Generalized high-dimensional trace regression via nuclear norm regularization*

Jianqing Fan[b,a], Wenyan Gong[a], Ziwei Zhu[a]

[a]Department of Operations Research and Financial Engineering, Princeton University, USA

[b] School of Economics, Fudan University, Shanghai China

August 11, 2018

Abstract

We study the generalized trace regression with a near low-rank regression coefficient matrix, which extends notion of sparsity for regression coefficient vectors. Specifically, given a matrix covariate $\mathbf{X}$, the probability density function of the response Y is $f(Y|\mathbf{X}) = c(Y)\exp\left(\phi^{-1}\left[-Y\eta^* + b(\eta^*)\right]\right)$, where $\eta^* = \mathrm{tr}(\mathbf{\Theta}^{*T}\mathbf{X})$. This model accommodates various types of responses and embraces many important problem setups such as reduced-rank regression, matrix regression that accommodates a panel of regressors, matrix completion, among others. We estimate $\mathbf{\Theta}^*$ through minimizing empirical negative log-likelihood plus nuclear norm penalty. We first establish a general theory and then for each specific problem, we derive explicitly the statistical rate of the proposed estimator. They all match the minimax rates in the linear trace regression up to logarithmic factors. Numerical studies confirm the rates we established and demonstrate the advantage of generalized trace regression over linear trace regression when the response is dichotomous. We also show the benefit of incorporating nuclear norm regularization in dynamic stock return prediction and in image classification.

1 Introduction

In modern data analytics, the parameters of interest often exhibit high ambient dimensions but low intrinsic dimensions that can be exploited to circumvent the curse

*This paper is supported by NSF grants DMS-1406266, DMS-1662139, and DMS-1712591

of dimensionality. One of the most illustrating examples is the sparse signal recovery through incorporating sparsity regularization into empirical risk minimization (Tibshirani (1996); Chen et al. (2001); Fan and Li (2001)). As shown in the profound works (Candes and Tao (2007); Fan and Lv (2008, 2011); Zou and Li (2008); Zhang et al. (2010), among others), the statistical rate of the appropriately regularized M-estimator has mere logarithmic dependence on the ambient dimension d. This implies that consistent signal recovery is feasible even when d grows exponentially with respect to the sample size n. In econometrics, sparse models and methods have also been intensively studied and are proven to be powerful. For example, Belloni et al. (2012) studied estimation of optimal instruments under sparse high-dimensional models and showed that the instrumental variable (IV) estimator based on Lasso and post-Lasso methods enjoys root-n consistency and asymptotic normality. Hansen and Kozbur (2014) and Caner and Fan (2015) investigated instrument selection using high-dimensional regularization methods. Kock and Callot (2015) established oracle inequalities for high dimensional vector autoregressions and Chan et al. (2015) applied group Lasso in threshold autoregressive models and established near-optimal rates in the estimation of threshold parameters. Belloni et al. (2017) employed high-dimensional techniques for program evaluation and causal inference.

When the parameter of interest arises in the matrix form, elementwise sparsity is not the sole way of constraining model complexity; another structure that is exclusive to matrices comes into play: the rank. Low-rank matrices have much fewer degrees of freedom than its ambient dimensions $d_1 \cdot d_2$. To determine a rank-r matrix $\mathbf{\Theta} \in \mathbb{R}^{d_1 \times d_2}$, we only need r left and right singular vectors and r singular values, which correspond to $r(d_1 + d_2 - 1)$ degrees of freedom, without accounting for the orthogonality. As a novel regularization approach, low-rankness motivates matrix representations of the parameters of interest in various statistical and econometric models. If we rearrange the coefficient in the traditional linear model as a matrix, we obtain the so-called trace regression model:

$$Y = \mathrm{tr}(\mathbf{\Theta}^{*T}\mathbf{X}) + \epsilon, \tag{1.1}$$

where $\mathrm{tr}(\cdot)$ denotes the trace, $\mathbf{X} \in \mathbb{R}^{d_1 \times d_2}$ is a matrix of explanatory variables, $\mathbf{\Theta}^* \in \mathbb{R}^{d_1 \times d_2}$ is the matrix of regression coefficients, $Y \in \mathbb{R}$ is the response and $\epsilon \in \mathbb{R}$ is the noise. In predictive econometric applications, $\mathbf{X}$ can be a large panel of time series data such as stock returns or macroeconomic variables (Stock and Watson, 2002; Ludvigson and Ng, 2009), whereas in statistical machine learning $\mathbf{X}$ can be images. The rank of a matrix is controlled by the ℓ_q-norm for $q \in [0, 1)$ of its singular values:

$$\mathcal{B}_q(\mathbf{\Theta}^*) := \sum_{j=1}^{d_1 \wedge d_2} \sigma_j(\mathbf{\Theta}^*)^q \leq \rho, \tag{1.2}$$

where $\sigma_j(\mathbf{\Theta}^*)$ is the jth largest singular value of $\mathbf{\Theta}^*$, and ρ is a positive constant that can grow to infinity. Note that when $q = 0$, it controls the rank of $\mathbf{\Theta}^*$ at ρ. Trace regression is a natural model for matrix-type covariates, such as the panel data, images, genomics microarrays, etc. In addition, particular forms of $\mathbf{X}$ can reduce trace regression to several well-known problem setups. For example, when $\mathbf{X}$ contains only a column and the response Y is multivariate, (1.1) becomes reduced-rank regression model (Anderson (1951), Izenman (1975b)). When $\mathbf{X} \in \mathbb{R}^{d_1 \times d_2}$ is a singleton in the sense that all entries of $\mathbf{X}$ are zeros except for one entry that equals one, (1.1) characterizes the matrix completion problem in item response problems and online recommendation systems. We will specify these problems later.

To explore the low rank structure of $\mathbf{\Theta}^*$ in (1.1), a natural approach is the penalized least-squares with the nuclear norm penalty. Specifically, consider the following optimization problem.

$$\widehat{\mathbf{\Theta}} = \operatorname{argmin} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(\langle \mathbf{\Theta}, \mathbf{X}_i \rangle - Y_i \right)^2 + \lambda \left\| \mathbf{\Theta} \right\|_N \right\}, \tag{1.3}$$

where $\langle A, B \rangle := \operatorname{tr}(A^T B)$ is the inner product of two matrices A and B which have the same dimension and $\|\mathbf{\Theta}\|_N = \sum_{j=1}^{d_1 \wedge d_2} \sigma_j(\mathbf{\Theta})$ is the nuclear norm of $\mathbf{\Theta}$. As ℓ_1-norm regularization yields sparse estimators, nuclear norm regularization enforces the solution to have sparse singular values, in other words, to be low-rank. Recent literatures have rigorously studied the statistical properties of $\widehat{\mathbf{\Theta}}$. Negahban and Wainwright (2011) and Koltchinskii et al. (2011) derived the statistical error rate of $\widehat{\mathbf{\Theta}}$ when ϵ is sub-Gaussian. Fan et al. (2016) introduced a shrinkage principle to handle heavy-tailed noise and achieved the same statistical error rate as Negahban and Wainwright (2011) when the noise has merely bounded second moments.

However, (1.1) does not accomodate categorical responses, which is ubiquitous in pragmatic settings. For example, in P2P microfinance, platforms like Kiva seek potential pairs of lenders and borrowers to create loans. The analysis is based on a large binary matrix with the rows correspondent to the lenders and columns correspondent to the borrowers. Entry (i, j) of the matrix is either checked, meaning that lender i endorses an loan to borrower j, or missing, meaning that lender i is not interested in borrower j or has not seen the request of borrower j. The specific amount of the loan is inaccessible due to privacy concern, thus leading to the binary nature of the response (Lee et al. (2014)). Another example is the famous Netflix Challenge. There, people are given a large rating matrix with the rows representing the customers and the columns representing the movies. Most of its entries are missing and the aim is to infer these missing ratings based on the observed ones. Since the Netflix adopts a five-star movie rating system, the response is categorical with only five levels. This kind of matrix completion problems for item response arise also frequently in other economic surveys, similar to the aforementioned P2P microfinance. These problem setups with categorical responses motivate us to consider the generalized trace regression model.

Suppose that the response Y follows a distribution from the following exponential family:

$$f_n(\mathbf{Y}; X, \beta^*) = \prod_{i=1}^{n} f(Y_i; \eta_i^*) = \prod_{i=1}^{n} \left\{ c(Y_i) \exp\left(\frac{Y_i \eta_i^* - b(\eta_i^*)}{\phi} \right) \right\}, \tag{1.4}$$

where $\eta_i^* = \mathrm{tr}(\mathbf{\Theta}^{*T}\mathbf{X}_i) = \langle \mathbf{\Theta}^*, \mathbf{X}_i \rangle$ is the linear predictor, ϕ is a constant and $c(\cdot)$ and $b(\cdot)$ are known functions. The negative log-likelihood corresponding to (1.4) is given, up to an affine transformation, by

$$\mathcal{L}_n(\mathbf{\Theta}) = \frac{1}{n} \sum_{i=1}^{n} \left[-Y_i \langle \mathbf{\Theta}, \mathbf{X}_i \rangle + b(\langle \mathbf{\Theta}, \mathbf{X}_i \rangle) \right] \tag{1.5}$$

and the gradient and Hessian of $\mathcal{L}_n(\mathbf{\Theta})$ are respectively

$$\begin{aligned} \nabla \mathcal{L}_n(\mathbf{\Theta}) &= \frac{1}{n} \sum_{i=1}^{n} \left[b'(\eta_i) - Y_i \right] \mathbf{X}_i = \frac{1}{n} \sum_{i=1}^{n} \left[b'(\langle \mathbf{\Theta}, \mathbf{X}_i \rangle) - Y_i \right] \mathbf{X}_i \\ \mathbf{H}_n(\mathbf{\Theta}) := \nabla^2 \mathcal{L}_n(\mathbf{\Theta}) &= \frac{1}{n} \sum_{i=1}^{n} b''(\langle \mathbf{\Theta}, \mathbf{X}_i \rangle) \mathrm{vec}(\mathbf{X}_i) \mathrm{vec}(\mathbf{X}_i)^T. \end{aligned} \tag{1.6}$$

For future convenience, we denote $\mathrm{E}[\mathbf{H}_n(\mathbf{\Theta})]$ by $\mathbf{H}(\mathbf{\Theta})$. To estimate $\mathbf{\Theta}^*$, we recruit the following M-estimator that minimizes the negative log-likelihood plus nuclear norm penalty.

$$\widehat{\mathbf{\Theta}} = \mathrm{argmin}_{\mathbf{\Theta} \in \mathbb{R}^{d_1 \times d_2}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left[b(\langle \mathbf{\Theta}, \mathbf{X}_i \rangle) - Y_i \langle \mathbf{\Theta}, \mathbf{X}_i \rangle \right] + \lambda \left\| \mathbf{\Theta} \right\|_N \right\}. \tag{1.7}$$

This is a high-dimensional convex optimization problem. We will discuss the algorithms for computing (1.7) in the simulation section.

Related to our work is the matrix completion problem with binary entry, i.e., 1-bit matrix completion, which is a specific example of our generalized trace regression and has direct application in predicting aforementioned P2P microfinance. Therein entry (i, j) of the matrix is modeled as a response from a logistic regression or probit regression with parameter $\mathbf{\Theta}^*_{ij}$ and information of each responded items is related through the low-rank assumption of $\mathbf{\Theta}^*$. Previous works studied the estimation of $\mathbf{\Theta}^*$ by minimizing the negative log-likelihood function under the constraint of max-norm (Cai and Zhou (2013)), nuclear norm (Davenport et al. (2014)) and rank (Bhaskar and Javanmard (2015)). There are also some works in 1-bit compressed sensing to recover sparse signal vectors (Gupta et al., 2010; Plan and Vershynin, 2013a,b). Nevertheless, we did not find any work in the generality that we are dealing with, which fits matrix-type explanatory variables and various types of dependent variables.

In this paper, we establish a unified framework for statistical analysis of $\widehat{\mathbf{\Theta}}$ in (1.7) under the generalized trace regression model. As showcases of the applications of our general theory, we focus on three problem setups: generalized matrix regression, reduced-rank regression and one-bit matrix completion. We explicitly derive statistical rate of $\widehat{\mathbf{\Theta}}$ under these three problem setups respectively. It is worth noting that for one-bit matrix completion, our statistical rate is sharper than that in Davenport et al. (2014). We also conduct numerical experiments on both simulated and real data to verify the established rate and illustrate the advantage of using the generalized trace regression over the vanilla trace regression when categorical responses occur.

The paper is organized as follows. In Section 2, we specify the problem setups and present the statistical rates of $\widehat{\mathbf{\Theta}}$ under generalized matrix regression, reduced-rank regression and one-bit matrix completion respectively. In Section 3, we present simulation results to back up our theoretical results from Section 2 and to demonstrate superiority of generalized trace regression over the standard one. In Section 4, we use real data to display the improvement brought by nuclear norm regularization in return prediction and image classification.

2 Main results

2.1 Notation

We use regular letters for random variables, bold lower case letters for random vectors and bold upper case letter for matrices. For a function $f(\cdot)$, we use $f'(\cdot)$, $f''(\cdot)$ and $f'''(\cdot)$ to denote its first, second and third order derivative. For sequences $\{a_i\}_{i=1}^{\infty}$ and $\{b_i\}_{i=1}^{\infty}$, we say $a_i = O(b_i)$ if there exists a constant $c > 0$ such that $a_i/b_i < c$ for $1 \leq i < \infty$, and we say $a_i = \Omega(b_i)$ if there exists a constant $c > 0$ such that $a_i/b_i \geq c$ for $1 \leq i < \infty$. For a random variable x, we denote its sub-Gaussian norm as $\|x\|_{\Psi_2} := \sup_{p\geq 1} (\mathbb{E}\,|x|^p)^{1/p}/\sqrt{p}$ and its sub-exponential norm as $\|x\|_{\Psi_1} = \sup_{p\geq 1} (\mathbb{E}\,|x|^p)^{1/p}/p$. For a random vector $\mathbf{x} \in \mathbb{R}^d$, we denote its sub-Gaussian norm as $\|\mathbf{x}\|_{\Psi_2} = \sup_{\mathbf{v}\in\mathcal{S}^{d-1}} \left\|\mathbf{v}^T\mathbf{x}\right\|_{\Psi_2}$ and its sub-exponential norm as $\|\mathbf{x}\|_{\Psi_1} = \sup_{\mathbf{v}\in\mathcal{S}^{d-1}} \left\|\mathbf{v}^T\mathbf{x}\right\|_{\Psi_1}$. Here, $\mathcal{S}^{d-1}$ denotes the unit sphere in $\mathbb{R}^d$. We use $\mathbf{e}_j$ to denote a vector whose elements are all 0 except that the jth one is 1. For a matrix $\mathbf{X} \in \mathbb{R}^{d_1\times d_2}$, we use $\text{vec}(\mathbf{X}) \in \mathbb{R}^{d_1 d_2}$ to represent the vector that consists of all the elements from $\mathbf{X}$ column by column. We use $r(\mathbf{X})$, $\|\mathbf{X}\|_\infty$, $\|\mathbf{X}\|_{\text{op}}$, $\|\mathbf{X}\|_N$ to denote the rank, elementwise max norm, operator norm and nuclear norm of $\mathbf{X}$ respectively. We call $\{\mathbf{X} : \|\mathbf{X}-\mathbf{Y}\|_\infty \leq r\}$ a L_∞-ball centered at $\mathbf{Y}$ with radius r for $r > 0$. Define $d_1 \wedge d_2 := \min(d_1, d_2)$ and $d_1 \vee d_2 := \max(d_1, d_2)$. For matrices $\mathbf{A}$ and $\mathbf{B}$, let $\langle\mathbf{A},\mathbf{B}\rangle = \text{tr}(\mathbf{A}^T\mathbf{B})$. For any subspace $\mathcal{M} \subset \mathbb{R}^{d\times d}$, define its orthogonal space $\mathcal{M}^\perp := \{\mathbf{A} : \forall\mathbf{M} \in \mathcal{M}, \langle\mathbf{A},\mathbf{M}\rangle = 0\}$.

2.2 General theory

In this section, we provide a general theorem on the statistical rate of $\widehat{\boldsymbol{\Theta}}$ in (1.7). As we shall see, the statistical consistency of $\widehat{\boldsymbol{\Theta}}$ essentially requires two conditions: i) sufficient penalization λ; ii) localized restricted strong convexity of $\mathcal{L}_n(\boldsymbol{\Theta})$ around $\boldsymbol{\Theta}^*$. In high-dimensional statistics, it is well known that the restricted strong convexity (RSC) of the loss function underpins the statistical rate of the M-estimator (Negahban et al., 2011; Raskutti et al., 2010). In generalized trace regression, however, the fact that the Hessian matrix $\mathbf{H}_n(\boldsymbol{\Theta})$ depends on $\boldsymbol{\Theta}$ creates technical difficulty for verifying RSC for the loss function. To address this issue, we apply the localized analysis due to Fan et al. (2015), where they only require local RSC (LRSC) of $\mathcal{L}_n(\boldsymbol{\Theta})$ around $\boldsymbol{\Theta}^*$ to derive statistical rates of $\widehat{\boldsymbol{\Theta}}$. Below we formulate the concept of LRSC. For simplicity, from now on we assume that $\boldsymbol{\Theta}^*$ is a d-by-d square matrix. We can easily extend our analysis to the case of rectangular $\boldsymbol{\Theta}^* \in \mathbb{R}^{d_1\times d_2}$; the only change in the result is a replacement of d with $\max(d_1, d_2)$ in the statistical rate.

Definition 1. *Given a constraint set $\mathcal{C} \subset \mathbb{R}^{d\times d}$, a local neighborhood $\mathcal{N}$ of $\boldsymbol{\Theta}^*$, a positive constants κ_ℓ and a tolerance term τ_ℓ, we say that the loss function $\mathcal{L}(\cdot)$ satisfies LRSC$(\mathcal{C}, \mathcal{N}, \kappa_\ell, \tau_\ell)$ if for all $\boldsymbol{\Delta} \in \mathcal{C}$ and $\boldsymbol{\Theta} \in \mathcal{N}$,*

$$\mathcal{L}(\boldsymbol{\Theta}+\boldsymbol{\Delta}) - \mathcal{L}(\boldsymbol{\Theta}) - \langle\nabla\mathcal{L}(\boldsymbol{\Theta}), \boldsymbol{\Delta}\rangle \geq \kappa_\ell \|\boldsymbol{\Delta}\|_F^2 - \tau_\ell. \tag{2.1}$$

Note that τ_ℓ is a tolerance term that will be specified in the main theorem. Now we introduce the constraint set $\mathcal{C}$ in our context. Let $\boldsymbol{\Theta}^* = \mathbf{U}\mathbf{D}\mathbf{V}^T$ be the SVD of $\boldsymbol{\Theta}^*$, where the diagonal of $\mathbf{D}$ is in the decreasing order. Denote the first r columns of $\mathbf{U}$

and $\mathbf{V}$ by $\mathbf{U}^r$ and $\mathbf{V}^r$ respectively, and define

$$\begin{aligned} &\mathcal{M} := \{\boldsymbol{\Theta} \in \mathbb{R}^{d\times d} \mid \mathrm{row}(\boldsymbol{\Theta}) \subseteq \mathrm{col}(\mathbf{V}^r), \mathrm{col}(\boldsymbol{\Theta}) \subseteq \mathrm{col}(\mathbf{U}^r)\}, \\ &\overline{\mathcal{M}}^{\perp} := \{\boldsymbol{\Theta} \in \mathbb{R}^{d\times d} \mid \mathrm{row}(\boldsymbol{\Theta}) \perp \mathrm{col}(\mathbf{V}^r), \mathrm{col}(\boldsymbol{\Theta}) \perp \mathrm{col}(\mathbf{U}^r)\}, \end{aligned} \tag{2.2}$$

where $\mathrm{col}(\cdot)$ and $\mathrm{row}(\cdot)$ denote the column space and row space respectively. For any $\boldsymbol{\Delta} \in \mathbb{R}^{d\times d}$ and Hilbert space $\mathcal{W} \subseteq \mathbb{R}^{d\times d}$, let $\boldsymbol{\Delta}_{\mathcal{W}}$ be the projection of $\boldsymbol{\Delta}$ onto $\mathcal{W}$. We first clarify here what $\boldsymbol{\Delta}_{\mathcal{M}}$, $\boldsymbol{\Delta}_{\overline{\mathcal{M}}}$ and $\boldsymbol{\Delta}_{\overline{\mathcal{M}}^{\perp}}$ are. Write $\boldsymbol{\Delta}$ as

$$\boldsymbol{\Delta} = [\mathbf{U}^r, \mathbf{U}^{r\perp}] \begin{bmatrix} \boldsymbol{\Gamma}_{11} & \boldsymbol{\Gamma}_{12} \\ \boldsymbol{\Gamma}_{21} & \boldsymbol{\Gamma}_{22} \end{bmatrix} [\mathbf{V}^r, \mathbf{V}^{r\perp}]^T,$$

then the following equalities hold:

$$\begin{aligned} &\boldsymbol{\Delta}_{\mathcal{M}} = \mathbf{U}^r \boldsymbol{\Gamma}_{11} (\mathbf{V}^r)^T, \quad \boldsymbol{\Delta}_{\overline{\mathcal{M}}^{\perp}} = \mathbf{U}^{r\perp} \boldsymbol{\Gamma}_{22} (\mathbf{V}^{r\perp})^T, \\ &\boldsymbol{\Delta}_{\overline{\mathcal{M}}} = [\mathbf{U}^r, \mathbf{U}^{r\perp}] \begin{bmatrix} \boldsymbol{\Gamma}_{11} & \boldsymbol{\Gamma}_{12} \\ \boldsymbol{\Gamma}_{21} & \mathbf{0} \end{bmatrix} [\mathbf{V}^r, \mathbf{V}^{r\perp}]^T. \end{aligned} \tag{2.3}$$

According to Negahban et al. (2012), when $\lambda \geq 2\|n^{-1} \sum_{i=1}^{n} [b'(\langle \mathbf{X}_i, \boldsymbol{\Theta}^* \rangle) - Y_i] \cdot \mathbf{X}_i\|_{op}$, regardless of what r is, $\widehat{\boldsymbol{\Delta}}$ falls in the following cone:

$$\mathcal{C}(\mathcal{M}, \overline{\mathcal{M}}^{\perp}, \boldsymbol{\Theta}^*) := \Big\{ \boldsymbol{\Delta} \in \mathbb{R}^{d\times d} : \|\boldsymbol{\Delta}_{\overline{\mathcal{M}}^{\perp}}\|_N \leq 3 \left\|\boldsymbol{\Delta}_{\overline{\mathcal{M}}}\right\|_N + 4 \sum_{j \geq r+1} \sigma_j(\boldsymbol{\Theta}^*) \Big\}.$$

Now we present the main theorem that serves as a roadmap to establish the statistical rate of convergence for $\widehat{\boldsymbol{\Theta}}$.

Theorem 1. *Consider the model* (1.4)*. Suppose $\mathcal{B}_q(\boldsymbol{\Theta}^*) \leq \rho$ and*

$$\lambda \geq 2 \|\frac{1}{n} \sum_{i=1}^{n} \big[b'(\langle \mathbf{X}_i, \boldsymbol{\Theta}^* \rangle) - Y_i \big] \cdot \mathbf{X}_i \|_{op}. \tag{2.4}$$

Define $\mathcal{N} := \{\boldsymbol{\Theta} \in \mathbb{R}^{d\times d} : \|\boldsymbol{\Theta} - \boldsymbol{\Theta}^*\|_F^2 \leq C_1 \rho \lambda^{2-q}, \boldsymbol{\Theta} - \boldsymbol{\Theta}^* \in \mathcal{C}(\mathcal{M}, \overline{\mathcal{M}}^{\perp}, \boldsymbol{\Theta}^*)\}$, where C_1 is a constant and $\mathcal{M}$ and $\overline{\mathcal{M}}$ are constructed as per (2.2)*. Suppose $\mathcal{L}_n(\boldsymbol{\Theta})$ satisfies LRSC($\mathcal{C}(\mathcal{M}, \overline{\mathcal{M}}^{\perp}, \boldsymbol{\Theta}^*), \mathcal{N}, \kappa_\ell, \tau_\ell$), where $\tau_\ell = C_0 \rho \lambda^{2-q}$ for some constant C_0 and κ_ℓ is a positive constant. Then it holds that*

$$\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\|_F^2 \leq C_1 \rho \left(\frac{\lambda}{\kappa_\ell} \right)^{2-q} \quad \text{and} \quad \|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\|_N \leq C_2 \rho \left(\frac{\lambda}{\kappa_\ell} \right)^{1-q}, \tag{2.5}$$

where C_1, C_2 are constants.

Theorem 1 points out two conditions that lead to the statistical rate of $\widehat{\mathbf{\Theta}}$. First, we need λ to be sufficiently large, which has an adverse impact on the rates. Therefore, the optimal choice of λ is the lower bound given in (2.4). The second requirement is LRSC of $\mathcal{L}_n(\mathbf{\Theta})$ around $\mathbf{\Theta}^*$. In the sequel, for each problem setup we will first derive the rate of the lower bound of λ as shown in (2.4) and then verify the LRSC of $\mathcal{L}_n(\mathbf{\Theta})$ so that we can establish the statistical rate. **Note that the LRSC property does not imply any constraint on the choice of the initial values for solving the optimization problem. It is a pure statistical assumption and used to show that the minimizer of the penalized likelihood possesses the established statistical property.**

For notational convenience, later on when we refer to certain quantities as constants, we mean they are independent of n, d, ρ. In the next subsections, we will apply the general theorem to analyze various specific problem setups and derive the explicit rates of convergence.

2.3 Generalized matrix regression

Generalized matrix regression can be regarded as a generalized linear model (GLM) with matrix covariates. Here we assume that $\text{vec}(\mathbf{X}_i)$, the vectorized version of $\mathbf{X}_i$, is a sub-Gaussian random vector with bounded Ψ_2-norm. Consider $\widehat{\mathbf{\Theta}}$ as defined in (1.7). To derive statistical rate of $\widehat{\mathbf{\Theta}}$, we first establish the rate of the lower bound of λ as characterized in (2.4).

Lemma 1. *Consider the following conditions:*

(C1) $\{vec(\mathbf{X}_i)\}_{i=1}^n$ *are i.i.d. sub-Gaussian vectors with* $\|vec(\mathbf{X}_i)\|_{\Psi_2} \leq \kappa_0 < \infty$;
(C2) $|b''(x)| \leq M < \infty$ *for any* $x \in \mathbb{R}$;

Then for any $\nu > 0$, *there exists a constant* $\gamma > 0$ *such that as long as* $d/n < \gamma$, *it holds that*

$$\mathbb{P}\left(\|\frac{1}{n}\sum_{i=1}^{n}(b'(\langle \mathbf{\Theta}^*, \mathbf{X}_i\rangle) - Y_i) \cdot \mathbf{X}_i\|_{op} > \nu\sqrt{\frac{d}{n}} \right) \leq C\exp(-cd), \tag{2.6}$$

where C *and* c *are constants.*

Next we verify the LRSC of $\mathcal{L}_n(\mathbf{\Theta})$.

Lemma 2. *Besides (C1) and (C2) in Lemma 1, assume that*

(C3) $\lambda_{min}(\mathbf{H}(\mathbf{\Theta}^*)) \geq \kappa_\ell > 0$;
(C4) $\|\mathbf{\Theta}^*\|_F \geq \alpha\sqrt{d}$ *for some constant* α;

(C5) $|b'''(x)| \leq |x|^{-1}$ *for* $|x| > 1$.

Suppose $\lambda \geq \nu\sqrt{d/n}$, *where* ν *is the same as in Lemma 1. Let* $\mathcal{N} = \{\mathbf{\Theta} \in \mathbb{R}^{d\times d} : \|\mathbf{\Theta} - \mathbf{\Theta}^*\|_F^2 \leq C_1\rho\lambda^{2-q}, \mathbf{\Theta} - \mathbf{\Theta}^* \in \mathcal{C}(\mathcal{M}, \overline{\mathcal{M}}^{\perp}, \mathbf{\Theta}^*)\}$. *As long as* $\rho\lambda^{1-q}$ *is sufficiently small,* $\mathcal{L}_n(\mathbf{\Theta})$ *satisfies LRSC*$(\mathcal{C}(\mathcal{M}, \overline{\mathcal{M}}^{\perp}, \mathbf{\Theta}^*), \mathcal{N}, \kappa/2, \tau_\ell)$ *with probability at least* $1 - C_1 \exp(-c_1 d)$, *where* $\tau_\ell = C_0\rho\lambda^{2-q}$ *and* c_1, C_0 *and* C_1 *are constants.*

Remark 1. *Since* $\langle \mathbf{\Theta}, \mathbf{X}\rangle$ *represents the signal in our model, the lower bound on* $\|\mathbf{\Theta}^*\|_F$ *in Condition (C4) guarantees sufficient strength of the signal. If* $\|\mathbf{\Theta}^*\|_F$ *is too small, the signal might be dominated by the noise. Condition (C4) is mild; even if* $\mathbf{\Theta}^*$ *is sparse and only has* $O(d)$ *non-zero entries, as long as they are of constant order, (C4) is satisfied. When* $\mathbf{\Theta}^*$ *is extremely sparse and only has* $O(1)$ *non-zero entries, Condition (C4) requires their magnitude to be comparable to* d *since otherwise the signal is too weak. In fact, if* $\mathbf{\Theta}^*$ *is extremely sparse,* L_1 *regularization shall be better than the nuclear norm regularization for accurate matrix recovery.*

Condition (C5) requires that the third order derivative of $b(\cdot)$ *decays sufficiently fast. In fact, except for Poisson regression, most members in the family of generalized linear models satisfy this condition, e.g., linear model, logistic regression, log-linear model, etc.*

Based on the above two lemmas, we apply Theorem 1 and establish the explicit statistical rate of $\widehat{\mathbf{\Theta}}$ as follows.

Theorem 2. *Under the conditions in Lemmas 1 and 2, choosing* $\lambda = 2\nu\sqrt{d/n}$, *where* ν *is the same as in Lemma 1, there exist constants* $\{c_i\}_{i=1}^2$ *and* $\{C_i\}_{i=1}^5$ *such that once* $\rho(d/n)^{(1-q)/2} \leq C_1$, *we have*

$$\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*\|_F^2 \leq C_2\rho\left(\frac{d}{n}\right)^{1-q/2}, \quad \|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*\|_N \leq C_3\rho\left(\frac{d}{n}\right)^{(1-q)/2} \tag{2.7}$$

with probability at least $1 - C_4\exp(-c_1 d) - C_5 \exp(-c_2 d)$.

When $q = 0$, ρ becomes the rank of $\mathbf{\Theta}^*$ and there are $O(\rho d)$ free parameters. Each of these parameters can be estimated at rate $O_P(1/\sqrt{n})$. Therefore, the sum of squared errors should at least be $O(\rho d/n)$. This is indeed the bound of $\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*\|_F^2$ given by (2.7), which depends on the effective dimension ρd rather than the ambient dimension d^2. The second result of (2.7) confirms this in the spectral "L_1-norm", the nuclear norm.

2.4 Generalized reduced-rank regression

Consider the conventional reduced-rank regression model (RRR)

$$\mathbf{y}_i = \mathbf{\Theta}^* \mathbf{x}_i + \boldsymbol{\varepsilon}_i,$$

where $\mathbf{x}_i \in \mathbb{R}^d$ is the covariate, $\mathbf{y}_i \in \mathbb{R}^d$ is the response, $\mathbf{\Theta}^* \in \mathbb{R}^{d\times d}$ is a near low-rank coefficient matrix and $\boldsymbol{\varepsilon}_i \in \mathbb{R}^d$ is the noise. Again, we set the number of covariates to be the same as the number of responses purely for simplicity of the presentation. Note that in each sample there are d responses correspondent to the same covariate vector. RRR aims to reduce the number of regression parameters in multivariate analysis. It was first studied in detail by Anderson (1951), where the author considered multi-response regression with linear constraints on the coefficient matrix and applied this model to obtain points estimation and confidence regions in "shock models" in econometrics (Marshak (1950)). Since then, there has been great amount of literature on RRR in econometrics (Ahn and Reinsel (1994), Geweke (1996), Kleibergen and Paap (2006)) and statistics (Izenman (1975a), Velu and Reinsel (2013), Chen et al. (2013)).

Now we generalize the above reduced-rank regression to accommodate various types of dependent variables. For any $1 \le i \le n$ and $1 \le j \le d$, y_{ij} is generated from the following density function.

$$f(y_{ij}; \mathbf{x}_i, \mathbf{\Theta}^*) = c(y_{ij}) \exp\Big(\frac{y_{ij}\eta^*_{ij} - b(\eta^*_{ij})}{\phi}\Big) = c(y_{ij}) \exp\Big(\frac{y_{ij}\boldsymbol{\theta}_j^{*T}\mathbf{x}_i - b(\boldsymbol{\theta}_j^{*T}\mathbf{x}_i)}{\phi}\Big), \quad (2.8)$$

where $\boldsymbol{\theta}^*_j$ is the jth row of $\mathbf{\Theta}^*$, $\eta^*_{ij} = \boldsymbol{\theta}_j^{*T}\mathbf{x}_i$, $c(\cdot)$ and $b(\cdot)$ are known functions. We further assume that for any $(i_1, j_1) \neq (i_2, j_2)$, $y_{i_1 j_1} \perp\!\!\!\perp y_{i_2 j_2}$. Note that we can recast this model as a generalized trace regression with $N = nd$ samples: $\{\mathbf{X}_{(i-1)d+j} = \mathbf{e}_j \mathbf{x}_i^T \in \mathbb{R}^{d\times d}, Y_{(i-1)d+j} = y_{ij} \in \mathbb{R} : 1 \le i \le n, 1 \le j \le d\}$. We emphasize here that throughout this paper we will use $(\mathbf{x}_i, \mathbf{y}_i)$ and $\{(\mathbf{X}_t, Y_t)\}_{t=(i-1)d+1}^{id}$ to denote the vector and matrix forms of the ith sample in RRR.

According to model (2.8), we solve for the nuclear norm regularized M-estimator $\widehat{\mathbf{\Theta}}$ as follows.

$$\begin{aligned}\widehat{\mathbf{\Theta}} &= \operatorname{argmin}_{\mathbf{\Theta}\in\mathbb{R}^{d\times d}} \frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{d} \big[b(\langle \mathbf{\Theta}, \mathbf{X}_{(i-1)d+j}\rangle) - Y_{(i-1)d+j}\cdot\langle \mathbf{\Theta}, \mathbf{X}_{(i-1)d+j}\rangle\big] + \lambda \left\|\mathbf{\Theta}\right\|_N \\ &= \operatorname{argmin}_{\mathbf{\Theta}\in\mathbb{R}^{d\times d}} \frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{d} \big[b(\boldsymbol{\theta}_j^T\mathbf{x}_i) - y_{ij}\cdot\boldsymbol{\theta}_j^T\mathbf{x}_i\big] + \lambda \left\|\mathbf{\Theta}\right\|_N .\end{aligned} \quad (2.9)$$

Under the sub-Gaussian design, we are able to derive the covergence rate of $\widehat{\mathbf{\Theta}}$ in RRR with the same tool as what we used in matrix regression. Notice that there is a change of notation introduced in this section. $\{\mathbf{x}_i\}_{i=1}^n$ and $\{\mathbf{y}_i\}_{i=1}^n$ are the original forms of reduced-rank regression while $\{\mathbf{X}_i\}_{i=1}^N$ and $\{Y_i\}_{i=1}^N$ are the rephrased forms that match our framework ($N = nd$). Again, we explicitly derive the rate of the lower bound of λ in the following lemma.

Lemma 3. *Suppose the following conditions hold:*

(C1) $\{\mathbf{x}_i\}_{i=1}^n$ *are i.i.d sub-Gaussian vectors with* $\|\mathbf{x}_i\|_{\Psi_2} \le \kappa_0 < \infty$*;*
(C2) $b''(\cdot) \le M < \infty$, $b'''(\cdot) \le L < \infty$.

Then for any $\nu > 0$*, there exists a constant* $\gamma > 0$ *such that as long as* $d/n < \gamma$*, it holds that*

$$P\big(\|\frac{1}{N}\sum_{i=1}^{N}(b'(\langle \mathbf{X}_i, \mathbf{\Theta}^*\rangle) - Y_i)\mathbf{X}_i\|_{op} \ge d^{-1}\sqrt{\frac{\phi M \kappa_0 d}{n}}\big) \le 2\exp(-cd), \tag{2.10}$$

where ϕ *is the same as in* (2.8) *and c is a universal constant.*

The following lemma establishes the LRSC of the loss function.

Lemma 4. *Besides conditions in Lemma 3, assume that*

(C3) $\lambda_{min}\left(\mathbf{H}(\mathbf{\Theta}^*)\right) \ge \kappa_\ell > 0$.

Choose $\lambda = d^{-1}\sqrt{\phi M \kappa_0 d/n}$ *as in* (2.10)*. Let* $\mathcal{N} := \{\mathbf{\Theta} : \|\mathbf{\Theta} - \mathbf{\Theta}^*\|_F^2 \le \rho\lambda^{2-q}\}$*. For any* $\delta > 4$*, when* $\rho(d/n)^{1-q/2}\log(nd)$ *is sufficiently small,* $\mathcal{L}_n(\mathbf{\Theta})$ *satisfies LRSC*$(\mathbb{R}^{d\times d}, \mathcal{N}, \kappa_\ell/(2d), 0)$ *with probability at least* $1 - 2(nd)^{2-\frac{\delta}{2}}$.

Combining the above lemmas with Theorem 1, we can derive the statistical rate of $\widehat{\mathbf{\Theta}}$ as defined in (2.9).

Theorem 3. *Suppose conditions in Lemmas 3 and 4 hold. Take* $\lambda = d^{-1}\sqrt{\phi M \kappa_0 d/n}$*. For any* $\delta > 4$*, there exist constants* $\{c_i\}_{i=1}^2$ *and* $\{C_i\}_{i=1}^2$ *such that once* $\rho(d/n)^{1-q/2}$ $\log(nd) < c_1$*, any solution to (2.9) satisfies*

$$\left\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*\right\|_F^2 \le C_1\rho\left(\frac{d}{n}\right)^{1-q/2}, \quad \left\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*\right\|_N \le C_2\rho\left(\frac{d}{n}\right)^{(1-q)/2} \tag{2.11}$$

with probability at least $1 - 2\exp(-c_2 d) - 2(nd)^{2-\frac{\delta}{2}}$.

Again, as remarked at the end of Section 2.3, the error depends on the effective dimension ρd rather than the ambient dimension d^2 for the case $q = 0$.

2.5 One-bit matrix completion

Another important example of the generalized trace regression is the one-bit matrix completion problem, which appears frequently in the online item response questionnaire and recommendation system. The showcase example is the aforementioned Kiva platform in P2P microfinance, in which we only observe sparse binary entries of lenders and borrowers. Suppose that we have d_1 users that answer a small fraction of d_2 binary questions. For simplicity of presentation, we again assume that $d_1 = d_2 = d$. Specifically, consider the following logistic regression model with $\mathbf{X}_i = \mathbf{e}_{a(i)}\mathbf{e}_{b(i)}^T \in \mathbb{R}^{d\times d}$. Namely, the ith data records the $a(i)$th user answers the binary question $b(i)$. The problem is also very similar to the aforementioned Netflix problem, except that only dichotomous responses are recorded here.

The logistic regression model assumes that

$$\log \frac{\mathbb{P}\left(Y_i = 1 \mid \mathbf{X}_i\right)}{\mathbb{P}\left(Y_i = 0 \mid \mathbf{X}_i\right)} = \operatorname{tr}(\boldsymbol{\Theta}^{*T}\mathbf{X}_i) = \Theta^*_{a(i),b(i)}. \tag{2.12}$$

Note that this model can be derived from generalized trace regression (1.4) with $b'(\eta_i^*) = (1+\exp(-\eta_i^*))^{-1}$. (2.12) says that given $\mathbf{X}_i = \mathbf{e}_{a(i)}\mathbf{e}_{b(i)}^T \in \mathbb{R}^{d\times d}$, Y_i is a Bernoulli random variable with $\mathbb{P}(Y_i = 1 \mid \mathbf{X}_i) = (1+\exp(-\Theta^*_{a(i),b(i)}))^{-1}$. We assume that $\{(a(i), b(i))\}_{i=1}^N$ are randomly and uniformly distributed over $\{(j,k)\}_{1\le j\le d, 1\le k\le d}$. We further require $\boldsymbol{\Theta}^*$ to be non-spiky in the sense that $\|\boldsymbol{\Theta}^*\|_\infty = O(1)$ and thus $\|\boldsymbol{\Theta}^*\|_F = O(d)$. This condition ensures consistent estimation as elucidated in Negahban and Wainwright (2012). For ease of theoretical reasoning, from now on we will rescale the design matrix $\mathbf{X}_i$ and the signal $\boldsymbol{\Theta}^*$ such that $\mathbf{X}_i = d\mathbf{e}_{a(i)}\mathbf{e}_{b(i)}^T$ and $\|\boldsymbol{\Theta}^*\|_F \le 1$. Based on such setting, we estimate $\boldsymbol{\Theta}^*$ through minimizing negative log-likelihood plus nuclear norm penalty under an element-wise max-norm constraint:

$$\widehat{\boldsymbol{\Theta}} = \operatorname{argmin}_{\|\boldsymbol{\Theta}\|_\infty \le R/d} \left\{ \frac{1}{n}\sum_{i=1}^n \left[\log(1+\exp(\langle \boldsymbol{\Theta}, \mathbf{X}_i\rangle)) - Y_i\langle \boldsymbol{\Theta}, \mathbf{X}_i\rangle\right] + \lambda \|\boldsymbol{\Theta}\|_N \right\}, \tag{2.13}$$

where λ and R are tuning parameters.

Again, we first derive the rate of the lower bound for λ as shown in Theorem 1. For this specific model, simple calculation shows that the lower bound (2.4) reduces to

$$\|n^{-1}\sum_{i=1}^n [\exp(\langle \boldsymbol{\Theta}^*, \mathbf{X}_i\rangle)/(1+\exp(\langle \boldsymbol{\Theta}^*, \mathbf{X}_i\rangle)) - Y_i] \cdot \mathbf{X}_i\|_{\mathrm{op}}.$$

Lemma 5. *Under the following conditions:*

(C1) $\|\boldsymbol{\Theta}^*\|_F \le 1$, $\|\boldsymbol{\Theta}^*\|_\infty \le R/d$ *where* $0 < R < \infty$*;*

(C2) $\{\mathbf{X}_i\}_{i=1}^n$ *are uniformly sampled from* $\{d\mathbf{e}_j\mathbf{e}_k^T\}_{1\le j,k\le d}$*;*

For any $\delta > 1$*, there exists* $\gamma > 0$ *such that as long as* $d\log d/n < \gamma$*, the following inequality holds for some constant* $\nu > 0$*:*

$$\mathbb{P}\left(\|\frac{1}{n}\sum_{i=1}^n\big(\frac{\exp\left(\langle\mathbf{\Theta}^*,\mathbf{X}_i\rangle\right)}{\exp\left(\langle\mathbf{\Theta}^*,\mathbf{X}_i\rangle\right)+1}-Y_i\big)\mathbf{X}_i\|_{op} > \nu\sqrt{\frac{\delta d\log d}{n}}\right) \le 2d^{1-\delta}. \tag{2.14}$$

Next we study the LRSC of the loss function. Following Negahban and Wainwright (2012), besides $\mathcal{C}(\mathcal{M},\overline{\mathcal{M}}^\perp,\mathbf{\Theta}^*)$, we define another constraint set

$$\mathcal{C}'(c_0) := \left\{\mathbf{\Delta}\in\mathbb{R}^{d\times d}, \mathbf{\Delta}\neq\mathbf{0} : \frac{\|\mathbf{\Delta}\|_\infty}{\|\mathbf{\Delta}\|_F}\cdot\frac{\|\mathbf{\Delta}\|_N}{\|\mathbf{\Delta}\|_F} \le \frac{1}{c_0 d}\sqrt{\frac{n}{d\log d}}\right\}. \tag{2.15}$$

Here $\|\mathbf{\Delta}\|_\infty/\|\mathbf{\Delta}\|_F$ and $\|\mathbf{\Delta}\|_N/\|\mathbf{\Delta}\|_F$ are measures of spikiness and low-rankness of $\mathbf{\Delta}$. Let $\mathcal{N} = \{\mathbf{\Theta} : \|\mathbf{\Theta}-\mathbf{\Theta}^*\|_\infty \le 2R/d\}$. Note that $\mathcal{N}$ is not the same as in Theorem 1 any more. As we shall see later, instead of directly applying Theorem 1, we need to adapt the proof of Theorem 1 to the matrix completion setting to derive statistical rate of $\widehat{\mathbf{\Theta}}$. The following lemma establishes $\mathrm{LRSC}(\mathcal{C}'(c_0),\mathcal{N},\kappa_\ell,0)$ of $\mathcal{L}_n(\mathbf{\Theta})$ for some $\kappa_\ell > 0$.

Lemma 6. *There exist constants* C_1, C_2, c_1, c_2 *such that as long as* $n > C_1 d\log d$ *and* $R \le c_1$*, it holds with probability greater than* $1 - C_2\exp\left(-c_2 d\log d\right)$ *that for all* $\mathbf{\Delta}\in\mathcal{C}'(c_0)$ *and* $\mathbf{\Theta}\in\mathcal{N}$*,*

$$vec(\mathbf{\Delta})^T\mathbf{H}_n(\mathbf{\Theta})vec(\mathbf{\Delta}) \ge \frac{\|\mathbf{\Delta}\|_F^2}{512(\exp(R)+\exp(-R)+2)}. \tag{2.16}$$

Now we are ready to establish the statistical rate of $\widehat{\mathbf{\Theta}}$ in (2.13).

Theorem 4. *Let* $\widehat{\mathbf{\Theta}}$ *be defined by (2.13). Suppose the conditions (C1) and (C2) in Lemma 5 hold for a sufficiently small* R *and* $\mathcal{B}_q(\mathbf{\Theta}^*) \le \rho$*. Consider any solution* $\widehat{\mathbf{\Theta}}$ *to (2.13) with parameter* $\lambda = 2\nu\sqrt{\delta d\log d/n}$*, where* $\delta > 1$*. There exist constants* $\{C_i\}_{i=0}^4$ *such that as long as* $n > C_0 d\log d$*,*

$$\begin{aligned}\left\|\widehat{\mathbf{\Theta}}-\mathbf{\Theta}^*\right\|_F^2 &\le C_1\max\Big\{\rho\left(\sqrt{\frac{d\log d}{n}}\right)^{2-q}, \frac{R^2}{n}\Big\}\\ \left\|\widehat{\mathbf{\Theta}}-\mathbf{\Theta}^*\right\|_N &\le C_2\max\Big\{\rho\left(\sqrt{\frac{d\log d}{n}}\right)^{1-q}, \left(\rho\left(\frac{R^2}{n}\right)^{1-q}\right)^{\frac{1}{2-q}}\Big\}\end{aligned} \tag{2.17}$$

with probability at least $1 - C_3\exp\left(-C_4 d\log d\right) - 2d^{1-\delta}$*.*

Remark 2. *In Davenport et al. (2014), they derived that* $\left\|\widehat{\mathbf{\Theta}}-\mathbf{\Theta}^*\right\|_F^2 = O_P(\sqrt{\rho d/n})$

when $\mathbf{\Theta}^*$ *is exactly low-rank. This is slower than our rate* $O_P(\rho d \log d/n)$*. Moreover, we provide an extra bound on the nuclear norm of the error.*

3 Simulation study

3.1 Generalized matrix regression

In this section, we verify the statistical rates derived in (2.7) through simulations. We let $d = 20, 40$ and 60. For each dimension, we take n to be $1800, 3600, 5400, 7200, 9000$ and 10800. We set $\mathbf{\Theta}^* \in \mathbb{R}^{d\times d}$ with $r(\mathbf{\Theta}^*) = 5$ and all the nonzero singular values of $\mathbf{\Theta}^*$ equal to 1. Each design matrix $\mathbf{X}_i$ has i.i.d. entries from $\mathcal{N}(0,1)$ and $Y_i \sim \text{Bin}(0, \exp(\eta_i^*)/(1+\exp(\eta_i^*)))$, where $\eta_i^* = \langle \mathbf{\Theta}^*, \mathbf{X}_i\rangle$. We choose $\lambda \asymp \sqrt{d/n}$ and tune the constant before the rate for optimal performance.

Our simulation is based on 100 independent replications, where we record the estimation error in terms of the logarithmic Frobenius norm $\log\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*\|_F$. The averaged statistical error is plotted against the logarithmic sample size in Figure 1. As we can observe from the plot, the slope of curve is almost $-1/2$, which is consistent with the order of n in the statistical rate we derived for $\widehat{\mathbf{\Theta}}$. The intercept also matches the order of d in our theory. For example, in the plot, the difference between the green and red lines predicted by the theory is $(\log(60) - \log(40))/2 = 0.20$, which is in line with the empirical plot. Similarly, the difference between the red and black lines should be around $(\log(40) - \log(20))/2 = 0.35$, which is also consistent with the plot.

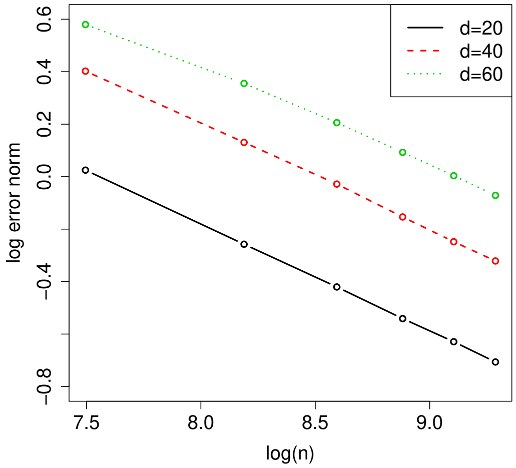

Figure 1: $\log\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*\|_F$ versus $\log(n)$ for different dimension d.

To solve the optimization problem (1.7), we exploit an iterative Peaceman-Rachford splitting method. We start from $\widehat{\mathbf{\Theta}}^{(0)} = \mathbf{0}$. In the kth step, we take the local quadratic approximation of $\mathcal{L}_n(\mathbf{\Theta})$ at $\mathbf{\Theta} = \mathbf{\Theta}^{(k-1)}$:

$$\begin{aligned}\mathcal{L}_n^{(k)}(\mathbf{\Theta}) =& \frac{1}{2}\text{vec}(\mathbf{\Theta} - \mathbf{\Theta}^{(k-1)})^T \nabla^2_{\mathbf{\Theta}} \mathcal{L}_n(\mathbf{\Theta}^{(k-1)})\text{vec}(\mathbf{\Theta} - \mathbf{\Theta}^{(k-1)}) \\ &+ \langle \nabla_{\mathbf{\Theta}} \mathcal{L}_n(\mathbf{\Theta}^{(k-1)}), \mathbf{\Theta} - \mathbf{\Theta}^{(k-1)}\rangle + \mathcal{L}_n(\mathbf{\Theta}^{(k-1)}).\end{aligned} \tag{3.1}$$

and then solve the following optimization problem to obtain $\widehat{\mathbf{\Theta}}^{(k)}$:

$$\widehat{\mathbf{\Theta}}^{(k)} = \text{argmin}_{\mathbf{\Theta}} \, \mathcal{L}_n^{(k)}(\mathbf{\Theta}) + \lambda \left\|\mathbf{\Theta}\right\|_N. \tag{3.2}$$

We borrow the algorithm from Fan et al. (2016) to solve the optimization problem (3.2). In Section 5.1 of Fan et al. (2016), they applied a contractive Peaceman-Rachford splitting method to solve a nuclear norm penalized least square problem:

$$\begin{aligned}\widehat{\mathbf{\Theta}} &= \text{argmin}_{\mathbf{\Theta}} \left\{ \frac{1}{n}\sum_{i=1}^n (Y_i - \langle \mathbf{\Theta}, \mathbf{X}_i\rangle)^2 + \lambda \left\|\mathbf{\Theta}\right\|_N \right\} \\ &= \text{argmin}_{\mathbf{\Theta}} \left\{ \text{vec}(\mathbf{\Theta})^T \frac{1}{n}\sum_{i=1}^n \text{vec}(\mathbf{X}_i)\text{vec}(\mathbf{X}_i)^T \text{vec}(\mathbf{\Theta}) + \langle \frac{2}{n}\sum_{i=1}^n Y_i \mathbf{X}_i, \mathbf{\Theta}\rangle + \lambda \left\|\mathbf{\Theta}\right\|_N \right\}.\end{aligned} \tag{3.3}$$

Construct

$$\tilde{\mathbf{X}}_i^{(k)} = \sqrt{b''(\langle \widehat{\mathbf{\Theta}}^{(k-1)}, \mathbf{X}_i\rangle)}\mathbf{X}_i$$

and

$$\tilde{Y}_i^{(k)} = b''(\langle \widehat{\mathbf{\Theta}}^{(k-1)}, \mathbf{X}_i\rangle)^{-\frac{1}{2}} \left[Y_i - b'(\langle \widehat{\mathbf{\Theta}}^{(k-1)}, \mathbf{X}_i\rangle)\right].$$

Some algebra shows that the following nuclear norm penalized least square problem is equivalent to (3.2)

$$\begin{aligned}\widehat{\mathbf{\Theta}}^{(k)} = \text{argmin}_{\mathbf{\Theta}} \Big\{ &\frac{1}{2}\text{vec}(\mathbf{\Theta} - \widehat{\mathbf{\Theta}}^{(k-1)})^T \frac{1}{n}\sum_{i=1}^n \text{vec}(\tilde{\mathbf{X}}_i^{(k)})\text{vec}(\tilde{\mathbf{X}}_i^{(k)})^T \text{vec}(\mathbf{\Theta} - \widehat{\mathbf{\Theta}}^{(k-1)}) \\ &+ \langle \frac{1}{n}\sum_{i=1}^n \tilde{Y}_i^{(k)} \tilde{\mathbf{X}}_i^{(k)}, \mathbf{\Theta} - \widehat{\mathbf{\Theta}}^{(k-1)}\rangle + \lambda \left\|\mathbf{\Theta}\right\|_N \Big\}.\end{aligned} \tag{3.4}$$

We can further write (3.4) as an optimization problem of minimizing the sum of two convex functions:

$$\begin{aligned} \underset{x}{\text{minimize}} \quad & \frac{1}{2n}\sum_{i=1}^{n}\Big(\tilde{Y}_i^{(k)} - \langle \mathbf{\Theta}_x, \tilde{\mathbf{X}}_i^{(k)}\rangle\Big)^2 + \lambda\,\|\mathbf{\Theta}_y\|_N \\ \text{subject to} \quad & \mathbf{\Theta}_x - \mathbf{\Theta}_y = -\mathbf{\Theta}^{(k-1)}. \end{aligned}$$

It has been explicitly explained in Fan et al. (2016) on how to solve the above optimization problem using the Peaceman-Rachford splitting method. We provide the algorithm that is specific to our problem here. Here we first define the singular value soft thresholding operator $\mathcal{S}_\tau(\cdot)$. For any $\mathbf{X} \in \mathbb{R}^{d\times d}$, let $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T$ be its SVD, where $\mathbf{U}$ and $\mathbf{V}$ are two orthonormal matrices and $\mathbf{D} = \mathrm{diag}(\sigma_1, \ldots, \sigma_d)$ with $\sigma_1 \geq \ldots \geq \sigma_d$. Then $\mathcal{S}_\tau(\mathbf{X}) := \mathbf{U}\widetilde{\mathbf{D}}\mathbf{V}^T$, where $\widetilde{\mathbf{D}} := \mathrm{diag}(\max(\sigma_1-\tau, 0), \max(\sigma_2-\tau, 0), \ldots, \max(\sigma_d-\tau, 0))$. Let $\mathbb{X}^{(k)}$ be an $n\times d^2$ matrix whose rows are $\mathrm{vec}(\tilde{\mathbf{X}}_i^{(k)})$ and $\mathbb{Y}^{(k)}$ be the response vector $\tilde{Y}^{(k)}$. For $\ell = 0, 1, \ldots,$

$$\begin{cases} \boldsymbol{\theta}_x^{(\ell+1)} = (2\mathbb{X}^{(k)\top}\mathbb{X}^{(k)}/n + \beta\cdot\mathbf{I})^{-1}(\beta\cdot(\boldsymbol{\theta}_y^{(\ell)} - \mathrm{vec}(\widehat{\mathbf{\Theta}}^{(k-1)})) + \boldsymbol{\rho}^{(\ell)} + 2\mathbb{X}^{(k)\top}\mathbb{Y}^{(k)}/n), \\ \boldsymbol{\rho}^{(\ell+\frac{1}{2})} = \boldsymbol{\rho}^{(\ell)} - \alpha\beta(\boldsymbol{\theta}_x^{(\ell+1)} - \boldsymbol{\theta}_y^{(\ell)} + \mathrm{vec}(\widehat{\mathbf{\Theta}}^{(k-1)})), \\ \boldsymbol{\theta}_y^{(\ell+1)} = \mathrm{vec}(\mathcal{S}_{2\lambda/\beta}(\mathrm{mat}(\boldsymbol{\theta}_x + \mathrm{vec}(\widehat{\mathbf{\Theta}}^{(k-1)}) - \boldsymbol{\rho}^{(\ell+\frac{1}{2})}/\beta))), \\ \boldsymbol{\rho}^{(\ell+1)} = \boldsymbol{\rho}^{(\ell+\frac{1}{2})} - \alpha\beta(\boldsymbol{\theta}_x^{(\ell+1)} + \mathrm{vec}(\widehat{\mathbf{\Theta}}^{(k-1)}) - \boldsymbol{\theta}_y^{(\ell+1)}), \end{cases} \tag{3.5}$$

where we choose $\alpha = 0.9$ and $\beta = 1$. $\boldsymbol{\theta}_x^{(\ell)}, \boldsymbol{\theta}_y^{(\ell)} \in \mathbb{R}^{d^2}$ for $\ell \geq 0$ and we can initialize them by $\boldsymbol{\theta}_x^{(0)} = \boldsymbol{\theta}_y^{(0)} = \mathbf{0}$. **Since both the objective function and the feasible set are convex, any initializer should work well theoretically. In practice, we can incorporate prior knowledge if any to choose the initializer for faster convergence.** When $\boldsymbol{\theta}_x^{(\ell)}$ and $\boldsymbol{\theta}_y^{(\ell)}$ converge, we reshape $\boldsymbol{\theta}_y^{(\ell)}$ as a $d\times d$ matrix and return it as $\widehat{\mathbf{\Theta}}^{(k)}$. We iterate this procedure until $\|\widehat{\mathbf{\Theta}}^{(k)} - \widehat{\mathbf{\Theta}}^{(k-1)}\|_F$ is smaller than 10^{-3} and return $\widehat{\mathbf{\Theta}}^{(k)}$ as the final estimator of $\mathbf{\Theta}^*$. The algorithm is concluded as follows in Algorithm 1.

3.2 Generalized reduced-rank regression

In this section, we let $d = 20, 40, 60, 80$ and 100. For each dimension, we take n to be $1800, 3600, 5400, 7200, 9000$ and 10800. We set the rank of $\mathbf{\Theta}^*$ to be 5 and let $\|\mathbf{\Theta}^*\|_F = 1$. For $1 \leq i \leq n$ and $1 \leq j \leq d$, we let the covariate $\mathbf{x}_i$ have i.i.d. entries from $\mathcal{N}(0, 1)$ and let y_{ij} follow $\mathrm{Bin}(0, \exp(\eta^*)/(1+\exp(\eta^*)))$ where $\eta^* = \langle \mathbf{\Theta}_j^*, \mathbf{x}_i\rangle$. We choose $\lambda \asymp \sqrt{d/n}$ and tune the constant before the rate for optimal performance. The experiment is repeated for 100 times and the logarithmic Frobenius norm of the

Algorithm 1 Deriving the estimator in generalized matrix regression

```
Take $\widehat{\mathbf{\Theta}}^{(0)} = \mathbf{0} \in \mathbb{R}^{d\times d}$, $k \leftarrow 1$
*loop 1*:
$\tilde{\mathbf{X}}_i^{(k)} = \sqrt{b''(\langle \widehat{\mathbf{\Theta}}^{(k-1)}, \mathbf{X}_i\rangle)}\mathbf{X}_i$ for $1 \le i \le n$
$\mathbb{X}^{(k)} = \left(\mathrm{vec}(\tilde{\mathbf{X}}_1^{(k)})^T, \mathrm{vec}(\tilde{\mathbf{X}}_2^{(k)})^T, ..., \mathrm{vec}(\tilde{\mathbf{X}}_n^{(k)})^T\right)^T \in \mathbb{R}^{n\times d^2}$
$\tilde{Y}_i^{(k)} = b''(\langle \widehat{\mathbf{\Theta}}^{(k-1)}, \mathbf{X}_i\rangle)^{-\frac{1}{2}}\left[Y_i - b'(\langle \widehat{\mathbf{\Theta}}^{(k-1)}, \mathbf{X}_i\rangle)\right]$ for $1 \le i \le n$
$\mathbb{Y}^{(k)} = \left(\tilde{Y}_1^{(k)}, \tilde{Y}_2^{(k)}, ..., \tilde{Y}_n^{(k)}\right)^T$
Take $\boldsymbol{\theta}_x^{(0)} = \boldsymbol{\theta}_y^{(0)} = \mathbf{0} \in \mathbb{R}^{d^2}$, $\alpha = 0.9$, $\beta = 1$, $\ell \leftarrow 0$
*loop 2*:
    $\boldsymbol{\theta}_x^{(\ell+1)} = (2\mathbb{X}^{(k)\top}\mathbb{X}^{(k)}/n + \beta\cdot\mathbf{I})^{-1}(\beta\cdot(\boldsymbol{\theta}_y^{(\ell)} - \mathrm{vec}(\widehat{\mathbf{\Theta}}^{(k-1)})) + \boldsymbol{\rho}^{(\ell)} + 2\mathbb{X}^{(k)\top}\mathbb{Y}^{(k)}/n)$
    $\boldsymbol{\rho}^{(\ell+\frac{1}{2})} = \boldsymbol{\rho}^{(\ell)} - \alpha\beta(\boldsymbol{\theta}_x^{(\ell+1)} - \boldsymbol{\theta}_y^{(\ell)} + \mathrm{vec}(\widehat{\mathbf{\Theta}}^{(k-1)}))$
    $\boldsymbol{\theta}_y^{(\ell+1)} = \mathrm{vec}(\mathcal{S}_{2\lambda/\beta}(\mathrm{mat}(\boldsymbol{\theta}_x + \mathrm{vec}(\widehat{\mathbf{\Theta}}^{(k-1)}) - \boldsymbol{\rho}^{(\ell+\frac{1}{2})}/\beta)))$
    $\boldsymbol{\rho}^{(\ell+1)} = \boldsymbol{\rho}^{(\ell+\frac{1}{2})} - \alpha\beta(\boldsymbol{\theta}_x^{(\ell+1)} + \mathrm{vec}(\widehat{\mathbf{\Theta}}^{(k-1)}) - \boldsymbol{\theta}_y^{(\ell+1)})$
    If $\left\|\boldsymbol{\theta}_y^{(\ell+1)} - \boldsymbol{\theta}_y^{(\ell)}\right\|_F < \epsilon_1$, **close**
    $\ell \leftarrow \ell + 1$, **goto** *loop 2*
Take $\widehat{\mathbf{\Theta}}^{(k)} = \mathrm{mat}(\theta_y^{(\ell)}) \in \mathbb{R}^{d\times d}$
If $\left\|\widehat{\mathbf{\Theta}}^{(k)} - \widehat{\mathbf{\Theta}}^{(k-1)}\right\|_F < \epsilon_2$, **close**
$k \leftarrow k + 1$, **goto** *loop 1*
**return** $\widehat{\mathbf{\Theta}}^{(k)}$.
```

estimation error is recorded in each repetition. We plot the averaged statistical error in Figure 2.

We can see from the figure that the logarithmic error decays as logarithmic sample size grows and the slope is almost $-1/2$.

As for the implementation, we again use the iterative Peaceman-Rachford splitting method to solve for the estimator. We start from $\widehat{\mathbf{\Theta}}^{(0)} = \mathbf{0}$. In the kth step ($k \ge 1$), let

$$\mathbf{S}^{(k)} = \frac{1}{nd}\sum_{i=1}^{n}\sum_{j=1}^{d}\frac{\exp(\langle \widehat{\mathbf{\Theta}}_j^{(k-1)}, \mathbf{x}_i\rangle)}{(1+\exp(\langle \widehat{\mathbf{\Theta}}_j^{(k-1)}, \mathbf{x}_i\rangle))^2}\mathbf{x}_i\mathbf{x}_i^T,$$

$$\tilde{y}_{ij}^{(k)} = y_{ij} - \frac{\exp(\langle \widehat{\mathbf{\Theta}}_j^{(k-1)}, \mathbf{x}_i\rangle)}{1+\exp(\langle \widehat{\mathbf{\Theta}}_j^{(k-1)}, \mathbf{x}_i\rangle)} \quad \text{and} \quad \mathbf{T}^{(k)} = \sum_{i=1}^{n}\mathbf{x}_i\tilde{\mathbf{y}}_i^T.$$

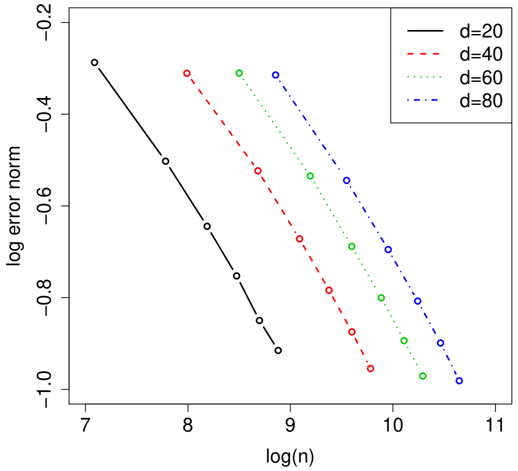

Figure 2: $\log\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*\|_F$ versus $\log(n)$ for different dimension d.

We iterate the following algorithm to solve for $\widehat{\mathbf{\Theta}}^{(k)}$. Here $\alpha = 0.9$ and $\beta = 1$.

$$\begin{cases} \mathbf{\Theta}_x^{(\ell+1)} = (2\mathbf{S}^{(k)}/n + \beta \cdot \mathbf{I})^{-1}(\beta \cdot (\mathbf{\Theta}_y^{(\ell)} - \widehat{\mathbf{\Theta}}^{(k-1)}) + \boldsymbol{\rho}^{(\ell)} + 2\mathbf{T}^{(k)}/n), \\ \boldsymbol{\rho}^{(\ell+\frac{1}{2})} = \boldsymbol{\rho}^{(\ell)} - \alpha\beta(\mathbf{\Theta}_x^{(\ell+1)} + \widehat{\mathbf{\Theta}}^{(k-1)} - \mathbf{\Theta}_y^{(\ell)}), \\ \mathbf{\Theta}_y^{(\ell+1)} = \mathcal{S}_{2\lambda/\beta}(\mathbf{\Theta}_x + \widehat{\mathbf{\Theta}}^{(k-1)} - \boldsymbol{\rho}^{(\ell+\frac{1}{2})}/\beta), \\ \boldsymbol{\rho}^{(\ell+1)} = \boldsymbol{\rho}^{(\ell+\frac{1}{2})} - \alpha\beta(\mathbf{\Theta}_x^{(\ell+1)} + \widehat{\mathbf{\Theta}}^{(k-1)} - \mathbf{\Theta}_y^{(\ell+1)}). \end{cases} \tag{3.6}$$

Here, $\mathcal{S}_\tau(\cdot)$ is the singular value soft thresholding function we introduced in Section 3.1. Note that $\mathbf{\Theta}_x^{(\ell)}, \mathbf{\Theta}_y^{(\ell)} \in \mathbb{R}^{d\times d}$ for all $\ell \geq 0$ and they are irrelevant to $\widehat{\mathbf{\Theta}}^{(k)}$ though they share similar notations. We start from $\mathbf{\Theta}_x^{(0)} = \mathbf{\Theta}_y^{(0)} = \mathbf{0}$ and iterate this procedure until they converge. We return the last $\mathbf{\Theta}_y^{(\ell)}$ to be $\widehat{\mathbf{\Theta}}^{(k)}$.

We repeat the above algorithm until $\|\widehat{\mathbf{\Theta}}^{(k)} - \widehat{\mathbf{\Theta}}^{(k-1)}\|_F$ is smaller than 10^{-3} and take $\widehat{\mathbf{\Theta}}^{(k)}$ as the final estimator of $\mathbf{\Theta}^*$. The algorithm is concluded in Algorithm 2.

3.3 One-bit matrix completion

3.3.1 Statistical consistency

We consider $\mathbf{\Theta}^* \in \mathbb{R}^{d\times d}$ with dimension $d = 20, 40, 60$ and 80. For each dimension, we consider 6 different values for n such that $n/(d\log d) = 30, 60, 90, 120, 150$ and 180. We let $\mathrm{r}(\mathbf{\Theta}^*) = 5$, $\|\mathbf{\Theta}^*\|_F = 1$ and $R = 2\,\|\mathbf{\Theta}^*\|_\infty$. The design matrix $\mathbf{X}_i$ is a singleton and it is uniformly sampled from $\{\mathbf{e}_j\mathbf{e}_k^T\}_{1\leq j,k\leq d}$. We choose $\lambda \asymp \sqrt{d\log(d)/n}$ and tune the

Algorithm 2 Deriving the estimator in reduced-rank regression

```
Take $\widehat{\mathbf{\Theta}}^{(0)} = \mathbf{0} \in \mathbb{R}^{d\times d}$, $k \leftarrow 1$
*loop 1*:
$\mathbf{S}^{(k)} = \frac{1}{nd}\sum_{i=1}^{n}\sum_{j=1}^{d}\frac{\exp(\langle\widehat{\mathbf{\Theta}}_j^{(k-1)},\mathbf{x}_i\rangle)}{(1+\exp(\langle\widehat{\mathbf{\Theta}}_j^{(k-1)},\mathbf{x}_i\rangle))^2}\mathbf{x}_i\mathbf{x}_i^T$
$\tilde{y}_{ij}^{(k)} = y_{ij} - \frac{\exp(\langle\widehat{\mathbf{\Theta}}_j^{(k-1)},\mathbf{x}_i\rangle)}{1+\exp(\langle\widehat{\mathbf{\Theta}}_j^{(k-1)},\mathbf{x}_i\rangle)}$
$\mathbf{T}^{(k)} = \sum_{i=1}^{n}\mathbf{x}_i\tilde{\mathbf{y}}_i^T$
Take $\mathbf{\Theta}_x^{(0)} = \mathbf{\Theta}_y^{(0)} = \mathbf{0} \in \mathbb{R}^{d\times d}$, $\alpha = 0.9$, $\beta = 1$, $\ell \leftarrow 0$
*loop 2*:
    $\mathbf{\Theta}_x^{(\ell+1)} = (2\mathbf{S}^{(k)}/n + \beta\cdot\mathbf{I})^{-1}(\beta\cdot(\mathbf{\Theta}_y^{(\ell)} - \widehat{\mathbf{\Theta}}^{(k-1)}) + \boldsymbol{\rho}^{(\ell)} + 2\mathbf{T}^{(k)}/n)$
    $\boldsymbol{\rho}^{(\ell+\frac{1}{2})} = \boldsymbol{\rho}^{(\ell)} - \alpha\beta(\mathbf{\Theta}_x^{(\ell+1)} + \widehat{\mathbf{\Theta}}^{(k-1)} - \mathbf{\Theta}_y^{(\ell)})$
    $\mathbf{\Theta}_y^{(\ell+1)} = \mathcal{S}_{2\lambda/\beta}(\mathbf{\Theta}_x + \widehat{\mathbf{\Theta}}^{(k-1)} - \boldsymbol{\rho}^{(\ell+\frac{1}{2})}/\beta)$
    $\boldsymbol{\rho}^{(\ell+1)} = \boldsymbol{\rho}^{(\ell+\frac{1}{2})} - \alpha\beta(\mathbf{\Theta}_x^{(\ell+1)} + \widehat{\mathbf{\Theta}}^{(k-1)} - \mathbf{\Theta}_y^{(\ell+1)})$
    If $\left\|\mathbf{\Theta}_y^{(\ell+1)} - \mathbf{\Theta}_y^{(\ell)}\right\|_F < \epsilon_1$, **close**
    $\ell \leftarrow \ell + 1$, **goto** *loop 2*
Take $\widehat{\mathbf{\Theta}}^{(k)} = \Theta_y^{(\ell)} \in \mathbb{R}^{d\times d}$
If $\left\|\widehat{\mathbf{\Theta}}^{(k)} - \widehat{\mathbf{\Theta}}^{(k-1)}\right\|_F < \epsilon_2$, **close**
$k \leftarrow k + 1$, **goto** *loop 1*
**return** $\widehat{\mathbf{\Theta}}^{(k)}$.
```

constant before the rate for optimal performance. The experiment is repeated for 100 times and the logarithmic Frobenius norm of the estimation error is recorded in each repetition. We plot the averaged statistical error against the logarithmic sample size in Figure 3.

We can see from the left panel in Figure 3 that $\log\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*\|_F$ decays as $\log n$ grows and the slope is almost $-1/2$. Meanwhile, Theorem 4 says that $\log\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*\|_F$ should be proportional to $\log(d\log d/n)$. The right panel of Figure 3 verifies this rate: it shows that the statistical error curves for different dimensions are well-aligned if we adjust the sample size to be $n/d\log d$.

To solve the optimization problem in (2.13), we exploit the ADMM method used in Section 5.2 in Fan et al. (2016). In Fan et al. (2016), they minimized a quadratic loss function with a nuclear norm penalty under elementwise max norm constraint. Our goal is to replace the quadratic loss therein with negative log-likelihood and solve the optimization problem. Here we iteratively call the ADMM method in Fan et al.

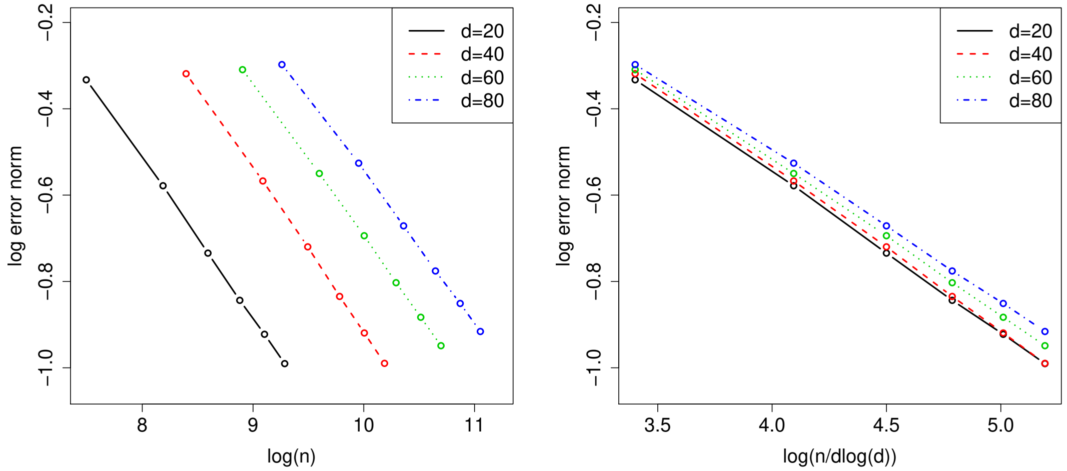

Figure 3: $\log\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*\|_F$ versus $\log(n)$ and $\log(n/d\log d)$.

(2016) to solve a series of optimization problems whose loss function is local quadratic approximation of the negative log-likelihood. We initialize $\mathbf{\Theta}$ with $\widehat{\mathbf{\Theta}}^{(0)} = \mathbf{0}$ and introduce the algorithm below.

In the kth step, we take the local quadratic approximation of $\mathcal{L}_n(\mathbf{\Theta})$ at $\mathbf{\Theta} = \widehat{\mathbf{\Theta}}^{(k-1)}$:

$$\begin{aligned}\mathcal{L}_n^{(k)}(\mathbf{\Theta}) =& \frac{1}{2}\text{vec}(\mathbf{\Theta} - \widehat{\mathbf{\Theta}}^{(k-1)})^T \nabla^2_{\mathbf{\Theta}} \mathcal{L}_n(\widehat{\mathbf{\Theta}}^{(k-1)})\text{vec}(\mathbf{\Theta} - \widehat{\mathbf{\Theta}}^{(k-1)}) \\ &+ \langle \nabla_{\mathbf{\Theta}} \mathcal{L}_n(\widehat{\mathbf{\Theta}}^{(k-1)}), \mathbf{\Theta} - \widehat{\mathbf{\Theta}}^{(k-1)}\rangle + \mathcal{L}_n(\widehat{\mathbf{\Theta}}^{(k-1)}).\end{aligned} \tag{3.7}$$

and solve the following optimization problem to obtain $\widehat{\mathbf{\Theta}}^{(k)}$:

$$\widehat{\mathbf{\Theta}}^{(k)} = \text{argmin}_{\mathbf{\Theta}}\, \mathcal{L}_n^{(k)}(\mathbf{\Theta}) + \lambda \|\mathbf{\Theta}\|_N. \tag{3.8}$$

To solve the above optimization problem, we borrow the algorithm proposed in Fang et al. (2015). Let $\mathbf{L}, \mathbf{R}, \mathbf{W} \in \mathbb{R}^{2d\times 2d}$ be the variables in our algorithm and let $\mathbf{L}^{(0)} = \mathbf{R}^{(0)} = \mathbf{0}$. Define

$$\mathbf{\Theta}^a_{jk} = \sum_{i=1}^n \frac{\exp(\langle \mathbf{\Theta}, \mathbf{X}_i\rangle)}{(1+\exp(\langle \mathbf{\Theta}, \mathbf{X}_i\rangle))^2} \mathbb{1}_{\{\mathbf{X}_i = \mathbf{e}_j\mathbf{e}_k^T\}},$$

$$\mathbf{\Theta}^b_{jk} = \sum_{i=1}^n \left[Y_i - \frac{\exp(\langle \mathbf{\Theta}, \mathbf{X}_i\rangle)}{1+\exp(\langle \mathbf{\Theta}, \mathbf{X}_i\rangle)}\right] \mathbb{1}_{\{\mathbf{X}_i = \mathbf{e}_j\mathbf{e}_k^T\}}.$$

We introduce the algorithms of the variables in our problem and interested readers can refer to Fang et al. (2015) for the technical details in the derivation and stopping criteria of the algorithm. For $\ell \geq 0$,

$$\begin{cases} \mathbf{L}^{(\ell+1)} = \Pi_{\mathcal{S}_+^{2d}} \left\{ \mathbf{R}^{(\ell)} + \begin{pmatrix} \mathbf{0} & \widehat{\boldsymbol{\Theta}}^{(k-1)} \\ \widehat{\boldsymbol{\Theta}}^{(k-1)} & \mathbf{0} \end{pmatrix} - \rho^{-1}(\mathbf{W}^{(\ell)} + 2\lambda \mathbf{I}) \right\} \\ = \begin{pmatrix} [\mathbf{L}^{(\ell+1)}]^{11} & [\mathbf{L}^{(\ell+1)}]^{12} \\ [\mathbf{L}^{(\ell+1)}]^{21} & [\mathbf{L}^{(\ell+1)}]^{22} \end{pmatrix}, \\ \mathbf{C} = \begin{pmatrix} \mathbf{C}^{11} & \mathbf{C}^{12} \\ \mathbf{C}^{21} & \mathbf{C}^{22} \end{pmatrix} = \mathbf{L}^{(\ell+1)} - \begin{pmatrix} \mathbf{0} & \widehat{\boldsymbol{\Theta}}^{(k-1)} \\ \widehat{\boldsymbol{\Theta}}^{(k-1)} & \mathbf{0} \end{pmatrix} + \mathbf{W}^{(\ell)}/\rho, \\ \mathbf{R}^{12}_{jk} = \Pi_{[-R,R]} \left\{ (\rho \mathbf{C}^{12}_{jk} + 2\boldsymbol{\Theta}^{b}_{jk}/n)/(\rho + 2\boldsymbol{\Theta}^{a}_{jk}/n) \right\}, 1 \leq j \leq d, 1 \leq k \leq d, \\ \mathbf{R}^{(\ell+1)} = \begin{pmatrix} \mathbf{C}^{11} & \mathbf{R}^{(12)} \\ (\mathbf{R}^{12})^T & \mathbf{C}^{22} \end{pmatrix}, \\ \mathbf{W}^{(\ell+1)} = \mathbf{W}^{(\ell)} + \gamma\rho(\mathbf{L}^{(\ell+1)} - \mathbf{R}^{(\ell+1)} - \begin{pmatrix} \mathbf{0} & \widehat{\boldsymbol{\Theta}}^{(k-1)} \\ \widehat{\boldsymbol{\Theta}}^{(k-1)} & \mathbf{0} \end{pmatrix}). \end{cases} \tag{3.9}$$

In the algorithm, $\Pi_{\mathcal{S}_+^{2d}}(\cdot)$ represents the projection operator onto the space of positive semidefinite matrices $\mathcal{S}_+^{2d}$, ρ is taken to be 0.1 and γ is the step length which is set to be 1.618. When the algorithm converges and stops, we elementwise truncate $\mathbf{L}^{12}$ at the level of R and return the truncated $\tilde{\mathbf{L}}^{12}$ as $\widehat{\boldsymbol{\Theta}}^{(k)}$. Specifically, $\tilde{\mathbf{L}}^{12}_{jk} = \mathrm{sgn}(\mathbf{L}^{12}_{jk})(|\mathbf{L}^{12}_{jk}| \wedge R)$ for $1 \leq j \leq d, 1 \leq k \leq d$.

When $\|\widehat{\boldsymbol{\Theta}}^{(k)} - \widehat{\boldsymbol{\Theta}}^{(k-1)}\|_F$ is smaller than 10^{-3}, we return $\widehat{\boldsymbol{\Theta}}^{(k)}$ as our final estimator of $\boldsymbol{\Theta}^*$. We summarize the algorithm in Algorithm 3.

3.3.2 Comparison between GLM and linear model

As we mentioned in the introduction, the motivation of generalizing trace regression is to accommodate the dichotomous response in recommending systems such as Netflix Challenge, Kiva, etc. In this section, we compare the performance of generalized trace regression and standard trace regression in predicting discrete ratings.

The setting is very similar to the last section. We set $\boldsymbol{\Theta}^*$ to be a square matrix with dimension $d = 20, 40, 60$ and 80. We let $r(\boldsymbol{\Theta}^*) = 5$ and its eigenspace be that of the sample covariance matrix of 100 random vectors following $\mathcal{N}(\mathbf{0}, \mathbf{I}_d)$. For each dimension, we consider 10 different values for n such that $n/d \log d = 1, 2, ..., 10$. and

Algorithm 3 Deriving the estimator in 1-bit matrix completion

```
Take $\widehat{\mathbf{\Theta}}^{(0)} = \mathbf{0} \in \mathbb{R}^{d\times d}$, $k \leftarrow 1$
*loop 1*:
$\mathbf{\Theta}^a_{jk} = \sum_{i=1}^{n} \frac{\exp(\langle \mathbf{\Theta}, \mathbf{X}_i\rangle)}{(1+\exp(\langle \mathbf{\Theta}, \mathbf{X}_i\rangle))^2} \mathbb{1}_{\{\mathbf{X}_i = \mathbf{e}_j \mathbf{e}_k^T\}}$
$\mathbf{\Theta}^b_{jk} = \sum_{i=1}^{n} \left[ Y_i - \frac{\exp(\langle \mathbf{\Theta}, \mathbf{X}_i\rangle)}{1+\exp(\langle \mathbf{\Theta}, \mathbf{X}_i\rangle)} \right] \mathbb{1}_{\{\mathbf{X}_i = \mathbf{e}_j \mathbf{e}_k^T\}}$
Take $\mathbf{L}^{(0)} = \mathbf{R}^{(0)} = \mathbf{0} \in \mathbb{R}^{2d\times 2d}$, $\rho = 0.1$, $\gamma = 1.618$, $\ell \leftarrow 0$
*loop 2*:
    $\mathbf{L}^{(\ell+1)} = \Pi_{\mathcal{S}^{2d}_+} \left\{ \mathbf{R}^{(\ell)} + \begin{pmatrix} \mathbf{0} & \widehat{\mathbf{\Theta}}^{(k-1)} \\ \widehat{\mathbf{\Theta}}^{(k-1)} & \mathbf{0} \end{pmatrix} - \rho^{-1}(\mathbf{W}^{(\ell)} + 2\lambda \mathbf{I}) \right\} = \begin{pmatrix} [\mathbf{L}^{(\ell+1)}]^{11} & [\mathbf{L}^{(\ell+1)}]^{12} \\ [\mathbf{L}^{(\ell+1)}]^{21} & [\mathbf{L}^{(\ell+1)}]^{22} \end{pmatrix}$
    $\mathbf{C} = \begin{pmatrix} \mathbf{C}^{11} & \mathbf{C}^{12} \\ \mathbf{C}^{21} & \mathbf{C}^{22} \end{pmatrix} = \mathbf{L}^{(\ell+1)} - \begin{pmatrix} \mathbf{0} & \widehat{\mathbf{\Theta}}^{(k-1)} \\ \widehat{\mathbf{\Theta}}^{(k-1)} & \mathbf{0} \end{pmatrix} + \mathbf{W}^{(\ell)}/\rho$
    $\mathbf{R}^{12}_{jk} = \Pi_{[-R,R]} \left\{ (\rho \mathbf{C}^{12}_{jk} + 2\mathbf{\Theta}^b_{jk}/n)/(\rho + 2\mathbf{\Theta}^a_{jk}/n) \right\}, 1 \le j \le d, 1 \le k \le d$
    $\mathbf{R}^{(\ell+1)} = \begin{pmatrix} \mathbf{C}^{11} & \mathbf{R}^{(12)} \\ (\mathbf{R}^{12})^T & \mathbf{C}^{22} \end{pmatrix}$
    $\mathbf{W}^{(\ell+1)} = \mathbf{W}^{(\ell)} + \gamma\rho(\mathbf{L}^{(\ell+1)} - \mathbf{R}^{(\ell+1)} - \begin{pmatrix} \mathbf{0} & \widehat{\mathbf{\Theta}}^{(k-1)} \\ \widehat{\mathbf{\Theta}}^{(k-1)} & \mathbf{0} \end{pmatrix})$
    If $\left\| \mathbf{L}^{(\ell+1)} - \mathbf{L}^{(\ell)} \right\|_F < \epsilon_1$, **close**
    $\ell \leftarrow \ell + 1$, **goto** *loop 2*
Take $\widehat{\mathbf{\Theta}}^{(k)} = [\tilde{\mathbf{L}}^{(\ell+1)}]^{12} \in \mathbb{R}^{d\times d}$
If $\left\| \widehat{\mathbf{\Theta}}^{(k)} - \widehat{\mathbf{\Theta}}^{(k-1)} \right\|_F < \epsilon_2$, **close**
$k \leftarrow k + 1$, **goto** *loop 1*
**return** $\widehat{\mathbf{\Theta}}^{(k)}$.
```

generate the true rating matrix $\mathbf{T}$ in the following way:

$$T_{i,j} = \begin{cases} 1 & \text{w.p.} \quad \frac{\exp(\mathbf{\Theta}^*_{ij})}{1+\exp{(\mathbf{\Theta}^*_{ij})}} \\ 0 & \text{w.p.} \quad \frac{1}{1+\exp{(\mathbf{\Theta}^*_{ij})}} \end{cases} \qquad 1 \le i \le d, 1 \le j \le d.$$

We will show that generalized trace regression outperforms the linear trace regression in prediction.

We predict the ratings in two different ways. We first estimate the underlying $\mathbf{\Theta}^*$

with nuclear norm regularized logistic regression model. We set $\lambda = 0.2\sqrt{d \log d/n}$ and derive the estimator $\widehat{\mathbf{\Theta}}^{(1)}$ according to (2.13). We estimate the rating matrix $\mathbf{T}$ by $\widehat{\mathbf{T}}^{(1)}$ as defined below:

$$\widehat{\mathbf{T}}_{ij}^{(1)} = \begin{cases} 1 & \text{if} \quad \widehat{\mathbf{\Theta}}_{ij}^{(1)} \geq 0 \\ 0 & \text{else} \end{cases} .$$

The second method is to estimate $\mathbf{\Theta}^*$ with nuclear norm regularized linear model. Again, we take the tuning parameter $\lambda = 0.2\sqrt{d \log d/n}$ and derive the estimator $\widehat{\mathbf{\Theta}}^{(2)}$ as follows:

$$\widehat{\mathbf{\Theta}}^{(2)} = \operatorname{argmin}_{\|\mathbf{\Theta}\|_\infty \leq R} \left\{ \frac{1}{n} \sum_{i=1}^{n} (Y_i - \langle \mathbf{\Theta}^*, \mathbf{X}_i \rangle)^2 + \lambda \left\| \mathbf{\Theta} \right\|_N \right\}. \tag{3.10}$$

To estimate the rating matrix $\mathbf{T}$, we use

$$\widehat{T}_{ij}^{(2)} = \begin{cases} 1 & \text{if} \quad \widehat{\mathbf{\Theta}}_{ij}^{(2)} \geq 0.5 \\ 0 & \text{else} \end{cases} .$$

The experiment is repeated for 100 times. In each repetition, we record the prediction accuracy as $1 - \|\widehat{\mathbf{T}}^{(k)} - \mathbf{T}\|_F^2/d^2$ for $k = 1$ and 2, which is the proportion of correct predictions. We plot the average prediction accuracy in Figure 4.

We use solid lines to denote the prediction accuracy achieved by regularized GLM and we use dotted lines to denote the accuracy achieved by regularized linear model. We can see from Figure 4 that no matter how the dimension changes, the solid lines are always above the dotted lines, showing that the generalized model always outperforms the linear model with categorical response. This validates our motivation to use the generalized model in matrix recovery problems with categorical outcomes.

4 Real data analysis

In this section, we apply generalized trace regression with nuclear norm regularization to stock return prediction and image classification. The former can be regarded as a reduced rank regression and the latter can be seen as the categorical responses with matrix inputs. The results demonstrate the advantage of recruiting nuclear norm penalty compared with no penalty or using ℓ_1-norm regularization.

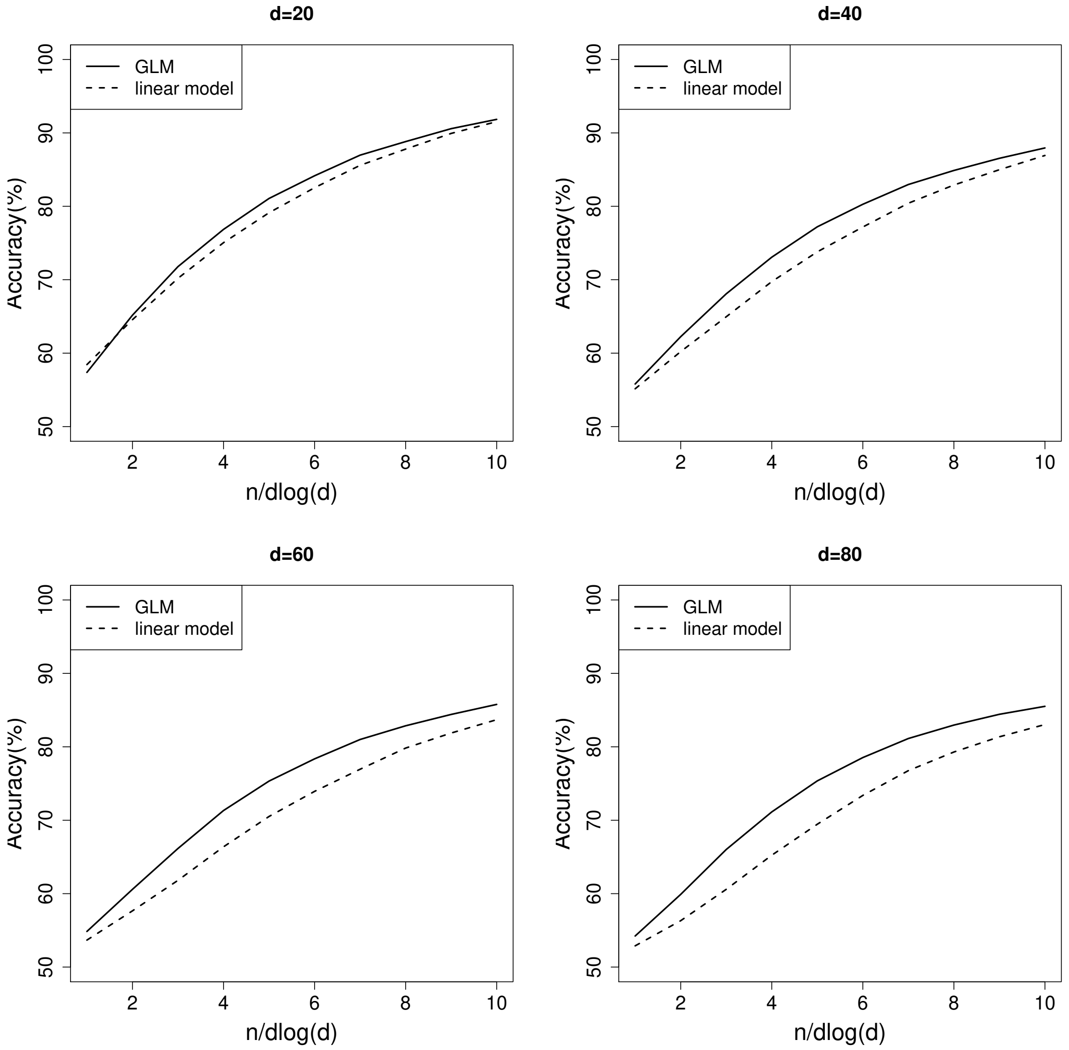

Figure 4: Prediction accuracy $1 - \|\widehat{\mathbf{T}} - \mathbf{T}\|_F^2/d^2$ in matrix completion for various dimension d.

4.1 Stock return prediction

In this subsection, we aim to predict the sign of the one-day forward stock return, i.e., whether the price of the stock will rise or fall in the next day. **Through nuclear norm regularization, we try to learn a small number of eigen-portfolios whose historical returns have prediction power in the future return direction of all the stocks of interest. For readers who are interested in predicting stock directions, either long-term or short-term, please also refer to Pesaran and Timmermann (2002, 2004); Lunde and Timmermann (2004); Huang et al.**

(2005); Kara et al. (2011), among others.

We pick 19 individual stocks as our objects of study: AAPL, BAC, BRK-B, C, COP, CVX, DIS, GE, GOOGL, GS, HON, JNJ, JPM, MRK, PFE, UNH, V, WFC and XOM. These are the largest holdings of Vanguard ETF in technology, health care, finance, energy, industrials and consumer. We also include S&P500 in our pool of stocks since it represents the market portfolio and should help the prediction. Therefore, we have $d_1 = 20$ stocks in total. We collect the daily returns of these stocks from 01/01/13 to 8/31/2017 and divide them into the training set (2013-2014), the evaluation set (2015) and the testing set (2016-2017). The sample sizes of the training, evaluation and testing sets are $n_1 = 504, n_2 = 252$ and $n_3 = 420$ respectively.

We fit a generalized reduced-rank regression model (2.8) based on the moving average (MA) of returns of each stock in the past 1 day, 3 days, 5 days, 10 days and 20 days. Hence, the dimension of $\mathbf{x}_i$ is $20 \times 5 = 100$. Let $\mathbf{y}_i \in \mathbb{R}^{20}$ be the sign of returns of the selected stocks on the $(i + 1)$th day. We assume that $\mathbf{\Theta}^* \in \mathbb{R}^{20\times100}$ is a near low-rank matrix, considering high correlations across the returns of the selected stocks. We tune λ for the best performance on the evaluation data. When we predict on the test set, we will update $\widehat{\mathbf{\Theta}}$ on a monthly basis, i.e., for each month in the testing set, we refit (2.8) based on the data in the most recent three years. Given an estimator $\widehat{\mathbf{\Theta}}$, our prediction $\widehat{\mathbf{y}}_j$ are the signs of $(\widehat{\mathbf{\Theta}}^T \mathbf{x}_j)$.

We have two baseline models in our analysis. The first one is the deterministic bet (DB): if a stock has more positive returns than negative ones in the training set, we always predict positive returns; otherwise, we always predict negative returns. The second one is the generalized RRR without any nuclear norm regularization. We use this baseline to demonstrate the advantage of incorporating nuclear norm regularization.

From Table 1, we can see that the nuclear norm penalized model yields an average accuracy of 53.89% while the accuracy of the unpenalized model and DB are 52.74% and 51.62%. Note that the penalized model performs the same as or better than the unpenalized model in 18 out of 20 stocks. When compared with the DB, the penalized model performs better in 15 out of the 20 stocks. The improvement in the overall performance illustrates the advantage of using generalized RRR with nuclear norm regularization.

4.2 CIFAR10 Dataset

Besides the application in finance, we also apply our model to the well-known CIFAR10 dataset in image classification. The CIFAR10 dataset has 60,000 colored 32×32 images in 10 classes: the airplane, automobile, bird, cat, dog, deer, dog, frog, horse, ship and truck. Each figure has three channels (red, green and blue) and hence is stored as

Stock	DB	Prediction Accuracy with Regularization	Prediction Accuracy without Regularization
AAPL	55.13	51.07	51.07
BAC	47.26	49.88	49.64
BRK-B	54.18	59.90	59.90
C	52.98	51.55	51.07
COP	47.49	54.18	54.18
CVX	48.69	55.37	54.18
DIS	49.40	56.80	56.80
GE	48.45	55.61	56.09
GOOGL	53.94	52.74	52.74
GS	52.74	53.22	47.49
HON	56.09	51.55	51.31
JNJ	51.79	54.65	53.70
JPM	52.27	53.94	47.02
MRK	51.55	51.31	51.31
PFE	49.40	52.27	49.40
UNH	52.74	53.70	52.74
V	56.09	58.00	58.23
WFC	49.16	52.74	50.12
XOM	48.21	54.42	53.46
SPY	54.89	54.89	54.42
Average	51.62	53.89	52.74

Table 1: Prediction Result of 20 selected stocks.(Unit: %)

a 32 × 96 matrix. We represent the 10 classes with the numbers 0,1, ..., 9. The training data contains 50,000 figures and the testing data contains 10,000 figures. **In our work, we only use 10,000 samples to train the model since we intend to illustrate how the regularizations alleviate the overfitting problem; after all, overfitting would not be a problem when the sample size was large.**

We construct and train a convolutional neural network (CNN) with ℓ_1 norm and nuclear norm regularization on $\boldsymbol{\Theta}$ respectively to learn the pattern of the figures. The naïve GLM is inappropriate for image classification, since pixel values are meaningless features as regard to the content of the picture. For example, two different images in the class "truck" might have trucks in different positions or colors, leading to dramatically different pixel values of the pictures. To extract useful features from pictures, we resort to the CNN. The structure of the CNN follows the online tutorial from

TensorFlow*. It extracts a 384-dimensional feature vector from each image and maps it to 10 categories through logistic regression with a 384×10 coefficient matrix. Here to exploit potential matrix structure of the features, we reshape this 384-dimensional feature vector into a 24×16 matrix and map it to one of the ten categories through generalized trace regression with ten 24×16 coefficient matrices. We penalize these coefficient matrices by their nuclear norm and ℓ_1-norm respectively and we summarize our results in Table 2 below.

λ	0	0.02	0.05	0.1	0.2	0.3
nuclear penalty	74.30%	76.04%	76.17%	75.29%	74.45%	73.46%
λ	0	0.001	0.002	0.005	0.008	0.01
ℓ_1 penalty	74.30%	75.70%	75.90%	75.53%	75.37%	75.22%

Table 2: Prediction accuracy in CIFAR10 under different λ with different penalties with convolutional neural network.

The results show that both regularization methods promote the prediction accuracy while nuclear norm regularization again outperforms ℓ_1 norm.

5 Discussion

Our theory is established upon assumptions of i.i.d. samples. It is possible to relax this i.i.d. assumption under the existing framework. As shown in Theorem 1, the statistical error rate of $\widehat{\boldsymbol{\Theta}}$ depends on two conditions on the tuning parameter λ and LRSC respectively. When the samples are not i.i.d., we need to verify these two conditions accordingly. For example, if we have Markov chain samples, we might recruit concentration results in Lezaud (1998), Paulin (2015) or Fan et al. (2018) to verify the required two conditions. However, to our best knowledge, probabilistic tools such as the matrix Bernstein's inequality with Markov Chain samples are not well-established yet. Therefore, we do not intend to discuss the non-i.i.d. case in this paper and we leave the problem to future work.

*The code can be downloaded from `https://github.com/tensorflow/models/tree/master/tutorials/image/cifar10`. The tutorial can be found at `https://www.tensorflow.org/tutorials/deep_cnn`

References

Sung K Ahn and Gregory C Reinsel. Estimation of partially nonstationary vector autoregressive models with seasonal behavior. *Journal of Econometrics*, 62(2):317–350, 1994.

Theodore Wilbur Anderson. Estimating linear restrictions on regression coefficients for multivariate normal distributions. *The Annals of Mathematical Statistics*, pages 327–351, 1951.

Alexandre Belloni, Daniel Chen, Victor Chernozhukov, and Christian Hansen. Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, 80(6):2369–2429, 2012.

Alexandre Belloni, Victor Chernozhukov, Ivan Fernández-Val, and Christian Hansen. Program evaluation and causal inference with high-dimensional data. *Econometrica*, 85(1):233–298, 2017.

Sonia A Bhaskar and Adel Javanmard. 1-bit matrix completion under exact low-rank constraint. In *Information Sciences and Systems (CISS), 2015 49th Annual Conference on*, pages 1–6. IEEE, 2015.

Tony Cai and Wen-Xin Zhou. A max-norm constrained minimization approach to 1-bit matrix completion. *Journal of Machine Learning Research*, 14(1):3619–3647, 2013.

Emmanuel Candes and Terence Tao. The dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, pages 2313–2351, 2007.

Mehmet Caner and Qingliang Fan. Hybrid generalized empirical likelihood estimators: Instrument selection with adaptive lasso. *Journal of Econometrics*, 187(1):256–274, 2015.

Ngai Hang Chan, Chun Yip Yau, and Rong-Mao Zhang. Lasso estimation of threshold autoregressive models. *Journal of Econometrics*, 189(2):285–296, 2015.

Kun Chen, Hongbo Dong, and Kung-Sik Chan. Reduced rank regression via adaptive nuclear norm penalization. *Biometrika*, 100(4):901–920, 2013.

Scott Shaobing Chen, David L Donoho, and Michael A Saunders. Atomic decomposition by basis pursuit. *SIAM review*, 43(1):129–159, 2001.

Mark A Davenport, Yaniv Plan, Ewout van den Berg, and Mary Wootters. 1-bit matrix completion. *Information and Inference*, 3(3):189–223, 2014.

J. Fan, W. Wang, and Z. Zhu. A Shrinkage Principle for Heavy-Tailed Data: High-Dimensional Robust Low-Rank Matrix Recovery. *ArXiv e-prints*, March 2016.

Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456): 1348–1360, 2001.

Jianqing Fan and Jinchi Lv. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849–911, 2008.

Jianqing Fan and Jinchi Lv. Nonconcave penalized likelihood with np-dimensionality. *IEEE Transactions on Information Theory*, 57(8):5467–5484, 2011.

Jianqing Fan, Han Liu, Qiang Sun, and Tong Zhang. Tac for sparse learning: Simultaneous control of algorithmic complexity and statistical error. *arXiv preprint arXiv:1507.01037*, 2015.

Jianqing Fan, Bai Jiang, and Qiang Sun. Hoeffding's lemma for markov chains and its applications to statistical learning. *arXiv preprint arXiv:1802.00211*, 2018.

Ethan X Fang, Han Liu, Kim-Chuan Toh, and Wen-Xin Zhou. Max-norm optimization for robust matrix recovery. *Mathematical Programming*, pages 1–31, 2015.

John Geweke. Bayesian reduced rank regression in econometrics. *Journal of econometrics*, 75(1):121–146, 1996.

Ankit Gupta, Robert Nowak, and Benjamin Recht. Sample complexity for 1-bit compressed sensing and sparse classification. In *Information Theory Proceedings (ISIT), 2010 IEEE International Symposium on*, pages 1553–1557. IEEE, 2010.

Christian Hansen and Damian Kozbur. Instrumental variables estimation with many weak instruments using regularized jive. *Journal of Econometrics*, 182(2):290–308, 2014.

Wei Huang, Yoshiteru Nakamori, and Shou-Yang Wang. Forecasting stock market movement direction with support vector machine. *Computers & Operations Research*, 32(10):2513–2522, 2005.

Alan Julian Izenman. Reduced-rank regression for the multivariate linear model. *Journal of multivariate analysis*, 5(2):248–264, 1975a.

Alan Julian Izenman. Reduced-rank regression for the multivariate linear model. *Journal of multivariate analysis*, 5(2):248–264, 1975b.

Yakup Kara, Melek Acar Boyacioglu, and Ömer Kaan Baykan. Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the istanbul stock exchange. *Expert systems with Applications*, 38(5):5311–5319, 2011.

Frank Kleibergen and Richard Paap. Generalized reduced rank tests using the singular value decomposition. *Journal of econometrics*, 133(1):97–126, 2006.

Anders Bredahl Kock and Laurent Callot. Oracle inequalities for high dimensional vector autoregressions. *Journal of Econometrics*, 186(2):325–344, 2015.

Vladimir Koltchinskii, Karim Lounici, and Alexandre B Tsybakov. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics*, pages 2302–2329, 2011.

Eric L Lee, Jing-Kai Lou, Wei-Ming Chen, Yen-Chi Chen, Shou-De Lin, Yen-Sheng Chiang, and Kuan-Ta Chen. Fairness-aware loan recommendation for microfinance services. In *Proceedings of the 2014 International Conference on Social Computing*, page 3. ACM, 2014.

Pascal Lezaud. Chernoff-type bound for finite markov chains. *Annals of Applied Probability*, pages 849–867, 1998.

Sydney C Ludvigson and Serena Ng. Macro factors in bond risk premia. *The Review of Financial Studies*, 22(12):5027–5067, 2009.

Asger Lunde and Allan Timmermann. Duration dependence in stock prices: An analysis of bull and bear markets. *Journal of Business & Economic Statistics*, 22(3): 253–273, 2004.

Jacob Marshak. *Statistical inference in economics: an introduction.* John Wiley & Sons, 1950.

Sahand Negahban and Martin J Wainwright. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *The Annals of Statistics*, pages 1069–1097, 2011.

Sahand Negahban and Martin J Wainwright. Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *Journal of Machine Learning Research*, 13(1):1665–1697, 2012.

Sahand Negahban, Pradeep Ravikumar, Martin J Wainwright, and Bin Yu. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. In *Adv. Neural Inf. Proc. Sys.(NIPS)*. Citeseer, 2011.

Daniel Paulin. Concentration inequalities for markov chains by marton couplings and spectral methods. *Electronic Journal of Probability*, 20, 2015.

M Hashem Pesaran and Allan Timmermann. Market timing and return prediction under model instability. *Journal of Empirical Finance*, 9(5):495–510, 2002.

M Hashem Pesaran and Allan Timmermann. How costly is it to ignore breaks when forecasting the direction of a time series? *International Journal of Forecasting*, 20 (3):411–425, 2004.

Yaniv Plan and Roman Vershynin. One-bit compressed sensing by linear programming. *Communications on Pure and Applied Mathematics*, 66(8):1275–1297, 2013a.

Yaniv Plan and Roman Vershynin. Robust 1-bit compressed sensing and sparse logistic regression: A convex programming approach. *IEEE Transactions on Information Theory*, 59(1):482–494, 2013b.

Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Restricted eigenvalue properties for correlated gaussian designs. *Journal of Machine Learning Research*, 11(Aug): 2241–2259, 2010.

James H Stock and Mark W Watson. Forecasting using principal components from a large number of predictors. *Journal of the American statistical association*, 97(460): 1167–1179, 2002.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

Raja Velu and Gregory C Reinsel. *Multivariate reduced-rank regression: theory and applications*, volume 136. Springer Science & Business Media, 2013.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

Cun-Hui Zhang et al. Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics*, 38(2):894–942, 2010.

Hui Zou and Runze Li. One-step sparse estimates in nonconcave penalized likelihood models. *Annals of statistics*, 36(4):1509, 2008.

6 Proofs and Technical Lemmas

6.1 Proof for Theorem 1

We follow the proof scheme of Lemma B.4 in Fan et al. (2015). We first construct a middle point $\widehat{\mathbf{\Theta}}_\eta = \mathbf{\Theta}^* + \eta(\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*)$ such that we choose $\eta = 1$ when $\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*\|_F \leq \ell$ and $\eta = \ell / \|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*\|_F$ when $\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*\|_F > \ell$. For simplicity, we let $\widehat{\mathbf{\Delta}} = \widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*$ and $\widehat{\mathbf{\Delta}}_\eta = \widehat{\mathbf{\Theta}}_\eta - \mathbf{\Theta}^*$ in the remainder of the proof.

According to Negahban et al. (2012), when $\lambda \geq 2\|n^{-1}\sum\limits_{i=1}^{n}[b'(\langle \mathbf{X}_i, \mathbf{\Theta}^* \rangle) - Y_i] \cdot \mathbf{X}_i\|_{op}$, $\widehat{\mathbf{\Delta}}$ falls in the following cone:

$$\mathcal{C}(\mathcal{M}, \overline{\mathcal{M}}^\perp, \mathbf{\Theta}^*) := \Big\{ \left\| \mathbf{\Delta}_{\overline{\mathcal{M}}^\perp} \right\|_N \leq 3 \left\| \mathbf{\Delta}_{\overline{\mathcal{M}}} \right\|_N + 4 \sum_{j \geq r+1} \sigma_j(\mathbf{\Theta}^*) \Big\}.$$

Since $\widehat{\mathbf{\Delta}}_\eta$ is parallel to $\widehat{\mathbf{\Delta}}$, $\widehat{\mathbf{\Delta}}_\eta$ also falls in this cone. Given $\|\widehat{\mathbf{\Delta}}_\eta\|_N \leq \ell$ and LRSC($\mathcal{C}, \mathcal{N}$, κ_ℓ, τ_ℓ) of $\mathcal{L}_n(\mathbf{\Theta})$, we have

$$\kappa_\ell \|\widehat{\mathbf{\Delta}}_\eta\|_F^2 - \tau_\ell \leq \langle \nabla \mathcal{L}_n(\widehat{\mathbf{\Theta}}_\eta) - \nabla \mathcal{L}_n(\mathbf{\Theta}), \widehat{\mathbf{\Delta}}_\eta \rangle =: D_{\mathcal{L}}(\widehat{\mathbf{\Theta}}_\eta, \mathbf{\Theta}^*), \tag{6.1}$$

where $D_{\mathcal{L}}(\mathbf{\Theta}_1, \mathbf{\Theta}_2) = \mathcal{L}_n(\mathbf{\Theta}_1) - \mathcal{L}_n(\mathbf{\Theta}_2) - \langle \nabla \mathcal{L}_n(\mathbf{\Theta}_2), \mathbf{\Theta}_1 - \mathbf{\Theta}_2 \rangle$ is the symmetric Bregman divergence. By Lemma F.4 in Fan et al. (2015), $D_{\mathcal{L}}(\widehat{\mathbf{\Theta}}_\eta, \mathbf{\Theta}^*) \leq \eta \cdot D_{\mathcal{L}}(\widehat{\mathbf{\Theta}}, \mathbf{\Theta}^*)$. We thus have

$$\kappa_\ell \|\widehat{\mathbf{\Delta}}_\eta\|_F^2 - \tau_\ell \leq D_{\mathcal{L}}(\widehat{\mathbf{\Theta}}_\eta, \mathbf{\Theta}^*) \leq \eta D_{\mathcal{L}}(\widehat{\mathbf{\Theta}}, \mathbf{\Theta}^*) = \langle \nabla \mathcal{L}_n(\widehat{\mathbf{\Theta}}) - \nabla \mathcal{L}_n(\mathbf{\Theta}^*), \widehat{\mathbf{\Delta}}_\eta \rangle. \tag{6.2}$$

Since $\widehat{\mathbf{\Theta}}$ is the minimizer of the loss, we shall have the optimality condition $\nabla \mathcal{L}(\widehat{\mathbf{\Theta}}) + \lambda \boldsymbol{\xi} = \mathbf{0}$ for some subgradient $\boldsymbol{\xi}$ of the $\|\mathbf{\Theta}\|_N$ at $\mathbf{\Theta} = \widehat{\mathbf{\Theta}}$. Therefore, (6.2) simplifies to

$$\begin{aligned} &\kappa_\ell \|\widehat{\mathbf{\Delta}}_\eta\|_F^2 - \tau_\ell \leq -\langle \nabla \mathcal{L}(\mathbf{\Theta}^*) + \lambda \boldsymbol{\xi}, \widehat{\mathbf{\Delta}}_\eta \rangle \leq 1.5\lambda \|\widehat{\mathbf{\Delta}}_\eta\|_N \\ &\quad \leq 6\lambda\sqrt{2r} \left\| (\widehat{\mathbf{\Delta}}_\eta)_{\overline{\mathcal{M}}} \right\|_F + 6\lambda \sum_{j \geq r+1} \sigma_j(\mathbf{\Theta}^*) \leq 6\lambda\sqrt{2r} \left\| \widehat{\mathbf{\Delta}}_\eta \right\|_F + 6\lambda \sum_{j \geq r+1} \sigma_j(\mathbf{\Theta}^*). \end{aligned} \tag{6.3}$$

For a threshold $\tau > 0$, we choose $r = \#\{j \in \{1, 2, \ldots, d\} | \sigma_j(\mathbf{\Theta}^*) \geq \tau\}$. Then it follows that

$$\sum_{j \geq r+1} \sigma_j(\mathbf{\Theta}^*) \leq \tau \sum_{j \geq r+1} \frac{\sigma_j(\mathbf{\Theta}^*)}{\tau} \leq \tau \sum_{j \geq r+1} \big(\frac{\sigma_j(\mathbf{\Theta}^*)}{\tau} \big)^q \leq \tau^{1-q} \sum_{j \geq r+1} \sigma_j(\mathbf{\Theta}^*)^q \leq \tau^{1-q} \rho. \tag{6.4}$$

On the other hand, $\rho \geq \sum\limits_{j \leq r} \sigma_j(\mathbf{\Theta}^*)^q \geq r\tau^q$, so $r \leq \rho\tau^{-q}$. Choose $\tau = \lambda/\kappa_\ell$. Given (6.3), (6.4) and $\tau_\ell = C_0 \rho \lambda^{2-q} / \kappa_\ell^{1-q}$ yields that for some constant C_1, $\|\widehat{\mathbf{\Delta}}_\eta\|_F \leq$

$C_1\sqrt{\rho}(\lambda/\kappa_\ell)^{1-q/2}$. If we choose $\ell > C_1\sqrt{\rho}(\lambda/\kappa_\ell)^{1-q/2}$ in advance, we have $\mathbf{\Delta}_\eta = \mathbf{\Delta}$. Note that $\mathrm{rank}(\widehat{\mathbf{\Delta}}_{\overline{\mathcal{M}}}) \le 2r$; we thus have

$$\begin{aligned}\|\widehat{\mathbf{\Delta}}\|_N &\le \|(\widehat{\mathbf{\Delta}})_{\overline{\mathcal{M}}}\|_N + \|(\widehat{\mathbf{\Delta}})_{\overline{\mathcal{M}}^\perp}\|_N \le 4\|(\widehat{\mathbf{\Delta}})_{\overline{\mathcal{M}}}\|_N + 4\sum_{j\ge r+1}\sigma_j(\mathbf{\Theta}^*)\\ &\le 4\sqrt{2r}\|\widehat{\mathbf{\Delta}}\|_F + 4\sum_{j\ge r+1}\sigma_j(\mathbf{\Theta}^*) \le 4\sqrt{\rho}\tau^{-\frac{q}{2}}\|\mathbf{\Delta}\|_F + 4\rho\Big(\frac{\lambda}{\kappa_\ell}\Big)^{1-q} \qquad (6.5)\\ &\le (4C_1+4)\rho\Big(\frac{\lambda}{\kappa_\ell}\Big)^{1-q}.\end{aligned}$$

6.2 Proof for Lemma 1

Let $\eta_i = \langle \mathbf{\Theta}^*, \mathbf{X}_i\rangle$ and $\eta = \langle \mathbf{\Theta}^*, \mathbf{X}\rangle$.

$$\begin{aligned}&\Big\|\frac{1}{n}\sum_{i=1}^n (b'(\eta_i)-Y_i)\mathbf{X}_i\Big\|_{\mathrm{op}}\\ =&\Big\|\frac{1}{n}\sum_{i=1}^n (b'(\eta_i)-Y_i)\mathbf{X}_i - \mathbb{E}[(b'(\eta)-Y)\mathbf{X}] + \mathbb{E}[(b'(\eta)-Y)\mathbf{X}]\Big\|_{\mathrm{op}}\\ =&\Big\|\frac{1}{n}\sum_{i=1}^n (b'(\eta_i)-Y_i)\mathbf{X}_i - \mathbb{E}[(b'(\eta-Y)\mathbf{X})] + \mathbb{E}[b'(\eta)-Y]\cdot\mathbb{E}\mathbf{X}\Big\|_{\mathrm{op}} \qquad (6.6)\\ =&\Big\|\frac{1}{n}\sum_{i=1}^n (b'(\eta_i)-Y_i)\mathbf{X}_i - \mathbb{E}[(b'(\eta)-Y)\mathbf{X}]\Big\|_{\mathrm{op}}\end{aligned}$$

The last step is true because $EY = b'(\eta)$, which is proved in Chapter 2 in McCullagh and Nelder (1989). Now, we use the covering argument to bound the above operator norm.

Let $\mathcal{S}^{d-1} = \big\{\mathbf{u}\in\mathbb{R}^d : \|\mathbf{u}\|_2 = 1\big\}$, $\mathcal{N}^d$ be the 1/4 covering on $\mathcal{S}^{d-1}$ and $\Phi(\mathbf{A}) = \sup\limits_{\substack{\mathbf{u}\in\mathcal{N}^d\\ \mathbf{v}\in\mathcal{N}^d}} \mathbf{u}^T\mathbf{A}\mathbf{v}$ for $\forall\mathbf{A}\in\mathbb{R}^{d\times d}$.

We claim that

$$\|\mathbf{A}\|_{\mathrm{op}} \le \frac{16}{7}\Phi(\mathbf{A}). \qquad (6.7)$$

To establish the above inequality, we shall notice that since $\mathcal{N}^{d_1-1}$ is a 1/4 covering, for any given $\mathbf{u}\in\mathcal{S}^{d-1}, \mathbf{v}\in\mathcal{S}^{d-1}$, there is a $\tilde{\mathbf{u}}\in\mathcal{N}^d$ and $\tilde{\mathbf{v}}\in\mathcal{N}^d$ such that $\|\mathbf{u}-\tilde{\mathbf{u}}\|\le 1/4$ and $\|\mathbf{v}-\tilde{\mathbf{v}}\|\le 1/4$. Therefore,

$$\begin{aligned}\mathbf{u}^T\mathbf{A}\mathbf{v} =&\tilde{\mathbf{u}}^T\mathbf{A}\tilde{\mathbf{v}} + \tilde{\mathbf{u}}^T\mathbf{A}(\mathbf{v}-\tilde{\mathbf{v}}) + (\mathbf{u}-\tilde{\mathbf{u}})^T\mathbf{A}\tilde{\mathbf{v}} + (\mathbf{u}-\tilde{\mathbf{u}})\mathbf{A}(\mathbf{v}-\tilde{\mathbf{v}})\\ \le&\Phi(\mathbf{A}) + \frac{1}{4}\|\mathbf{A}\|_{\mathrm{op}} + \frac{1}{4}\|\mathbf{A}\|_{\mathrm{op}} + \frac{1}{16}\|\mathbf{A}\|_{\mathrm{op}}\end{aligned}$$

$$=\Phi(\mathbf{A}) + \frac{9}{16}\left\|\mathbf{A}\right\|_{\mathrm{op}}$$

Take the supremum over all possible $\mathbf{u} \in \mathcal{S}^{d-1}, \mathbf{v} \in \mathcal{S}^{d-1}$, we have

$$\left\|\mathbf{A}\right\|_{\mathrm{op}} = \sup_{\substack{\mathbf{u}\in\mathcal{S}^{d-1}\\ \mathbf{v}\in\mathcal{S}^{d-1}}} \mathbf{u}^T\mathbf{A}\mathbf{v} \le \Phi(\mathbf{A}) + \frac{9}{16}\left\|\mathbf{A}\right\|_{\mathrm{op}}$$

and this leads to (6.7).

In the remaining of this proof, for fixed $\mathbf{u} \in \mathcal{N}^d$ and $\mathbf{v} \in \mathcal{N}^d$, denote $\mathbf{u}^T\mathbf{X}_i\mathbf{v}$ by Z_i and $\mathbf{u}^T X\mathbf{v}$ by Z for convenience. According to the definition of sub-gaussian norm and sub-exponential norm, given the independence between the two terms, we have $\left\|\left[b'(\eta_i) - Y_i\right] Zi\right\|_{\Psi_1} \le \left\|b'(\eta_i) - Y_i\right\|_{\Psi_2}\left\|Z_i\right\|_{\Psi_2} \le M\kappa_0$. By Proposition 5.16 (Bernstein-type inequality) in Vershynin (2010), it follows that for sufficiently small t,

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}(b'(\eta_i) - Y_i)Z_i - \mathbb{E}[(b'(\eta) - Y_i)Z]\right| > t\right) \le 2\exp\left(-\frac{c_1 n t^2}{M^2\kappa_0^2}\right) \tag{6.8}$$

where c_1 is a positive constant. Here M is an upper bound for $\left\|b'(\eta_i) - Y_i\right\|_{\Psi_2}$. It is upper bounded since the variance of the response Y is bounded according to condition (C5).

Then the combination of the union bound over all points on $\mathcal{N}^d \times \mathcal{N}^d$ and (6.7) delivers

$$\mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^{n}(b'(\eta_i) - Y_i)Z_i - \mathbb{E}[(b'(\eta) - Y)Z]\right\|_{\mathrm{op}} > \frac{16}{7}t\right) \le 2\exp\left(d\log 8 - \frac{c_1 n t^2}{M^2\kappa_0^2}\right). \tag{6.9}$$

In conclusion, if we choose $t \asymp \sqrt{d/n}$, we can find a constant $\gamma > 0$ such that as long as $d/n < \gamma$, it holds that

$$P\left(\left\|\frac{1}{n}\sum_{i=1}^{n}(b'(\eta_i) - Y_i)\mathbf{X}_i\right\|_{\mathrm{op}} > \nu\sqrt{\frac{d}{n}}\right) \le c_1 \cdot e^{-c_2 d}. \tag{6.10}$$

where c_1 and c_2 are constants.

6.3 Proof for Lemma 2

In this proof, we will first show the RSC of $\mathcal{L}_n(\mathbf{\Theta})$ at $\mathbf{\Theta} = \mathbf{\Theta}^*$ over the cone

$$\mathcal{C}(\mathcal{M}_r, \overline{\mathcal{M}}_r^{\perp}, \mathbf{\Theta}^*) = \left\{\mathbf{\Delta} \in \mathbb{R}^{d\times d} : \left\|\mathbf{\Delta}_{\overline{\mathcal{M}}_r^{\perp}}\right\|_N \le 3\left\|\mathbf{\Delta}_{\overline{\mathcal{M}}_r}\right\|_N + 4\sum_{j\ge r+1}\sigma_j(\mathbf{\Theta}^*)\right\}$$

for some $1 \leq r \leq d$. Then, we will prove the LRSC of $\mathcal{L}_n(\mathbf{\Theta})$ in a nuclear-norm neighborhood of $\mathbf{\Theta}^*$ with respect to the same cone.

1. An important inequality that leads to RSC of $\mathcal{L}_n(\mathbf{\Theta})$ at $\mathbf{\Theta} = \mathbf{\Theta}^*$.

   We first prove that the following inequality holds for all $\mathbf{\Delta} \in \mathbb{R}^{d\times d}$ with probability greater than $1 - \exp(-c_1 d)$:

   $$\operatorname{vec}(\mathbf{\Delta})^T \cdot \widehat{\mathbf{H}}(\mathbf{\Theta}^*) \cdot \operatorname{vec}(\mathbf{\Delta}) \geq \kappa \cdot \|\mathbf{\Delta}\|_F^2 - C_0\sqrt{\frac{d}{n}}\,\|\mathbf{\Delta}\|_N^2 \,. \tag{6.11}$$

   Let $\mathbf{\Delta} = \mathbf{U}\mathbf{D}\mathbf{V}^T$ be the SVD of $\mathbf{\Delta}$. Then $\|\operatorname{vec}(\mathbf{D})\|_2 = \|\mathbf{\Delta}\|_F$ and $\|\operatorname{vec}(\mathbf{D})\|_1 = \|\mathbf{\Delta}\|_N$. It follows that

   $$\begin{aligned}
   &\operatorname{vec}(\mathbf{\Delta})^T \cdot \widehat{\mathbf{H}}(\mathbf{\Theta}^*) \cdot \operatorname{vec}(\mathbf{\Delta}) \\
   =&\frac{1}{n}\sum_{i=1}^n \operatorname{vec}(\mathbf{\Delta})^T \cdot b''(\langle \mathbf{\Theta}^*, \mathbf{X}_i\rangle) \cdot \operatorname{vec}(\mathbf{X}_i) \cdot \operatorname{vec}(\mathbf{X}_i)^T \cdot \operatorname{vec}(\mathbf{\Delta}) \\
   =&\frac{1}{n}\sum_{i=1}^n \operatorname{tr}(\sqrt{b''(\langle \mathbf{\Theta}^*, \mathbf{X}_i\rangle)}\mathbf{X}_i^T\mathbf{\Delta})^2 = \frac{1}{n}\sum_{i=1}^n \operatorname{tr}(\sqrt{b''(\langle \mathbf{\Theta}^*, \mathbf{X}_i\rangle)}\mathbf{X}_i^T\mathbf{U}\mathbf{D}\mathbf{V}^T)^2 \qquad (6.12)\\
   =&\frac{1}{n}\sum_{i=1}^n \operatorname{tr}(\sqrt{b''(\langle \mathbf{\Theta}^*, \mathbf{X}_i\rangle)}\mathbf{V}^T\mathbf{X}_i^T\mathbf{U}\mathbf{D})^2 = \frac{1}{n}\sum_{i=1}^n \operatorname{tr}(\tilde{\mathbf{X}}_i^T\mathbf{D})^2 \\
   =&\operatorname{vec}(\mathbf{D})^T \cdot \mathbf{\Sigma}_{\tilde{\mathbf{X}}\tilde{\mathbf{X}}} \cdot \operatorname{vec}(\mathbf{D}) + \operatorname{vec}(\mathbf{D})^T \cdot (\widehat{\mathbf{\Sigma}}_{\tilde{\mathbf{X}}\tilde{\mathbf{X}}} - \mathbf{\Sigma}_{\tilde{\mathbf{X}}\tilde{\mathbf{X}}}) \cdot \operatorname{vec}(\mathbf{D}).
   \end{aligned}$$

   Here, $\tilde{\mathbf{X}}_i = \sqrt{b''(\langle \mathbf{\Theta}^*, \mathbf{X}_i\rangle)}\mathbf{U}^T\mathbf{X}_i\mathbf{V}$, $\widehat{\mathbf{\Sigma}}_{\tilde{\mathbf{X}}\tilde{\mathbf{X}}} = n^{-1}\sum_{i=1}^n \operatorname{vec}(\tilde{\mathbf{X}}_i)\cdot\operatorname{vec}(\tilde{\mathbf{X}}_i)^T$ and $\mathbf{\Sigma}_{\tilde{\mathbf{X}}\tilde{\mathbf{X}}} = \mathbb{E}\widehat{\mathbf{\Sigma}}_{\tilde{\mathbf{X}}\tilde{\mathbf{X}}}$.

   To derive a lower bound for (6.12), we bound the first term from below and bound the second one from above.

   $$\begin{aligned}
   \lambda_{\min}(\mathbf{\Sigma}_{\tilde{\mathbf{X}}\tilde{\mathbf{X}}}) =& \inf_{\substack{\mathbf{W}_1,\mathbf{W}_2\in\mathbb{R}^{d\times d}\\ \|\mathbf{W}_1\|_F=\|\mathbf{W}_2\|_F=1}} \operatorname{vec}(\mathbf{W}_1)^T \cdot \mathbf{\Sigma}_{\tilde{\mathbf{X}}\tilde{\mathbf{X}}} \cdot \operatorname{vec}(\mathbf{W}_2) \\
   =& \inf_{\substack{\mathbf{W}_1,\mathbf{W}_2\in\mathbb{R}^{d\times d}\\ \|\mathbf{W}_1\|_F=\|\mathbf{W}_2\|_F=1}} \mathbb{E}\left[b''(\langle \mathbf{\Theta}^*, \mathbf{X}_i\rangle) \cdot \operatorname{tr}(\mathbf{W}_1^T\mathbf{U}^T\mathbf{X}_i\mathbf{V}) \cdot \operatorname{tr}(\mathbf{W}_2^T\mathbf{U}^T\mathbf{X}_i\mathbf{V})\right] \\
   =& \inf_{\substack{\mathbf{W}_1,\mathbf{W}_2\in\mathbb{R}^{d\times d}\\ \|\mathbf{W}_1\|_F=\|\mathbf{W}_2\|_F=1}} \mathbb{E}\left[b''(\langle \mathbf{\Theta}^*, \mathbf{X}_i\rangle) \cdot \operatorname{tr}(\mathbf{V}\mathbf{W}_1^T\mathbf{U}^T\mathbf{X}_i) \cdot \operatorname{tr}(\mathbf{V}\mathbf{W}_2^T\mathbf{U}^T\mathbf{X}_i)\right] \qquad (6.13)\\
   =& \inf_{\substack{\mathbf{W}_1,\mathbf{W}_2\in\mathbb{R}^{d\times d}\\ \|\mathbf{W}_1\|_F=\|\mathbf{W}_2\|_F=1}} \operatorname{vec}(\mathbf{U}\mathbf{W}_1\mathbf{V}) \cdot \mathbf{H}(\mathbf{\Theta}^*) \cdot \operatorname{vec}(\mathbf{U}\mathbf{W}_2\mathbf{V}) \\
   =&\lambda_{\min}(\mathbf{H}(\mathbf{\Theta}^*)) = \kappa
   \end{aligned}$$

Hence,

$$\operatorname{vec}(\boldsymbol{\Delta})^T \cdot \widehat{\boldsymbol{\Sigma}}_{\mathbf{X}\mathbf{X}} \cdot \operatorname{vec}(\boldsymbol{\Delta}) \geq \kappa \left\|\boldsymbol{\Delta}\right\|_F^2 - \left\|\widehat{\boldsymbol{\Sigma}}_{\tilde{\mathbf{X}}\tilde{\mathbf{X}}} - \boldsymbol{\Sigma}_{\tilde{\mathbf{X}}\tilde{\mathbf{X}}}\right\|_\infty \left\|\boldsymbol{\Delta}\right\|_N^2. \tag{6.14}$$

Meanwhile, for some appropriate constants c_3, c_4 and C_1, we establish the following inequality, which serves as the key step to bound $\|\widehat{\boldsymbol{\Sigma}}_{\tilde{\mathbf{X}}\tilde{\mathbf{X}}} - \boldsymbol{\Sigma}_{\tilde{\mathbf{X}}\tilde{\mathbf{X}}}\|_\infty$.

$$\mathbb{P}\left(\Big|\sup_{\substack{\mathbf{u}_1,\mathbf{u}_2\in\mathcal{S}^d\\ \mathbf{v}_1,\mathbf{v}_2\in\mathcal{S}^d}} \operatorname{vec}(\mathbf{u}_1\mathbf{v}_1^T)^T(\widehat{\boldsymbol{\Sigma}}_{\tilde{\mathbf{X}}\tilde{\mathbf{X}}} - \boldsymbol{\Sigma}_{\tilde{\mathbf{X}}\tilde{\mathbf{X}}})\operatorname{vec}(\mathbf{u}_2\mathbf{v}_2^T)\Big| > C_1\sqrt{\frac{d}{n}}\right) \leq c_3\exp(-c_4 d). \tag{6.15}$$

We apply the covering argument to prove the claim above. Denote the $1/8-$net of $\mathcal{S}^d$ by $\mathcal{N}^d$. For any $\mathbf{A} \in \mathbb{R}^{d^2\times d^2}$, define

$$\Phi(\mathbf{A}) := \sup_{\substack{\mathbf{u}_1,\mathbf{u}_2\in\mathcal{S}^d\\ \mathbf{v}_1,\mathbf{v}_2\in\mathcal{S}^d}} \operatorname{vec}(\mathbf{u}_1\mathbf{v}_1^T)^T\mathbf{A}\operatorname{vec}(\mathbf{u}_2\mathbf{v}_2^T)$$

and

$$\Phi_{\mathcal{N}}(\mathbf{A}) := \sup_{\substack{\mathbf{u}_1,\mathbf{u}_2\in\mathcal{N}^d\\ \mathbf{v}_1,\mathbf{v}_2\in\mathcal{N}^d}} \operatorname{vec}(\mathbf{u}_1\mathbf{v}_1^T)^T\mathbf{A}\operatorname{vec}(\mathbf{u}_2\mathbf{v}_2^T).$$

Note that for any $\mathbf{u}_1, \mathbf{v}_1, \mathbf{u}_2, \mathbf{v}_2 \in \mathcal{S}^d$, there exist $\overline{\mathbf{u}}_1, \overline{\mathbf{v}}_1, \overline{\mathbf{u}}_2, \overline{\mathbf{v}}_2 \in \mathcal{N}^d$ such that $\|\mathbf{u}_i - \overline{\mathbf{u}}_i\|_2 \leq 1/8$ and $\|\mathbf{v}_i - \overline{\mathbf{v}}_i\|_2 \leq 1/8$ for $i = 1, 2$. Then it follows that

$$\begin{aligned}
&\operatorname{vec}(\mathbf{u}_1\mathbf{v}_1^T)^T\mathbf{A}\operatorname{vec}(\mathbf{u}_2\mathbf{v}_2^T)\\
&= \operatorname{vec}(\overline{\mathbf{u}}_1\overline{\mathbf{v}}_1^T)^T\mathbf{A}\operatorname{vec}(\overline{\mathbf{u}}_2\overline{\mathbf{v}}_2^T) + \operatorname{vec}(\mathbf{u}_1(\mathbf{v}_1 - \overline{\mathbf{v}}_1)^T)^T\mathbf{A}\operatorname{vec}(\overline{\mathbf{u}}_2\overline{\mathbf{v}}_2^T) + \operatorname{vec}((\mathbf{u}_1 - \overline{\mathbf{u}}_1)\overline{\mathbf{v}}_1^T)^T\\
&\mathbf{A}\operatorname{vec}(\overline{\mathbf{u}}_2\overline{\mathbf{v}}_2^T) + \operatorname{vec}(\mathbf{u}_1\mathbf{v}_1^T)^T\mathbf{A}\operatorname{vec}(\mathbf{u}_2(\mathbf{v}_2 - \overline{\mathbf{v}}_2)^T) + \operatorname{vec}(\mathbf{u}_1\mathbf{v}_1^T)^T\mathbf{A}\operatorname{vec}((\mathbf{u}_2 - \overline{\mathbf{u}}_2)\overline{\mathbf{v}}_2^T)\\
&+ \operatorname{vec}((\mathbf{u}_1 - \overline{\mathbf{u}}_1)\mathbf{v}_1^T)^T\mathbf{A}\operatorname{vec}((\mathbf{u}_2 - \overline{\mathbf{u}}_2)\overline{\mathbf{v}}_2^T) + \operatorname{vec}(\mathbf{u}_1(\mathbf{v}_1 - \overline{\mathbf{v}}_1)^T)^T\mathbf{A}\operatorname{vec}((\mathbf{u}_2 - \overline{\mathbf{u}}_2)\overline{\mathbf{v}}_2^T)\\
&+ \operatorname{vec}((\mathbf{u}_1 - \overline{\mathbf{u}}_1)\mathbf{v}_1^T)^T\mathbf{A}\operatorname{vec}(\mathbf{u}_2(\mathbf{v}_2 - \overline{\mathbf{v}}_2)^T) + \operatorname{vec}(\mathbf{u}_1(\mathbf{v}_1 - \overline{\mathbf{v}}_1)^T)^T\mathbf{A}\operatorname{vec}(\mathbf{u}_2(\mathbf{v}_2 - \overline{\mathbf{v}}_2)^T)\\
&\leq \Phi_{\mathcal{N}}(\mathbf{A}) + \frac{1}{2}\Phi(\mathbf{A}) + \frac{1}{16}\Phi(\mathbf{A}).
\end{aligned} \tag{6.16}$$

So we have $\Phi(\mathbf{A}) \leq (16/7)\Phi_{\mathcal{N}}(\mathbf{A})$. For any $\mathbf{u}_1, \mathbf{u}_2 \in \mathcal{S}^d$ and $\mathbf{v}_1, \mathbf{v}_2 \in \mathcal{S}^d$, we know from Lemma 5.14 in Vershynin (2010) that

$$\begin{aligned}
&\|\langle \mathbf{u}_1\mathbf{v}_1', \tilde{\mathbf{X}}_i\rangle\langle \mathbf{u}_2\mathbf{v}_2', \tilde{\mathbf{X}}_i\rangle\|_{\Psi_1} \le \frac{1}{2}(\|\langle \mathbf{u}_1\mathbf{v}_1', \tilde{\mathbf{X}}_i\rangle^2\|_{\Psi_1} + \|\langle \mathbf{u}_2\mathbf{v}_2', \tilde{\mathbf{X}}_i\rangle^2\|_{\Psi_1}) \\
&\le \left\|\langle \mathbf{u}_1\mathbf{v}_1^T, \sqrt{b''(\langle \mathbf{\Theta}^*, \mathbf{X}_i\rangle)}\mathbf{U}^T\mathbf{X}_i\mathbf{V}\rangle\right\|_{\Psi_2}^2 + \left\|\langle \mathbf{u}_2\mathbf{v}_2^T, \sqrt{b''(\langle \mathbf{\Theta}^*, \mathbf{X}_i\rangle)}\mathbf{U}^T\mathbf{X}_i\mathbf{V}\rangle\right\|_{\Psi_2}^2 \\
&\le 2M\kappa_0^2.
\end{aligned} \tag{6.17}$$

Applying Bernstein Inequality yields

$$\mathbb{P}\left(\left|\mathrm{vec}(\mathbf{u}_1\mathbf{v}_1^T)^T(\widehat{\mathbf{\Sigma}}_{\tilde{\mathbf{X}}\tilde{\mathbf{X}}} - \mathbf{\Sigma}_{\tilde{\mathbf{X}}\tilde{\mathbf{X}}})\mathrm{vec}(\mathbf{u}_2\mathbf{v}_2^T)\right| > t\right) \le 2\exp\left(-c\min\Big(\frac{nt^2}{M^2\kappa_0^4}, \frac{nt}{M\kappa_0^2}\Big)\right).$$

Finally, by the union bound over $(\mathbf{u}_1, \mathbf{u}_2, \mathbf{v}_1, \mathbf{v}_2) \in \mathcal{N}^d \times \mathcal{N}^d \times \mathcal{N}^d \times \mathcal{N}^d$, we have

$$\begin{aligned}
&\mathbb{P}\left(\Big|\sup_{\substack{\mathbf{u}_1,\mathbf{u}_2\in\mathcal{S}^d \\ \mathbf{v}_1,\mathbf{v}_2\in\mathcal{S}^d}} \mathrm{vec}(\mathbf{u}_1\mathbf{v}_1^T)^T(\widehat{\mathbf{\Sigma}}_{\tilde{\mathbf{X}}\tilde{\mathbf{X}}} - \mathbf{\Sigma}_{\tilde{\mathbf{X}}\tilde{\mathbf{X}}})\mathrm{vec}(\mathbf{u}_2\mathbf{v}_2^T)\Big| > t\right) \\
&\le \exp\left(2d\log 8 - c\min\Big(\frac{nt^2}{M^2\kappa_0^4}, \frac{nt}{M\kappa_0^2}\Big)\right).
\end{aligned} \tag{6.18}$$

Take $t \asymp \sqrt{d/n}$, we derive the inequality (6.15). By combining (6.14) and (6.15), we successfully prove (6.11).

2. RSC at $\mathcal{L}_n(\mathbf{\Theta}^*)$ over $\mathcal{C}(\mathcal{M}_r, \overline{\mathcal{M}}_r^\perp, \mathbf{\Theta}^*)$

For all

$$\mathbf{\Delta} \in \mathcal{C}(\mathcal{M}_r, \overline{\mathcal{M}}_r^\perp, \mathbf{\Theta}^*) = \left\{\mathbf{\Delta} \in \mathbb{R}^{d\times d} : \left\|\mathbf{\Delta}_{\overline{\mathcal{M}}_r^\perp}\right\|_N \le 3\left\|\mathbf{\Delta}_{\overline{\mathcal{M}}_r}\right\|_N + 4\sum_{j\ge r+1}\sigma_j(\mathbf{\Theta}^*)\right\},$$

where $1 \le r \le d$, we have

$$\begin{aligned}
\|\mathbf{\Delta}\|_N &\le \left\|\mathbf{\Delta}_{\overline{\mathcal{M}}_r}\right\|_N + \left\|\mathbf{\Delta}_{\overline{\mathcal{M}}_r^\perp}\right\|_N \le 4\left\|\mathbf{\Delta}_{\overline{\mathcal{M}}_r}\right\|_N + 4\sum_{j\ge r+1}\sigma_j(\mathbf{\Theta}^*) \\
&\le 4\sqrt{2r}\,\|\mathbf{\Delta}\|_F + 4\sum_{j\ge r+1}\sigma_j(\mathbf{\Theta}^*).
\end{aligned} \tag{6.19}$$

Let $\tilde{\kappa} = (1/8)\kappa$. As we did in the proof for Theorem 1, we take $\tau = \lambda/\tilde{\kappa}$ and let

$r = \#\{j \in \{1, 2, ..., d\} | \sigma_j(\boldsymbol{\Theta}^*) > \tau\}$. Then,

$$\sum_{j \geq r+1} \sigma_j(\boldsymbol{\Theta}^*) = \tau \cdot \sum_{j \geq r+1} \frac{\sigma_j(\boldsymbol{\Theta}^*)}{\tau} \leq \tau \cdot \sum_{j \geq r+1} \frac{\sigma_j(\boldsymbol{\Theta}^*)}{\tau}^q \leq \tau^{1-q}\rho = \lambda^{1-q}\tilde{\kappa}^{q-1}\rho \tag{6.20}$$

On the other hand, $\rho > \sum_{j \leq r} \sigma(\boldsymbol{\Theta}^*)^q \geq r\tau^q$ so that $r \leq \rho\tau^{-q} = \rho\tilde{\kappa}^q\lambda^{-q}$. Plugging these results into (6.19), we have

$$\|\boldsymbol{\Delta}\|_N \leq 4\sqrt{2\rho}\lambda^{-q/2}\tilde{\kappa}^{q/2} \left\|\boldsymbol{\Delta}\right\|_F + 4\lambda^{1-q}\tilde{\kappa}^{q-1}\rho. \tag{6.21}$$

Since $\lambda = 2\nu\sqrt{d/n}$, there exist constants c_5 and c_6 such that as long as $\rho(d/n)^{(1-q)/2}$ $\leq c_4$, combining (6.14) and (6.21) we have

$$\text{vec}(\boldsymbol{\Delta})^T\widehat{\mathbf{H}}(\boldsymbol{\Theta}^*)\text{vec}(\boldsymbol{\Delta}) \geq \tilde{\kappa} \left\|\boldsymbol{\Delta}\right\|_F^2 - c_5\rho\lambda^{2-q}. \tag{6.22}$$

with high probability.

In the first two parts of this proof, we not only verify the RSC of $\mathcal{L}_n(\boldsymbol{\Theta}^*)$, but also provide the complete procedure of how to verify the RSC of the empirical loss given the RSC of the population loss. This is very important in Part 3 of this proof.

3. LRSC of $\mathcal{L}_n(\boldsymbol{\Theta})$ around $\boldsymbol{\Theta}^*$

In the remaining proof, we verify the LRSC by showing that there exists a positive constant $\tilde{\kappa}'$ such that

$$\text{vec}(\widehat{\boldsymbol{\Delta}})^T\widehat{\mathbf{H}}(\boldsymbol{\Theta})\text{vec}(\widehat{\boldsymbol{\Delta}}) \geq \tilde{\kappa}' \left\|\widehat{\boldsymbol{\Delta}}\right\|_F^2 - c_6\rho\lambda^{2-q}. \tag{6.23}$$

holds for all $\widehat{\boldsymbol{\Delta}} \in \mathcal{C}(\mathcal{M}_r, \overline{\mathcal{M}}_r^{\perp}, \boldsymbol{\Theta}^*)$ and $\boldsymbol{\Theta}$ such that $\|\boldsymbol{\Theta} - \boldsymbol{\Theta}^*\|_F \leq c_7\sqrt{\rho}\lambda^{(1-q)/2}$ for some positive constant c_7. Note that given $\boldsymbol{\Theta} - \boldsymbol{\Theta}^* \in \mathcal{C}(\mathcal{M}, \overline{\mathcal{M}}^{\perp}, \boldsymbol{\Theta}^*)$, by (6.21) we have $\|\boldsymbol{\Theta} - \boldsymbol{\Theta}^*\|_N \leq c_8\rho\lambda^{1-q} =: \ell$ for some constant c_8.

Define functions

$$\widehat{\mathbf{h}}(\boldsymbol{\Theta}) := n^{-1}\sum_{i=1}^n b''(\langle\boldsymbol{\Theta}, \mathbf{X}_i\rangle) \cdot \mathbb{1}_{\{|\langle\boldsymbol{\Theta}^*, \mathbf{X}_i\rangle| > \tau\|\mathbf{X}_i\|_{\text{op}} \geq \tau\gamma\}} \cdot \text{vec}(\mathbf{X}_i)\text{vec}(\mathbf{X}_i)^T$$

and

$$\mathbf{h}(\boldsymbol{\Theta}) := \mathbb{E}(\widehat{\mathbf{h}}(\boldsymbol{\Theta}))$$

for constants τ and γ to be determined. Recall that $\widehat{\mathbf{H}}(\boldsymbol{\Theta}^*) = n^{-1}\sum\limits_{i=1}^{n} b''(\langle\boldsymbol{\Theta}^*, \mathbf{X}_i\rangle)$ $\text{vec}(\mathbf{X}_i)\text{vec}(\mathbf{X}_i)^T$. The only difference between $\mathbf{h}(\cdot)$ and $\mathbf{H}(\cdot)$ is the indicator function so that $\widehat{\mathbf{H}}(\cdot) \succeq \widehat{\mathbf{h}}(\cdot)$.

We will finish the proof of LRSC in two steps. Firstly, we show that $\mathbf{h}(\mathbf{\Theta}^*)$ is positive definite over the restricted cone. Then by following the procedure of showing (6.22), we can prove that $\widehat{\mathbf{h}}(\mathbf{\Theta}^*)$ is positive definite over the cone with high probability. Secondly, we bound the difference between $\mathrm{vec}(\widehat{\mathbf{\Delta}})^T\widehat{\mathbf{h}}(\mathbf{\Theta})\mathrm{vec}(\widehat{\mathbf{\Delta}})$ and $\mathrm{vec}(\widehat{\mathbf{\Delta}})^T\widehat{\mathbf{h}}(\mathbf{\Theta}^*)\mathrm{vec}(\widehat{\mathbf{\Delta}})$ and show that $\widehat{\mathbf{h}}(\mathbf{\Theta})$ is locally positive definite around $\mathbf{\Theta}^*$. This naturally lead to the LRSC of $\mathcal{L}_n(\mathbf{\Theta})$ around $\mathbf{\Theta}^*$.

We establish the following lemma before proceeding.

Lemma 7. *When $\|\mathbf{\Theta}^*\|_F \geq \alpha\sqrt{d}$ and $\{vec(\mathbf{X}_i)\}_{i=1}^n$ are sub-Gaussian, there exist universal constants $\tau > 0$ and $\gamma > 0$ such that $\lambda_{min}(\mathbf{h}(\mathbf{\Theta}^*)) \geq \kappa_1$ where κ_1 is a positive constant.*

We select appropriate τ and γ to make $\mathbf{h}(\mathbf{\Theta}^*)$ positive definite. Follow the same procedure in Part 1 and Part 2 of this proof, we derive that

$$\mathrm{vec}(\widehat{\mathbf{\Delta}})^T \cdot \widehat{\mathbf{h}}_1(\mathbf{\Theta}) \cdot \mathrm{vec}(\widehat{\mathbf{\Delta}}) \geq \tilde{\kappa_1} \left\|\widehat{\mathbf{\Delta}}\right\|_F^2 - c_6\rho\lambda^{2-q}. \tag{6.24}$$

for a positive $\tilde{\kappa_1}$ with high probability.

Meanwhile,

$$\begin{aligned}
&\left|\mathrm{vec}(\widehat{\mathbf{\Delta}})^T \cdot \widehat{\mathbf{h}}(\mathbf{\Theta}^*) \cdot \mathrm{vec}(\widehat{\mathbf{\Delta}}) - \mathrm{vec}(\widehat{\mathbf{\Delta}})^T \cdot \widehat{\mathbf{h}}(\mathbf{\Theta}) \cdot \mathrm{vec}(\widehat{\mathbf{\Delta}})\right| \\
\leq &\cdot \frac{1}{n}\sum_{i=1}^n \left|b''(\langle\mathbf{\Theta}^*, \mathbf{X}_i\rangle) - b''(\langle\mathbf{\Theta}, \mathbf{X}_i\rangle)\right| \cdot \mathbb{1}_{\{|\langle\mathbf{\Theta}^*,\mathbf{X}_i\rangle| > \tau\|\mathbf{X}_i\|_{\mathrm{op}} \geq \tau\gamma\}} \cdot (\mathrm{vec}(\mathbf{X}_i)^T\mathrm{vec}(\widehat{\mathbf{\Delta}}))^2 \\
= &\frac{1}{n}\sum_{i=1}^n \left|b'''(\langle\tilde{\mathbf{\Theta}}, \mathbf{X}_i\rangle)\langle\mathbf{\Theta} - \mathbf{\Theta}^*, \mathbf{X}_i\rangle\right| \cdot \mathbb{1}_{\{|\langle\mathbf{\Theta}^*,\mathbf{X}_i\rangle| > \tau\|\mathbf{X}_i\|_{\mathrm{op}} \geq \tau\gamma\}}(\mathrm{vec}(\mathbf{X}_i)^T\mathrm{vec}(\widehat{\mathbf{\Delta}}))^2
\end{aligned} \tag{6.25}$$

Here $\tilde{\mathbf{\Theta}}$ is a middle point between $\mathbf{\Theta}^*$ and $\mathbf{\Theta}$, thus it is also in the nuclear ball centered at $\mathbf{\Theta}^*$ with radius ℓ. We know that $\left|\langle\tilde{\mathbf{\Theta}}, \mathbf{X}_i\rangle\right| \geq |\langle\mathbf{\Theta}^*, \mathbf{X}_i\rangle| - \left|\langle\mathbf{\Theta}^* - \tilde{\mathbf{\Theta}}, \mathbf{X}_i\rangle\right| \geq (\tau - \ell)\|\mathbf{X}_i\|_{\mathrm{op}}$ when the indicator function equals to 1. If $(\tau - \ell)\|\mathbf{X}_i\|_{\mathrm{op}} > 1$,

$$\left|b'''(\langle\tilde{\mathbf{\Theta}}, \mathbf{X}_i\rangle)\langle\mathbf{\Theta} - \mathbf{\Theta}^*, \mathbf{X}_i\rangle\right| \leq \frac{1}{(\tau - \ell)\|\mathbf{X}_i\|_{\mathrm{op}}}\|\mathbf{X}_i\|_{\mathrm{op}}\|\mathbf{\Theta} - \mathbf{\Theta}^*\|_N \leq \frac{\ell}{\tau - \ell}.$$

Otherwise, $\|\mathbf{X}_i\|_{\mathrm{op}}$ is bounded by $1/(\tau - \ell)$ and $\left|b'''(\langle\tilde{\mathbf{\Theta}}, \mathbf{X}_i\rangle)\langle\mathbf{\Theta} - \mathbf{\Theta}^*, \mathbf{X}_i\rangle\right| \leq C \cdot \frac{\ell}{\tau-\ell}$ where C is the upper bound of $b'''(x)$ for $|x| > (\tau - \ell)\|\mathbf{X}_i\|_{\mathrm{op}} > (\tau - \ell)\gamma$.

In summary,

$$(6.25) \leq \mathrm{vec}(\widehat{\mathbf{\Delta}})^T \cdot \frac{C\ell}{n(\tau-\ell)} \sum_{i=1}^{n} \mathrm{vec}(\mathbf{X}_i)\mathrm{vec}(\mathbf{X}_i)^T \cdot \mathrm{vec}(\widehat{\mathbf{\Delta}}) \tag{6.26}$$

Denote $\widehat{\mathbf{\Sigma}}_{\mathbf{XX}} = n^{-1} \sum_{i=1}^{n} \mathrm{vec}(\mathbf{X}_i)\mathrm{vec}(\mathbf{X}_i)^T$ and $\mathbf{\Sigma}_{\mathbf{XX}} = \mathbb{E}\widehat{\mathbf{\Sigma}}_{\mathbf{XX}}$. Suppose the eigenvalues of $\mathbf{\Sigma}_{\mathbf{XX}}$ is upper bounded by $K < \infty$, with a similar result to (6.11) and (6.21), as long as $\rho(d/n)^{1-q/2} \leq c_5$, we shall have

$$\begin{aligned} &\mathrm{vec}(\widehat{\mathbf{\Delta}})^T \cdot \frac{C\ell}{n(\tau-\ell)} \sum_{i=1}^{n} \mathrm{vec}(\mathbf{X}_i)\mathrm{vec}(\mathbf{X}_i)^T \cdot \mathrm{vec}(\widehat{\mathbf{\Delta}}) \\ \leq & \frac{C\ell}{(\tau-\ell)} \left(K \left\| \widehat{\mathbf{\Delta}} \right\|_F^2 + C\sqrt{\frac{d}{n}} \left\| \widehat{\mathbf{\Delta}} \right\|_N^2 \right) \\ \leq & \frac{2KC\ell}{\tau-\ell} \left\| \widehat{\mathbf{\Delta}} \right\|_F^2 \end{aligned} \tag{6.27}$$

As long as the constant ℓ is sufficiently small such that $2KC\ell/(\tau-\ell) < \tilde{\kappa_1}/2$, $\mathrm{vec}(\widehat{\mathbf{\Delta}})^T \cdot \widehat{\mathbf{h}}(\mathbf{\Theta}) \cdot \mathrm{vec}(\widehat{\mathbf{\Delta}}) \geq \tilde{\kappa_2} \|\widehat{\mathbf{\Delta}}\|_F^2$ holds with $\tilde{\kappa_2} = \tilde{\kappa_1}/2$. This delivers that $\widehat{\mathbf{h}}(\mathbf{\Theta})$ is locally positive definite around $\mathbf{\Theta}^*$. Recall that $\mathbf{H}(\cdot) \succeq \mathbf{h}(\cdot)$, we have verified that $\widehat{\mathbf{H}}(\mathbf{\Theta})$ is also locally positive definite around $\mathbf{\Theta}^*$. In summary, there exist some constant $\ell > 0$ such that for any $\|\mathbf{\Theta} - \mathbf{\Theta}^*\|_N \leq \ell$,

$$\mathrm{vec}(\widehat{\mathbf{\Delta}})^T \cdot \frac{1}{n} \sum_{i=1}^{n} b''(\langle \mathbf{\Theta}, \mathbf{X}_i \rangle)\mathrm{vec}(\mathbf{X}_i)\mathrm{vec}(\mathbf{X}_i)^T \cdot \mathrm{vec}(\widehat{\mathbf{\Delta}}) \geq \tilde{\kappa_2} \left\| \widehat{\mathbf{\Delta}} \right\|_F^2 - c_6 \rho \lambda^{2-q}. \tag{6.28}$$

for all $\widehat{\mathbf{\Delta}} \in \mathcal{C}(\mathcal{M}_r, \overline{\mathcal{M}}_r^{\perp}, \mathbf{\Theta}^*)$. This finalized our proof of the LRSC of $\mathcal{L}_n(\mathbf{\Theta})$ around $\mathbf{\Theta}^*$.

Below we provide the proof of Lemma 7.

Proof for Lemma 7

We first show that for any $p_0 \in (0,1)$, there exist constants τ and γ such that $\mathbb{P}(|\langle \mathbf{\Theta}^*, \mathbf{X}_i \rangle| > \tau \|\mathbf{X}_i\|_{\mathrm{op}} \geq \tau\gamma) \geq p_0$.

It is sufficient to show that $\mathbb{P}(|\langle \mathbf{\Theta}, \mathbf{X}_i \rangle| > \tau \|\mathbf{X}_i\|_{\mathrm{op}}) \geq (p_0+1)/2$ and $\mathbb{P}(\|\mathbf{X}_i\|_{\mathrm{op}} > \gamma) \geq (p_0+1)/2$ for some positive constants τ and γ. Then according to Bonferroni Inequality, $\mathbb{P}(|\langle \mathbf{\Theta}, \mathbf{X}_i \rangle| > \tau \|\mathbf{X}_i\|_{\mathrm{op}} > \tau\gamma) \geq (p_0+1)/2 + (p_0+1)/2 - 1 = p_0$.

The second inequality is easy to show.

$$\|\mathbf{X}_i\|_{\mathrm{op}} = \max_{\mathbf{u} \in \mathcal{S}^{d-1}, \mathbf{v} \in \mathcal{S}^{d-1}} \left| \mathbf{u}^T \mathbf{X}_i \mathbf{v} \right| \geq \|\mathbf{X}_i\|_{\infty} \tag{6.29}$$

Hence $\mathbb{P}(\|\mathbf{X}_i\|_{\text{op}} > \gamma) \geq \mathbb{P}(\|\mathbf{X}_i\|_\infty > \gamma)$. Since $\text{vec}(\mathbf{X}_i)$ is a sub-Gaussian vector with dimension d^2, $\mathbb{P}(\|\mathbf{X}_i\|_\infty > \gamma)$ monotonically goes to 1 as d grows. Let γ be sufficiently small so that $\mathbb{P}(\|\mathbf{X}_i\|_\infty > \gamma) = (p_0+1)/2$ when $\mathbf{X}_i \in \mathbb{R}^{2\times 2}$, it would be true that $\mathbb{P}(\|\mathbf{X}_i\|_\infty > \gamma) \geq (p_0+1)/2$ for all $d \geq 2$.

To prove the first inequality, we again divide it into two inequalities and combine them with Bonferroni Inequality. We would show that $\mathbb{P}(|\langle\mathbf{\Theta}, \mathbf{X}_i\rangle| > c_1\sqrt{d}) \geq (p_0+3)/4$ and $\mathbb{P}(\|\mathbf{X}_i\|_{\text{op}} \leq c_2\sqrt{d}) \geq (p_0+3)/4$ for some positive constants c_1 and c_2. Then

$$\mathbb{P}(|\langle\mathbf{\Theta}, \mathbf{X}_i\rangle| > c_1/c_2 \, \|\mathbf{X}_i\|_{\text{op}}) \geq (p_0+3)/4 + (p_0+3)/4 - 1 = (p_0+1)/2 \tag{6.30}$$

On one hand, $\langle\mathbf{\Theta}, \mathbf{X}_i\rangle$ is a sub-Gaussian variable since it is a linear transformation of a sub-Gaussian vector. Its mean is 0 and its sub-Gaussian norm is bounded by $\kappa_0 \|\mathbf{\Theta}\|_F$. Since $\|\mathbf{\Theta}\|_F \geq \alpha\sqrt{d}$, take c_1 to be sufficiently small, we have

$$\mathbb{P}(|\langle\mathbf{\Theta}, \mathbf{X}_i\rangle| > c_1\sqrt{d}) \geq \mathbb{P}(|x| > c_1/\alpha) \geq \frac{p_0+3}{4} \tag{6.31}$$

where x is a sub-Gaussian variable and $\|x\|_{\Psi_2} \leq \kappa_0$.

On the other hand,

$$\begin{aligned}\|\mathbf{X}_i\|_{\text{op}} &= \max_{\mathbf{u}\in\mathcal{S}^{d-1},\mathbf{v}\in\mathcal{S}^{d-1}} \left|\mathbf{u}^T\mathbf{X}_i\mathbf{v}\right| = \max_{\mathbf{u}\in\mathcal{S}^{d-1},\mathbf{v}\in\mathcal{S}^{d-1}} \left|\text{tr}(\mathbf{u}^T\mathbf{X}_i\mathbf{v})\right| \\ &= \max_{\mathbf{u}\in\mathcal{S}^{d-1},\mathbf{v}\in\mathcal{S}^{d-1}} \left|\text{tr}(\mathbf{v}\mathbf{u}^T\mathbf{X}_i)\right| = \max_{\mathbf{u}\in\mathcal{S}^{d-1},\mathbf{v}\in\mathcal{S}^{d-1}} \left|\langle\mathbf{u}\mathbf{v}^T, \mathbf{X}_i\rangle\right|.\end{aligned} \tag{6.32}$$

Recall the covering argument in the proof of Lemma 1. Denote $\mathcal{N}^d$ as a 1/4-net on $\mathcal{S}^{d-1}$, then

$$\max_{\mathbf{u}\in\mathcal{S}^{d-1},\mathbf{v}\in\mathcal{S}^{d-1}} \left|\langle\mathbf{u}\mathbf{v}^T, \mathbf{X}_i\rangle\right| \leq \frac{16}{7} \max_{\mathbf{u}\in\mathcal{N}^d,\mathbf{v}\in\mathcal{N}^d} \left|\langle\mathbf{u}\mathbf{v}^T, \mathbf{X}_i\rangle\right| \tag{6.33}$$

For any $\mathbf{u}_1 \in \mathcal{N}^d$, $\mathbf{v}_1 \in \mathcal{N}^d$, given $\|\mathbf{X}_i\|_{\Psi_2} \leq \kappa_0$, we have $\left\|\langle\mathbf{u}_1\mathbf{v}_1^T, \mathbf{X}_i\rangle\right\|_{\Psi_1} \leq \kappa_0$. According to Bernstein-type inequality in Vershynin (2010), it follows that for sufficiently small t and some positive constant C,

$$\mathbb{P}(\left|\langle\mathbf{u}_1\mathbf{v}_1^T, \mathbf{X}_i\rangle\right| > t) \leq 2\exp\left(-\frac{Ct^2}{\kappa_0^2}\right) \tag{6.34}$$

Therefore, the overall union bound follows:

$$\mathbb{P}(\max_{\mathbf{u}\in\mathcal{S}^{d-1},\mathbf{v}\in\mathcal{S}^{d-1}} \left|\langle \mathbf{u}\mathbf{v}^T, \mathbf{X}_i\rangle\right| > t) \leq 2\exp\left(2d\log 4 - \frac{Ct^2}{\kappa_0^2}\right) \tag{6.35}$$

Let $t = c_2\sqrt{d}$ for some positive constant $c_2 > \sqrt{4\log 4\kappa_0^2/C}$, the above probability decays. This means that with high probability (which is greater than $(p_0+3)/4$) $\|\mathbf{X}_i\|_{\text{op}}$ is less than $c_2\sqrt{d}$. This finalize our proof of (6.30).

Now we look at

$$\mathbf{h}(\mathbf{\Theta}) = n^{-1}\mathbb{E}\left[\sum_{i=1}^n b''(\langle\mathbf{\Theta}, \mathbf{X}_i\rangle)\cdot \mathbb{1}_{\{|\langle\mathbf{\Theta}^*,\mathbf{X}_i\rangle|>\tau\|\mathbf{X}_i\|_{\text{op}}\geq\tau\gamma\}}\cdot \text{vec}(\mathbf{X}_i)\text{vec}(\mathbf{X}_i)^T\right].$$

Denote $\{|\langle\mathbf{\Theta}^*,\mathbf{X}_i\rangle| > \tau\left\|\mathbf{X}_i\right\|_{\text{op}} \geq \tau\gamma\}$ as an event A_i with probability sufficiently close to 1. For any $\mathbf{v}\in\mathbb{R}^{d^2}$,

$$\begin{aligned}
n\mathbf{v}^T\mathbf{h}(\mathbf{\Theta}^*)\mathbf{v} =& \mathbb{E}\left[\sum_{i=1}^n b''(\langle\mathbf{\Theta}^*,\mathbf{X}_i\rangle)(\text{vec}(\mathbf{X}_i)^T\mathbf{v})^2\right] \\
&-\mathbb{E}\left[\sum_{i=1}^n b''(\langle\mathbf{\Theta}^*,\mathbf{X}_i\rangle)\cdot\mathbb{1}_{A_i^c}\cdot(\text{vec}(\mathbf{X}_i)^T\mathbf{v})^2\right] \\
\geq& n\kappa\left\|\mathbf{v}\right\|_2^2 - \sqrt{\mathbb{E}\left[\sum_{i=1}^n b''(\langle\mathbf{\Theta}^*,\mathbf{X}_i\rangle)^2\left(\text{vec}(\mathbf{X}_i)^T\mathbf{v}\right)^4\right]}\cdot\sqrt{\mathbb{E}\sum_{i=1}^n\mathbb{1}_{A_i^c}} \\
\geq& n\kappa\left\|\mathbf{v}\right\|_2^2 - nMK\sqrt{1-p_0}\left\|\mathbf{v}\right\|_2^2
\end{aligned} \tag{6.36}$$

Here, M is an global upper bound of $b''(\cdot)$ and K is the largest eigenvalue of the fourth moment of $\mathbf{X}_i$. Since $\mathbf{X}_i$ is sub-Gaussian, the fourth moment is bounded. We let $1-p_0$ be sufficiently small so that $nMK\sqrt{1-p_0}\leq\kappa/2$, then we proved that $\lambda_{\min}(\mathbf{h}(\mathbf{\Theta}^*)) \geq \kappa/2 > 0$ and thus $\mathbf{h}(\mathbf{\Theta}^*)$ is positive definite.

6.4 Proof of Lemma 3

$$\frac{1}{N}\sum_{i=1}^N(b'(\langle\mathbf{X}_i,\mathbf{\Theta}^*\rangle)-Y_i)\mathbf{X}_i = \frac{1}{n}\sum_{i=1}^n\frac{1}{d}\sum_{j=1}^d(b'(\boldsymbol{\theta}_j^{*T}\mathbf{x}_i)-y_{ij})\mathbf{x}_i\mathbf{e}_j^T = \frac{1}{d}\cdot\frac{1}{n}\sum_{i=1}^n\mathbf{x}_i\mathbf{z}_i^T,$$

where $\mathbf{z}_i$ satisfies that $z_{ij} = b'(\boldsymbol{\theta}_j^{*T}\mathbf{x}_i) - y_{ij}$. Note that given $\mathbf{x}_i$, $\|z_{ij}\|_{\Psi_2}\leq\phi M$. To see why, let $\eta_{ij} = \mathbf{x}_i^T \boldsymbol{\theta}_j^*$. We have

$$
\begin{aligned}
\mathrm{E}\exp(tz_{ij} \mid \mathbf{x}_i) &= \int_{y\in\mathcal{Y}} c(y)\exp\big(\frac{\eta_{ij}y - b(\eta_{ij})}{\phi}\big)\exp(t(y - b'(\eta_{ij})))dy \\
&= \int_{y\in\mathcal{Y}} c(y)\exp\big(\frac{(\eta_{ij}+\phi t)y - b(\eta_{ij}+\phi t) + b(\eta_{ij}+\phi t) - b(\eta_{ij}) - \phi t b'(\eta_{ij})}{\phi}\big)dy \\
&= \exp\big(\frac{b(\eta_{ij}+\phi t) - b(\eta_{ij}) - \phi t b'(\eta_{ij})}{\phi}\big) \le \exp\big(\frac{\phi M t^2}{2}\big).
\end{aligned}
$$

Besides, $y_{ij} \perp\!\!\!\perp y_{ik}$ for $j \neq k$ given $\mathbf{x}_i$. Therefore, $\|\mathbf{z}_i\|_{\Psi_2} \le \phi M$. Since $\mathrm{E}\,\mathbf{z}_i\mathbf{x}_i^T = \mathbf{0}$, by the standard covering argument (Theorem 5.39 and Remark 5.40 in Vershynin (2010)), there exists $\gamma > 0$ such that when $n > \gamma d$, it holds for some universal constant $c > 0$,

$$
\mathbb{P}\big(\|\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{z}_i^T\|_{op} \ge \sqrt{\frac{\phi M\kappa_0 d}{n}}\big) \le 2\exp(-cd).
$$

6.5 Proof of Lemma 4

$$
\begin{aligned}
\mathrm{vec}(\widehat{\boldsymbol{\Delta}})^T\widehat{\mathbf{H}}(\boldsymbol{\Theta}^*)\mathrm{vec}(\widehat{\boldsymbol{\Delta}}) &= \frac{1}{N}\sum_{i=1}^{N} b''(\langle\mathbf{X}_i,\boldsymbol{\Theta}^*\rangle)\langle\widehat{\boldsymbol{\Delta}},\mathbf{X}_i\rangle^2 = \frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{d} b''(\mathbf{x}_i^T\boldsymbol{\theta}_j^*)\langle\widehat{\boldsymbol{\Delta}},\mathbf{x}_j\mathbf{e}_i^T\rangle^2 \\
&= \frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{d} b''(\mathbf{x}_i^T\boldsymbol{\theta}_j^*)\,\mathrm{tr}(\mathbf{x}_i^T\widehat{\boldsymbol{\Delta}}\mathbf{e}_j)^2 = \frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{d} b''(\mathbf{x}_i^T\boldsymbol{\theta}_j^*)(\mathbf{x}_i^T\widehat{\boldsymbol{\Delta}}_j)^2.
\end{aligned}
\tag{6.37}
$$

Note that for any $1 \le j \le d$, $\|\sqrt{b''(\mathbf{x}_i^T\boldsymbol{\theta}_j)\mathbf{x}_i}\|_{\Psi_2} \le \sqrt{M}\kappa_0$. By Theorem 5.39 in Vershynin (2010), there exists some $\gamma > 0$ such that if $n > \gamma d$, we have for some universal constant $c > 0$,

$$
\mathbb{P}\Big(\|\frac{1}{n}\sum_{i=1}^{n} b''(\mathbf{x}_i^T\boldsymbol{\theta}_j^*)\mathbf{x}_i\mathbf{x}_i^T - \mathrm{E}(b''(\mathbf{x}_i^T\boldsymbol{\theta}_j^*)\mathbf{x}_i\mathbf{x}_i^T)\|_{op} \ge \kappa_0\sqrt{\frac{Md}{n}}\Big) \le 2\exp(-cd). \tag{6.38}
$$

Denote this event by $\mathcal{E}_0$. By the union bound, it holds that

$$
\mathbb{P}\Big(\max_{1\le j\le d}\|\frac{1}{n}\sum_{i=1}^{n} b''(\mathbf{x}_i^T\boldsymbol{\theta}_j^*)\mathbf{x}_i\mathbf{x}_i^T - \mathbf{H}(\boldsymbol{\Theta}^*)\|_{op} \ge \kappa_0\sqrt{\frac{Md}{n}}\Big) \le 2d\exp(-cd).
$$

In addition, for any $\boldsymbol{\Theta} \in \mathbb{R}^{d\times d}$ such that $\|\boldsymbol{\Theta} - \boldsymbol{\Theta}^*\|_F \le r$, $\|\boldsymbol{\theta}_j - \boldsymbol{\theta}_j^*\|_2 \le r$ holds for all $1 \le j \le d$. Given that $\|\mathbf{x}_i\|_{\Psi_2} \le \kappa_0$,

$$
\mathbb{P}(\max_{1\le i\le n, 1\le j\le d}|\mathbf{x}_i^T(\boldsymbol{\theta}_j - \boldsymbol{\theta}_j^*)| \ge t) \le 2nd\exp\Big(-\frac{t^2}{2\kappa_0^2 r^2}\Big).
$$

Substituting $t = \kappa_0 r\sqrt{\delta \log(nd)}$ into the inequality above, we have

$$\mathbb{P}(\max_{1\le i\le n, 1\le j\le d} |\mathbf{x}_i^T(\boldsymbol{\theta}_j - \boldsymbol{\theta}_j^*)| \ge \kappa_0 r\sqrt{\delta \log(nd)}) \le 2(nd)^{1-\frac{\delta}{2}}.$$

Denote the above event by $\mathcal{E}_1$. Therefore, under $\mathcal{E}_1^c$,

$$\begin{aligned} \|\frac{1}{n}\sum_{i=1}^n (b''(\mathbf{x}_i^T\boldsymbol{\theta}_j) - b''(\mathbf{x}_i^T\boldsymbol{\theta}_j^*))\mathbf{x}_i\mathbf{x}_i^T\|_{op} &\le L\|\frac{1}{n}\sum_{i=1}^n (\mathbf{x}_i^T(\boldsymbol{\theta}_j - \boldsymbol{\theta}_j^*))\mathbf{x}_i\mathbf{x}_i^T\|_{op} \\ &\le L\kappa_0 r\sqrt{\delta\log(nd)} \cdot \|\frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^T\|_{op}. \end{aligned} \tag{6.39}$$

Again by Theorem 5.39 in Vershynin (2010), when n/d is sufficiently large,

$$\mathbb{P}\Big(\|\frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^T - \boldsymbol{\Sigma}_{\mathbf{x}\mathbf{x}}\|_{op} \ge \kappa_0\sqrt{\frac{d}{n}}\Big) \le 2\exp(-cd).$$

Therefore, when n/d is sufficiently large, $\|n^{-1}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^T\|_{op} \le 2\kappa_0$ with high probability. Denote this event by $\mathcal{E}_2$. Combining this with (6.38) and (6.39), we have under $\mathcal{E}_1^c \cap \mathcal{E}_2^c$,

$$\|\frac{1}{n}\sum_{i=1}^n (b''(\mathbf{x}_i^T\boldsymbol{\theta}_j) - b''(\mathbf{x}_i^T\boldsymbol{\theta}_j^*))\mathbf{x}_i\mathbf{x}_i^T\|_{op} \le 2L\kappa_0^2 r\sqrt{\delta\log(nd)}.$$

Finally, for sufficiently large n/d, it holds with probability at least $1 - 2(nd)^{1-\frac{\delta}{2}}$ for all $\boldsymbol{\theta}$ such that $\|\boldsymbol{\Theta} - \boldsymbol{\Theta}^*\|_F \le r$,

$$\lambda_{\min}\Big(\frac{1}{n}\sum_{i=1}^n b''(\mathbf{x}_i^T\boldsymbol{\theta}_j)\mathbf{x}_i\mathbf{x}_i^T\Big) \ge \kappa_\ell - 2L\kappa_0^2 r\sqrt{\delta\log(nd)}.$$

By a union bound across $j = 1, \ldots, d$, we can deduce that for any $\delta > 4$, it holds with probability at least $1 - 2(nd)^{2-\frac{\delta}{2}}$ that for all $\boldsymbol{\Delta} \in \mathbb{R}^{d\times d}$ and all $\boldsymbol{\Theta} \in \mathcal{N}$,

$$\text{vec}(\boldsymbol{\Delta})^T\widehat{\mathbf{H}}(\boldsymbol{\Theta})\text{vec}(\boldsymbol{\Delta}) \ge \frac{1}{d}(\kappa_\ell - 2L\kappa_0^2 r\sqrt{\delta\log(nd)})\|\boldsymbol{\Delta}\|_F^2.$$

Since $r \asymp \sqrt{\rho}\lambda^{1-q/2}$, as long as $\rho(d/n)^{1-q/2}\log(nd)$ is sufficiently small, LRSC($\mathcal{C}$,$\mathcal{N}$, $(1/2)\kappa_\ell, 0$) holds.

6.6 Proof for Lemma 5

Here, we take advantage of the singleton design of X and apply the Matrix Bernstein inequality (Theorem 6.1.1 in Tropp(2015)) to bound the operator norm of the gradient of the loss function.

Denote $\mathbf{Z}_i = [\exp(\langle \mathbf{\Theta}^*, \mathbf{X}_i \rangle)/(1+\exp(\langle \mathbf{\Theta}^*, \mathbf{X}_i \rangle)) - Y_i] \cdot \mathbf{X}_i \in \mathbb{R}^{d \times d}$. $\forall \mathbf{u} \in \mathcal{S}^{d-1}, \mathbf{v} \in \mathcal{S}^{d-1}$,

$$\mathbf{u}^T \mathbf{Z}_i \mathbf{v} \le \left| \frac{e^{\langle \mathbf{\Theta}^*, X_i \rangle}}{e^{\langle \mathbf{\Theta}^*, X_i \rangle} + 1} - Y_i \right| \cdot d \le d.$$

Thus $\|\mathbf{Z}_i\|_{\text{op}} \le d$. Meanwhile,

$$\begin{aligned} \left\|\mathbb{E}\mathbf{Z}_i\mathbf{Z}_i^T\right\|_{\text{op}} =& \left\| \mathbb{E}\left[\left(\frac{e^{\langle \mathbf{\Theta}^*, \mathbf{X}_i \rangle}}{e^{\langle \mathbf{\Theta}^*, \mathbf{X}_i \rangle} + 1} - Y_i \right)^2 \mathbf{X}_i \mathbf{X}_i^T \right] \right\|_{\text{op}} \le \left\| \mathbb{E}\left[\mathbf{X}_i \mathbf{X}_i^T \right] \right\|_{\text{op}} \\ =& d^2 \cdot \left\| \mathbb{E}\left[\mathbf{e}_{a(i)} \mathbf{e}_{a(i)}^T \right] \right\|_{\text{op}} = d^2 \cdot \frac{1}{d} = d \end{aligned} \tag{6.40}$$

Similarly, we have $\left\|\mathbb{E}\mathbf{Z}_i^T\mathbf{Z}_i\right\|_{\text{op}} \le d$. Therefore, $\max\left\{ \left\|\mathbb{E}\mathbf{Z}_i\mathbf{Z}_i^T\right\|_{\text{op}}, \left\|\mathbb{E}\mathbf{Z}_i^T\mathbf{Z}_i\right\|_{\text{op}} \right\} \le d$.

According to Matrix Bernstein inequality,

$$P\left(\left\| \frac{1}{n} \sum_{i=1}^n \mathbf{Z}_i \right\|_{\text{op}} \ge t \right) \le 2d \cdot \exp\left(\frac{-nt^2/2}{d + dt/3}\right) \tag{6.41}$$

Let $t = \nu\sqrt{\delta d \log d / n}$, then

$$\begin{aligned} P\left(\left\| \frac{1}{n} \sum_{i=1}^n \mathbf{Z}_i \right\|_{\text{op}} \ge \nu\sqrt{\frac{\delta d \log d}{n}} \right) \le& 2d \cdot \exp\left(\frac{-\nu^2 \delta d \log d}{2d + 2\nu\sqrt{\frac{d^2 \delta d \log d}{n}}/3}\right) \\ =& 2d^{1 - \frac{\nu^2 \delta}{2 + 2\nu\sqrt{d \cdot \delta \cdot \log d}/3\sqrt{n}}} \\ \le& 2d^{1-\delta} \end{aligned} \tag{6.42}$$

for some constant ν as long as $d \log d / n \le \gamma$ for some constant γ.

6.7 Proof for Lemma 6

We aim to show that the loss function has LRSC property in a L_∞-ball centered at $\mathbf{\Theta}^*$ with radius $2R/d$.

For all $\tilde{\mathbf{\Theta}} \in \mathbb{R}^{d\times d}$ satisfying $\left\|\tilde{\mathbf{\Theta}} - \mathbf{\Theta}^*\right\|_\infty \leq 2R/d$, let us denote $f(\mathbf{\Theta}) = \exp\left(\langle \mathbf{\Theta}, \mathbf{X}_i\rangle\right) / (1+\exp\left(\langle \mathbf{\Theta}, \mathbf{X}_i\rangle\right))^2$. Then

$$\begin{aligned}
&\mathrm{vec}(\mathbf{\Delta})^T[\widehat{\mathbf{H}}(\tilde{\mathbf{\Theta}}) - \widehat{\mathbf{H}}(\mathbf{\Theta}^*)]\mathrm{vec}(\mathbf{\Delta}) \\
=&\mathrm{vec}(\mathbf{\Delta})^T \cdot \frac{1}{n}\sum_{i=1}^n \left[f\left(\langle \tilde{\mathbf{\Theta}}, \mathbf{X}_i\rangle\right) - f\left(\langle \mathbf{\Theta}^*, \mathbf{X}_i\rangle\right)\right]\mathrm{vec}(\mathbf{X}_i)\mathrm{vec}(\mathbf{X}_i)^T \cdot \mathrm{vec}(\mathbf{\Delta}) \\
\leq&\mathrm{vec}(\mathbf{\Delta})^T \cdot \frac{1}{n}\sum_{i=1}^n f'\left(\langle \bar{\mathbf{\Theta}}_i, \mathbf{X}_i\rangle\right)\langle \tilde{\mathbf{\Theta}} - \mathbf{\Theta}^*, \mathbf{X}_i\rangle \mathrm{vec}(\mathbf{X}_i)\mathrm{vec}(\mathbf{X}_i)^T \cdot \mathrm{vec}(\mathbf{\Delta})
\end{aligned} \tag{6.43}$$

Here $\bar{\mathbf{\Theta}}_i$ is a middle point between $\tilde{\mathbf{\Theta}}$ and $\mathbf{\Theta}^*$. Due to the singleton design of $\mathbf{X}_i$, $\langle \tilde{\mathbf{\Theta}} - \mathbf{\Theta}^*, \mathbf{X}_i\rangle \leq d \cdot \left\|\tilde{\mathbf{\Theta}} - \mathbf{\Theta}^*\right\|_\infty \leq 2R$. Given that the derivative of $f(\cdot)$ is bounded by 0.1, we have

$$\begin{aligned}
\mathrm{vec}(\mathbf{\Delta})^T[\widehat{\mathbf{H}}(\tilde{\mathbf{\Theta}}) - \widehat{\mathbf{H}}(\mathbf{\Theta}^*)]\mathrm{vec}(\mathbf{\Delta}) \leq& \frac{R}{5} \cdot \mathrm{vec}(\mathbf{\Delta})^T \cdot \frac{1}{n}\sum_{i=1}^n \mathrm{vec}(\mathbf{X}_i)\mathrm{vec}(\mathbf{X}_i)^T \cdot \mathrm{vec}(\mathbf{\Delta}) \\
=:& \frac{R}{5n}\left\|\tilde{\mathfrak{X}}_n(\mathbf{\Delta})\right\|_2^2
\end{aligned} \tag{6.44}$$

It is proved in the proof of Theorem 1 in Negahban and Wainwright (2012) that as long as $n > c_6 d\log d$,

$$\left|\frac{\left\|\tilde{\mathfrak{X}}_n(\mathbf{\Delta})\right\|_2}{\sqrt{n}} - \|\mathbf{\Delta}\|_F\right| \geq \frac{7}{8}\|\mathbf{\Delta}\|_F + \frac{16d\|\mathbf{\Delta}\|_\infty}{\sqrt{n}} \tag{6.45}$$

for all $\mathbf{\Delta} \in \mathcal{C}'(c_0)$ with probability at most $c_7\exp(-c_8 d\log d)$. Therefore, since $\mathbf{\Delta} \in \mathcal{C}'(c_0)$ and $128d\|\mathbf{\Delta}\|_\infty/\sqrt{n}\|\mathbf{\Delta}\|_F \leq 1/2$, we shall have

$$\frac{\left\|\tilde{\mathfrak{X}}_n(\mathbf{\Delta})\right\|_2}{\sqrt{n}} \leq \frac{15}{8}\|\mathbf{\Delta}\|_F + \frac{16d\|\mathbf{\Delta}\|_\infty}{\sqrt{n}} \leq \left(\frac{15}{8} + \frac{1}{16}\right)\|\mathbf{\Delta}\|_F \leq 2\|\mathbf{\Delta}\|_F \tag{6.46}$$

with probability greater than $1 - c_7\exp(-c_8 d\log d)$. When (6.46) holds, plug it into (6.44), we shall have

$$\mathrm{vec}(\mathbf{\Delta})^T[\widehat{\mathbf{H}}(\tilde{\mathbf{\Theta}}) - \widehat{\mathbf{H}}(\mathbf{\Theta}^*)]\mathrm{vec}(\mathbf{\Delta}) \leq \frac{R}{5} \cdot 4\|\mathbf{\Delta}\|_F^2 \leq \frac{\|\mathbf{\Delta}\|_F^2}{512(e^R + e^{-R} + 2)} \tag{6.47}$$

for sufficiently small $R > 0$. The following inequality thus holds for all $\tilde{\mathbf{\Theta}}$ satisfying $\left\|\tilde{\mathbf{\Theta}} - \mathbf{\Theta}^*\right\|_\infty \leq 2R/d$:

$$\text{vec}(\mathbf{\Delta})^T \widehat{\mathbf{H}}(\tilde{\mathbf{\Theta}})\text{vec}(\mathbf{\Delta}) \geq \frac{\|\mathbf{\Delta}\|_F^2}{512(e^R + e^{-R} + 2)} \tag{6.48}$$

6.8 Proof for Theorem 4

In this proof, we define an operator $\tilde{\mathfrak{X}}_n : \mathbb{R}^{d\times d} \to \mathbb{R}^n$ such that $[\tilde{\mathfrak{X}}_n(\mathbf{\Gamma})]_i = \langle \mathbf{\Gamma}, \mathbf{X}_i\rangle$ for all $\mathbf{\Gamma} \in \mathbb{R}^{d\times d}$.

Denote $\widehat{\mathbf{\Delta}} = \widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*$. If $\widehat{\mathbf{\Delta}} \notin \mathcal{C}'(c_0)$, according to Case 1 in the proof for Theorem 2 in Negahban and Wainwright (2012), we shall have

$$\left\|\widehat{\mathbf{\Delta}}\right\|_F^2 \leq 2c_0 R\sqrt{\frac{d\log d}{n}} \cdot \left\{ 8\sqrt{r}\left\|\widehat{\mathbf{\Delta}}\right\|_F + 4\sum_{j=r+1}^{d} \sigma_j(\mathbf{\Theta}^*)\right\} \tag{6.49}$$

for any $1 \leq r \leq d$. Following the same strategy we used in the proof for Theorem 1, we will have

$$\left\|\widehat{\mathbf{\Delta}}\right\|_F \leq C_1\sqrt{\rho}\left(2C_1 R\sqrt{\frac{d\log d}{n}}\right)^{1-q/2}$$

for some constant C_1.

If $\widehat{\mathbf{\Delta}} \in \mathcal{C}'(c_0)$, when (2.16) in Lemma 1 holds, on one hand, if $128d\left\|\widehat{\mathbf{\Delta}}\right\|_\infty/\sqrt{n}\left\|\widehat{\mathbf{\Delta}}\right\|_F$ $> 1/2$, we have

$$\left\|\widehat{\mathbf{\Delta}}\right\|_F \leq \frac{256d\left\|\widehat{\mathbf{\Delta}}\right\|_\infty}{\sqrt{n}} \leq \frac{512R}{\sqrt{n}} \tag{6.50}$$

As what we did in the proof for Theorem 1, we take $\tau = \left(R^2/\rho n\right)^{\frac{1}{2-q}}$ and we have

$$\left\|\widehat{\mathbf{\Delta}}\right\|_N \leq C_2\left(\rho\left(\frac{R^2}{n}\right)^{1-q}\right)^{\frac{1}{2-q}} \tag{6.51}$$

for some constant C_2.

On the other hand, if $128d\left\|\widehat{\mathbf{\Delta}}\right\|_\infty/\sqrt{n}\left\|\widehat{\mathbf{\Delta}}\right\|_F \leq 1/2$, we have

$$\frac{\left\|\mathfrak{X}_n(\widehat{\mathbf{\Delta}})\right\|_2}{\sqrt{n}} \geq \frac{\left\|\widehat{\mathbf{\Delta}}\right\|_F}{16(e^{R/2} + e^{-R/2})} \quad \text{i.e.,} \quad \frac{\left\|\mathfrak{X}_n(\widehat{\mathbf{\Delta}})\right\|_2^2}{n} \geq \frac{\left\|\widehat{\mathbf{\Delta}}\right\|_F^2}{256(e^R + e^{-R} + 2)} \tag{6.52}$$

Thus by Lemma 1 and 2 it naturally holds that

$$\left\|\widehat{\boldsymbol{\Theta}}-\boldsymbol{\Theta}\right\|_F^2 \leq C_3\rho\left(\sqrt{\frac{d\log d}{n}}\right)^{2-q}, \quad \left\|\widehat{\boldsymbol{\Theta}}-\boldsymbol{\Theta}\right\|_N \leq C_4\rho\left(\sqrt{\frac{d\log d}{n}}\right)^{1-q}.$$

In summary, as long as $n/(d\log d)$ is sufficiently large, we shall have

$$\begin{aligned} \left\|\widehat{\boldsymbol{\Theta}}-\boldsymbol{\Theta}^*\right\|_F^2 &\leq C_5\max\left\{\rho\left(\sqrt{\frac{d\log d}{n}}\right)^{2-q}, \frac{R^2}{n}\right\}, \\ \left\|\widehat{\boldsymbol{\Theta}}-\boldsymbol{\Theta}^*\right\|_N &\leq C_6\max\left\{\rho\left(\sqrt{\frac{d\log d}{n}}\right)^{1-q}, \left(\rho\left(\frac{R^2}{n}\right)^{1-q}\right)^{\frac{1}{2-q}}\right\} \end{aligned} \tag{6.53}$$

with probability greater than $1-C_7\exp\left(-c_1 d\log d\right)-2d^{1-\delta}$, where $\{C_i\}_{i=5}^7$ and c_1 are constants.