsilon|\langle \mathrm{sign}(S), \hat{S} - S\rangle| \leq \tag{4.23}$$
$$(c_\gamma + 1)\varphi(\bar{\varepsilon}^{-1})m^2\varepsilon^2 + \frac{1}{4}\|\hat{S} - S\|^2_{L_2(\Pi^2)} + \frac{\bar{\varepsilon}}{4}\|W^{1/2}(\hat{S} - S)\|^2_{L_2(\Pi^2)}.$$

To bound the second term in the right hand side of (4.19), note that

$$\bar{\varepsilon}|\langle W^{1/2}S, W^{1/2}(\hat{S} - S)\rangle_{L_2(\Pi^2)}| \leq \bar{\varepsilon}\|W^{1/2}S\|_{L_2(\Pi^2)}\|W^{1/2}(\hat{S} - S)\|_{L_2(\Pi^2)} \leq \tag{4.24}$$
$$\bar{\varepsilon}\|W^{1/2}S\|^2_{L_2(\Pi^2)} + \frac{\bar{\varepsilon}}{4}\|W^{1/2}(\hat{S} - S)\|^2_{L_2(\Pi^2)}.$$

The main part of the proof deals with bounding the stochastic term

$$2\langle \Xi, \hat{S} - S_*\rangle + 2(P - P_n)(S - S_*)(\hat{S} - S) + 2(P - P_n)(\hat{S} - S)^2.$$

in the right hand side of (4.19). To this end, define (for fixed S, S_*)

$$f_A(y, u, v) := (y - S_*(u, v))(A - S)(u, v) - (S - S_*)(u, v)(A - S)(u, v) - (A - S)^2(u, v) =$$
$$(y - S(u, v))(A - S)(u, v) - (A - S)^2(u, v)$$

and consider the following empirical process

$$\alpha_n(\delta_1, \delta_2, \delta_3) := \sup\Big\{|(P_n - P)(f_A)| : A \in \mathcal{T}(\delta_1, \delta_2, \delta_3)\Big\},$$

where

$$\mathcal{T}(\delta_1, \delta_2, \delta_3) := \Big\{\|A - S\|_{L_2(\Pi^2)} \leq \delta_1, \|\mathcal{P}_L^\perp A\|_1 \leq \delta_2, \|W^{1/2}(A - S)\|_{L_2(\Pi^2)} \leq \delta_3\Big\}.$$

Clearly, we have

$$2\langle \Xi, \hat{S} - S_*\rangle + 2(P - P_n)(S - S_*)(\hat{S} - S) + 2(P - P_n)(\hat{S} - S)^2 \leq \tag{4.25}$$
$$2\alpha_n\Big(\|\hat{S} - S\|_{L_2(\Pi^2)}, \|\mathcal{P}_L^\perp \hat{S}\|_1, \|W^{1/2}(\hat{S} - S)\|_{L_2(\Pi^2)}\Big)$$

and it remains to provide an upper bound on $\alpha_n(\delta_1, \delta_2, \delta_3)$ that is uniform in some intervals of the parameters $\delta_1, \delta_2, \delta_3$ (such that either the norms $\|\hat{S} - S\|_{L_2(\Pi^2)}, \|\mathcal{P}_L^\perp \hat{S}\|_1, \|W^{1/2}(\hat{S} - S)\|_{L_2(\Pi^2)}$ belong to these intervals with a high probability, or bound of the theorem trivially holds). Note that the functions f_A are uniformly bounded by a numerical constant (under the assumptions that $a = 1$, $|Y| \leq a$ and all the kernels are also bounded by a) and we have $Pf_A^2 \leq c_1\|A - S\|^2_{L_2(\Pi)}$ with some numerical constant $c_1 > 0$. Using

Talagrand's concentration inequality for empirical processes we conclude that for fixed $\delta_1, \delta_2, \delta_3$ with probability at least $1 - e^{-t}$ and with some constant $c_2 > 0$

$$\alpha_n(\delta_1, \delta_2, \delta_3) \le 2\mathbb{E}\alpha_n(\delta_1, \delta_2, \delta_3) + c_2\Big(\delta_1\sqrt{\frac{t}{n}} + \frac{t}{n}\Big).$$

We will make this bound uniform in $\delta_k \in [\delta_k^-, \delta_k^+], \delta_k^- < \delta_k^+, k = 1, 2, 3$ (these intervals will be chosen later). Define $\delta_k^j := \delta_k^+ 2^{-j}, j = 0, \dots, [\log_2(\delta_k^+/\delta_k^-)] + 1, k = 1, 2, 3$ and let $\bar{t} := t + \sum_{k=1}^3 \log\Big([\log_2(\delta_k^+/\delta_k^-)] + 2\Big)$. By the union bound, with probability at least $1 - e^{-t}$ and for all $j_k = 0, \dots, [\log_2(\delta_k^+/\delta_k^-)] + 1, k = 1, 2, 3$,

$$\alpha_n(\delta_1^{j_1}, \delta_2^{j_2}, \delta_3^{j_3}) \le 2\mathbb{E}\alpha_n(\delta_1^{j_1}, \delta_2^{j_2}, \delta_3^{j_3}) + c_2\Big(\delta_1^{j_1}\sqrt{\frac{\bar{t}}{n}} + \frac{\bar{t}}{n}\Big).$$

By monotonicity of α_n and of the right hand side of the bound with respect to each of the variables $\delta_1, \delta_2, \delta_3$, we conclude that with the same probability and with some numerical constant $c_3 > 0$, for all $\delta_k \in [\delta_k^-, \delta_k^+], k = 1, 2, 3$,

$$\alpha_n(\delta_1, \delta_2, \delta_3) \le 2\mathbb{E}\alpha_n(2\delta_1, 2\delta_2, 2\delta_3) + c_3\Big(\delta_1\sqrt{\frac{\bar{t}}{n}} + \frac{\bar{t}}{n}\Big). \tag{4.26}$$

To bound the expectation $\mathbb{E}\alpha_n(2\delta_1, 2\delta_2, 2\delta_3)$ in the right hand side of (4.26), note that, by the definition of function f_A,

$$\begin{aligned}
&\mathbb{E}\alpha_n(\delta_1, \delta_2, \delta_3) \le \mathbb{E}\sup\Big\{\Big|(P_n - P)(y - S)(A - S)\Big| : A \in \mathcal{T}(\delta_1, \delta_2, \delta_3)\Big\} + \\
&\mathbb{E}\sup\Big\{\Big|(P_n - P)(A - S)^2\Big| : A \in \mathcal{T}(\delta_1, \delta_2, \delta_3)\Big\}.
\end{aligned} \tag{4.27}$$

A standard application of symmetrization inequality followed by contraction inequality for Rademacher sums (see, e.g., Koltchinskii (2011b), Chapter 2) yields

$$\begin{aligned}
&\mathbb{E}\sup\Big\{\Big|(P_n - P)(A - S)^2\Big| : A \in \mathcal{T}(\delta_1, \delta_2, \delta_3)\Big\} \le \\
&16\mathbb{E}\sup\Big\{\Big|R_n(A - S)\Big| : A \in \mathcal{T}(\delta_1, \delta_2, \delta_3)\Big\}.
\end{aligned} \tag{4.28}$$

It easily follows from (4.27) and (4.28) that

$$\begin{aligned}
&\mathbb{E}\alpha_n(\delta_1, \delta_2, \delta_3) \le \mathbb{E}\sup\Big\{|\langle \Xi_1, A - S\rangle| : A \in \mathcal{T}(\delta_1, \delta_2, \delta_3)\Big\} + \\
&16\mathbb{E}\sup\Big\{|\langle \Xi_2, A - S\rangle| : A \in \mathcal{T}(\delta_1, \delta_2, \delta_3)\Big\},
\end{aligned} \tag{4.29}$$

where

$$\Xi_1 := \frac{1}{n}\sum_{j=1}^n (Y_j - S(X_j, X_j'))E_{X_j, X_j'} - \mathbb{E}(Y - S(X, X'))E_{X, X'}$$

and $\Xi_2 := \frac{1}{n}\sum_{j=1}^n \varepsilon_j E_{X_j,X_j'}$, $\{\varepsilon_j\}$ being i.i.d. Rademacher random variables independent of $(X_1, X_1', Y_1), \ldots, (X_n, X_n', Y_n)$. We will upper bound the expectations in the right hand side of (4.29), which reduces to bounding $\mathbb{E}\sup\Big\{|\langle \Xi_i, A-S\rangle| : A \in \mathcal{T}(\delta_1,\delta_2,\delta_3)\Big\}$ for each of the random matrices Ξ_1, Ξ_2. For $i = 1, 2$ and $A \in \mathcal{T}(\delta_1,\delta_2,\delta_3)$, we have

$$
\begin{aligned}
&|\langle \Xi_i, A-S\rangle| \le |\langle \Xi_i, \mathcal{P}_L(A-S)\rangle| + |\langle \Xi_i, \mathcal{P}_L^{\perp}(A)\rangle| \\
&\le |\langle \mathcal{P}_L\Xi_i, A-S\rangle| + \|\Xi_i\|\|\mathcal{P}_L^{\perp}(A)\|_1 \le |\langle \mathcal{P}_L\Xi_i, A-S\rangle| + \delta_2\|\Xi_i\|. \qquad (4.30)
\end{aligned}
$$

To bound $\|\Xi_i\|$, we use a version of noncommutative Bernstein inequality of Ahlswede and Winter (2002) (see also Tropp (2010), Koltchinskii (2011a, 2011b, 2011c) for other versions of such inequalities).

Lemma 5 *Let Z be a random symmetric matrix with $\mathbb{E}Z = 0$, $\sigma_Z^2 := \|\mathbb{E}Z^2\|$ and $\|Z\| \le U$ for some $U > 0$. Let $Z_1, \ldots, Z_n$ be n i.i.d. copies of Z. Then for all $t > 0$, with probability at least $1 - e^{-t}$*

$$
\left\|\frac{1}{n}\sum_{i=1}^n Z_i\right\| \le 2\left(\sigma_Z\sqrt{\frac{t+\log(2m)}{n}} \bigvee U\frac{t+\log(2m)}{n}\right).
$$

This also implies that

$$
\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^n Z_i\right\| \le 4\left(\sigma_Z\sqrt{\frac{\log(2m)}{n}} \bigvee U\frac{\log(2m)}{n}\right).
$$

It is applied to i.i.d. random matrices

$$
Z_j := (Y_j - S(X_j, X_j'))E_{X_j,X_j'} - \mathbb{E}(Y - S(X,X')E_{X,X'}
$$

in the case of matrix Ξ_1 and to i.i.d. random matrices $Z_j := \varepsilon_j E_{X_j,X_j'}$ in the case of matrix Ξ_2. In both cases, $\|Z_j\| \le 4$ and, by a simple computation, $\sigma_{Z_j}^2 := \|\mathbb{E}Z_j^2\| \le 4/m$ (see, e.g., Koltchinskii (2011b), Section 9.4), Lemma 5 implies that, for $i = 1, 2$,

$$
\mathbb{E}\|\Xi_i\| \le 16\left[\sqrt{\frac{\log(2m)}{nm}} \bigvee \frac{\log(2m)}{n}\right] =: \varepsilon^*. \qquad (4.31)
$$

To control the term $|\langle \mathcal{P}_L\Xi_i, A-S\rangle|$ in bound (4.30), we will use the following lemma.

Lemma 6 *For all $\delta > 0$,*

$$
\mathbb{E}\sup_{\|M\|_2\le\delta, \|W^{1/2}M\|_2\le 1} |\langle \mathcal{P}_L\Xi_i, M\rangle| \le 4\sqrt{2}\sqrt{c_\gamma + 1}\sqrt{\frac{1}{nm}}\delta\sqrt{\varphi(\delta^{-2})}.
$$

Proof. For all symmetric $m \times m$ matrices M,

$$\langle \mathcal{P}_L \Xi_i, M\rangle = \sum_{k,j=1}^m \langle \mathcal{P}_L \Xi, \phi_k \otimes \phi_j\rangle \langle M, \phi_k \otimes \phi_j\rangle.$$

Assuming that

$$\|M\|_2^2 = \sum_{k,j=1}^m |\langle M, \phi_k \otimes \phi_j\rangle|^2 \le \delta^2 \quad \text{and} \quad \|W^{1/2} M\|_2^2 = \sum_{k,j=1}^m \lambda_k |\langle M, \phi_k \otimes \phi_j\rangle|^2 \le 1.$$

it is easy to conclude that $\sum_{k,j=1}^m \frac{|\langle M, \phi_k \otimes \phi_j\rangle|^2}{\lambda_k^{-1} \wedge \delta^2} \le 2$. It follows

$$\begin{aligned} &|\langle \mathcal{P}_L \Xi_i, M\rangle| \le \\ &\left(\sum_{k,j=1}^m (\lambda_k^{-1} \wedge \delta^2) |\langle \mathcal{P}_L \Xi, \phi_k \otimes \phi_j\rangle|^2 \right)^{1/2} \left(\sum_{k,j=1}^m \frac{|\langle M, \phi_k \otimes \phi_j\rangle|^2}{\lambda_k^{-1} \wedge \delta^2} \right)^{1/2} \le \\ &\sqrt{2} \left(\sum_{k,j=1}^m (\lambda_k^{-1} \wedge \delta^2) |\langle \mathcal{P}_L \Xi, \phi_k \otimes \phi_j\rangle|^2 \right)^{1/2}. \end{aligned} \tag{4.32}$$

Consider the following inner product:

$$\langle M_1, M_2\rangle_w := \sum_{k,j=1}^m (\lambda_k^{-1} \wedge \delta^2) \langle M_1, \phi_k \otimes \phi_j\rangle \langle M_2, \phi_k \otimes \phi_j\rangle$$

and let $\|\cdot\|_w$ be the corresponding norm. We will provide an upper bound on

$$\mathbb{E}\|\mathcal{P}_L \Xi_i\|_w = \mathbb{E}\left(\sum_{k,j=1}^m (\lambda_k^{-1} \wedge \delta^2) |\langle \mathcal{P}_L \Xi, \phi_k \otimes \phi_j\rangle|^2 \right)^{1/2}.$$

Recall that

$$\Xi_i = n^{-1} \sum_{j=1}^n \zeta_j E_{X_j, X_j'} - \mathbb{E}(\zeta E_{X, X'}),$$

where $\zeta_j = Y_j - S(X_j, X_j')$ for $i = 1$ and $\zeta_j = \varepsilon_j$ for $i = 2$. Note that in the first case $|\zeta_j| \le 2$ and in the second case $|\zeta_j| \le 1$. Therefore,

$$\mathbb{E}\|\mathcal{P}_L \Xi_i\|_w \le \mathbb{E}^{1/2} \|\mathcal{P}_L \Xi_i\|_w^2 \le \sqrt{\frac{\mathbb{E}\zeta^2 \|\mathcal{P}_L E_{X, X'}\|_w^2}{n}} \le 2\sqrt{\frac{\mathbb{E}\|\mathcal{P}_L E_{X, X'}\|_w^2}{n}}. \tag{4.33}$$

It remains to bound $\mathbb{E}\|\mathcal{P}_L E_{X,X'}\|_w^2$:

$$\mathbb{E}\|\mathcal{P}_L(E_{X,X'})\|_w^2 = \tag{4.34}$$
$$\mathbb{E}\sum_{k,j=1}^{m}(\lambda_k^{-1}\wedge\delta^2)|\langle\mathcal{P}_L(E_{X,X'}),\phi_k\otimes\phi_j\rangle|^2 = \sum_{k,j=1}^{m}(\lambda_k^{-1}\wedge\delta^2)\mathbb{E}|\langle E_{X,X'},\mathcal{P}_L(\phi_k\otimes\phi_j)\rangle|^2 =$$
$$\sum_{k,j=1}^{m}(\lambda_k^{-1}\wedge\delta^2)m^{-2}\sum_{u,v\in V}|\langle E_{u,v},\mathcal{P}_L(\phi_k\otimes\phi_j)\rangle|^2 \leq$$
$$m^{-2}\sum_{k,j=1}^{m}(\lambda_k^{-1}\wedge\delta^2)\|\mathcal{P}_L(\phi_k\otimes\phi_j)\|_2^2 \leq 2m^{-2}\sum_{k,j=1}^{m}(\lambda_k^{-1}\wedge\delta^2)(\|P_L\phi_k\|^2+\|P_L\phi_j\|^2) =$$
$$2m^{-1}\sum_{k=1}^{m}(\lambda_k^{-1}\wedge\delta^2)\|P_L\phi_k\|^2 + 2m^{-2}\sum_{k=1}^{m}(\lambda_k^{-1}\wedge\delta^2)\sum_{j=1}^{m}\|P_L\phi_j\|^2 =$$
$$2m^{-1}\sum_{k=1}^{m}(\lambda_k^{-1}\wedge\delta^2)\|P_L\phi_k\|^2 + 2m^{-2}\sum_{k=1}^{m}(\lambda_k^{-1}\wedge\delta^2)\|P_L\|_2^2 =$$
$$2m^{-1}\sum_{k=1}^{m}(\lambda_k^{-1}\wedge\delta^2)\|P_L\phi_k\|^2 + 2m^{-2}r\sum_{k=1}^{m}(\lambda_k^{-1}\wedge\delta^2).$$

Note that

$$\sum_{k=1}^{m}(\lambda_k^{-1}\wedge\delta^2)\|P_L\phi_k\|^2 \leq \delta^2\sum_{\lambda_k\leq\delta^{-2}}\|P_L\phi_k\|^2 + \sum_{\lambda_k>\delta^{-2}}\lambda_k^{-1}\|P_L\phi_k\|^2. \tag{4.35}$$

Using the first bound of Lemma 4, we get from (4.35) that

$$\sum_{k=1}^{m}(\lambda_k^{-1}\wedge\delta^2)\|P_L\phi_k\|^2 \leq \delta^2\varphi(\delta^{-2}) + c_\gamma\delta^2\varphi(\delta^{-2}) = (c_\gamma+1)\delta^2\varphi(\delta^{-2}). \tag{4.36}$$

We also have $\sum_{k=1}^{m}(\lambda_k^{-1}\wedge\delta^2) \leq \sum_{\lambda_k\leq\delta^{-2}}\delta^2 + \sum_{\lambda_k>\delta^{-2}}\lambda_k^{-1}$, which, by the second bound of Lemma 4, implies that

$$\sum_{k=1}^{m}(\lambda_k^{-1}\wedge\delta^2) \leq \delta^2\bar{F}(\delta^{-2}) + c_\gamma\delta^2\bar{F}(\delta^{-2}) \leq (c_\gamma+1)\delta^2\bar{F}(\delta^{-2}). \tag{4.37}$$

Using bounds (4.34), (4.36) and (4.37) and the fact that $\varphi(\lambda) \geq \frac{r}{m}\bar{F}(\lambda)$, we get,

$$\mathbb{E}\|\mathcal{P}_L(E_{X,X'})\|_w^2 \leq \tag{4.38}$$
$$2m^{-1}(c_\gamma+1)\delta^2\varphi(\delta^{-2}) + 2m^{-2}r(c_\gamma+1)\delta^2\bar{F}(\delta^{-2}) \leq 4m^{-1}(c_\gamma+1)\delta^2\varphi(\delta^{-2}).$$

The proof follows from (4.32), (4.33) and (4.38).

□

Let $\delta := \frac{\delta_1}{\delta_3}$. Using Lemma 6, we get

$$\mathbb{E}\sup\Big\{|\langle \mathcal{P}_L \Xi_i, A-S\rangle| : A \in \mathcal{T}(\delta_1,\delta_2,\delta_3)\Big\} \le$$
$$\mathbb{E}\sup\Big\{|\langle \mathcal{P}_L \Xi_i, A-S\rangle| : \|A-S\|_{L_2(\Pi^2)} \le \delta_1, \|W^{1/2}(A-S)\|_{L_2(\Pi^2)} \le \delta_3\Big\} =$$
$$\mathbb{E}\sup\Big\{|\langle \mathcal{P}_L \Xi_i, A-S\rangle| : \|A-S\|_2 \le \delta_1 m, \|W^{1/2}(A-S)\|_2 \le \delta_3 m\Big\} \le$$
$$\delta_3 m \mathbb{E}\sup\Big\{|\langle \mathcal{P}_L \Xi_i, A-S\rangle| : \|A-S\|_2 \le \delta, \|W^{1/2}(A-S)\|_{L_2(\Pi^2)} \le 1\Big\} \le$$
$$4\sqrt{2}\delta_3 m\sqrt{c_\gamma+1}\sqrt{\frac{1}{nm}}\delta\sqrt{\varphi(\delta^{-2})} = 4\sqrt{2}\sqrt{c_\gamma+1}\sqrt{\frac{m}{n}}\delta_1\sqrt{\varphi(\delta^{-2})}.$$

In the case when $\delta^2 \ge \bar{\varepsilon}$, we get

$$\mathbb{E}\sup\Big\{|\langle \mathcal{P}_L \Xi_i, A-S\rangle| : A \in \mathcal{T}(\delta_1,\delta_2,\delta_3)\Big\} \le 4\sqrt{2}\sqrt{c_\gamma+1}\delta_1\sqrt{\frac{m\varphi(\bar{\varepsilon}^{-1})}{n}}.$$

In the opposite case, when $\delta^2 < \bar{\varepsilon}$, we use the fact that the function $\frac{\varphi(\lambda)}{\lambda} = \frac{\varphi(\lambda)}{F(\lambda)}\frac{\bar{F}(\lambda)}{\lambda}$ is nonincreasing. This implies that $\delta^2\varphi(\delta^{-2}) \le \bar{\varepsilon}\varphi(\bar{\varepsilon}^{-1})$ and we get

$$\mathbb{E}\sup\Big\{|\langle \mathcal{P}_L \Xi_i, A-S\rangle| : A \in \mathcal{T}(\delta_1,\delta_2,\delta_3)\Big\} \le$$
$$4\sqrt{2}\sqrt{c_\gamma+1}\sqrt{\frac{m}{n}}\delta_1\sqrt{\varphi(\delta^{-2})} = 4\sqrt{2}\sqrt{c_\gamma+1}\sqrt{\frac{m}{n}}\delta_3\sqrt{\delta^2\varphi(\delta^{-2})} \le$$
$$4\sqrt{2}\sqrt{c_\gamma+1}\sqrt{\frac{m}{n}}\delta_3\sqrt{\bar{\varepsilon}\varphi(\bar{\varepsilon}^{-1})} = 4\sqrt{2}\sqrt{c_\gamma+1}\sqrt{\bar{\varepsilon}}\delta_3\sqrt{\frac{m\varphi(\bar{\varepsilon}^{-1})}{n}}.$$

We can conclude that

$$\mathbb{E}\sup\Big\{|\langle \mathcal{P}_L \Xi_i, A-S\rangle| : A \in \mathcal{T}(\delta_1,\delta_2,\delta_3)\Big\} \le$$
$$4\sqrt{2}\sqrt{c_\gamma+1}\delta_1\sqrt{\frac{m\varphi(\bar{\varepsilon}^{-1})}{n}} + 4\sqrt{2}\sqrt{c_\gamma+1}\sqrt{\bar{\varepsilon}}\delta_3\sqrt{\frac{m\varphi(\bar{\varepsilon}^{-1})}{n}}.$$

This bound will be combined with (4.30) and (4.31) to get that, for $i = 1, 2$,

$$\mathbb{E}\sup\Big\{|\langle \Xi_i, A-S\rangle| : A \in \mathcal{T}(\delta_1,\delta_2,\delta_3)\Big\} \le \varepsilon^*\delta_2 +$$
$$4\sqrt{2}\sqrt{c_\gamma+1}\delta_1\sqrt{\frac{m\varphi(\bar{\varepsilon}^{-1})}{n}} + 4\sqrt{2}\sqrt{c_\gamma+1}\sqrt{\bar{\varepsilon}}\delta_3\sqrt{\frac{m\varphi(\bar{\varepsilon}^{-1})}{n}}.$$

In view of (4.29), this yields the bound

$$\mathbb{E}\alpha_n(\delta_1,\delta_2,\delta_3) \le C'\varepsilon^*\delta_2 + C'\delta_1\sqrt{\frac{m\varphi(\bar{\varepsilon}^{-1})}{n}} + C'\sqrt{\bar{\varepsilon}}\delta_3\sqrt{\frac{m\varphi(\bar{\varepsilon}^{-1})}{n}}$$

that holds with some constant $C' > 0$ for all $\delta_1, \delta_2, \delta_3 > 0$. Using (4.26), we conclude that for some constants C and for all $\delta_k \in [\delta_k^-, \delta_k^+], k = 1, 2, 3$,

$$\alpha_n(\delta_1, \delta_2, \delta_3) \le C\left[\delta_1\sqrt{\frac{m\varphi(\bar\varepsilon^{-1})}{n}} + \delta_1\sqrt{\frac{\bar t}{n}} + \frac{\bar t}{n} + \varepsilon^*\delta_2 + \sqrt{\bar\varepsilon}\delta_3\sqrt{\frac{m\varphi(\bar\varepsilon^{-1})}{n}}\right]$$

that holds with probability at least $1 - e^{-t}$. This yields the following upper bound on the stochastic term in (4.19) (see also (4.25)):

$$\begin{aligned} &2\langle \Xi, \hat S - S_*\rangle + 2(P - P_n)(S - S_*)(\hat S - S) + 2(P - P_n)(\hat S - S)^2 \le \qquad (4.39)\\ &2C\Bigg[\|\hat S - S\|_{L_2(\Pi^2)}\sqrt{\frac{m\varphi(\bar\varepsilon^{-1})}{n}} + \|\hat S - S\|_{L_2(\Pi^2)}\sqrt{\frac{\bar t}{n}} + \frac{\bar t}{n} + \\ &\varepsilon^*\|\mathcal{P}_L\hat S\|_1 + \sqrt{\bar\varepsilon}\|W^{1/2}(\hat S - S)\|_{L_2(\Pi^2)}\sqrt{\frac{m\varphi(\bar\varepsilon^{-1})}{n}}\Bigg] \end{aligned}$$

that holds provided that

$$\|\hat S - S\|_{L_2(\Pi^2)} \in [\delta_1^-, \delta_1^+], \|\mathcal{P}_L^\perp\hat S\|_1 \in [\delta_2^-, \delta_2^+], \|W^{1/2}(\hat S - S)\|_{L_2(\Pi^2)} \in [\delta_3^-, \delta_3^+]. \quad (4.40)$$

We substitute bound (4.39) in (4.19) and further bound some of its terms as follows:

$$2C\|\hat S - S\|_{L_2(\Pi^2)}\sqrt{\frac{m\varphi(\bar\varepsilon^{-1})}{n}} \le \frac{1}{8}\|\hat S - S\|^2_{L_2(\Pi^2)} + 8C^2\frac{m\varphi(\bar\varepsilon^{-1})}{n},$$

$$2C\|\hat S - S\|_{L_2(\Pi^2)}\sqrt{\frac{\bar t}{n}} \le \frac{1}{8}\|\hat S - S\|^2_{L_2(\Pi^2)} + 8C^2\frac{\bar t}{n},$$

and

$$2C\sqrt{\bar\varepsilon}\|W^{1/2}(\hat S - S)\|_{L_2(\Pi^2)}\sqrt{\frac{m\varphi(\bar\varepsilon^{-1})}{n}} \le \frac{1}{4}\bar\varepsilon\|W^{1/2}(\hat S - S)\|^2_{L_2(\Pi^2)} + 4C^2\frac{m\varphi(\bar\varepsilon^{-1})}{n}.$$

We will also use (4.23) to control the term $\varepsilon|\langle \mathrm{sign}(S), \hat S - S\rangle|$ in (4.19) and (4.24) to control the term $\bar\varepsilon|\langle W^{1/2}S, W^{1/2}(\hat S - S)\rangle|$. If condition (4.4) holds with $D \ge 32C$, then $\varepsilon \ge 2C\varepsilon^*$. It follows from (4.19) by a simple algebra that

$$\|\hat S - S_*\|^2_{L_2(\Pi^2)} \le \|S - S_*\|^2_{L_2(\Pi^2)} + C_1m^2\varepsilon^2\varphi(\bar\varepsilon^{-1}) + C_1\frac{m\varphi(\bar\varepsilon^{-1})}{n} + \bar\varepsilon\|W^{1/2}S\|^2_{L_2(\Pi^2)} + \frac{\bar t}{n}$$

with some constant $C_1 > 0$. Since, under condition (4.4) with $a = 1$, $m^2\varepsilon^2 \ge D^2\frac{m\log(2m)}{n} \ge D^2\frac{m}{n}$, we can conclude that

$$\|\hat S - S_*\|^2_{L_2(\Pi^2)} \le \|S - S_*\|^2_{L_2(\Pi^2)} + C_2m^2\varepsilon^2\varphi(\bar\varepsilon^{-1}) + \bar\varepsilon\|W^{1/2}S\|^2_{L_2(\Pi^2)} + \frac{\bar t}{n} \quad (4.41)$$

with some constant $C_2 > 0$.

We still have to choose the values of δ_k^-, δ_k^+ and to handle the case when conditions (4.40) do not hold. First note that due to the assumption that $\|S\|_{L_\infty} \leq 1, S \in \mathbb{D}$, we have $\|\hat{S} - S\|_{L_2(\Pi)} \leq 2$, $\|\mathcal{P}_L^\perp \hat{S}\|_1 \leq \|\hat{S}\|_1 \leq \sqrt{m}\|\hat{S}\|_2 \leq m^{3/2}$ and $\|W^{1/2}(\hat{S} - S)\|_{L_2(\Pi^2)} \leq 2\sqrt{\lambda_m}$. Thus, we can set $\delta_1^+ := 2, \delta_2^+ := m^{3/2}, \delta_3^+ := 2\sqrt{\lambda_m}$, which guarantees that the upper bounds of (4.40) are satisfied. We will also set $\delta_1^- = \delta_2^- := n^{-1/2}$, $\delta_3^- := \sqrt{\frac{\bar{\lambda}}{n}}$. In the case when one of the lower bounds of (4.40) does not hold, we can still use inequality (4.39), but we have to replace each of the norms $\|\hat{S} - S\|_{L_2(\Pi)}, \|\mathcal{P}_L^\perp \hat{S}\|_1, \|W^{1/2}(\hat{S} - S)\|_{L_2(\Pi^2)}$ which are smaller than the corresponding δ_k^- by the quantity δ_k^-. Then it is straightforward to check that inequality (4.41) still holds for some value of constant $C_2 > 0$. With the above choice of δ_k^-, δ_k^+, we have $\bar{t} \leq t + 3\log\left(2\log_2 n + \frac{1}{2}\log_2 \frac{\lambda_m}{\bar{\lambda}} + 2\right) = t_{n,m}$. This completes the proof.

□

References

[1] Ahlswede, R. and Winter, A. (2002) Strong converse for identification via quantum channels. *IEEE Transactions on Information Theory,* 48, 3, pp. 569–679.

[2] Aubin, J.-P. and Ekeland, I. (1984) Applied Nonlinear Analysis. J. Wiley&Sons, New York.

[3] Balcan, M.-F., Blum, A. and Srebro, N. (2008) Improved Guarantees for Learning via Similarity Functions. In: *Proc. 21st Annual Conference on Learning Theory*, COLT 2008.

[4] Candes, E. and Recht, B. (2009) Exact matrix completion via convex optimization. *Foundations of Computational Mathematics,* 9(6), 717–772.

[5] Candes, E. and Tao, T. (2010) The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56, 2053–2080.

[6] Candes, E. and Plan, Y. (2011) Tight Oracle Bounds for Low-Rank Matrix Recovery from a Minimal Number of Random Measurements. *IEEE Transactions on Information Theory*, 57(4), 2342–2359.

[7] Gaifas, S. and Lecué, G. (2011) Hyper-Sparse Optimal Aggregation. *J. Machine Learning Research*, 12, 1813–1833.

[8] de la Pena, V. and Giné, E. (1998) Decoupling: From Dependence to Independence, Springer, New York.

[9] Chen, Y., Garcia, E.K., Gupta, M.R., Rahimi, A. and Cazzanti, L. (2009) Similarity-based Classification: Concepts and Algorithms. *J. Machine Learning Research*, 10, 747–776.

[10] Gross, D. (2011) Recovering Low-Rank Matrices From Few Coefficients in Any Basis. *IEEE Transactions on Information Theory*, 57, 3, 1548–1566.

[11] Klopp, O. (2012) Noisy low-rank matrix completion with general sampling distribution. Preprint.

[12] Koltchinskii, V. (2011a) Von Neumann Entropy Penalization and Low Rank Matrix Estimation. *Annals of Statistics,* 39, 6, 2936–2973.

[13] Koltchinskii, V. (2011b) Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems, *Ecole d'ete de Probabilités de Saint-Flour 2008,* Lecture Notes in Mathematics, Springer.

[14] Koltchinskii, V. (2011c) A remark on low rank matrix recovery and noncommutative Bernstein type inequalities. Preprint.

[15] Koltchinskii, V. and Rangel, P. (2012) Low Rank Estimation of Similarities on Graphs. Preprint.

[16] Koltchinskii, V., Lounici, K. and Tsybakov, A. (2011) Nuclear norm penalization and optimal rates for noisy matrix completion. *Annals of Statistics,* 39, 5, 2302–2329.

[17] Leskovec, J., Huttenlocher, D. and Kleinberg, J. (2010) Predicting Positive and Negative Links in Online Social Networks. In: *Proc. 19th International Conference on World Wide Web*, ACM, New York.

[18] Maurer, A. (2008) Learning Similarity with Operator-valued Large-margin Classifiers. *J. Machine Learning Research*, 9, 1049–1082.

[19] Massart, P. (2000) Some applications of concentration inequalities to Statistics. *Annales de la Faculté des Sciences de Toulouse (6)*, 9, 2, 863–884.

[20] Massart, P. (2007) Concentration Inequalities and Model Selection. *Ecole d'ete de Probabilités de Saint-Flour 2003,* Lecture Notes in Mathematics, Springer.

[21] Negahban, S. and Wainwright, M.J. (2010) Restricted strong convexity and weighted matrix completion with noise. Preprint.

[22] Recht, B., Fazel, M. and Parrilo, P. (2010) Guaranteed minimum rank solutions of matrix equations via nuclear norm minimization. *SIAM Review*, 52, 3, 471–501.

[23] Rohde, A. and Tsybakov, A. (2011) Estimation of high-dimensional low rank matrices. *Annals of Statistics*, 39, 2, 887–930.

[24] Tropp, J.A. (2010) User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, to appear.

[25] Tsybakov, A.B. (2009) Introduction to Nonparametric Estimation. Springer.