Nonparametric estimation of low rank matrix valued function

Fan Zhou[*]

School of Mathematics
Georgia Institute of Technology
Atlanta, GA 30332-0160
e-mail: fzhou40@math.gatech.edu

Abstract: Let $A : [0,1] \to \mathbb{H}_m$ (the space of Hermitian matrices) be a matrix valued function which is low rank with entries in Hölder class $\Sigma(\beta, L)$. The goal of this paper is to study statistical estimation of A based on the regression model $\mathbb{E}(Y_j|\tau_j, X_j) = \langle A(\tau_j), X_j \rangle$, where τ_j are i.i.d. uniformly distributed in $[0,1]$, X_j are i.i.d. matrix completion sampling matrices, Y_j are independent bounded responses. We propose an innovative nuclear norm penalized local polynomial estimator and establish an upper bound on its point-wise risk measured by Frobenius norm. Then we extend this estimator globally and prove an upper bound on its integrated risk measured by L_2-norm. We also propose another new estimator based on bias-reducing kernels to study the case when A is not necessarily low rank and establish an upper bound on its risk measured by L_∞-norm. We show that the obtained rates are all optimal up to some logarithmic factor in minimax sense. Finally, we propose an adaptive estimation procedure based on Lepskii's method and model selection with data splitting technique, which is computationally efficient and can be easily implemented and parallelized on distributed systems.

MSC 2010 subject classifications: Primary 62G05, 62G08; secondary 62H12.
Keywords and phrases: Nonparametric estimation, low rank, matrix completion, nuclear norm penalization, local polynomial estimator, model selection, minimax lower bound.

Received March 2018.

1. Introduction

Let $A : [0,1] \to \mathbb{H}_m$ (the space of Hermitian matrices)[1] be a matrix valued function. The goal of this paper is to study the problem of statistical estimation of a matrix valued function A based on the regression model

$$\mathbb{E}(Y_j|\tau_j, X_j) = \langle A(\tau_j), X_j \rangle, \quad j = 1, \ldots, n, \tag{1.1}$$

where τ_j are i.i.d. random univariates uniformly distributed on $[0,1]$, X_j are i.i.d. matrix completion sampling matrices, Y_j are independent bounded random

---

[1]Note that we use $\mathbb{H}_m$ for the simplicity of presentation, and our results can be trivially generalize to regular matrix spaces such as $\mathbb{R}^{m_1 \times m_2}$.

responses. Sometimes, it is convenient to write model (1.1) in the form

$$Y_j = \langle A(\tau_j), X_j \rangle + \xi_j, \quad j = 1, \ldots, n, \tag{1.2}$$

where the noise variables $\xi_j = Y_j - \mathbb{E}(Y_j | \tau_j, X_j)$ are independent and have zero means. In particular, we are interested in the case where A is low rank and its entries belong to a standard function class $\Sigma(\beta, L)$ which is called Hölder class, see Definition 1. When $A(t) = A_0$ with some fixed $A_0 \in \mathbb{H}_m$ for any $t \in [0, 1]$, such a problem coincides with the well known matrix completion/recovery problem that has drawn a lot of attention in the statistics community during the past few years, see [8, 6, 7, 9, 18, 15, 20, 31, 29, 10] and the references therein. The low rank assumption in matrix completion/estimation problems has profound practical background. In the following, we discuss several simple examples of general low rank matrix valued functions that fit in our problem.

Example 1. Dynamic Collaborative Filtering Model. Let $P : [0, 1] \to \mathbb{R}^{m_1 \times r}$ and $Q : [0, 1] \to \mathbb{R}^{m_2 \times r}$, then $A = PQ^T$ is apparently a low rank matrix valued function when $r \ll (m_1 \wedge m_2)$. This kind of dynamic collaborative filtering model was initially introduced by [21], which generalized their well known work [22] to tackle the Netflix Prize on building dynamic recommender systems.

Example 2. Matrix Function Multiplication Model. Let $f : [0, 1] \to \mathbb{R}$ and $\tilde{A} \in \mathbb{R}^{m_1 \times m_2}$, where $\text{rank}(\tilde{A}) = r \ll (m_1 \wedge m_2)$. Then $A = \tilde{A} * f$ is a low rank matrix valued function. Such applications can be found in biology, chemistry and signal processing (see [38, 28]) where the underlying information $\tilde{A}$ is diffused via certain uniform diffusion function f.

Example 3. Euclidean Distance Matrix Model. Given the trajectory vectors of m points in $\mathbb{R}^d$, $p_1, ..., p_m : [0, 1] \to \mathbb{R}^d$. Then the Euclidean distance matrix (EDM) $D : [0, 1] \to \mathbb{R}^{m \times m}$ with $D_{ij} = \|p_i - p_j\|_2^2$ is a matrix valued function with rank at most $d + 2$ regardless of its size m. Clearly, when $m \gg d$, D falls into the low rank realm. In molecular biology, such points are typically in a low dimensional space such as $\mathbb{R}^2$ or $\mathbb{R}^3$. Similar topics in cases when points are fixed (see [39]) or in rigid motion (see [32]) have been studied.

An appealing way to address the low rank issue in matrix recovery problems is through nuclear norm minimization, see [30]. In section 3, we inherit this idea and propose a local polynomial estimator (see [13]) with nuclear norm penalization:

$$\widehat{S}^h = \arg\min_{S \in \mathbb{D}} \frac{1}{nh} \sum_{j=1}^{n} K\left(\frac{\tau_j - t_0}{h}\right) \left(Y_j - \left\langle \sum_{i=0}^{\ell} S_i p_i \left(\frac{\tau_j - t_0}{h}\right), X_j \right\rangle\right)^2 + \varepsilon \|S\|_1,$$
$$\tag{1.3}$$

where $\mathbb{D} \subset \mathbb{H}_{(\ell+1)m}$ is a closed subset of block diagonal matrices with $S_j \in \mathbb{H}_m$ on its diagonal, and $\{p_i(t)\}_{i=0}^{\infty}$ is a sequence of orthogonal polynomials with nonnegative weight function K. The solution to the convex optimization

problem (1.3) induces a pointwise estimator of $A(t_0)$:

$$\widehat{S}^h(t_0) := \sum_{i=0}^{\ell} \widehat{S}_i^h p_i(0)$$

where $\widehat{S}_i^h$ are the blocks on the diagonal of $\widehat{S}^h$ and $\ell = \lfloor \beta \rfloor$. We prove that under mild conditions, the pointwise risk measured by $m^{-2}\big\|\widehat{S}^h(t_0) - A(t_0)\big\|_2^2$ of $\widehat{S}^h(t_0)$ over Hölder class $\Sigma(\beta, L)$ satisfies the following upper bound

$$m^{-2}\big\|\widehat{S}^h(t_0) - A(t_0)\big\|_2^2 = O_p\Big(\Big[\frac{mr \log n}{n}\Big]^{\frac{2\beta}{2\beta+1}}\Big) \tag{1.4}$$

where r is the low rank parameter and $\|\cdot\|_2$ denotes the Frobenius norm of a matrix.

In section 4, we propose a new global estimator $\widehat{A}$ based on local polynomial smoothing and prove that the integrated risk of $\widehat{A}$ measured by L_2-norm satisfies the following upper bound

$$m^{-2}\int_0^1 \big\|\widehat{A}(t) - A(t)\big\|_2^2 dt = O_p\Big(\Big[\frac{mr \log n}{n}\Big]^{\frac{2\beta}{2\beta+1}}\Big). \tag{1.5}$$

Then we study another naive kernel estimator $\tilde{A}$ which can be used to estimate matrix valued functions which are not necessarily low rank. This estimator is associated with another popular approach to deal with low rank recovery which is called singular value thresholding, see [6, 20, 10]. We prove that the sup-norm risk of $\tilde{A}$ satisfies the following upper bound

$$\sup_{t\in[h,1-h]} m^{-2}\big\|\tilde{A}(t) - A(t)\big\|^2 = O_p\Big(\Big[\frac{m \log n}{n}\Big]^{\frac{2\beta}{2\beta+1}}\Big), \tag{1.6}$$

where $\|\cdot\|$ denotes the matrix spectral norm. Note that those rates coincide with that of classical matrix recovery setting when the smoothness parameter β goes to infinity.

An immediate question to ask is whether the above rates are optimal. In section 5, we prove that the rates in (1.4), (1.5) and (1.6) are all optimal up to some logarithmic factor in minimax sense, which essentially verified the effectiveness of our methodology.

As one may have noticed, there is an adaptation issue involved in (1.3). Namely, one needs to choose a proper bandwidth h and a proper order of degree ℓ of polynomials. Both parameters are closely related to the smoothness parameter β of A which is unknown to us in advance. In section 6, we propose a model selection procedure based on Lepskii's method ([25]) and the work of [3] and [37]. We prove that this procedure adaptively selects an estimator $\widehat{A}^*$ such that the integrated risk of $\widehat{A}^*$ measured by L_2-norm has the following upper bound

$$m^{-2}\int_0^1 \big\|\widehat{A}^*(t) - A(t)\big\|_2^2 dt = O_p\Big(\Big[\frac{mr \log n}{n}\Big]^{\frac{2\beta}{2\beta+1}}\Big) \tag{1.7}$$

which is still near optimal. What is more important, such a procedure is computationally efficient, feasible when m is comparable to n, and can be easily parallelized.

The major contribution of this paper is on the theory front. We generalized the recent developments of low rank matrix completion theory to nonparametric estimation setting by proposing an innovative optimal estimation procedure. To our best knowledge, no one has ever thoroughly studied such problems from a theoretical point of view.

2. Preliminaries

In this section, we introduce some important definitions, basic facts, and notations for the convenience of presentation.

2.1. Notations

For any Hermitian matrices $A, B \in \mathbb{H}_m$, we denote by $\langle A, B \rangle := \text{tr}(AB)$ the Hilbert-Schmidt inner product. Denote $\langle A, B \rangle_{L_2(\Pi)} = \mathbb{E}\langle A, X \rangle \langle B, X \rangle$, where Π denotes the distribution of X. The corresponding norm $\|A\|_{L_2(\Pi)}^2$ is given by $\|A\|_{L_2(\Pi)}^2 = \mathbb{E}\langle A, X \rangle^2$.

We use $\|\cdot\|_2$ to denote the Hilbert-Schmidt norm (Frobenius norm or Schatten 2-norm) induced by the inner product $\langle \cdot, \cdot \rangle$; $\| \cdot \|$ to denote the operator norm (spectral norm) of a matrix: the largest singular value; $\| \cdot \|_1$ to denote the trace norm (Schatten 1-norm or nuclear norm), i.e. the sum of singular values; $|A|$ to denote the nonnegative matrix with entries $|A_{ij}|$ corresponding to A.

Given $X_1,...,X_n$ as the i.i.d. copies of the random measurement matrix X, denote

$$\sigma_X^2 := \left\| n^{-1} \sum_{j=1}^{n} \mathbb{E} X_i^2 \right\|, \ U_X := \left\| \|X\| \right\|_{L_\infty},$$

where U_X denotes the L_∞-norm of the random variable $\|X\|$.

2.2. Matrix completion and statistical learning setting

The matrix completion setting refers to that the random sampling matrices X_j are i.i.d. uniformly distributed on the following orthonormal basis $\mathcal{X}$ of $\mathbb{H}_m$:

$$\mathcal{X} := \{E_{kj} : k, j = 1, ..., m\},$$

where $E_{kk} := e_k \otimes e_k$, $k = 1, ..., m$; $E_{jk} := \frac{1}{\sqrt{2}}(e_k \otimes e_j + e_j \otimes e_k)$, $1 \le k < j \le m$; $E_{kj} := \frac{i}{\sqrt{2}}(e_k \otimes e_j - e_j \otimes e_k)$, $1 \le k < j \le m$ with $\{e_j\}_{j=1}^{m}$ being the canonical basis of $\mathbb{R}^m$. The following identities are easy to check when the design matrices are under matrix completion setting:

$$\|A\|_{L_2(\Pi)}^2 = \frac{1}{m^2} \|A\|_2^2, \quad \sigma_X^2 \le \frac{2}{m}, \quad U_X = 1. \tag{2.1}$$

The statistical learning setting refers to the bounded response case: there exists a constant a such that

$$\max_{j=1,\dots n} |Y_j| \leq a, \ a.s. \tag{2.2}$$

In this paper, we will consider model (1.1) under both matrix completion and statistical learning setting.

2.3. Matrix valued function

Let $A : [0,1] \to \mathbb{H}_m$ be a matrix valued function. One should notice that we consider the image space to be Hermitian matrix space for the convenience of presentation. Our methods and results can be readily extended to general rectangular matrix space. Now we define the rank of a matrix valued function. Let $\mathrm{rank}_A(t) := \mathrm{rank}(A(t)), \ \forall t \in [0,1]$.

Definition 1. *Let β and L be two positive real numbers. The **Hölder class** $\Sigma(\beta, L)$ on $[0,1]$ is defined as the set of $\ell = \lfloor \beta \rfloor$ times differentiable functions $f : [0,1] \to \mathbb{R}$ with derivative $f^{(\ell)}$ satisfying*

$$|f^{(\ell)}(x) - f^{(\ell)}(x')| \leq L|x - x'|^{\beta - \ell}, \quad \forall x, x' \in [0,1]. \tag{2.3}$$

The parameters β and ℓ characterize the smoothness of Hölder class $\Sigma(\beta, L)$. They are the most important parameters in our problem just like the dimension of the matrix m and sample size n. Throughout this paper, we only consider the case when ℓ is a fixed constant, or in other words $\ell \ll m$. The reason is that in the asymptotic theory of low rank matrix recovery, the size of m is often considered to be comparable to the sample size n, say $m = O(n)$. If ℓ is also comparable to m, then our theory in this paper can be problematic.

In particular, we are interested in matrix valued functions satisfying the following assumptions:

A1 Given a measurement matrix X and for some constant a_1,

$$\sup_{t \in [0,1]} \left| \langle A(t), X \rangle \right| \leq a_1.$$

A2 Given a measurement matrix X and for some constant a_2, the derivative matrices $A^{(k)}$ of A satisfy

$$\sup_{t \in [0,1]} \left| \langle A^{(k)}(t), X \rangle \right| \leq a_2, \quad k = 1, \dots, \ell.$$

A3 The rank of $A, A', \dots, A^{(\ell)}$ are uniformly bounded by a constant r,

$$\sup_{t \in [0,1]} \mathrm{rank}_{A^{(k)}}(t) \leq r, \quad k = 0, 1, \dots, \ell.$$

A4 Assume that for $\forall i, j$, the entry A_{ij} is in the Hölder class $\Sigma(\beta, L)$.

3. A local polynomial Lasso estimator

In this section, we study the pointwise estimation of a low rank matrix valued function A in $\Sigma(\beta, L)$ with $\ell = \lfloor \beta \rfloor$. The construction of our estimator is inspired by local polynomial smoothing and nuclear norm penalization. The intuition of the localization technique originates from classical local polynomial estimators, see [13]. The intuition behind nuclear norm penalization is that whereas rank function counts the number of non-vanishing singular values, the nuclear norm sums their amplitude. The theoretical foundations behind using nuclear norm heuristic for the rank minimization were proved by [30]. Instead of using the trivial basis $\{1, t, t^2, ..., t^\ell\}$ to generate an estimator, we use orthogonal polynomials for some technical resasons that we will specify in the proof of Theorem 3.1. Let $\{p_i(t)\}_{i=0}^\infty$ be a sequence of orthogonal polynomials with nonnegative weight function K compactly supported on $[-1, 1]$, then

$$\int_{-1}^{1} K(u)p_i(u)p_j(u)du = \delta_{ij}$$

with $\delta_{ij} = \mathbf{1}\{i = j\}$ and $\mathbf{1}\{\cdot\}$ being the indicator function. It is easy to see that there exists an invertible linear transformation $T \in \mathbb{R}^{(\ell+1) \times (\ell+1)}$ such that

$$(1, t, t^2/2!, ..., t^\ell/\ell!)^T = T(p_0(t), p_1(t), ..., p_\ell(t))^T.$$

Apparently, T is lower triangular, and set $R(T) := \max_{1 \leq j \leq \ell+1} \sum_{i=1}^{\ell+1} |T_{ij}|$.

Let
$$\mathbb{D} := \left\{ \mathrm{Diag} \begin{bmatrix} S_0 & S_1 & \cdots & S_{\ell-1} & S_\ell \end{bmatrix} \right\} \subset \mathbb{H}_{m(\ell+1)}$$

be the set of block diagonal matrices with $S_k \in \mathbb{H}_m$ satisfying $|S_{ij}| \leq R(T)a$. With observations (τ_j, X_j, Y_j), $j = 1, ..., n$ from model (1.1), we define $\widehat{S}^h$ as

$$\widehat{S}^h := \arg\min_{S \in \mathbb{D}} \frac{1}{nh} \sum_{j=1}^{n} K\left(\frac{\tau_j - t_0}{h}\right)\left(Y_j - \left\langle \sum_{i=0}^{\ell} S_i p_i\left(\frac{\tau_j - t_0}{h}\right), X_j \right\rangle\right)^2 + \varepsilon \|S\|_1.$$

$$(3.1)$$

Remark 1. Note that one can rewrite (3.1) as

$$\widehat{S}^h = \arg\min_{S \in \mathbb{D}} \frac{1}{n} \sum_{j=1}^{n} \left(\tilde{Y}_j - \left\langle S, \tilde{X}_j \right\rangle\right)^2 + \varepsilon \|S\|_1, \qquad (3.2)$$

where $\tilde{X}_j = \mathrm{Diag}\left[\sqrt{\frac{1}{h}K\left(\frac{\tau_j - t_0}{h}\right)}p_0\left(\frac{\tau_j - t_0}{h}\right)X_j, ..., \sqrt{\frac{1}{h}K\left(\frac{\tau_j - t_0}{h}\right)}p_\ell\left(\frac{\tau_j - t_0}{h}\right)X_j\right]$, and $\tilde{Y}_j = \sqrt{\frac{1}{h}K\left(\frac{\tau_j - t_0}{h}\right)}Y_j$. Then (3.1) is a matrix Lasso type estimator.

$\widehat{S}^h$ naturally induces a local polynomial estimator of order ℓ around t_0:

$$\widehat{S}^h(\tau) := \sum_{i=0}^{\ell} \widehat{S}_i^h p_i\left(\frac{\tau - t_0}{h}\right)\mathbf{1}\left\{\left|\frac{\tau - t_0}{h}\right| \leq 1\right\}. \qquad (3.3)$$

The point estimate of A at t_0 is given by

$$\widehat{S}^h(t_0) := \sum_{i=0}^{\ell} \widehat{S}_i^h p_i(0). \tag{3.4}$$

Remark 2. Note that (3.1) only guarantees that each $\widehat{S}_i^h$ is approximately low rank and may not exactly recover the rank of $A^{(i)}(t_0)$. However, under our assumption that as long as ℓ is small compared with the matrix size m, then $\widehat{S}^h(t_0)$ is still approximately low rank.

In the following theorem, we establish an upper bound on the pointwise risk of $\widehat{S}^h(t_0)$ when $A(t)$ is in the Hölder class $\Sigma(\beta, L)$ with $\ell = \lfloor \beta \rfloor$. The proof of Theorem 3.1 can be found in section 8.1.

Theorem 3.1. *Under model (1.1), let (τ_j, X_j, Y_j), $j = 1, ..., n$ be i.i.d. copies of the random triplet (τ, X, Y) with X uniformly distributed in $\mathcal{X}$, τ uniformly distributed in $[0, 1]$, X and τ are independent, and $|Y| \leq a$, a.s. for some constant $a > 0$. Let A be a matrix valued function satisfying A1, A2, A3, and A4. Denote $\Phi = \max_{i=0,...,\ell} \|\sqrt{K} p_i\|_\infty$, and $\ell = \lfloor \beta \rfloor$. Take*

$$\widehat{h}_n = C_1 \Big(\frac{(\ell^3 (\ell!)^2 \Phi^2 R(T)^2 a^2 mr \log n}{L^2 n} \Big)^{\frac{1}{2\beta+1}}, \ \varepsilon = D\ell a\Phi \sqrt{\frac{\log 2m}{nm\widehat{h}_n}},$$

for some numerical constants C_1 and D. Then for any $\widehat{h}_n \leq t_0 \leq 1 - \widehat{h}_n$, the following bound holds with probability at least $1 - n^{-mr}$,

$$\frac{1}{m^2} \Big\| \widehat{S}^h(t_0) - A(t_0) \Big\|_2^2 \leq C_1(a, \Phi, \ell, L) \Big(\frac{mr \log n}{n} \Big)^{\frac{2\beta}{2\beta+1}}, \tag{3.5}$$

where $C_1(a, \Phi, \ell, L)$ is a constant depending on a, Φ, ℓ and L.

Remark 3. One should notice that when $\beta \to \infty$, bound (3.5) coincides with a similar result in classical matrix completion of which the rate is $O_p\big(\frac{mr \log m}{n} \big)$, see [20]. As long as n is of the polynomial order of m which is typical in reality, there is only up to a constant between $\log n$ and $\log m$. In section 5, we prove that bound (3.5) is minimax optimal up to a logarithmic factor. The logarithmic factor in bound (3.5) and bound of classical matrix completion is introduced by matrix Bernstein inequality, see [34]. In the case of nonparametric estimation of real valued function, it is unnecessary, see [35]. Currently, it still remains as an open problem whether this logarithmic factor is necessary or not for our problem as well as for classical low rank matrix estimation problem.

4. Global estimators and upper bounds on integrated risk

In this section, we propose two global estimators and study their integrated risk measured by L_2-norm and L_∞-norm.

4.1. From localization to globalization

Firstly, we construct a global estimator based on (3.3). Take

$$\widehat{h}_n = C_1\Big(\frac{\ell^3(\ell!)^2\Phi^2 R(T)^2 a^2 mr\log n}{L^2 n}\Big)^{\frac{1}{2\beta+1}},\ M = \lceil 1/\widehat{h}_n\rceil.$$

Without loss of generality, assume that M is even. Denote $\widehat{S}_k^h(t)$ the local polynomial estimator around t_{2k-1} as in (3.3) by using orthogonal polynomials with $K(t) = \mathbf{1}\{-1 \le t \le 1\}$, where $t_{2k-1} = \frac{2k-1}{M}$, $k = 1, 2, ..., M/2$ and $\mathbf{1}\{\cdot\}$ is the indicator function. Denote

$$\widehat{A}(t) = \sum_{k=1}^{M/2} \widehat{S}_k^h(t)\mathbf{1}\{t_{2k-1} - \widehat{h}_n < t \le t_{2k-1} + \widehat{h}_n\},\ \ t \in (0,1). \tag{4.1}$$

Note that the weight function K is not necessary to be $\mathbf{1}\{-1 \le t \le 1\}$. It can be replaced by any K that satisfies $K \ge K_0 > 0$ on $[-1, 1]$. The following result characterizes the integrated risk of estimator (4.1) under matrix completion setting measured by L_2-norm. The proof of Theorem 4.1 can be found in section 8.2.

Theorem 4.1. *Assume that the conditions of Theorem 3.1 hold, and let $\widehat{A}$ be an estimator defined as in (4.1). Then with probability at least $1 - n^{-(mr-1)}$,*

$$\frac{1}{m^2}\int_0^1 \big\|\widehat{A}(t) - A(t)\big\|_2^2 dt \le C_2(a, \Phi, \ell, L)\Big(\frac{mr\log n}{n}\Big)^{\frac{2\beta}{2\beta+1}}, \tag{4.2}$$

where $C_2(a, \Phi, \ell, L)$ is a constant depending on a, Φ, ℓ, L.

Remark 4. When the dimension m degenerates to 1, bound (4.2) matches the nonparametric minimax rate $O(n^{-2\beta/(2\beta+1)})$ for real valued functions over Hölder class (see [35]) up to some logarithmic factor, which again is introduced by the matrix Bernstein inequality, see [34]. In section 5, we show that bound (4.2) is minimax optimal up to a logarithmic factor.

4.2. Bias reduction through higher order kernels

If $A(t)$ is not necessarily low rank, we propose an estimator which is easy to implement and prove an upper bound on its risk measured by L_∞-norm. Such estimators are related to another popular approach parallel to local polynomial estimators for bias reduction, namely, using higher order kernels to reduce bias. They can also be applied to another important technique of low rank estimation or approximation via singular value thresholding, see [6] and [10]. The estimator through nuclear norm penalization is shown by [20] to be equivalent to soft singular value thresholding of such type of estimators.

The kernels we are interested in satisfy the following conditions:

K1 $K(\cdot)$ is symmetric, i.e. $K(u) = K(-u)$.

K2 $K(\cdot)$ is compactly supported on $[-1, 1]$.

K3 $R_K := \int_{-\infty}^{\infty} K^2(u) du < \infty$.

K4 $K(\cdot)$ is of order ℓ, where $\ell = \lfloor \beta \rfloor$.

K5 $K(\cdot)$ is Lipschitz continuous with Lipschitz constant $0 < L_K < \infty$.

Consider

$$\tilde{A}(t) = \frac{m^2}{nh} \sum_{j=1}^{n} K\Big(\frac{\tau_j - t}{h}\Big) Y_j X_j. \tag{4.3}$$

Note that when $K \geq 0$, (4.3) is the solution to the following convex optimization problem

$$\tilde{A}(t) = \arg\min_{S \in \mathbb{D}} \frac{1}{nh} \sum_{j=1}^{n} K\Big(\frac{\tau_j - t}{h}\Big) (Y_j - \langle S, X_j \rangle)^2. \tag{4.4}$$

In the following theorem we prove an upper bound on its global performance measured by L_∞-norm over $\Sigma(\beta, L)$. Such kind of bounds is much harder to obtain even for classical matrix lasso problems. The proof of Theorem 4.2 can be found in section 8.3.

Theorem 4.2. *Under model (1.1), let* (τ_j, X_j, Y_j), $j = 1, ..., n$ *be i.i.d. copies of the random triplet* (τ, X, Y) *with* X *uniformly distributed in* $\mathcal{X}$, τ *uniformly distributed in* $[0, 1]$, X *and* τ *are independent, and* $|Y| \leq a$ *a.s. for some constant* $a > 0$; *let* A *be any matrix valued function satisfying A1 and A4, and kernel* K *satisfies K1–K5. Denote* $\ell = \lfloor \beta \rfloor$. *Take*

$$\tilde{h}_n := c_*(K)\Big(\frac{a^2(\ell!)^2 m \log n}{2\beta L^2 n}\Big)^{\frac{1}{2\beta+1}}. \tag{4.5}$$

Then with probability at least $1 - n^{-2}$, *the estimator defined in (4.3) satisfies*

$$\sup_{t \in [\tilde{h}_n, 1-\tilde{h}_n]} \frac{1}{m^2} \big\| \tilde{A}(t) - A(t) \big\|^2 \leq C^*(K)\Big(\frac{a^2(\ell!)^2 m \log n}{2\beta L^2 n}\Big)^{\frac{2\beta}{2\beta+1}}, \tag{4.6}$$

where $C^*(K)$ *and* $c_*(K)$ *are constants depending on* K.

Remark 5. When m degenerates to 1, bound (4.6) coincides with that of real valued case over Hölder class, which is $O((\frac{\log n}{n})^{2\beta/(2\beta+1)})$, see [35]. Note that the logarithmic factor under such metric for real valued Hölder class is necessary. In section 5, we show that bound (4.6) is minimax optimal up to a logarithmic factor when $m \gg \log n$.

5. Lower bounds under matrix completion setting

In this section, we prove the minimax lower bounds corresponding to estimators (3.4), (4.1) and (4.3). In the realm of classical low rank matrix estimation, [29] studied the optimality issue measured by the Frobenius norm on the classes defined in terms of a "spikeness index" of the true matrix; [31] derived optimal

rates in noisy matrix completion on different classes of matrices for the empirical prediction error; [20] established the minimax rates of noisy matrix completion problems up to a logarithmic factor measured by the Frobenius norm. Based on the ideas of [20], standard methods to prove minimax lower bounds in real valued nonparametric estimation in [35], and some fundamental results in coding theory, we establish the corresponding minimax lower bounds of (3.5), (4.2) and (4.6) which essentially shows that the upper bounds we get are all optimal up to some logarithmic factor.

For the convenience of presentation, we denote by $\inf_{\widehat{A}}$ the infimum over all estimators of A. We denote by $\mathcal{A}(r, a)$ the set of matrix valued functions satisfying A1, A2, A3, and A4. We denote by $\mathcal{P}(r, a)$ the class of distributions of random triplet (τ, X, Y) that satisfies model (1.1) with any $A \in \mathcal{A}(r, a)$.

In the following theorem, we show the minimax lower bound on the pointwise risk. The proof of Theorem 5.1 can be found in section 8.4.

Theorem 5.1. *Under model (1.1), let (τ_j, X_j, Y_j), $j = 1, ..., n$ be i.i.d. copies of the random triplet (τ, X, Y) with X uniformly distributed in $\mathcal{X}$, τ uniformly distributed in $[0, 1]$, X and τ are independent, and $|Y| \leq a$, a.s. for some constant $a > 0$; let A be any matrix valued function in $\mathcal{A}(r, a)$. Then there is an absolute constant $\eta \in (0, 1)$ such that for all $t_0 \in [0, 1]$*

$$\inf_{\widehat{A}} \sup_{P^A_{\tau, X, Y} \in \mathcal{P}(r, a)} \mathbb{P}_{P^A_{\tau, X, Y}} \left\{ \frac{1}{m^2} \left\| \widehat{A}(t_0) - A(t_0) \right\|_2^2 > C \left(\frac{mr}{n} \right)^{\frac{2\beta}{2\beta + 1}} \right\} \geq \eta, \qquad (5.1)$$

where $C := C(\beta, L, a)$ is a constant depending on β, L and a.

Remark 6. Note that compared with the upper bound (3.5), the lower bound (5.1) matches it up to a logarithmic factor. As a consequence, it shows that the estimator (3.4) achieves a near optimal minimax rate of pointwise estimation. Although, the result of Theorem 5.1 is under bounded response condition, it can be readily extended to the case when the noise in (1.2) is Gaussian.

In the following theorem, we show the minimax lower bound on the integrated risk measured by L_2-norm. The proof of Theorem 5.2 can be found in section 8.5.

Theorem 5.2. *Under model (1.1), let (τ_j, X_j, Y_j), $j = 1, ..., n$ be i.i.d. copies of the random triplet (τ, X, Y) with X uniformly distributed in $\mathcal{X}$, τ uniformly distributed in $[0, 1]$, X and τ are independent, and $|Y| \leq a$, a.s. for some constant $a > 0$; let A be any matrix valued function in $\mathcal{A}(r, a)$. Then there is an absolute constant $\eta \in (0, 1)$ such that*

$$\inf_{\widehat{A}} \sup_{P^A_{\tau, X, Y} \in \mathcal{P}(r, a)} \mathbb{P}_{P^A_{\tau, X, Y}} \left\{ \frac{1}{m^2} \int_0^1 \left\| \widehat{A}(t) - A(t) \right\|_2^2 dt > \tilde{C} \left(\frac{mr}{n} \right)^{\frac{2\beta}{2\beta + 1}} \right\} \geq \eta, \ (5.2)$$

where $\tilde{C} := \tilde{C}(\beta, L, a)$ is a constant depending on L, β and a.

Remark 7. The lower bound in (5.2) matches the upper bound we get in (4.2) up to a logarithmic factor. Therefore, it means that the estimator (4.1) achieves

a near optimal minimax rate on the integrated risk measured by L_2-norm. The result of Theorem 5.2 can be readily extended to the case when the noise in (1.2) is Gaussian.

Now we consider the minimax lower bound on integrated risk measured by L_∞-norm for general matrix valued functions without any rank information. Denote

$$\mathcal{A}(a) := \big\{ A(t) \in \mathbb{H}_m, \ \forall t \in [0,1] : \ |A_{ij}(t)| \le a, \ A_{ij} \in \Sigma(\beta, L) \big\}.$$

We denote by $\mathcal{P}(a)$ the class of distributions of random triplet (τ, X, Y) that satisfies model (1.1) with any $A \in \mathcal{A}(a)$.

In the following theorem, we show the minimax lower bound over $\mathcal{P}(a)$ and $\mathcal{A}(a)$ measured by L_∞-norm. The proof of Theorem 5.3 can be found in section 8.6.

Theorem 5.3. *Under model (1.1), let* (τ_j, X_j, Y_j), $j = 1, ..., n$ *be i.i.d. copies of the random triplet* (τ, X, Y) *with* X *uniformly distributed in* $\mathcal{X}$, τ *uniformly distributed in* $[0,1]$, X *and* τ *are independent, and* $|Y| \le a$, *a.s. for some constant* $a > 0$; *let* A *be any matrix valued function in* $\mathcal{A}(a)$. *Then there exist an absolute constant* $\eta \in (0,1)$ *such that*

$$\inf_{\widehat{A}} \ \sup_{P^A_{\tau,X,Y} \in \mathcal{P}(a)} \mathbb{P}_{P^A_{\tau,X,Y}} \bigg\{ \sup_{t \in (0,1)} \frac{1}{m^2} \big\| \widehat{A}(t) - A(t) \big\|^2 > \bar{C} \Big(\frac{m \vee \log n}{n} \Big)^{\frac{2\beta}{2\beta+1}} \bigg\} \ge \eta,$$

(5.3)

where $\bar{C} := \bar{C}(\beta, L, a)$ *is a constant depending on* β, L *and* a.

Remark 8. Recall that in the real valued case, the minimax lower bound measured by L_∞-norm over Hölder class is $O((\frac{\log n}{n})^{2\beta/(2\beta+1)})$, see [35]. According to bound (5.3), if dimension m degenerates to 1, we get the same result as in real valued case and it is optimal. While the dimension m is large enough such that $m \gg \log n$, the lower bound (5.3) shows that the estimator (4.3) achieves a near optimal minimax optimal rate up to a logarithmic factor.

6. Model selection

Despite the fact that estimators (3.4) and (4.1) achieve near optimal minimax rates in theory with properly chosen bandwidth h and order of degree ℓ, such parameters depend on quantities like β and L which are unknown to us in advance. In this section, we propose an adaptive estimation procedure to choose h and ℓ adaptively.

Two popular methods to address such problems are proposed in the past few decades. One is Lepskii's method, and the other is aggregation method. In the 1990s, many data-driven procedures for selecting the smoothing parameter h emerged. Among them, a series of papers stood out and shaped a method what is now called Lepskii's method. This method has been described in its general form and in great detail in [25]. Later, [24] proposed a bandwidth selection

procedure based on pointwise adaptation of a kernel estimator that achieves optimal minimax rate of pointwise estimation over Hölder class, and [23] proposed a new bandwidth selector that achieves optimal rates of convergence over Besov classes with spatially imhomogeneous smoothness. The basic idea of Lepskii's method is to choose a bandwidth from a geometric grid to get an estimator not very different from those indexed by smaller bandwidths on the grid. Although Lepskii's method is shown to give optimal rates in pointwise estimation over Hölder class in [24], it has a major defect when applied to our problem: the procedure already requires a huge amount of computational cost when real valued functions are replaced by matrix valued functions. Indeed, with Lepskii's method, in order to get a good bandwidth, one needs to compare all candidates indexed by smaller bandwidth with the target one, which leads to dramatically growing computational cost. Still, we have an extra parameter ℓ that needs to fit with h. As a result, we turn to aggregation method to choose a bandwidth from the geometric grid introduced by Lepskii's method, which is more computationally efficient for our problem. The idea of aggregation method can be briefly summarized as follows: one splits the data set into two parts; the first is used to build all candidate estimators and the second is used to aggregate the estimates to build a new one (aggregation) or select one (model selection) which is at least as good as the best among all candidates.

The model selection procedure we use was initially introduced by [3] in classical nonparametric estimation with bounded response. [37] generalized this method to the case where the noise can be unbounded but with a finite p-th moment for some $p > 2$. One can find a more detailed review on such penalization methods in [16].

Firstly, we introduce the geometric grid created by [24] where to conduct our model selection procedure. Assume that the bandwidth falls into the range $[h_{min}, h_{max}]$. Recall that the optimal bandwidth $\widehat{h}_n$ in theory is given as

$$\widehat{h}_n = C_1 \Big(\frac{\ell^3 \big(\ell!\Phi R(T)a\big)^2 mr \log n}{L^2 n}\Big)^{\frac{1}{2\beta+1}}. \tag{6.1}$$

Assume that $[\beta_*, \beta^*]$ and $[L_*, L^*]$ are the ranges of β, L to be considered respectively. Then $h_{\max}$ and $h_{\min}$ can be chosen as

$$h_{\max} = C_1 \Big(\frac{\ell^{*3} \big(\ell^*!\Phi R(T)a\big)^2 mr \log n}{L_*^2 n}\Big)^{\frac{1}{2\beta^*+1}},$$

and

$$h_{\min} = C_1 \Big(\frac{\ell_*^3 \big(\ell_*!\Phi R(T)a\big)^2 mr \log n}{L^{*2} n}\Big)^{\frac{1}{2\beta_*+1}}$$

where $\ell^* = \lfloor \beta^* \rfloor$ and $\ell_* = \lfloor \beta_* \rfloor$. When those ranges are not given, a natural upper bound of $h_{\max}$ is 1, and a typical choice of $h_{\min}$ can be set to $n^{-1/2}$.

Denote

$$d(h) = \sqrt{1 \vee 2\log\Big(\frac{h_{max}}{h}\Big)}, \quad d_n = \sqrt{2\log\Big(\frac{h_{\max}}{h_{\min}}\Big)}, \quad \alpha(h) = \frac{1}{\sqrt{d(h)}}.$$

Apparently, $d_n = O(\sqrt{\log n})$. Define grid $\mathcal{H}$ inductively by

$$\mathcal{H} := \left\{ h_k \in [h_{min}, h_{max}] : h_0 = h_{max}, h_{k+1} = \frac{h_k}{1 + \alpha(h_k)}, k = 0, 1, 2, ... \right\}. \quad (6.2)$$

$\{h_k\}$ on the grid $\mathcal{H}$ is a decreasing sequence and the sequence becomes denser as k grows.

Now, we consider possible choices of ℓ. A trivial candidate set is

$$\mathcal{L} := \{\lfloor \beta_* \rfloor, \lfloor \beta_* \rfloor + 1, ..., \lfloor \beta^* \rfloor\} \subset \mathbb{N}^*.$$

If the size of this set is large, one can shrink it through the correspondence (6.1) for each h_k. For example, if $n = \Theta(m^d)$ for some $d > 1$, one can choose ℓ_i such that $\lfloor \frac{(1-d^{-1})\log n^{-1}}{2\log h_k} - 0.5 \rfloor \le \ell_i \le \lfloor \frac{\log n^{-1}}{2\log h_k} - 0.5 \rfloor$, which indicates the more the data, the narrower the range. We denote the candidate set for ℓ as $\mathcal{L}$. Then the set

$$\tilde{\mathcal{H}} = \mathcal{H} \times \mathcal{L} := \{(h, \ell) : h \in \mathcal{H}, \ \ell \in \mathcal{L}\}$$

indexed a countable set of candidate estimators.

Remark 9. In general, selecting h is considered to be more challenging and important than selecting ℓ and ε. On one hand, one needs to select h from an interval which is an uncountable set compared with selecting ℓ from only a finite set of integers. On the other hand, the performance of the estimator is much more sensitive to different choices of h, namely, a very small change of h can lead to huge performance degradation. We shall see this through our simulation study in section 7.2. Once h and ℓ are chosen, one can get ε_i by plug in the value of (h_i, ℓ_i) to get the corresponding $\varepsilon_i = (\ell_i + 1)R(T)\Phi\sqrt{\frac{\log 2m}{nmh_i}}$.

Now we introduce our model selection procedure based on $\tilde{\mathcal{H}}$. We split the data (τ_j, X_j, Y_j), $j = 1, ..., 2n$, into two parts with equal size. The first part of the observations $\{(\tau_j, X_j, Y_j) : j \in \hbar_n\}$ contains n data points, which are randomly drawn without replacement from the original data set. We construct a sequence of estimators $\widehat{A}^k$, $k = 1, 2, ...$ based on the training data set $\hbar_n$ through (4.1) for each pair in $\tilde{\mathcal{H}}$. Our main goal is to select an estimator $\widehat{A}$ among $\{\widehat{A}^k\}$, which is as good as the one that has the smallest mean square error. We introduce an quantity π_k associated with each estimator $\widehat{A}^k$ which serves as a penalty term. We use the remaining part of the data set $\{(\tau_j, X_j, Y_j) : j \in \tau_n^\dagger\}$ to perform the selection procedure:

$$k^* = \arg\min_k \frac{1}{n} \sum_{j \in \tau_n^\dagger} (Y_j - \langle \widehat{A}^k(\tau_j), X_j \rangle)^2 + \frac{\pi_k}{n}. \quad (6.3)$$

Denote $\widehat{A}^* = \widehat{A}^{k^*}$ as the adaptive estimator. In practice, we suggest one to rank all estimators $\widehat{A}^k$ according to the following rule: 1. pairs with bigger h always have smaller index; 2. if two pairs have the same h, the one with smaller ℓ has smaller index. Our selection procedure can be summarized as Algorithm 1.

Algorithm 1: Model Selection Procedure.

1. Construct the geometric grid $\mathcal{H}$ defined in (6.2) and the candidate set $\tilde{\mathcal{H}}$;

2. Equally split the data set (τ_j, X_j, Y_j), $j = 1, ..., N$ into two parts $\hbar_n$ and $\tau_n^\dagger$ by randomly drawing without replacement;

3. For each pair in $\tilde{\mathcal{H}}$, construct an estimator $\widehat{A}^k$ defined in (4.1) using data in $\hbar_n$;

4. Perform the selection procedure in (6.3) using data in $\tau_n^\dagger$.

The selection procedure described in Algorithm 1 have several advantages: firstly, it chooses a global bandwidth instead of a local one; secondly, since our selection procedure as in (6.3) is only based on computation of entries of $\widehat{A}^k$, no matrix computation is involved in the last step, which can efficiently save computational cost when m is large; finally, step 3 and 4 can be easily parallelized on distributed platforms.

The following theorem shows that the integrated risk of $\widehat{A}^*$ measured by L_2-norm can be bounded by the smallest one among all candidates plus an extra term of order $O(n^{-1})$ which is negligible. The proof of Theorem 6.1 can be found in section 8.7.

Theorem 6.1. *Under model (1.1), let (τ_j, X_j, Y_j), $j = 1, ..., 2n$ be i.i.d. copies of the random triplet (τ, X, Y) with X uniformly distributed in $\mathcal{X}$, τ uniformly distributed in $[0, 1]$, X and τ are independent, and $|Y| \leq a$, a.s. for some constant $a > 0$; let A be a matrix valued function satisfying A1, A2, A3, and A4; let $\{\widehat{A}^k\}$ be a sequence of estimators constructed from $\tilde{\mathcal{H}}$; let $\widehat{A}^*$ be the adaptive estimator selected through Algorithm 1. Then with probability at least $1 - n^{-(mr-1)}$*

$$\frac{1}{m^2} \int_0^1 \left\| \widehat{A}^*(t) - A(t) \right\|_2^2 dt \leq 3 \min_k \left\{ \frac{1}{m^2} \int_0^1 \left\| \widehat{A}^k(t) - A(t) \right\|_2^2 dt + \frac{\pi_k}{n} \right\} + \frac{C(a)}{n},$$
(6.4)

where $C(a)$ is a constant depending on a.

Recall that $\mathrm{Card}(\mathcal{H}) = O(\log n)$, one can take $\pi_k = kmr$. Then $\pi_k \leq c_1 mr \log n$ uniformly for all k with some numerical constant c_1. According to Lepskii's method that at least one candidate in $\mathcal{H}$ gives the optimal bandwidth associated with the unknown smoothness parameter β, the following corollary is a direct consequence of Theorem 4.1 and 6.1, which shows that $\widehat{A}^*$ is adaptive.

Corollary 6.1. Assume that the conditions of Theorem 6.1 hold with $\pi_k = kmr$, and $n > mr \log n$. Then with probability at least $1 - n^{-(mr-1)}$

$$\frac{1}{m^2} \int_0^1 \left\| \widehat{A}^*(t) - A(t) \right\|_2^2 dt \leq C(a, \ell, L) \left(\frac{mr \log n}{n} \right)^{\frac{2\beta}{2\beta+1}}$$
(6.5)

where $C(a, \ell, L)$ is a constant depending on a, ℓ, and L.

7. Numerical simulation

In this section, we present numerical simulation results of the estimators (3.1) and (4.1) to validate the theoretical bounds in (3.5), (4.2), (5.1), and (5.2). Then we present the simulation results of the model selection procedure shown in Algorithm 1. Recall that the key optimization problem we need to solve is (3.1). We develop a solver based on the well known alternating direction method of multipliers (ADMM) algorithm [5] and its applications to matrix recovery problems, see [27, 11]. The algorithm can be summarized as in Algorithm 2.

Algorithm 2: ADMM Algorithm.

Set up the values of $max\_Iteration$ and tolerance $\varepsilon_{tol} > 0$; Initialize $S^{(0)}, \bar{S}^{(0)} \in \mathbb{D}$ and $Z^{(0)} = \mathbf{0}$; **while** $k < max\_Iteration$ **do**

$\quad$ $S^{(k+1)} = \arg\min\limits_{S \in \mathbb{D}} \frac{1}{nh} \sum\limits_{j=1}^{n} K\left(\frac{\tau_j - t_0}{h}\right)\left(Y_j - \left\langle \sum\limits_{i=0}^{\ell} S_i p_i\left(\frac{\tau_j - t_0}{h}\right), X_j \right\rangle\right)^2 + \frac{\rho}{2}\|S -$

$\quad$ $\bar{S}^{(k)}\|_2^2 + \langle Z^{(k)}, S - \bar{S}^{(k)}\rangle;$

$\quad$ $\bar{S}^{(k+1)} = \arg\min\limits_{\bar{S} \in \mathbb{D}} \varepsilon\|\bar{S}\|_1 + \frac{\rho}{2}\|S^{(k+1)} - \bar{S}\|_2^2 + \langle Z^{(k)}, S^{(k+1)} - \bar{S}\rangle;$

$\quad$ $Z^{(k+1)} = Z^{(k)} + \rho(S^{(k+1)} - \bar{S}^{(k+1)});$

$\quad$ **if** $\|\bar{S}^{(k+1)} - \bar{S}^{(k)}\|_2^2 \leq \varepsilon_{tol}$ or $\|Z^{(k+1)} - Z^{(k)}\|_2^2 \leq \rho^2\varepsilon_{tol}$ **then**

$\quad$ $\quad$ | Reaching the tolerance;

$\quad$ **end**

$\quad$ Return $\bar{S}^{(k+1)}$. $k = k + 1;$

end

Return $\bar{S}^{(k+1)}$.

The underlying matrix valued function we create is in Hölder class $\Sigma(\beta, L)$ with $\beta = 3/2$, $L = 24$ and rank constraint $r \leq 3$. The orthogonal polynomial we choose is Chebyshev polynomials of the second kind.

7.1. Simulation results of theoretical bounds

We present the numerical simulation results to validates the theoretical bounds that we proved in section 3, 4 and 5. By plug in the optimal bandwidth in Theorem 3.1, we run Algorithm 2 to solve the pointwise estimator at $t_0 = 0.5$ with $m = 150$. Fig. 1a–Fig. 1g show different levels of recovery of the underlying true data matrix as in Fig. 1h. As we can see, the recovery quality increases evidently as sample size n grows.

In Fig. 2a, we display the comparison of pointwise risk between our theoretical bounds proved in (3.5), (5.1) and our simulation results. In Fig. 2b, we display the comparison of integrated risk measured by the L_2-norm between the theoretical bounds proved in (4.2), (5.2) and our simulation results. Since $\beta = 3/2$ and $\ell = 1$, we use piecewise linear polynomials to approximate the underlying matrix-valued function. Fig. 2a and 2b show that the simulation results match well with the minimax lower bound (5.1) and (5.2). One should notice that sometimes our simulated error rate is smaller than the theoretical minimax
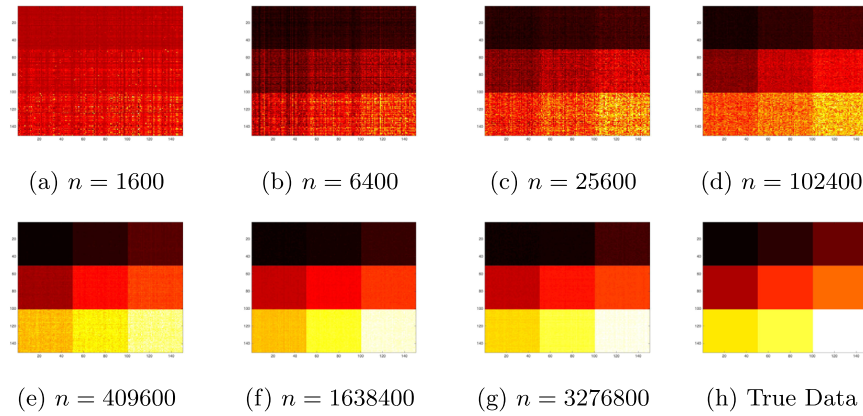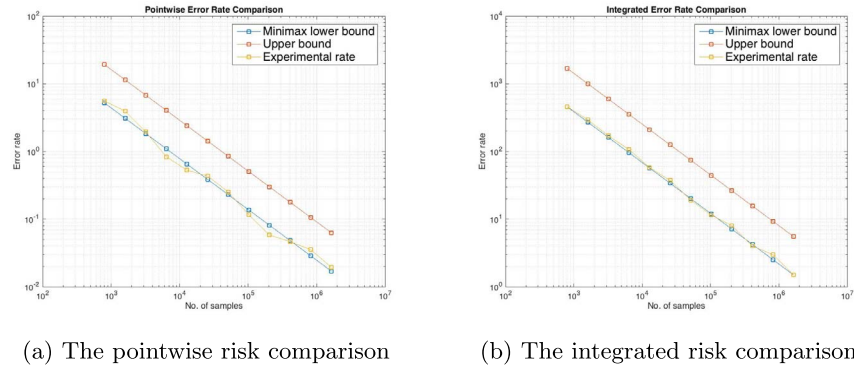
(a) $n = 1600$ (b) $n = 6400$ (c) $n = 25600$ (d) $n = 102400$

(e) $n = 409600$ (f) $n = 1638400$ (g) $n = 3276800$ (h) True Data

FIG 1. *Pointwise estimator at $t_0 = 0.5$ with different sample size.*

(a) The pointwise risk comparison   (b) The integrated risk comparison

FIG 2. *Error Rate Comparison.*

lower bound. We think the discrepency is due to the constant factors depending on a, L in the minimax lower bound that we computed are not very accurate.
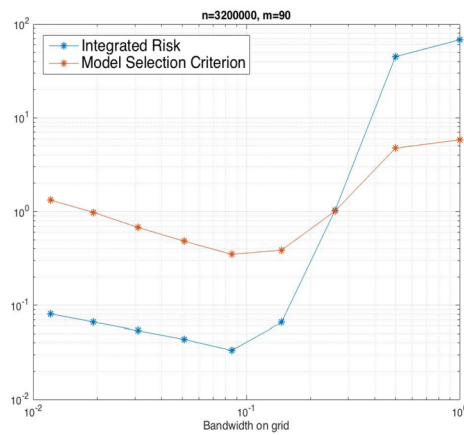
7.2. Simulation results of model selection

Recall that in section 6, we developed Algorithm 1 to adaptively choose parameters h and ℓ. Since the choice of ℓ is made through simply choosing one from a set of integers and is quite straight forward, and choosing a good bandwidth h is more critical and complicated, we focus on the choice of the smoothing parameter h in our simulation study. We set $\ell = 1$ which is the true parameter and focus on the selection of h.

We implement Algorithm 1 in this section, and perform simulation with $m = 90$ and $n = 3200000$. The theorectially optimal bandwidth h^* is around 0.09. We choose $h_{\max} = 1.0$ and $h_{\min} = 0.01$ to construct the geometric grid $\mathcal{H}$ as in (6.2). We display the simulation results in Table 1. To be more specific, we

TABLE 1
Model Selection.

Bandwidth on grid $\mathcal{H}$	Integrated risk	Model selection criterion
1.0000	68.1239	5.8238
0.5000	45.0275	4.7442
0.2602	1.0207	1.0100
0.1461	0.0657	0.3862
0.0853	**0.0333**	**0.3490**
0.0510	0.0437	0.4821
0.0311	0.0538	0.6741
0.0192	0.0663	0.9771
0.0121	0.0807	1.3199

FIG 3. *Model Selection on Grid $\mathcal{H}$.*

compute each global estimator as in (4.1) with each bandwidth on the $\mathcal{H}$. The corresponding integrated risks measured by L_2-norm are displayed in second column and our model selection criterion computed as in (6.3) are displayed in the third column. One should expect better integrated risk with smaller value of the third column. The data are plotted in Fig. 3. As we can see, our model selection procedure selects $\widehat{h} = 0.0853$ with the smallest criterion value of 0.3490, which shows that $\widehat{h}$ is very close to the optimal value of h. The corresponding integrated risk is also the smallest among all candidates on the grid and stays very close to the global minimum.

8. Proofs

8.1. Proof of Theorem 3.1

Proof. Firstly, we introduce a sharp oracle inequality of locally integrated risk of estimator (3.3) in the following lemma. The proof of Lemma 1 can be found in the appendix, which follows the same derivision as the proof of Theorem 19.1

in [19]. To be more specific, one just needs to rewrite (3.1) as

$$\widehat{S}^h = \arg\min_{S \in \mathbb{D}} \frac{1}{n} \sum_{j=1}^n \left(\tilde{Y}_j - \left\langle S, \tilde{X}_j \right\rangle \right)^2 + \varepsilon \|S\|_1, \tag{8.1}$$

where $\tilde{X}_j = \mathrm{Diag}\left[\sqrt{\frac{1}{h}K\left(\frac{\tau_j - t_0}{h}\right)}p_0\left(\frac{\tau_j - t_0}{h}\right)X_j, ..., \sqrt{\frac{1}{h}K\left(\frac{\tau_j - t_0}{h}\right)}p_\ell\left(\frac{\tau_j - t_0}{h}\right)X_j \right]$,

and $\tilde{Y}_j = \sqrt{\frac{1}{h}K\left(\frac{\tau_j - t_0}{h}\right)}Y_j$. Then the proof of Lemma 1 can be reproduced from the original proof with minor modifications. Since it is mostly tedious repeated arguments, we omit it here. One should notice that in the original proof of Theorem 19.1 in [19], a matrix isometry condition needs to be satisfied. That is $\|A\|_{L_2(\tilde{\Pi})}^2 = \mu_0 \|A\|_2^2$ for some constant $\mu_0 > 0$ and any $A \in \mathbb{H}_m$ with $\tilde{\Pi}$ being the distribution of $\tilde{X}$. One can easily check that it is true with (8.1). It is also the primary reason why we used orthogonal polynomials instead of the trivial basis $\{1, t, t^2, ...\}$.

Lemma 1. Assume that the condition of Theorem 3.1 holds. Then there exist a numerical constants $D > 0$ such that for all

$$\varepsilon \geq D(\ell + 1)R(T)\Phi a\left(\sqrt{\frac{\log 2m}{nmh}} \bigvee \frac{(\log 2m)\Phi}{nh} \right),$$

and for arbitrary $\eta > 0$, the estimator (3.4) satisfies with probability at least $1 - e^{-\eta}$

$$\frac{1}{h}\mathbb{E}K\left(\frac{\tau - t_0}{h}\right)\left\langle A(\tau) - \widehat{S}^h(\tau), X \right\rangle^2$$

$$\leq \inf_{S \in \mathbb{D}} \left\{ \frac{1}{h}\mathbb{E}K\left(\frac{\tau - t_0}{h}\right)\left\langle A(\tau) - S(\tau), X \right\rangle^2 \right. \tag{8.2}$$

$$\left. + \frac{D^2(\ell+1)^2\Phi^2 R(T)^2 a^2(\mathrm{rank}(S)m\log 2m + \eta)}{nh} \right\},$$

where $S(\tau) := \sum_{i=0}^\ell S_i p_i\left(\frac{\tau - t_0}{h}\right)$.

Then we consider

$$\frac{1}{h}\mathbb{E}K\left(\frac{\tau - t_0}{h}\right)\left\langle A(\tau) - \sum_{i=0}^\ell \widehat{S}_i^h p_i\left(\frac{\tau - t_0}{h}\right), X \right\rangle^2$$

$$= \frac{1}{h}\mathbb{E}K\left(\frac{\tau - t_0}{h}\right)\left\langle A(\tau) - \sum_{i=0}^\ell S_i p_i\left(\frac{\tau - t_0}{h}\right) + \sum_{i=0}^\ell (S_i - \widehat{S}_i^h) p_i\left(\frac{\tau - t_0}{h}\right), X \right\rangle^2$$

$$= \frac{1}{h}\mathbb{E}K\left(\frac{\tau - t_0}{h}\right)\left\langle \sum_{i=0}^\ell (S_i - \widehat{S}_i^h) p_i\left(\frac{\tau - t_0}{h}\right), X \right\rangle^2$$

$$+ \frac{1}{h}\mathbb{E}K\left(\frac{\tau - t_0}{h}\right)\left\langle A(\tau) - \sum_{i=0}^\ell S_i p_i\left(\frac{\tau - t_0}{h}\right), X \right\rangle^2$$

$$+ \frac{2}{h} \mathbb{E} K\Big(\frac{\tau - t_0}{h}\Big) \Big\langle A(\tau) - \sum_{i=0}^{\ell} S_i p_i\Big(\frac{\tau - t_0}{h}\Big), X \Big\rangle \Big\langle \sum_{i=0}^{\ell} (S_i - \widehat{S}_i^h) p_i\Big(\frac{\tau - t_0}{h}\Big), X \Big\rangle$$

$$(8.3)$$

Therefore, from (8.2) and (8.3), we have for any $S \in \mathbb{D}$

$$\frac{1}{h} \mathbb{E} K\Big(\frac{\tau - t_0}{h}\Big) \Big\langle \sum_{i=0}^{\ell} (S_i - \widehat{S}_i^h) p_i\Big(\frac{\tau - t_0}{h}\Big), X \Big\rangle^2$$

$$\leq \frac{2}{h} \mathbb{E} K\Big(\frac{\tau - t_0}{h}\Big) \Big| \Big\langle A(\tau) - \sum_{i=0}^{\ell} S_i p_i\Big(\frac{\tau - t_0}{h}\Big), X \Big\rangle \Big\langle \sum_{i=0}^{\ell} (S_i - \widehat{S}_i^h) p_i\Big(\frac{\tau - t_0}{h}\Big), X \Big\rangle \Big|$$

$$+ \frac{D^2 (\ell + 1)^2 \Phi^2 R(T)^2 a^2 (\mathrm{rank}(S) m \log 2m + \eta)}{nh}.$$

$$\leq \Big(\frac{c^4}{c^2 - 1}\Big) \frac{1}{h} \mathbb{E} K\Big(\frac{\tau - t_0}{h}\Big) \Big\langle A(\tau) - \sum_{i=0}^{\ell} S_i p_i\Big(\frac{\tau - t_0}{h}\Big), X \Big\rangle^2$$

$$+ \Big(\frac{c^2}{c^2 - 1}\Big) \Big\{ \frac{D^2 (\ell + 1)^2 \Phi^2 R(T)^2 a^2 (\mathrm{rank}(S) m \log 2m + \eta)}{nh} \Big\},$$

$$(8.4)$$

where we used the fact that for any positive constants a and b, $2ab \leq \frac{1}{c^2} a^2 + c^2 b^2$ for some $c > 1$. Take S such that

$$\sum_{i=0}^{\ell} S_i p_i\Big(\frac{\tau - t_0}{h}\Big) = A(t_0) + A'(t_0) h\Big(\frac{\tau - t_0}{h}\Big) + \ldots + \frac{A^{(\ell)}(t_0) h^\ell}{\ell!}\Big(\frac{\tau - t_0}{h}\Big)^\ell. \quad (8.5)$$

Note that this is possible since the right hand side is a matrix valued polynomial of $\frac{\tau - t_0}{h}$ up to order ℓ, and $\mathrm{span}\{p_0(t), p_1(t), \ldots, p_\ell(t)\} = \mathrm{span}\{1, t, \ldots, t^\ell\}$. Under the condition that all entries of $A^{(k)}(t)$ are bounded by a, then entries of S_k are bounded by $R(T)a$. Thus, the corresponding $S \in \mathbb{D}$. Obviously, $\mathrm{rank}(S_i) \leq (\ell + 1 - i)r$. Since $A \in \Sigma(\beta, L)$, we consider ℓ-th order Taylor expansion of A at t_0 to get

$$A(\tau) = A(t_0) + A'(t_0)(\tau - t_0) + \ldots + \frac{\tilde{A}(\tau - t_0)^\ell}{\ell!}, \quad (8.6)$$

where $\tilde{A}$ is the matrix with $\tilde{A}_{ij} = A_{ij}^{(\ell)}(t_0 + \alpha_{ij}(\tau - t_0))$ for some $\alpha_{ij} \in [0, 1]$. Then we apply the Taylor expansion (8.6) and identity (8.5) to get

$$\frac{1}{h} \mathbb{E} K\Big(\frac{\tau - t_0}{h}\Big) \Big\langle A(\tau) - \sum_{i=0}^{\ell} S_i p_i\Big(\frac{\tau - t_0}{h}\Big), X \Big\rangle^2$$

$$\leq \frac{1}{h} \mathbb{E} K\Big(\frac{\tau - t_0}{h}\Big) \frac{1}{m^2} \Big\| \frac{LU(\tau - t)^\beta}{\ell!} \Big\|_2^2 \leq \frac{L^2 h^{2\beta}}{(\ell!)^2},$$

$$(8.7)$$

where U denotes the matrix with all entries being 1. The first inequality is due to $A_{ij} \in \Sigma(\beta, L)$, and the second is due to $|\tau - t_0| \leq h$. Under the condition

that X is uniformly distributed in $\mathcal{X}$, and the orthogonality of $\{p_i(t)\}_{i=0}^{\ell}$, it is easy to check that

$$\frac{1}{h}\mathbb{E}K\Big(\frac{\tau-t_0}{h}\Big)\Big\langle\sum_{i=0}^{\ell}(S_i-\widehat{S}_i^h)p_i\Big(\frac{\tau-t_0}{h}\Big),X\Big\rangle^2 = \frac{1}{m^2}\sum_{i=0}^{\ell}\|\widehat{S}_i^h-S_i\|_2^2 \quad (8.8)$$

Note that

$$\left\|\widehat{S}^h(t_0)-S(t_0)\right\|_2^2 = \left\|\sum_{i=0}^{\ell}(\widehat{S}_i^h-S_i)p_i(0)\right\|_2^2 \le (\ell+1)\Phi^2\sum_{i=0}^{\ell}\left\|\widehat{S}_i^h-S_i\right\|_2^2, \quad (8.9)$$

where the second inequality is due to Cauchy-Schwarz inequality and $p_i(t)$ are uniformly bounded on $[-1,1]$. Combining (8.4), (8.7), (8.8), and (8.9), we get with probability at least $1-e^{-\eta}$

$$\frac{1}{m^2}\|\widehat{S}^h(t_0)-A(t_0)\|_2^2 \le \Big(\frac{c^4}{c^2-1}\Big)\frac{2L^2h^{2\beta}}{(\ell!)^2}$$
$$+\Big(\frac{c^2}{c^2-1}\Big)\Big\{\frac{D^2(\ell+1)^2\Phi^2R(T)^2a^2(\mathrm{rank}(S)m\log 2m+\eta)}{nh}\Big\}.$$

By optimizing the right hand side with respect to h and take $\eta = mr\log n$, we take

$$\widehat{h}_n = C\Big(\frac{\ell^3(\ell!)^2\Phi^2R(T)^2a^2mr\log n}{L^2n}\Big)^{\frac{1}{2\beta+1}},$$

where C is a numerical constant. This completes the proof of the theorem. $\quad\square$

8.2. Proof of Theorem 4.1

Proof. It is easy to see that

$$\int_0^1\|\widehat{A}(t)-A(t)\|_2^2dt \le \sum_{k=1}^{M/2}\int_{t_{2k-1}-\widehat{h}_n}^{t_{2k-1}+\widehat{h}_n}\|\widehat{S}_k^h(t)-A(t)\|_2^2dt. \quad (8.10)$$

For each k,

$$\frac{1}{m^2}\int_{t_{2k-1}-\widehat{h}_n}^{t_{2k-1}+\widehat{h}_n}\left\|\widehat{S}_k^h(t)-A(t)\right\|_2^2dt$$
$$= \mathbb{E}_{\tau,X}\mathbf{1}\{(t_{2k-1}-\widehat{h}_n,t_{2k-1}+\widehat{h}_n]\}\Big\langle A(\tau)-\widehat{S}^h(\tau),X\Big\rangle^2$$

By (8.2), (8.7) and arguments used to prove Theorem 3.1, we have with probability at least $1-\frac{1}{n^{mr}}$,

$$\frac{1}{m^2\widehat{h}_n}\int_{t_{2k-1}-\widehat{h}_n}^{t_{2k-1}+\widehat{h}_n}\left\|\widehat{S}_k^h(t)-A(t)\right\|_2^2dt \le C_1(a,\Phi,\ell,L)\Big(\frac{mr\log n}{n}\Big)^{\frac{2\beta}{2\beta+1}}.$$

Then take the union bound over k, from (8.10) we get with probability at least $1 - \frac{1}{n^{mr-1}}$,

$$\frac{1}{m^2} \int_0^1 \|\widehat{A}(t) - A(t)\|_2^2 dt \leq C_2(a, \Phi, \ell, L) \Big(\frac{mr \log n}{n}\Big)^{\frac{2\beta}{2\beta+1}},$$

where $C_2(a, \Phi, \ell, L)$ is a constant depending on a, Φ, ℓ, L. $\qquad\square$

8.3. *Proof of Theorem 4.2*

Proof. In this proof, we use $C(K)$ to denote any constant depending on K which may vary from place to place. This simplifies the presentation while does no harm to the soundness of our proof.

Consider

$$\sup_{t \in [\tilde{h}_n, 1-\tilde{h}_n]} \|\tilde{A}(t) - A(t)\| \leq \sup_{t \in [\tilde{h}_n, 1-\tilde{h}_n]} \|\tilde{A}(t) - \mathbb{E}\tilde{A}(t)\| + \sup_{t \in [\tilde{h}_n, 1-\tilde{h}_n]} \|\mathbb{E}\tilde{A}(t) - A(t)\|.$$
(8.11)

The first term on the right hand side is recognized as the variance and the second is the bias. Firstly, we deal with the bias term. Denote $B(t_0) := \mathbb{E}\tilde{A}(t_0) - A(t_0)$, $t_0 \in [\tilde{h}_n, 1 - \tilde{h}_n]$. Recall from (1.2), $\mathbb{E}(\xi_j | \tau_j, X_j) = 0$ for any $t_0 \in [\tilde{h}_n, 1 - \tilde{h}_n]$ we have

$$\mathbb{E}\tilde{A}(t_0) = \mathbb{E}\frac{m^2}{nh} \sum_{j=1}^n K\Big(\frac{\tau_j - t_0}{h}\Big)(\langle A(\tau_j), X_j \rangle + \xi_j)X_j$$

$$= \frac{m^2}{h}\mathbb{E}K\Big(\frac{\tau - t_0}{h}\Big)\langle A(\tau), X \rangle X.$$

By applying the Taylor expansion of $A(\tau)$ as in (8.6) and the fact that K is a kernel of order ℓ, we get

$$\mathbb{E}\tilde{A}(t_0) = \mathbb{E}\frac{m^2}{h}K\Big(\frac{\tau - t_0}{h}\Big)\langle A(t_0), X \rangle X + \mathbb{E}\frac{m^2}{h}K\Big(\frac{\tau - t_0}{h}\Big)\frac{(\tau - t_0)^\ell}{\ell!}\langle \tilde{A}, X \rangle X,$$

where $\tilde{A}$ is the same as in (8.6). It is easy to check that the first term on the right hand side is $A(t_0)$. Therefore we rewrite $B(t_0)$ as

$$B(t_0) = \mathbb{E}\frac{m^2}{h}K\Big(\frac{\tau - t_0}{h}\Big)\frac{(\tau - t_0)^\ell}{\ell!}\langle \tilde{A}, X \rangle X$$

$$= \mathbb{E}\frac{m^2}{h}K\Big(\frac{\tau - t_0}{h}\Big)\frac{(\tau - t_0)^\ell}{\ell!}\langle \tilde{A} - A^{(\ell)}(t_0), X \rangle X,$$

where the second equity is due to the fact that each element of $A(t)$ is in $\Sigma(\beta, L)$ and K is a kernel of order ℓ. Then we can bound each element of matrix $B(t_0)$

as

$$|B_{ij}(t_0)| \le \int_0^1 \frac{1}{h} K\Big(\frac{\tau - t_0}{h}\Big) \frac{|\tau - t_0|^\ell}{\ell!} |a_{ij}^{(\ell)}(t_0 + \alpha(\tau - t_0)) - a_{ij}^{(\ell)}(t_0)| d\tau$$

$$\le L \int_0^1 |K(u)| \frac{|uh|^\beta}{\ell!} du$$

$$\le C(K) \frac{Lh^\beta}{\ell!}.$$

Thus

$$\sup_{t \in [\tilde{h}_n, 1-\tilde{h}_n]} \|B(t)\| \le C(K) \frac{Lmh^\beta}{\ell!}. \tag{8.12}$$

On the other hand, for the variance term $\sup_{t \in [\tilde{h}_n, 1-\tilde{h}_n]} \|\tilde{A}(t) - \mathbb{E}\tilde{A}(t)\|^2$, we construct a $\delta - net$ on the interval $[0,1]$ with $\delta = 1/M$, and

$$M = n^2, \quad t_j = \frac{2j-1}{2M}, \ j = 1, ..., M.$$

Denote $S_n(t) := \tilde{A}(t) - \mathbb{E}\tilde{A}(t)$, then we have

$$\sup_{t \in [\tilde{h}_n, 1-\tilde{h}_n]} \|S_n(t)\| \le \sup_{t \in [0,1]} \|S_n(t)\| \le \max_i \|S_n(t_i)\| + \sup_{|t-t'| \le \delta} \|S_n(t) - S_n(t')\|. \tag{8.13}$$

Next, we bound both terms on the right hand side respectively. For each t_i,

$$S_n(t_i) = \frac{m^2}{nh} \sum_{j=1}^n \Big(K\Big(\frac{\tau_j - t_i}{h}\Big) Y_j X_j - \mathbb{E}K\Big(\frac{\tau_j - t_i}{h}\Big) Y_j X_j \Big).$$

The right hand side is a sum of zero mean random matrices, we apply the matrix Bernstein inequality, see [34]. Under the assumption of Theorem 4.2, one can easily check that with probability at least $1 - e^{-\eta}$,

$$\|S_n(t_i)\| \le C(K)m^2 \Big(\sqrt{\frac{a^2(\eta + \log 2m)}{mnh}} \bigvee \frac{a(\eta + \log 2m)}{nh} \Big).$$

Indeed, by setting $\bar{X} = m^2/hK\big(\frac{\tau-t}{h}\big)YX - \mathbb{E}m^2/hK\big(\frac{\tau-t}{h}\big)YX$, it is easy to check that $U_{\bar{X}} \lesssim \|K\|_\infty am^2/h$ and $\sigma_{\bar{X}}^2 \lesssim R_K a^2 m^3/h$. By taking the union bound over all i and setting $\eta = 4 \log n$, we get with probability at least $1 - n^{-2}$,

$$\max_i \|S_n(t_i)\|^2 \le C(K) \frac{a^2 m^3 \log n}{nh}.$$

As for the second term on the right hand side of (8.13), by the assumption that K is a Lipschitz function with Lipschitz constant L_K, we have

$$\sup_{|t-t'| \le \delta} \|S_n(t) - S_n(t')\| \le \sup_{|t-t'| \le \delta} \|(\tilde{A}(t) - \tilde{A}(t'))\| + \sup_{|t-t'| \le \delta} \|\mathbb{E}(\tilde{A}(t) - \tilde{A}(t'))\|$$

$$\le \frac{L_K am^3}{n^2 h^2} + \frac{L_K am}{n^2 h^2}.$$

Thus with probability at least $1 - n^{-2}$,

$$\sup_{t \in [\tilde{h}_n, 1 - \tilde{h}_n]} \|S_n(t)\|^2 \le C(K) \frac{a^2 m^3 \log n}{nh}$$

Together with the upper bound we get on the bias in (8.12), we have with probability at least $1 - n^{-2}$,

$$\sup_{t \in [\tilde{h}_n, 1 - \tilde{h}_n]} \frac{1}{m^2} \|\tilde{A}(t) - A(t)\|^2 \le C(K) \Big(\frac{a^2 m \log n}{nh} + \frac{L^2 h^{2\beta}}{\ell!^2} \Big).$$

Choose

$$\tilde{h}_n = C(K) \Big(\frac{a^2 (\ell!)^2 m \log n}{2\beta L^2 n} \Big)^{\frac{1}{2\beta+1}},$$

we get

$$\sup_{t \in [\tilde{h}_n, 1 - \tilde{h}_n]} \frac{1}{m^2} \|\tilde{A}(t) - A(t)\|^2 \le C(K) \Big(\frac{a^2 (\ell!)^2 m \log n}{2\beta L^2 n} \Big)^{\frac{2\beta}{2\beta+1}}. \qquad \Box$$

8.4. Proof of Theorem 5.1

Proof. Without loss of generality, we assume that both m and r are even numbers. We introduce several notations which are key to construct the hypothesis set. For some constant $\gamma > 0$, denote

$$\mathcal{C} = \big\{ \tilde{A} = (a_{ij}) \in \mathbb{C}^{\frac{m}{2} \times \frac{r}{2}} : a_{ij} \in \{0, \gamma\}, \forall 1 \le i \le m/2, 1 \le j \le r/2 \big\},$$

and consider the set of block matrices

$$\mathcal{B}(\mathcal{C}) = \left\{ \begin{bmatrix} \tilde{A} & \tilde{A} & \dots & \tilde{A} & O \end{bmatrix} \in \mathbb{C}^{\frac{m}{2} \times \frac{m}{2}} : \tilde{A} \in \mathcal{C} \right\}, \qquad (8.14)$$

where O denotes the $m/2 \times (m/2 - r\lfloor m/r \rfloor/2)$ zero matrix. Then we consider a subset of Hermitian matrices $\mathcal{S}_m \subset \mathbb{H}_m$,

$$\mathcal{S}_m = \left\{ \begin{bmatrix} \tilde{O} & \hat{A} \\ \hat{A}^* & \tilde{O} \end{bmatrix} \in \mathbb{C}^{m \times m} : \hat{A} \in \mathcal{B}(\mathcal{C}) \right\}. \qquad (8.15)$$

An immediate observation is that for any matrix $A \in \mathcal{S}_m$, $\text{rank}(A) \le r$.

Due to the Varshamov-Gilbert bound (see Lemma 2.9 in [35]), there exists a subset $\mathcal{A}^0 \subset \mathcal{S}_m$ with cardinality $\text{Card}(\mathcal{A}^0) \ge 2^{mr/32} + 1$ containing the zero $m \times m$ matrix $\mathbf{0}$ such that for any two distinct elements A_1 and A_2 of $\mathcal{A}^0$,

$$\|A_1 - A_2\|_2^2 \ge \frac{mr}{16} \lfloor \frac{m}{r} \rfloor \gamma^2 \ge \gamma^2 \frac{m^2}{32}. \qquad (8.16)$$

Let $f_n(t)$ denote the function $f_n(t) := Lh_n^\beta f\left(\frac{t-t_0}{h_n}\right)$, $t \in [0,1]$, where $h_n = c_0\left(\frac{mr}{n}\right)^{1/(2\beta+1)}$, with some $c_0 > 0$, and $f \in \Sigma(\beta, 1/2) \cap C^\infty$ and $\text{Supp}(f) = [-1/2, 1/2]$. Note that there exist functions f satisfying this condition. For instance, one can take

$$f(t) = \alpha e^{-\frac{1}{1-4u^2}} \mathbb{I}(|u| < 1/2), \tag{8.17}$$

for some sufficient small $\alpha > 0$. It is easy to check that $f_n(t) \in \Sigma(\beta, L)$ on $[0,1]$.

We consider the following hypotheses of A at t_0:

$$\mathcal{A}_0^\beta := \left\{ \widehat{A}(t) = Af_n(t), t \in [0,1] : A \in \mathcal{A}^0 \right\}.$$

The following claims are easy to check: firstly, any element in $\mathcal{A}_0^\beta$ together with its derivative have rank uniformly bounded by r, and the difference of any two elements of $\mathcal{A}_0^\beta$ satisfies the same property for fixed t_0; secondly, the entries of any element of $\mathcal{A}_0^\beta$ together with its derivative are uniformly bounded by some constant for sufficiently small chosen γ; finally, each element of $A(t) \in \mathcal{A}_0^\beta$ belongs to $\Sigma(\beta, L)$. Therefore, $\mathcal{A}_0^\beta \subset \mathcal{A}(r, a)$ with some chosen γ.

According to (8.16), for any two distinct elements $\widehat{A}_1(t)$ and $\widehat{A}_2(t)$ of $\mathcal{A}_0^\beta$, the difference between $\widehat{A}_1(t)$ and $\widehat{A}_2(t)$ at point t_0 is given by

$$\|\widehat{A}_1(t_0) - \widehat{A}_2(t_0)\|_2^2 \geq \frac{\gamma^2 L^2 c_0^{2\beta} f^2(0)}{32} m^2 \left(\frac{mr}{n}\right)^{\frac{2\beta}{2\beta+1}}. \tag{8.18}$$

On the other hand, we consider the joint distributions $P_{\tau,X,Y}^A$ such that $\tau \sim U[0,1]$, $X \sim \Pi_0$ where Π_0 denotes the uniform distribution on $\mathcal{X}$, τ and X are independent, and

$$\mathbb{P}_A(Y|\tau, X) = \begin{cases} \frac{1}{2} + \frac{\langle A(\tau), X\rangle}{4a}, & Y = a, \\ \frac{1}{2} - \frac{\langle A(\tau), X\rangle}{4a}, & Y = -a. \end{cases}$$

One can easily check that as long as $A(\tau) \in \mathcal{A}_0^\beta$, such $P_{\tau,X,Y}^A$ belongs to the distribution class $\mathcal{P}(r, a)$. We denote the corresponding n-product probability measure by $\mathbb{P}_A$. Then for any $A(\tau) \in \mathcal{A}_0^\beta$, the Kullback-Leibler Divergence between $\mathbb{P}_0$ and $\mathbb{P}_A$ is

$$K(\mathbb{P}_0, \mathbb{P}_A) = n\mathbb{E}\left(p_0(\tau, X) \log \frac{p_0(\tau, X)}{p_A(\tau, X)} + (1 - p_0(\tau, X)) \log \frac{1 - p_0(\tau, X)}{1 - p_A(\tau, X)}\right),$$

where $p_A(\tau, X) = 1/2 + \langle A(\tau), X\rangle/4a$. Note that $\mathbb{P}_A(Y = a|\tau, X) \in [1/4, 3/4]$ is guaranteed provided that $|\langle A(t), X\rangle| \leq a$. Thus by the inequality $-\log(1+u) \leq -u + u^2/2$, $\forall u > -1$, and the fact that $\mathbb{P}_A(Y = a|\tau, X) \in [1/4, 3/4]$, we have

$$K(\mathbb{P}_0, \mathbb{P}_A) \leq n\mathbb{E}2(p_0(\tau, X) - p_A(\tau, X))^2 \leq \frac{n}{8a^2}\mathbb{E}\langle A(\tau), X\rangle^2.$$

Recall that $A(\tau) = Af_n(\tau) \in \mathcal{A}_0^\beta$, by $\tau \sim U[0,1]$ and $X \sim \Pi_0$, we have

$$K(\mathbb{P}_0, \mathbb{P}_A) \leq \frac{n}{8a^2} \frac{1}{m^2} L^2 \|f\|_2^2 h_n^{2\beta+1} m^2 \gamma^2 \leq \frac{L^2 \|f\|_2^2 c_0^{2\beta+1} \gamma^2}{8a^2} mr. \tag{8.19}$$

Therefore, provided the fact that $\mathrm{Card}(\mathcal{A}^0) \geq 2^{mr/32} + 1$, together with (8.19), we have

$$\frac{1}{\mathrm{Card}(\mathcal{A}_0^\beta) - 1} \sum_{A \in \mathcal{A}_0^\beta} K(\mathbb{P}_0, \mathbb{P}_A) \leq \alpha \log(\mathrm{Card}(\mathcal{A}_0^\beta) - 1) \tag{8.20}$$

is satisfied for any $\alpha > 0$ if γ is chosen as a sufficiently small constant. In view of (8.18) and (8.20), the lower bound (5.1) follows from Theorem 2.5 in [35]. $\quad\square$

8.5. Proof of Theorem 5.2

Proof. Without loss of generality, we assume that both m and r are even numbers. Take a real number $c_1 > 0$, define

$$M = \left\lceil c_1 \left(\frac{n}{mr}\right)^{\frac{1}{2\beta+1}} \right\rceil, \quad h_n = \frac{1}{2M}, \quad t_j = \frac{2j - 1}{2M},$$

and

$$\phi_j(t) = L h_n^\beta f\left(\frac{t - t_j}{h_n}\right), \quad j = 1, ...M, \quad t \in [0, 1],$$

where f is defined the same as in (8.17). Meanwhile, we consider the set of all binary sequences of length M: $\Omega = \left\{\omega = (\omega_1, ..., \omega_M), \omega_i \in \{0, 1\}\right\} = \{0, 1\}^M$. By Varshamov-Gilbert bound, there exists a subset $\Omega_0 = \{\omega^0, ..., \omega^N\}$ of Ω such that $\omega^0 = (0, ..., 0) \in \Omega_0$, and $d(\omega^j, \omega^k) \geq \frac{M}{8}$, $\forall\, 0 \leq j < k \leq N$, and $N \geq 2^{\frac{M}{8}}$, where $d(\cdot, \cdot)$ denotes the Hamming distance of two binary sequences. Then we define a collection of functions based on Ω_0: $\mathcal{E} = \left\{f_\omega(t) = \sum_{j=1}^{M} \omega_j \phi_j(t) : \omega \in \Omega_0\right\}$.

From the result of Varshamov-Gilbert bound, we know that $S := \mathrm{Card}(\mathcal{E}) = \mathrm{Card}(\Omega_0) \geq 2^{\frac{M}{8}} + 1$. It is also easy to check that for all $f_\omega,\, f_{\omega'} \in \mathcal{E}$,

$$\begin{aligned}
\int_0^1 (f_\omega(t) - f_{\omega'}(t))^2 dt &= \sum_{j=1}^{M} (\omega_j - \omega_j')^2 \int_{\Delta_j} \phi_j^2(t) dt \\
&= L^2 h_n^{2\beta+1} \|f\|_2^2 \sum_{j=1}^{M} (\omega_j - \omega_j')^2 \\
&\geq L^2 h_n^{2\beta} \|f\|_2^2 / 16,
\end{aligned} \tag{8.21}$$

where $\Delta_j = [(j-1)/M, j/M]$.

In what follows, we combine two fundamental results in coding theory: one is Varshamov-Gilbert bound ([14, 36]) in its general form of a q-ary code, the other is the volume estimate of Hamming balls. Let $A_q(n, d)$ denote the largest size of a q-ary code of block length n with minimal Hamming distance d.

Proposition 8.1. The maximal size of a $q - ary$ code of block length n with minimal Hamming distance $d = pn$, satisfies

$$A_q(n, d+1) \geq q^{n(1 - h_q(p))}, \tag{8.22}$$

where $p \in [0, 1 - 1/q]$, $h_q(p) = p \log_q(q - 1) - p \log_q p - (1 - p) \log_q(1 - p)$ is the $q - ary$ entropy function.

We now have all the elements needed in hand to construct our hypotheses set. Denote $\Omega_1 = \{\omega^1, ..., \omega^N\}$, which is a subset of Ω_0 without ω^0. We then consider a subset $\mathcal{E}_1$ of $\mathcal{E}$ which is given by $\mathcal{E}_1 := \left\{ f_\omega(t) = \sum\limits_{j=1}^{M} \omega_j \phi_j(t) : \omega \in \Omega_1 \right\}$. Clearly, $S_1 := \mathrm{Card}(\mathcal{E}_1) \geq 2^{M/8}$. Then we define a new collection of matrix valued functions as

$$\mathcal{C} = \left\{ \tilde{A} = (a_{ij}) \in \mathbb{C}^{\frac{m}{2} \times \frac{r}{2}} : a_{ij} \in \{\delta f_\omega : \omega \in \Omega_1\}, \right.$$
$$\left. \delta \in \mathbb{C}, \forall 1 \leq i \leq m/2, 1 \leq j \leq r/2 \right\}.$$

Obviously, the collection $\mathcal{C}$ is a S_1-ary code of block length $mr/4$. Thus, we can apply the result of Proposition 8.1. It is easy to check that for $p = 1/4$, and $q \geq 4$

$$1 - h_q(p) = 1 - p \log_q \frac{q - 1}{p} + (1 - p) \log_q(1 - p) \geq \frac{1}{4}. \tag{8.23}$$

In our case, $q = S_1 \geq 2^{M/8}$ and $n = mr/4$. If we take $p = 1/4$, we know that

$$A_{S_1}(mr/4, mr/16) \geq A_{S_1}(mr/4, mr/16 + 1) \geq S_1^{mr/16}. \tag{8.24}$$

In other words, (8.24) guarantees that there exists a subset $\mathcal{H}^0 \subset \mathcal{C}$ with $\mathrm{Card}(\mathcal{H}^0) \geq 2^{Mmr/128}$ such that for any $A_1, A_2 \in \mathcal{H}^0$, the Hamming distance between A_1 and A_2 is at least $mr/16$. Now we define the building blocks of our hypotheses set

$$\mathcal{H} := \mathcal{H}^0 \cup \left\{ O_{\frac{m}{2} \times \frac{r}{2}} \right\},$$

where $O_{\frac{m}{2} \times \frac{r}{2}}$ is the $\frac{m}{2} \times \frac{r}{2}$ zero matrix. Obviously, $\mathcal{H}$ has size $\mathrm{Card}(\mathcal{H}) \geq 2^{Mmr/64} + 1$, and for any $A_1(t), A_2(t) \in \mathcal{H}$, the minimum Hamming distance is still greater than $mr/16$. We consider the set of matrix valued functions

$$\mathcal{B}(\mathcal{H}) = \left\{ \begin{bmatrix} \tilde{A} & \tilde{A} & \dots & \tilde{A} & O \end{bmatrix} : \tilde{A} \in \mathcal{H} \right\},$$

where O denotes the $m/2 \times (m/2 - r\lfloor m/r \rfloor/2)$ zero matrix. Finally, our hypotheses set of matrix valued functions $\mathcal{H}_m$ is defined as

$$\mathcal{H}_m = \left\{ \begin{bmatrix} \tilde{O} & \widehat{A} \\ \widehat{A}^* & \tilde{O} \end{bmatrix} \in \mathbb{C}^{m \times m} : \widehat{A} \in \mathcal{B}(\mathcal{H}) \right\}.$$

By the definition of $\mathcal{H}_m$ and similar to the arguments in proof of Theorem 5.1, it is easy to check that $\mathcal{H}_m \subset \mathcal{A}(r, a)$, and also

$$\mathrm{Card}(\mathcal{H}_m) \geq 2^{Mmr/64} + 1. \tag{8.25}$$

Now we consider any two different hypotheses $A_j(t), A_k(t) \in \mathcal{H}_m$.

$$\int_0^1 \|A_j(t) - A_k(t)\|_2^2 dt \geq \gamma^2 \frac{mr}{16} 2\left\lfloor \frac{m}{r} \right\rfloor \int_0^1 (f_\omega(t) - f_{\omega'}(t))^2 dt, \qquad (8.26)$$

where $\omega \neq \omega'$. Based on (8.21), we have

$$\frac{1}{m^2} \int_0^1 \|A_j(t) - A_k(t)\|_2^2 dt \geq \frac{\gamma^2 L^2 h_n^{2\beta} \|f\|_2^2}{256} \geq c_* \left(\frac{mr}{n}\right)^{\frac{2\beta}{2\beta+1}}, \qquad (8.27)$$

where c_* is a constant depending on $\|f\|_2$, L, c_1 and γ.

On the other hand, we repeat the same analysis on the Kullback-Leibler divergence $K(\mathbb{P}_0, \mathbb{P}_A)$ as in the proof of Theorem 5.1. One can get

$$K(\mathbb{P}_0, \mathbb{P}_A) \leq \frac{n}{8a^2} \mathbb{E}\langle A(\tau), X\rangle^2 \leq \frac{n}{8a^2} \gamma^2 \sum_{j=1}^M \int_0^1 \phi_j^2(\tau) d\tau \leq \frac{\gamma^2 c_1^{2\beta+1} L^2 Mmr \|f\|_2^2}{8a^2},$$

$$(8.28)$$

where $A(\tau) \in \mathcal{H}_m$. Combine (8.25) and (8.28) we know that

$$\frac{1}{\mathrm{Card}(\mathcal{H}_m) - 1} \sum_{A(t) \in \mathcal{H}_m} K(\mathbb{P}_0, \mathbb{P}_A) \leq \alpha \log(\mathrm{Card}(\mathcal{H}_m) - 1) \qquad (8.29)$$

is satisfied for any $\alpha > 0$ if γ is chosen as a sufficiently small constant. In view of (8.27) and (8.29), the lower bound follows from Theorem 2.5 in [35]. $\qquad \square$

8.6. Proof of Theorem 5.3

Proof. Without loss of generality, assume that m is an even number. For some constant $\gamma > 0$, denote $\mathcal{V} = \left\{ v \in \mathbb{C}^{\frac{m}{2}} : a_i \in \{0, \gamma\}, \forall\, 1 \leq i \leq m/2 \right\}$. Due to the Varshamov-Gilbert bound (see Lemma 2.9 in [35]), there exists a subset $\mathcal{V}^0 \subset \mathcal{V}$ with cardinality $\mathrm{Card}(\mathcal{V}^0) \geq 2^{m/16} + 1$ containing the zero vector $\mathbf{0} \in \mathbb{C}^{\frac{m}{2}}$, and such that for any two distinct elements v_1 and v_2 of $\mathcal{V}^0$,

$$\|v_1 - v_2\|_2^2 \geq \frac{m}{16}\gamma^2. \qquad (8.30)$$

Consider the set of matrices

$$\mathcal{B}(\mathcal{V}) = \left\{ \begin{bmatrix} v & v & \dots & v \end{bmatrix} \in \mathbb{C}^{\frac{m}{2} \times \frac{m}{2}} : v \in \mathcal{V}_0 \right\}.$$

Clearly, $\mathcal{B}(\mathcal{V})$ is a collection of rank one matrices. Then we construct another matrix set $\mathcal{V}_m$,

$$\mathcal{V}_m = \left\{ \begin{bmatrix} \tilde{O} & V \\ V^* & \tilde{O} \end{bmatrix} \in \mathbb{C}^{m \times m} : V \in \mathcal{B}(\mathcal{V}) \right\}$$

where $\tilde{O}$ is the $m/2 \times m/2$ zero matrix. Apparently, $\mathcal{V}_m \subset \mathbb{H}_m$.

On the other hand, we define the grid on $[0,1]$

$$M = \left\lceil c_2 \left(\frac{n}{m + \log n}\right)^{\frac{1}{2\beta+1}} \right\rceil, \quad h_n = \frac{1}{2M}, \quad t_j = \frac{2j-1}{2M},$$

and

$$\phi_j(t) = Lh_n^\beta f\left(\frac{t - t_j}{h_n}\right), \quad j = 1, ...M, \quad t \in [0,1]$$

where f is defined the same as in (8.17), and c_2 is some constant. Denote $\Phi := \left\{\phi_j : j = 1, ...M\right\}$. We consider the following set of hypotheses: $\mathcal{A}_\mathcal{B}^\beta := \{\widehat{A}(t) = V\phi_j(t) : V \in \mathcal{V}_m, \ \phi_j \in \Phi\}$. One can immediately get that the size of $\mathcal{A}_\mathcal{B}^\beta$ satisfies

$$\mathrm{Card}(\mathcal{A}_\mathcal{B}^\beta) \geq (2^{m/16} + 1)M. \tag{8.31}$$

By construction, the following claims are obvious: any element $\widehat{A}(t)$ of $\mathcal{A}_\mathcal{B}^\beta$ has rank at most 2; the entries of $\widehat{A}(t) \in \mathcal{A}_\mathcal{B}^\beta$ are uniformly bounded for some sufficiently small γ, and $\widehat{A}_{ij}(t) \in \Sigma(\beta, L)$. Thus $\mathcal{A}_\mathcal{B}^\beta \subset \mathcal{A}(a)$.

Now we consider the distance between two distinct elements $A(t)$ and $A'(t)$ of $\mathcal{A}_\mathcal{B}^\beta$. An immediate observation is that

$$\sup_{t\in[0,1]} \|A(t) - A'(t)\|^2 \geq \frac{1}{4} \sup_{t\in[0,1]} \|A(t) - A'(t)\|_2^2,$$

due to the fact that $\forall t \in (0,1)$, $\mathrm{rank}(A(t) - A'(t)) \leq 4$. Then we turn to get lower bound on $\sup_{t\in(0,1)} \|A(t) - A'(t)\|_2^2$. Recall that by construction of $\mathcal{A}_\mathcal{B}^\beta$, we have for any $A \neq A'$, $A(t) = A_1\phi_j(t)$, $A'(t) = A_2\phi_k(t)$, where $A_1, A_2 \in \mathcal{V}_m$. There are three cases need to be considered: 1). $A_1 \neq A_2$ and $j = k$; 2). $A_1 = A_2 \neq 0$ and $j \neq k$; 3). $A_1 \neq A_2$ and $j \neq k$.

For case 1,

$$\sup_{t\in[0,1]} \|A(t) - A'(t)\|_2^2 = \|A_1 - A_2\|_2^2 \|\phi_j\|_\infty^2$$

$$\geq \frac{m^2}{16}\gamma^2 L^2 h_n^{2\beta} \|f\|_\infty^2 \geq c_1^* m^2 \left(\frac{m + \log n}{n}\right)^{\frac{2\beta}{2\beta+1}},$$

where c_1^* is a constant depending on $\|f\|_\infty^2$, β, L and γ.

For case 2,

$$\sup_{t\in[0,1]} \|A(t) - A'(t)\|_2^2 = \|A_1\|_2^2 \|\phi_j - \phi_k\|_\infty^2$$

$$\geq \frac{m^2}{16}\gamma^2 L^2 h_n^{2\beta} \|f\|_\infty^2 \geq c_2^* m^2 \left(\frac{m + \log n}{n}\right)^{\frac{2\beta}{2\beta+1}},$$

where c_2^* is a constant depending on $\|f\|_\infty^2$, β, L and γ.

For case 3,

$$\sup_{t\in[0,1]} \|A(t) - A'(t)\|_2^2 \geq (\|A_1\|_2^2 \|\phi_j\|_\infty^2 \vee \|A_2\|_2^2 \|\phi_k\|_\infty^2)$$

$$\geq \frac{m^2}{16}\gamma^2 L^2 h_n^{2\beta} \|f\|_\infty^2 \geq c_3^* m^2 \Big(\frac{m + \log n}{n}\Big)^{\frac{2\beta}{2\beta+1}},$$

where c_3^* is a constant depending on $\|f\|_\infty^2$, β, L and γ.

Therefore, by the analysis above we conclude that for any two distinct elements $A(t)$ and $A'(t)$ of $\mathcal{A}_\mathcal{B}^\beta$,

$$\sup_{t\in[0,1]} \|A(t) - A'(t)\|^2 \geq \frac{1}{4}\sup_{t\in[0,1]} \|A(t) - A'(t)\|_2^2 \geq c_* m^2 \Big(\frac{m + \log n}{n}\Big)^{\frac{2\beta}{2\beta+1}},$$

$$(8.32)$$

where c_* is a constant depending on $\|f\|_\infty^2$, L, γ and β.

Meanwhile, we repeat the same analysis on the Kullback-Leibler divergence $K(\mathbb{P}_0, \mathbb{P}_A)$ as in the proof of Theorem 5.1. One can get that for any $A \in \mathcal{A}_\mathcal{B}^\beta$, the Kullback-Leibler divergence $K(\mathbb{P}_0, \mathbb{P}_A)$ between $\mathbb{P}_0$ and $\mathbb{P}_A$ satisfies

$$K(\mathbb{P}_0, \mathbb{P}_A) \leq \frac{n}{8a^2}\mathbb{E}|\langle A(\tau), X\rangle|^2 \leq \frac{n}{8a^2}\gamma^2 \int_0^1 \phi_j^2(\tau)d\tau$$

$$\leq \frac{\gamma^2 c_2^{2\beta+1} L^2 (m + \log n)\|f\|_2^2}{8a^2}.$$

$$(8.33)$$

Combine (8.31) and (8.33) we know that

$$\frac{1}{\mathrm{Card}(\mathcal{A}_\mathcal{B}^\beta) - 1}\sum_{A\in\mathcal{A}_\mathcal{B}^\beta} K(\mathbb{P}_0, \mathbb{P}_A) \leq \alpha \log(\mathrm{Card}(\mathcal{A}_\mathcal{B}^\beta) - 1) \qquad (8.34)$$

is satisfied for any $\alpha > 0$ if γ is chosen as a sufficiently small constant. In view of (8.32) and (8.34), the lower bound follows from Theorem 2.5 in [35]. $\qquad\square$

8.7. Proof of Theorem 6.1

Proof. For any $\widehat{A}^k$, denote the difference in empirical loss between $\widehat{A}^k$ and A by

$$r_n(\widehat{A}^k, A) := \frac{1}{n}\sum_{j=1}^n (Y_j - \langle \widehat{A}^k(\tau_j), X_j\rangle)^2 - \frac{1}{n}\sum_{j=1}^n (Y_j - \langle A(\tau_j), X_j\rangle)^2$$

$$= -\frac{1}{n}\sum_{j=1}^n U_j,$$

where $U_j = (Y_j - \langle A(\tau_j), X_j\rangle)^2 - (Y_j - \langle \widehat{A}^k(\tau_j), X_j\rangle)^2$. It is easy to check that

$$U_j = 2(Y_j - \langle A(\tau_j), X_j\rangle)\langle \widehat{A}^k(\tau_j) - A(\tau_j), X_j\rangle - \langle \widehat{A}^k(\tau_j) - A(\tau_j), X_j\rangle^2. \quad (8.35)$$

We denote $r(\widehat{A}^k, A) := \mathbb{E}\langle \widehat{A}^k(\tau) - A(\tau), X\rangle^2$. The following concentration inequality developed by [12] to prove Bernstein's inequality is key to our proof.

Lemma 2. Let U_j, $j = 1, ..., n$ be independent bounded random variables satisfying $|U_j - \mathbb{E}U_j| \le M$ with $h = M/3$. Set $\bar{U} = n^{-1}\sum_{j=1}^n U_j$. Then for all $t > 0$

$$\mathbb{P}\left\{\bar{U} - \mathbb{E}\bar{U} \ge \frac{t}{n\varepsilon} + \frac{n\varepsilon\text{var}(\bar{U})}{2(1-c)}\right\} \le e^{-t},$$

with $0 < \varepsilon h \le c < 1$.

Firstly, we bound the variance of U_j. Under the assumption that $|Y|$ and $|\langle A(\tau), X\rangle|$ are bounded by a constant a, one can easily check that $h = 8a^2/3$. Given $\mathbb{E}(Y_j|\tau_j, X_j) = \langle A(\tau_j), X_j\rangle$, we know that the covariance between the two terms on the right hand side of (8.35) is zero. Conditionally on (τ, X), the second order moment of the first term satisfies

$$4\mathbb{E}\sigma^2_{Y|\tau,X}\langle\widehat{A}^k(\tau_j) - A(\tau_j), X_j\rangle^2 \le 4a^2 r(\widehat{A}^k, A).$$

To see why, one can consider the random variable $\tilde{Y}$ with the distribution $\mathbb{P}\{\tilde{Y} = a\} = \mathbb{P}\{\tilde{Y} = -a\} = 1/2$. The variance of Y is always bounded by the variance of $\tilde{Y}$ which is a^2 under the assumption that $|Y_j|$ and $|\langle\widehat{A}^k(\tau_j), X_j\rangle|$ are bounded by a constant $a > 0$. Similarly, we can get that the variance of the second term conditioned on (τ, X) is also bounded by $4a^2\mathbb{E}\langle\widehat{A}^k(\tau_j) - A(\tau_j), X_j\rangle^2$. As a result, $n\text{var}(\bar{U}) \le 8a^2 r(\widehat{A}^k, A)$. By the result of Lemma 2, we have for any $\widehat{A}^k$ with probability at least $1 - e^{-t}$

$$r(\widehat{A}^k, A) - r_n(\widehat{A}^k, A) < \frac{t}{n\varepsilon} + \frac{4a^2\varepsilon r(\widehat{A}^k, A)}{1-c}.$$

Set $t = \varepsilon\pi_k + \log 1/\delta$, we get with probability at least $1 - \delta/e^{\varepsilon\pi_k}$

$$(1-\alpha)r(\widehat{A}^k, A) < r_n(\widehat{A}^k, A) + \frac{\pi_k}{n} + \frac{4a^2}{(1-c)\alpha}\left(\frac{\log 1/\delta}{n}\right),$$

where $\alpha = 4a^2\varepsilon/(1-c) < 1$. Denote

$$\tilde{k}^* = \arg\min_k\left\{r(\widehat{A}^k, A) + \frac{\pi_k}{n}\right\}.$$

By the definition of $\widehat{A}^*$, we have with probability at least $1 - \delta/e^{\varepsilon\widehat{\pi}^*}$

$$(1-\alpha)r(\widehat{A}^*, A) < r_n(\widehat{A}^{\tilde{k}^*}, A) + \frac{\pi_{\tilde{k}^*}}{n} + \frac{4a^2}{(1-c)\alpha}\left(\frac{\log 1/\delta}{n}\right), \tag{8.36}$$

where $\widehat{\pi}^*$ is the penalty terms associated with $\widehat{A}^*$.

Now we apply the result of Lemma 2 one more time and set $t = \log 1/\delta$, we get with probability at least $1 - \delta$

$$r_n(\widehat{A}^{\tilde{k}^*}, A) \le (1+\alpha)r(\widehat{A}^{\tilde{k}^*}, A) + \frac{4a^2}{(1-c)\alpha}\frac{\log 1/\delta}{n}. \tag{8.37}$$

Apply the union bound of (8.36) and (8.37), we get with probability at least $1 - \delta(1 + e^{-\varepsilon\widehat{\pi}^*})$

$$r(\widehat{A}^*, A) \le \frac{(1+\alpha)}{(1-\alpha)}\left(r(\widehat{A}^{\tilde{k}^*}, A) + \frac{\pi_{\tilde{k}^*}}{n}\right) + \frac{4a^2}{(1-c)\alpha(1-\alpha)}\frac{\log 1/\delta}{n}.$$

By taking $\varepsilon = 3/32a^2$ and $c = \varepsilon h$,

$$r(\widehat{A}^*, A) \le 3\left(r(\widehat{A}^{\tilde{k}^*}, A) + \frac{\pi_{\tilde{k}^*}}{n}\right) + \frac{64a^2}{3}\frac{\log 1/\delta}{n}.$$

By taking $\delta = 1/n^{mr}$ and adjusting the constant, we have with probability at least $1 - 1/n^{mr}$

$$\frac{1}{m^2}\int_0^1 \left\|\widehat{A}^*(t) - A(t)\right\|_2^2 dt$$

$$\le 3\min_k\left\{\frac{1}{m^2}\int_0^1 \left\|\widehat{A}^k(t) - A(t)\right\|_2^2 dt + \frac{\pi_k}{n}\right\} + C(a)\frac{mr\log n}{n}$$

where $C(a)$ is a constant depending on a. $\qquad\square$

Appendix: Proof of Lemma 1

The proof of Lemma 1 follows from a similar approach introduced by [19].

Proof. For any $S \in \mathbb{H}_m$ of rank r, $S = \sum_{j=1}^r \lambda_i(e_j \otimes e_j)$, where λ_j are non-zero eigenvalues of S (repeated with their multiplicities) and $e_j \in \mathbb{C}^m$ are the corresponding orthonormal eigenvectors. Denote $\text{sign}(S) := \sum_{j=1}^r \text{sign}(\lambda_i)(e_j \otimes e_j)$.

Let $\mathcal{P}_L$, $\mathcal{P}_L^\perp$ be the following orthogonal projectors in the space $(\mathbb{H}_m, \langle\cdot,\cdot\rangle)$:

$$\mathcal{P}_L(A) := A - P_{L^\perp}AP_{L^\perp}, \ \mathcal{P}_L^\perp(A) := P_{L^\perp}AP_{L^\perp}, \ \forall A \in \mathbb{H}_m$$

where P_L denotes the orthogonal projector on the linear span of $\{e_1, ..., e_r\}$, and $P_{L^\perp}$ is its orthogonal complement. Clearly, this formulation provides a decomposition of a matrix A into a "low rank part" $\mathcal{P}_L(A)$ and a "high rank part" $\mathcal{P}_L^\perp(A)$ if $\text{rank}(S) = r$ is small. Given $b > 0$, define the following cone in the space $\mathbb{H}_m$:

$$\mathcal{K}(\mathbb{D}; L; b) := \{A \in \mathbb{D} : \|\mathcal{P}_L^\perp A\|_1 \le b\|\mathcal{P}_L(A)\|_1\}$$

which consists of matrices with a "dominant" low rank part if S is low rank.

Firstly, we can rewrite (3.1) as

$$\widehat{S}^h = \arg\min_{S \in \mathbb{D}} \frac{1}{n}\sum_{j=1}^n \left(\tilde{Y}_j - \left\langle S, \tilde{X}_j \right\rangle\right)^2 + \varepsilon\|S\|_1, \tag{A.1}$$

where $\tilde{X}_j = \text{Diag}\left[\sqrt{\frac{1}{h}K\left(\frac{\tau_j - t_0}{h}\right)}p_0\left(\frac{\tau_j - t_0}{h}\right)X_j, ..., \sqrt{\frac{1}{h}K\left(\frac{\tau_j - t_0}{h}\right)}p_\ell\left(\frac{\tau_j - t_0}{h}\right)X_j\right],$

and $\tilde{Y}_j = \sqrt{\frac{1}{h}K\left(\frac{\tau_j - t_0}{h}\right)}Y_j$.

Denote the loss function as

$$\mathcal{L}\big(\tilde{Y}; \langle S(\tau), \tilde{X}\rangle\big) := \left(\tilde{Y}_j - \left\langle S, \tilde{X}_j\right\rangle\right)^2,$$

and the risk

$$PL\big(\tilde{Y}; \langle S(\tau), \tilde{X}\rangle\big) := \mathbb{E}\mathcal{L}\big(\tilde{Y}; \langle S(\tau), \tilde{X}\rangle\big) = \sigma^2 + \mathbb{E}\frac{1}{h}K\left(\frac{\tau - t_0}{h}\right)\big(Y - \langle S(\tau), X\rangle\big)^2$$

Since $\widehat{S}^h$ is a solution of the convex optimization problem (A.1), there exists a $\widehat{V} \in \partial\big\|\widehat{S}^h\big\|_1$, such that for $\forall S \in \mathbb{D}$ (see [2] Chap. 2)

$$\frac{2}{n}\sum_{j=1}^n\left(\left\langle\widehat{S}^h, \tilde{X}_j\right\rangle - \tilde{Y}_j\right)\langle\widehat{S}^h - S, \tilde{X}_j\rangle + \varepsilon\langle\widehat{V}, \widehat{S}^h - S\rangle \le 0.$$

This implies that, for all $S \in \mathbb{D}$,

$$\begin{aligned}
&\mathbb{E}\mathcal{L}'(\tilde{Y}; \langle\widehat{S}^h, \tilde{X}\rangle)\langle\widehat{S}^h - S, \tilde{X}\rangle + \varepsilon\langle\widehat{V}, \widehat{S}^h - S\rangle \\
&\le \mathbb{E}\mathcal{L}'(\tilde{Y}; \langle\widehat{S}^h, \tilde{X}\rangle)\langle\widehat{S}^h - S, \tilde{X}\rangle - \frac{2}{n}\sum_{j=1}^n(\langle\widehat{S}^h, \tilde{X}_j\rangle - \tilde{Y}_j)\langle\widehat{S}^h - S, \tilde{X}_j\rangle,
\end{aligned} \tag{A.2}$$

where $\mathcal{L}'$ denotes the partial derivative of $\mathcal{L}(y; u)$ with respect to u. One can easily check that for $\forall S \in \mathbb{D}$,

$$\begin{aligned}
&\mathbb{E}\mathcal{L}'(\tilde{Y}; \langle\widehat{S}^h, \tilde{X}\rangle)\langle\widehat{S}^h - S, \tilde{X}\rangle \\
&\ge \mathbb{E}(\mathcal{L}(\tilde{Y}; \langle\widehat{S}^h, \tilde{X}\rangle) - \mathcal{L}(\tilde{Y}; \langle S, \tilde{X}\rangle)) + \|\widehat{S}^h - S\|_{L_2(\tilde{\Pi})}^2,
\end{aligned} \tag{A.3}$$

where $\tilde{\Pi}$ denotes the distribution of $\tilde{X}$. If $\mathbb{E}\mathcal{L}(\tilde{Y}; \langle\widehat{S}^h, \tilde{X}\rangle) \le \mathbb{E}\mathcal{L}(\tilde{Y}; \langle S, \tilde{X}\rangle)$ for $\forall S \in \mathbb{D}$, then the oracle inequality in Lemma 1 holds trivially. So we assume that $\mathbb{E}\mathcal{L}(\tilde{Y}; \langle\widehat{S}^h, \tilde{X}\rangle) > \mathbb{E}\mathcal{L}(\tilde{Y}; \langle S, \tilde{X}\rangle)$ for some $S \in \mathbb{D}$. Thus, inequalities (A.2) and (A.3) imply that

$$\begin{aligned}
&\mathbb{E}\mathcal{L}(\tilde{Y}; \langle\widehat{S}^h, \tilde{X}\rangle) + \|\widehat{S}^h - S\|_{L_2(\tilde{\Pi})}^2 + \varepsilon\langle\widehat{V}, \widehat{S}^h - S\rangle \\
&\le \mathbb{E}\mathcal{L}(\tilde{Y}; \langle S, \tilde{X}\rangle) + \mathbb{E}\mathcal{L}'(\tilde{Y}; \langle\widehat{S}^h, \tilde{X}\rangle)\langle\widehat{S}^h - S, \tilde{X}\rangle \\
&- \frac{2}{n}\sum_{j=1}^n(\langle\widehat{S}^h, \tilde{X}_j\rangle - \tilde{Y}_j)\langle\widehat{S}^h - S, \tilde{X}_j\rangle.
\end{aligned} \tag{A.4}$$

According to the well known representation of subdifferential of nuclear norm, see [17] Sec. A.4, for any $V \in \partial\|S\|_1$, we have

$$V := \text{sign}(S) + \mathcal{P}_L^\perp(W), \ W \in \mathbb{H}_m \ , \|W\| \le 1.$$

By the duality between nuclear norm and operator norm

$$\langle \mathcal{P}_L^\perp(W), \widehat{S}^h - S \rangle = \langle \mathcal{P}_L^\perp(W), \widehat{S}^h \rangle = \langle W, \mathcal{P}_L^\perp(\widehat{S}^h) \rangle = \|\mathcal{P}_L^\perp(\widehat{S}^h)\|_1.$$

Therefore, by the monotonicity of subdifferentials of convex function $\| \cdot \|_1$, for any $V := \text{sign}(S) + \mathcal{P}_L^\perp(W) \in \partial\|S\|_1$, we have

$$\langle V, \widehat{S}^h - S \rangle = \langle \text{sign}(S), \widehat{S}^h - S \rangle + \|\mathcal{P}_L^\perp(\widehat{S}^h)\|_1 \leq \langle \widehat{V}, \widehat{S}^h - S \rangle, \qquad (A.5)$$

we can use (A.5) to change the bound in (A.4) to get

$$\begin{aligned}
&\mathbb{E}\mathcal{L}(\tilde{Y}; \langle \widehat{S}^h, \tilde{X} \rangle) + \|S - \widehat{S}^h\|_{L_2(\tilde{\Pi})}^2 + \varepsilon\|\mathcal{P}_L^\perp(\widehat{S}^h)\|_1 \\
&\leq \mathbb{E}\mathcal{L}(\tilde{Y}; \langle S, \tilde{X} \rangle) + \varepsilon\langle \text{sign}(S), S - \widehat{S}^h \rangle + \mathbb{E}\mathcal{L}'(\tilde{Y}; \langle \widehat{S}^h, \tilde{X} \rangle)\langle \widehat{S}^h - S, \tilde{X} \rangle \\
&- \frac{2}{n}\sum_{j=1}^n (\langle \widehat{S}^h, \tilde{X}_j \rangle - \tilde{Y}_j)\langle \widehat{S}^h - S, \tilde{X}_j \rangle.
\end{aligned} \qquad (A.6)$$

For the simplicity of representation, we use the following notation to denote the empirical process:

$$\begin{aligned}
&(P - P_n)(\mathcal{L}'(\tilde{Y}; \langle \widehat{S}^h, \tilde{X} \rangle))\langle \widehat{S}^h - S, \tilde{X} \rangle := \\
&\mathbb{E}\mathcal{L}'(\tilde{Y}; \langle \widehat{S}^h, \tilde{X} \rangle)\langle \widehat{S}^h - S, \tilde{X} \rangle - \frac{2}{n}\sum_{j=1}^n (\langle \widehat{S}^h, \tilde{X}_j \rangle - \tilde{Y}_j)\langle \widehat{S}^h - S, \tilde{X}_j \rangle.
\end{aligned} \qquad (A.7)$$

The following part of the proof is to derive an upper bound on the empirical process (A.7). Before we start with the derivation, let us present several vital ingredients that will be used in the following literature. For a given $S \in \mathbb{D}$ and for $\delta_1, \delta_2, \delta_3, \delta_4 \geq 0$, denote

$$\mathcal{A}(\delta_1, \delta_2) := \{A \in \mathbb{D} : A - S \in \mathcal{K}(\mathbb{D}; L; b), \|A - S\|_{L_2(\tilde{\Pi})} \leq \delta_1, \|\mathcal{P}_L^\perp A\|_1 \leq \delta_2\},$$

$$\tilde{\mathcal{A}}(\delta_1, \delta_2, \delta_3) := \{A \in \mathbb{D} : \|A - S\|_{L_2(\tilde{\Pi})} \leq \delta_1, \|\mathcal{P}_L^\perp A\|_1 \leq \delta_2, \|\mathcal{P}_L(A - S)\|_1 \leq \delta_3\},$$

$$\check{\mathcal{A}}(\delta_1, \delta_4) := \{A \in \mathbb{D} : \|A - S\|_{L_2(\tilde{\Pi})} \leq \delta_1, \|A - S\|_1 \leq \delta_4\},$$

and

$$\alpha_n(\delta_1, \delta_2) := \sup\{|(P - P_n)(\mathcal{L}'(\tilde{Y}; \langle A, \tilde{X} \rangle))\langle A - S, \tilde{X} \rangle| : A \in \mathcal{A}(\delta_1, \delta_2)\},$$

$$\tilde{\alpha}_n(\delta_1, \delta_2, \delta_3) := \sup\{|(P - P_n)(\mathcal{L}'(\tilde{Y}; \langle A, \tilde{X} \rangle))\langle A - S, \tilde{X} \rangle| : A \in \tilde{\mathcal{A}}(\delta_1, \delta_2, \delta_3)\},$$

$$\check{\alpha}_n(\delta_1, \delta_4) := \sup\{|(P - P_n)(\mathcal{L}'(\tilde{Y}; \langle A, \tilde{X} \rangle))\langle A - S, \tilde{X} \rangle| : A \in \check{\mathcal{A}}(\delta_1, \delta_4)\}.$$

Given the definitions above, Lemma 3 below shows upper bounds on the three quantities $\alpha_n(\delta_1, \delta_2)$, $\tilde{\alpha}_n(\delta_1, \delta_2, \delta_3)$, $\check{\alpha}_n(\delta_1, \delta_4)$. The proof of Lemma 3 can be found in section A.1. Denote

$$\Xi := n^{-1}\sum_{j=1}^n \varepsilon_j \tilde{X}_j \qquad (A.8)$$

where ε_j are i.i.d. Rademacher random variables.

Lemma 3. Suppose $0 < \delta_k^- < \delta_k^+$, $k = 1, 2, 3, 4$. Let $\eta > 0$ and

$$\bar{\eta} := \eta + \sum_{k=1}^{2} \log([\log_2(\frac{\delta_k^+}{\delta_k^-})] + 2) + \log 3,$$

$$\tilde{\eta} := \eta + \sum_{k=1}^{3} \log([\log_2(\frac{\delta_k^+}{\delta_k^-})] + 2) + \log 3,$$

$$\check{\eta} := \eta + \sum_{k=1, k=4} \log([\log_2(\frac{\delta_k^+}{\delta_k^-})] + 2) + \log 3.$$

Then with probability at least $1 - e^{-\eta}$, for all $\delta_k \in [\delta_k^-, \delta_k^+]$, $k = 1, 2, 3$

$$\alpha_n(\delta_1, \delta_2) \leq$$
$$\frac{C_1(\ell+1)R(T)\Phi a}{\sqrt{h}}\left\{\mathbb{E}\|\Xi\|(\sqrt{\text{rank}(S)}m\delta_1 + \delta_2) + \frac{2(\ell+1)R(T)\Phi a\bar{\eta}}{n\sqrt{h}} + \delta_1\sqrt{\frac{\bar{\eta}}{n}}\right\} \tag{A.9}$$

$$\tilde{\alpha}_n(\delta_1, \delta_2, \delta_3) \leq$$
$$\frac{C_2(\ell+1)R(T)\Phi a}{\sqrt{h}}\left\{\mathbb{E}\|\Xi\|(\delta_2 + \delta_3) + \frac{2(\ell+1)R(T)\Phi a\tilde{\eta}}{n\sqrt{h}} + \delta_1\sqrt{\frac{\tilde{\eta}}{n}}\right\} \tag{A.10}$$

$$\check{\alpha}_n(\delta_1, \delta_4) \leq$$
$$\frac{C_3(\ell+1)R(T)\Phi a}{\sqrt{h}}\left\{\mathbb{E}\|\Xi\|\delta_4 + \frac{2(\ell+1)R(T)\Phi a\check{\eta}}{n\sqrt{h}} + \delta_1\sqrt{\frac{\check{\eta}}{n}}\right\}, \tag{A.11}$$

where C_1, C_2, and C_3 are numerical constants.

Since both $\widehat{S}^h$ and S are in $\mathbb{D}$, by the definition of $\tilde{\alpha}$ and $\check{\alpha}$, we have

$$(P - P_n)(\mathcal{L}'(\tilde{Y}; \langle \widehat{S}^h, \tilde{X} \rangle))\langle \widehat{S}^h - S, \tilde{X} \rangle$$
$$\leq \tilde{\alpha}(\|\widehat{S}^h - S\|_{L_2(\tilde{\Pi})}; \|\mathcal{P}_L^\perp \widehat{S}^h\|_1; \|\mathcal{P}_L(\widehat{S}^h - S)\|_1), \tag{A.12}$$

and

$$(P - P_n)(\mathcal{L}'(\tilde{Y}; \langle \widehat{S}^h, \tilde{X} \rangle))\langle \widehat{S}^h - S, \tilde{X} \rangle \leq \check{\alpha}(\|\widehat{S}^h - S\|_{L_2(\tilde{\Pi})}; \|\widehat{S}^h - S\|_1). \tag{A.13}$$

If $\widehat{S}^h - S \in \mathcal{K}(\mathbb{D}; L; b)$, by the definition of α, we have

$$(P - P_n)(\mathcal{L}'(\tilde{Y}; \langle \widehat{S}^h, \tilde{X} \rangle))\langle \widehat{S}^h - S, \tilde{X} \rangle \leq \alpha(\|\widehat{S}^h - S\|_{L_2(\tilde{\Pi})}; \|\mathcal{P}_L^\perp \widehat{S}^h\|_1). \tag{A.14}$$

Assume for a while that

$$\|\widehat{S}^h - S\|_{L_2(\tilde{\Pi})} \in [\delta_1^-, \delta_1^+], \ \|\mathcal{P}_L^\perp \widehat{S}^h\|_1 \in [\delta_2^-, \delta_2^+], \ \|\mathcal{P}_L^\perp(\widehat{S}^h - S)\|_1 \in [\delta_3^-, \delta_3^+]. \tag{A.15}$$

By the definition of subdifferential, for any $\widehat{V} \in \partial\|\widehat{S}^h\|_1$,

$$\langle \widehat{V}, S - \widehat{S}^h \rangle \leq \|S\|_1 - \|\widehat{S}^h\|_1.$$

Then we apply (A.13) in bound (A.4) and use the upper bound on $\breve{\alpha}_n(\delta_1, \delta_4)$ of Lemma 3, and get with probability at least $1 - e^{-\eta}$,

$$
\begin{aligned}
&P(\mathcal{L}(\tilde{Y}; \langle \widehat{S}^h, \tilde{X} \rangle)) + \|\widehat{S}^h - S\|_{L_2(\tilde{\Pi})}^2 \\
&\leq P(\mathcal{L}(\tilde{Y}; \langle S, \tilde{X} \rangle)) + \varepsilon(\|S\|_1 - \|\widehat{S}^h\|_1) + \breve{\alpha}_n(\|\widehat{S}^h - S\|_{L_2(\tilde{\Pi})}, \|\widehat{S}^h - S\|_1) \\
&\leq P(\mathcal{L}(\tilde{Y}; \langle S, \tilde{X} \rangle)) + \varepsilon(\|S\|_1 - \|\widehat{S}^h\|_1) \\
&\quad + \frac{C_3(\ell+1)R(T)\Phi a}{\sqrt{h}}\Big\{ \mathbb{E}\|\Xi\|\|\widehat{S}^h - S\|_1 + \frac{2(\ell+1)R(T)\Phi a\breve{\eta}}{n\sqrt{h}} \\
&\quad + \|\widehat{S}^h - S\|_{L_2(\tilde{\Pi})}\sqrt{\frac{\breve{\eta}}{n}} \Big\}.
\end{aligned}
$$
(A.16)

Assuming that

$$
\varepsilon > \frac{C(\ell+1)R(T)\Phi a}{\sqrt{h}}\mathbb{E}\|\Xi\|,
$$
(A.17)

where $C = C_1 \vee 4C_2 \vee C_3$. From (A.16)

$$
P(\mathcal{L}(\tilde{Y}; \langle \widehat{S}^h, \tilde{X} \rangle)) \leq P(\mathcal{L}(\tilde{Y}; \langle S, \tilde{X} \rangle)) + 2\varepsilon\|S\|_1 + \frac{C_3(\ell+1)^2 R(T)^2 \Phi^2 a^2 \tilde{\eta}}{nh}.
$$
(A.18)

We now apply the upper bound on $\tilde{\alpha}_n(\|\widehat{S}^h - S\|_{L_2(\tilde{\Pi})}, \|\mathcal{P}_L^\perp \widehat{S}^h)\|_1, \|\mathcal{P}_L(\widehat{S}^h - S)\|_1)$ to (A.6) and get with probability at least $1 - e^{-\eta}$,

$$
\begin{aligned}
&P(\mathcal{L}(\tilde{Y}; \langle \widehat{S}^h, \tilde{X} \rangle)) + \|\widehat{S}^h - S\|_{L_2(\tilde{\Pi})}^2 + \varepsilon\|\mathcal{P}_L^\perp(\widehat{S}^h)\|_1 \\
&\leq P(\mathcal{L}(\tilde{Y}; \langle S, \tilde{X} \rangle)) + \varepsilon\|\mathcal{P}_L(\widehat{S}^h - S)\|_1 \\
&\quad + \tilde{\alpha}_n(\|\widehat{S}^h - S\|_{L_2(\tilde{\Pi})}, \|\mathcal{P}_L^\perp \widehat{S}^h)\|_1, \|\mathcal{P}_L(\widehat{S}^h - S)\|_1) \\
&\leq P(\mathcal{L}(\tilde{Y}; \langle S, \tilde{X} \rangle)) + \varepsilon\|\mathcal{P}_L(\widehat{S}^h - S)\|_1 \\
&\quad + \frac{C_2(\ell+1)R(T)\Phi a}{\sqrt{h}}\Big\{ \mathbb{E}\|\Xi\|(\|\mathcal{P}_L^\perp \widehat{S}^h)\|_1 + \|\mathcal{P}_L(\widehat{S}^h - S)\|_1) \Big\} \\
&\quad + \frac{C_2(\ell+1)^2 R(T)^2 \Phi^2 a^2 \tilde{\eta}}{nh},
\end{aligned}
$$
(A.19)

where the first inequality is due to the fact that

$$
\begin{aligned}
|\langle \text{sign}(S), S - \widehat{S}^h \rangle| &= |\langle \text{sign}(S), \mathcal{P}_L(S - \widehat{S}^h) \rangle| \\
&\leq \|\text{sign}(S)\|\|\mathcal{P}_L(S - \widehat{S}^h)\|_1 \leq \|\mathcal{P}_L(S - \widehat{S}^h)\|_1.
\end{aligned}
$$

With assumption (A.17) holds, we get from (A.19)

$$
\begin{aligned}
&P\mathcal{L}(\tilde{Y}; \langle \widehat{S}^h, \tilde{X} \rangle) + \varepsilon\|\mathcal{P}_L^\perp(\widehat{S}^h)\|_1 \\
&\leq P\mathcal{L}(\tilde{Y}; \langle S, \tilde{X} \rangle) + \frac{5\varepsilon}{4}\|\mathcal{P}_L(\widehat{S}^h - S)\|_1 + \frac{\varepsilon}{4}\|\mathcal{P}_L^\perp(\widehat{S}^h)\|_1 \\
&\quad + \frac{C_2(\ell+1)^2 R(T)^2 \Phi^2 a^2 \tilde{\eta}}{nh}.
\end{aligned}
$$
(A.20)

If the following is satisfied:

$$\frac{C_2(\ell+1)^2 R(T)^2 \Phi^2 a^2 \tilde{\eta}}{nh} \geq \frac{5\varepsilon}{4}\|\mathcal{P}_L(\widehat{S}^h - S)\|_1 + \frac{\varepsilon}{4}\|\mathcal{P}_L^{\perp}(\widehat{S}^h)\|_1, \qquad \text{(A.21)}$$

we can just conclude that

$$P(\mathcal{L}(\tilde{Y}; \langle \widehat{S}^h, \tilde{X} \rangle)) \leq P(\mathcal{L}(\tilde{Y}; \langle S, \tilde{X} \rangle)) + \frac{C_2(\ell+1)^2 R(T)^2 \Phi^2 a^2 \tilde{\eta}}{nh}, \qquad \text{(A.22)}$$

which is sufficient to meet the bound of Lemma 1. Otherwise, by the assumption that $P(\mathcal{L}(\tilde{Y}; \langle \widehat{S}^h, \tilde{X} \rangle)) > P(\mathcal{L}(\tilde{Y}; \langle S, \tilde{X} \rangle))$, one can easily check that

$$\|\mathcal{P}_L^{\perp}(\widehat{S}^h - S)\|_1 \leq 5\|\mathcal{P}_L(\widehat{S}^h - S)\|_1,$$

which implies that $\widehat{S}^h - S \in \mathcal{K}(\mathbb{D}; L; 5)$. This fact allows us to use the bound on $\alpha_n(\delta_1, \delta_2)$ of Lemma 3. We get from (A.6)

$$\begin{aligned}
&P(\mathcal{L}(\tilde{Y}; \langle \widehat{S}^h, \tilde{X} \rangle)) + \|\widehat{S}^h - S\|_{L_2(\tilde{\Pi})}^2 + \varepsilon\|\mathcal{P}_L^{\perp}(\widehat{S}^h)\|_1 \\
&\leq P(\mathcal{L}(\tilde{Y}; \langle S, \tilde{X} \rangle)) + \varepsilon\langle \text{sign}(S), S - \widehat{S}^h \rangle \\
&\quad + \frac{C_1(\ell+1)R(T)\Phi a}{\sqrt{h}}\mathbb{E}\|\Xi\|(\sqrt{\text{rank}(S)}m\|\widehat{S}^h - S\|_{L_2(\tilde{\Pi})} + \|\mathcal{P}_L^{\perp}(\widehat{S}^h)\|_1) \\
&\quad + \frac{C_1(\ell+1)^2 R(T)^2 \Phi^2 a^2 \bar{\eta}}{nh}.
\end{aligned} \qquad \text{(A.23)}$$

By applying the inequality

$$\left|\langle \text{sign}(S), \widehat{S}^h - S \rangle\right| \leq m\sqrt{\text{rank}(S)}\|\widehat{S}^h - S\|_{L_2(\tilde{\Pi})},$$

and the assumption (A.17), we have with probability at least $1 - e^{-\eta}$,

$$P(\mathcal{L}(\tilde{Y}; \langle \widehat{S}^h, \tilde{X} \rangle)) \leq P(\mathcal{L}(\tilde{Y}; \langle S, \tilde{X} \rangle)) + \varepsilon^2 m^2 \text{rank}(S) + \frac{C_1(\ell+1)^2 R(T)^2 \Phi^2 a^2 \bar{\eta}}{nh}. \qquad \text{(A.24)}$$

To sum up, the bound of Lemma 1 follows from (A.18), (A.22) and (A.24) provided that condition (A.17) and condition (A.15) hold.

We still need to specify δ_k^-, δ_k^+, $k = 1, 2, 3, 4$ to establish the bound of the theorem. By the definition of $\widehat{S}^h$, we have

$$P_n(\mathcal{L}(\tilde{Y}; \langle X, \widehat{S}^h \rangle)) + \varepsilon\|\widehat{S}^h\|_1 \leq P_n(\mathcal{L}(\tilde{Y}; \langle X, 0 \rangle)) \leq Q,$$

implying that $\|\widehat{S}^h\|_1 \leq \frac{Q}{\varepsilon}$. Next, $\|\mathcal{P}_L^{\perp}\widehat{S}^h\|_1 \leq \|\widehat{S}^h\|_1 \leq \frac{Q}{\varepsilon}$ and $\|\mathcal{P}_L(\widehat{S}^h - S)\|_1 \leq 2\|\widehat{S}^h - S\|_1 \leq \frac{2Q}{\varepsilon} + 2\|S\|_1$. Finally, we have $\|\widehat{S}^h - S\|_{L_2(\tilde{\Pi})} \leq 2a$. Thus, we can take $\delta_1^+ := 2a$, $\delta_2^+ := \frac{Q}{\varepsilon}$, $\delta_3^+ := \frac{2Q}{\varepsilon} + 2\|S\|_1$, $\delta_4^+ := \frac{Q}{\varepsilon} + \|S\|_1$. With these choices, δ_k^+, $k = 1, 2, 3, 4$ are upper bounds on the corresponding norms in condition (A.15). We choose $\delta_1^- := \frac{a}{\sqrt{n}}$, $\delta_2^- := \frac{a^2}{n\varepsilon} \wedge \frac{\delta_2^+}{2}$, $\delta_3^- := \frac{a^2}{n\varepsilon} \wedge \frac{\delta_3^+}{2}$, $\delta_4^- := \frac{a^2}{n\varepsilon} \wedge \frac{\delta_4^+}{2}$. Let $\eta^* := \eta + 3\log(B\log_2(\|S\|_1 \vee n \vee \varepsilon \vee a^{-1} \vee Q))$. It is easy to verify that

$\bar{\eta} \vee \tilde{\eta} \vee \tilde{\eta} \leq \eta^*$, for a proper choice of numerical constant B in the definition of η^*. When condition (A.15) does not hold, which means at least one of the numbers δ_k^-, $k = 1, 2, 3, 4$ we chose is not a lower bound on the corresponding norm, we can still use the bounds

$$
\begin{aligned}
&(P - P_n)(\mathcal{L}'(Y; \langle \widehat{S}^h, \tilde{X} \rangle)) \langle \widehat{S}^h - S, \tilde{X} \rangle \\
&\quad \leq \tilde{\alpha}(\|\widehat{S}^h - S\|_{L_2(\tilde{\Pi})} \vee \delta_1^-; \|\mathcal{P}_L^\perp \widehat{S}^h\|_1 \vee \delta_2^-; \|\mathcal{P}_L(\widehat{S}^h - S)\|_1 \vee \delta_3^-),
\end{aligned}
\tag{A.25}
$$

and

$$
\begin{aligned}
&(P - P_n)(\mathcal{L}'(Y; \langle \widehat{S}^h, \tilde{X} \rangle)) \langle \widehat{S}^h - S, \tilde{X} \rangle \\
&\hspace{4cm} \leq \check{\alpha}(\|\widehat{A}(t)^\varepsilon - S\|_{L_2(\tilde{\Pi})} \vee \delta_1^-; \|\widehat{S}^h - S\|_1 \vee \delta_4^-),
\end{aligned}
\tag{A.26}
$$

instead of (A.12), (A.13). In the case when $\widehat{S}^h - S \in \mathcal{K}(\mathbb{D}; L; 5)$, we can use the bound

$$
(P - P_n)(\mathcal{L}'(Y; \langle \widehat{S}^h, \tilde{X} \rangle)) \langle \widehat{S}^h - S, \tilde{X} \rangle \leq \alpha(\|\widehat{S}^h - S\|_{L_2(\tilde{\Pi})} \vee \delta_1^-; \|\mathcal{P}_L^\perp \widehat{S}^h\|_1 \vee \delta_2^-),
\tag{A.27}
$$

instead of bound (A.14). Then one can repeat the arguments above with only minor modifications. By the adjusting the constants, the result of Lemma 1 holds.

The last thing we need to specify is the size of ε which controls the nuclear norm penalty. Recall that from condition (A.17), the essence is to control $\mathbb{E}\|\Xi\|$. Here we use a simple but powerful noncommutative matrix Bernstein inequalities. The original approach was introduced by [1]. Later, the result was improved by [34] based on the classical result of [26]. We give the following lemma which is a direct consequence of the result proved by [34], and we omit the proof here.

Lemma 4. Under the model (1.1), Ξ is defined as in (A.8) with τ_j are i.i.d. uniformly distributed in [0,1], and ε_j are i.i.d. Rademacher random variables, and X_j are i.i.d uniformly distributed in $\mathcal{X}$. Then for any $\eta > 0$, with probability at least $1 - e^{-\eta}$

$$
\|\Xi\| \leq 4\left(\sqrt{\frac{(\eta + \log 2m)}{nm}} \vee \frac{(\eta + \log 2m)\Phi}{n\sqrt{h}}\right),
$$

and by integrating its exponential tail bounds

$$
\mathbb{E}\|\Xi\| \leq C\left(\sqrt{\frac{\log 2m}{nm}} \vee \frac{(\log 2m)\Phi}{n\sqrt{h}}\right)
$$

where C is a numerical constant.

Together with (A.17), we know for some numerical constant $D > 0$,

$$
\varepsilon \geq D \frac{\Phi a(\ell + 1)R(T)}{\sqrt{h}}\left(\sqrt{\frac{\log 2m}{nm}} \vee \frac{(\log 2m)\Phi}{n\sqrt{h}}\right),
$$

which completes the proof of Lemma 1. $\qquad \square$

A.1. Proof of Lemma 3

Proof. We only prove the first bound in detail, and proofs of the rest two bounds are similar with only minor modifications. By Talagrand's concentration inequality [33], and its Bousquet's form [4], with probability at least $1 - e^{-\eta}$,

$$\alpha_n(\delta_1, \delta_2) \le 2\mathbb{E}\alpha_n(\delta_1, \delta_2) + \frac{24(\ell+1)^2 R(T)^2 \Phi^2 a^2 \eta}{nh} + \frac{12(\ell+1)R(T)\Phi a \delta_1 \sqrt{\eta}}{\sqrt{nh}}.$$
(A.28)

By standard Rademacher symmetrization inequalities, see [17], Sec. 2.1, we can get

$$\mathbb{E}\alpha_n(\delta_1, \delta_2) \le 4\mathbb{E}\sup\left\{\left|\frac{1}{n}\sum_{j=1}^{n}\varepsilon_j(\langle A, \tilde{X}_j\rangle - \tilde{Y}_j)\langle A - S, \tilde{X}_j\rangle\right| : A \in \mathcal{A}(\delta_1, \delta_2)\right\},$$
(A.29)

where $\{\varepsilon_j\}$ are i.i.d. Rademacher random variables independent of $\{(\tau_j, X_j, \tilde{Y}_j)\}$. Then we consider the function $f(u) = (u - y + v)u$, where $|y| \le \frac{2\Phi a}{\sqrt{h}}$ and $|v|, |u| \le \frac{2(\ell+1)R(T)\Phi a}{\sqrt{h}}$. Clearly, this function has a Lipschitz constant $\frac{6(\ell+1)R(T)\Phi a}{\sqrt{h}}$. Thus by comparison inequality, see [17], Sec. 2.2, we can get

$$\mathbb{E}\sup\left\{\left|n^{-1}\sum_{j=1}^{n}\varepsilon_j(\langle A, \tilde{X}_j\rangle - \tilde{Y}_j)\langle A - S, \tilde{X}_j\rangle\right| : A \in \mathcal{A}(\delta_1, \delta_2)\right\}$$
$$\le \frac{6(\ell+1)R(T)\Phi a}{\sqrt{h}}\mathbb{E}\sup\left\{n^{-1}\left|\sum_{j=1}^{n}\varepsilon_i\langle A - S, \tilde{X}_j\rangle\right| : A \in \mathcal{A}(\delta_1, \delta_2)\right\}.$$
(A.30)

As a consequence of (A.29) and (A.30), we have

$$\mathbb{E}\alpha_n(\delta_1, \delta_2) \le \frac{12(\ell+1)R(T)\Phi a}{\sqrt{h}}\mathbb{E}\sup\left\{n^{-1}\left|\sum_{j=1}^{n}\varepsilon_i\langle A - S, \tilde{X}_j\rangle\right| : A \in \mathcal{A}(\delta_1, \delta_2)\right\}.$$
(A.31)

The next step is to get an upper bound on $\left|n^{-1}\sum_{j=1}^{n}\varepsilon_i\langle A - S, \tilde{X}_j\rangle\right|$. Recall that $\Xi := n^{-1}\sum_{j=1}^{n}\varepsilon_j\tilde{X}_j$, then we have $n^{-1}\sum_{j=1}^{n}\varepsilon_i\langle A - S, \tilde{X}_j\rangle = \langle A - S, \Xi\rangle$. One can check that

$$|\langle A - S, \Xi\rangle| \le |\langle \mathcal{P}_L(A - S), \mathcal{P}_L\Xi\rangle| + |\langle \mathcal{P}_L^\perp(A - S), \Xi\rangle|$$
$$\le \|\mathcal{P}_L\Xi\|_2\|\mathcal{P}_L(A - S)\|_2 + \|\Xi\|\|\mathcal{P}_L^\perp A\|_1$$
$$\le m\sqrt{2\text{rank}(S)}\|\Xi\|\|A - S\|_{L_2(\tilde{\Pi})} + \|\Xi\|\|\mathcal{P}_L^\perp A\|_1.$$

The second line of this inequality is due to Hölder's inequality and the third line is due to the facts that $(A - S) \in \mathcal{K}(\mathbb{D}; L; 5)$, $\text{rank}(\mathcal{P}_L(\Xi)) \le 2\text{rank}(S)$,

$\|\mathcal{P}_L \Xi\|_2 \le 2\sqrt{\mathrm{rank}(\mathcal{P}_L(\Xi))}\|\Xi\|$, and $\left\|A - S\right\|_{L_2(\tilde{\Pi})}^2 = \frac{1}{m^2}\left\|A - S\right\|_2$. Therefore,

$$
\begin{aligned}
&\frac{12(\ell+1)R(T)\Phi a}{\sqrt{h}}\mathbb{E}\sup\left\{\left|n^{-1}\sum_{j=1}^n \varepsilon_i\langle A - S, \tilde{X}_j\rangle\right| : A \in \mathcal{A}(\delta_1, \delta_2)\right\} \\
&\le \frac{12(\ell+1)R(T)\Phi a}{\sqrt{h}}\mathbb{E}\|\Xi\|(2\sqrt{2\mathrm{rank}(S)}m\delta_1 + \delta_2).
\end{aligned}
\tag{A.32}
$$

It follows from (A.28), (A.31) and (A.32) that with probability at least $1 - e^{-\eta}$,

$$
\begin{aligned}
\alpha_n(\delta_1, \delta_2) &\le \left(\frac{12(\ell+1)R(T)\Phi a}{\sqrt{h}}\mathbb{E}\|\Xi\|(\sqrt{\mathrm{rank}(S)}m\delta_1 + \delta_2)\right) \\
&\quad + \frac{24(\ell+1)^2 R(T)^2 \Phi^2 a^2 \eta}{nh} + \frac{12(\ell+1)R(T)\Phi a\delta_1\sqrt{\eta}}{\sqrt{nh}}.
\end{aligned}
$$

Now similar to the approach in [19], we make this bound uniform in $\delta_k \in [\delta_k^-, \delta_k^+]$. Let $\delta_k^{j_k} = \delta_k^+ 2^{-j_k}, j_k = 0, ..., [\log_2(\delta_k^+/\delta_k^-)] + 1, \ k = 1, 2$. By the union bound, with probability at least $1 - e^{-\eta}/3$, for all $j_k = 0, ..., [\log_2(\delta_k^+/\delta_k^-)] + 1, \ k = 1, 2$, we have

$$
\begin{aligned}
\alpha_n(\delta_1, \delta_2) &\le \left(\frac{12(\ell+1)R(T)\Phi a}{\sqrt{h}}\mathbb{E}\|\Xi\|(\sqrt{\mathrm{rank}(S)}m\delta_1^{j_1} + \delta_2^{j_2})\right) \\
&\quad + \frac{24(\ell+1)^2 R(T)^2 \Phi^2 a^2 \eta}{nh} + \frac{12(\ell+1)R(T)\Phi a\delta_1^{j_1}\sqrt{\eta}}{\sqrt{nh}},
\end{aligned}
$$

which implies that for all $\delta_k \in [\delta_k^-, \delta_k^+], \ k = 1, 2$,

$$
\begin{aligned}
\alpha_n(\delta_1, \delta_2) &\le \left(\frac{12(\ell+1)R(T)\Phi a}{\sqrt{h}}\mathbb{E}\|\Xi\|(\sqrt{\mathrm{rank}(S)}m\delta_1 + \delta_2)\right) \\
&\quad + \frac{24(\ell+1)^2 R(T)^2 \Phi^2 a^2 \bar{\eta}}{nh} + \frac{12(\ell+1)R(T)\Phi a\delta_1\sqrt{\bar{\eta}}}{\sqrt{nh}}.
\end{aligned}
$$

The proofs of the second and the third bounds are similar to this one, we omit the repeated arguments. □

Acknowledgements

References

[1] R. Ahlswede and A. Winter. Strong converse for identification via quantum channels. *IEEE Transactions on Information Theory*, 48(3):569–579, 2002. MR1889969

[2] J.-P. Aubin and I. Ekeland. *Applied nonlinear analysis.* Courier Corporation, 2006. MR2303896

[3] A. R. Barron. Complexity regularization with application to artificial neural networks. *Nonparametric Functional Estimation and Related Topics*, 335:561–576, 1991. MR1154352

[4] O. Bousquet. A Bennett concentration inequality and its application to suprema of empirical processes. *Comptes Rendus Mathematique*, 334(6):495–500, 2002. MR1890640

[5] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1):1–122, 2011.

[6] J.-F. Cai, E. J. Candès, and Z. Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4):1956–1982, 2010. MR2600248

[7] E. J. Candes and Y. Plan. Matrix completion with noise. *Proceedings of the IEEE*, 98(6):925–936, 2010.

[8] E. J. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717, 2009. MR2565240

[9] E. J. Candès and T. Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080, 2010. MR2723472

[10] S. Chatterjee. Matrix estimation by universal singular value thresholding. *The Annals of Statistics*, 43(1):177–214, 2015. MR3285604

[11] C. Chen, B. He, and X. Yuan. Matrix completion via an alternating direction method. *IMA Journal of Numerical Analysis*, 32(1):227–245, 2012. MR2875250

[12] C. C. Craig. On the Tchebychef inequality of Bernstein. *The Annals of Mathematical Statistics*, 4(2):94–102, 1933.

[13] J. Fan and I. Gijbels. *Local polynomial modelling and its applications.* Monographs on Statistics and Applied Probability, volume 66. CRC Press, 1996. MR1383587

[14] E. N. Gilbert. A comparison of signalling alphabets. *Bell Labs Technical Journal*, 31(3):504–522, 1952.

[15] D. Gross. Recovering low-rank matrices from few coefficients in any basis. *IEEE Transactions on Information Theory*, 57(3):1548–1566, 2011. MR2815834

[16] V. Koltchinskii. Local Rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, 34(6):2593–2656, 2006. MR2329442

[17] V. Koltchinskii. Oracle inequalities in empirical risk minimization and sparse recovery problems. 2011. MR2829871

[18] V. Koltchinskii. Von Neumann entropy penalization and low-rank matrix estimation. *The Annals of Statistics*, pages 2936–2973, 2011. MR3012397

[19] V. Koltchinskii. Sharp oracle inequalities in low rank estimation. In *Empirical Inference*, pages 217–230. Springer, 2013. MR3236868

[20] V. Koltchinskii, K. Lounici, and A. B. Tsybakov. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics*, 39(5):2302–2329, 2011. MR2906869

[21] Y. Koren. Collaborative filtering with temporal dynamics. *Communications of the ACM*, 53(4):89–97, 2010.

[22] Y. Koren, R. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8), 2009.

[23] O. V. Lepski, E. Mammen, and V. G. Spokoiny. Optimal spatial adaptation to inhomogeneous smoothness: an approach based on kernel estimates with variable bandwidth selectors. *The Annals of Statistics*, 25(3):929–947, 1997. MR1447734

[24] O. V. Lepski and V. G. Spokoiny. Optimal pointwise adaptive methods in nonparametric estimation. *The Annals of Statistics*, 25(6):2512–2546, 1997. MR1604408

[25] O. V. Lepskii. On a problem of adaptive estimation in gaussian white noise. *Theory of Probability & Its Applications*, 35(3):454–466, 1991. MR1091202

[26] E. H. Lieb. Convex trace functions and the Wigner-Yanase-Dyson conjecture. *Advances in Mathematics*, 11(3):267–288, 1973. MR0332080

[27] Z. Lin, M. Chen, and Y. Ma. The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices. *arXiv preprint arXiv:1009.5055*, 2010.

[28] W. H. Miller. The classical s-matrix in molecular collisions. *Advances in Chemical Physics*, 30:77, 1973.

[29] S. Negahban and M. J. Wainwright. Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *Journal of Machine Learning Research*, 13(May):1665–1697, 2012. MR2930649

[30] B. Recht, M. Fazel, and P. A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review*, 52(3):471–501, 2010. MR2680543

[31] A. Rohde and A. B. Tsybakov. Estimation of high-dimensional low-rank matrices. *The Annals of Statistics*, 39(2):887–930, 2011. MR2816342

[32] A. Singer and M. Cucuringu. Uniqueness of low-rank matrix completion by rigidity theory. *SIAM Journal on Matrix Analysis and Applications*, 31(4):1621–1641, 2010. MR2595541

[33] M. Talagrand. New concentration inequalities in product spaces. *Inventiones Mathematicae*, 126(3):505–563, 1996. MR1419006

[34] J. A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012. MR2946459

[35] A. B. Tsybakov. Introduction to nonparametric estimation. revised and extended from the 2004 french original. Translated by Vladimir Zaiats,

2009. MR2724359

[36] R. R. Varshamov. Estimate of the number of signals in error correcting codes. *Dokl. Akad. Nauk SSSR*, 117:739–741, 1957. MR0095090

[37] M. Wegkamp. Model selection in nonparametric regression. *The Annals of Statistics*, 31(1):252–273, 2003. MR1962506

[38] J. Weickert and T. Brox. Diffusion and regularization of vector-and matrix-valued. *Inverse Problems, Image Analysis, and Medical Imaging: AMS Special Session on Interaction of Inverse Problems and Image Analysis, January 10-13, 2001, New Orleans, Louisiana*, 313:251, 2002. MR1940999

[39] L. Zhang, G. Wahba, and M. Yuan. Distance shrinkage and euclidean embedding via regularized kernel estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(4):849–867, 2016. MR3534353