Journal of Machine Learning Research 23 (2022) 1-62 Submitted 7/20; Revised 4/22; Published 7/22

Matrix Completion with Covariate Information and Informative Missingness

Huaqing Jin* HUAQING.JIN@UCSF.EDU
Department of Statistics and Actuarial Science
The University of Hong Kong, Hong Kong

Yanyuan Ma YZM63@PSU.EDU
Department of Statistics
Pennsylvania State University

Fei Jiang† FEI.JIANG@UCSF.EDU
Department of Epidemiology and Biostatistics
The University of California, San Francisco

Editor: Benjamin Marlin

Abstract

We study the problem of matrix completion when the missingness of the matrix entries is dependent on the unobserved response values themselves and hence the missingness itself is informative. Furthermore, we allow to take into account the covariate information to establish its relation with the response and hence enable prediction. We devise a novel procedure to simultaneously complete the partially observed matrix and assess the covariate effect. Allowing the matrix dimensions as well as the number of covariates to grow ultra-high, under the classic low-rank matrix and sparse covariate effect assumptions, we rigorously establish the statistical guarantee of our procedure and the algorithmic convergence. The method is demonstrated via simulation studies and is used to analyze a Yelp data set and a MovieLens data set.

keywords: matrix completion, informative missing, low rank, sparse, tensor

1. Motivation and Introduction

Matrix completion has become a popular research topic in both statistics and computer science, mainly driven by the commercial need from online providers. Consider a concrete example from Yelp, where the scores from a total of n customers evaluating m restaurants are of interest. Let Y_{ij} be an indicator denoting whether the evaluation of the ith subject regarding the jth restaurant is positive, where $i = 1, \ldots, n, j = 1, \ldots, m$. Let $\mathbf{Y}$ be the collection of Y_{ij}'s, i.e. $\mathbf{Y}$ is a $n \times m$ matrix. Obviously, not everyone will evaluate every restaurant, so many elements in $\mathbf{Y}$ are missing. Let R_{ij} denote the corresponding missingness index and let $\mathbf{R}$ be the corresponding matrix formed by the collection of R_{ij}'s. It is sensible to suspect that the very fact of missingness is related to the potential evalua-

∗. Currently in Department of Radiology and Biomedical Imaging, UCSF.
†. Corresponding Author.

tion score itself—in fact, some individuals may dislike some restaurants so much that they refuse to visit them hence will naturally not evaluate them. Thus, if they had visited and evaluated, the feedback would be likely negative. This naturally leads to the nonignorable missing mechanism, where Y_{ij} and R_{ij} are dependent on each other. To recover the missing entries enables us to predict the missing evaluating scores, but is challenging because it is impossible to infer the distribution of Y_{ij} based on the biased sample formed by the observed entries. The issue can be resolved when there exist additional baseline covariates for each observation, while these covariates are independent of R_{ij} given Y_{ij}. When such covariate information is available, we aim to predict the missing entries of $\mathbf{Y}$ based on the relation between Y_{ij} and the covariates.

Luckily, in the Yelp data, a large number of covariates have been collected, which include but are not limited to the covariates of a restaurant such as size, price, open hours, and traits of a customer such as the number of his/her friends and the number of his/her restaurant reviews. These covariates are fully observed, which makes the parameter estimation feasible even though Y_{ij} and R_{ij} are correlated. On the other hand, the covariates are of the ultra-high dimension, which brings difficulties in parameter estimation. To take into account both the high dimensional covariates and the nonignorable missingness nature of such data, we impose the familiar sparsity and low-rank constraints on the relation between Y_{ij} and the covariate vector $\mathbf{X}_{ij}$. Furthermore, in practice, the missing-data mechanism is often not well understood. Thus, we consider statistical methods that do not require specification of the mechanism. This leads us to the matrix completion problem with ultra-high dimensional covariate and nonignorable missingness.

In the remaining text, we describe our statistical procedure in Section 2 and establish the finite sample statistical properties in Section 3. We consider the computational issue both algorithmically and theoretically in Section 4. Simulated examples are demonstrated in Section 5 and we analyze both a Yelp data set and a MovieLens data set in Section 6. We conclude the paper in Section 8, and relegate the proof details to an Appendix.

2. Methodology

We now present the probability model used to construct likelihood for the parameter estimation. Let $\mathbf{X}_{ij}$ be a p-dimensional covariate vector. Furthermore let the kth element of $\mathbf{X}_{ij}$ be X_{ijk} and let $\mathbf{X}$ be the $n \times m \times p$ tensor whose (i, j, k)th element is X_{ijk}. Let $\mathbf{Y}$ be the $n \times m$ matrix whose (i, j)th element is Y_{ij}. We work in a very flexible model class, the generalized linear model, where the conditional density of Y_{ij} given $\mathbf{X}_{ij}$ is $f(Y_{ij}, \mathbf{\Theta}_{0ij} + \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{X}_{ij})$, where $\boldsymbol{\beta}_0$ is the true effect from $\mathbf{X}_{ij}$ on Y_{ij}, $\mathbf{\Theta}_{0ij}$ is the (i, j)th element of $\mathbf{\Theta}_0$. Here $\mathbf{\Theta}_0$ is the true intercept matrix which is a low rank $n \times m$ matrix with rank r. Because it is expected that only a small number of covariates may affect the outcome, we assume the widely adopted sparsity assumption on $\boldsymbol{\beta}_0$, and assume the sparseness of $\boldsymbol{\beta}_0$ is s, i.e. $\|\boldsymbol{\beta}_0\|_0 = s$. When we focus on the estimation of the parameter $\boldsymbol{\beta}_0$, we also assume $\sum_k \beta_{0k}\mathbf{X}^k$ to be sparse to ensure identifiability, where $\mathbf{X}^k$ is a $n \times m$ matrix whose (i, j)th element is X_{ijk}. Of course, when we only aim at prediction, then we do not make this assumption. Here despite of the sparsity and the low-rank assumptions, we do not restrict s, r to be finite. In other words, we allow s, r to grow with the sample size, although at a slower rate than p and n, m. For notational simplicity, we denote $d = \sqrt{mn}$.

Under the informative missing setting, Y_{ij} and R_{ij} are dependent. This implies that the expectation of the logarithm of the observed data likelihood $\sum_{i=1}^{n}\sum_{j=1}^{m} R_{ij}\log f(Y_{ij}, \boldsymbol{\Theta}_{ij} + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_{ij})$ is no longer maximized at the true values $\boldsymbol{\beta}_0, \boldsymbol{\Theta}_0$, because $E\{R_{ij}\log f(Y_{ij}, \boldsymbol{\Theta}_{ij} + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_{ij})\}$ does not equal to $E(R_{ij}|\mathbf{X}_{ij})E\{\log f(Y_{ij}, \boldsymbol{\Theta}_{ij} + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_{ij})\}$, which would have been the case for the noninformative missing. Hence when estimating $\boldsymbol{\Theta}_{ij}$ and $\boldsymbol{\beta}$, the information in R_{ij} cannot be ignored, and the loss function constructed from this likelihood cannot be used to estimate the parameters. On the other hand, because $\mathbf{X}_{ij}$ is fully observed, under the assumption that $\mathbf{X}_{ij}$ is independent of R_{ij} when Y_{ij} is given, even though Y_{ij} may not be available, we still can use the conditional distribution of $\mathrm{pr}(\mathbf{X}_{ij}|Y_{ij}, \boldsymbol{\Theta}_0, \boldsymbol{\beta}_0)$ to construct a pseudo-likelihood. In fact, the conditional independence between $\mathbf{X}_{ij}$ and R_{ij} given Y_{ij} is not essential and can be relaxed. As long as part of the covariates in $\mathbf{X}_{ij}$ are independent of R_{ij} given Y_{ij} and the remaining part in $\mathbf{X}_{ij}$, we can still construct a pseudo-likelihood. The essential benefit of considering a pseudo-likelihood is in eliminating the sampling bias. Intuitively, the observed portion of the data, i.e. the complete data, form a biased sample of the hypothetical full data, caused by the sampling bias in Y_{ij}'s. Through conditioning on Y_{ij}'s, the problem on the surface becomes to study the dependence of $\mathbf{X}_{ij}$ given Y_{ij}, which does not relate to how Y_{ij} is sampled any more since the sampling of Y_{ij} now becomes a design issue hence could be done in any way we like. To help better understanding this issue, consider a standard regression problem where the response variable is named $\mathbf{X}$ and the covariates named Y. In estimating the parameter involved in the model of $\mathbf{X} \mid Y$, we can construct various estimators based on the pairs of data $(Y, \mathbf{X})$'s without worrying how the covariates Y's are sampled. In fact, even if the covariates Y's are collected by design, say we only collected the covariates on a grid in a fixed region, it does not prevent us from obtaining a valid estimator based on the collected data. Here, in our context, we can view the non-missing Y_{ij}'s as the collected covariates, and the complete data $(Y_{ij}, \mathbf{X}_{ij})$'s as the observations in a standard regression problem to help grasp the intuition behind the pseudo likelihood method. We name this view point a covariate-dependent design scheme.

Specifically, starting from the conditional distribution of $\mathbf{X}_{ij}$ given Y_{ij}, the average of the logarithm of the pseudo-likelihood, i.e. the conditional likelihood of the complete data, is written as

$$(mn)^{-1}\log\left[\prod_{i=1}^{n}\prod_{j=1}^{m}\{\mathrm{pr}(\mathbf{X}_{ij}|Y_{ij}, \boldsymbol{\Theta}_0, \boldsymbol{\beta}_0)\}^{R_{ij}}\right] \propto -(mn)^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m}\mathcal{L}_{ij}(\boldsymbol{\Theta}_0, \boldsymbol{\beta}_0),$$

where $\mathcal{L}_{ij}(\boldsymbol{\Theta}, \boldsymbol{\beta}) = R_{ij}\ell_{ij}(\boldsymbol{\Theta}, \boldsymbol{\beta})$,

$$\ell_{ij}(\boldsymbol{\Theta}, \boldsymbol{\beta}) = -\left[\log\{f(Y_{ij}, \boldsymbol{\Theta}_{ij} + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_{ij})\} - \log\left\{\int f(Y_{ij}, \boldsymbol{\Theta}_{ij} + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})g(\mathbf{X})d\mathbf{X}\right\}\right],$$

and $g(\cdot)$ is the joint Radon-Nikodym probability density function of $\mathbf{X}_{ij}$, i.e., the Radon-Nikodym derivative of the probability distribution of interest with respect to the dominating measure. We require that $g(\cdot)$ is not a Dirac function. The dominating measure for the continuous component in $\mathbf{X}_{ij}$ is the Lebesgue measure. The dominating measure for the discrete component in $\mathbf{X}_{ij}$ is the counting measure, and therefore the integration with respect to the discrete component is the sum over its domain. Clearly, because $\Pr(R_{ij} = 1|Y_{ij}, \mathbf{X}_{ij}) = \Pr(R_{ij} = 1|Y_{ij})$, $E\{R_{ij}\ell_{ij}(\boldsymbol{\Theta}, \boldsymbol{\beta})|Y_{ij}\} = E(R_{ij}|Y_{ij})E\{\ell_{ij}(\boldsymbol{\Theta}, \boldsymbol{\beta})|Y_{ij}\}$. Therefore

the minimizer of $E\{\ell_{ij}(\boldsymbol{\Theta}, \boldsymbol{\beta}) \mid Y_{ij}\}$ is the minimizer of $E\{\mathcal{L}_{ij}(\boldsymbol{\Theta}, \boldsymbol{\beta}) \mid Y_{ij}\}$. Furthermore, $\Pr(R_{ij} = 1|Y_{ij})$ and the marginal distribution of Y_{ij} do not contain information regarding the unknown parameters. Therefore taking expectation with respect to Y_{ij}, we minimize $E\{\ell_{ij}(\boldsymbol{\Theta}, \boldsymbol{\beta})\}$ through minimizing $E\{\mathcal{L}_{ij}(\boldsymbol{\Theta}, \boldsymbol{\beta})\}$.

Now, let $\mathcal{L}(\boldsymbol{\Theta}, \boldsymbol{\beta}) \equiv (mn)^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m} \mathcal{L}_{ij}(\boldsymbol{\Theta}, \boldsymbol{\beta})$ be the negative pseudo-likelihood. Conditional on the observed Y_{ij}'s, the observed covariates form a conditional random sample even when the response missingness depends on the response value Y_{ij} itself. Furthermore, the expected logarithm of the pseudo-likelihood, i.e. $E\{\mathcal{L}_{ij}(\boldsymbol{\Theta}, \boldsymbol{\beta})\}$, is maximized at $\boldsymbol{\Theta}_0$ and $\boldsymbol{\beta}_0$. This can be verified in two ways. The first way continues from the covariate design point of view. Because the pseudo-likelihood is the same as the conditional likelihood of $\mathbf{X}_{ij} \mid Y_{ij}$ restricted to the complete data, while the complete data are obtained from a covariate-dependent (i.e. Y dependent) design scheme, hence the maximizer is the true regression parameters $\boldsymbol{\Theta}_0$ and $\boldsymbol{\beta}_0$ at infinite samples. The second way is through detailed mathematical computation. We can verify that the derivative of the log pseudo-likelihood has expectation zero at $\boldsymbol{\Theta}_0$ and $\boldsymbol{\beta}_0$. Further, we can also verify that the log pseudo-likelihood is a convex function, so it is indeed maximized at Θ_0 and $\boldsymbol{\beta}_0$ at infinite samples. Hence, following standard M-estimation theory (Van Der Vaart and Wellner, 2000), maximizing the pseudo-likelihood will indeed lead to a valid statistical estimation procedure. Note that the function we maximize is not the same as the log-likelihood of the observed data, regardless we write the likelihood in terms of $\mathbf{X}_{ij} \mid Y_{ij}$ or $Y_{ij} \mid \mathbf{X}_{ij}$, but is the conditional likelihood of the complete data, hence it is named pseudo-likelihood.

When the dimensions of $\boldsymbol{\Theta}$ and $\boldsymbol{\beta}$ are ultra-high, the value of the regression function can diverge to infinity if the entries of $\boldsymbol{\Theta}$ and $\boldsymbol{\beta}$ are not bounded. This is unreasonable when the response has finite mean. To control the magnitude of the regression function, we assume that $\boldsymbol{\Theta}_0$ and $\boldsymbol{\beta}_0$ have finite entries, and search the estimators in the feasible sets that $\|\boldsymbol{\Theta}\|_{\max} \leq a$ and $\|\boldsymbol{\beta}\|_\infty \leq a$ for a constant $a > 0$. Combining with the low-rank and sparse structures of $\boldsymbol{\Theta}$ and $\boldsymbol{\beta}$ respectively, we propose to estimate $\boldsymbol{\Theta}_0, \boldsymbol{\beta}_0$, through

$$(\widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\beta}}) = \underset{\|\boldsymbol{\Theta}\|_{\max} \leq a, \|\boldsymbol{\beta}\|_\infty \leq a}{\operatorname{argmin}} \mathcal{L}(\boldsymbol{\Theta}, \boldsymbol{\beta}) + \lambda_{\boldsymbol{\Theta}} \|\boldsymbol{\Theta}\|_* + \lambda_{\boldsymbol{\beta}} \|\boldsymbol{\beta}\|_1. \tag{1}$$

Here, $\|\boldsymbol{\Theta}\|_*$ is the nuclear norm of the matrix $\boldsymbol{\Theta}$, which drives the resulting estimator towards a low-rank matrix. Furthermore, $\|\boldsymbol{\beta}\|_1$ is the L_1 norm of $\boldsymbol{\beta}$, which induces the sparseness of the estimated $\boldsymbol{\beta}$. In general, for any matrix $\mathbf{A}$, we use $\|\mathbf{A}\|_F$, $\|\mathbf{A}\|_{op}$, $\|\mathbf{A}\|_*$, $\|\mathbf{A}\|_1$, $\|\mathbf{A}\|_\infty$ and $\|\mathbf{A}\|_{\max}$ to denote the Frobenius norm, spectral norm, nuclear norm, matrix 1-norm, matrix sup-norm and the element-wise maximal norm of the matrix $\mathbf{A}$, respectively. We also use $\|\mathbf{a}\|_2$, $\|\mathbf{a}\|_\infty$, $\|\mathbf{a}\|_1$ to denote the L_2, L_∞ and L_1 norm of the vector $\mathbf{a}$, respectively. We use boldface letters to denote vectors or matrices throughout the text.

3. Statistical Properties

3.1 Additional Notation

We introduce some additional notations to facilitate the presentation of the theoretic properties. Some of their explicit forms are derived in Appendix A. Let f_2 and f_{22} be the first and second derivative of f with respect to the second argument, respectively. Let $\mathbf{S}(Y_{ij}, \mathbf{X}_{ij}|\boldsymbol{\Theta}, \boldsymbol{\beta})$ be the derivative of $-\log\{f(Y_{ij}, \boldsymbol{\Theta}_{ij} + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_{ij})\}$ with respect to $\mathbf{M} =$

Notation	Definition
f_2	The first derivative of f with respect to the second argument.
f_{22}	The second derivative of f with respect to the second argument.
$\mathbf{z}_{ij}$	An $n \times m$ matrix with its (i,j)th entry 1 and all other entries 0.
d_{H_2}	Upper bound of $\lvert H_2(y, \mathbf{x}\lvert\boldsymbol{\Theta}, \boldsymbol{\beta})\rvert$.
d_{S_2}	Upper bound of $\lvert S_2(y, \mathbf{x}\lvert\boldsymbol{\Theta}, \boldsymbol{\beta})\rvert$.
a	Upper bound of $\lVert\boldsymbol{\beta}\rVert_\infty$ and $\lVert\boldsymbol{\Theta}\rVert_{\max}$.
c_0	Upper bound of $\lVert\mathbf{X}_{ij}\rVert_1$.
$\mathbf{X}^k$	$n \times m$ matrix whose (i,j)th element is X_{ijk}.
σ_{1F}	$32c_0^2a^2(d_{H_2} + d_{S_2}^2)$.
σ_{dF}	$32a^2(d_{H_2} + d_{S_2}^2)$.
$W_{ijk}(\boldsymbol{\Theta}, \boldsymbol{\beta})$	$\mathbf{e}_k^{\mathrm{T}}\partial^2\ell_{ij}(\boldsymbol{\Theta}, \boldsymbol{\beta})/\partial\boldsymbol{\beta}\partial\langle\boldsymbol{\Theta}, \mathbf{z}_{ij}\rangle$.
$d_{\mathbf{W}}$	Upper bound of $\sup_k \lVert\mathrm{vec}\{\mathbf{R} \circ \mathbf{W}_k(\boldsymbol{\Theta}, \boldsymbol{\beta}_0)\}\rVert_1$.
$d_{E\mathbf{X}}$	Upper bound of $\sup_k \lvert\sum_{i=1}^n\sum_{j=1}^m R_{ij}E\{S_2(Y_{ij}, \mathbf{X}\lvert\boldsymbol{\Theta}_0, \boldsymbol{\beta}_0)X_k\lvert Y_{ij}\}\rvert$.
$\mathbf{E}_{ij}, E_{ijk}$	$-R_{ij}\partial\ell_{ij}(\boldsymbol{\Theta}, \boldsymbol{\beta})/\partial\boldsymbol{\beta}$, $\mathbf{e}_k^{\mathrm{T}}\mathbf{E}_{ij}$.
$\mathbf{S}_{\boldsymbol{\Theta}}(Y_{ij}, \mathbf{X}_{ij}\lvert\boldsymbol{\Theta}, \boldsymbol{\beta})$	Partial derivatives of $-\log\{f(Y_{ij}, \boldsymbol{\Theta}_{ij} + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_{ij})\}$ with respect to $\boldsymbol{\Theta}$.
$\mathbf{S}_{\boldsymbol{\beta}}(Y_{ij}, \mathbf{X}_{ij}\lvert\boldsymbol{\Theta}, \boldsymbol{\beta})$	Partial derivatives of $-\log\{f(Y_{ij}, \boldsymbol{\Theta}_{ij} + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_{ij})\}$ with respect to $\boldsymbol{\beta}$.
$\mathbf{S}(Y_{ij}, \mathbf{X}_{ij}\lvert\boldsymbol{\Theta}, \boldsymbol{\beta})$	$[\mathrm{vec}\{\mathbf{S}_{\boldsymbol{\Theta}}(Y_{ij}, \mathbf{X}_{ij}\lvert\boldsymbol{\Theta}, \boldsymbol{\beta})\}^{\mathrm{T}}, \mathbf{S}_{\boldsymbol{\beta}}(Y_{ij}, \mathbf{X}_{ij}\lvert\boldsymbol{\Theta}, \boldsymbol{\beta})^{\mathrm{T}}]^{\mathrm{T}}$.
$S_2(Y_{ij}, \mathbf{X}_{ij}\lvert\boldsymbol{\Theta}, \boldsymbol{\beta})$	The first derivative of $-\log f(Y_{ij}, \boldsymbol{\Theta}_{ij} + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_{ij})$ with respect to $\boldsymbol{\Theta}_{ij} + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_{ij}$.
$H_2(Y_{ij}, \mathbf{X}_{ij}\lvert\boldsymbol{\Theta}, \boldsymbol{\beta})$	The second derivative of $-\log f(Y_{ij}, \boldsymbol{\Theta}_{ij} + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_{ij})$ with respect to $\boldsymbol{\Theta}_{ij} + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_{ij}$.
$\mathbf{H}(Y_{ij}, \mathbf{X}_{ij}\lvert\boldsymbol{\Theta}, \boldsymbol{\beta})$	The Hessian matrix of $-\log\{f(Y_{ij}, \boldsymbol{\Theta}_{ij} + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_{ij})\}$ with respect to $\{\mathrm{vec}(\boldsymbol{\Theta})^{\mathrm{T}}, \boldsymbol{\beta}^{\mathrm{T}}\}^{\mathrm{T}}$.
$\mathbf{H}_{\boldsymbol{\Theta}}(Y_{ij}\mathbf{X}_{ij}\lvert\boldsymbol{\Theta}, \boldsymbol{\beta})$	The second derivatives of $-\log f(Y_{ij}, \boldsymbol{\Theta}_{ij} + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})$ with respect to $\mathrm{vec}(\boldsymbol{\Theta})$.
$\mathbf{H}_{\boldsymbol{\beta}}(Y_{ij}, \mathbf{X}_{ij}\lvert\boldsymbol{\Theta}, \boldsymbol{\beta})$	The second derivatives of $-\log f(Y_{ij}, \boldsymbol{\Theta}_{ij} + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})$ with respect to $\mathrm{vec}(\boldsymbol{\beta})$.
$\mathbf{F}_{\boldsymbol{\beta}}(\mathbf{X}, \mathbf{Y}\lvert\boldsymbol{\Theta}, \boldsymbol{\beta})$	$\partial^2\mathcal{L}(\boldsymbol{\Theta}, \boldsymbol{\beta})/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^{\mathrm{T}}$.
$\mathbf{F}_{\boldsymbol{\Theta}}(\mathbf{X}, \mathbf{Y}\lvert\boldsymbol{\Theta}, \boldsymbol{\beta})$	$\partial^2\mathcal{L}(\boldsymbol{\Theta}, \boldsymbol{\beta})/\partial\mathrm{vec}(\boldsymbol{\Theta})\partial\mathrm{vec}(\boldsymbol{\Theta})^{\mathrm{T}}$.
$E\{\mathbf{F}_{\boldsymbol{\beta}}(\mathbf{X}, \mathbf{Y}\lvert\widehat{\boldsymbol{\Theta}}, \boldsymbol{\beta}^*)\}$	$E\{\mathbf{F}_{\boldsymbol{\beta}}(\mathbf{X}, \mathbf{Y}\lvert\boldsymbol{\Theta}, \boldsymbol{\beta})\}\lvert_{\boldsymbol{\Theta}=\widehat{\boldsymbol{\Theta}}, \boldsymbol{\beta}=\boldsymbol{\beta}^*}$.
$E\{\mathbf{F}_{\boldsymbol{\Theta}}(\mathbf{X}, \mathbf{Y}\lvert\boldsymbol{\Theta}^*, \widehat{\boldsymbol{\beta}})\}$	$E\{\mathbf{F}_{\boldsymbol{\Theta}}(\mathbf{X}, \mathbf{Y}\lvert\boldsymbol{\Theta}, \boldsymbol{\beta})\}\lvert_{\boldsymbol{\Theta}=\boldsymbol{\Theta}^*, \boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}}$.

Table 1: Notation.

$\{\mathrm{vec}(\boldsymbol{\Theta})^{\mathrm{T}}, \boldsymbol{\beta}^{\mathrm{T}}\}^{\mathrm{T}}$. Obviously $\mathbf{S}(Y_{ij}, \mathbf{X}_{ij}|\boldsymbol{\Theta}, \boldsymbol{\beta})$ is the score function. We further decompose the score function as $\mathbf{S}(Y_{ij}, \mathbf{X}_{ij}|\boldsymbol{\Theta}, \boldsymbol{\beta}) = [\mathrm{vec}\{\mathbf{S}_{\boldsymbol{\Theta}}(Y_{ij}, \mathbf{X}_{ij}|\boldsymbol{\Theta}, \boldsymbol{\beta})\}^{\mathrm{T}}, \mathbf{S}_{\boldsymbol{\beta}}(Y_{ij}, \mathbf{X}_{ij}|\boldsymbol{\Theta}, \boldsymbol{\beta})^{\mathrm{T}}]^{\mathrm{T}}$, where $\mathbf{S}_{\boldsymbol{\Theta}}(Y_{ij}, \mathbf{X}_{ij}|\boldsymbol{\Theta}, \boldsymbol{\beta})$ and $\mathbf{S}_{\boldsymbol{\beta}}(Y_{ij}, \mathbf{X}_{ij}|\boldsymbol{\Theta}, \boldsymbol{\beta})$ are the partial derivatives of $-\log\{f(Y_{ij}, \boldsymbol{\Theta}_{ij} + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_{ij})\}$ with respect to $\boldsymbol{\Theta}$ and $\boldsymbol{\beta}$ respectively. Let $\mathbf{S}_{\boldsymbol{\Theta},k,l}(Y_{ij}, \mathbf{X}_{ij}|\boldsymbol{\Theta}, \boldsymbol{\beta})$, $\mathbf{S}_{\boldsymbol{\beta},k}(Y_{ij}, \mathbf{X}_{ij}|\boldsymbol{\Theta}, \boldsymbol{\beta})$ be the (k,l)th element and kth element of $\mathbf{S}_{\boldsymbol{\Theta}}(Y_{ij}, \mathbf{X}_{ij}|\boldsymbol{\Theta}, \boldsymbol{\beta})$ and $\mathbf{S}_{\boldsymbol{\beta}}(Y_{ij}, \mathbf{X}_{ij}|\boldsymbol{\Theta}, \boldsymbol{\beta})$, respectively.

Further let $S_2(Y_{ij}, \mathbf{X}_{ij}|\boldsymbol{\Theta}, \boldsymbol{\beta})$, $H_2(Y_{ij}, \mathbf{X}_{ij}|\boldsymbol{\Theta}, \boldsymbol{\beta})$ be the first and second derivative of the negative log-likelihood $-\log f(Y_{ij}, \boldsymbol{\Theta}_{ij} + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_{ij})$ with respect to $\boldsymbol{\Theta}_{ij} + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_{ij}$ and let $\mathbf{H}(Y_{ij}, \mathbf{X}_{ij}|\boldsymbol{\Theta}, \boldsymbol{\beta})$ be the Hessian matrix of $-\log\{f(Y_{ij}, \boldsymbol{\Theta}_{ij} + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_{ij})\}$ with respect to $\{\mathrm{vec}(\boldsymbol{\Theta})^{\mathrm{T}}, \boldsymbol{\beta}^{\mathrm{T}}\}^{\mathrm{T}}$. Let $\mathbf{H}_{\boldsymbol{\Theta}}(Y_{ij}\mathbf{X}_{ij}|\boldsymbol{\Theta}, \boldsymbol{\beta})$ and $\mathbf{H}_{\boldsymbol{\beta}}(Y_{ij}, \mathbf{X}_{ij}|\boldsymbol{\Theta}, \boldsymbol{\beta})$ be the diagonal blocks of $\mathbf{H}(Y_{ij}, \mathbf{X}_{ij}|\boldsymbol{\Theta}, \boldsymbol{\beta})$, i.e. they are the second derivatives of $-\log f(Y_{ij}, \boldsymbol{\Theta}_{ij} + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})$ with respect to $\mathrm{vec}(\boldsymbol{\Theta})$ and $\boldsymbol{\beta}$ respectively.

Let $\mathbf{z}_{ij}$ be an $n \times m$ matrix with its (i,j)th entry 1 and all other entries 0. Note that we can extract the (i,j)th entry of $\boldsymbol{\Theta}$ using $\mathbf{z}_{ij}$ and $\boldsymbol{\Theta}$. To simplify the notation, we define $W_{ijk}(\boldsymbol{\Theta}, \boldsymbol{\beta}) \equiv \mathbf{e}_k^{\mathrm{T}}\partial^2\ell_{ij}(\boldsymbol{\Theta}, \boldsymbol{\beta})/\partial\boldsymbol{\beta}\partial\langle\boldsymbol{\Theta}, \mathbf{z}_{ij}\rangle$. Then $\mathbf{e}_k^{\mathrm{T}}\partial^2\mathcal{L}_{ij}(\boldsymbol{\Theta}, \boldsymbol{\beta})/\partial\boldsymbol{\beta}\partial\langle\boldsymbol{\Theta}, \mathbf{z}_{ij}\rangle = R_{ij}W_{ijk}(\boldsymbol{\Theta}, \boldsymbol{\beta})$. Let $\mathbf{W}_k(\boldsymbol{\Theta}, \boldsymbol{\beta})$ be the $n \times m$ matrix with its (i,j)th element $W_{ijk}(\boldsymbol{\Theta}, \boldsymbol{\beta})$. Let $\mathbf{E}_{ij} \equiv -R_{ij}\partial\ell_{ij}(\boldsymbol{\Theta}, \boldsymbol{\beta})/\partial\boldsymbol{\beta}$ and $E_{ijk} = \mathbf{e}_k^{\mathrm{T}}\mathbf{E}_{ij}$. We also write $\mathbf{X}^k$ as the $n \times m$ matrix whose (i,j)th element is X_{ijk}, i.e. $\mathbf{X}^k$ is the kth slice of the tensor $\mathbf{X}$. Define $\mathbf{F}_{\boldsymbol{\beta}}(\mathbf{X}, \mathbf{Y}|\boldsymbol{\Theta}, \boldsymbol{\beta}) =$

$\partial^2\mathcal{L}(\boldsymbol{\Theta},\boldsymbol{\beta})/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^{\mathrm{T}}$, and $\mathbf{F}_{\boldsymbol{\Theta}}(\mathbf{X},\mathbf{Y}|\boldsymbol{\Theta},\boldsymbol{\beta}) = \partial^2\mathcal{L}(\boldsymbol{\Theta},\boldsymbol{\beta})/\partial\mathrm{vec}(\boldsymbol{\Theta})\partial\mathrm{vec}(\boldsymbol{\Theta})^{\mathrm{T}}$. Specifically,

$$\begin{aligned}
&\mathbf{F}_{\boldsymbol{\beta}}(\mathbf{X},\mathbf{Y}|\boldsymbol{\Theta},\boldsymbol{\beta}) \\
= \;& (mn)^{-1}\Bigg(\sum_{i=1}^{n}\sum_{j=1}^{m} R_{ij}[H_2(Y_{ij},\mathbf{X}_{ij}|\boldsymbol{\Theta},\boldsymbol{\beta})\mathbf{X}_{ij}\mathbf{X}_{ij}^{\mathrm{T}} - E\{H_2(Y_{ij},\mathbf{X}_{ij}|\boldsymbol{\Theta},\boldsymbol{\beta})\mathbf{X}_{ij}\mathbf{X}_{ij}^{\mathrm{T}}|Y_{ij}\} \\
& + E\{S_2^2(Y_{ij},\mathbf{X}_{ij}|\boldsymbol{\Theta},\boldsymbol{\beta})\mathbf{X}_{ij}\mathbf{X}_{ij}^{\mathrm{T}}|Y_{ij}\} - E\{S_2(Y_{ij},\mathbf{X}_{ij}|\boldsymbol{\Theta},\boldsymbol{\beta})\mathbf{X}_{ij}|Y_{ij}\}^{\otimes 2}]\Bigg),
\end{aligned}$$

and

$$\begin{aligned}
&\mathbf{F}_{\boldsymbol{\Theta}}(\mathbf{X},\mathbf{Y}|\boldsymbol{\Theta},\boldsymbol{\beta}) \\
= \;& (mn)^{-1}\Bigg(\sum_{i=1}^{n}\sum_{j=1}^{m} R_{ij}[H_2(Y_{ij},\mathbf{X}_{ij}|\boldsymbol{\Theta},\boldsymbol{\beta})\mathrm{vec}(\mathbf{z}_{ij})\mathrm{vec}(\mathbf{z}_{ij})^{\mathrm{T}} \\
& - E\{H_2(Y_{ij},\mathbf{X}_{ij}|\boldsymbol{\Theta},\boldsymbol{\beta})\mathrm{vec}(\mathbf{z}_{ij})\mathrm{vec}(\mathbf{z}_{ij})^{\mathrm{T}}|Y_{ij}\} \\
& + E\{S_2^2(Y_{ij},\mathbf{X}_{ij}|\boldsymbol{\Theta},\boldsymbol{\beta})\mathrm{vec}(\mathbf{z}_{ij})\mathrm{vec}(\mathbf{z}_{ij})^{\mathrm{T}}|Y_{ij}\} \\
& - E\{S_2(Y_{ij},\mathbf{X}_{ij}|\boldsymbol{\Theta},\boldsymbol{\beta})\mathrm{vec}(\mathbf{z}_{ij})|Y_{ij}\}^{\otimes 2}]\Bigg),
\end{aligned}$$

where all the expectations are with respect to the true distribution of $\mathbf{X}_{ij}$ given Y_{ij}.

We use $\boldsymbol{\beta}^*$ to denote a point on the line between $\widehat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}_0$ and $\boldsymbol{\Theta}^*$ to denote a point between $\widehat{\boldsymbol{\Theta}}$ and $\boldsymbol{\Theta}_0$, respectively. We summarize the notations in Table 1.

3.2 Conditions

To obtain the statistical convergence properties of our estimator, we assume the following regularity conditions.

(C1) $\|\boldsymbol{\Theta}_0\|_{\max} \le a$, $\|\boldsymbol{\beta}_0\|_\infty \le a$ for a constant $a > 0$.

(C2) The matrix $\sum_{k=1}^{p}\beta_{0k}\mathbf{X}^k$ is sparse and the matrix $\boldsymbol{\Theta}$ has low rank. Specifically, $\inf_{\rho>0}\omega_1(\rho)\omega_2(\rho) < 1$. Here,

$$\begin{aligned}
\omega_1(\rho) &\equiv \max\left\{\rho\|\mathrm{sign}\left(\sum_{k=1}^{p}\beta_{0k}\mathbf{X}^k\right)\|_{1\to 1}, \rho^{-1}\|\mathrm{sign}\left(\sum_{k=1}^{p}\beta_{0k}\mathbf{X}^k\right)\|_{\infty\to\infty}\right\}, \\
\omega_2(\rho) &\equiv \rho^{-1}\|\mathbf{U}\mathbf{U}^{\mathrm{T}}\|_{\max} + \rho\|\mathbf{V}\mathbf{V}^{\mathrm{T}}\|_{\max} + \|\mathbf{U}\|_{2\to\infty}\|\mathbf{V}\|_{2\to\infty},
\end{aligned}$$

where for any $n \times m$ matrix $\mathbf{M}$,

$$\|\mathbf{M}\|_{p\to q} \equiv \max\left(\|\mathbf{M}\mathbf{v}\|_q, \mathbf{v}\in\mathbb{R}^m, \|\mathbf{v}\|_p \le 1\right),$$

and $\mathbf{U},\mathbf{V}$ are the left and right singular vectors of $\boldsymbol{\Theta}_0$.

(C3) $\sup_k \|\mathrm{vec}\{\mathbf{R}\circ\mathbf{W}_k(\boldsymbol{\Theta},\boldsymbol{\beta}_0)\}\|_1 \le d_{\mathbf{W}}$ for $\boldsymbol{\Theta}$ that satisfies $\|\boldsymbol{\Theta}\|_{\max} \le a$. Here, $d_{\mathbf{W}} > 0$ and a is the constant defined in Condition (C1).

(C4) Assume

$$\sup_k |\sum_{i=1}^{n}\sum_{j=1}^{m} R_{ij}E\{S_2(Y_{ij},\mathbf{X}|\boldsymbol{\Theta}_0,\boldsymbol{\beta}_0)X_k|Y_{ij}\}| \leq d_{E\mathbf{X}}.$$

Here $d_{E\mathbf{X}} > 0$. Here $d_{E\mathbf{X}}$ does not need to be bounded.

(C5) Assume $f(Y_{ij}, \boldsymbol{\Theta}_{0ij} + \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{X}_{ij})$ is a continuously differentiable function with respect to $\boldsymbol{\Theta}_{0ij}$ and $\boldsymbol{\beta}_0$. Furthermore, $S_2(Y_{ij}, \mathbf{X}_{ij}|\boldsymbol{\Theta}_0, \boldsymbol{\beta}_0)$ is a sub-Gaussian random variable. $\|\mathbf{X}_{ij}\|_1 \leq c_0$ for $c_0 > 0$. This implies $\|\mathbf{X}_{ij}\|_2 \leq c_0$.

(C6) $|H_2(y, \mathbf{x}|\boldsymbol{\Theta}, \boldsymbol{\beta})|$ is bounded uniformly by d_{H_2} and $|S_2(y, \mathbf{x}|\boldsymbol{\Theta}, \boldsymbol{\beta})|$ is bounded uniformly by d_{S_2} for all $y, \mathbf{x}$, and $\boldsymbol{\Theta}, \boldsymbol{\beta}$ in the feasible set that $\|\boldsymbol{\Theta}\|_{\max} \leq a$, $\|\boldsymbol{\beta}\|_\infty \leq a$. Here $d_{H_2} > 0$ and $d_{S_2} > 0$ satisfy $c_0^2 a^2 (d_{H_2} + d_{S_2}^2)\sqrt{\log\{\max(p, mn)\}/(mn)} \to 0$ and $a^2(d_{H_2} + d_{S_2}^2)\sqrt{d\log(d)/(mn)} \to 0$.

Condition (C1) bounds the supnorms of $\boldsymbol{\Theta}_0$ and $\boldsymbol{\beta}_0$ so that the true parameters fall in the feasible set defined in (1). Let $\mathbf{E} = \sum_{k=1}^{p} \beta_{0k}\mathbf{X}^k$, then Condition (C2) is a standard identifiability condition for the matrix completion problem to identify $\boldsymbol{\Theta}_0$ and $\mathbf{E}$, where the regression function is the sum of a low-rank matrix and a sparse matrix (Hsu et al., 2011). Here, we set $\boldsymbol{\Theta}$ to be low rank and allow $\mathbf{E} = \sum_{k=1}^{p} \beta_{0k}\mathbf{X}^k$ to be sparse because the covariate matrices in the social media data are often naturally sparse, where the majority of them are dictionary variables with a small number of nonzero entries (Robin et al., 2018). Without this condition, there will not be a clear definition of $\boldsymbol{\beta}_0$, hence it is impossible to derive the distance $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2$. Condition (C2) needs to hold only when we are interested in estimating $\boldsymbol{\Theta}_0$ and $\boldsymbol{\beta}_0$. If the goal is to predict the missing entries in $\mathbf{Y}$, Condition (C2) is not necessary. Note that we also assume sparsity on $\boldsymbol{\beta}_0$, which is customary in treating high dimensional regression and does not pertain to the matrix completion problem alone. Condition (C3) ensures that the objective function has bounded second derivative with respect to $\boldsymbol{\beta}$ in L_1 norm. Conditions (C4), (C5) and (C6) provide the upper bounds of the high dimensional covariate, score functions and Hessian matrix, respectively. It is important to note that c_0, d_{H_2} and d_{S_2} are not necessarily bounded. Their growing rates jointly determine the convergence of $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\Theta}}$ as shown in Theorems 1 and 2.

3.3 Statistical Guarantee for $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\Theta}}$

We utilize the profiling procedure to show that $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\Theta}}$ converge to the true values. For $\widehat{\boldsymbol{\beta}}$, we show that $\widehat{\boldsymbol{\beta}}$ approaches $\boldsymbol{\beta}_0$ when $\boldsymbol{\Theta}$ in the loss function is fixed at $\widehat{\boldsymbol{\Theta}}$. To reach this result, we first derive an upper bound of $\|\partial\mathcal{L}(\widehat{\boldsymbol{\Theta}}, \boldsymbol{\beta}_0)/\partial\boldsymbol{\beta}\|_\infty$ in terms of $d_{E\mathbf{X}}$ and $d_{\mathbf{W}}$. The upper bound vanishes as long as $d_{E\mathbf{X}}$ and $d_{\mathbf{W}}$ grow slower than mn. Here because we consider $\widehat{\boldsymbol{\Theta}}$ instead of $\boldsymbol{\Theta}_0$ in the loss function, the upper bound is slightly larger than the standard result in high dimensional generalized linear models, where only a sparse high dimensional parameter is of interest. Furthermore, to assess the convexity of the objective function, we start with verifying that a type of the restricted eigenvalue condition (e.g. Bickel et al. (2009); Van De Geer et al. (2009)) is satisfied on a high sparsity (low L_1/L_2) and low spikiness (low L_∞/L_2) parameter set, which contains our true parameter that

satisfies $\|\boldsymbol{\beta}\|_0 = s$ and $\|\boldsymbol{\beta}\|_\infty \leq a$. To do that, we carefully define a set $\mathbb{B}(D)$ (Section C), which has the following two properties: (1) the covering number of $\mathbb{B}(D)$, that is the number of spherical balls of a given size that cover $\mathbb{B}(D)$, grows slower than $O\{\exp(mn)\}$; (2) its union over all values of D covers the entire space of $\boldsymbol{\beta}$. In such $\mathbb{B}(D)$, we show that the probability of $|\boldsymbol{\beta}^{\mathrm{T}}\mathbf{F}_{\boldsymbol{\beta}}(\mathbf{X},\mathbf{Y}|\widetilde{\boldsymbol{\Theta}},\widetilde{\boldsymbol{\beta}})\boldsymbol{\beta} - \boldsymbol{\beta}^{\mathrm{T}}E\{\mathbf{F}_{\boldsymbol{\beta}}(\mathbf{X},\mathbf{Y}|\widetilde{\boldsymbol{\Theta}},\widetilde{\boldsymbol{\beta}})\}\boldsymbol{\beta}| \to 0$ at given $\widetilde{\boldsymbol{\Theta}},\widetilde{\boldsymbol{\beta}}$ for any $\boldsymbol{\beta}$ approaches one when $mn \to \infty$. In addition, we show that $\boldsymbol{\beta}^{\mathrm{T}}E\{\mathbf{F}_{\boldsymbol{\beta}}(\mathbf{X},\mathbf{Y}|\widetilde{\boldsymbol{\Theta}},\widetilde{\boldsymbol{\beta}})\}\boldsymbol{\beta} > 0$, which leads to the result that $\mathbf{F}_{\boldsymbol{\beta}}(\mathbf{X},\mathbf{Y}|\widetilde{\boldsymbol{\Theta}},\widetilde{\boldsymbol{\beta}})$ is asymptotically strictly positive definite as described in Lemma A.10. This establishes that $\mathbf{F}_{\boldsymbol{\beta}}(\mathbf{X},\mathbf{Y}|\widetilde{\boldsymbol{\Theta}},\widetilde{\boldsymbol{\beta}})$ indeed satisfies the restricted eigenvalue condition. This, together with the boundedness of $\|\partial\mathcal{L}(\widehat{\boldsymbol{\Theta}},\boldsymbol{\beta}_0)/\partial\boldsymbol{\beta}\|_\infty$, leads to the convergence of $\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|_2$. We use a similar procedure to derive the upper bound for $\|\widehat{\boldsymbol{\Theta}}-\boldsymbol{\Theta}_0\|_F$ when $\boldsymbol{\beta}$ in the loss function is fixed at $\widehat{\boldsymbol{\beta}}$. Specifically, to study the convexity of the objective function, we verify that the restricted eigenvalue property is satisfied on a low rank (low nuclear norm/Frobenius norm ratio) and low spikiness (low $L_{\max}/L_2$) set, which contains the true parameter that satisfies $\mathrm{rank}(\boldsymbol{\Theta}) = r$ and $\|\boldsymbol{\Theta}\|_{\max} \leq a$.

Lemma 1 *Assume Conditions (C1), (C4) and (C5) to hold. Then there is a constant $\omega > 0$ so that*

$$\left\|\frac{\partial\mathcal{L}(\widehat{\boldsymbol{\Theta}},\boldsymbol{\beta}_0)}{\partial\boldsymbol{\beta}}\right\|_\infty \leq \sqrt{\frac{\omega\log\{\max(p,mn)\}}{mn}} + \frac{d_{E\mathbf{X}}}{mn} + \frac{2ad_{\mathbf{W}}}{mn}$$

with probability at least $1-2\{\max(p,mn)\}^{-1}$.

Lemma 1 is a direct consequence of Lemmas A.5 and A.6 in the Appendix. Here the term $\sqrt{\omega\log\{\max(p,mn)\}/(mn)}+(mn)^{-1}d_{E\mathbf{X}}$ is the order of $\|\partial\mathcal{L}(\boldsymbol{\Theta}_0,\boldsymbol{\beta}_0)/\partial\boldsymbol{\beta}\|_\infty$, while the remaining term on the right hand side represents the order of the error $\|\partial\mathcal{L}(\boldsymbol{\Theta}_0,\boldsymbol{\beta}_0)/\partial\boldsymbol{\beta} - \partial\mathcal{L}(\widehat{\boldsymbol{\Theta}},\boldsymbol{\beta}_0)/\partial\boldsymbol{\beta}\|_\infty$.

Theorem 1 *Assume Conditions (C1)–(C6) hold. Let*

$$\lambda_{\boldsymbol{\beta}} \geq 2\sqrt{\frac{\omega\log\{\max(p,mn)\}}{mn}} + \frac{2d_{E\mathbf{X}}}{mn} + \frac{4ad_{\mathbf{W}}}{mn}.$$

Then

$$\begin{aligned}\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|_2 \;\leq\; & \max\left(\left[\frac{10\sigma_{1F}}{\alpha_{0\boldsymbol{\beta}}\sqrt{mn}} + \left\{\frac{3\lambda_{\boldsymbol{\beta}}\sqrt{s}}{\alpha_{0\boldsymbol{\beta}}}\right\}^2\right]^{1/2} + 3\frac{\lambda_{\boldsymbol{\beta}}\sqrt{s}}{\alpha_{0\boldsymbol{\beta}}},\right. \\ & \left. 8a\sqrt{s}\gamma\sqrt{\frac{\log\{\max(p,mn)\}}{mn}}, (4\alpha_{0\boldsymbol{\beta}})^{-1/2}\left\{2\sigma_{1F}^2\frac{\log\{\max(p,mn)\}}{mn}\right\}^{1/4}\right) \quad (2)\end{aligned}$$

with probability at least $1-4\max(p,mn)^{-1}-2(mn)^{-1}-2\{\max(p,mn)\}^{-C}$ for some positive constant C, where $\sigma_{1F} = 32c_0^2a^2(d_{H_2}+d_{S_2}^2)$, and

$$\alpha_{0\boldsymbol{\beta}} \equiv \alpha_{\min}(E[R_{ij}E\{S_2^2(Y_{ij},\mathbf{X}_{ij}|\widehat{\boldsymbol{\Theta}},\boldsymbol{\beta}^*)\mathbf{X}_{ij}\mathbf{X}_{ij}^{\mathrm{T}}|Y_{ij}\} - R_{ij}E\{S_2(Y_{ij},\mathbf{X}_{ij}|\widehat{\boldsymbol{\Theta}},\boldsymbol{\beta}^*)\mathbf{X}_{ij}|Y_{ij}\}^{\otimes 2}])/4,$$

where $\alpha_{\min}(\mathbf{M})$ here and throughout the text is the minimal eigenvalue of a given matrix $\mathbf{M}$.

The proof of Theorem 1 is lengthy and is detailed in Appendix C and D. The convergence of $\widehat{\boldsymbol{\beta}}$ depends on many things, including the growing rate of p, m, n, the sparseness parameter s, the bounds $d_{H_2}, d_{S_2}, d_{E\mathbf{X}}, d_{\mathbf{W}}, c_0, a, \lambda_{\boldsymbol{\beta}}$ and the missingness related quantity $\alpha_{0\boldsymbol{\beta}}$. If s and $\alpha_{0\boldsymbol{\beta}}$ are finite, $d_{\mathbf{W}}$ and $d_{E\mathbf{X}}$ grow slower than mn, and $\log(p) = o(mn)$, together with the condition $c_0^2 a^2 (d_{H_2} + d_{S_2}^2)\sqrt{\log\{\max(p, mn)\}/(mn)} \to 0$ in Condition (C6), then the upper bound of $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2$ in Theorem 1 will go to zero as mn increases. It is worth mentioning that $\alpha_{0\boldsymbol{\beta}}$ always appears in the denominator of the upper bound of $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2$. It indicates that although $\text{pr}(R_{ij} = 1|Y_{ij})$ cannot be zero, it is allowed to go to zero at a certain rate. In fact, as long as its vanishing speed is sufficiently slow so that $\lambda_{\boldsymbol{\beta}}\sqrt{s}/\alpha_{0\boldsymbol{\beta}}$ and $\sigma_{1F}[\log\{\max(p, mn)\}/(mn)]^{1/2}/\alpha_{0\boldsymbol{\beta}}$ converge to zero, then the upper bound of $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2$ will still go to zero.

It is worth mentioning that, for the identifiability of the parameters, we do not allow a dense covariate matrix in the model. To see that, our Condition (C5) requires $\|\mathbf{X}_{ij}\|_1 \leq c_0$, where c_0 is a quantity that determines the growing rate of $\sigma_{1F} \equiv 32c_0^2 a^2 (d_{H_2} + d_{S_2}^2)$. As shown in Theorem 1, the upper bound of the estimation error satisfies (2) in probability. Thus, for $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2$ to vanish to zero, we would want σ_{1F} to grow slower than $\sqrt{(mn)/\log\{\max(p, mn)\}}$. Suppose we have bounded d_{H_2} and d_{S_2}, which is the case for logistic regression, then c_0^2 cannot grow faster than $\sqrt{(mn)/\log\{\max(p, mn)\}}$. This prevents us from having a dense covariate matrix. In fact, if the covariate vector is dense, it is likely to violate the identifiability Condition (C2) that $\sum_{k=1}^{p} \beta_{0k}\mathbf{X}^k$ is a sparse matrix.

In practice, it is very rare that people can collect high dimensional dense Gaussian matrices in the social media data. In the social media data, the majority of the features are categorical, such as gender, race, address, etc, which are named dictionary in the literature (See, for example, Robin et al. (2018)). Such features are then transformed to dummy variables and result in only very few nonzero values in the covariate matrices. Some features, such as address, can have a large number of categories, and lead to a series of high dimensional but highly sparse matrices $\mathbf{X}^k$'s formed by dummy variables. The matrix $\mathbf{X}^k$ thus has value one only at the entry corresponding to a specific user and a specific restaurant. In all the above situations, we are likely to have a sparse matrix $\sum_{k=1}^{p} \beta_{0k}\mathbf{X}^k$.

Our method can be applied to the majority of the social media data with the dictionary as the covariates, which is a very challenging problem even under missing-at-random settings (Robin et al., 2018). The proposed technique can be very useful in analyzing social media data.

Lemma 2 *Assume Conditions (C1)–(C6) hold. Let*

$$\lambda_{\boldsymbol{\beta}} \geq 2\sqrt{\frac{\omega\log\{\max(p, mn)\}}{mn}} + \frac{2d_{E\mathbf{X}}}{mn} + \frac{4ad_{\mathbf{W}}}{mn}.$$

Then for some $c_d, \gamma > 0$,

$$\left\|\frac{\partial\mathcal{L}(\boldsymbol{\Theta}_0, \widehat{\boldsymbol{\beta}})}{\partial\boldsymbol{\Theta}}\right\|_{op} \leq c_d\sqrt{\frac{d\log(d)}{mn}} + (2d_{H_2} + 2d_{S_2}^2)\max\left(\left[\frac{10\sigma_{1F}}{\alpha_{0\boldsymbol{\beta}}\sqrt{mn}} + \left\{\frac{3\lambda_{\boldsymbol{\beta}}\sqrt{s}}{\alpha_{0\boldsymbol{\beta}}}\right\}^2\right]^{1/2} + \frac{3\lambda_{\boldsymbol{\beta}}\sqrt{s}}{\alpha_{0\boldsymbol{\beta}}}, 8a\sqrt{s}\gamma\sqrt{\frac{\log\{\max(p, mn)\}}{mn}},$$

$$(4\alpha_{0\boldsymbol{\beta}})^{-1/2}\left\{\frac{2\sigma_{1F}^2\log\{\max(p,mn)\}}{mn}\right\}^{1/4}\Bigg)$$

with probability at least $1-1/d$.

This lemma is a direct consequence of Lemmas A.13 and A.14 in the Appendix. Here $c_d\sqrt{d\log(d)/(mn)}$ on the right hand side captures the order of $\|\partial\mathcal{L}(\boldsymbol{\Theta}_0,\boldsymbol{\beta}_0)/\partial\boldsymbol{\Theta}\|_\infty$ from Lemma A.13, and the remaining terms on the right hand side describe the order of $\|\partial\mathcal{L}(\boldsymbol{\Theta}_0,\boldsymbol{\beta}_0)/\partial\boldsymbol{\Theta}-\partial\mathcal{L}(\boldsymbol{\Theta}_0,\widehat{\boldsymbol{\beta}})/\partial\boldsymbol{\Theta}\|_\infty$. We further show that the estimator $\widehat{\boldsymbol{\Theta}}$ described in (1) converges to the true parameter value in probability as well.

Theorem 2 *Assume Conditions (C1)–(C6) hold. Let*

$$\begin{aligned}\lambda_{\boldsymbol{\Theta}} \;\geq\; & 2c_d\sqrt{\frac{d\log(d)}{mn}}+2(2d_{H_2}+2d_{S_2}^2)\max\Bigg(\left[\frac{10\sigma_{1F}}{\alpha_{0\boldsymbol{\beta}}\sqrt{mn}}+\left\{\frac{3\lambda_{\boldsymbol{\beta}}\sqrt{s}}{\alpha_{0\boldsymbol{\beta}}}\right\}^2\right]^{1/2}\\ & +\frac{3\lambda_{\boldsymbol{\beta}}\sqrt{s}}{\alpha_{0\boldsymbol{\beta}}},8a\sqrt{s}\gamma\sqrt{\frac{\log\{\max(p,mn)\}}{mn}},(4\alpha_{0\boldsymbol{\beta}})^{-1/2}\left\{\frac{2\sigma_{1F}^2\log\{\max(p,mn)\}}{mn}\right\}^{1/4}\Bigg).\end{aligned}$$

Then

$$\begin{aligned}\|\widehat{\boldsymbol{\Theta}}-\boldsymbol{\Theta}_0\|_F \;\leq\; & \max\Bigg(\left[\frac{137\sigma_{dF}}{32\alpha_{0\boldsymbol{\Theta}}\sqrt{mn}}+\frac{36\lambda_{\boldsymbol{\Theta}}^2 r}{\alpha_{0\boldsymbol{\Theta}}^2}\right]^{1/2}+\frac{6\lambda_{\boldsymbol{\Theta}}\sqrt{r}}{\alpha_{0\boldsymbol{\Theta}}},\\ & 16a\sqrt{r}\nu\sqrt{\frac{d\log(d)}{mn}},(4\alpha_{0\boldsymbol{\Theta}})^{-1/2}\left\{\frac{2\sigma_{dF}^2 d\log(d)}{mn}\right\}^{1/4}\Bigg)\end{aligned}$$

with probability at least $1-\exp\{-Cd\log(d)\}-2\exp\{-d\log(d)\}-d^{-1}$, *where* C *is a constant. Here* $\sigma_{dF}\equiv 32a^2(d_{H_2}+d_{S_2}^2)$ *and*

$$\begin{aligned}\alpha_{0\boldsymbol{\Theta}} \;\equiv\; & \alpha_{\min}\Bigg((mn)^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m}E[R_{ij}E\{S_2^2(Y_{ij},\mathbf{X}_{ij}|\boldsymbol{\Theta}^*,\widehat{\boldsymbol{\beta}})\mathrm{vec}(\mathbf{z}_{ij})\mathrm{vec}(\mathbf{z}_{ij})^{\mathrm{T}}|Y_{ij}\}\\ & -R_{ij}E\{S_2(Y_{ij},\mathbf{X}_{ij}|\boldsymbol{\Theta}^*,\widehat{\boldsymbol{\beta}})\mathrm{vec}(\mathbf{z}_{ij})|Y_{ij}\}^{\otimes 2}]\Bigg)/4.\end{aligned}$$

We provide the proof of Theorem 2 in Appendix E and F. The convergence of $\widehat{\boldsymbol{\Theta}}$ depends on the order of m,n,r,p. Assume that $\boldsymbol{\Theta}_0$ is a matrix with finite rank, $d\log(d)/\log\{\max(p,mn)\}\to\infty$, and $d\log(d)/(mn)\to 0$. Combining with the order of $\|\partial\mathcal{L}(\boldsymbol{\Theta}_0,\widehat{\boldsymbol{\beta}})/\partial\boldsymbol{\Theta}\|_{op}$ in Lemma 2, the upper bound of $\|\widehat{\boldsymbol{\Theta}}-\boldsymbol{\Theta}_0\|_F$ is of order $O_p\{d\log(d)/(mn)^{1/4}\}$. Similar to Theorem 1, we can see that $\alpha_{0\boldsymbol{\Theta}}$ always appears in the denominator in the upper bound of $\|\widehat{\boldsymbol{\Theta}}-\boldsymbol{\Theta}_0\|_2$. This implies $\mathrm{pr}(R_{ij}=1|Y_{ij})$ is allowed to go to zero as long as the vanishing speed is sufficiently slow so that $\lambda_{\boldsymbol{\Theta}}\sqrt{r}/\alpha_{0\boldsymbol{\Theta}}$ and $\sigma_{dF}\{d\log(d)/(mn)\}^{1/2}/\alpha_{0\boldsymbol{\Theta}}$ converge to zero.

Due to the high dimensionality of $\boldsymbol{\beta}$ and $\boldsymbol{\Theta}$, the convergence of $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\Theta}}$ is difficult to achieve. The sample version of the loss function can be non-convex, which may lead to non-unique solution of the estimation procedure. To show statistical convergence, we show

that the second derivative of the loss function with respect to the parameters satisfies the restricted eigenvalue conditions as shown in Lemmas A.10 and A.19 for $\boldsymbol{\beta}$ and $\boldsymbol{\Theta}$, respectively. These restricted eigenvalue conditions guarantee the convergence of our estimators in the feasible set. Unlike standard results, the convergence of $\widehat{\boldsymbol{\Theta}}$ depends on that of $\widehat{\boldsymbol{\beta}}$ and vice verse, and hence we require $\log\{\max(p, mn)\} = o_p(mn)$ and $d\log(d) = o_p(mn)$ to achieve the consistency of both $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\Theta}}$. The detailed theoretical derivations are in the Appendix. We first establish a series of necessary lemmas in Sections C and E of the Appendix, and then provide the detailed proofs in Sections D and F of the Appendix.

4. Computational Algorithm and Convergence

We now describe the computational algorithm and show its convergence rate theoretically. In the computational aspect, we treat the two parameters $\boldsymbol{\Theta}_0$ and $\boldsymbol{\beta}_0$ differently. In updating $\boldsymbol{\beta}$ we use the proximal gradient algorithm, and in updating $\boldsymbol{\Theta}$ we use soft-impute. We name the combination of the proximal gradient algorithm and the soft-impute treatment as the Soft Impute Proximal Gradient (SIPG) algorithm. In computing the integrals, we can view the integrations as mean and approximate it using sample averages across different $\mathbf{X}_{ij}$'s when this brings computational gain.

Remark 1 *Note that* $\partial\mathcal{L}(\boldsymbol{\Theta}, \boldsymbol{\beta})/\partial\boldsymbol{\beta} - \partial\mathcal{L}(\boldsymbol{\Theta}, \boldsymbol{\beta}')/\partial\boldsymbol{\beta} = \langle \partial\mathcal{L}(\boldsymbol{\Theta}, \boldsymbol{\beta}'')^2/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\beta} - \boldsymbol{\beta}' \rangle$, *by Conditions (C5) and (C6), where* $\boldsymbol{\beta}''$ *is a point on the line connecting* $\boldsymbol{\beta}$ *and* $\boldsymbol{\beta}'$. *Therefore,* $\|\partial\mathcal{L}(\boldsymbol{\Theta}, \boldsymbol{\beta})/\partial\boldsymbol{\beta} - \partial\mathcal{L}(\boldsymbol{\Theta}, \boldsymbol{\beta}')/\partial\boldsymbol{\beta}\|_2 \leq \sigma_{\boldsymbol{\beta}}\|\boldsymbol{\beta} - \boldsymbol{\beta}'\|_2$ *with* $\sigma_{\boldsymbol{\beta}} \equiv \{2c_0^2(d_{H_2} + d_{S_2}^2)\}$. *We say that* $\mathcal{L}(\boldsymbol{\Theta}, \boldsymbol{\beta})$ *is* $\sigma_{\boldsymbol{\beta}}$ *smooth with respect to* $\boldsymbol{\beta}$. *Similarity let* $\boldsymbol{\Theta}$ *and* $\boldsymbol{\Theta}'$ *be two* $m \times n$ *matrices, and* $\Theta'_{lk} = \Theta_{lk}$ *if* $(l, k) \neq (i, j)$, *and* $\Theta'_{ij} \neq \Theta_{ij}$. *It is easy to show that*

$$|\partial\mathcal{L}(\boldsymbol{\Theta}, \boldsymbol{\beta})/\partial\Theta_{ij} - \partial\mathcal{L}(\boldsymbol{\Theta}', \boldsymbol{\beta})/\partial\Theta_{ij}| \leq \sigma_{\boldsymbol{\Theta}}|\Theta_{ij} - \Theta'_{ij}|,$$

where $\sigma_{\boldsymbol{\Theta}} \equiv 2(d_{H_2} + d_{S_2}^2)$. *We say that* $\mathcal{L}(\boldsymbol{\Theta}, \boldsymbol{\beta})$ *is* $\sigma_{\boldsymbol{\Theta}}$ *smooth with respect to* $\boldsymbol{\Theta}$. *Similarly we can show that*

$$\|\partial\mathcal{L}(\boldsymbol{\Theta}, \boldsymbol{\beta})/\partial\boldsymbol{\beta} - \partial\mathcal{L}(\boldsymbol{\Theta}', \boldsymbol{\beta})/\partial\boldsymbol{\beta}\|_2 \leq \sigma_{\boldsymbol{\beta}\boldsymbol{\Theta}}\|\boldsymbol{\Theta} - \boldsymbol{\Theta}'\|_F,$$

for $\sigma_{\boldsymbol{\beta}\boldsymbol{\Theta}} \equiv 2c_0(d_{H_2} + d_{S_2}^2)$. *We say that* $\partial\mathcal{L}(\boldsymbol{\Theta}, \boldsymbol{\beta})/\partial\boldsymbol{\beta}$ *is* $\sigma_{\boldsymbol{\beta}\boldsymbol{\Theta}}$ *Lipschitz continuous.*

Using $\boldsymbol{\beta}^t, \boldsymbol{\Theta}^t$ to denote the corresponding estimators at the tth iteration, and assuming in the tth iteration $\boldsymbol{\beta}^{t-1}, \boldsymbol{\Theta}^{t-1}$ are given, we estimate $\boldsymbol{\beta}^t$ using the proximal gradient descent. That is, we obtain

$$\boldsymbol{\beta}^t = \mathrm{argmin}_{\boldsymbol{\beta}}\frac{1}{2}\|\boldsymbol{\beta} - \boldsymbol{\beta}^{t-1} + \eta\frac{\partial\mathcal{L}(\boldsymbol{\Theta}^{t-1}, \boldsymbol{\beta}^{t-1})}{\partial\boldsymbol{\beta}}\|_2^2 + \eta\lambda_{\boldsymbol{\beta}}\|\boldsymbol{\beta}\|_1$$

through letting

$$\boldsymbol{\beta}^t = \rho_{\eta\lambda_{\boldsymbol{\beta}}}\left(\boldsymbol{\beta}^{t-1} - \eta\frac{\partial\mathcal{L}\{\boldsymbol{\Theta}^{t-1}, \boldsymbol{\beta}^{t-1})\}}{\partial\boldsymbol{\beta}}\right),$$

where ρ_a is the component-wise soft thresholding operator so that the ith element of $\rho_a(\mathbf{x})$ is $x_i - a$ if $x_i > a$, $x_i + a$ if $x_i < -a$ or 0 if $|x_i| < a$. Here η is a step size. To update $\boldsymbol{\Theta}$, we

consider

$$\mathbf{\Theta}^t = \text{argmin}_{\mathbf{\Theta}} \frac{1}{2} \left\| \mathbf{\Theta} - \mathbf{\Theta}^{t-1} + \eta_1 \frac{\partial \mathcal{L}(\mathbf{\Theta}^{t-1}, \boldsymbol{\beta}^t)}{\partial \mathbf{\Theta}} \right\|_F^2 + \eta_1 \lambda_{\mathbf{\Theta}} \|\mathbf{\Theta}\|_*,$$

where η_1 is the step size, and use Soft-Impute (Mazumder et al., 2010) algorithm to implement the optimization. Specifically, let $\mathbf{UDV}^{\mathrm{T}}$ be the singular value decomposition of $\mathbf{\Theta}^{t-1} - \eta_1 \partial \mathcal{L}(\mathbf{\Theta}^{t-1}, \boldsymbol{\beta}^t)/\partial \mathbf{\Theta}$, we obtain

$$\mathbf{\Theta}^t = \mathbf{U}\text{Soft}_{\eta_1 \lambda_{\mathbf{\Theta}}}(\mathbf{D})\mathbf{V}^{\mathrm{T}},$$

where $\text{Soft}_{\eta_1 \lambda_{\mathbf{\Theta}}}(\cdot)$ operates element-wise on the diagonal matrix $\mathbf{D}$ by replacing D_{ii} with $(D_{ii} - \eta_1 \lambda_{\mathbf{\Theta}})_+$.

Let $F(\mathbf{\Theta}, \boldsymbol{\beta}) = \mathcal{L}(\mathbf{\Theta}, \boldsymbol{\beta}) - L + \lambda_{\mathbf{\Theta}} \|\mathbf{\Theta}\|_* + \lambda_{\boldsymbol{\beta}} \|\boldsymbol{\beta}\|_1$ for $|L| < \infty$ such that $\mathcal{L}(\mathbf{\Theta}, \boldsymbol{\beta}) > L$. In Theorem 3, we establish the convergence of the proposed SIPG algorithm.

To prepare for Theorem 3, define

$$\begin{aligned} g_{\boldsymbol{\beta}}(\mathbf{\Theta}, \boldsymbol{\beta}, Q) &= \max_{\|\widetilde{\boldsymbol{\beta}}\| \le Q} \langle \partial \mathcal{L}(\mathbf{\Theta}, \boldsymbol{\beta})/\partial \boldsymbol{\beta}, \boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}} \rangle + \lambda_{\boldsymbol{\beta}}(\|\boldsymbol{\beta}\|_1 - \|\widetilde{\boldsymbol{\beta}}\|_1), \\ g_{\mathbf{\Theta}}(\mathbf{\Theta}, \boldsymbol{\beta}, R) &= \max_{\|\widetilde{\mathbf{\Theta}}\|_* \le R} \langle \partial \mathcal{L}(\mathbf{\Theta}, \boldsymbol{\beta})/\partial \mathbf{\Theta}, \mathbf{\Theta} - \widetilde{\mathbf{\Theta}} \rangle + \lambda_{\mathbf{\Theta}}(\|\mathbf{\Theta}\|_* - \|\widetilde{\mathbf{\Theta}}\|_*). \end{aligned}$$

Let $Q^t = F(\mathbf{\Theta}^t, \boldsymbol{\beta}^t)/\lambda_{\boldsymbol{\beta}}$, $R^t = F(\mathbf{\Theta}^t, \boldsymbol{\beta}^{t+1})/\lambda_{\mathbf{\Theta}}$, $\sigma_{\boldsymbol{\beta}} = C_1(2d_{H_2} + 2d_{S_2}^2)c_0^2$ for $C_1 > 1$, $\sigma_{\mathbf{\Theta}} = 2d_{H_2} + 2d_{S_2}^2$.

Theorem 3 *Assume that for all $\boldsymbol{\beta}, \widetilde{\boldsymbol{\beta}}$ that satisfy $\|\boldsymbol{\beta}\|_\infty \le 2a$, $\|\widetilde{\boldsymbol{\beta}}\|_\infty \le a$ and all $\mathbf{\Theta}, \widetilde{\mathbf{\Theta}}$ that satisfy $\|\mathbf{\Theta}\|_{\max} \le 2a$, $\|\widetilde{\mathbf{\Theta}}\|_{\max} \le a$,*

$$\begin{aligned} \boldsymbol{\beta}^{\mathrm{T}} E\{\mathbf{F}_{\boldsymbol{\beta}}(\mathbf{X}, \mathbf{Y} | \widetilde{\mathbf{\Theta}}, \widetilde{\boldsymbol{\beta}})\}\boldsymbol{\beta} &\ge 4c_0 \sqrt{(d_{H_2} + d_{S_2}^2)\frac{\log\{\max(p, mn)\}}{mn}} \|\boldsymbol{\beta}\|_2^2, \\ \text{vec}(\mathbf{\Theta})^{\mathrm{T}} E\{\mathbf{F}_{\mathbf{\Theta}}(\mathbf{X}, \mathbf{Y} | \widetilde{\mathbf{\Theta}}, \widetilde{\boldsymbol{\beta}})\}\text{vec}(\mathbf{\Theta}) &\ge 4\sqrt{(d_{H_2} + d_{S_2}^2)\frac{\log(mn)}{mn}} \|\mathbf{\Theta}\|_F^2, \\ E\left(\begin{Bmatrix} \boldsymbol{\beta} \\ \text{vec}(\mathbf{\Theta}) \end{Bmatrix}^{\mathrm{T}} \frac{\partial^2 \mathcal{L}(\mathbf{\Theta}, \boldsymbol{\beta})}{\partial [\{\boldsymbol{\beta}^{\mathrm{T}}, \text{vec}(\mathbf{\Theta})^{\mathrm{T}}\}^{\mathrm{T}}]^{\otimes 2}} \begin{Bmatrix} \boldsymbol{\beta} \\ \text{vec}(\mathbf{\Theta}) \end{Bmatrix} \right) &\ge 4(c_0 + 1)\sqrt{(d_{H_2} + d_{S_2}^2)\frac{\log(mn)}{mn}} \left\| \begin{Bmatrix} \boldsymbol{\beta} \\ \text{vec}(\mathbf{\Theta}) \end{Bmatrix} \right\|_2^2 \end{aligned} \quad (3)$$

where the expectations are taken over $\mathbf{X}, \mathbf{Y}$. Select $0 < \eta < 1/\sigma_{\boldsymbol{\beta}}$ and select $0 < \eta_1 < 1/\sigma_{\mathbf{\Theta}}$. Then when

$$\sum_{t=0}^{T-1} 1/C(t) \ge \left\{ \frac{1}{\epsilon} - \frac{1}{F(\mathbf{\Theta}^0, \boldsymbol{\beta}^0) - F(\widehat{\mathbf{\Theta}}, \widehat{\boldsymbol{\beta}})} \right\} \quad (4)$$

for $C(t) = \max\{32(R^{t-1})^2/\eta_1 + 16\sigma_{\boldsymbol{\beta}\mathbf{\Theta}}^2 (Q^t)^2/(1/\eta_1 - \sigma_{\mathbf{\Theta}}), 32(Q^t)^2/\eta\}$, we have

$$F(\mathbf{\Theta}^T, \boldsymbol{\beta}^T) - F(\widehat{\mathbf{\Theta}}, \widehat{\boldsymbol{\beta}}) \le \epsilon$$

with probability at least $1 - 6(mn + p)^{-1}$. Hence $F(\mathbf{\Theta}^T, \boldsymbol{\beta}^T)$ is the ϵ-optimal solution for (1) with probability at least $1 - 6(mn + p)^{-1}$.

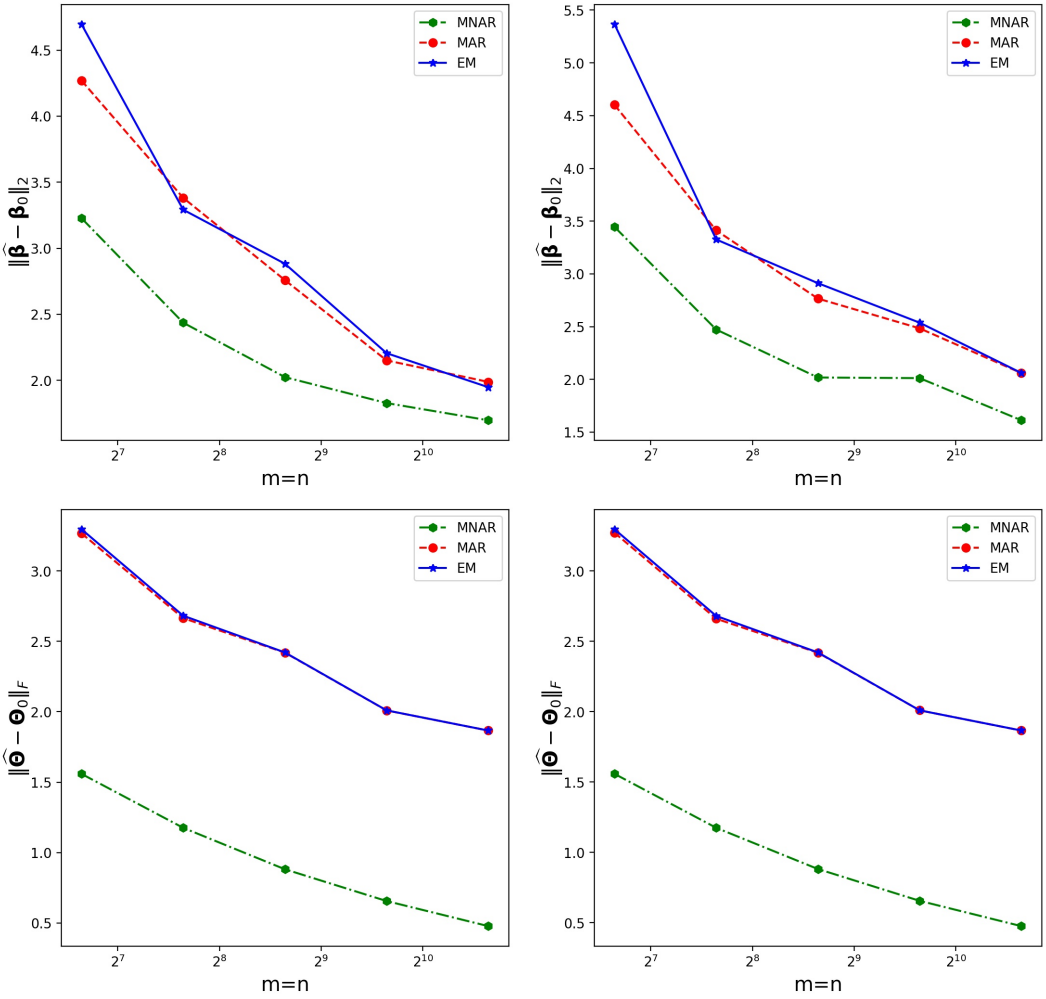

Figure 1: Linear Model : The averages of $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2$ and $\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_0\|_F$ of MNAR, EM and MAR over 50 simulations when $p = 50$ (left) and $p = 100$ (right).

The proof of Theorem 3 is given in Appendix G and H. Theorem 3 shows that when $mn + p \to \infty$, with probability approaching 1, the SIPG algorithm indeed converges to the optimizer described in (1) as long as sufficiently many iterations are carried out. Here the left hand side of (4) depends on the number of iterations T.

Remark 2 *In Theorem 3, (3) leads to the convexity of the objective function when the parameter search is implemented in the feasible set of* $\|\boldsymbol{\beta}\|_\infty \leq a$ *and* $\|\boldsymbol{\Theta}\|_{\max} \leq a$ *and satisfies (3). By the definition of* Q^t, R^t*, we have* $Q^t \leq Q^0$ *and* $R^t \leq R^0$ *for* $t \geq 1$*. Therefore,*

$$C(t) \leq \overline{C} \equiv \max\{32(R^0)^2/\eta_1 + 16\sigma^2_{\boldsymbol{\beta\Theta}}(Q^0)^2/(1/\eta_1 - \sigma_{\boldsymbol{\Theta}}), 32(Q^0)^2/\eta\},$$

i.e., $\sum_{t=0}^{T}\{1/C(t)\} \geq T\overline{C}^{-1}$. *Thus, a sufficient condition for (4) is*

$$T \geq \overline{C}\left\{\frac{1}{\epsilon} - \frac{1}{F(\boldsymbol{\Theta}^0, \boldsymbol{\beta}^0) - F(\widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\beta}})}\right\}.$$

This indicates that regardless what is the initial value, as long as the convexity on the feasible set holds, we can always perform sufficiently many iterations to obtain the ϵ-optimal solution with probability approaching to one. Of course, a closer initial value $(\boldsymbol{\Theta}^0, \boldsymbol{\beta}^0)$ to the solution will require fewer iterations.We point out that our conditions in (3) are very mild and are weaker than the standard literature. In fact, because the population version of $\mathcal{L}(\boldsymbol{\Theta}, \boldsymbol{\beta})$ is a convex function, the expectations in (3) are naturally nonnegative. In the existing literature (see, for example, the restrict eigenvalue condition in Loh and Wainwright (2012)), it is standard to further require the expectations to be strictly positive, i.e., the expectations in (3) are usually required to be larger than a positive constant. Here, because we allow the missing probability to go to 1, we relax the strictly positive condition by replacing the positive constant with a positive value that goes to zero.

5. Simulations

We perform three simulation studies to evaluate the MNAR method. The simulations are repeated 50 times. We first generate $X_{ijk} \sim \text{Bernoulli}(q)$ for $k = 1, \ldots, p$ independently, where $q = 0.2$. Thus, $g(\mathbf{x}_{ij}) = \prod_{k=1}^{p} q^{x_{ijk}}(1-q)^{1-x_{ijk}}$. Then we generate Y_{ij} from a linear model with mean $\boldsymbol{\Theta}_{0ij} + \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{X}_{ij}$ and standard deviation 5; We design $\boldsymbol{\Theta}_0$ to be a rank 5 matrix with singular values $(10, 1.8, 1.6, 1.4, 1.2)$, and $\boldsymbol{\beta}_0 = (1, 0, 2, 0, -3, -4, 5, 0, \ldots, 0)^{\mathrm{T}}$. We set $m = n$, and vary m, n from 100 to 1600. Furthermore, we generate R_{ij} from $\Pr(R_{ij} = 1|Y_{ij}) = \text{expit}(Y_{ij} - \bar{Y} - D)$, where $\bar{Y} = \sum_{i,j} Y_{ij}/(mn)$ and D is chosen to achieve 90% missingness in the data. As specified in Theorems 1 and 2, we select $\lambda_{\boldsymbol{\beta}} = C_{\boldsymbol{\beta}}\sqrt{\log\{\max(p, mn)\}/mn}$ and $\lambda_{\boldsymbol{\Theta}} = C_{\boldsymbol{\Theta}} \max\left\{\sqrt{\log(d)/d}, \log\{\max(p, mn)\}^{1/4}\sqrt{d}\right\}$, where $C_{\boldsymbol{\beta}}$ and $C_{\boldsymbol{\Theta}}$ are constants chosen to achieve similar sparseness of the estimators across all situations.

In Figure 1, we compare our method with the likelihood method which uses only the observed data and assume missing at random (referred to as MAR method), and the expectation–maximization (EM) method which imputes the missing outcomes by their expected values based on the previous estimators. The results show the MNAR method outperforms the EM and MAR methods consistently over all settings.

In the second simulation, we generate the response Y_{ij}'s from a logistic model with mean $\text{expit}(\boldsymbol{\Theta}_{0ij} + \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{X}_{ij})$, and we generate R_{ij} from $\Pr(R_{ij} = 1 \mid Y_{ij}) = 0.19I(Y_{ij} = 1) + 0.01I(Y_{ij} = 0)$, so the missing probability is around 0.90 marginally. Figure 2 shows that the MNAR method also outperforms the EM and MAR methods consistently under the logistic model.

Furthermore, we conduct additional simulations when the covariates are row (column)-specific, meaning that the covariates values are the same for the observations in the same column (row). We adopt the same simulation setting as the ones used in Figure 1, except that we change the covariate structure, where the first half covariates are row-specific while

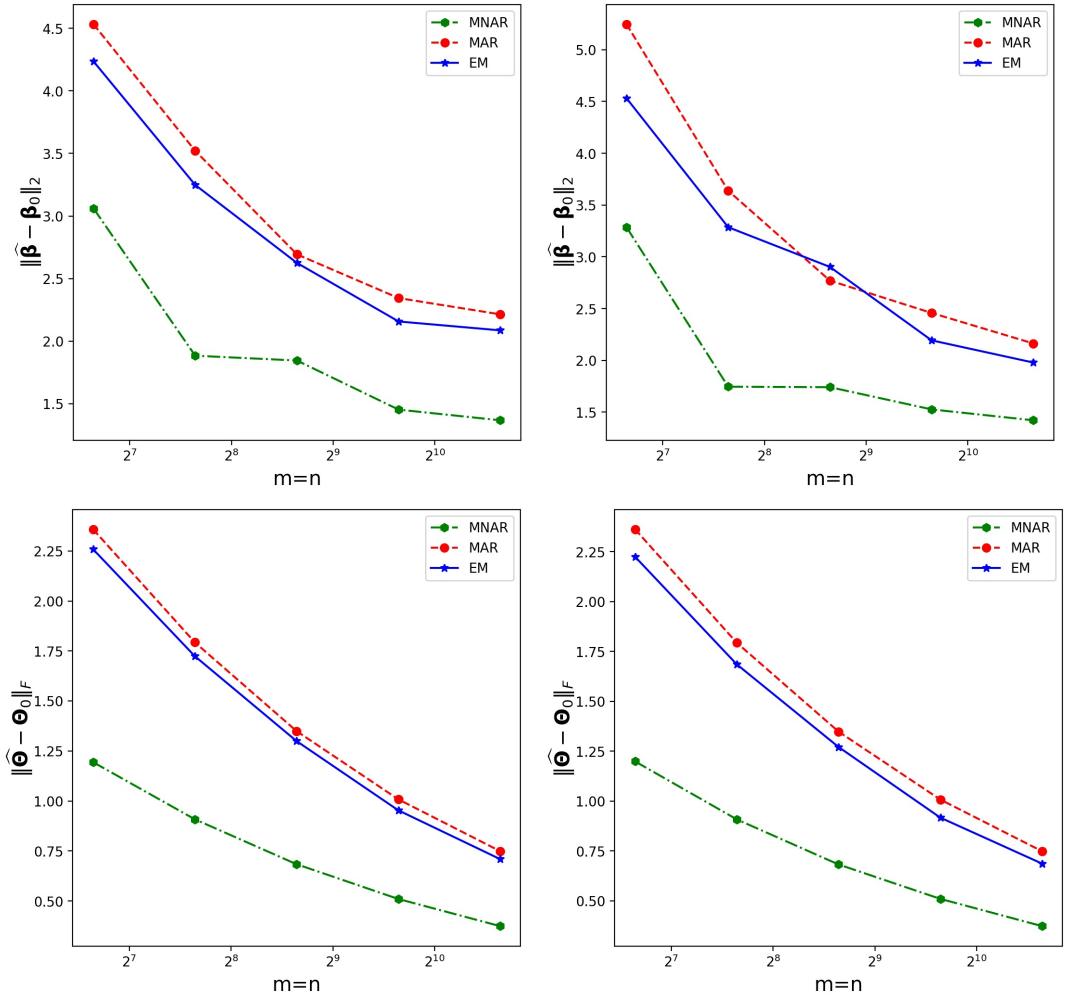

Figure 2: Logistic Model: The averages of $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2$ and $\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_0\|_F$ of MNAR, EM and MAR over 50 simulations when $p = 50$ (left) and $p = 100$ (right).

the last half covariates are columnn-specific. As shown in Figure 3, MNAR still outperforms the other two methods under this setting.

Lastly, to evaluate the performance of MNAR in larger sample settings, we generate data when $p = 2$, $\boldsymbol{\beta}_0 = (1, -2)^{\mathrm{T}}$ and vary m, n from 100 to 12800 with $m = n$. We compare the averages of $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2$ and $\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_0\|_F$ of MNAR, EM and MAR over 50 simulations in Figure 4 . The results show that the MNAR method outperforms the EM and MAR methods in this setting. The experiment also demonstrates the ability of MNAR to handle large data sets.

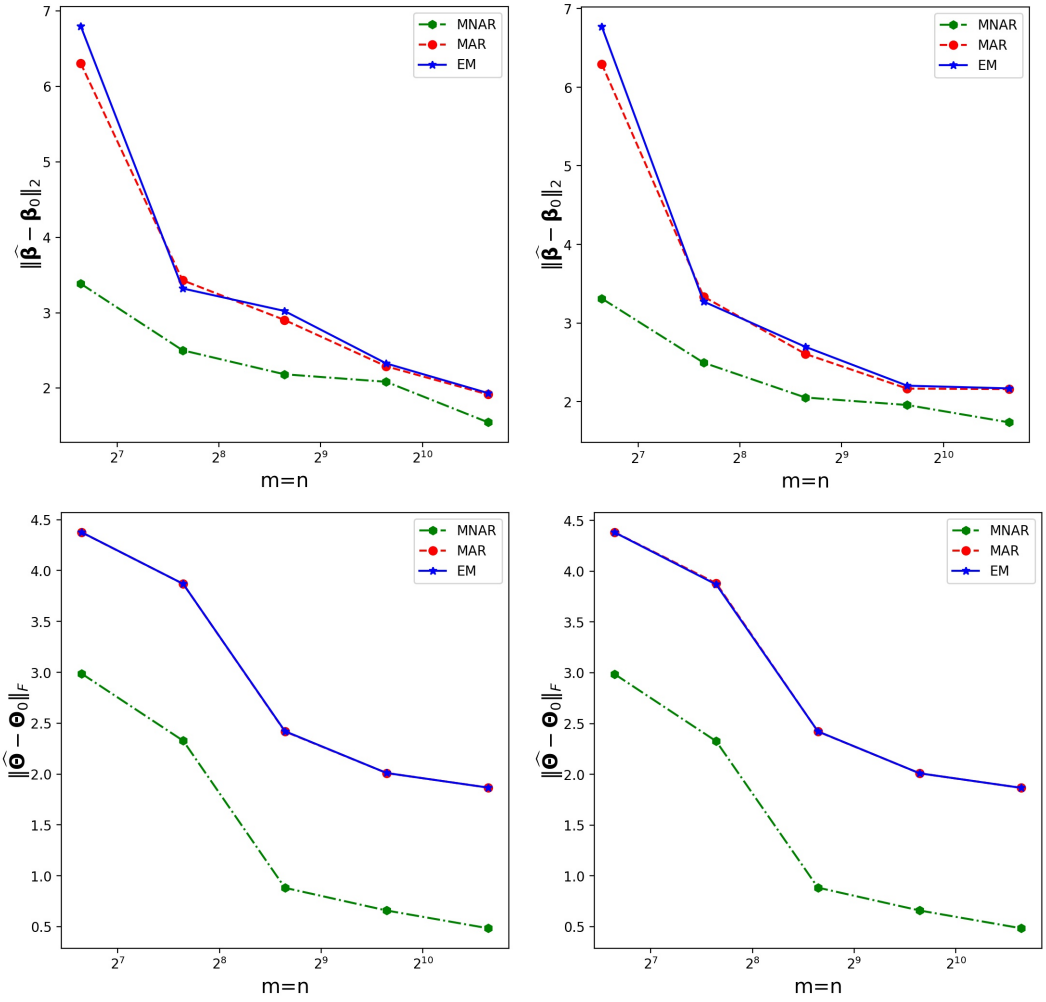

Figure 3: Row (column) specific covariates: The averages of $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2$ and $\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_0\|_F$ of MNAR, EM and MAR over 50 simulations when $p = 50$ (left) and $p = 100$ (right).

6. Real Data Analysis

We evaluate the performance of the MNAR, MAR, and EM methods on the real data from Yelp. Furthermore, we compare MNAR with a benchmark weighted collaborative filtering algorithm on the MovieLens (`https://grouplens.org/datasets/movielens/`) and the Yelp data.

6.1 Yelp Data Analysis

We apply the proposed method to analyze the data from Yelp, and the data set is available at `https://www.yelp.com/dataset/documentation/main`. The full data set is huge, which

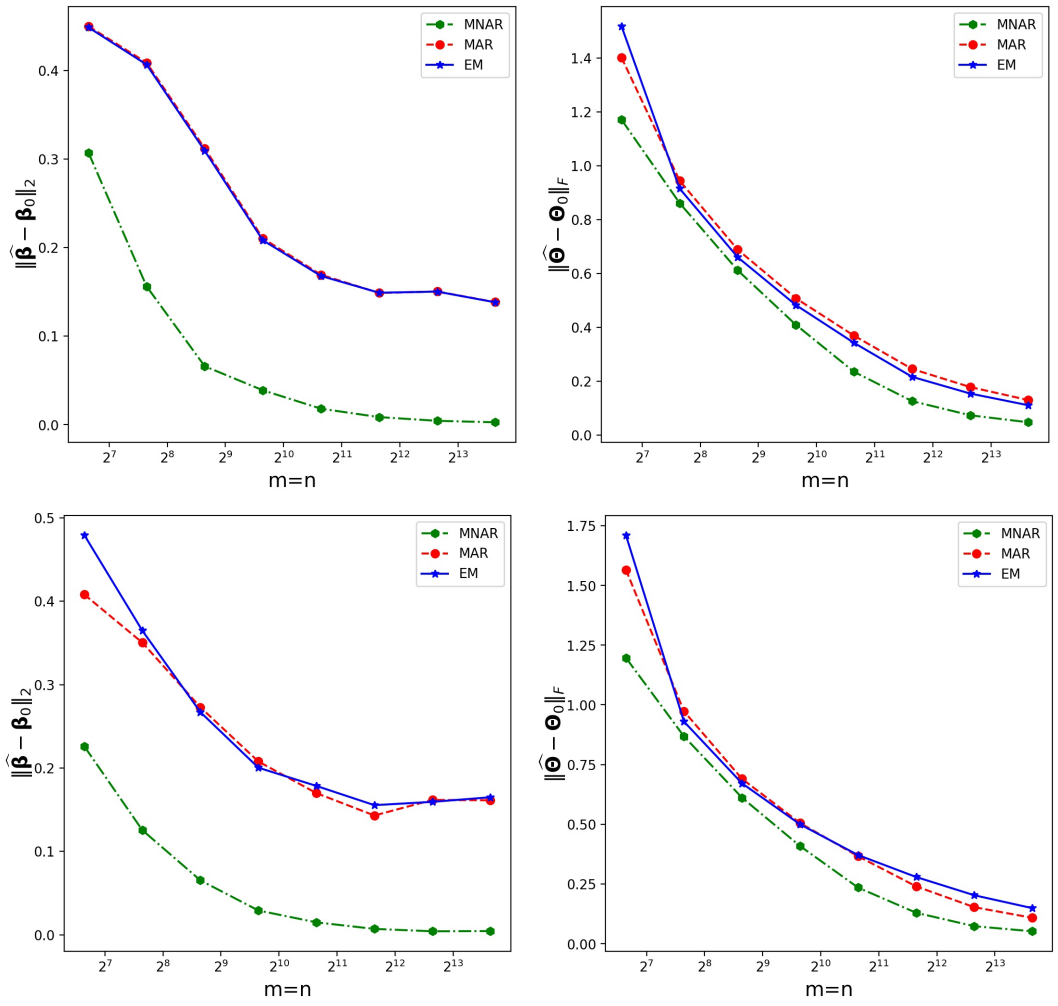

Figure 4: Large matrix size: The averages of $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2$ and $\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_0\|_F$ of MNAR, EM and MAR over 50 simulations under linear (upper) and logistic models (lower) when $p = 2$.

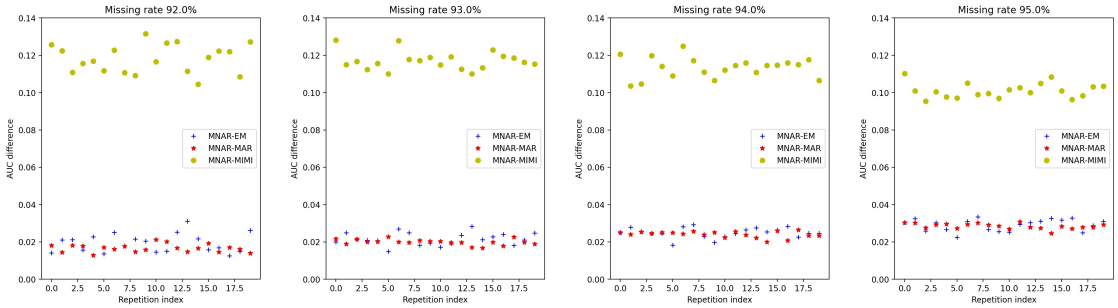

Figure 5: The AUC differences between MNAR and the other three models for 20 repetitions under different missing rates.

includes 8635403 reviews on 160585 businesses from 8 metropolitan areas. In combination with a large number of covariates, it leads to prohibitive computation. Thus, we first select 500 restaurants which received most ratings in Las Vegas. Then we choose 1000 users who gave most ratings for these 500 restaurants. The resulting $\mathbf{Y}$ matrix contains the reviews from 1000 ($n = 1000$) customers at 500 ($m = 500$) restaurants with 90.9% missingness rate. In addition, each Y_{ij}'s corresponds to a 22 ($p = 22$) dimensional covariates vector $\mathbf{X}_{ij}$, which includes baseline features of the ith customer and the jth restaurant such as the restaurant star, the open date, the customer review count, etc. We further standardize these covariates by subtracting the mean and dividing the standard deviation. We dichotomize the responses so that $Y_{ij} = 1$ if the review score from the ith customer at the jth restaurant is greater than 3.5. Furthermore, we introduce an evaluation procedure specifically for this missing not at random data set as follows. We first remove the observed Y_{ij}'s with probabilities p_1 and p_0 for $Y_{ij} = 1$ and $Y_{ij} = 0$, respectively, where p_0 and p_1 with $p_0 \neq p_1$ are chosen so that an additional $\alpha 100\%$ missingness is introduced into the data. Then we perform the proposed method on the remaining data, and use the estimation results to predict the removed entries. We perform the procedure 20 times, and compare the area under the receiver operating characteristic curve (AUC) of the MNAR, MAR, EM and MIMI methods proposed by Robin et al. (2020) under the settings with $\alpha = 0.011, 0.021, 0.031, 0.041$. In the implementation, we use the Monte Carlo method to approximate the integration in the loss function, while the distribution of $\mathbf{X}_{ij}$ is estimated empirically. Because $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_{ij}$ contributes to the conditional distribution of Y_{ij}, we re-estimate the distribution of $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_{ij}$ when a new $\boldsymbol{\beta}$ is obtained. At the tth iteration, we sample M copies of $\boldsymbol{\beta}^{t\mathrm{T}}\mathbf{X}_{ij}$ from the empirical distribution, and approximate the integration as

$$(mn)^{-1}\sum_{s=1}^{m}\sum_{u=1}^{n} f(Y_{ij}, \boldsymbol{\Theta}_{ij} + \boldsymbol{\beta}^{t\mathrm{T}}\mathbf{X}_{su}) \approx \int f(Y_{ij}, \boldsymbol{\Theta}_{ij} + \boldsymbol{\beta}^{t\mathrm{T}}\mathbf{X}) g(\mathbf{X}) d\mathbf{X}.$$

The penalty parameters in the four methods are selected by grid search, i.e., under each α, we generate a missing matrix $\mathbf{R}$ to tune the penalty parameters which yield largest AUC for the four methods.

We plot the 20 AUC differences between MNAR and the other three methods in Figure 5. In addition, we plot the average AUCs of the four methods over the 20 repetitions in

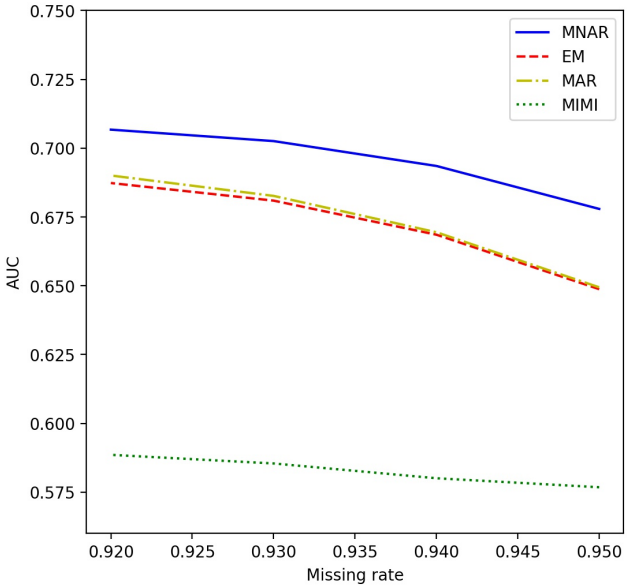

Figure 6: The means of AUC for the MNAR, EM, MAR and MIMI based on 20 repetitions under different missing rates.

Figure 6. The results show that MNAR clearly outperforms all competing methods with the largest AUC consistently across all settings.

6.2 Comparison with the weighted collaborative filtering method

Weighted collaborative filtering (WCF) is arguably the first approach that tackles the problem of nonignorable missing by assuming the missing data are the places where most negative feedbacks are expected to be found (Hu et al., 2008). We compare the performance of our method with that of the WCF method proposed in Hu et al. (2008) on MovieLens 1M data set, which includes one million ratings from 6040 users and 3952 movies. The data set also contains information on the age of a user and genre of a movie, which we use as the covariates in the MNAR method. There are 7 age groups and 302 different movie genres, so the covariate size is $p = 7 \times 302 = 2114$.

The missingness rate is over 95.7% in the full data. To evaluate the two methods, we first remove the observed Y_{ij}'s with probabilities p_1 and p_0 for $Y_{ij} = 5$ and $Y_{ij} \neq 5$ respectively, where p_0 and p_1 ($p_0 \neq p_1$) are chosen so that an additional 1% missingness is introduced into the training data. We evaluate the performance of the two methods using the expected percentile ranking proposed by Hu et al. (2008). Specifically, after obtaining the predicted ratings of all the movies by all the users, we obtain the percentile-ranking r_{ui} of movie i in all movies (including both rated and unrated movies) for user u.

Now let Y_{ui}^t be the observed rating of movie i by user u in the testing data set. Then the expected percentile ranking is calculated as

$$\bar{r} = \frac{\sum_{ui} Y_{ui}^t r_{ui}}{\sum_{ui} Y_{ui}^t}.$$

Note that a smaller $\bar{r}$ indicates better performance. The same procedure is repeated 20 times, and we provide the average of the 20 $\bar{r}$'s as the final performance measure.

For the WCF method, we obtain the weights using the BM25 function (Christopher et al., 2008) with two tuning parameters b, k_1 selected as $b = 0.75$ and $k_1 = 2$. This selection gives us the best performance among those recommended in Christopher et al. (2008) ($b = 0.75$ and $k_1 \in [1.2, 2.0]$). We also plot the mean of the expected percentile ranking over different choices of factors on 20 testing data sets in the left panel of Figure 7. The results show that WCF is sensitive to the selection of the number of factors with the optimal number of factors being 16 among all selections. Moreover, we plot the boxplots of the expected percentile rankings over the 20 testing data sets from the MNAR and WCF methods in the right panel of Figure 7. The results show that MNAR outperforms WCF with significantly smaller expected percentile rankings.

We also apply MNAR and WCF on the Yelp data set discussed in Section 6.1. We again plot the mean of the expected percentile ranking over different choices of factors in the left panel of Figure 8. The results show that the WCF is sensitive to the selection of the number of factors with the optimal number of factors being 4 among all selections. Moreover, we plot the boxplots of the expected percentile rankings from MNAR and WCF methods in the right panel of Figure 8. The results show that MNAR outperforms WCF with significantly smaller expected percentile rankings.

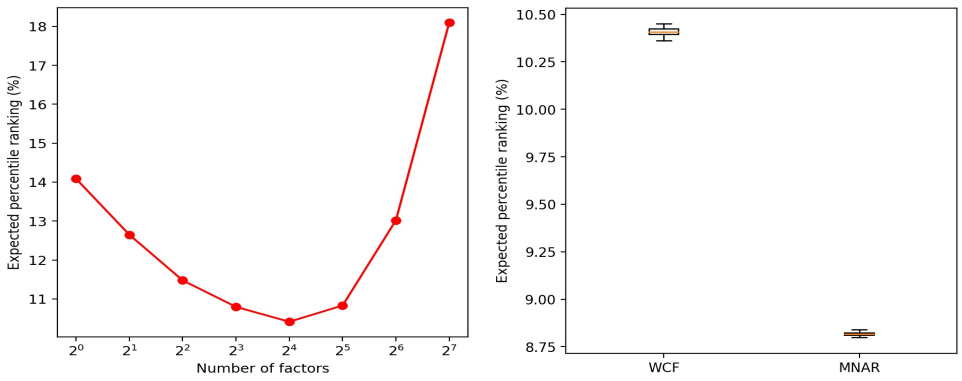

Figure 7: MovieLens data analysis: Left: The expected percentile ranking versus the number of factors from the WCF method. The expected percentile ranking decreases when the number of factors is less than 16, and starts to increase when the number of factors is greater than 16. Right: The boxplots of the expected percentile rankings from MNAR and WCF. MNAR has significantly smaller expected percentile rankings.

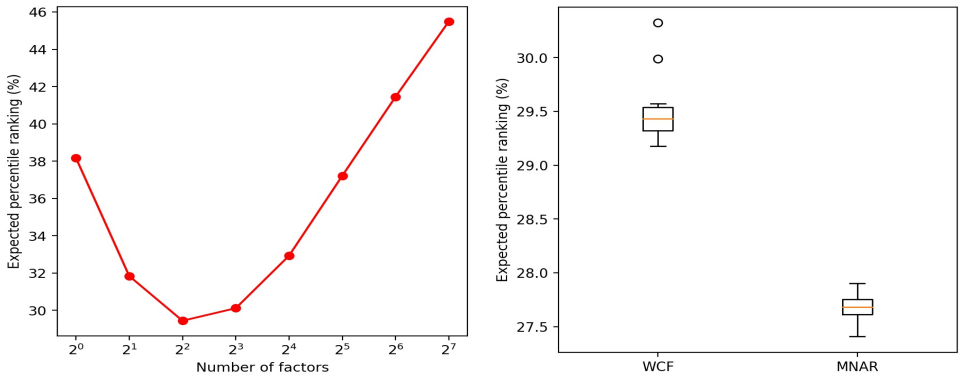

Figure 8: Yelp data analysis: Left: The expected percentile ranking versus the number of factors from the WCF method. The expected percentile ranking decreases when the number of factors is less than 4, and starts to increase when the number of factors is greater than 4. Right: The boxplots of the expected percentile rankings from MNAR and WCF methods. MNAR has significantly smaller expected percentile rankings.

7. Related work

Matrix completion problems have caught researchers' attention in recent years. The earlier works in this area do not consider any noise in the response (Candés and Recht, 2009; Recht, 2011), and only study the theoretical properties of perfect matrix completion. Noise issue is later taken into account in Candés and Plan (2010); Koltchinskii et al. (2011); Rohde and Tsybakov (2011); Negahban and Wainwright (2012), although they still do not include covariates or study missingness mechanism. In more recent works, nonuniform missing mechanism has also raised attention (Srebro and Salakhutdinov, 2010; Negahban and Wainwright, 2012; Klopp, 2014; Cai and Zhou, 2016; Cai et al., 2016; Bi et al., 2016; Mao et al., 2019). In addition, covariate information is taken into account. For example, Abernethy et al. (2009); Xu et al. (2013); Chiang et al. (2015); Mao et al. (2018) consider additional finite dimensional covariate effect; Zhu et al. (2016) cast the matrix completion problem in a sparse regression setting and estimate the high dimensional parameters with the conventional lasso type penalty; Robin et al. (2018) consider matrix completion with high dimensional covariate in the generalized linear model when the loss function is convex. Nevertheless, even the most sophisticated missingness feature studied in Mao et al. (2019) and the most flexible covariate structure studied in Robin et al. (2018) still assume a specific missingness mechanism model and require that the missingness does not depend on the potential response value given the covariates, i.e. they require modeling the missingness mechanism and they are limited to the framework of missing at random. These assumptions conveniently facilitate the parameter estimation, because the missingness does not affect the convexity of the loss function, and the estimation consistency can be easily achieved when the proportion of missing is not overwhelmingly large. However, this is restrictive and

is somewhat unrealistic in many customer evaluation system such as the Yelp restaurant evaluation problem considered here.

Missingness is usually classified into three categories, missing completely at random, missing at random and missing not at random. The first two classes have received extensive studies and a vast amount of literature is available. However, the third class, which is sometimes also referred to as informative or nonignorable missing, is much harder to treat and has not been well studied until very recent years. Some attempts are made, see, for example, Zhao and Shao (2015); Miao and Tchetgen Tchetgen (2016). The nonignorable missing problem has also caught attention in the collaborative filtering literature. Marlin and Zemel (2009), Hernández-Lobato et al. (2014), and Liang et al. (2016) discuss the collaborative prediction and ranking with nonignorable missing data while assuming the missing distributions are given; Hu et al. (2008) and Pan et al. (2008) tackle the problem of MNAR data in the case of implicit feedback-data; and Steck (2010) discusses how to evaluate the algorithm under MNAR settings. But none of these works directly predicts the product rating under the MNAR setting for the unknown missing distribution as we do, and none of them is in the high dimensional covariates setting as we consider here. Our work is in the third category, with the additional complexity of ultra-high dimensional covariates. To handle the nonignorable missing, we establish a penalized pseudo-likelihood framework, where despite of the goal of predicting Y_{ij} based on the covariate information $\mathbf{X}_{ij}$, we work with the reverse relation of $\mathbf{X}_{ij}$ given Y_{ij}, hence bypassing the difficulties caused by the nonignorable missingness. However, in the ultra-high dimensional covariate situation, this leads to a non-convexity issue even if the original full data likelihood is convex. We address and overcome these issues caused by the ultra-high dimension by simultaneously incorporating the low-rank restriction on the baseline evaluation matrix and the sparseness assumption on the potentially high dimensional covariates. Moreover, we develop efficient computational algorithms to obtain the low-rank and sparse parameter estimators. To the best of our knowledge, our work is the first complete framework to allow the missingness in the matrix to depend on the evaluation value that is itself missing, while simultaneously considering ultra-high dimensional covariates. The penalized pseudo-likelihood strategy may also have wide application in supervised learning settings when outcomes are missing with unknown mechanisms, or more generally when the sample is biased due to various sampling issues. We aim at tackling the complex case where both the high dimensional covariate effect and matrix completion are of interest. The proposed method is easily generalizable to simpler settings when only the covariate effect is of interests or only the matrix completion with low dimensional covariate effects is considered.

8. Conclusion

We have considered a matrix completion problem where not only the missingness mechanism of the entries in the matrix belongs to the missing not at random context, but also the covariate information is taken into account. The diverging size of the matrix is regularized through a rank constraint enforced through penalizing its nuclear norm, and the diverging dimension of the covariates is regularized through a sparsity assumption imposed via the L_1 penalization. It will be of interest to investigate if other practically justifiable constraints can be used to replace or enrich these assumptions. It will also be of interest to investigate

if better procedures can be developed if one is willing to make more concrete assumptions on the missingness mechanism. These are challenging problems but can be rewarding to study.

Acknowledgments

We are grateful to the referees and the editor for valuable comments and constructive suggestions. The authors are partially supported by grants from NSF and NIH.

Appendix

Appendix A. The Expression of Notations in Section 3.1

To facilitate the theoretic derivation, we write out the explicit expression for some important quantities. Recall the definition of $\mathbf{z}_{ij}$ defined in Section 3.1. Now

$$\begin{aligned}
&\frac{\partial \mathcal{L}(\boldsymbol{\Theta}, \boldsymbol{\beta})}{\partial \boldsymbol{\Theta}} \\
= \; & -(mn)^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m} R_{ij} \left[\left\{ \frac{f_2(Y_{ij}, \boldsymbol{\Theta}_{ij} + \boldsymbol{\beta}^{\mathrm{T}} \mathbf{X}_{ij})}{f(Y_{ij}, \boldsymbol{\Theta}_{ij} + \boldsymbol{\beta}^{\mathrm{T}} \mathbf{X}_{ij})} - \frac{\int f_2(Y_{ij}, \boldsymbol{\Theta}_{ij} + \boldsymbol{\beta}^{\mathrm{T}} \mathbf{X}) g(\mathbf{X}) d\mathbf{X}}{\int f(Y_{ij}, \boldsymbol{\Theta}_{ij} + \boldsymbol{\beta}^{\mathrm{T}} \mathbf{X}) g(\mathbf{X}) d\mathbf{X}} \right\} \mathbf{z}_{ij} \right]
\end{aligned}$$

and

$$\begin{aligned}
&\frac{\partial^2 \mathcal{L}(\boldsymbol{\Theta}, \boldsymbol{\beta})}{\partial \mathrm{vec}(\boldsymbol{\Theta}) \partial \mathrm{vec}(\boldsymbol{\Theta})^{\mathrm{T}}} \\
= \; & -(mn)^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m} R_{ij} \left[\left\{ \frac{f_{22}(Y_{ij}, \boldsymbol{\Theta}_{ij} + \boldsymbol{\beta}^{\mathrm{T}} \mathbf{X}_{ij})}{f(Y_{ij}, \boldsymbol{\Theta}_{ij} + \boldsymbol{\beta}^{\mathrm{T}} \mathbf{X}_{ij})} - \frac{f_2^2(Y_{ij}, \boldsymbol{\Theta}_{ij} + \boldsymbol{\beta}^{\mathrm{T}} \mathbf{X}_{ij})}{f^2(Y_{ij}, \boldsymbol{\Theta}_{ij} + \boldsymbol{\beta}^{\mathrm{T}} \mathbf{X}_{ij})} \right. \right. \\
& \left. \left. - \frac{\int f_{22}(Y_{ij}, \boldsymbol{\Theta}_{ij} + \boldsymbol{\beta}^{\mathrm{T}} \mathbf{X}) g(\mathbf{X}) d\mathbf{X}}{\int f(Y_{ij}, \boldsymbol{\Theta}_{ij} + \boldsymbol{\beta}^{\mathrm{T}} \mathbf{X}) g(\mathbf{X}) d\mathbf{X}} + \frac{\{\int f_2(Y_{ij}, \boldsymbol{\Theta}_{ij} + \boldsymbol{\beta}^{\mathrm{T}} \mathbf{X}) g(\mathbf{X}) d\mathbf{X}\}^2}{\left\{ \int f(Y_{ij}, \boldsymbol{\Theta}_{ij} + \boldsymbol{\beta}^{\mathrm{T}} \mathbf{X}) g(\mathbf{X}) d\mathbf{X} \right\}^2} \right\} \mathrm{vec}(\mathbf{z}_{ij})^{\otimes 2} \right].
\end{aligned}$$

Further, we have

$$\begin{aligned}
&\frac{\partial \mathcal{L}(\boldsymbol{\Theta}, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \\
= \; & -(mn)^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m} R_{ij} \left[\left\{ \frac{f_2(Y_{ij}, \boldsymbol{\Theta}_{ij} + \boldsymbol{\beta}^{\mathrm{T}} \mathbf{X}_{ij})}{f(Y_{ij}, \boldsymbol{\Theta}_{ij} + \boldsymbol{\beta}^{\mathrm{T}} \mathbf{X}_{ij})} \mathbf{X}_{ij} - \frac{\int f_2(Y_{ij}, \boldsymbol{\Theta}_{ij} + \boldsymbol{\beta}^{\mathrm{T}} \mathbf{X}) g(\mathbf{X}) \mathbf{X} d\mathbf{X}}{\int f(Y_{ij}, \boldsymbol{\Theta}_{ij} + \boldsymbol{\beta}^{\mathrm{T}} \mathbf{X}) g(\mathbf{X}) d\mathbf{X}} \right\} \right], \\
&\frac{\partial^2 \mathcal{L}(\boldsymbol{\Theta}, \boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^{\mathrm{T}}} \\
= \; & -(mn)^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m} R_{ij} \left[\left\{ \frac{f_{22}(Y_{ij}, \boldsymbol{\Theta}_{ij} + \boldsymbol{\beta}^{\mathrm{T}} \mathbf{X}_{ij})}{f(Y_{ij}, \boldsymbol{\Theta}_{ij} + \boldsymbol{\beta}^{\mathrm{T}} \mathbf{X}_{ij})} - \frac{f_2^2(Y_{ij}, \boldsymbol{\Theta}_{ij} + \boldsymbol{\beta}^{\mathrm{T}} \mathbf{X}_{ij})}{f^2(Y_{ij}, \boldsymbol{\Theta}_{ij} + \boldsymbol{\beta}^{\mathrm{T}} \mathbf{X}_{ij})} \right\} \mathbf{X}_{ij}^{\otimes 2} \right. \\
& \left. - \left\{ \frac{\int f_{22}(Y_{ij}, \boldsymbol{\Theta}_{ij} + \boldsymbol{\beta}^{\mathrm{T}} \mathbf{X}) g(\mathbf{X}) \mathbf{X}^{\otimes 2} d\mathbf{X}}{\int f(Y_{ij}, \boldsymbol{\Theta}_{ij} + \boldsymbol{\beta}^{\mathrm{T}} \mathbf{X}) g(\mathbf{X}) d\mathbf{X}} - \frac{\{\int f_2(Y_{ij}, \boldsymbol{\Theta}_{ij} + \boldsymbol{\beta}^{\mathrm{T}} \mathbf{X}) g(\mathbf{X}) \mathbf{X} d\mathbf{X}\}^{\otimes 2}}{\left\{ \int f(Y_{ij}, \boldsymbol{\Theta}_{ij} + \boldsymbol{\beta}^{\mathrm{T}} \mathbf{X}) g(\mathbf{X}) d\mathbf{X} \right\}^2} \right\} \right],
\end{aligned}$$

and

$$\begin{aligned}
&\frac{\partial^2 \mathcal{L}(\boldsymbol{\Theta}, \boldsymbol{\beta})}{\partial \mathrm{vec}(\boldsymbol{\Theta}) \partial \boldsymbol{\beta}^{\mathrm{T}}} \\
= \; & -(mn)^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m} R_{ij} \left[\left\{ \frac{f_{22}(Y_{ij}, \boldsymbol{\Theta}_{ij} + \boldsymbol{\beta}^{\mathrm{T}} \mathbf{X}_{ij})}{f(Y_{ij}, \boldsymbol{\Theta}_{ij} + \boldsymbol{\beta}^{\mathrm{T}} \mathbf{X}_{ij})} - \frac{f_2^2(Y_{ij}, \boldsymbol{\Theta}_{ij} + \boldsymbol{\beta}^{\mathrm{T}} \mathbf{X}_{ij})}{f^2(Y_{ij}, \boldsymbol{\Theta}_{ij} + \boldsymbol{\beta}^{\mathrm{T}} \mathbf{X}_{ij})} \right\} \mathrm{vec}(\mathbf{z}_{ij}) \mathbf{X}_{ij}^{\mathrm{T}} \right. \\
& - \left\{ \frac{\int f_{22}(Y_{ij}, \boldsymbol{\Theta}_{ij} + \boldsymbol{\beta}^{\mathrm{T}} \mathbf{X}) g(\mathbf{X}) \mathrm{vec}(\mathbf{z}_{ij}) \mathbf{X}^{\mathrm{T}} d\mathbf{X}}{\int f(Y_{ij}, \boldsymbol{\Theta}_{ij} + \boldsymbol{\beta}^{\mathrm{T}} \mathbf{X}) g(\mathbf{X}) d\mathbf{X}} \right. \\
& \left. \left. - \frac{\{\int f_2(Y_{ij}, \boldsymbol{\Theta}_{ij} + \boldsymbol{\beta}^{\mathrm{T}} \mathbf{X}) g(\mathbf{X}) d\mathbf{X} \mathrm{vec}(\mathbf{z}_{ij})\} \{\int f_2(Y_{ij}, \boldsymbol{\Theta}_{ij} + \boldsymbol{\beta}^{\mathrm{T}} \mathbf{X}) g(\mathbf{X}) \mathbf{X} d\mathbf{X}\}^{\mathrm{T}}}{\left\{ \int f(Y_{ij}, \boldsymbol{\Theta}_{ij} + \boldsymbol{\beta}^{\mathrm{T}} \mathbf{X}) g(\mathbf{X}) d\mathbf{X} \right\}^2} \right\} \right].
\end{aligned}$$

Let $\mathbf{M} = (\text{vec}(\boldsymbol{\Theta})^{\mathrm{T}}, \boldsymbol{\beta}^{\mathrm{T}})^{\mathrm{T}}$, we can write

$$\frac{\partial \mathcal{L}(\boldsymbol{\Theta}, \boldsymbol{\beta})}{\partial \mathbf{M}} = -(mn)^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m} R_{ij} \left[\mathbf{S}(Y_{ij}, \mathbf{X}_{ij} | \boldsymbol{\Theta}, \boldsymbol{\beta}) - E\{\mathbf{S}(Y_{ij}, \mathbf{X}_{ij} | \boldsymbol{\Theta}, \boldsymbol{\beta}) | Y_{ij}\} \right].$$

Furthermore because

$$\mathbf{H}(Y_{ij}, \mathbf{X}_{ij}, \boldsymbol{\Theta}, \boldsymbol{\beta}) = -\left\{ \frac{f_{22}(Y_{ij}, \boldsymbol{\Theta}_{ij} + \boldsymbol{\beta}^{\mathrm{T}} \mathbf{X}_{ij})}{f(Y_{ij}, \boldsymbol{\Theta}_{ij} + \boldsymbol{\beta}^{\mathrm{T}} \mathbf{X}_{ij})} - \frac{f_2^2(Y_{ij}, \boldsymbol{\Theta}_{ij} + \boldsymbol{\beta}^{\mathrm{T}} \mathbf{X}_{ij})}{f^2(Y_{ij}, \boldsymbol{\Theta}_{ij} + \boldsymbol{\beta}^{\mathrm{T}} \mathbf{X}_{ij})} \right\} [\{\text{vec}(\mathbf{z}_{ij})^{\mathrm{T}}, \mathbf{X}_{ij}^{\mathrm{T}}\}^{\mathrm{T}}]^{\otimes 2},$$

we get

$$\begin{aligned} E\{\mathbf{H}(Y_{ij}, \mathbf{X}_{ij} | \boldsymbol{\Theta}, \boldsymbol{\beta}) | Y_{ij}\} &= -\frac{\int f_{22}(Y_{ij}, \boldsymbol{\Theta}_{ij} + \boldsymbol{\beta}^{\mathrm{T}} \mathbf{X}) g(\mathbf{X}) \{(\text{vec}(\mathbf{z}_{ij})^{\mathrm{T}}, \mathbf{X}^{\mathrm{T}})^{\mathrm{T}}\}^{\otimes 2} d\mathbf{X}}{\int f(Y_{ij}, \boldsymbol{\Theta}_{ij} + \boldsymbol{\beta}^{\mathrm{T}} \mathbf{X}) g(\mathbf{X}) d\mathbf{X}} \\ &\quad + E\left\{ \frac{f_2^2(Y_{ij}, \boldsymbol{\Theta}_{ij} + \boldsymbol{\beta}^{\mathrm{T}} \mathbf{X})}{f^2(Y_{ij}, \boldsymbol{\Theta}_{ij} + \boldsymbol{\beta}^{\mathrm{T}} \mathbf{X})} \{(\text{vec}(\mathbf{z}_{ij})^{\mathrm{T}}, \mathbf{X}^{\mathrm{T}})^{\mathrm{T}}\}^{\otimes 2} | Y_{ij} \right\}, \end{aligned}$$

and

$$\begin{aligned} \frac{\partial^2 \mathcal{L}(\boldsymbol{\Theta}, \boldsymbol{\beta})}{\partial \mathbf{M} \partial \mathbf{M}^{\mathrm{T}}} &= (mn)^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m} R_{ij} \Big(\mathbf{H}(Y_{ij}, \mathbf{X}_{ij} | \boldsymbol{\Theta}, \boldsymbol{\beta}) - E\{\mathbf{H}(Y_{ij}, \mathbf{X}_{ij} | \boldsymbol{\Theta}, \boldsymbol{\beta}) | Y_{ij}\} \\ &\quad + E\left\{ \mathbf{S}(Y_{ij}, \mathbf{X}_{ij} | \boldsymbol{\Theta}, \boldsymbol{\beta})^{\otimes 2} | Y_{ij} \right\} - [E\{\mathbf{S}(Y_{ij}, \mathbf{X}_{ij} | \boldsymbol{\Theta}, \boldsymbol{\beta}) | Y_{ij}\}]^{\otimes 2} \Big). \end{aligned} \tag{5}$$

Now for any functions $h_1(R_{ij}), h_2(Y_{ij}, \mathbf{X}_{ij})$, because $\mathbf{X}_{ij}$ and R_{ij} are independent given Y_{ij}, we have

$$E\{h_1(R_{ij}) h_2(Y_{ij}, \mathbf{X}_{ij}) | Y_{ij}\} = E\{h_1(R_{ij}) | Y_{ij}\} E\{h_2(Y_{ij}, \mathbf{X}_{ij}) | Y_{ij}\}.$$

We also have

$$E\left(R_{ij} \left[\mathbf{H}(Y_{ij}, \mathbf{X}_{ij} | \boldsymbol{\Theta}, \boldsymbol{\beta}) - E\{\mathbf{H}(Y_{ij}, \mathbf{X}_{ij} | \boldsymbol{\Theta}, \boldsymbol{\beta}) | Y_{ij}\} \right] \right) = \mathbf{0}.$$

Combining with (5), we obtain that $E\{\partial^2 \mathcal{L}(\boldsymbol{\Theta}, \boldsymbol{\alpha}) / \partial \mathbf{M} \partial \mathbf{M}^{\mathrm{T}}\}$ is semi-positive definite. That is

$$\alpha_{\min}[E\{\partial^2 \mathcal{L}(\boldsymbol{\Theta}, \boldsymbol{\alpha}) / \partial \mathbf{M} \partial \mathbf{M}^{\mathrm{T}}\}] \geq 0, \tag{6}$$

where $\alpha_{\min}(\mathbf{A})$ is the smallest singular value of matrix $\mathbf{A}$. In addition, we write

$$\begin{aligned} W_{ijk}(\boldsymbol{\Theta}, \boldsymbol{\beta}) &\equiv \mathbf{e}_k^{\mathrm{T}} \frac{\partial \ell_{ij}(\boldsymbol{\Theta}, \boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \langle \boldsymbol{\Theta}, \mathbf{z}_{ij} \rangle} \\ &= \left[\left\{ \frac{f_{22}(Y_{ij}, \boldsymbol{\Theta}_{ij} + \boldsymbol{\beta}^{\mathrm{T}} \mathbf{X}_{ij})}{f(Y_{ij}, \boldsymbol{\Theta}_{ij} + \boldsymbol{\beta}^{\mathrm{T}} \mathbf{X}_{ij})} - \frac{f_2^2(Y_{ij}, \boldsymbol{\Theta}_{ij} + \boldsymbol{\beta}^{\mathrm{T}} \mathbf{X}_{ij})}{f^2(Y_{ij}, \boldsymbol{\Theta}_{ij} + \boldsymbol{\beta}^{\mathrm{T}} \mathbf{X}_{ij})} \right\} X_{ijk} \right. \\ &\quad - \left\{ \frac{\int f_{22}(Y_{ij}, \boldsymbol{\Theta}_{ij} + \boldsymbol{\beta}^{\mathrm{T}} \mathbf{X}) g(\mathbf{X}) X_k d\mathbf{X}}{\int f(Y_{ij}, \boldsymbol{\Theta}_{ij} + \boldsymbol{\beta}^{\mathrm{T}} \mathbf{X}) g(\mathbf{X}) d\mathbf{X}} \right. \\ &\quad \left. \left. - \frac{\{\int f_2(Y_{ij}, \boldsymbol{\Theta}_{ij} + \boldsymbol{\beta}^{\mathrm{T}} \mathbf{X}) g(\mathbf{X}) X_k d\mathbf{X}\} \{\int f_2(Y_{ij}, \boldsymbol{\Theta}_{ij} + \boldsymbol{\beta}^{\mathrm{T}} \mathbf{X}) g(\mathbf{X}) d\mathbf{X}\}}{\left\{ \int f(Y_{ij}, \boldsymbol{\Theta}_{ij} + \boldsymbol{\beta}^{\mathrm{T}} \mathbf{X}) g(\mathbf{X}) d\mathbf{X} \right\}^2} \right\} \right]. \end{aligned}$$

Appendix B. Some Lemmas

Lemma A.1 *Suppose the independent random variables $X_i, i = 1, \ldots, n$ are bounded in the interval $[a, b]$ with mean μ. Then X_i is a sub-Gaussian with parameter $\sigma = |b - a|$. Further,*

$$\Pr\left\{ |\sum_{i=1}^{n}(X_i - \mu)| \geq t \right\} \leq 2\exp\left\{ -\frac{t^2}{2n\sigma^2} \right\}.$$

Proof: This is the direct consequence of Example 2.3 and Proposition 2.1 in Chapter 2 of Wainwright (2019).

The lemma shows that a bounded variable is a sub-Gaussian distributed random variable.

Lemma A.2 *Consider the independent random variables $Y_1, \ldots, Y_n$ such that there are infinitely many constants u_i, v_i, $u_i \leq Y_i \leq v_i$, $i = 1, \ldots, n$. Let $Z = \sup_{\mathbf{t}\in\mathcal{T}} \sum_{i=1}^{n} t_i Y_i$, where $\mathcal{T}$ is a set of vectors $\mathbf{t} = (t_1, \ldots, t_n)$ and $\sigma = \sup_{\mathbf{t}\in\mathcal{T}} \{\sum_{i=1}^{n} t_i^2 (v_i - u_i)^2\}^{1/2} < \infty$. Let m_Z be the median of Z. Then for $\delta \geq 0$, we have*

$$\Pr(|Z - m_Z| \geq \delta) \leq 4\exp\{-\delta^2/(4\sigma^2)\}. \tag{7}$$

Furthermore

$$\begin{aligned} |E(Z) - m_Z| &\leq 4\sqrt{\pi}\sigma \\ var(Z) &\leq 16\sigma^2. \end{aligned} \tag{8}$$

Proof: This lemma is the Corollary 4.8 in Ledoux (2001).

The lemma demonstrates that a random variable with bounded variation has similar tail property as a sub-Gaussian distributed random variable does. It also provides bounds on the deviation from median and the distance between mean and median.

Lemma A.3 *Let $\mathbf{W}^i$ be independent $d_r \times d_c$ zero mean random matrix such that $\|\mathbf{W}^i\|_{op} \leq M$, and define*

$$\sigma_i^2 \equiv \max\left\{ \|E(\mathbf{W}^i \mathbf{W}^{i\mathrm{T}})\|_{op}, \|E(\mathbf{W}^{i\mathrm{T}} \mathbf{W}^i)\|_{op} \right\}$$

and $\sigma^2 = \sum_{i=1}^{n} \sigma_i^2$. Then we have

$$\Pr\left(\|\sum_{i=1}^{n} \mathbf{W}^i\|_{op} \geq t \right) \leq d_r d_c \max\left[\exp\{-t^2/(4\sigma^2)\}, \exp\{-t/(2M)\} \right].$$

Proof: This lemma follows Lemma 7 in Negahban and Wainwright (2012).

Lemma A.4 *(Hoeffding bound). Let $X_1, \ldots, X_N$ be independent centered sub-Gaussian random variables, let $K = \max_i \|X_i\|_{\psi_2}$, where $\|X_i\|_{\psi_2} \equiv \sup_{k\geq 1} k^{-1/2} E(|X|^k)^{1/k}$. is bounded for the sub-Gaussian random variable. Then for $t > 0$, we have*

$$\Pr\left\{ |\sum_{i=1}^{N} X_i| > t \right\} \leq 2\exp\left(-\frac{t^2}{2NK^2} \right).$$

Proof: This lemma is from Proposition 2.1 in Wainwright (2019).

The lemma demonstrates the tail property for the operation norm of the sum of random matrices.

Appendix C. Lemmas for Theorem 1

Define

$$\begin{aligned}\mathcal{C}(\gamma) &= \{\boldsymbol{\beta} \in \mathbf{R}^p : \|\boldsymbol{\beta}\|_\infty \|\boldsymbol{\beta}\|_1 / \|\boldsymbol{\beta}\|_2^2 \leq \gamma^{-1}\{mn/\log\{\max(p, mn)\}\}^{1/2}\},\\ \mathbb{B}(D) &= \{\boldsymbol{\Delta}_{\boldsymbol{\beta}} \in \mathcal{C}(\gamma) | \|\boldsymbol{\Delta}_{\boldsymbol{\beta}}\|_2 \leq D, \|\boldsymbol{\Delta}_{\boldsymbol{\beta}}\|_1 \leq D^2/\gamma\sqrt{mn/\log\{\max(p, mn)\}}, \|\boldsymbol{\Delta}_{\boldsymbol{\beta}}\|_\infty \leq a\}.\end{aligned}$$

Lemma A.5 *Assume Condition (C1) and (C3) hold, and let $\widehat{\boldsymbol{\Theta}}$ be the solution for (1), then*

$$\|\frac{\partial \mathcal{L}(\widehat{\boldsymbol{\Theta}}, \boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}} - \frac{\partial \mathcal{L}(\boldsymbol{\Theta}_0, \boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}}\|_\infty \leq 2a d_{\mathbf{W}} (mn)^{-1}$$

Proof: First note that by the mean value theorem, we have

$$\frac{\partial \mathcal{L}(\widehat{\boldsymbol{\Theta}}, \boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}} - \frac{\partial \mathcal{L}(\boldsymbol{\Theta}_0, \boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}} = (mn)^{-1} \sum_{i=1}^{n}\sum_{j=1}^{m} \frac{\partial^2 \mathcal{L}_{ij}(\boldsymbol{\Theta}^*, \boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}_0 \partial \langle \boldsymbol{\Theta}, \mathbf{z}_{ij}\rangle} \langle \mathbf{z}_{ij}, \widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_0\rangle,$$

where $\boldsymbol{\Theta}^*$ is the point on the line connecting $\widehat{\boldsymbol{\Theta}}$ and $\boldsymbol{\Theta}_0$. Hence

$$\begin{aligned}&\|\frac{\partial \mathcal{L}(\widehat{\boldsymbol{\Theta}}, \boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}} - \frac{\partial \mathcal{L}(\boldsymbol{\Theta}_0, \boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}}\|_\infty\\ =\ & \sup_k |(mn)^{-1} \sum_{i=1}^{n}\sum_{j=1}^{m} R_{ij} \langle W_{ijk}(\boldsymbol{\Theta}^*, \boldsymbol{\beta}_0)\mathbf{z}_{ij}, \widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_0\rangle|\\ =\ & \sup_k |(mn)^{-1} \langle \mathbf{R} \circ \mathbf{W}_k(\boldsymbol{\Theta}^*, \boldsymbol{\beta}_0), \widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_0\rangle|\\ \leq\ & (mn)^{-1} \sup_k \|\mathrm{vec}\{\mathbf{R} \circ \mathbf{W}_k(\boldsymbol{\Theta}^*, \boldsymbol{\beta}_0)\}\|_1 \|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_0\|_{\max}\\ \leq\ & 2a d_{\mathbf{W}} (mn)^{-1}.\end{aligned}$$

The last inequality holds by Condition (C1) and (C3). This proves the result.

Lemma A.6 *Assume Conditions (C1), (C4) and (C5) hold, there is a constant $\omega > 0$ such that*

$$\|\frac{\partial \mathcal{L}(\boldsymbol{\Theta}_0, \boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}}\|_\infty \leq \sqrt{\omega \log\{\max(p, mn)\}/(mn)} + (mn)^{-1} d_{E\mathbf{X}}$$

with probability at least $1 - 2\{\max(p, mn)\}^{-1}$.

Proof: Recall that $\mathbf{X}^k$ is the $n \times m$ matrix with the (i, j)th element X_{ijk}, and $S_2(Y_{ij}, \mathbf{X}_{ij}|\boldsymbol{\Theta}, \boldsymbol{\beta})$ is the partial derivative of $-\log\{f(Y_{ij}, \boldsymbol{\Theta}_{ij} + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_{ij})\}$ with respect to $\boldsymbol{\Theta}_{ij} + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_{ij}$. Let $\mathbf{S}_2(\mathbf{Y}, \mathbf{X}|\boldsymbol{\Theta}, \boldsymbol{\beta})$ be the $n \times m$ matrix with the (i, j)th element $S_2(Y_{ij}, \mathbf{X}_{ij}|\boldsymbol{\Theta}, \boldsymbol{\beta})$, we write

$$\begin{aligned}&\sup_k |\left\{\frac{\partial \mathcal{L}(\boldsymbol{\Theta}_0, \boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}}\right\}_k|\\ =\ & \sup_k |(mn)^{-1} \sum_{i=1}^{n}\sum_{j=1}^{m} R_{ij} \left[\left\{\frac{f_2(Y_{ij}, \boldsymbol{\Theta}_{0ij} + \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{X}_{ij})}{f(Y_{ij}, \boldsymbol{\Theta}_{0ij} + \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{X}_{ij})} X_{ijk} - \frac{\int f_2(Y_{ij}, \boldsymbol{\Theta}_{0ij} + \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{X}) g(\mathbf{X}) X_k d\mathbf{X}}{\int f(Y_{ij}, \boldsymbol{\Theta}_{0ij} + \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{X}) g(\mathbf{X}) d\mathbf{X}}\right\}\right]|\\ =\ & \sup_k |(mn)^{-1} \sum_{i=1}^{n}\sum_{j=1}^{m} R_{ij} S_2(Y_{ij}, \mathbf{X}_{ij}, \boldsymbol{\Theta}_0, \boldsymbol{\beta}_0) X_{ijk} - R_{ij} E\{S_2(Y_{ij}, \mathbf{X}|\boldsymbol{\Theta}_0, \boldsymbol{\beta}_0) X_k | Y_{ij}\}|\end{aligned}$$

$$\begin{aligned}
&\leq \sup_k |(mn)^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m} R_{ij}S_2(Y_{ij},\mathbf{X}_{ij}|\boldsymbol{\Theta}_0,\boldsymbol{\beta}_0)X_{ijk}| \\
&\quad + \sup_k |(mn)^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m} R_{ij}E\{S_2(Y_{ij},\mathbf{X}|\boldsymbol{\Theta}_0,\boldsymbol{\beta}_0)X_k|Y_{ij}\}| \\
&\leq \sup_k |(mn)^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m} R_{ij}S_2(Y_{ij},\mathbf{X}_{ij}|\boldsymbol{\Theta}_0,\boldsymbol{\beta}_0)X_{ijk}| + (mn)^{-1}d_{E\mathbf{X}}
\end{aligned}$$

Here we used Condition (C4) in step 4 and the last step. Now because $S_2(Y_{ij},\mathbf{X}_{ij}|\boldsymbol{\Theta}_0,\boldsymbol{\beta}_0)$ is sub-Gaussian and X_{ijk} are bounded as assumed in Conditions (C5), $E_{sij} = R_{ij}S_2(Y_{ij},\mathbf{X}_{ij}|\boldsymbol{\Theta}_0,\boldsymbol{\beta}_0)X_{ijk}$ is a sub-Gaussian random variable, by Lemma A.4, we have

$$\Pr\left(|(mn)^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m} E_{sij}| > t\right) \leq 2\exp(-2mnt^2/\omega)$$

for some constant $\omega > 0$. Therefore, we have

$$\begin{aligned}
&\Pr\left[\sup_k |(mn)^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m} R_{ij}S_2(Y_{ij},\mathbf{X}_{ij}|\boldsymbol{\Theta}_0,\boldsymbol{\beta}_0)X_{ijk}| \geq t\right] \\
\leq\ & p\Pr(|(mn)^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m} E_{sij}| > t) \\
\leq\ & 2\exp\{-2mnt^2/\omega + \log(p)\}.
\end{aligned}$$

Let $t = \sqrt{\omega\log\{\max(p,mn)\}/(mn)}$, we get

$$\Pr\left[\sup_k |(mn)^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m} R_{ij}S_2(Y_{ij},\mathbf{X}_{ij}|\boldsymbol{\Theta}_0,\boldsymbol{\beta}_0)X_{ijk}| \geq \sqrt{\omega\log\{\max(p,mn)\}/(mn)}\right] \leq 2\{\max(p,mn)\}^{-1}.$$

Plug in the above result to (9), we get

$$\left\|\frac{\partial\mathcal{L}(\boldsymbol{\Theta}_0,\boldsymbol{\beta}_0)}{\partial\boldsymbol{\beta}}\right\|_\infty \geq \sqrt{\omega\log\{\max(p,mn)\}/(mn)} + (mn)^{-1}d_{E\mathbf{X}}$$

with probability at most $2\{\max(p,mn)\}^{-1}$.

Lemma A.7 *Let $N(\delta)$ be the δ-covering number of $\mathbb{B}(D)$. Then there is a constant $c_1 > 0$ such that*

$$\log\{N(\delta)\} \leq \frac{9}{\delta^2}c_1^2D^4mn/\gamma^2.$$

Here, δ-covering number is defined as the number of disks with radius δ and center in $\mathbb{B}(D)$ needed to cover $\mathbb{B}(D)$.

Proof: Define $\mathbb{B}_1(D) = \{\boldsymbol{\Delta}_{\boldsymbol{\beta}} \in \mathbb{R}^p | \|\boldsymbol{\Delta}_{\boldsymbol{\beta}}\|_1 \leq D^2\sqrt{mn/\log\{\max(p,mn)\}}/\gamma\}$ and let $\widetilde{N}(\delta)$ be the δ-covering number of $\mathbb{B}_1(D)$. We have $\mathbb{B}(D) \subseteq \mathbb{B}_1(D)$ and $N(\delta) \leq \widetilde{N}(\delta)$. Now by the Sudakov minoration (Theorem 5.6 in Pisier (1999)) for a p-dimensional vector $\mathbf{G}$ containing independent identically distributed standard normal random variables $G_i, i = 1,\ldots,p$

$$\sqrt{\log\{\widetilde{N}(\delta)\}} \leq \frac{3}{\delta}E\left(\sup_{\|\boldsymbol{\Delta}_{\boldsymbol{\beta}}\|_1 \leq D^2\{mn/\log\{\max(p,mn)\}\}^{1/2}/\gamma} \langle\mathbf{G},\boldsymbol{\Delta}_{\boldsymbol{\beta}}\rangle\right)$$

$$\begin{aligned} &\leq \frac{3}{\delta} E(\max_{i\in[1,p]}|G_i|) D^2\sqrt{mn/\log(p)}/\gamma \\ &\leq \frac{3}{\delta} c_1 D^2\sqrt{\log(p)}\sqrt{mn/\log(p)}/\gamma \\ &= \frac{3}{\delta} c_1 D^2\sqrt{mn}/\gamma, \end{aligned}$$

for some constant $c_1 > 0$. The second line holds by the duality of L_1 and L_∞ norm. The third line holds because $E(\sup_{i\in[1,n]}|Z_i|) = O_p\{\sqrt{\log(n)}\}$ for an independent identically distributed sequence of normal random variables $Z_1, \ldots, Z_n$. Therefore, we have

$$\log\{N(\delta)\} \leq \frac{9}{\delta^2} c_1^2 D^4 mn/\gamma^2.$$

In the following we first show for $\widetilde{\boldsymbol{\Theta}}$ and $\widetilde{\boldsymbol{\beta}}$ in the feasible set that $\|\widetilde{\boldsymbol{\Theta}}\|_{\max} \leq a$ and $\|\widetilde{\boldsymbol{\beta}}\|_\infty \leq a$, we have

$$\begin{aligned} &\sup_{\boldsymbol{\beta}\in\mathbb{B}(D)} |\boldsymbol{\beta}^{\mathrm{T}}\mathbf{F}_{\boldsymbol{\beta}}(\mathbf{X},\mathbf{Y}|\widetilde{\boldsymbol{\Theta}},\widetilde{\boldsymbol{\beta}})\boldsymbol{\beta} - \boldsymbol{\beta}^{\mathrm{T}} E[R_{ij} E\{S_2^2(Y_{ij},\mathbf{X}_{ij}|\widetilde{\boldsymbol{\Theta}},\widetilde{\boldsymbol{\beta}})\mathbf{X}_{ij}\mathbf{X}_{ij}^{\mathrm{T}}|Y_{ij}\} \\ &-R_{ij} E\{S_2(Y_{ij},\mathbf{X}_{ij}|\widetilde{\boldsymbol{\Theta}},\widetilde{\boldsymbol{\beta}})\mathbf{X}_{ij}|Y_{ij}\}^{\otimes 2}]\boldsymbol{\beta}| \end{aligned}$$

is upper bounded by a $O(1)$ constants. Then we use the peeling arguments (Negahban and Wainwright, 2012) to show the boundedness of

$$\begin{aligned} &\sup_{\boldsymbol{\beta}\in\mathcal{C}(\gamma)} |\boldsymbol{\beta}^{\mathrm{T}}\mathbf{F}_{\boldsymbol{\beta}}(\mathbf{X},\mathbf{Y}|\widetilde{\boldsymbol{\Theta}},\widetilde{\boldsymbol{\beta}})\boldsymbol{\beta} - \boldsymbol{\beta}^{\mathrm{T}} E[R_{ij} E\{S_2^2(Y_{ij},\mathbf{X}_{ij}|\widetilde{\boldsymbol{\Theta}},\widetilde{\boldsymbol{\beta}})\mathbf{X}_{ij}\mathbf{X}_{ij}^{\mathrm{T}}|Y_{ij}\} \\ &-R_{ij} E\{S_2(Y_{ij},\mathbf{X}_{ij}|\widetilde{\boldsymbol{\Theta}},\widetilde{\boldsymbol{\beta}})\mathbf{X}_{ij}|Y_{ij}\}^{\otimes 2}]\boldsymbol{\beta}|. \end{aligned}$$

Lemma A.8 and Lemma A.9 are the auxiliary results to establish the first relation.

Lemma A.8 *Assume Conditions (C1), (C5), (C6). Let $\delta = D/\xi$, where ξ is an absolute constant, and c_1 is defined in Lemma A.7. Let $\boldsymbol{\beta} \in \mathbb{B}(D)$, and $\boldsymbol{\beta}^k, k = 1, \ldots N(\delta)$ be a δ-covering of $\mathbb{B}(D)$ in L_2 norm. By definition, given an arbitrary $\boldsymbol{\beta} \in \mathbb{B}(D)$, there is some index $k \in \{1, \ldots, N(\delta)\}$ and a difference matrix $\boldsymbol{\Delta}_{\boldsymbol{\beta}}$ with $\|\boldsymbol{\Delta}_{\boldsymbol{\beta}}\|_2 \leq \delta$ such that $\boldsymbol{\beta} = \boldsymbol{\beta}^k + \boldsymbol{\Delta}_{\boldsymbol{\beta}}$. Recall that for $\widetilde{\boldsymbol{\Theta}}$ and $\widetilde{\boldsymbol{\beta}}$ in the feasible set that $\|\widetilde{\boldsymbol{\Theta}}\|_{\max} \leq a$ and $\|\widetilde{\boldsymbol{\beta}}\|_\infty \leq a$*

$$\begin{aligned} &\mathbf{a}^{\mathrm{T}}\mathbf{F}_{\boldsymbol{\beta}}(\mathbf{X},\mathbf{Y}|\widetilde{\boldsymbol{\Theta}},\widetilde{\boldsymbol{\beta}})\mathbf{b} \\ = \;& (mn)^{-1}\left(\sum_{i=1}^n\sum_{j=1}^m \mathbf{a}^{\mathrm{T}} R_{ij}[H_2(Y_{ij},\mathbf{X}_{ij}|\widetilde{\boldsymbol{\Theta}},\widetilde{\boldsymbol{\beta}})\mathbf{X}_{ij}\mathbf{X}_{ij}^{\mathrm{T}} - E\{H_2(Y_{ij},\mathbf{X}_{ij}|\widetilde{\boldsymbol{\Theta}},\widetilde{\boldsymbol{\beta}})\mathbf{X}_{ij}\mathbf{X}_{ij}^{\mathrm{T}}|Y_{ij}\}\right. \\ &\left. +E\{S_2^2(Y_{ij},\mathbf{X}_{ij}|\widetilde{\boldsymbol{\Theta}},\widetilde{\boldsymbol{\beta}})\mathbf{X}_{ij}\mathbf{X}_{ij}^{\mathrm{T}}|Y_{ij}\} - E\{S_2(Y_{ij},\mathbf{X}_{ij}|\widetilde{\boldsymbol{\Theta}},\widetilde{\boldsymbol{\beta}})\mathbf{X}_{ij}|Y_{ij}\}^{\otimes 2}]\mathbf{b}\right). \end{aligned}$$

Then there are positive constants c_2, c_3 such that

$$\begin{aligned} &\Pr\left(\sup_{k=1,\ldots,N(\delta)} |\boldsymbol{\beta}^{k^{\mathrm{T}}}\mathbf{F}_{\boldsymbol{\beta}}(\mathbf{X},\mathbf{Y}|\widetilde{\boldsymbol{\Theta}},\widetilde{\boldsymbol{\beta}})\boldsymbol{\beta}^k - \boldsymbol{\beta}^{k^{\mathrm{T}}} E[R_{ij} E\{S_2^2(Y_{ij},\mathbf{X}_{ij}|\widetilde{\boldsymbol{\Theta}},\widetilde{\boldsymbol{\beta}})\mathbf{X}_{ij}\mathbf{X}_{ij}^{\mathrm{T}}|Y_{ij}\}\right. \\ &-R_{ij} E\{S_2(Y_{ij},\mathbf{X}_{ij}|\widetilde{\boldsymbol{\Theta}},\widetilde{\boldsymbol{\beta}})\mathbf{X}_{ij}|Y_{ij}\}^{\otimes 2}]\boldsymbol{\beta}^k| \\ &\left.\geq \delta^2 + \frac{32(d_{H_2} + d_{S_2}^2)a^2c_0^2}{\sqrt{mn}}\right) \\ \leq \;& 4\exp\left(-c_2 D^2 mn\right), \end{aligned}$$

$$
\begin{aligned}
&\Pr\Bigg(\sup_{k=1,\ldots,N(\delta)} |{\boldsymbol\beta^k}^{\mathrm{T}}\mathbf{F}_{\boldsymbol\beta}(\mathbf{X},\mathbf{Y}|\widetilde{\boldsymbol\Theta},\widetilde{\boldsymbol\beta})(\boldsymbol\beta-\boldsymbol\beta^k)-\boldsymbol\beta^{k\mathrm{T}}E[R_{ij}E\{S_2^2(Y_{ij},\mathbf{X}_{ij}|\widetilde{\boldsymbol\Theta},\widetilde{\boldsymbol\beta})\mathbf{X}_{ij}\mathbf{X}_{ij}^{\mathrm{T}}|Y_{ij}\}\\
&-R_{ij}E\{S_2(Y_{ij},\mathbf{X}_{ij}|\widetilde{\boldsymbol\Theta},\widetilde{\boldsymbol\beta})\mathbf{X}_{ij}|Y_{ij}\}^{\otimes 2}](\boldsymbol\beta-\boldsymbol\beta^k)|\\
&\geq \delta^2+\frac{64(d_{H_2}+d_{S_2}^2)a^2c_0^2}{\sqrt{mn}}\Bigg)\\
\leq\ & 4\exp\left(-c_3D^2mn\right).
\end{aligned}
$$

Proof: First we write

$$
\begin{aligned}
&|{\boldsymbol\beta^k}^{\mathrm{T}}R_{ij}[H_2(Y_{ij},\mathbf{X}_{ij}|\widetilde{\boldsymbol\Theta},\widetilde{\boldsymbol\beta})\mathbf{X}_{ij}\mathbf{X}_{ij}^{\mathrm{T}}-E\{H_2(Y_{ij},\mathbf{X}|\widetilde{\boldsymbol\Theta},\widetilde{\boldsymbol\beta})\mathbf{X}\mathbf{X}^{\mathrm{T}}|Y_{ij}\}\\
&+E\{S_2^2(Y_{ij},\mathbf{X}|\widetilde{\boldsymbol\Theta},\widetilde{\boldsymbol\beta})\mathbf{X}\mathbf{X}^{\mathrm{T}}|Y_{ij}\}-E\{S_2(Y_{ij},\mathbf{X}|\widetilde{\boldsymbol\Theta},\widetilde{\boldsymbol\beta})\mathbf{X}|Y_{ij}\}^{\otimes 2}]\boldsymbol\beta^k|\\
\leq\ & \sup_{\mathbf{X}_{ij}}\{2d_{H_2}\|\boldsymbol\beta^k\|_\infty\|\mathbf{X}_{ij}\|_1\|\boldsymbol\beta^k\|_\infty\|\mathbf{X}_{ij}\|_1+2d_{S_2}^2\|\boldsymbol\beta^k\|_\infty\|\mathbf{X}_{ij}\|_1\|\boldsymbol\beta^k\|_\infty\|\mathbf{X}_{ij}\|_1\}\\
=\ & 2(d_{H_2}+d_{S_2}^2)(ac_0)^2.
\end{aligned}
$$

Also we have

$$
\begin{aligned}
&|{\boldsymbol\beta^k}^{\mathrm{T}}R_{ij}[H_2(Y_{ij},\mathbf{X}_{ij}|\widetilde{\boldsymbol\Theta},\widetilde{\boldsymbol\beta})\mathbf{X}_{ij}\mathbf{X}_{ij}^{\mathrm{T}}-E\{H_2(Y_{ij},\mathbf{X}|\widetilde{\boldsymbol\Theta},\widetilde{\boldsymbol\beta})\mathbf{X}\mathbf{X}^{\mathrm{T}}|Y_{ij}\}\\
&+E\{S_2^2(Y_{ij},\mathbf{X}|\widetilde{\boldsymbol\Theta},\widetilde{\boldsymbol\beta})\mathbf{X}\mathbf{X}^{\mathrm{T}}|Y_{ij}\}-E\{S_2(Y_{ij},\mathbf{X}|\widetilde{\boldsymbol\Theta},\widetilde{\boldsymbol\beta})\mathbf{X}|Y_{ij}\}^{\otimes 2}]\boldsymbol\beta^k|\\
\leq\ & \sup_{\mathbf{X}_{ij}}\{2d_{H_2}\|\boldsymbol\beta^k\|_2\|\mathbf{X}_{ij}\|_2\|\boldsymbol\beta^k\|_\infty\|\mathbf{X}_{ij}\|_1+2d_{S_2}^2\|\boldsymbol\beta^k\|_2\|\mathbf{X}_{ij}\|_2\|\boldsymbol\beta^k\|_\infty\|\mathbf{X}_{ij}\|_1\}\\
=\ & 2(d_{H_2}+d_{S_2}^2)(ac_0^2)D,
\end{aligned}
$$

by Conditions (C1), (C5) and (C6). We now use Lemma A.2, where we treat each summand in ${\boldsymbol\beta^k}^{\mathrm{T}}\mathbf{F}_{\boldsymbol\beta}(\mathbf{X},\mathbf{Y}|\widetilde{\boldsymbol\Theta},\widetilde{\boldsymbol\beta})\boldsymbol\beta^k$ as Y_i in Lemma A.2, and set $(u_i,v_i)=\{-2(d_{H_2}+d_{S_2}^2)ac_0^2D,2(d_{H_2}+d_{S_2}^2)ac_0^2D\}$ and $(u_{1i},v_{1i})=\{-2(d_{H_2}+d_{S_2}^2)a^2c_0^2,2(d_{H_2}+d_{S_2}^2)a^2c_0^2\}$. Consider $\mathcal{T}$ to contain only one vector $\mathbf{t}$, where $\mathbf{t}$ is the mn dimensional vector with element $(mn)^{-1}$. Then Z in Lemma A.2 is ${\boldsymbol\beta^k}^{\mathrm{T}}\mathbf{F}_{\boldsymbol\beta}(\mathbf{X},\mathbf{Y}|\widetilde{\boldsymbol\Theta},\widetilde{\boldsymbol\beta})\boldsymbol\beta^k$ and we set σ in Lemma A.2 (7) as $\sigma=[\sum_{i=1}^n\sum_{j=1}^m(mn)^{-2}\{4(d_{H_2}+d_{S_2}^2)ac_0^2D\}^2]^{1/2}=4(d_{H_2}+d_{S_2}^2)ac_0^2D/\sqrt{mn}$, while we set σ in (8) as $\sigma_1=[\sum_{i=1}^n\sum_{j=1}^m(mn)^{-2}\{4(d_{H_2}+d_{S_2}^2)a^2c_0^2\}^2]^{1/2}=4(d_{H_2}+d_{S_2}^2)a^2c_0^2/\sqrt{mn}$. Let m_F be the median of ${\boldsymbol\beta^k}^{\mathrm{T}}F_{\boldsymbol\beta}(\mathbf{X},\mathbf{Y}|\widetilde{\boldsymbol\Theta},\widetilde{\boldsymbol\beta})\boldsymbol\beta^k$, we have

$$
\begin{aligned}
&\Pr\left(|{\boldsymbol\beta^k}^{\mathrm{T}}\mathbf{F}_{\boldsymbol\beta}(\mathbf{X},\mathbf{Y}|\widetilde{\boldsymbol\Theta},\widetilde{\boldsymbol\beta})\boldsymbol\beta^k-{\boldsymbol\beta^k}^{\mathrm{T}}E\{\mathbf{F}_{\boldsymbol\beta}(\mathbf{X},\mathbf{Y}|\widetilde{\boldsymbol\Theta},\widetilde{\boldsymbol\beta})\}\boldsymbol\beta^k|\geq\delta^2+\frac{32(d_{H_2}+d_{S_2}^2)a^2c_0^2}{\sqrt{mn}}\right)\\
\leq\ & \Pr\left(|{\boldsymbol\beta^k}^{\mathrm{T}}\mathbf{F}_{\boldsymbol\beta}(\mathbf{X},\mathbf{Y}|\widetilde{\boldsymbol\Theta},\widetilde{\boldsymbol\beta})\boldsymbol\beta^k-m_F\}|+|{\boldsymbol\beta^k}^{\mathrm{T}}E\{\mathbf{F}_{\boldsymbol\beta}(\mathbf{X},\mathbf{Y}|\widetilde{\boldsymbol\Theta},\widetilde{\boldsymbol\beta})\}\boldsymbol\beta^k-m_F\}|\geq\delta^2+\frac{32(d_{H_2}+d_{S_2}^2)a^2c_0^2}{\sqrt{mn}}\right)\\
\leq\ & \Pr\left(|{\boldsymbol\beta^k}^{\mathrm{T}}\mathbf{F}_{\boldsymbol\beta}(\mathbf{X},\mathbf{Y}|\widetilde{\boldsymbol\Theta},\widetilde{\boldsymbol\beta})\boldsymbol\beta^k-m_F\}|\geq\delta^2+\frac{32(d_{H_2}+d_{S_2}^2)a^2c_0^2}{\sqrt{mn}}-\frac{16\sqrt{\pi}(d_{H_2}+d_{S_2}^2)a^2c_0^2}{\sqrt{mn}}\right)\\
\leq\ & \Pr\left(|{\boldsymbol\beta^k}^{\mathrm{T}}\mathbf{F}_{\boldsymbol\beta}(\mathbf{X},\mathbf{Y}|\widetilde{\boldsymbol\Theta},\widetilde{\boldsymbol\beta})\boldsymbol\beta^k-m_F|\geq\delta^2\right)\\
\leq\ & 4\exp\left\{\frac{-mn\delta^4}{64(d_{H_2}+d_{S_2}^2)^2a^2c_0^4D^2}\right\}.
\end{aligned}
$$

Now for $\boldsymbol\beta^k\in\mathbb{B}(D)$, $k=1,\ldots,N(\delta)$, we have

$$
\Pr\left\{\sup_{k=1,\ldots,N(\delta)}|{\boldsymbol\beta^k}^{\mathrm{T}}\mathbf{F}_{\boldsymbol\beta}(\mathbf{X},\mathbf{Y}|\widetilde{\boldsymbol\Theta},\widetilde{\boldsymbol\beta})\boldsymbol\beta^k-\boldsymbol\beta^{k\mathrm{T}}E[R_{ij}E\{S_2^2(Y_{ij},\mathbf{X}_{ij}|\widetilde{\boldsymbol\Theta},\widetilde{\boldsymbol\beta})\mathbf{X}_{ij}\mathbf{X}_{ij}^{\mathrm{T}}|Y_{ij}\}\right.
$$

$$
\begin{aligned}
& -R_{ij}E\{S_2(Y_{ij},\mathbf{X}_{ij}|\widetilde{\boldsymbol{\Theta}},\widetilde{\boldsymbol{\beta}})\mathbf{X}_{ij}|Y_{ij}\}^{\otimes 2}]\boldsymbol{\beta}^k| \geq \delta^2 + \frac{32(d_{H_2}+d_{S_2}^2)a^2c_0^2}{\sqrt{mn}}\Bigg\} \\
\leq\ & 4\exp\left\{\frac{-mn\delta^4}{64(d_{H_2}+d_{S_2}^2)^2a^2c_0^4D^2}+\log\{N(\delta)\}\right\} \\
\leq\ & 4\exp\left\{\frac{-mn\delta^4}{64(d_{H_2}+d_{S_2}^2)^2a^2c_0^4D^2}+\frac{9}{\delta^2}c_1^2D^4mn/\gamma^2\right\}.
\end{aligned}
$$

Now because $\delta = D/\xi$, we can select γ sufficiently large so that $\delta^4\{64(d_{H_2}+d_{S_2}^2)^2a^2c_0^4D^2\}^{-1} \geq 18\delta^{-2}c_1^2D^4/\gamma^2$. Thus we have

$$
\begin{aligned}
& \Pr\Bigg\{\sup_{k=1,\ldots,N(\delta)}|\boldsymbol{\beta}^{k\mathrm{T}}\mathbf{F}_{\boldsymbol{\beta}}(\mathbf{X},\mathbf{Y}|\widetilde{\boldsymbol{\Theta}},\widetilde{\boldsymbol{\beta}})\boldsymbol{\beta}^k - \boldsymbol{\beta}^{k\mathrm{T}}E\{R_{ij}E\{S_2^2(Y_{ij},\mathbf{X}_{ij}|\widetilde{\boldsymbol{\Theta}},\widetilde{\boldsymbol{\beta}})\mathbf{X}_{ij}\mathbf{X}_{ij}^{\mathrm{T}}|Y_{ij}\} \\
& -R_{ij}E\{S_2(Y_{ij},\mathbf{X}_{ij}|\widetilde{\boldsymbol{\Theta}},\widetilde{\boldsymbol{\beta}})\mathbf{X}_{ij}|Y_{ij}\}^{\otimes 2}\}\boldsymbol{\beta}^k| \geq \delta^2 + \frac{32(d_{H_2}+d_{S_2}^2)a^2c_0^2}{\sqrt{mn}}\Bigg\} \\
\leq\ & 4\exp\left(-c_2D^2mn\right),
\end{aligned}
$$

where $c_2 = -9c_1^2\xi^2\gamma^{-2}$. Using the similar argument we can show that there is a $c_3 > 0$ such that

$$
\begin{aligned}
& \Pr\Bigg\{\sup_{k=1,\ldots,N(\delta)}|\boldsymbol{\beta}^{k\mathrm{T}}\mathbf{F}_{\boldsymbol{\beta}}(\mathbf{X},\mathbf{Y})(\boldsymbol{\beta}-\boldsymbol{\beta}^k) - \boldsymbol{\beta}^{k\mathrm{T}}E\{R_{ij}E\{S_2^2(Y_{ij},\mathbf{X}_{ij}|\widetilde{\boldsymbol{\Theta}},\widetilde{\boldsymbol{\beta}})\mathbf{X}_{ij}\mathbf{X}_{ij}^{\mathrm{T}}|Y_{ij}\} \\
& -R_{ij}E\{S_2(Y_{ij},\mathbf{X}_{ij}|\widetilde{\boldsymbol{\Theta}},\widetilde{\boldsymbol{\beta}})\mathbf{X}_{ij}|Y_{ij}\}^{\otimes 2}\}(\boldsymbol{\beta}-\boldsymbol{\beta}^k)| \\
& \geq D^2/\xi^2 + \frac{64(d_{H_2}+d_{S_2}^2)a^2c_0^2}{\sqrt{mn}}\Bigg\} \\
& \leq 4\exp\left(-c_3D^2mn\right).
\end{aligned}
$$

This proves the result.

Lemma A.9 *Assume Conditions (C5) and (C6) hold, $\delta = D/\xi$. Let $\mathbf{F}_{\boldsymbol{\beta}}(\mathbf{X},\mathbf{Y}|\widetilde{\boldsymbol{\Theta}},\widetilde{\boldsymbol{\beta}})$ be as defined in Lemma A.8. Define $\mathcal{D}_{\boldsymbol{\beta}}(D) \equiv \{\boldsymbol{\Delta}_{\boldsymbol{\beta}} \in \mathbb{R}^p | \|\boldsymbol{\Delta}_{\boldsymbol{\beta}}\|_2 \leq \delta, \|\boldsymbol{\Delta}_{\boldsymbol{\beta}}\|_1 \leq 2D^2\sqrt{mn/\log\{\max(p,mn)\}}/\gamma, \|\boldsymbol{\Delta}_{\boldsymbol{\beta}}\|_\infty \leq 2a\}$. Then*

$$
\sup_{\boldsymbol{\Delta}_{\boldsymbol{\beta}}\in\mathcal{D}_{\boldsymbol{\beta}}(D)}|\boldsymbol{\Delta}_{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{F}_{\boldsymbol{\beta}}(\mathbf{X},\mathbf{Y}|\widetilde{\boldsymbol{\Theta}},\widetilde{\boldsymbol{\beta}})\boldsymbol{\Delta}_{\boldsymbol{\beta}}| \leq 2(d_{H_2}+d_{S_2}^2)c_0^2\delta^2.
$$

Proof: First note that

$$
\sup_{\boldsymbol{\Delta}_{\boldsymbol{\beta}}\in\mathcal{D}_{\boldsymbol{\beta}}(D)}(mn)^{-1}\sum_{i=1}^n\sum_{j=1}^m\boldsymbol{\Delta}_{\boldsymbol{\beta}}^{\mathrm{T}}H_2(Y_{ij},\mathbf{X}_{ij}|\widetilde{\boldsymbol{\Theta}},\widetilde{\boldsymbol{\beta}})\mathbf{X}_{ij}\mathbf{X}_{ij}^{\mathrm{T}}\boldsymbol{\Delta}_{\boldsymbol{\beta}} \leq d_{H_2}c_0^2\delta^2.
$$

and first inequality holds by Conditions (C5) and (C6).

Follow the same arguments we have

$$
\sup_{\boldsymbol{\Delta}_{\boldsymbol{\beta}}\in\mathcal{D}_{\boldsymbol{\beta}}(D)}\left\{(mn)^{-1}\sum_{i=1}^n\sum_{j=1}^m\boldsymbol{\Delta}_{\boldsymbol{\beta}}^{\mathrm{T}}E\{H_2(Y_{ij},\mathbf{X}_{ij}|\widetilde{\boldsymbol{\Theta}},\widetilde{\boldsymbol{\beta}})\mathbf{X}_{ij}\mathbf{X}_{ij}^{\mathrm{T}}|Y_{ij}\}\boldsymbol{\Delta}_{\boldsymbol{\beta}}\right\} \leq d_{H_2}c_0^2\delta^2,
$$

$$
\sup_{\boldsymbol{\Delta}_{\boldsymbol{\beta}}\in\mathcal{D}_{\boldsymbol{\beta}}(D)}\left\{(mn)^{-1}\sum_{i=1}^n\sum_{j=1}^m\boldsymbol{\Delta}_{\boldsymbol{\beta}}^{\mathrm{T}}E\{S_2^2(Y_{ij},\mathbf{X}_{ij}|\widetilde{\boldsymbol{\Theta}},\widetilde{\boldsymbol{\beta}})\mathbf{X}_{ij}\mathbf{X}_{ij}^{\mathrm{T}}|Y_{ij}\}\boldsymbol{\Delta}_{\boldsymbol{\beta}}\right\} \leq d_{S_2}^2c_0^2\delta^2, \tag{9}
$$

and

$$\sup_{\boldsymbol{\Delta}_{\boldsymbol{\beta}}\in\mathcal{D}_{\boldsymbol{\beta}}(D)}\left\{(mn)^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m}\boldsymbol{\Delta}_{\boldsymbol{\beta}}^{\mathrm{T}}E\{S_2(Y_{ij},\mathbf{X}_{ij}|\widetilde{\boldsymbol{\Theta}},\widetilde{\boldsymbol{\beta}})\mathbf{X}_{ij}|Y_{ij}\}^{\otimes 2}\boldsymbol{\Delta}_{\boldsymbol{\beta}}\right\}\leq d_{S_2}^2c_0^2\delta^2 \tag{10}$$

Hence

$$\sup_{\boldsymbol{\Delta}_{\boldsymbol{\beta}}\in\mathcal{D}_{\boldsymbol{\beta}}D)}|\boldsymbol{\Delta}_{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{F}_{\boldsymbol{\beta}}(\mathbf{X},\mathbf{Y}|\widetilde{\boldsymbol{\Theta}},\widetilde{\boldsymbol{\beta}})\boldsymbol{\Delta}_{\boldsymbol{\beta}}|\leq 2(d_{H_2}+d_{S_2}^2)c_0^2\delta^2.$$

Lemma A.10 *Assume Conditions (C1)–(C6) hold. Let δ be as defined in Lemma A.9. For $\boldsymbol{\beta}\in\mathcal{C}(\gamma)$, we have*

$$\begin{aligned}
&|\boldsymbol{\beta}^{\mathrm{T}}\mathbf{F}_{\boldsymbol{\beta}}(\mathbf{X},\mathbf{Y}|\widetilde{\boldsymbol{\Theta}},\widetilde{\boldsymbol{\beta}})\boldsymbol{\beta}-\boldsymbol{\beta}^{\mathrm{T}}E[R_{ij}E\{S_2^2(Y_{ij},\mathbf{X}_{ij}|\widetilde{\boldsymbol{\Theta}},\widetilde{\boldsymbol{\beta}})\mathbf{X}_{ij}\mathbf{X}_{ij}^{\mathrm{T}}|Y_{ij}\}\\
&-R_{ij}E\{S_2(Y_{ij},\mathbf{X}_{ij}|\widetilde{\boldsymbol{\Theta}},\widetilde{\boldsymbol{\beta}})\mathbf{X}_{ij}|Y_{ij}\}^{\otimes 2}]\boldsymbol{\beta}|\\
\geq\ &\alpha_{\min}(E[R_{ij}E\{S_2(Y_{ij},\mathbf{X}_{ij}|\widetilde{\boldsymbol{\Theta}},\widetilde{\boldsymbol{\beta}})\mathbf{X}_{ij}\mathbf{X}_{ij}^{\mathrm{T}}|Y_{ij}\}\\
&-R_{ij}E\{S_2(Y_{ij},\mathbf{X}_{ij}|\widetilde{\boldsymbol{\Theta}},\widetilde{\boldsymbol{\beta}})\mathbf{X}_{ij}|Y_{ij}\}^{\otimes 2}])/2\|\boldsymbol{\beta}\|_2^2+160(d_{H_2}+d_{S_2}^2)a^2c_0^2/\sqrt{mn}
\end{aligned}$$

with probability at most $1-\exp[-C\log\{\max(p,mn)\}]$ for some positive constant C.

Proof: For any $\boldsymbol{\beta}\in\mathbb{B}(D)$, there is a $k\in\{1,\ldots,N(\delta)\}$, so that $\widetilde{\boldsymbol{\Delta}}_{\boldsymbol{\beta}}\equiv\boldsymbol{\beta}-\boldsymbol{\beta}^k$ satisfies $\|\widetilde{\boldsymbol{\Delta}}_{\boldsymbol{\beta}}\|_2\leq\delta$, $\|\widetilde{\boldsymbol{\Delta}}_{\boldsymbol{\beta}}\|_1\leq 2D^2\sqrt{mn/\log\{\max(p,mn)\}}/\gamma$, and $\|\widetilde{\boldsymbol{\Delta}}_{\boldsymbol{\beta}}\|_\infty\leq 2a$, hence $\widetilde{\boldsymbol{\Delta}}_{\boldsymbol{\beta}}\in\mathcal{D}_{\boldsymbol{\beta}}(D)$, where $\mathcal{D}_{\boldsymbol{\beta}}(D)$ is defined in Lemma A.9. In addition,

$$\begin{aligned}
&\boldsymbol{\beta}^{\mathrm{T}}\mathbf{F}_{\boldsymbol{\beta}}(\mathbf{X},\mathbf{Y}|\widetilde{\boldsymbol{\Theta}},\widetilde{\boldsymbol{\beta}})\boldsymbol{\beta}-\boldsymbol{\beta}^{\mathrm{T}}E[R_{ij}E\{S_2^2(Y_{ij},\mathbf{X}_{ij}|\widetilde{\boldsymbol{\Theta}},\widetilde{\boldsymbol{\beta}})\mathbf{X}_{ij}\mathbf{X}_{ij}^{\mathrm{T}}|Y_{ij}\}\\
&-R_{ij}E\{S_2(Y_{ij},\mathbf{X}_{ij}|\widetilde{\boldsymbol{\Theta}},\widetilde{\boldsymbol{\beta}})\mathbf{X}_{ij}|Y_{ij}\}^{\otimes 2}]\boldsymbol{\beta}\\
=\ &(\boldsymbol{\beta}^k+\widetilde{\boldsymbol{\Delta}}_{\boldsymbol{\beta}})^{\mathrm{T}}\mathbf{F}_{\boldsymbol{\beta}}(\mathbf{X},\mathbf{Y}|\widetilde{\boldsymbol{\Theta}},\widetilde{\boldsymbol{\beta}})(\boldsymbol{\beta}^k+\widetilde{\boldsymbol{\Delta}}_{\boldsymbol{\beta}})-(\boldsymbol{\beta}^k+\widetilde{\boldsymbol{\Delta}}_{\boldsymbol{\beta}})^{\mathrm{T}}E[R_{ij}E\{S_2^2(Y_{ij},\mathbf{X}_{ij}|\widetilde{\boldsymbol{\Theta}},\widetilde{\boldsymbol{\beta}})\mathbf{X}_{ij}\mathbf{X}_{ij}^{\mathrm{T}}|Y_{ij}\}\\
&-R_{ij}E\{S_2(Y_{ij},\mathbf{X}_{ij}|\widetilde{\boldsymbol{\Theta}},\widetilde{\boldsymbol{\beta}})\mathbf{X}_{ij}|Y_{ij}\}^{\otimes 2}](\boldsymbol{\beta}^k+\widetilde{\boldsymbol{\Delta}}_{\boldsymbol{\beta}})\\
=\ &{\boldsymbol{\beta}^k}^{\mathrm{T}}\mathbf{F}_{\boldsymbol{\beta}}(\mathbf{X},\mathbf{Y}|\widetilde{\boldsymbol{\Theta}},\widetilde{\boldsymbol{\beta}})\boldsymbol{\beta}^k-(\boldsymbol{\beta}^k)^{\mathrm{T}}E[R_{ij}E\{S_2^2(Y_{ij},\mathbf{X}_{ij}|\widetilde{\boldsymbol{\Theta}},\widetilde{\boldsymbol{\beta}})\mathbf{X}_{ij}\mathbf{X}_{ij}^{\mathrm{T}}|Y_{ij}\}\\
&-R_{ij}E\{S_2(Y_{ij},\mathbf{X}_{ij}|\widetilde{\boldsymbol{\Theta}},\widetilde{\boldsymbol{\beta}})\mathbf{X}_{ij}|Y_{ij}\}^{\otimes 2}]\boldsymbol{\beta}^k\\
&+2{\boldsymbol{\beta}^k}^{\mathrm{T}}\mathbf{F}_{\boldsymbol{\beta}}(\mathbf{X},\mathbf{Y}|\widetilde{\boldsymbol{\Theta}},\widetilde{\boldsymbol{\beta}})\widetilde{\boldsymbol{\Delta}}_{\boldsymbol{\beta}},-(\boldsymbol{\beta}^k)^{\mathrm{T}}E[R_{ij}E\{S_2^2(Y_{ij},\mathbf{X}_{ij}|\widetilde{\boldsymbol{\Theta}},\widetilde{\boldsymbol{\beta}})\mathbf{X}_{ij}\mathbf{X}_{ij}^{\mathrm{T}}|Y_{ij}\}\\
&-2R_{ij}E\{S_2(Y_{ij},\mathbf{X}_{ij}|\widetilde{\boldsymbol{\Theta}},\widetilde{\boldsymbol{\beta}})\mathbf{X}_{ij}|Y_{ij}\}^{\otimes 2}]\widetilde{\boldsymbol{\Delta}}_{\boldsymbol{\beta}}\\
&+\widetilde{\boldsymbol{\Delta}}_{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{F}_{\boldsymbol{\beta}}(\mathbf{X},\mathbf{Y}|\widetilde{\boldsymbol{\Theta}},\widetilde{\boldsymbol{\beta}})\widetilde{\boldsymbol{\Delta}}_{\boldsymbol{\beta}}-\widetilde{\boldsymbol{\Delta}}_{\boldsymbol{\beta}}^{\mathrm{T}}E[R_{ij}E\{S_2^2(Y_{ij},\mathbf{X}_{ij}|\widetilde{\boldsymbol{\Theta}},\widetilde{\boldsymbol{\beta}})\mathbf{X}_{ij}\mathbf{X}_{ij}^{\mathrm{T}}|Y_{ij}\}\\
&-R_{ij}E\{S_2(Y_{ij},\mathbf{X}_{ij},\widetilde{\boldsymbol{\Theta}},\widetilde{\boldsymbol{\beta}})\mathbf{X}_{ij}|Y_{ij}\}^{\otimes 2}]\widetilde{\boldsymbol{\Delta}}_{\boldsymbol{\beta}}.
\end{aligned}$$

Hence by Lemmas A.8 and A.9,

$$\begin{aligned}
&\sup_{\boldsymbol{\beta}\in\mathbb{B}(D)}|\boldsymbol{\beta}^{\mathrm{T}}\mathbf{F}_{\boldsymbol{\beta}}(\mathbf{X},\mathbf{Y}|\widetilde{\boldsymbol{\Theta}},\widetilde{\boldsymbol{\beta}})\boldsymbol{\beta}-\boldsymbol{\beta}^{\mathrm{T}}E[R_{ij}E\{S_2^2(Y_{ij},\mathbf{X}_{ij}|\widetilde{\boldsymbol{\Theta}},\widetilde{\boldsymbol{\beta}})\mathbf{X}_{ij}\mathbf{X}_{ij}^{\mathrm{T}}|Y_{ij}\}\\
&-R_{ij}E\{S_2(Y_{ij},\mathbf{X}_{ij}|\widetilde{\boldsymbol{\Theta}},\widetilde{\boldsymbol{\beta}})\mathbf{X}_{ij}|Y_{ij}\}^{\otimes 2}]\boldsymbol{\beta}|\\
\leq\ &\sup_{k=1,\ldots,N(\delta)}|{\boldsymbol{\beta}^k}^{\mathrm{T}}\mathbf{F}_{\boldsymbol{\beta}}(\mathbf{X},\mathbf{Y}|\widetilde{\boldsymbol{\Theta}},\widetilde{\boldsymbol{\beta}})\boldsymbol{\beta}^k-(\boldsymbol{\beta}^k)^{\mathrm{T}}E[R_{ij}E\{S_2^2(Y_{ij},\mathbf{X}_{ij}|\widetilde{\boldsymbol{\Theta}},\widetilde{\boldsymbol{\beta}})\mathbf{X}_{ij}\mathbf{X}_{ij}^{\mathrm{T}}|Y_{ij}\}\\
&-E\{S_2(Y_{ij},\mathbf{X}_{ij}|\widetilde{\boldsymbol{\Theta}},\widetilde{\boldsymbol{\beta}})\mathbf{X}_{ij}|Y_{ij}\}^{\otimes 2}]\boldsymbol{\beta}^k|\\
&+2\sup_{k=1,\ldots,N(\delta)}|{\boldsymbol{\beta}^k}^{\mathrm{T}}\mathbf{F}_{\boldsymbol{\beta}}(\mathbf{X},\mathbf{Y}|\widetilde{\boldsymbol{\Theta}},\widetilde{\boldsymbol{\beta}})\widetilde{\boldsymbol{\Delta}}_{\boldsymbol{\beta}}-(\boldsymbol{\beta}^k)^{\mathrm{T}}E[E\{S_2^2(Y_{ij},\mathbf{X}_{ij}|\widetilde{\boldsymbol{\Theta}},\widetilde{\boldsymbol{\beta}})\mathbf{X}_{ij}\mathbf{X}_{ij}^{\mathrm{T}}|Y_{ij}\}
\end{aligned}$$

$$
\begin{aligned}
& -R_{ij}E\{S_2(Y_{ij},\mathbf{X}_{ij}|\widetilde{\boldsymbol{\Theta}},\widetilde{\boldsymbol{\beta}})\mathbf{X}_{ij}|Y_{ij}\}^{\otimes 2}]\widetilde{\boldsymbol{\Delta}}_{\boldsymbol{\beta}}| \\
& + \sup_{\widetilde{\boldsymbol{\Delta}}_{\boldsymbol{\beta}}\in\mathcal{D}_{\boldsymbol{\beta}}(D)} |\widetilde{\boldsymbol{\Delta}}_{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{F}_{\boldsymbol{\beta}}(\mathbf{X},\mathbf{Y}|\widetilde{\boldsymbol{\Theta}},\widetilde{\boldsymbol{\beta}})\widetilde{\boldsymbol{\Delta}}_{\boldsymbol{\beta}}| + \sup_{\widetilde{\boldsymbol{\Delta}}_{\boldsymbol{\beta}}\in\mathcal{D}_{\boldsymbol{\beta}}(D)} |\widetilde{\boldsymbol{\Delta}}_{\boldsymbol{\beta}}^{\mathrm{T}}E[E\{S_2^2(Y_{ij},\mathbf{X}_{ij}|\widetilde{\boldsymbol{\Theta}},\widetilde{\boldsymbol{\beta}})\mathbf{X}_{ij}\mathbf{X}_{ij}^{\mathrm{T}}|Y_{ij} \\
& -E\{S_2(Y_{ij},\mathbf{X}_{ij}|\widetilde{\boldsymbol{\Theta}},\widetilde{\boldsymbol{\beta}})\mathbf{X}_{ij}|Y_{ij}\}^{\otimes 2}]\widetilde{\boldsymbol{\Delta}}_{\boldsymbol{\beta}}| \\
\leq\ & 3\delta^2 + \frac{160(d_{H_2}+d_{S_2}^2)a^2c_0^2}{\sqrt{mn}} + (2d_{H_2}c_0^2 + 2d_{S_2}^2c_0^2)\delta^2 + 2d_{S_2}^2c_0^2\delta^2 \\
=\ & (3 + 2d_{H_2}c_0^2 + 4d_{S_2}^2c_0^2)D^2/\xi^2 + \frac{160(d_{H_2}+d_{S_2}^2)a^2c_0^2}{\sqrt{mn}} \\
=\ & D_5D^2/\xi^2 + \frac{160(d_{H_2}+d_{S_2}^2)a^2c_0^2}{\sqrt{mn}} \qquad (11)
\end{aligned}
$$

with probability at least $1 - b_3\{\exp(-b_2D^2mn)\}$, where b_2, b_3 are constants not depending on D, and

$$D_5 = 3 + 2d_{H_2}c_0^2 + 4d_{S_2}^2c_0^2.$$

The second to the last equality holds by Lemma A.8, A.9 and (9), (10) in Lemma A.9. Now note for $\boldsymbol{\beta}\in\mathcal{C}(\gamma)$ with $\|\boldsymbol{\beta}\|_\infty = b$ for a constant $b \leq a$, we have

$$\|\boldsymbol{\beta}\|_2^2 \geq b\gamma\|\boldsymbol{\beta}\|_1\sqrt{\frac{\log\{\max(p,mn)\}}{mn}} \geq b\gamma\|\boldsymbol{\beta}\|_2\sqrt{\frac{\log\{\max(p,mn)\}}{mn}},$$

which implies $\|\boldsymbol{\beta}\|_2 \geq b\gamma\sqrt{\log\{\max(p,mn)\}/(mn)}$. Define $\mu(b) = b^2\gamma^2\log\{\max(p,mn)\}/(mn)$. Further let

$$\alpha_{0\boldsymbol{\beta}}(\widetilde{\boldsymbol{\Theta}},\widetilde{\boldsymbol{\beta}}) = \alpha_{\min}(E[R_{ij}E\{S_2^2(Y_{ij},\mathbf{X}_{ij}|\widetilde{\boldsymbol{\Theta}},\widetilde{\boldsymbol{\beta}})\mathbf{X}_{ij}\mathbf{X}_{ij}^{\mathrm{T}}|Y_{ij}\} - R_{ij}E\{S_2(Y_{ij},\mathbf{X}_{ij}|\widetilde{\boldsymbol{\Theta}},\widetilde{\boldsymbol{\beta}})\mathbf{X}_{ij}|Y_{ij}\}^{\otimes 2}]),$$

$\alpha = \alpha_{0\boldsymbol{\beta}}(\widetilde{\boldsymbol{\Theta}},\widetilde{\boldsymbol{\beta}})\xi^2/D_5$ and $\xi \geq \sqrt{D_5/\alpha_{0\boldsymbol{\beta}}}$ so that $\alpha > 1$. Define

$$
\begin{aligned}
& \mathcal{S}_l(b) = [\boldsymbol{\beta}\in\mathcal{C}(\gamma)|\|\boldsymbol{\beta}\|_\infty = b, \\
& \sqrt{\alpha^{l-1}\mu(b)} \leq \|\boldsymbol{\beta}\|_2 \leq \sqrt{\alpha^l\mu(b)}, \text{ and} \\
& \|\boldsymbol{\beta}\|_1 \leq b^{-1}\alpha^l\mu(b)\gamma^{-1}\sqrt{mn/\log\{\max(p,mn)\}}].
\end{aligned}
$$

Then $\mathcal{S}_l(b) \subseteq \mathbb{B}(\sqrt{\alpha^l\mu(b)})$ and $\{\boldsymbol{\beta} : \|\boldsymbol{\beta}\|_\infty = b\}\cap\mathcal{C}(\boldsymbol{\gamma}) \subset \cup_{l=1}^{\infty}\mathcal{S}_l(b)$.

For $\boldsymbol{\beta}\in\mathcal{S}_l(b)$, we have

$$
\begin{aligned}
\alpha_{0\boldsymbol{\beta}}(\widetilde{\boldsymbol{\Theta}},\widetilde{\boldsymbol{\beta}})\|\boldsymbol{\beta}\|_2^2 + \frac{160(d_{H_2}+d_{S_2}^2)a^2c_0^2}{\sqrt{mn}} &\geq \alpha_{0\boldsymbol{\beta}}(\widetilde{\boldsymbol{\Theta}},\widetilde{\boldsymbol{\beta}})\alpha^{l-1}\mu(b) + \frac{160(d_{H_2}+d_{S_2}^2)a^2c_0^2}{\sqrt{mn}} \\
&= \alpha_{0\boldsymbol{\beta}}\alpha^{-1}\alpha^l\mu(b) + \frac{160(d_{H_2}+d_{S_2}^2)a^2c_0^2}{\sqrt{mn}} \\
&= D_5\alpha^l\mu(b)/\xi^2 + \frac{160(d_{H_2}+d_{S_2}^2)a^2c_0^2}{\sqrt{mn}}.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
& \Pr\Big\{|\boldsymbol{\beta}^{\mathrm{T}}\mathbf{F}_{\boldsymbol{\beta}}(\mathbf{X},\mathbf{Y}|\widetilde{\boldsymbol{\Theta}},\widetilde{\boldsymbol{\beta}})\boldsymbol{\beta} - \boldsymbol{\beta}^{\mathrm{T}}E[R_{ij}E\{S_2^2(Y_{ij},\mathbf{X}_{ij}|\widetilde{\boldsymbol{\Theta}},\widetilde{\boldsymbol{\beta}})\mathbf{X}_{ij}\mathbf{X}_{ij}^{\mathrm{T}}|Y_{ij}\} \\
& -R_{ij}E\{S_2(Y_{ij},\mathbf{X}_{ij}|\widetilde{\boldsymbol{\Theta}},\widetilde{\boldsymbol{\beta}})\mathbf{X}_{ij}|Y_{ij}\}^{\otimes 2}]\boldsymbol{\beta}| \\
& \geq \alpha_{\min}(E[R_{ij}E\{S_2^2(Y_{ij},\mathbf{X}_{ij}|\widetilde{\boldsymbol{\Theta}},\widetilde{\boldsymbol{\beta}})\mathbf{X}_{ij}\mathbf{X}_{ij}^{\mathrm{T}}|Y_{ij}\}
\end{aligned}
$$

$$
\begin{aligned}
& -R_{ij}E\{S_2(Y_{ij},\mathbf{X}_{ij}|\widetilde{\boldsymbol{\Theta}},\widetilde{\boldsymbol{\beta}})\mathbf{X}_{ij}|Y_{ij}\}^{\otimes 2}])\|\boldsymbol{\beta}\|_2^2+\frac{160(d_{H_2}+d_{S_2}^2)a^2c_0^2}{\sqrt{mn}},\boldsymbol{\beta}\in\mathcal{C}(\gamma)\Big\}\\
\leq\ & \int_0^a \Pr\Big\{|\boldsymbol{\beta}^{\mathrm{T}}\mathbf{F}_{\boldsymbol{\beta}}(\mathbf{X},\mathbf{Y}|\widetilde{\boldsymbol{\Theta}},\widetilde{\boldsymbol{\beta}})\boldsymbol{\beta}-\boldsymbol{\beta}^{\mathrm{T}}E[R_{ij}E\{S_2^2(Y_{ij},\mathbf{X}_{ij}|\widetilde{\boldsymbol{\Theta}},\widetilde{\boldsymbol{\beta}})\mathbf{X}_{ij}\mathbf{X}_{ij}^{\mathrm{T}}|Y_{ij}\}\\
& -R_{ij}E\{S_2(Y_{ij},\mathbf{X}_{ij}|\widetilde{\boldsymbol{\Theta}},\widetilde{\boldsymbol{\beta}})\mathbf{X}_{ij}|Y_{ij}\}^{\otimes 2}]\boldsymbol{\beta}|\\
& \geq \alpha_{\min}(E[R_{ij}E\{S_2^2(Y_{ij},\mathbf{X}_{ij}|\widetilde{\boldsymbol{\Theta}},\widetilde{\boldsymbol{\beta}})\mathbf{X}_{ij}\mathbf{X}_{ij}^{\mathrm{T}}|Y_{ij}\}\\
& -R_{ij}E\{S_2(Y_{ij},\mathbf{X}_{ij}|\widetilde{\boldsymbol{\Theta}},\widetilde{\boldsymbol{\beta}})\mathbf{X}_{ij}|Y_{ij}\}^{\otimes 2}])\|\boldsymbol{\beta}\|_2^2+\frac{160(d_{H_2}+d_{S_2}^2)a^2c_0^2}{\sqrt{mn}},\boldsymbol{\beta}\in\mathcal{C}(\gamma),\|\boldsymbol{\beta}\|_\infty=b\Big\}db\\
\leq\ & \int_0^a \sum_{l=1}^{\infty}\Pr\Big\{|\boldsymbol{\beta}^{\mathrm{T}}\mathbf{F}_{\boldsymbol{\beta}}(\mathbf{X},\mathbf{Y}|\widetilde{\boldsymbol{\Theta}},\widetilde{\boldsymbol{\beta}})\boldsymbol{\beta}-\boldsymbol{\beta}^{\mathrm{T}}E[R_{ij}E\{S_2^2(Y_{ij},\mathbf{X}_{ij}|\widetilde{\boldsymbol{\Theta}},\widetilde{\boldsymbol{\beta}})\mathbf{X}_{ij}\mathbf{X}_{ij}^{\mathrm{T}}|Y_{ij}\}\\
& -R_{ij}E\{S_2(Y_{ij},\mathbf{X}_{ij}|\widetilde{\boldsymbol{\Theta}},\widetilde{\boldsymbol{\beta}})\mathbf{X}_{ij}|Y_{ij}\}^{\otimes 2}]\boldsymbol{\beta}|\\
& \geq \alpha_{\min}(E[R_{ij}E\{S_2^2(Y_{ij},\mathbf{X}_{ij},\widetilde{\boldsymbol{\Theta}},\widetilde{\boldsymbol{\beta}})\mathbf{X}_{ij}\mathbf{X}_{ij}^{\mathrm{T}}|Y_{ij}\}\\
& -R_{ij}E\{S_2(Y_{ij},\mathbf{X}_{ij}|\widetilde{\boldsymbol{\Theta}},\widetilde{\boldsymbol{\beta}})\mathbf{X}_{ij}|Y_{ij}\}^{\otimes 2}])\|\boldsymbol{\beta}\|_2^2+\frac{160(d_{H_2}+d_{S_2}^2)a^2c_0^2}{\sqrt{mn}},\boldsymbol{\beta}\in\mathcal{S}_l(b)\Big\}db\\
=\ & \sum_{l=1}^{\infty}\Pr\Big\{|\boldsymbol{\beta}^{\mathrm{T}}\mathbf{F}_{\boldsymbol{\beta}}(\mathbf{X},\mathbf{Y}|\widetilde{\boldsymbol{\Theta}},\widetilde{\boldsymbol{\beta}})\boldsymbol{\beta}-\boldsymbol{\beta}^{\mathrm{T}}E[R_{ij}E\{S_2^2(Y_{ij},\mathbf{X}_{ij}|\widetilde{\boldsymbol{\Theta}},\widetilde{\boldsymbol{\beta}})\mathbf{X}_{ij}\mathbf{X}_{ij}^{\mathrm{T}}|Y_{ij}\}\\
& -R_{ij}E\{S_2(Y_{ij},\mathbf{X}_{ij}|\widetilde{\boldsymbol{\Theta}},\widetilde{\boldsymbol{\beta}})\mathbf{X}_{ij}|Y_{ij}\}^{\otimes 2}]\boldsymbol{\beta}|\geq D_5\alpha^l\mu(b^*)/\xi^2+\frac{160(d_{H_2}+d_{S_2}^2)a^2c_0^2}{\sqrt{mn}},\boldsymbol{\beta}\in\mathcal{S}_l(b^*)\Big\}\\
\leq\ & b_3\sum_{l=1}^{\infty}[\exp(-b_2\alpha^l\mu(b^*)mn)]\\
\leq\ & b_3\sum_{l=1}^{\infty}[\exp(-b_2 l\mu(b^*)\log(\alpha)mn)]\\
\leq\ & b_3\frac{\exp(-b_2\log(\alpha)\mu(b^*)mn)}{1-\exp(-b_2\log(\alpha)\mu(b^*)mn)}\\
\leq\ & \exp[-C\log\{\max(p,mn)\}],
\end{aligned}
$$

where b^* is a point on the line connecting 0 and a. The fourth line holds by (11) with $D=\sqrt{\alpha^l\mu(b^*)}$, the fifth line holds because $\alpha^l\geq l\log(\alpha)$ for $\alpha>1$. The last equality holds because $\mu(b^*)=b^{*2}\gamma^2\{\max(p,mn)\}/(mn)$. This proves the result.

Lemma A.11 *Assume Conditions (C5) and (C6) hold, suppose $\boldsymbol{\Delta}_{\boldsymbol{\beta}}=\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0$, let $\widetilde{\boldsymbol{\Theta}}=\widehat{\boldsymbol{\Theta}}$ and $\widetilde{\boldsymbol{\beta}}=\boldsymbol{\beta}^*$ a point connecting $\widehat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}_0$, then there is a $\sigma_{1F}>0$ such that*

$$\boldsymbol{\Delta}_{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{F}_{\boldsymbol{\beta}}(\mathbf{X},\mathbf{Y}|\widehat{\boldsymbol{\Theta}},\boldsymbol{\beta}^*)\boldsymbol{\Delta}_{\boldsymbol{\beta}}\geq\boldsymbol{\Delta}_{\boldsymbol{\beta}}^{\mathrm{T}}E\{\mathbf{F}_{\boldsymbol{\beta}}(\mathbf{X},\mathbf{Y}|\widehat{\boldsymbol{\Theta}},\boldsymbol{\beta}^*)\}\boldsymbol{\Delta}_{\boldsymbol{\beta}}-\sigma_{1F}\sqrt{2\log\{\max(p,mn)\}/(mn)},$$

with probability at least $1-2\max(p,mn)^{-1}$, where $\sigma_{1F}=32c_0^2a^2(d_{H_2}+d_{S_2}^2)$ as defined in Theorem 1 and $E\{\mathbf{F}_{\boldsymbol{\beta}}(\mathbf{X},\mathbf{Y}|\widehat{\boldsymbol{\Theta}},\boldsymbol{\beta}^*)\}=E\{\mathbf{F}_{\boldsymbol{\beta}}(\mathbf{X},\mathbf{Y}|\boldsymbol{\Theta},\boldsymbol{\beta})\}_{\boldsymbol{\Theta}=\widehat{\boldsymbol{\Theta}},\boldsymbol{\beta}=\boldsymbol{\beta}^*}$.

Proof: First note that for $\boldsymbol{\Delta}_{\boldsymbol{\beta}}$ with $\|\boldsymbol{\Delta}_{\boldsymbol{\beta}}\|_\infty\leq 2a$, we have

$$
\begin{aligned}
& |(\boldsymbol{\Delta}_{\boldsymbol{\beta}})^{\mathrm{T}}R_{ij}H_2(Y_{ij},\mathbf{X}_{ij}|\widehat{\boldsymbol{\Theta}},\boldsymbol{\beta}^*)\mathbf{X}_{ij}\mathbf{X}_{ij}^{\mathrm{T}})^{\mathrm{T}}\boldsymbol{\Delta}_{\boldsymbol{\beta}}|\\
\leq\ & \{\|\boldsymbol{\Delta}_{\boldsymbol{\beta}}\|_{\max}\|\mathbf{X}_{ij}\|_1\}^2H_2(Y_{ij},\mathbf{X}_{ij}|\widehat{\boldsymbol{\Theta}},\boldsymbol{\beta}^*)\leq 4c_0^2a^2d_{H_2},
\end{aligned}
$$

Similarly,

$$|\boldsymbol{\Delta}_{\boldsymbol{\beta}}^{\mathrm{T}}R_{ij}E\{H_2(Y_{ij},\mathbf{X}_{ij}|\widehat{\boldsymbol{\Theta}},\boldsymbol{\beta}^*)\mathbf{X}_{ij}\mathbf{X}_{ij}^{\mathrm{T}}|Y_{ij}\}\boldsymbol{\Delta}_{\boldsymbol{\beta}}|\leq 4c_0^2a^2d_{H_2},$$

$$|(\boldsymbol{\Delta}_{\boldsymbol{\beta}})^{\mathrm{T}} R_{ij} E\{S_2^2(Y_{ij}, \mathbf{X}_{ij}|\widehat{\boldsymbol{\Theta}}, \boldsymbol{\beta}^*)\mathbf{X}_{ij}\mathbf{X}_{ij}^{\mathrm{T}})^{\mathrm{T}}|Y_{ij}\}(\boldsymbol{\Delta}_{\boldsymbol{\beta}})| \leq 4c_0^2 a^2 d_{S_2}^2, \tag{12}$$

and

$$|(\boldsymbol{\Delta}_{\boldsymbol{\beta}})^{\mathrm{T}} R_{ij} E\{S_2(Y_{ij}, \mathbf{X}_{ij}|\widehat{\boldsymbol{\Theta}}, \boldsymbol{\beta}^*)\mathbf{X}_{ij}|Y_{ij}\}^{\otimes 2}(\boldsymbol{\Delta}_{\boldsymbol{\beta}})| \leq 4c_0^2 a^2 d_{S_2}^2. \tag{13}$$

Hence we have each summand in $\boldsymbol{\Delta}_{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{F}_{\boldsymbol{\beta}}(\mathbf{X}, \mathbf{Y}|\widehat{\boldsymbol{\Theta}}, \boldsymbol{\beta}^*)\boldsymbol{\Delta}_{\boldsymbol{\beta}} - \boldsymbol{\Delta}_{\boldsymbol{\beta}}^{\mathrm{T}} E\{\mathbf{F}_{\boldsymbol{\beta}}(\mathbf{X}, \mathbf{Y}|\widehat{\boldsymbol{\Theta}}, \boldsymbol{\beta}^*)\}\boldsymbol{\Delta}_{\boldsymbol{\beta}}$ is in the range of $\{-16(c_0^2 a^2 d_{H_2} + c_0^2 a^2 d_{S_2}^2), 16(c_0^2 a^2 d_{H_2} + c_0^2 a^2 d_{S_2}^2)\}$, and in turn is sub-Gaussian with parameter $\sigma_{1F}^2 \equiv \{32(c_0^2 a^2 d_{H_2} + c_0^2 a^2 d_{S_2}^2)\}^2$ by Lemma A.1. Here the expectation is taken over $\mathbf{X}$ and $\mathbf{Y}$. Therefore, by Lemma A.1, we have

$$\begin{aligned}
& \Pr\left\{|\boldsymbol{\Delta}_{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{F}_{\boldsymbol{\beta}}(\mathbf{X}, \mathbf{Y}|\widetilde{\boldsymbol{\Theta}}, \widetilde{\boldsymbol{\beta}})\boldsymbol{\Delta}_{\boldsymbol{\beta}} - \boldsymbol{\Delta}_{\boldsymbol{\beta}}^{\mathrm{T}} E\{\mathbf{F}_{\boldsymbol{\beta}}(\mathbf{X}, \mathbf{Y}|\widetilde{\boldsymbol{\Theta}}, \widetilde{\boldsymbol{\beta}})\}\boldsymbol{\Delta}_{\boldsymbol{\beta}}| > t\right\} \\
= & \Pr\left\{|mn\boldsymbol{\Delta}_{\boldsymbol{\beta}}^{\mathrm{T}} F(\mathbf{X}, \mathbf{Y}|\widetilde{\boldsymbol{\Theta}}, \widetilde{\boldsymbol{\beta}})\boldsymbol{\Delta}_{\boldsymbol{\beta}} - mn\boldsymbol{\Delta}_{\boldsymbol{\beta}}^{\mathrm{T}} E\{\mathbf{F}_{\boldsymbol{\beta}}(\mathbf{X}, \mathbf{Y}|\widetilde{\boldsymbol{\Theta}}, \widetilde{\boldsymbol{\beta}})\}\boldsymbol{\Delta}_{\boldsymbol{\beta}}| > mnt\right\} \\
\leq & 2\exp\left\{-\frac{(mn)^2 t^2}{2mn\sigma_{1F}^2}\right\} \\
= & 2\exp\left(-\frac{mnt^2}{2\sigma_{1F}^2}\right).
\end{aligned}$$

Let $t = \sigma_{1F}\sqrt{2\log\{\max(p, mn)\}/(mn)}$, we obtain

$$\begin{aligned}
& \Pr\left\{|\boldsymbol{\Delta}_{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{F}_{\boldsymbol{\beta}}(\mathbf{X}, \mathbf{Y}|\widehat{\boldsymbol{\Theta}}, \boldsymbol{\beta}^*)\boldsymbol{\Delta}_{\boldsymbol{\beta}} - \boldsymbol{\Delta}_{\boldsymbol{\beta}}^{\mathrm{T}} E\{\mathbf{F}_{\boldsymbol{\beta}}(\mathbf{X}, \mathbf{Y}|\widehat{\boldsymbol{\Theta}}, \boldsymbol{\beta}^*)\}\boldsymbol{\Delta}_{\boldsymbol{\beta}}| > \sigma_{1F}\sqrt{2\log\{\max(p, mn)\}/(mn)}\right\} \\
& \leq 2\max(p, mn)^{-1}.
\end{aligned}$$

Hence we have

$$\boldsymbol{\Delta}_{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{F}_{\boldsymbol{\beta}}(\mathbf{X}, \mathbf{Y}|\widehat{\boldsymbol{\Theta}}, \boldsymbol{\beta}^*)\boldsymbol{\Delta}_{\boldsymbol{\beta}} \geq \boldsymbol{\Delta}_{\boldsymbol{\beta}}^{\mathrm{T}} E\{\mathbf{F}_{\boldsymbol{\beta}}(\mathbf{X}, \mathbf{Y}|\widehat{\boldsymbol{\Theta}}, \boldsymbol{\beta}^*)\}\boldsymbol{\Delta}_{\boldsymbol{\beta}} - \sigma_{1F}\sqrt{2\log\{\max(p, mn)\}/(mn)},$$

with probability at least $1 - 2\max(p, mn)^{-1}$.

Lemma A.12 *Assume Condition (C5) and (C6) hold, suppose $\boldsymbol{\Delta}_{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0$. Let $\lambda_{\boldsymbol{\beta}} \geq 2\|\partial\mathcal{L}(\widehat{\boldsymbol{\Theta}}, \boldsymbol{\beta}_0)/\partial\boldsymbol{\beta}\|_{op}$ and σ_{1F} be as defined in Theorem 1. Then either*

$$\begin{aligned}
\|\boldsymbol{\Delta}_{\boldsymbol{\beta}}\|_2 \leq & \{\alpha_{\min}(E[R_{ij}E\{S_2^2(Y_{ij}, \mathbf{X}_{ij}|\widehat{\boldsymbol{\Theta}}, \boldsymbol{\beta}^*)\mathbf{X}_{ij}\mathbf{X}_{ij}^{\mathrm{T}}|Y_{ij}\} \\
& - R_{ij}E\{S_2(Y_{ij}, \mathbf{X}_{ij}|\widehat{\boldsymbol{\Theta}}, \boldsymbol{\beta}^*)\mathbf{X}_{ij}|Y_{ij}\}^{\otimes 2}])\}^{-1/2}\{2\sigma_{1F}^2\log\{\max(p, mn)\}/(mn)\}^{1/4},
\end{aligned}$$

or $\|\boldsymbol{\Delta}_{\boldsymbol{\beta}}\|_1 \leq 4\sqrt{s}\|\boldsymbol{\Delta}_{\boldsymbol{\beta}}\|_2$ with probability at least $1 - 2\{\max(p, mn)\}^{-1}$.

Proof. First consider $\boldsymbol{\Delta}_{\boldsymbol{\beta}}^{\mathrm{T}} E\{\mathbf{F}_{\boldsymbol{\beta}}(\mathbf{X}, \mathbf{Y}|\widehat{\boldsymbol{\Theta}}, \boldsymbol{\beta}^*)\}\boldsymbol{\Delta}_{\boldsymbol{\beta}} \leq \sigma_{1F}\sqrt{2\log\{\max(p, mn)\}/(mn)}$. Because

$$\begin{aligned}
E\{\mathbf{F}_{\boldsymbol{\beta}}(\mathbf{X}, \mathbf{Y}|\widehat{\boldsymbol{\Theta}}, \boldsymbol{\beta}^*)\} = & E[R_{ij}E\{S_2^2(Y_{ij}, \mathbf{X}_{ij}|\widehat{\boldsymbol{\Theta}}, \boldsymbol{\beta}^*)\mathbf{X}_{ij}\mathbf{X}_{ij}^{\mathrm{T}}|Y_{ij}\} \\
& - R_{ij}E\{S_2(Y_{ij}, \mathbf{X}_{ij}|\widehat{\boldsymbol{\Theta}}, \boldsymbol{\beta}^*)\mathbf{X}_{ij}|Y_{ij}\}^{\otimes 2}],
\end{aligned}$$

we have

$$\begin{aligned}
& \alpha_{\min}(E[R_{ij}E\{S_2^2(Y_{ij}, \mathbf{X}_{ij}|\widehat{\boldsymbol{\Theta}}, \boldsymbol{\beta}^*)\mathbf{X}_{ij}\mathbf{X}_{ij}^{\mathrm{T}}|Y_{ij}\} \\
& - R_{ij}E\{S_2(Y_{ij}, \mathbf{X}_{ij}|\widehat{\boldsymbol{\Theta}}, \boldsymbol{\beta}^*)\mathbf{X}_{ij}|Y_{ij}\}^{\otimes 2}])\|\boldsymbol{\Delta}_{\boldsymbol{\beta}}\|_2^2 \leq \sigma_{1F}\sqrt{2\log\{\max(p, mn)\}/(mn)},
\end{aligned}$$

and hence

$$\begin{aligned}
\|\boldsymbol{\Delta}_{\boldsymbol{\beta}}\|_2 \leq & \{\alpha_{\min}(E[R_{ij}E\{S_2^2(Y_{ij}, \mathbf{X}_{ij}|\widehat{\boldsymbol{\Theta}}, \boldsymbol{\beta}^*)\mathbf{X}_{ij}\mathbf{X}_{ij}^{\mathrm{T}}|Y_{ij}\}
\end{aligned}$$

$$-R_{ij}E\{S_2(Y_{ij},\mathbf{X}_{ij}|\widehat{\boldsymbol{\Theta}},\boldsymbol{\beta}^*)\mathbf{X}_{ij}|Y_{ij}\}^{\otimes 2}])\}^{-1/2}\sigma_{1F}^{1/2}\{2\log\{\max(p,mn)\}/(mn)\}^{1/4}.$$

Now consider $\boldsymbol{\Delta}_{\boldsymbol{\beta}}^{\mathrm{T}}E\{\mathbf{F}_{\boldsymbol{\beta}}(\mathbf{X},\mathbf{Y}|\widehat{\boldsymbol{\Theta}},\boldsymbol{\beta}^*)\}\boldsymbol{\Delta}_{\boldsymbol{\beta}} > \sigma_{1F}\sqrt{2\log\{\max(p,mn)\}/(mn)}$. We have

$$0 < \boldsymbol{\Delta}_{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{F}_{\boldsymbol{\beta}}(\mathbf{X},\mathbf{Y}|\widehat{\boldsymbol{\Theta}},\boldsymbol{\beta}^*)\boldsymbol{\Delta}_{\boldsymbol{\beta}}/2 = -\frac{\partial\mathcal{L}(\widehat{\boldsymbol{\Theta}},\boldsymbol{\beta}_0)}{\partial\boldsymbol{\beta}^{\mathrm{T}}}\boldsymbol{\Delta}_{\boldsymbol{\beta}} + \lambda_{\boldsymbol{\beta}}\|\boldsymbol{\beta}_0\|_1 - \lambda_{\boldsymbol{\beta}}\|\widehat{\boldsymbol{\beta}}\|_1 \tag{14}$$

with probability at least $1-2\max(p,mn)^{-1}$. Let S be the set of indices that $\beta_{0j}\neq 0$, and $\mathbf{v}_S$ be the sub-vector of $\mathbf{v}$ with the elements $j\in S$. Then we have

$$\begin{aligned}\|\boldsymbol{\beta}_0+\boldsymbol{\Delta}_{\boldsymbol{\beta}}\|_1 + \|\boldsymbol{\Delta}_{\boldsymbol{\beta}S}\|_1 &\geq \|\boldsymbol{\beta}_0+\boldsymbol{\Delta}_{\boldsymbol{\beta}}-\boldsymbol{\Delta}_{\boldsymbol{\beta}S}\|_1\\ &= \|\boldsymbol{\beta}_0+\boldsymbol{\Delta}_{\boldsymbol{\beta}S^c}\|_1\\ &= \|\boldsymbol{\beta}_{0S}\|_1+\|\boldsymbol{\Delta}_{\boldsymbol{\beta}S^c}\|_1.\end{aligned}$$

Hence

$$\begin{aligned}\|\boldsymbol{\beta}_0+\boldsymbol{\Delta}_{\boldsymbol{\beta}}\|_1 - \|\boldsymbol{\beta}_0\|_1 &\geq \{\|\boldsymbol{\beta}_0\|_1 - \|\boldsymbol{\Delta}_{\boldsymbol{\beta}S}\|_1\} + \|\boldsymbol{\Delta}_{\boldsymbol{\beta}S^c}\|_1 - \|\boldsymbol{\beta}_0\|_1\\ &= \|\boldsymbol{\Delta}_{\boldsymbol{\beta}S^c}\|_1 - \|\boldsymbol{\Delta}_{\boldsymbol{\beta}S}\|_1.\end{aligned} \tag{15}$$

Combine with (14), we have

$$\begin{aligned}0 &\leq -\frac{\partial\mathcal{L}(\widehat{\boldsymbol{\Theta}},\boldsymbol{\beta}_0)}{\partial\boldsymbol{\beta}^{\mathrm{T}}}\boldsymbol{\Delta}_{\boldsymbol{\beta}} + \lambda_{\boldsymbol{\beta}}(\|\boldsymbol{\Delta}_{\boldsymbol{\beta}S}\|_1 - \|\boldsymbol{\Delta}_{\boldsymbol{\beta}S^c}\|_1)\\ &\leq \|\frac{\partial\mathcal{L}(\widehat{\boldsymbol{\Theta}},\boldsymbol{\beta}_0)}{\partial\boldsymbol{\beta}^{\mathrm{T}}}\|_{op}(\|\boldsymbol{\Delta}_{\boldsymbol{\beta}S}\|_1 + \|\boldsymbol{\Delta}_{\boldsymbol{\beta}S^c}\|_1) + \lambda_{\boldsymbol{\beta}}(\|\boldsymbol{\Delta}_{\boldsymbol{\beta}S}\|_1 - \|\boldsymbol{\Delta}_{\boldsymbol{\beta}S^c}\|_1)\\ &\leq \lambda_{\boldsymbol{\beta}}/2(\|\boldsymbol{\Delta}_{\boldsymbol{\beta}S}\|_1 + \|\boldsymbol{\Delta}_{\boldsymbol{\beta}S^c}\|_1) + \lambda_{\boldsymbol{\beta}}(\|\boldsymbol{\Delta}_{\boldsymbol{\beta}S}\|_1 - \|\boldsymbol{\Delta}_{\boldsymbol{\beta}S^c}\|_1)\\ &= 3\lambda_{\boldsymbol{\beta}}/2\|\boldsymbol{\Delta}_{\boldsymbol{\beta}S}\|_1 - \lambda_{\boldsymbol{\beta}}/2\|\boldsymbol{\Delta}_{\boldsymbol{\beta}S^c}\|_1\end{aligned}$$

which implies $\|\boldsymbol{\Delta}_{\boldsymbol{\beta}S^c}\|_1 \leq 3\|\boldsymbol{\Delta}_{\boldsymbol{\beta}S}\|_1$ and in turn $\|\boldsymbol{\Delta}_{\boldsymbol{\beta}}\|_1 \leq 4\|\boldsymbol{\Delta}_{\boldsymbol{\beta}S}\|_1 \leq 4\sqrt{s}\|\boldsymbol{\Delta}_{\boldsymbol{\beta}}\|_2$ with probability at least $1-2p^{-1}$. This proves the result.

Appendix D. Proof of Theorem 1

First, we note that from $\lambda_{\boldsymbol{\beta}} \geq 2\sqrt{\omega\log\{\max(p,mn)\}/(mn)} + 2(mn)^{-1}d_{E\mathbf{X}} + 4ad_{\mathbf{W}}(mn)^{-1}$ in the theorem statement and Lemma A.5 and A.6, we obtain that

$$\|\frac{\partial\mathcal{L}(\widehat{\boldsymbol{\Theta}},\boldsymbol{\beta}_0)}{\partial\boldsymbol{\beta}^{\mathrm{T}}}\|_{\infty} \leq \frac{\lambda_{\boldsymbol{\beta}}}{2}$$

with probability at least $1-2\{\max(p,mn)\}^{-1}$. Further, because $\|\mathbf{A}\|_{op}\leq\|\mathbf{A}\|_{\infty}$ for any matrix $\mathbf{A}$, we get

$$\|\frac{\partial\mathcal{L}(\widehat{\boldsymbol{\Theta}},\boldsymbol{\beta}_0)}{\partial\boldsymbol{\beta}^{\mathrm{T}}}\|_{op} \leq \frac{\lambda_{\boldsymbol{\beta}}}{2}.$$

When $\boldsymbol{\Delta}_{\boldsymbol{\beta}}^{\mathrm{T}}E\{\mathbf{F}_{\boldsymbol{\beta}}(\mathbf{X},\mathbf{Y})\}\boldsymbol{\Delta}_{\boldsymbol{\beta}} \leq \sigma_{1F}\sqrt{2\log\{\max(p,mn)\}/(mn)}$, from the proof of Lemma A.12, we know

$$\|\boldsymbol{\Delta}_{\boldsymbol{\beta}}\|_2 \leq (4\alpha_{0\boldsymbol{\beta}})^{-1/2}\{2\sigma_{1F}^2\log\{\max(p,mn)\}/(mn)\}^{1/4}.$$

We discuss two cases when $\boldsymbol{\Delta}_{\boldsymbol{\beta}}^{\mathrm{T}} E\{\mathbf{F}_{\boldsymbol{\beta}}(\mathbf{X},\mathbf{Y})\}\boldsymbol{\Delta}_{\boldsymbol{\beta}} \geq \sigma_{1F}\sqrt{2\log\{\max(p,mn)\}/(mn)}$. Case I: $\boldsymbol{\Delta}_{\boldsymbol{\beta}} \notin \mathcal{C}(\gamma)$. Then

$$\|\boldsymbol{\Delta}_{\boldsymbol{\beta}}\|_2^2 \leq \|\boldsymbol{\Delta}_{\boldsymbol{\beta}}\|_\infty \|\boldsymbol{\Delta}_{\boldsymbol{\beta}}\|_1 \gamma \sqrt{\frac{\log\{\max(p,mn)\}}{mn}} \leq 8a\sqrt{s}\|\boldsymbol{\Delta}_{\boldsymbol{\beta}}\|_2 \gamma \sqrt{\frac{\log\{\max(p,mn)\}}{mn}}$$

with probability at least $1-2\max(p,mn)^{-1}$, which implies $\|\boldsymbol{\Delta}_{\boldsymbol{\beta}}\|_2 \leq 8a\sqrt{s}\gamma\sqrt{\log\{\max(p,mn)\}/(mn)}$ with probability at least $1-2\max(p,mn)^{-1}$.

Case II: $\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 \in \mathcal{C}(\gamma)$. Because $\widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\beta}}$ is the minimizer of $\mathcal{L}(\boldsymbol{\Theta},\boldsymbol{\beta}) + \lambda_{\boldsymbol{\Theta}}\|\boldsymbol{\Theta}\|_* + \lambda_{\boldsymbol{\beta}}\|\boldsymbol{\beta}\|_1$, we have

$$\mathcal{L}(\widehat{\boldsymbol{\Theta}},\widehat{\boldsymbol{\beta}}) - \mathcal{L}(\widehat{\boldsymbol{\Theta}},\boldsymbol{\beta}_0) = \frac{\partial \mathcal{L}(\widehat{\boldsymbol{\Theta}},\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}^{\mathrm{T}}}\boldsymbol{\Delta}_{\boldsymbol{\beta}} + \frac{1}{2}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)^{\mathrm{T}} \frac{\partial \mathcal{L}(\widehat{\boldsymbol{\Theta}},\boldsymbol{\beta}^*)}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^{\mathrm{T}}}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)$$

and

$$\begin{aligned}
\boldsymbol{\Delta}_{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{F}_{\boldsymbol{\beta}}(\mathbf{X},\mathbf{Y}|\widehat{\boldsymbol{\Theta}},\boldsymbol{\beta}^*)\boldsymbol{\Delta}_{\boldsymbol{\beta}}/2 &= (\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)^{\mathrm{T}} \frac{\partial^2 \mathcal{L}(\widehat{\boldsymbol{\Theta}},\boldsymbol{\beta}^*)}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^{\mathrm{T}}}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)/2 \\
&= \mathcal{L}(\widehat{\boldsymbol{\Theta}},\widehat{\boldsymbol{\beta}}) - \mathcal{L}(\widehat{\boldsymbol{\Theta}},\boldsymbol{\beta}_0) - \frac{\partial \mathcal{L}(\widehat{\boldsymbol{\Theta}},\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}^{\mathrm{T}}}\boldsymbol{\Delta}_{\boldsymbol{\beta}} \\
&\leq -\frac{\partial \mathcal{L}(\widehat{\boldsymbol{\Theta}},\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}^{\mathrm{T}}}\boldsymbol{\Delta}_{\boldsymbol{\beta}} + \lambda_{\boldsymbol{\beta}}\|\boldsymbol{\beta}_0\|_1 - \lambda_{\boldsymbol{\beta}}\|\widehat{\boldsymbol{\beta}}\|_1 \\
&\leq \|\frac{\partial \mathcal{L}(\widehat{\boldsymbol{\Theta}},\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}^{\mathrm{T}}}\|_\infty \|\boldsymbol{\Delta}_{\boldsymbol{\beta}}\|_1 + \lambda_{\boldsymbol{\beta}}\|\boldsymbol{\beta}_0\|_1 - \lambda_{\boldsymbol{\beta}}\|\widehat{\boldsymbol{\beta}}\|_1 \\
&\leq \lambda_{\boldsymbol{\beta}}/2\|\boldsymbol{\Delta}_{\boldsymbol{\beta}}\|_1 + \lambda_{\boldsymbol{\beta}}\|\boldsymbol{\beta}_0 - \widehat{\boldsymbol{\beta}}\|_1 \\
&\leq 6\lambda_{\boldsymbol{\beta}}\sqrt{s}\|\boldsymbol{\Delta}_{\boldsymbol{\beta}}\|_2 \qquad (16)
\end{aligned}$$

with probability at least $1-2\max(p,mn)^{-1}-2(mn)^{-1}$. The first inequality holds by the second order mean value theorem for $\mathcal{L}(\widehat{\boldsymbol{\Theta}},\boldsymbol{\beta})$ on $\boldsymbol{\beta}$. The third inequality holds by the fact that $\lambda_{\boldsymbol{\beta}} \geq 2\sqrt{\omega\log\{\max(p,mn)\}/(mn)} + 2(mn)^{-1}d_{E\mathbf{X}} + 4ad_{\mathbf{W}}(mn)^{-1}$ in the theorem statement and Lemma A.5 and A.6, and triangular inequality. In the last inequality, we used Lemma A.12. Further, by Lemma A.10, with probability at least $1-\exp\{-C\log\{\max(p,mn)\}\}$,

$$\begin{aligned}
&|\boldsymbol{\Delta}_{\boldsymbol{\beta}}^{\mathrm{T}}/2\mathbf{F}_{\boldsymbol{\beta}}(\mathbf{X},\mathbf{Y}|\widehat{\boldsymbol{\Theta}},\boldsymbol{\beta}^*)\boldsymbol{\Delta}_{\boldsymbol{\beta}}/2 - \boldsymbol{\Delta}_{\boldsymbol{\beta}}^{\mathrm{T}}/2E\{\mathbf{F}_{\boldsymbol{\beta}}(\mathbf{X},\mathbf{Y}|\widehat{\boldsymbol{\Theta}},\boldsymbol{\beta}^*)\}\boldsymbol{\Delta}_{\boldsymbol{\beta}}/2| \\
\leq\ & 2\alpha_{0\boldsymbol{\beta}}\|\boldsymbol{\Delta}_{\boldsymbol{\beta}}/2\|_2^2 + 160(d_{H_2}+d_{S_2}^2)a^2c_0^2/\sqrt{mn},
\end{aligned}$$

so

$$\begin{aligned}
&\boldsymbol{\Delta}_{\boldsymbol{\beta}}^{\mathrm{T}}/2\mathbf{F}_{\boldsymbol{\beta}}(\mathbf{X},\mathbf{Y}|\widehat{\boldsymbol{\Theta}},\boldsymbol{\beta}^*)\boldsymbol{\Delta}_{\boldsymbol{\beta}}/2 - \boldsymbol{\Delta}_{\boldsymbol{\beta}}^{\mathrm{T}}/2E\{\mathbf{F}_{\boldsymbol{\beta}}(\mathbf{X},\mathbf{Y}|\widehat{\boldsymbol{\Theta}},\boldsymbol{\beta}^*)\}\boldsymbol{\Delta}_{\boldsymbol{\beta}}/2 \\
\geq\ & -2\alpha_{0\boldsymbol{\beta}}\|\boldsymbol{\Delta}_{\boldsymbol{\beta}}/2\|_2^2 - 160(d_{H_2}+d_{S_2}^2)a^2c_0^2/\sqrt{mn}.
\end{aligned}$$

Thus,

$$\begin{aligned}
&\boldsymbol{\Delta}_{\boldsymbol{\beta}}^{\mathrm{T}}/2\mathbf{F}_{\boldsymbol{\beta}}(\mathbf{X},\mathbf{Y}|\widehat{\boldsymbol{\Theta}},\boldsymbol{\beta}^*)\boldsymbol{\Delta}_{\boldsymbol{\beta}}/2 \\
\geq\ & \alpha_{\min}(E[R_{ij}E\{S_2^2(Y_{ij},\mathbf{X}_{ij}|\widehat{\boldsymbol{\Theta}},\boldsymbol{\beta}^*)\mathbf{X}_{ij}\mathbf{X}_{ij}^{\mathrm{T}}|Y_{ij}\} - R_{ij}E\{S_2(Y_{ij},\mathbf{X}_{ij}|\widehat{\boldsymbol{\Theta}},\boldsymbol{\beta}^*)\mathbf{X}_{ij}|Y_{ij}\}^{\otimes 2}])\|\boldsymbol{\Delta}_{\boldsymbol{\beta}}/2\|_2^2 \\
& -2^{-1}\alpha_{0\boldsymbol{\beta}}\|\boldsymbol{\Delta}_{\boldsymbol{\beta}}\|_2^2 - 160(d_{H_2}+d_{S_2}^2)a^2c_0^2/\sqrt{mn} \\
=\ & 2^{-1}\alpha_{0\boldsymbol{\beta}}\|\boldsymbol{\Delta}_{\boldsymbol{\beta}}\|_2^2 - 160(d_{H_2}+d_{S_2}^2)a^2c_0^2/\sqrt{mn}
\end{aligned}$$

with probability at least $1-\exp\{-C\log\{\max(p,mn)\}\}$, which implies

$$\boldsymbol{\Delta}_{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{F}_{\boldsymbol{\beta}}(\mathbf{X},\mathbf{Y}|\widehat{\boldsymbol{\Theta}},\boldsymbol{\beta}^*)\boldsymbol{\Delta}_{\boldsymbol{\beta}}/2 \geq \alpha_{0\boldsymbol{\beta}}\|\boldsymbol{\Delta}_{\boldsymbol{\beta}}\|_2^2 - 320(d_{H_2}+d_{S_2}^2)a^2c_0^2/\sqrt{mn},$$

with probability at least $1-\exp\{-C\log\{\max(p,mn)\}\}$. Combining with (16), we have

$$\alpha_{0\boldsymbol{\beta}}\|\boldsymbol{\Delta}_{\boldsymbol{\beta}}\|_2^2 \le 6\lambda_{\boldsymbol{\beta}}\sqrt{s}\|\boldsymbol{\Delta}_{\boldsymbol{\beta}}\|_2 + 320(d_{H_2}+d_{S_2}^2)a^2c_0^2/(\sqrt{mn})$$

with probability at least $1-2\max(p,mn)^{-1}-2(mn)^{-1}-\exp\{-C\log\{\max(p,mn)\}\}$. Then

$$\|\boldsymbol{\Delta}_{\boldsymbol{\beta}}\|_2^2 - 6\lambda_{\boldsymbol{\beta}}\sqrt{s}\|\boldsymbol{\Delta}_{\boldsymbol{\beta}}\|_2/\alpha_{0\boldsymbol{\beta}} \le 320(d_{H_2}+d_{S_2}^2)a^2c_0^2/(\alpha_{0\boldsymbol{\beta}}\sqrt{mn}).$$

This leads to

$$(\|\boldsymbol{\Delta}_{\boldsymbol{\beta}}\|_2^2 - 3\lambda_{\boldsymbol{\beta}}\sqrt{s}/\alpha_{0\boldsymbol{\beta}})^2 \le 320(d_{H_2}+d_{S_2}^2)a^2c_0^2/(\alpha_{0\boldsymbol{\beta}}\sqrt{mn}) + (3\lambda_{\boldsymbol{\beta}}\sqrt{s}/\alpha_{0\boldsymbol{\beta}})^2,$$

Hence

$$\|\boldsymbol{\Delta}_{\boldsymbol{\beta}}\|_2 \quad\le\quad \{320(d_{H_2}+d_{S_2}^2)a^2c_0^2/(\alpha_{0\boldsymbol{\beta}}\sqrt{mn}) + (3\lambda_{\boldsymbol{\beta}}\sqrt{s}/\alpha_{0\boldsymbol{\beta}})^2\}^{1/2} + 3\lambda_{\boldsymbol{\beta}}\sqrt{s}/\alpha_{0\boldsymbol{\beta}},$$

with probability at least $1-2\max(p,mn)^{-1}-2(mn)^{-1}-\exp\{-C\log(\max(p,mn))\}$. Combine with the order in Case I and before Case I, we have

$$\begin{aligned}
&\|\boldsymbol{\Delta}_{\boldsymbol{\beta}}\|_2 \\
\le\ & \max\Big(\{320(d_{H_2}+d_{S_2}^2)a^2c_0^2/(\alpha_{0\boldsymbol{\beta}}\sqrt{mn}) + (3\lambda_{\boldsymbol{\beta}}\sqrt{s}/\alpha_{0\boldsymbol{\beta}})^2\}^{1/2} + 3\lambda_{\boldsymbol{\beta}}\sqrt{s}/\alpha_{0\boldsymbol{\beta}}, 8a\sqrt{s}\gamma\sqrt{\log\{\max(p,mn)\}/(mn)}, \\
& (4\alpha_{0\boldsymbol{\beta}})^{-1/2}\{2\sigma_{1F}^2\log\{\max(p,mn)\}/(mn)\}^{1/4}\Big)
\end{aligned}$$

with probability at least $1-4\max(p,mn)^{-1}-2(mn)^{-1}-2\exp\{-C\log\{\max(p,mn)\}\}$. This proves the result.

Appendix E. Lemmas for Theorem 2

Let $\rho(\nu,D)\equiv D^2/\{\nu\sqrt{d\log(d)/(mn)}\} = D^2/\{\nu\sqrt{\log(d)/d}\}$. Define sets

$$\begin{aligned}
\mathcal{C}_{\boldsymbol{\Theta}}(\nu) &\equiv \left\{0\neq\boldsymbol{\Theta}\in\mathbb{R}^{m\times n}\Big|\frac{\|\boldsymbol{\Theta}\|_{\max}}{\|\boldsymbol{\Theta}\|_F}\frac{\|\boldsymbol{\Theta}\|_*}{\|\boldsymbol{\Theta}\|_F}\le\frac{1}{\nu}\sqrt{\frac{mn}{d\log(d)}}\right\} \\
&= \left\{0\neq\boldsymbol{\Theta}\in\mathbb{R}^{m\times n}\Big|\frac{\|\boldsymbol{\Theta}\|_{\max}}{\|\boldsymbol{\Theta}\|_F}\frac{\|\boldsymbol{\Theta}\|_*}{\|\boldsymbol{\Theta}\|_F}\le\frac{1}{\nu}\sqrt{\frac{d}{\log(d)}}\right\}, \\
\mathbb{B}_{\boldsymbol{\Theta}}(D) &\equiv \{\boldsymbol{\Theta}\in\mathcal{C}_{\boldsymbol{\Theta}}(\nu)|\|\boldsymbol{\Theta}\|_{\max}=a,\|\boldsymbol{\Theta}\|_F\le D,\|\boldsymbol{\Theta}\|_*\le\rho(\nu,D)\}, \\
\bar{\mathbb{B}}_{\boldsymbol{\Theta}}(D) &\equiv \{\boldsymbol{\Theta}\in\mathcal{C}_{\boldsymbol{\Theta}}(\nu)|\|\boldsymbol{\Theta}\|_{\max}\le a,\|\boldsymbol{\Theta}\|_F\le D,\|\boldsymbol{\Theta}\|_*\le\rho(\nu,D)\}.
\end{aligned}$$

Let $\boldsymbol{\Theta}^1,\ldots,\boldsymbol{\Theta}^{N_{\boldsymbol{\Theta}}(\delta)}$ be a δ-covering of $\bar{\mathbb{B}}_{\boldsymbol{\Theta}}(D)$ in Frobenius norm. By definition, given an arbitrary $\boldsymbol{\Theta}\in\bar{\mathbb{B}}_{\boldsymbol{\Theta}}(D)$, there is some index $k\in\{1,\ldots,N_{\boldsymbol{\Theta}}(\delta)\}$ and a difference matrix $\boldsymbol{\Delta}_{\boldsymbol{\Theta}}$ with $\|\boldsymbol{\Delta}_{\boldsymbol{\Theta}}\|_F\le\delta$ such that $\boldsymbol{\Theta}=\boldsymbol{\Theta}^k+\boldsymbol{\Delta}_{\boldsymbol{\Theta}}$.

Lemma A.13 *Assume Conditions (C1) and (C6) hold. Then there is a positive constant c_d such that for sufficiently large n,m,d,*

$$\Pr\left(\|\frac{\partial\mathcal{L}(\boldsymbol{\Theta}_0,\boldsymbol{\beta}_0)}{\partial\boldsymbol{\Theta}}\|_{op}\ge c_d\sqrt{d\log(d)/(mn)}\right)<d^{-1},$$

or equivalently,

$$\Pr\left(\|\frac{\partial\mathcal{L}(\boldsymbol{\Theta}_0,\boldsymbol{\beta}_0)}{\partial\boldsymbol{\Theta}}\|_{op}\ge c_d\sqrt{\log(d)/d}\right)<d^{-1}.$$

Proof: First note that

$$\begin{aligned}
&\frac{\partial \mathcal{L}(\boldsymbol{\Theta}_0, \boldsymbol{\beta}_0)}{\partial \boldsymbol{\Theta}} \\
= \; & -(mn)^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m} R_{ij} \left[\left\{ \frac{f_2(Y_{ij}, \boldsymbol{\Theta}_{0ij} + \boldsymbol{\beta}_0^{\mathrm{T}} \mathbf{X}_{ij})}{f(Y_{ij}, \boldsymbol{\Theta}_{0ij} + \boldsymbol{\beta}_0^{\mathrm{T}} \mathbf{X}_{ij})} - \frac{\int f_2(Y_{ij}, \boldsymbol{\Theta}_{0ij} + \boldsymbol{\beta}^{\mathrm{T}} \mathbf{X}) g(\mathbf{X}) d\mathbf{X}}{\int f(Y_{ij}, \boldsymbol{\Theta}_{0ij} + \boldsymbol{\beta}_0^{\mathrm{T}} \mathbf{X}) g(\mathbf{X}) d\mathbf{X}} \right\} \mathbf{e}_i \mathbf{e}_j^{\mathrm{T}} \right] \\
= \; & (mn)^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m} R_{ij} [S_2(Y_{ij}, \mathbf{X}_{ij} | \boldsymbol{\Theta}_0, \boldsymbol{\beta}_0) - E\{S_2(Y_{ij}, \mathbf{X}_{ij} | \boldsymbol{\Theta}_0, \boldsymbol{\beta}_0) | Y_{ij}\}] \mathbf{e}_i \mathbf{e}_j^{\mathrm{T}}.
\end{aligned}$$

Let $\mathbf{W}^{ij} = R_{ij}[S_2(Y_{ij}, \mathbf{X}_{ij}|\boldsymbol{\Theta}_0, \boldsymbol{\beta}_0) - E\{S_2(Y_{ij}, \mathbf{X}_{ij}|\boldsymbol{\Theta}_0, \boldsymbol{\beta}_0)|Y_{ij}\}]\mathbf{e}_i\mathbf{e}_j^{\mathrm{T}}/(mn)$. Then $\mathbf{W}^{ij}$ is a mean zero random matrix with $\|\mathbf{W}^{ij}\|_{op} \leq \sqrt{\|\mathbf{W}^{ij}\|_1 \|\mathbf{W}^{ij}\|_\infty} \leq 2d_{S_2}/(mn)$ by Condition (C6). Further

$$\begin{aligned}
& \|E(\mathbf{W}^{ij} \mathbf{W}^{ij\mathrm{T}})\|_{op} \\
= \; & \|E\left(R_{ij}[S_2(Y_{ij}, \mathbf{X}_{ij}|\boldsymbol{\Theta}_0, \boldsymbol{\beta}_0) - E\{S_2(Y_{ij}, \mathbf{X}_{ij}|\boldsymbol{\Theta}_0, \boldsymbol{\beta}_0)|Y_{ij}\}]^2 \mathbf{e}_i \mathbf{e}_i^{\mathrm{T}}/(mn)^2 \right) \|_{op} \\
\leq \; & \|4d_{S_2}^2/(mn)^2 \mathbf{e}_i \mathbf{e}_i^{\mathrm{T}}\|_{op} \\
= \; & 4d_{S_2}^2/(mn)^2.
\end{aligned}$$

Similarly,

$$\|E(\mathbf{W}^{ij\mathrm{T}} \mathbf{W}^{ij})\|_{op} \leq 4d_{S_2}^2/(mn)^2.$$

Hence

$$\sigma_W^2 \equiv mn \max(\|E(\mathbf{W}^{ij\mathrm{T}} \mathbf{W}^{ij})\|_{op}, \|E(\mathbf{W}^{ij} \mathbf{W}^{ij\mathrm{T}})\|_{op}) \leq 4d_{S_2}^2/(mn).$$

Therefore, by Lemma A.3 we have

$$\begin{aligned}
& \Pr(\|\frac{\partial \mathcal{L}(\boldsymbol{\Theta}_0, \boldsymbol{\beta}_0)}{\partial \boldsymbol{\Theta}}\|_{op} \geq t) \\
\leq \; & \Pr(\|\sum_{i=1}^{n} \sum_{j=1}^{m} \mathbf{W}^{ij}\|_{op} \geq t) \\
\leq \; & mn \max\left(\exp[-t^2/\{16d_{S_2}^2/(mn)\}], \exp[-t/\{4d_{S_2}/(mn)\}] \right).
\end{aligned}$$

Now let $t = c_d\sqrt{d\log(d)/(mn)}$ for $c_d = 4d_{S_2}$. Then we have

$$\begin{aligned}
& \Pr(\|\frac{\partial \mathcal{L}(\boldsymbol{\Theta}_0, \boldsymbol{\beta}_0)}{\partial \boldsymbol{\Theta}}\|_{op} \geq t) \\
\leq \; & mn \max\left(\exp[-c_d^2 d\log(d)/(mn)/\{16d_{S_2}^2/(mn)\}], \exp[-c_d\sqrt{d\log(d)/(mn)}/\{4d_{S_2}/(mn)\}] \right) \\
\leq \; & mn \max\left(\exp\{-d\log(d)\}, \exp\{-mn\sqrt{d\log(d)mn}\} \right) \\
= \; & \exp\{-d\log(d) + \log(mn)\} \\
\leq \; & d^{-1}
\end{aligned}$$

for $d, m, n \to \infty$.

Lemma A.14 *Assume Conditions (C1)–(C6) hold. Let*

$$\lambda_{\boldsymbol{\beta}} \geq 2\sqrt{\omega \log\{\max(p, mn)\}/mn} + 2(mn)^{-1} d_{E\mathbf{X}} + 4a d_{\mathbf{W}} (mn)^{-1}.$$

Then

$$
\begin{aligned}
& \|\frac{\partial \mathcal{L}(\mathbf{\Theta}_0, \widehat{\boldsymbol{\beta}})}{\partial \mathbf{\Theta}} - \frac{\partial \mathcal{L}(\mathbf{\Theta}_0, \boldsymbol{\beta}_0)}{\partial \mathbf{\Theta}}\|_{op} \\
\leq \quad & (2d_{H_2} + 2d_{S_2}^2) \max\Big([320(d_{H_2} + d_{S_2}^2) a^2 c_0^2 / (\alpha_{0\boldsymbol{\beta}} \sqrt{mn}) + \{3\lambda_{\boldsymbol{\beta}} \sqrt{s} / \alpha_{0\boldsymbol{\beta}}\}^2]^{1/2} \\
& + 3\lambda_{\boldsymbol{\beta}} \sqrt{s} / \alpha_{0\boldsymbol{\beta}}, 8a\sqrt{s}\gamma \sqrt{\log\{\max(p, mn)\}/(mn)}, (4\alpha_{0\boldsymbol{\beta}})^{-1/2} \{2\sigma_{1F}^2 \log\{\max(p, mn)\}/(mn)\}^{1/4} \Big)
\end{aligned}
$$

Proof: First note that

$$
\begin{aligned}
& \|\frac{\partial \mathcal{L}(\mathbf{\Theta}_0, \widehat{\boldsymbol{\beta}})}{\partial \mathbf{\Theta}} - \frac{\partial \mathcal{L}(\mathbf{\Theta}_0, \boldsymbol{\beta}_0)}{\partial \mathbf{\Theta}}\|_{op} \\
= \quad & \|\frac{\partial^2 \mathcal{L}(\mathbf{\Theta}_0, \boldsymbol{\beta}^*)}{\partial \mathbf{\Theta} \partial \boldsymbol{\beta}^{\mathrm{T}}} (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\|_{op} \\
= \quad & \|(mn)^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m} R_{ij} \left(H_2(Y_{ij}, \mathbf{X}_{ij} | \mathbf{\Theta}_0, \boldsymbol{\beta}^*) - E\{H_2(Y_{ij}, \mathbf{X}_{ij} | \mathbf{\Theta}_0, \boldsymbol{\beta}^*) | Y_{ij}\} \right. \\
& \left. + E\left\{ S_2(Y_{ij}, \mathbf{X}_{ij} | \mathbf{\Theta}_0, \boldsymbol{\beta}^*)^{\otimes 2} | Y_{ij} \right\} - [E\{S_2(Y_{ij}, \mathbf{X}_{ij} | \mathbf{\Theta}_0, \boldsymbol{\beta}^*) | Y_{ij}\}]^{\otimes 2} \right) \mathbf{e}_i \mathbf{e}_j^{\mathrm{T}} \{\mathbf{X}_{ij}^{\mathrm{T}} (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\}\|_{op} \\
\leq \quad & (mn)^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m} \|R_{ij} \left(H_2(Y_{ij}, \mathbf{X}_{ij} | \mathbf{\Theta}_0, \boldsymbol{\beta}^*) - E\{H_2(Y_{ij}, \mathbf{X}_{ij} | \mathbf{\Theta}_0, \boldsymbol{\beta}^*) | Y_{ij}\} \right. \\
& \left. + E\left\{ S_2(Y_{ij}, \mathbf{X}_{ij} | \mathbf{\Theta}_0, \boldsymbol{\beta}^*)^{\otimes 2} | Y_{ij} \right\} - [E\{S_2(Y_{ij}, \mathbf{X}_{ij} | \mathbf{\Theta}_0, \boldsymbol{\beta}^*) | Y_{ij}\}]^{\otimes 2} \right) \mathbf{e}_i \mathbf{e}_j^{\mathrm{T}} \|_{op} \max_{ij} |\mathbf{X}_{ij}^{\mathrm{T}} (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)| \\
\leq \quad & (mn)^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m} \|(2d_{H_2} + 2d_{S_2}^2) \mathbf{e}_i \mathbf{e}_j^{\mathrm{T}} \|_{op} \max_{ij} |\mathbf{X}_{ij}^{\mathrm{T}} (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)| \\
= \quad & (2d_{H_2} + 2d_{S_2}^2) \max\Big([320(d_{H_2} + d_{S_2}^2) a^2 c_0^2 / (\alpha_{0\boldsymbol{\beta}} \sqrt{mn}) + \{3\lambda_{\boldsymbol{\beta}} \sqrt{s} / \alpha_{0\boldsymbol{\beta}}\}^2]^{1/2} \\
& + 3\lambda_{\boldsymbol{\beta}} \sqrt{s} / \alpha_{0\boldsymbol{\beta}}, 8a\sqrt{s}\gamma \sqrt{\log\{\max(p, mn)\}/(mn)}, (4\alpha_{0\boldsymbol{\beta}})^{-1/2} \{2\sigma_{1F}^2 \log\{\max(p, mn)\}/(mn)\}^{1/4} \Big),
\end{aligned}
$$

where $\boldsymbol{\beta}^*$ is on the line in between $\widehat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}_0$. The last inequality holds by Condition (C5) and Theorem 1. This proves the result.

Lemma A.15 *The δ-covering number of $\bar{\mathbb{B}}_{\mathbf{\Theta}}(\delta)$ satisfies*

$$\log N_{\mathbf{\Theta}}(\delta) \leq \frac{144\rho(\nu, D)^2}{\delta^2} \max(n, m).$$

Proof: First using the same arguments as those lead to (39) in Negahban and Wainwright (2012), we have

$$\sqrt{\log N_{\mathbf{\Theta}}(\delta)} \leq \frac{3\rho(\nu, D)}{\delta} E(\|\mathbf{G}\|_2),$$

where $\mathbf{G}$ is a random matrix containing independent identically distribution standard normal entries. Further,

$$E(\|\mathbf{G}\|_2) \leq 4 \max(\sqrt{n}, \sqrt{m})$$

by the results in Section 3.1 in Bandeira et al. (2016). We have

$$\log N_{\mathbf{\Theta}}(\delta) \leq \frac{144\rho(\nu, D)^2}{\delta^2} \max(n, m).$$

This proves the result.

Lemma A.16 *Assume Conditions (C1), (C5), (C6). Let $\delta = D/\xi$. Let $\mathbf{\Theta}^l \in \bar{\mathbb{B}}_{\mathbf{\Theta}}(D)$, and $\mathbf{b}^l, l = 1, \dots N_{\mathbf{\Theta}}(\delta)$ be δ-covering of $\bar{\mathbb{B}}_{\mathbf{\Theta}}(D)$ in Frobenius norm. Assume*

$$\frac{mn\delta^4}{128(d_{H_2}+d_{S_2}^2)^2a^2D^2} \geq 2\frac{144\rho(\nu,D)^2}{\delta^2}\max(n,m).$$

Recall that

$$\begin{aligned}
&\mathrm{vec}(\mathbf{A})^{\mathrm{T}}\mathbf{F}_{\mathbf{\Theta}}(\mathbf{X},\mathbf{Y}|\widetilde{\mathbf{\Theta}},\widetilde{\boldsymbol{\beta}})\mathrm{vec}(\mathbf{B}) \\
= \;& (mn)^{-1}\left(\sum_{i=1}^{n}\sum_{j=1}^{m}\mathrm{vec}(\mathbf{A})^{\mathrm{T}}R_{ij}[H_2(Y_{ij},\mathbf{X}_{ij}|\widetilde{\mathbf{\Theta}},\widetilde{\boldsymbol{\beta}})\mathrm{vec}(\mathbf{z}_{ij})\mathrm{vec}(\mathbf{z}_{ij})^{\mathrm{T}}\right. \\
&-E\{H_2(Y_{ij},\mathbf{X}_{ij}|\widetilde{\mathbf{\Theta}},\widetilde{\boldsymbol{\beta}})\mathrm{vec}(\mathbf{z}_{ij})\mathrm{vec}(\mathbf{z}_{ij})^{\mathrm{T}}|Y_{ij}\} \\
&\left.+E\{S_2^2(Y_{ij},\mathbf{X}_{ij},\widetilde{\mathbf{\Theta}},\widetilde{\boldsymbol{\beta}})\mathrm{vec}(\mathbf{z}_{ij})\mathrm{vec}(\mathbf{z}_{ij})^{\mathrm{T}}|Y_{ij}\} - E\{S_2(Y_{ij},\mathbf{X}_{ij}|\widetilde{\mathbf{\Theta}},\widetilde{\boldsymbol{\beta}})\mathrm{vec}(\mathbf{z}_{ij})|Y_{ij}\}^{\otimes 2}]\mathrm{vec}(\mathbf{B})\right).
\end{aligned}$$

Then there are positive constants c_{d2}, c_{d3} such that

$$\begin{aligned}
&\Pr\left(\sup_{k=1,\dots,N(\delta)}\left|\mathrm{vec}(\mathbf{\Theta}^k)^{\mathrm{T}}\mathbf{F}_{\mathbf{\Theta}}(\mathbf{X},\mathbf{Y}|\widetilde{\mathbf{\Theta}},\widetilde{\boldsymbol{\beta}})\mathrm{vec}(\mathbf{\Theta}^k)\right.\right. \\
&-\mathrm{vec}(\mathbf{\Theta}^k)^{\mathrm{T}}(mn)^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m}E[R_{ij}E\{S_2^2(Y_{ij},\mathbf{X}_{ij}|\widetilde{\mathbf{\Theta}},\widetilde{\boldsymbol{\beta}})\mathbf{z}_{ij}\mathbf{z}_{ij}^{\mathrm{T}}|Y_{ij}\} \\
&\left.\left.-R_{ij}E\{S_2(Y_{ij},\mathbf{X}_{ij}|\widetilde{\mathbf{\Theta}},\widetilde{\boldsymbol{\beta}})\mathbf{z}_{ij}|Y_{ij}\}^{\otimes 2}]\mathrm{vec}(\mathbf{\Theta}^k)\right| \geq \delta^2 + \frac{32(d_{H_2}+d_{S_2}^2)a^2}{\sqrt{mn}}\right) \\
\leq \;& 4\exp\left(-c_{d2}D^2mn\right),
\end{aligned}$$

$$\begin{aligned}
&\Pr\left(\sup_{k=1,\dots,N(\delta)}|\mathrm{vec}(\mathbf{\Theta}^k)^{\mathrm{T}}\mathbf{F}_{\mathbf{\Theta}}(\mathbf{X},\mathbf{Y}|\widetilde{\mathbf{\Theta}},\widetilde{\boldsymbol{\beta}})\mathrm{vec}(\mathbf{\Theta}-\mathbf{\Theta}^k)\right. \\
&-\mathrm{vec}(\mathbf{\Theta}^k)^{\mathrm{T}}(mn)^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m}E[R_{ij}E\{S_2^2(Y_{ij},\mathbf{X}_{ij}|\widetilde{\mathbf{\Theta}},\widetilde{\boldsymbol{\beta}})\mathbf{z}_{ij}\mathbf{z}_{ij}^{\mathrm{T}}|Y_{ij}\} \\
&\left.-R_{ij}E\{S_2(Y_{ij},\mathbf{X}_{ij}|\widetilde{\mathbf{\Theta}},\widetilde{\boldsymbol{\beta}})\mathbf{z}_{ij}|Y_{ij}\}^{\otimes 2}]\mathrm{vec}(\mathbf{\Theta}-\mathbf{\Theta}^k)| \geq \delta^2 + \frac{64(d_{H_2}+d_{S_2}^2)a^2}{\sqrt{mn}}\right) \\
\leq \;& 4\exp\left(-c_{d3}D^2mn\right).
\end{aligned}$$

Proof: First we have

$$\begin{aligned}
&|\mathrm{vec}(\mathbf{\Theta})^{k^{\mathrm{T}}}R_{ij}[H_2(Y_{ij},\mathbf{X}_{ij}|\widetilde{\mathbf{\Theta}},\widetilde{\boldsymbol{\beta}})\mathrm{vec}(\mathbf{z}_{ij})\mathrm{vec}(\mathbf{z}_{ij})^{\mathrm{T}} - E\{H_2(Y_{ij},\mathbf{X}|\widetilde{\mathbf{\Theta}},\widetilde{\boldsymbol{\beta}})\mathrm{vec}(\mathbf{z}_{ij})\mathrm{vec}(\mathbf{z}_{ij})^{\mathrm{T}}|Y_{ij}\} \\
&+E\{S_2^2(Y_{ij},\mathbf{X}|\widetilde{\mathbf{\Theta}},\widetilde{\boldsymbol{\beta}})\mathrm{vec}(\mathbf{z}_{ij})\mathrm{vec}(\mathbf{z}_{ij})^{\mathrm{T}}|Y_{ij}\} - E\{S_2(Y_{ij},\mathbf{X}|\widetilde{\mathbf{\Theta}},\widetilde{\boldsymbol{\beta}})\mathrm{vec}(\mathbf{z}_{ij})|Y_{ij}\}^{\otimes 2}]\mathrm{vec}(\mathbf{\Theta})^k| \\
\leq \;& \{2d_{H_2}\|\mathbf{\Theta}^k\|_F\|\mathrm{vec}(\mathbf{z}_{ij})\|_2\|\mathbf{\Theta}^k\|_{\max}\|\mathrm{vec}(\mathbf{z}_{ij})\|_1 \\
&+2d_{S_2}^2\|\mathbf{\Theta}^k\|_F\|\mathrm{vec}(\mathbf{z}_{ij})\|_2\|\mathbf{\Theta}^k\|_{\max}\|\mathrm{vec}(\mathbf{z}_{ij})\|_1\} \\
= \;& 2(d_{H_2}+d_{S_2}^2)aD
\end{aligned}$$

and

$$|\mathrm{vec}(\mathbf{\Theta})^{k^{\mathrm{T}}}R_{ij}[H_2(Y_{ij},\mathbf{X}_{ij}|\widetilde{\mathbf{\Theta}},\widetilde{\boldsymbol{\beta}})\mathrm{vec}(\mathbf{z}_{ij})\mathrm{vec}(\mathbf{z}_{ij})^{\mathrm{T}} - E\{H_2(Y_{ij},\mathbf{X}|\widetilde{\mathbf{\Theta}},\widetilde{\boldsymbol{\beta}})\mathrm{vec}(\mathbf{z}_{ij})\mathrm{vec}(\mathbf{z}_{ij})^{\mathrm{T}}|Y_{ij}\}$$

$$
\begin{aligned}
& + E\{S_2^2(Y_{ij}, \mathbf{X}|\widetilde{\boldsymbol{\Theta}}, \widetilde{\boldsymbol{\beta}})\mathrm{vec}(\mathbf{z}_{ij})\mathrm{vec}(\mathbf{z}_{ij})^{\mathrm{T}}|Y_{ij}\} - E\{S_2(Y_{ij}, \mathbf{X}|\widetilde{\boldsymbol{\Theta}}, \widetilde{\boldsymbol{\beta}})\mathrm{vec}(\mathbf{z}_{ij})|Y_{ij}\}^{\otimes 2}]\mathrm{vec}(\boldsymbol{\Theta})^k| \\
\leq\ & \{2d_{H_2}\|\boldsymbol{\Theta}^k\|_{\max}\|\mathrm{vec}(\mathbf{z}_{ij})\|_1\|\boldsymbol{\Theta}^k\|_{\max}\|\mathrm{vec}(\mathbf{z}_{ij})\|_1 \\
& + 2d_{S_2}^2\|\boldsymbol{\Theta}^k\|_{\max}\|\mathrm{vec}(\mathbf{z}_{ij})\|_1\|\boldsymbol{\Theta}^k\|_{\max}\|\mathrm{vec}(\mathbf{z}_{ij})\|_1\} \\
=\ & 2(d_{H_2} + d_{S_2}^2)a^2.
\end{aligned}
$$

by Condition (C6). We now use Lemma A.2, where we treat each summand in $\mathrm{vec}(\boldsymbol{\Theta}^k)^{\mathrm{T}}\mathbf{F}_{\boldsymbol{\Theta}}(\mathbf{X}, \mathbf{Y}|\widetilde{\boldsymbol{\Theta}}, \widetilde{\boldsymbol{\beta}})\mathrm{vec}(\boldsymbol{\Theta}^k)$ as Y_i in Lemma A.2, and set $(u_i, v_i) = \{-2(d_{H_2}+d_{S_2}^2)aD, 2(d_{H_2}+d_{S_2}^2)aD\}$ and $(u_{1i}, v_{1i}) = \{-2(d_{H_2}+d_{S_2}^2)a^2, 2(d_{H_2}+d_{S_2}^2)a^2\}$. Consider $\mathcal{T}$ that contains only one vector $\mathbf{t}$, where $\mathbf{t}$ is the mn dimensional vector with element $(mn)^{-1}$. Then Z in Lemma A.2 is $\mathrm{vec}(\boldsymbol{\Theta}^k)^{\mathrm{T}}\mathbf{F}_{\boldsymbol{\Theta}}(\mathbf{X}, \mathbf{Y}|\widetilde{\boldsymbol{\Theta}}, \widetilde{\boldsymbol{\beta}})\mathrm{vec}(\boldsymbol{\Theta}^k)$ and the σ in Lemma A.2 (7) is $\sigma = [\sum_{i=1}^n \sum_{j=1}^m (mn)^{-2}\{4(d_{H_2}+d_{S_2}^2)aD\}^2]^{1/2} = 4(d_{H_2}+d_{S_2}^2)aD/\sqrt{mn} = O(1)$ by Condition (C6). Further, the σ in Lemma A.2 (8) is $\sigma_1 = [\sum_{i=1}^n \sum_{j=1}^m (mn)^{-2}\{4(d_{H_2}+d_{S_2}^2)a^2\}^2]^{1/2} = 4(d_{H_2} + d_{S_2}^2)a^2/\sqrt{mn}$. Hence let m_F be the median of $\mathrm{vec}(\boldsymbol{\Theta}^k)^{\mathrm{T}}\mathbf{F}_{\boldsymbol{\Theta}}(\mathbf{X}, \mathbf{Y}|\widetilde{\boldsymbol{\Theta}}, \widetilde{\boldsymbol{\beta}})\mathrm{vec}(\boldsymbol{\Theta}^k)$, we have

$$
\begin{aligned}
& \Pr\left(|\mathrm{vec}(\boldsymbol{\Theta}^k)^{\mathrm{T}}\mathbf{F}_{\boldsymbol{\Theta}}(\mathbf{X}, \mathbf{Y}|\widetilde{\boldsymbol{\Theta}}, \widetilde{\boldsymbol{\beta}})\mathrm{vec}(\boldsymbol{\Theta}^k) - \mathrm{vec}(\boldsymbol{\Theta}^k)^{\mathrm{T}}E\{\mathbf{F}_{\boldsymbol{\Theta}}(\mathbf{X}, \mathbf{Y}|\widetilde{\boldsymbol{\Theta}}, \widetilde{\boldsymbol{\beta}})\}\mathrm{vec}(\boldsymbol{\Theta}^k)| \geq \delta^2 + \frac{32(d_{H_2} + d_{S_2}^2)a^2}{\sqrt{mn}}\right) \\
\leq\ & \Pr\Big(|\mathrm{vec}(\boldsymbol{\Theta}^k)^{\mathrm{T}}\mathbf{F}_{\boldsymbol{\Theta}}(\mathbf{X}, \mathbf{Y}|\widetilde{\boldsymbol{\Theta}}, \widetilde{\boldsymbol{\beta}})\mathrm{vec}(\boldsymbol{\Theta}^k) - m_F\}| + |\mathrm{vec}(\boldsymbol{\Theta}^k)^{\mathrm{T}}E\{\mathbf{F}_{\boldsymbol{\Theta}}(\mathbf{X}, \mathbf{Y}|\widetilde{\boldsymbol{\Theta}}, \widetilde{\boldsymbol{\beta}})\}\mathrm{vec}(\boldsymbol{\Theta}^k) - m_F\}| \geq \delta^2 \\
& + \frac{32(d_{H_2} + d_{S_2}^2)a^2}{\sqrt{mn}}\Big) \\
\leq\ & \Pr\left(|\mathrm{vec}(\boldsymbol{\Theta}^k)^{\mathrm{T}}\mathbf{F}_{\boldsymbol{\Theta}}(\mathbf{X}, \mathbf{Y}|\widetilde{\boldsymbol{\Theta}}, \widetilde{\boldsymbol{\beta}})\mathrm{vec}(\boldsymbol{\Theta}^k) - m_F\}| \geq \delta^2 + \frac{32(d_{H_2} + d_{S_2}^2)a^2}{\sqrt{mn}} - \frac{16\sqrt{\pi}(d_{H_2} + d_{S_2}^2)a^2}{\sqrt{mn}}\right) \\
\leq\ & \Pr\Big(|\mathrm{vec}(\boldsymbol{\Theta}^k)^{\mathrm{T}}\mathbf{F}_{\boldsymbol{\Theta}}(\mathbf{X}, \mathbf{Y}|\widetilde{\boldsymbol{\Theta}}, \widetilde{\boldsymbol{\beta}})\mathrm{vec}(\boldsymbol{\Theta}^k) - m_F| \geq \delta^2\Big) \\
\leq\ & 4\exp\left\{\frac{-mn\delta^4}{64(d_{H_2} + d_{S_2}^2)^2a^2D^2}\right\}.
\end{aligned}
$$

Now for $\boldsymbol{\Theta}^k \in \bar{\mathbb{B}}(D)$, $k = 1, \ldots, N_{\boldsymbol{\Theta}}(\delta)$, we have

$$
\begin{aligned}
& \Pr\Bigg\{\sup_{k=1,\ldots,N(\delta)} |\mathrm{vec}(\boldsymbol{\Theta}^k)^{\mathrm{T}}\mathbf{F}_{\boldsymbol{\Theta}}(\boldsymbol{\Theta}^k, \mathbf{X}, \mathbf{Y}|\widetilde{\boldsymbol{\Theta}}, \widetilde{\boldsymbol{\beta}})\mathrm{vec}(\boldsymbol{\Theta}^k) \\
& -(mn)^{-1}\sum_{i=1}^n\sum_{j=1}^m \mathrm{vec}(\boldsymbol{\Theta})^{k^{\mathrm{T}}} E[R_{ij}E\{S_2^2(Y_{ij}, \mathbf{X}_{ij}|\widetilde{\boldsymbol{\Theta}}, \widetilde{\boldsymbol{\beta}})\mathrm{vec}(\mathbf{z}_{ij})\mathrm{vec}(\mathbf{z}_{ij})^{\mathrm{T}}|Y_{ij}\} \\
& -R_{ij}E\{S_2(Y_{ij}, \mathbf{X}_{ij}|\widetilde{\boldsymbol{\Theta}}, \widetilde{\boldsymbol{\beta}})\mathrm{vec}(\mathbf{z}_{ij})|Y_{ij}\}^{\otimes 2}]\mathrm{vec}(\boldsymbol{\Theta}^k)| \geq \delta^2 + \frac{32(d_{H_2} + d_{S_2}^2)a^2}{\sqrt{mn}}\Bigg\} \\
\leq\ & 4\exp\left\{\frac{-mn\delta^4}{64(d_{H_2} + d_{S_2}^2)^2a^2D^2} + \log\{N_{\boldsymbol{\Theta}}(\delta)\}\right\} \\
\leq\ & 4\exp\left\{\frac{-mn\delta^4}{64(d_{H_2} + d_{S_2}^2)^2a^2D^2} + \frac{144\rho(\nu, D)^2}{\delta^2}\max(n, m)\right\}.
\end{aligned}
$$

Now recall $\delta = D/\xi$, and

$$
\frac{mn\delta^4}{64(d_{H_2} + d_{S_2}^2)^2a^2D^2} \geq 2\frac{144\rho(\nu, D)^2}{\delta^2}\max(n, m).
$$

Thus we have

$$
\Pr\left\{\sup_{k=1,\ldots,N_{\boldsymbol{\Theta}}(\delta)} |\mathrm{vec}(\boldsymbol{\Theta}^k)^{\mathrm{T}}\mathbf{F}_{\boldsymbol{\Theta}}(\mathbf{X}, \mathbf{Y}|\widetilde{\boldsymbol{\Theta}}, \widetilde{\boldsymbol{\beta}})\mathrm{vec}(\boldsymbol{\Theta}^k)\right.
$$

$$
\begin{aligned}
&-\mathrm{vec}(\boldsymbol{\Theta}^k)^{\mathrm{T}}(mn)^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m}E\{R_{ij}E\{S_2^2(Y_{ij},\mathbf{X}_{ij}|\widetilde{\boldsymbol{\Theta}},\widetilde{\boldsymbol{\beta}})\mathrm{vec}(\mathbf{z}_{ij})\mathrm{vec}(\mathbf{z}_{ij})^{\mathrm{T}}|Y_{ij}\}\\
&-R_{ij}E\{S_2(Y_{ij},\mathbf{X}_{ij}|\widetilde{\boldsymbol{\Theta}},\widetilde{\boldsymbol{\beta}})\mathrm{vec}(\mathbf{z}_{ij})|Y_{ij}\}^{\otimes 2}\}\mathrm{vec}(\boldsymbol{\Theta})^k|\geq\delta^2+\frac{32(d_{H_2}+d_{S_2}^2)a^2}{\sqrt{mn}}\Bigg\}\\
\leq\;& 4\exp\left(-c_{d2}D^2mn\right),
\end{aligned}
$$

where $c_{d2}=1/\{128a^2\xi^4(d_{H_2}+d_{S_2}^2)^2\}$. Using the similar argument we can show that there is a $c_{d3}>0$ such that

$$
\begin{aligned}
&\Pr\Bigg\{\sup_{k=1,\ldots,N_{\boldsymbol{\Theta}}(\delta)}|\mathrm{vec}(\boldsymbol{\Theta}^k)^{\mathrm{T}}\mathbf{F}_{\boldsymbol{\Theta}}(\mathbf{X},\mathbf{Y}|\widetilde{\boldsymbol{\Theta}},\widetilde{\boldsymbol{\beta}})\mathrm{vec}(\boldsymbol{\Theta}-\boldsymbol{\Theta}^k)\\
&-(\mathrm{vec}\boldsymbol{\Theta}^k)^{\mathrm{T}}(mn)^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m}E\{R_{ij}E\{S_2^2(Y_{ij},\mathbf{X}_{ij}|\widetilde{\boldsymbol{\Theta}},\widetilde{\boldsymbol{\beta}})\mathrm{vec}(\mathbf{z}_{ij})\mathrm{vec}(\mathbf{z}_{ij})^{\mathrm{T}}|Y_{ij}\}\\
&-R_{ij}E\{S_2(Y_{ij},\mathbf{X}_{ij}|\widetilde{\boldsymbol{\Theta}},\widetilde{\boldsymbol{\beta}})\mathrm{vec}(\mathbf{z}_{ij})|Y_{ij}\}^{\otimes 2}\}\mathrm{vec}(\boldsymbol{\Theta}-\boldsymbol{\Theta}^k)|\geq\delta^2+\frac{64(d_{H_2}+d_{S_2}^2)a^2}{\sqrt{mn}}\Bigg\}\\
&\leq 4\exp\left(-c_{d3}D^2mn\right).
\end{aligned}
$$

with $c_{d3}=1/\{256a^2\xi^4(d_{H_2}+d_{S_2}^2)^2\}$. This proves the result.

Define

$$
\begin{aligned}
Q_{ij}\;\equiv\;&|H_2(Y_{ij},\mathbf{X}_{ij}|\widetilde{\boldsymbol{\Theta}},\widetilde{\boldsymbol{\beta}})-E\{H_2(Y_{ij},\mathbf{X}_{ij}|\widetilde{\boldsymbol{\Theta}},\widetilde{\boldsymbol{\beta}})|Y_{ij}\}\\
&+E\{S_2^2(Y_{ij},\mathbf{X}_{ij}|\widetilde{\boldsymbol{\Theta}},\widetilde{\boldsymbol{\beta}})|Y_{ij}\}-E\{S_2(Y_{ij},\mathbf{X}_{ij}|\widetilde{\boldsymbol{\Theta}},\widetilde{\boldsymbol{\beta}})|Y_{ij}\}^2|, \qquad (17)
\end{aligned}
$$

and by Condition (C6), we have $|Q_{ij}|\leq 2(d_{H_2}+d_{S_2}^2)$.

Lemma A.17 *Define*

$$
\mathcal{D}(D)=\{\boldsymbol{\Delta}_{\boldsymbol{\Theta}}\in\mathbb{R}^{n\times m}|\|\boldsymbol{\Delta}_{\boldsymbol{\Theta}}\|_F\leq\delta,\|\boldsymbol{\Delta}_{\boldsymbol{\Theta}}\|_*\leq 2\rho(\nu,D),\textit{ and }\|\boldsymbol{\Delta}_{\boldsymbol{\Theta}}\|_{\max}\leq 2a\}.
$$

The δ-covering number of $\mathcal{D}(D)$ satisfies

$$
\log N_{\mathcal{D}}(\delta)\leq\frac{576\rho(\nu,D)^2}{\delta^2}\max(n,m).
$$

Proof: Because $\mathcal{D}(D)\subset\widetilde{\mathcal{D}}(D)\equiv\{\boldsymbol{\Delta}_{\boldsymbol{\Theta}}\in\mathbb{R}^{n\times m}|\|\boldsymbol{\Delta}_{\boldsymbol{\Theta}}\|_*\leq 2\rho(\nu,D)\}$. Then δ-covering number of $\widetilde{\mathcal{D}}(D)$ defined by $N_{\widetilde{\mathcal{D}}}(\delta)$ satisfies $N_{\mathcal{D}}(\delta)\leq N_{\widetilde{\mathcal{D}}}(\delta)$. Using the same arguments as those lead to (39) in Negahban and Wainwright (2012), we have

$$
\sqrt{\log N_{\widetilde{\mathcal{D}}}(\delta)}\leq\frac{6\rho(\nu,D)}{\delta}E(\|\mathbf{G}\|_2),
$$

where $\mathbf{G}$ is a random matrix containing independent identically distribution standard normal entries. Further,

$$
E(\|\mathbf{G}\|_2)\leq 4\max(\sqrt{n},\sqrt{m})
$$

by the results in Section 3.1 in Bandeira et al. (2016). We have

$$
\log N_{\mathcal{D}}(\delta)\leq\log N_{\widetilde{\mathcal{D}}}(\delta)\leq\frac{576\rho(\nu,D)^2}{\delta^2}\max(n,m).
$$

This proves the result.

Lemma A.18 *Assume Conditions (C4), (C5) and (C6) hold. Define*

$$\mathcal{D}(D) = \{\mathbf{\Delta_\Theta} \in \mathbb{R}^{n\times m} | \|\mathbf{\Delta_\Theta}\|_F \le \delta, \|\mathbf{\Delta_\Theta}\|_* \le 2\rho(\nu, D), and \|\mathbf{\Delta_\Theta}\|_{\max} \le 2a\},$$

and let $\delta = D/\xi$. Assume ν satisfies

$$\frac{mn\delta^4}{256(d_{H_2} + d_{S_2}^2)^2 a^2 \delta^2} \ge 2\frac{576\rho(\nu, D)^2}{\delta^2}\max(n, m).$$

Further, let $\mathbf{F_\Theta}(\mathbf{X}, \mathbf{Y}|\widetilde{\mathbf{\Theta}}, \widetilde{\boldsymbol{\beta}})$ be as defined in Lemma A.16. Then

$$\begin{aligned}
&\Pr\left\{\sup_{\mathbf{\Delta_\Theta}\in\mathcal{D}(D)} |\mathrm{vec}(\mathbf{\Delta_\Theta})^{\mathrm{T}}\mathbf{F_\Theta}(\ \mathbf{X}, \mathbf{Y}|\widetilde{\mathbf{\Theta}}, \widetilde{\boldsymbol{\beta}})\mathrm{vec}(\mathbf{\Delta_\Theta}) - \mathrm{vec}(\mathbf{\Delta_\Theta})^{\mathrm{T}} E\{\mathbf{F_\Theta}(\mathbf{X}, \mathbf{Y}|\widetilde{\mathbf{\Theta}}, \widetilde{\boldsymbol{\beta}})\}\mathrm{vec}(\mathbf{\Delta_\Theta})| > D^2/\xi^2 \right.\\
&\left. + \frac{114(d_{H_2} + d_{S_2}^2)a^2}{\sqrt{mn}}\right\}\\
\le\ & 4\exp\left(\frac{-mnD^2}{512(d_{H_2} + d_{S_2}^2)^2 a^2 \xi^2}\right).
\end{aligned}$$

Proof: First note that for $\mathbf{\Delta_\Theta} \in \mathcal{D}(D)$ we have

$$\begin{aligned}
&|\mathrm{vec}(\mathbf{\Delta_\Theta})^{\mathrm{T}} R_{ij} H_2(Y_{ij}, \mathbf{X}_{ij}|\widetilde{\mathbf{\Theta}}, \widetilde{\boldsymbol{\beta}})\mathrm{vec}(\mathbf{z}_{ij})\mathrm{vec}(\mathbf{z}_{ij})^{\mathrm{T}}\mathrm{vec}(\mathbf{\Delta_\Theta})|\\
\le\ & \{\|\mathbf{\Delta_\Theta}\|_{\max}\|\mathrm{vec}(\mathbf{z}_{ij})\|_1\}^2 |H_2(Y_{ij}, \mathbf{X}_{ij}, \widetilde{\mathbf{\Theta}}, \widetilde{\boldsymbol{\beta}})| \le 4a^2 d_{H_2}.
\end{aligned}$$

Similarly,

$$\begin{aligned}
|\mathrm{vec}(\mathbf{\Delta}_\Theta)^{\mathrm{T}} R_{ij} E\{H_2(Y_{ij}, \mathbf{X}_{ij}|\widetilde{\mathbf{\Theta}}, \widetilde{\boldsymbol{\beta}})\mathrm{vec}(\mathbf{z}_{ij})\mathrm{vec}(\mathbf{z}_{ij})^{\mathrm{T}}|Y_{ij}\}\mathrm{vec}(\mathbf{\Delta_\Theta})| &\le 4a^2 d_{H_2},\\
|\mathrm{vec}(\mathbf{\Delta_\Theta})^{\mathrm{T}} R_{ij} E\{S_2^2(Y_{ij}, \mathbf{X}_{ij}|\widetilde{\mathbf{\Theta}}, \widetilde{\boldsymbol{\beta}})\mathrm{vec}(\mathbf{z}_{ij})\mathrm{vec}(\mathbf{z}_{ij})^{\mathrm{T}}|Y_{ij}\}\mathrm{vec}(\mathbf{\Delta_\Theta})| &\le 4a^2 d_{S_2}^2,\\
|\mathrm{vec}(\mathbf{\Delta_\Theta})^{\mathrm{T}} R_{ij} E\{S_2(Y_{ij}, \mathbf{X}_{ij}|\widetilde{\mathbf{\Theta}}, \widetilde{\boldsymbol{\beta}})\mathbf{z}_{ij}|Y_{ij}\}^{\otimes 2}\mathrm{vec}(\mathbf{\Delta_\Theta})| &\le 4a^2 d_{S_2}^2.
\end{aligned}$$

Also we have

$$\begin{aligned}
&|\mathrm{vec}(\mathbf{\Delta_\Theta})^{\mathrm{T}} R_{ij} H_2(Y_{ij}, \mathbf{X}_{ij}|\widetilde{\mathbf{\Theta}}, \widetilde{\boldsymbol{\beta}})\mathrm{vec}(\mathbf{z}_{ij})\mathrm{vec}(\mathbf{z}_{ij})^{\mathrm{T}}\mathrm{vec}(\mathbf{\Delta_\Theta})|\\
\le\ & \{\|\mathbf{\Delta_\Theta}\|_{\max}\|\mathrm{vec}(\mathbf{z}_{ij})\|_1\}\{\|\mathbf{\Delta_\Theta}\|_F\|\mathbf{z}_{ij}\|_F\}|H_2(Y_{ij}, \mathbf{X}_{ij}, \mathbf{\Theta}, \boldsymbol{\beta}^*)| \le 2a\delta d_{H_2}.
\end{aligned}$$

Similarly,

$$\begin{aligned}
|\mathrm{vec}(\mathbf{\Delta}_\Theta)^{\mathrm{T}} R_{ij} E\{H_2(Y_{ij}, \mathbf{X}_{ij}|\widetilde{\mathbf{\Theta}}, \widetilde{\boldsymbol{\beta}})\mathrm{vec}(\mathbf{z}_{ij})\mathrm{vec}(\mathbf{z}_{ij})^{\mathrm{T}}|Y_{ij}\}\mathrm{vec}(\mathbf{\Delta_\Theta})| &\le 2a\delta d_{H_2},\\
|\mathrm{vec}(\mathbf{\Delta_\Theta})^{\mathrm{T}} R_{ij} E\{S_2^2(Y_{ij}, \mathbf{X}_{ij}|\widetilde{\mathbf{\Theta}}, \widetilde{\boldsymbol{\beta}})\mathrm{vec}(\mathbf{z}_{ij})\mathrm{vec}(\mathbf{z}_{ij})^{\mathrm{T}}|Y_{ij}\}\mathrm{vec}(\mathbf{\Delta_\Theta})| &\le 2a\delta d_{S_2}^2,\\
|\mathrm{vec}(\mathbf{\Delta_\Theta})^{\mathrm{T}} R_{ij} E\{S_2(Y_{ij}, \mathbf{X}_{ij}|\widetilde{\mathbf{\Theta}}, \widetilde{\boldsymbol{\beta}})\mathbf{z}_{ij}|Y_{ij}\}^{\otimes 2}\mathrm{vec}(\mathbf{\Delta_\Theta})| &\le 2a\delta d_{S_2}^2.
\end{aligned}$$

Hence each summand in $\mathrm{vec}(\mathbf{\Delta_\Theta})^{\mathrm{T}}\mathbf{F_\Theta}(\mathbf{X}, \mathbf{Y}|\widetilde{\mathbf{\Theta}}, \widetilde{\boldsymbol{\beta}})\mathrm{vec}(\mathbf{\Delta_\Theta})$ is in the range of $[-8a^2(d_{H_2}+d_{S_2}^2), 8a^2(d_{H_2}+d_{S_2}^2)]$, and also in the range of $[-4a\delta(d_{H_2} + d_{S_2}^2), 4a\delta(d_{H_2} + d_{S_2}^2)]$. Define $\sigma_d \equiv 8a\delta(d_{H_2} + d_{S_2}^2)$ be the σ in Lemma A.2 (7) and $\sigma_{1d} \equiv 16a^2(d_{H_2} + d_{S_2}^2)$ be the σ in Lemma A.2 (8). By Lemma A.2, let m_F be the median of $\mathrm{vec}(\mathbf{\Delta_\Theta})^{\mathrm{T}}\mathbf{F_\Theta}(\mathbf{X}, \mathbf{Y}|\widetilde{\mathbf{\Theta}}, \widetilde{\boldsymbol{\beta}})\mathrm{vec}(\mathbf{\Delta_\Theta})$, we have

$$\begin{aligned}
&\Pr\Big(|\mathrm{vec}(\mathbf{\Delta_\Theta})^{\mathrm{T}}\mathbf{F_\Theta}(\mathbf{X}, \mathbf{Y}|\widetilde{\mathbf{\Theta}}, \widetilde{\boldsymbol{\beta}})\mathrm{vec}(\mathbf{\Delta_\Theta}) - \mathrm{vec}(\mathbf{\Delta_\Theta})^{\mathrm{T}} E\{\mathbf{F_\Theta}(\mathbf{X}, \mathbf{Y}|\widetilde{\mathbf{\Theta}}, \widetilde{\boldsymbol{\beta}})\}\mathrm{vec}(\mathbf{\Delta_\Theta})| \ge\\
&\delta^2 + \frac{114(d_{H_2} + d_{S_2}^2)a^2}{\sqrt{mn}}\Big)
\end{aligned}$$

$$
\begin{aligned}
&\leq \Pr\Big(|\mathrm{vec}(\boldsymbol{\Delta}_{\boldsymbol{\Theta}})^{\mathrm{T}}\mathbf{F}_{\boldsymbol{\Theta}}(\mathbf{X},\mathbf{Y}|\widetilde{\boldsymbol{\Theta}},\widetilde{\boldsymbol{\beta}})\mathrm{vec}(\boldsymbol{\Delta}_{\boldsymbol{\Theta}}) - m_F\}| + |\mathrm{vec}(\boldsymbol{\Delta}_{\boldsymbol{\Theta}})^{\mathrm{T}}E\{\mathbf{F}_{\boldsymbol{\Theta}}(\mathbf{X},\mathbf{Y}|\widetilde{\boldsymbol{\Theta}},\widetilde{\boldsymbol{\beta}})\}\mathrm{vec}(\boldsymbol{\Delta}_{\boldsymbol{\Theta}}) - m_F\}| \geq \\
&\quad \delta^2 + \frac{114(d_{H_2}+d_{S_2}^2)a^2}{\sqrt{mn}}\Big) \\
&\leq \Pr\Big(|\mathrm{vec}(\boldsymbol{\Delta}_{\boldsymbol{\Theta}})^{\mathrm{T}}\mathbf{F}_{\boldsymbol{\Theta}}(\mathbf{X},\mathbf{Y}|\widetilde{\boldsymbol{\Theta}},\widetilde{\boldsymbol{\beta}})\mathrm{vec}(\boldsymbol{\Delta}_{\boldsymbol{\Theta}}) - m_F\}| \geq \delta^2 \\
&\quad + \frac{114(d_{H_2}+d_{S_2}^2)a^2}{\sqrt{mn}} - \frac{64\sqrt{\pi}(d_{H_2}+d_{S_2}^2)a^2}{\sqrt{mn}}\Big) \\
&\leq \Pr\Big(\sqrt{mn}|\mathrm{vec}(\boldsymbol{\Delta}_{\boldsymbol{\Theta}})^{\mathrm{T}}\mathbf{F}_{\boldsymbol{\Theta}}(\mathbf{X},\mathbf{Y}|\widetilde{\boldsymbol{\Theta}},\widetilde{\boldsymbol{\beta}})\mathrm{vec}(\boldsymbol{\Delta}_{\boldsymbol{\Theta}}) - m_F| \geq \sqrt{mn}\delta^2\Big) \\
&\leq 4\exp\left\{\frac{-mn\delta^4}{256(d_{H_2}+d_{S_2}^2)^2a^2\delta^2}\right\}.
\end{aligned}
$$

Therefore combine Lemma A.17,

$$
\begin{aligned}
&\Pr\left\{\sup_{\boldsymbol{\Delta}_{\boldsymbol{\Theta}}\in\mathcal{D}(D)}|\mathrm{vec}(\boldsymbol{\Delta}_{\boldsymbol{\Theta}})^{\mathrm{T}}\mathbf{F}_{\boldsymbol{\Theta}}(\mathbf{X},\mathbf{Y}|\widetilde{\boldsymbol{\Theta}},\widetilde{\boldsymbol{\beta}})\mathrm{vec}(\boldsymbol{\Delta}_{\boldsymbol{\Theta}}) - \mathrm{vec}(\boldsymbol{\Delta}_{\boldsymbol{\Theta}})^{\mathrm{T}}E\{\mathbf{F}_{\boldsymbol{\Theta}}(\mathbf{X},\mathbf{Y}|\widetilde{\boldsymbol{\Theta}},\widetilde{\boldsymbol{\beta}})\}\mathrm{vec}(\boldsymbol{\Delta}_{\boldsymbol{\Theta}})| > \delta^2\right. \\
&\quad \left. + \frac{114(d_{H_2}+d_{S_2}^2)a^2}{\sqrt{mn}}\right\} \\
&\leq 4\exp\left(\frac{-mn\delta^4}{256(d_{H_2}+d_{S_2}^2)^2a^2\delta^2} + \log\{N_{\mathcal{D}}(\delta)\}\right) \\
&\leq 4\exp\left(\frac{-mn\delta^4}{256(d_{H_2}+d_{S_2}^2)^2a^2\delta^2} + \frac{576\rho(\nu,D)^2}{\delta^2}\max(n,m)\right).
\end{aligned}
$$

Now because $\delta = D/\xi$ and

$$
\frac{mn\delta^4}{256(d_{H_2}+d_{S_2}^2)^2a^2\delta^2} \geq 2\frac{576\rho(\nu,D)^2}{\delta^2}\max(n,m).
$$

Therefore,

$$
\begin{aligned}
&\Pr\left\{\sup_{\boldsymbol{\Delta}_{\boldsymbol{\Theta}}\in\mathcal{D}(D)}|\mathrm{vec}(\boldsymbol{\Delta}_{\boldsymbol{\Theta}})^{\mathrm{T}}\mathbf{F}_{\boldsymbol{\Theta}}(\mathbf{X},\mathbf{Y}|\widetilde{\boldsymbol{\Theta}},\widetilde{\boldsymbol{\beta}})\mathrm{vec}(\boldsymbol{\Delta}_{\boldsymbol{\Theta}}) - \mathrm{vec}(\boldsymbol{\Delta}_{\boldsymbol{\Theta}})^{\mathrm{T}}E\{\mathbf{F}_{\boldsymbol{\Theta}}(\mathbf{X},\mathbf{Y}|\widetilde{\boldsymbol{\Theta}},\widetilde{\boldsymbol{\beta}})\}\mathrm{vec}(\boldsymbol{\Delta}_{\boldsymbol{\Theta}})| > D^2/\xi^2\right. \\
&\quad \left. + \frac{114(d_{H_2}+d_{S_2}^2)a^2}{\sqrt{mn}}\right\} \\
&\leq 4\exp\left(\frac{-mnD^2}{512(d_{H_2}+d_{S_2}^2)^2a^2\xi^2}\right).
\end{aligned}
$$

This proves the result.

Lemma A.19 *Assume Conditions (C1)–(C5) hold. Let $\delta = D/\xi$. Then we have*

$$
\begin{aligned}
&\Pr\left\{|\mathrm{vec}(\boldsymbol{\Theta})^{\mathrm{T}}\mathbf{F}_{\boldsymbol{\Theta}}(\mathbf{X},\mathbf{Y}|\boldsymbol{\Theta}^*,\widehat{\boldsymbol{\beta}})\mathrm{vec}(\boldsymbol{\Theta}) - \mathrm{vec}(\boldsymbol{\Theta})^{\mathrm{T}}(mn)^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m}E[R_{ij}E\{S_2^2(Y_{ij},\mathbf{X}_{ij}|\boldsymbol{\Theta}^*,\widehat{\boldsymbol{\beta}})\mathrm{vec}(\mathbf{z}_{ij})\mathrm{vec}(\mathbf{z}_{ij})^{\mathrm{T}}|Y_{ij}\}\right. \\
&\quad \left. - R_{ij}E\{S_2(Y_{ij},\mathbf{X}_{ij}|\boldsymbol{\Theta}^*,\widehat{\boldsymbol{\beta}})\mathrm{vec}(\mathbf{z}_{ij})|Y_{ij}\}^{\otimes 2}]\mathrm{vec}(\boldsymbol{\Theta})|\right. \\
&\geq \alpha_{\min}\left((mn)^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m}E[R_{ij}E\{S_2^2(Y_{ij},\mathbf{X}_{ij}|\boldsymbol{\Theta}^*,\widehat{\boldsymbol{\beta}})\mathrm{vec}(\mathbf{z}_{ij})\mathrm{vec}(\mathbf{z}_{ij})^{\mathrm{T}}|Y_{ij}\}\right.
\end{aligned}
$$

$$-R_{ij}E\{S_2(Y_{ij},\mathbf{X}_{ij}|\boldsymbol{\Theta}^*,\widehat{\boldsymbol{\beta}})\mathrm{vec}(\mathbf{z}_{ij})|Y_{ij}\}^{\otimes 2}]\Big)/2\|\boldsymbol{\Theta}\|_F^2+\frac{274(d_{H_2}+d_{S_2}^2)a^2}{\sqrt{mn}},\boldsymbol{\Theta}\in\mathcal{C}_{\boldsymbol{\Theta}}(\nu)\Big\}$$

$$= \exp\{-Cd\log(d)\}$$

for some positive constant C, where the expectations are taken over $\mathbf{X},\mathbf{Y}$.

Proof: For any $\boldsymbol{\Theta}\in\bar{\mathbb{B}}_{\boldsymbol{\Theta}}(D)$, for any $k\in\{1,\ldots,N(\delta)\}$, let $\widetilde{\boldsymbol{\Delta}}_{\boldsymbol{\Theta}}=\boldsymbol{\Theta}-\boldsymbol{\Theta}^k$, then $\widetilde{\boldsymbol{\Delta}}_{\boldsymbol{\Theta}}\in\mathcal{D}(D)$. In addition,

$$\begin{aligned}
&\mathrm{vec}(\boldsymbol{\Theta})^{\mathrm{T}}\mathbf{F}_{\boldsymbol{\Theta}}(\mathbf{X},\mathbf{Y}|\boldsymbol{\Theta}^*,\widehat{\boldsymbol{\beta}})\mathrm{vec}(\boldsymbol{\Theta})-\mathrm{vec}(\boldsymbol{\Theta})^{\mathrm{T}}(mn)^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m}E[R_{ij}E\{S_2^2(Y_{ij},\mathbf{X}_{ij}|\boldsymbol{\Theta}^*,\widehat{\boldsymbol{\beta}})\mathrm{vec}(\mathbf{z}_{ij})\mathrm{vec}(\mathbf{z}_{ij})^{\mathrm{T}}|Y_{ij}\}\\
&-R_{ij}E\{S_2(Y_{ij},\mathbf{X}_{ij}|\boldsymbol{\Theta}^*,\widehat{\boldsymbol{\beta}})\mathrm{vec}(\mathbf{z}_{ij})|Y_{ij}\}^{\otimes 2}]\mathrm{vec}(\boldsymbol{\Theta})\\
=\;&\mathbf{F}_{\boldsymbol{\beta}}(\boldsymbol{\Theta}^k+\widetilde{\boldsymbol{\Delta}}_{\boldsymbol{\Theta}},\boldsymbol{\Theta}^k+\widetilde{\boldsymbol{\Delta}}_{\boldsymbol{\Theta}},\mathbf{X},\mathbf{Y}|\boldsymbol{\Theta}^*,\widehat{\boldsymbol{\beta}})\\
&-\mathrm{vec}(\boldsymbol{\Theta}^k+\widetilde{\boldsymbol{\Delta}}_{\boldsymbol{\Theta}})^{\mathrm{T}}(mn)^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m}E[R_{ij}E\{S_2^2(Y_{ij},\mathbf{X}_{ij}|\boldsymbol{\Theta}^*,\widehat{\boldsymbol{\beta}})\mathrm{vec}(\mathbf{z}_{ij})\mathrm{vec}(\mathbf{z}_{ij})^{\mathrm{T}}|Y_{ij}\}\\
&-R_{ij}E\{S_2(Y_{ij},\mathbf{X}_{ij}|\boldsymbol{\Theta}^*,\widehat{\boldsymbol{\beta}})\mathrm{vec}(\mathbf{z}_{ij})|Y_{ij}\}^{\otimes 2}]\mathrm{vec}(\boldsymbol{\Theta}^k+\widetilde{\boldsymbol{\Delta}}_{\boldsymbol{\Theta}})\\
=\;&\mathrm{vec}(\boldsymbol{\Theta}^k)^{\mathrm{T}}\mathbf{F}_{\boldsymbol{\Theta}}(\mathbf{X},\mathbf{Y}|\boldsymbol{\Theta}^*,\widehat{\boldsymbol{\beta}})\mathrm{vec}(\boldsymbol{\Theta}^k)-\mathrm{vec}(\boldsymbol{\Theta}^k)^{\mathrm{T}}(mn)^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m}E[R_{ij}E\{S_2^2(Y_{ij},\mathbf{X}_{ij}|\boldsymbol{\Theta}^*,\widehat{\boldsymbol{\beta}})\mathrm{vec}(\mathbf{z}_{ij})\mathrm{vec}(\mathbf{z}_{ij})^{\mathrm{T}}|Y_{ij}\}\\
&-R_{ij}E\{S_2(Y_{ij},\mathbf{X}_{ij}|\boldsymbol{\Theta}^*,\widehat{\boldsymbol{\beta}})\mathrm{vec}(\mathbf{z}_{ij})|Y_{ij}\}^{\otimes 2}]\mathrm{vec}(\boldsymbol{\Theta}^k)\\
&+2\mathrm{vec}(\boldsymbol{\Theta}^k)^{\mathrm{T}}\mathbf{F}_{\boldsymbol{\Theta}}(\mathbf{X},\mathbf{Y}|\boldsymbol{\Theta}^*,\widehat{\boldsymbol{\beta}})\widetilde{\boldsymbol{\Delta}}_{\boldsymbol{\Theta}}-2\mathrm{vec}(\boldsymbol{\Theta}^k)^{\mathrm{T}}(mn)^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m}E[R_{ij}E\{S_2^2(Y_{ij},\mathbf{X}_{ij}|\boldsymbol{\Theta}^*,\widehat{\boldsymbol{\beta}})\mathrm{vec}(\mathbf{z}_{ij})\mathrm{vec}(\mathbf{z}_{ij})^{\mathrm{T}}|Y_{ij}\}\\
&-R_{ij}E\{S_2(Y_{ij},\mathbf{X}_{ij}|\widehat{\boldsymbol{\Theta}},\boldsymbol{\beta}^*)\mathrm{vec}(\mathbf{z}_{ij})|Y_{ij}\}^{\otimes 2}]\mathrm{vec}(\widetilde{\boldsymbol{\Delta}}_{\boldsymbol{\Theta}})\\
&+\mathrm{vec}(\widetilde{\boldsymbol{\Delta}}_{\boldsymbol{\Theta}})^{\mathrm{T}}\mathbf{F}_{\boldsymbol{\Theta}}(\mathbf{X},\mathbf{Y}|\boldsymbol{\Theta}^*,\widehat{\boldsymbol{\beta}})\widetilde{\boldsymbol{\Delta}}_{\boldsymbol{\Theta}}-\mathrm{vec}(\widetilde{\boldsymbol{\Delta}}_{\boldsymbol{\Theta}})^{\mathrm{T}}(mn)^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m}E[R_{ij}E\{S_2^2(Y_{ij},\mathbf{X}_{ij}|\boldsymbol{\Theta}^*,\widehat{\boldsymbol{\beta}})\mathrm{vec}(\mathbf{z}_{ij})\mathrm{vec}(\mathbf{z}_{ij})^{\mathrm{T}}|Y_{ij}\}\\
&-R_{ij}E\{S_2(Y_{ij},\mathbf{X}_{ij}|\boldsymbol{\Theta}^*,\widehat{\boldsymbol{\beta}})\mathrm{vec}(\mathbf{z}_{ij})|Y_{ij}\}^{\otimes 2}]\mathrm{vec}(\widetilde{\boldsymbol{\Delta}}_{\boldsymbol{\Theta}}).
\end{aligned}$$

Hence by Lemmas A.16 and A.18, while replacing $\widetilde{\boldsymbol{\Theta}}$ by $\boldsymbol{\Theta}^*$ and $\widetilde{\boldsymbol{\beta}}$ by $\widehat{\boldsymbol{\beta}}$, there are constants b_{d3} such that

$$\begin{aligned}
&\sup_{\boldsymbol{\Theta}\in\mathbb{B}_{\boldsymbol{\Theta}}(D)}|\mathrm{vec}(\boldsymbol{\Theta})^{\mathrm{T}}\mathbf{F}_{\boldsymbol{\Theta}}(\mathbf{X},\mathbf{Y}|\boldsymbol{\Theta}^*,\widehat{\boldsymbol{\beta}})\mathrm{vec}(\boldsymbol{\Theta}) && (18)\\
&-\mathrm{vec}(\boldsymbol{\Theta})^{\mathrm{T}}(mn)^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m}E[R_{ij}E\{S_2^2(Y_{ij},\mathbf{X}_{ij}|\boldsymbol{\Theta}^*,\widehat{\boldsymbol{\beta}})\mathrm{vec}(\mathbf{z}_{ij})\mathrm{vec}(\mathbf{z}_{ij})^{\mathrm{T}}|Y_{ij}\}\\
&-R_{ij}E\{S_2(Y_{ij},\mathbf{X}_{ij}|\boldsymbol{\Theta}^*,\widehat{\boldsymbol{\beta}})\mathrm{vec}(\mathbf{z}_{ij})|Y_{ij}\}^{\otimes 2}]\mathrm{vec}(\boldsymbol{\Theta})|\\
\leq\;&\sup_{\boldsymbol{\Theta}\in\bar{\mathbb{B}}_{\boldsymbol{\Theta}}(D)}|\mathrm{vec}(\boldsymbol{\Theta})^{\mathrm{T}}\mathbf{F}_{\boldsymbol{\Theta}}(\mathbf{X},\mathbf{Y}|\boldsymbol{\Theta}^*,\widehat{\boldsymbol{\beta}})\mathrm{vec}(\boldsymbol{\Theta}) && (19)\\
&-\mathrm{vec}(\boldsymbol{\Theta})^{\mathrm{T}}(mn)^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m}E[R_{ij}E\{S_2^2(Y_{ij},\mathbf{X}_{ij}|\boldsymbol{\Theta}^*,\widehat{\boldsymbol{\beta}})\mathrm{vec}(\mathbf{z}_{ij})\mathrm{vec}(\mathbf{z}_{ij})^{\mathrm{T}}|Y_{ij}\}\\
&-R_{ij}E\{S_2(Y_{ij},\mathbf{X}_{ij}|\boldsymbol{\Theta}^*,\widehat{\boldsymbol{\beta}})\mathrm{vec}(\mathbf{z}_{ij})|Y_{ij}\}^{\otimes 2}]\mathrm{vec}(\boldsymbol{\Theta})|\\
\leq\;&\sup_{k=1,\ldots,N_{\boldsymbol{\Theta}}(\delta)}|\mathrm{vec}(\boldsymbol{\Theta}^k)^{\mathrm{T}}\mathbf{F}_{\boldsymbol{\Theta}}(\mathbf{X},\mathbf{Y}|\boldsymbol{\Theta}^*,\widehat{\boldsymbol{\beta}})\boldsymbol{\Theta}^k && (20)\\
&-\mathrm{vec}(\boldsymbol{\Theta}^k)^{\mathrm{T}}(mn)^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m}E[R_{ij}E\{S_2^2(Y_{ij},\mathbf{X}_{ij}|\boldsymbol{\Theta}^*,\widehat{\boldsymbol{\beta}})\mathrm{vec}(\mathbf{z}_{ij})\mathrm{vec}(\mathbf{z}_{ij})^{\mathrm{T}}|Y_{ij}\}
\end{aligned}$$

$$
\begin{aligned}
&-R_{ij}E\{S_2(Y_{ij},\mathbf{X}_{ij}|\boldsymbol{\Theta}^*,\widehat{\boldsymbol{\beta}})\mathrm{vec}(\mathbf{z}_{ij})|Y_{ij}\}^{\otimes 2}]\mathrm{vec}(\boldsymbol{\Theta}^k)| \\
&+2\sup_{k=1,\ldots,N_{\boldsymbol{\Theta}}(\delta)}|\mathrm{vec}(\boldsymbol{\Theta}^k)^{\mathrm{T}}\mathbf{F}_{\boldsymbol{\Theta}}(\mathbf{X},\mathbf{Y}|\boldsymbol{\Theta}^*,\widehat{\boldsymbol{\beta}})\widetilde{\boldsymbol{\Delta}}_{\boldsymbol{\Theta}} \qquad (21)\\
&-\mathrm{vec}(\boldsymbol{\Theta}^k)^{\mathrm{T}}(mn)^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m}E[R_{ij}E\{S_2^2(Y_{ij},\mathbf{X}_{ij}|\boldsymbol{\Theta}^*,\widehat{\boldsymbol{\beta}})\mathrm{vec}(\mathbf{z}_{ij})\mathrm{vec}(\mathbf{z}_{ij})^{\mathrm{T}}|Y_{ij}\} \\
&-R_{ij}E\{S_2(Y_{ij},\mathbf{X}_{ij}|\boldsymbol{\Theta},\widehat{\boldsymbol{\beta}})\mathrm{vec}(\mathbf{z}_{ij})|Y_{ij}\}^{\otimes 2}]\mathrm{vec}(\widetilde{\boldsymbol{\Delta}}_{\boldsymbol{\Theta}})| \\
&+\sup_{\widetilde{\boldsymbol{\Delta}}_{\boldsymbol{\Theta}}\in\mathcal{D}(D)}|\mathrm{vec}(\widetilde{\boldsymbol{\Delta}}_{\boldsymbol{\Theta}})^{\mathrm{T}}\mathbf{F}_{\boldsymbol{\Theta}}(\mathbf{X},\mathbf{Y}|\boldsymbol{\Theta}^*,\widehat{\boldsymbol{\beta}})\mathrm{vec}(\widetilde{\boldsymbol{\Delta}}_{\boldsymbol{\Theta}}) \qquad (22)\\
&-\mathrm{vec}(\widetilde{\boldsymbol{\Delta}}_{\boldsymbol{\Theta}})^{\mathrm{T}}(mn)^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m}E[R_{ij}E\{S_2^2(Y_{ij},\mathbf{X}_{ij}|\boldsymbol{\Theta}^*,\widehat{\boldsymbol{\beta}})\mathrm{vec}(\mathbf{z}_{ij})\mathrm{vec}(\mathbf{z}_{ij})^{\mathrm{T}}|Y_{ij}\} \\
&-R_{ij}E\{S_2(Y_{ij},\mathbf{X}_{ij}|\boldsymbol{\Theta}^*,\widehat{\boldsymbol{\beta}})\mathrm{vec}(\mathbf{z}_{ij})|Y_{ij}\}^{\otimes 2}]\mathrm{vec}(\widetilde{\boldsymbol{\Delta}}_{\boldsymbol{\Theta}})| \\
\leq\ & 4D^2/\xi^2+\frac{274(d_{H_2}+d_{S_2}^2)a^2}{\sqrt{mn}} \qquad (23)
\end{aligned}
$$

with probability at least $1-12\{\exp(-b_{d3}D^2mn)\}$. The last inequality holds by Lemma A.16 and A.18. Denote

$$
\begin{aligned}
\alpha_0 \equiv\ & \alpha_{\min}\Bigg((mn)^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m}E[R_{ij}E\{S_2^2(Y_{ij},\mathbf{X}_{ij}|\boldsymbol{\Theta}^*,\widehat{\boldsymbol{\beta}})\mathrm{vec}(\mathbf{z}_{ij})\mathrm{vec}(\mathbf{z}_{ij})^{\mathrm{T}}|Y_{ij}\} \\
&-R_{ij}E\{S_2(Y_{ij},\mathbf{X}_{ij}|\boldsymbol{\Theta}^*,\widehat{\boldsymbol{\beta}})\mathrm{vec}(\mathbf{z}_{ij})|Y_{ij}\}^{\otimes 2}]\Bigg),
\end{aligned}
$$

We choose $\xi>\sqrt{8/\alpha_0}$ and define $\alpha=\xi^2\alpha_0/8$, then $\alpha>1$.

Now note for $\boldsymbol{\Theta}\in\mathcal{C}_{\boldsymbol{\Theta}}(\nu)$ with $\|\boldsymbol{\Theta}\|_{\max}=b$ for a constant b, we have

$$
\|\boldsymbol{\Theta}\|_F^2\geq b\nu\|\boldsymbol{\Theta}\|_*\sqrt{\frac{\log(d)}{d}}\geq b\nu\|\boldsymbol{\Theta}\|_F\sqrt{\frac{\log(d)}{d}},
$$

which implies $\|\boldsymbol{\Theta}\|_F\geq b\nu\sqrt{\log(d)/d}$. Define $\mu(b)=b^2\nu^2\log(d)/d$. Moreover, we define

$$
\mathcal{S}_l(b)\equiv\{\boldsymbol{\Theta}\in\mathcal{C}_{\boldsymbol{\Theta}}(\nu)|\|\boldsymbol{\Theta}\|_{\max}=b,\sqrt{\alpha^{l-1}\mu(b)}\leq\|\boldsymbol{\Theta}\|_F\leq\sqrt{\alpha^l\mu(b)},\|\boldsymbol{\Theta}\|_*\leq\rho(\nu,\sqrt{\alpha^l\mu(b)})\},
$$

then $\mathcal{S}_l(b)\subset\bar{\mathbb{B}}_{\boldsymbol{\Theta}}(\sqrt{\alpha^l\mu(b)})$. For $\boldsymbol{\Theta}\in\mathcal{S}_l$ $D_{d7}=274(d_{H_2}+d_{S_2}^2)a^2$, we have

$$
\begin{aligned}
(\alpha_0/2)\|\boldsymbol{\Theta}\|_F^2+\frac{D_{d7}}{\sqrt{mn}} &\geq (\alpha_0/2)\alpha^{l-1}\mu(b)+\frac{D_{d7}}{\sqrt{mn}} \\
&= \frac{4\alpha^l\mu(b)}{\xi^2}+\frac{D_{d7}}{\sqrt{mn}}.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
&\Pr\Bigg\{|\mathrm{vec}(\boldsymbol{\Theta})^{\mathrm{T}}\mathbf{F}_{\boldsymbol{\Theta}}(\mathbf{X},\mathbf{Y}|\boldsymbol{\Theta}^*,\widehat{\boldsymbol{\beta}})\mathrm{vec}(\boldsymbol{\Theta}) \\
&-\mathrm{vec}(\boldsymbol{\Theta})^{\mathrm{T}}(mn)^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m}E[R_{ij}E\{S_2^2(Y_{ij},\mathbf{X}_{ij}|\boldsymbol{\Theta}^*,\widehat{\boldsymbol{\beta}})\mathrm{vec}(\mathbf{z}_{ij})\mathrm{vec}(\mathbf{z}_{ij})^{\mathrm{T}}|Y_{ij}\} \\
&-R_{ij}E\{S_2(Y_{ij},\mathbf{X}_{ij}|\boldsymbol{\Theta}^*,\widehat{\boldsymbol{\beta}})\mathrm{vec}(\mathbf{z}_{ij})|Y_{ij}\}^{\otimes 2}]\mathrm{vec}(\boldsymbol{\Theta})|
\end{aligned}
$$

$$
\begin{aligned}
&\geq (\alpha_0/2)\|\boldsymbol{\Theta}\|_F^2 + \frac{D_{d7}}{\sqrt{mn}}, \boldsymbol{\Theta} \in \mathcal{C}_{\boldsymbol{\Theta}}(\nu)\Big\} \\
\leq\ & \int_0^a \Pr\Big\{|\mathrm{vec}(\boldsymbol{\Theta})^{\mathrm{T}}\mathbf{F}_{\boldsymbol{\Theta}}(\mathbf{X},\mathbf{Y}|\boldsymbol{\Theta}^*,\widehat{\boldsymbol{\beta}})\mathrm{vec}(\boldsymbol{\Theta}) \\
&-\mathrm{vec}(\boldsymbol{\Theta})^{\mathrm{T}}(mn)^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m} E[R_{ij}E\{S_2^2(Y_{ij},\mathbf{X}_{ij}|\boldsymbol{\Theta}^*,\widehat{\boldsymbol{\beta}})\mathrm{vec}(\mathbf{z}_{ij})\mathrm{vec}(\mathbf{z}_{ij})^{\mathrm{T}}|Y_{ij}\} \\
&-R_{ij}E\{S_2(Y_{ij},\mathbf{X}_{ij}|\boldsymbol{\Theta}^*,\widehat{\boldsymbol{\beta}})\mathrm{vec}(\mathbf{z}_{ij})|Y_{ij}\}^{\otimes 2}]\mathrm{vec}(\boldsymbol{\Theta})| \\
&\geq (\alpha_0/2)\|\boldsymbol{\Theta}\|_F^2 + \frac{D_{d7}}{\sqrt{mn}}, \boldsymbol{\Theta} \in \mathcal{C}_{\boldsymbol{\Theta}}(\nu), \|\boldsymbol{\Theta}\|_{\max} = b\Big\}\, db \\
\leq\ & \int_0^a \sum_{l=1}^{\infty}\Pr\Big\{|\mathrm{vec}(\boldsymbol{\Theta})^{\mathrm{T}}\mathbf{F}_{\boldsymbol{\Theta}}(\mathbf{X},\mathbf{Y}|\boldsymbol{\Theta}^*,\widehat{\boldsymbol{\beta}})\mathrm{vec}(\boldsymbol{\Theta}) \\
&-\mathrm{vec}(\boldsymbol{\Theta})^{\mathrm{T}}(mn)^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m} E[R_{ij}E\{S_2^2(Y_{ij},\mathbf{X}_{ij}|\boldsymbol{\Theta}^*,\widehat{\boldsymbol{\beta}})\mathrm{vec}(\mathbf{z}_{ij})\mathrm{vec}(\mathbf{z}_{ij})^{\mathrm{T}}|Y_{ij}\} \\
&-R_{ij}E\{S_2(Y_{ij},\mathbf{X}_{ij}|\boldsymbol{\Theta}^*,\widehat{\boldsymbol{\beta}})\mathrm{vec}(\mathbf{z}_{ij})|Y_{ij}\}^{\otimes 2}]\mathrm{vec}(\boldsymbol{\Theta})| \\
&\geq (\alpha_0/2)\|\boldsymbol{\Theta}\|_F^2 + \frac{D_{d7}}{\sqrt{mn}}, \boldsymbol{\Theta} \in \mathcal{S}_l(b)\Big\}\, db \\
=\ & \sum_{l=1}^{\infty}\Pr\Big\{|\mathrm{vec}(\boldsymbol{\Theta})^{\mathrm{T}}\mathbf{F}_{\boldsymbol{\Theta}}(\mathbf{X},\mathbf{Y}|\boldsymbol{\Theta}^*,\widehat{\boldsymbol{\beta}})\mathrm{vec}(\boldsymbol{\Theta}) \\
&-\mathrm{vec}(\boldsymbol{\Theta})^{\mathrm{T}}(mn)^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m} E[R_{ij}E\{S_2^2(Y_{ij},\mathbf{X}_{ij}|\boldsymbol{\Theta}^*,\widehat{\boldsymbol{\beta}})\mathrm{vec}(\mathbf{z}_{ij})\mathrm{vec}(\mathbf{z}_{ij})^{\mathrm{T}}|Y_{ij}\} \\
&-R_{ij}E\{S_2(Y_{ij},\mathbf{X}_{ij}|\boldsymbol{\Theta}^*,\widehat{\boldsymbol{\beta}})\mathrm{vec}(\mathbf{z}_{ij})|Y_{ij}\}^{\otimes 2}]\mathrm{vec}(\boldsymbol{\Theta})| \\
&\geq (\alpha_0/2)\|\boldsymbol{\Theta}\|_F^2 + \frac{D_{d7}}{\sqrt{mn}}, \boldsymbol{\Theta} \in \mathcal{S}_l(b^*)\Big\} \\
\leq\ & \sum_{l=1}^{\infty}\Pr\Big\{|\mathrm{vec}(\boldsymbol{\Theta})^{\mathrm{T}}\mathbf{F}_{\boldsymbol{\Theta}}(\mathbf{X},\mathbf{Y}|\boldsymbol{\Theta}^*,\widehat{\boldsymbol{\beta}})\mathrm{vec}(\boldsymbol{\Theta}) \\
&-\mathrm{vec}(\boldsymbol{\Theta})^{\mathrm{T}}(mn)^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m} E[R_{ij}E\{S_2^2(Y_{ij},\mathbf{X}_{ij}|\boldsymbol{\Theta}^*,\widehat{\boldsymbol{\beta}})\mathrm{vec}(\mathbf{z}_{ij})\mathrm{vec}(\mathbf{z}_{ij})^{\mathrm{T}}|Y_{ij}\} \\
&-R_{ij}E\{S_2(Y_{ij},\mathbf{X}_{ij}|\boldsymbol{\Theta}^*,\widehat{\boldsymbol{\beta}})\mathrm{vec}(\mathbf{z}_{ij})|Y_{ij}\}^{\otimes 2}]\mathrm{vec}(\boldsymbol{\Theta})| \geq 4\alpha^l\mu(b^*)/\xi^2 + \frac{D_{d7}}{\sqrt{mn}}, \boldsymbol{\Theta} \in \mathcal{S}_l(b^*)\Big\} \\
\leq\ & 12\sum_{l=1}^{\infty}\exp(-b_{d3}\mu(b^*)\alpha^l mn) \\
\leq\ & 12\sum_{l=1}^{\infty}\exp\{-b_{d3}l\log(\alpha)\mu(b^*)mn\} \\
\leq\ & 12\frac{\exp\{-b_{d3}\log(\alpha)mn\mu(b^*)\}}{1-\exp\{-b_{d3}\log(\alpha)mn\mu(b^*)\}} \\
\leq\ & \exp\{-Cd\log(d)\},
\end{aligned}
$$

for some positive constant C, where b^* is a point on the line connecting 0 and a. The third inequality holds by (18), the fourth inequality holds because $\alpha^l \geq l\log(\alpha) > 0$ for $\alpha > 1$. This proves the result.

Lemma A.20 *Assume Conditions (C5) and (C6) hold, and select* $\lambda_{\mathbf{\Theta}} \geq 2\|\partial\mathcal{L}(\mathbf{\Theta}_0, \widehat{\boldsymbol{\beta}})/\partial\mathbf{\Theta}\|_{op}$. *If*

$$\mathrm{vec}(\mathbf{\Delta}_{\mathbf{\Theta}})^{\mathrm{T}} E\{\mathbf{F}_{\mathbf{\Theta}}(\mathbf{X}, \mathbf{Y}|\mathbf{\Theta}^*, \widehat{\boldsymbol{\beta}})\}\mathrm{vec}(\mathbf{\Delta}_{\mathbf{\Theta}}) \leq \sigma_{dF}\sqrt{2d\log(d)/(mn)},$$

then

$$\begin{aligned}\|\mathbf{\Delta}_{\mathbf{\Theta}}\|_F \quad \leq \quad & \alpha_{\min}\Bigg((mn)^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m} E[R_{ij}E\{S_2^2(Y_{ij}, \mathbf{X}_{ij}|\mathbf{\Theta}^*, \widehat{\boldsymbol{\beta}})\mathrm{vec}(\mathbf{z}_{ij})\mathrm{vec}(\mathbf{z}_{ij})^{\mathrm{T}}|Y_{ij}\} \\ & -R_{ij}E\{S_2(Y_{ij}, \mathbf{X}_{ij}|\mathbf{\Theta}^*, \widehat{\boldsymbol{\beta}})\mathrm{vec}(\mathbf{z}_{ij})|Y_{ij}\}^{\otimes 2}]\Bigg)^{-1/2}\{2\sigma_{dF}^2 d\log(d)/(mn)\}^{1/4}.\end{aligned}$$

Otherwise

$$\|\mathbf{\Delta}_{\mathbf{\Theta}}\|_* \leq 8\sqrt{r}\|\mathbf{\Delta}_{\mathbf{\Theta}}\|_F$$

with probability at least $1 - 2\exp\{-d\log(d)\}$, *where* $\sigma_{dF} = 32a^2(d_{H_2} + d_{S_2}^2)$ *as defined in Theorem 2.*

Proof: First as shown in Lemma A.18 that for $\mathbf{\Delta}_{\mathbf{\Theta}} \in \mathcal{D}(D)$ we have

$$\begin{aligned}& |\mathrm{vec}(\mathbf{\Delta}_{\mathbf{\Theta}})^{\mathrm{T}} R_{ij} H_2(Y_{ij}, \mathbf{X}_{ij}|\mathbf{\Theta}^*, \widehat{\boldsymbol{\beta}})\mathrm{vec}(\mathbf{z}_{ij})\mathrm{vec}(\mathbf{z}_{ij})^{\mathrm{T}}\mathrm{vec}(\mathbf{\Delta}_{\mathbf{\Theta}})| \\ \leq \quad & \{\|\mathbf{\Delta}_{\mathbf{\Theta}}\|_{\max}\|\mathrm{vec}(\mathbf{z}_{ij})\|_1\}^2 |H_2(Y_{ij}, \mathbf{X}_{ij}|\mathbf{\Theta}^*, \widehat{\boldsymbol{\beta}})| \leq 4a^2 d_{H_2}.\end{aligned}$$

Similarly,

$$\begin{aligned}|\mathrm{vec}(\mathbf{\Delta}_{\Theta})^{\mathrm{T}} R_{ij} E\{H_2(Y_{ij}, \mathbf{X}_{ij}|\mathbf{\Theta}^*, \widehat{\boldsymbol{\beta}})\mathrm{vec}(\mathbf{z}_{ij})\mathrm{vec}(\mathbf{z}_{ij})^{\mathrm{T}}|Y_{ij}\}\mathrm{vec}(\mathbf{\Delta}_{\mathbf{\Theta}})| \quad &\leq \quad 4a^2 d_{H_2}, \\ |\mathrm{vec}(\mathbf{\Delta}_{\mathbf{\Theta}})^{\mathrm{T}} R_{ij} E\{S_2^2(Y_{ij}, \mathbf{X}_{ij}|\mathbf{\Theta}^*, \widehat{\boldsymbol{\beta}})\mathrm{vec}(\mathbf{z}_{ij})\mathrm{vec}(\mathbf{z}_{ij})^{\mathrm{T}}|Y_{ij}\}\mathrm{vec}(\mathbf{\Delta}_{\mathbf{\Theta}})| \quad &\leq \quad 4a^2 d_{S_2}^2, \\ |\mathrm{vec}(\mathbf{\Delta}_{\mathbf{\Theta}})^{\mathrm{T}} R_{ij} E\{S_2(Y_{ij}, \mathbf{X}_{ij}|\mathbf{\Theta}^*, \widehat{\boldsymbol{\beta}})\mathbf{z}_{ij}|Y_{ij}\}^{\otimes 2}\mathrm{vec}(\mathbf{\Delta}_{\mathbf{\Theta}})| \quad &\leq \quad 4a^2 d_{S_2}^2.\end{aligned}$$

Hence each summand in $\mathrm{vec}(\mathbf{\Delta}_{\mathbf{\Theta}})^{\mathrm{T}}\mathbf{F}_{\mathbf{\Theta}}(\mathbf{X}, \mathbf{Y}|\mathbf{\Theta}^*, \widehat{\boldsymbol{\beta}})\mathrm{vec}(\mathbf{\Delta}_{\mathbf{\Theta}}) - \mathrm{vec}(\mathbf{\Delta}_{\mathbf{\Theta}})^{\mathrm{T}} E\{\mathbf{F}_{\mathbf{\Theta}}(\mathbf{X}, \mathbf{Y}|\mathbf{\Theta}^*, \widehat{\boldsymbol{\beta}})\}\mathrm{vec}(\mathbf{\Delta}_{\mathbf{\Theta}})$ is in the range of $[-16a^2(d_{H_2} + d_{S_2}^2), 16a^2(d_{H_2} + d_{S_2}^2)]$, and in turn is sub-Gaussian with parameter σ_{dF} by Lemma A.1. Lemma A.1 further leads to

$$\begin{aligned}& \Pr\left\{|\mathrm{vec}(\mathbf{\Delta}_{\mathbf{\Theta}})^{\mathrm{T}}\mathbf{F}_{\mathbf{\Theta}}(\mathbf{X}, \mathbf{Y}|\mathbf{\Theta}^*, \widehat{\boldsymbol{\beta}})\mathrm{vec}(\mathbf{\Delta}_{\mathbf{\Theta}}) - \mathrm{vec}(\mathbf{\Delta}_{\mathbf{\Theta}})^{\mathrm{T}} E\{\mathbf{F}_{\mathbf{\Theta}}(\mathbf{X}, \mathbf{Y}|\mathbf{\Theta}^*, \widehat{\boldsymbol{\beta}})\}\mathrm{vec}(\mathbf{\Delta}_{\mathbf{\Theta}})| > t\right\} \\ = \quad & \Pr\left\{|mn\mathrm{vec}(\mathbf{\Delta}_{\mathbf{\Theta}})\mathbf{F}_{\mathbf{\Theta}}(\mathbf{X}, \mathbf{Y}|\mathbf{\Theta}^*, \widehat{\boldsymbol{\beta}})\mathrm{vec}(\mathbf{\Delta}_{\mathbf{\Theta}}) - mn\mathrm{vec}(\mathbf{\Delta}_{\mathbf{\Theta}})^{\mathrm{T}} E\{\mathbf{F}_{\mathbf{\Theta}}(\mathbf{X}, \mathbf{Y}|\mathbf{\Theta}^*, \widehat{\boldsymbol{\beta}})\}\mathrm{vec}(\mathbf{\Delta}_{\mathbf{\Theta}})| > mnt\right\} \\ \leq \quad & 2\exp\left\{-\frac{(mn)^2 t^2}{2mn\sigma_{dF}^2}\right\} \\ = \quad & 2\exp\left(-\frac{mnt^2}{2\sigma_{dF}^2}\right) \qquad (24)\end{aligned}$$

for any $t > 0$. By (24), let $t = \sigma_{dF}\sqrt{2d\log(d)/(mn)}$, we have

$$\mathrm{vec}(\mathbf{\Delta}_{\mathbf{\Theta}})^{\mathrm{T}}\mathbf{F}_{\mathbf{\Theta}}(\mathbf{X}, \mathbf{Y}|\mathbf{\Theta}^*, \widehat{\boldsymbol{\beta}})\mathrm{vec}(\mathbf{\Delta}_{\mathbf{\Theta}}) \geq \mathrm{vec}(\mathbf{\Delta}_{\mathbf{\Theta}})^{\mathrm{T}} E\{\mathbf{F}_{\mathbf{\Theta}}(\mathbf{X}, \mathbf{Y}|\mathbf{\Theta}^*, \widehat{\boldsymbol{\beta}})\}\mathrm{vec}(\mathbf{\Delta}_{\mathbf{\Theta}}) - \sigma_{dF}\sqrt{2d\log(d)/(mn)} \quad (25)$$

with probability at least $1 - 2\exp\{-d\log(d)\}$.

Now we discuss two cases. Case I:

$$\mathrm{vec}(\mathbf{\Delta}_{\mathbf{\Theta}})^{\mathrm{T}} E\{\mathbf{F}_{\mathbf{\Theta}}(\mathbf{X}, \mathbf{Y}|\mathbf{\Theta}^*, \widehat{\boldsymbol{\beta}})\}\mathrm{vec}(\mathbf{\Delta}_{\mathbf{\Theta}}) \leq \sigma_{dF}\sqrt{2d\log(d)/(mn)}$$

This leads to

$$\alpha_{\min}(E[R_{ij}E\{S_2^2(Y_{ij}, \mathbf{X}_{ij}|\mathbf{\Theta}^*, \widehat{\boldsymbol{\beta}})\mathrm{vec}(\mathbf{z}_{ij})\mathrm{vec}(\mathbf{z}_{ij})^{\mathrm{T}}|Y_{ij}\}$$

$$-R_{ij}E\{S_2(Y_{ij},\mathbf{X}_{ij}|\boldsymbol{\Theta}^*,\widehat{\boldsymbol{\beta}})\mathrm{vec}(\mathbf{z}_{ij})|Y_{ij}\}^{\otimes 2}])\|\boldsymbol{\Delta}_{\boldsymbol{\Theta}}\|_F^2 \le \sigma_{dF}\sqrt{2d\log(d)/(mn)},$$

hence

$$\begin{aligned}\|\boldsymbol{\Delta}_{\boldsymbol{\Theta}}\|_F \;\le\; & \alpha_{\min}(E[R_{ij}E\{S_2^2(Y_{ij},\mathbf{X}_{ij}|\boldsymbol{\Theta}^*,\widehat{\boldsymbol{\beta}})\mathrm{vec}(\mathbf{z}_{ij})\mathrm{vec}(\mathbf{z}_{ij})^{\mathrm{T}}|Y_{ij}\}\\ & -R_{ij}E\{S_2(Y_{ij},\mathbf{X}_{ij},\boldsymbol{\Theta}^*,\widehat{\boldsymbol{\beta}})\mathrm{vec}(\mathbf{z}_{ij})|Y_{ij}\}^{\otimes 2}])^{-1/2}\{2\sigma_{dF}^2 d\log(d)/(mn)\}^{1/4}.\end{aligned}$$

Case II:

$$\mathrm{vec}(\boldsymbol{\Delta}_{\boldsymbol{\Theta}})^{\mathrm{T}}E\{\mathbf{F}_{\boldsymbol{\Theta}}(\mathbf{X},\mathbf{Y}|\boldsymbol{\Theta}^*,\widehat{\boldsymbol{\beta}})\}\mathrm{vec}(\boldsymbol{\Delta}_{\boldsymbol{\Theta}}) > \sigma_{dF}\sqrt{2d\log(d)/(mn)}.$$

Under (25), this leads to

$$\mathrm{vec}(\boldsymbol{\Delta}_{\boldsymbol{\Theta}})^{\mathrm{T}}\mathbf{F}_{\boldsymbol{\Theta}}(\mathbf{X},\mathbf{Y}|\boldsymbol{\Theta}^*,\widehat{\boldsymbol{\beta}})\mathrm{vec}(\boldsymbol{\Delta}_{\boldsymbol{\Theta}}) > 0. \tag{26}$$

By Lemma 1 and (21) in Negahban and Wainwright (2012), when $\lambda_{\boldsymbol{\Theta}} \ge 2\|\partial\mathcal{L}(\boldsymbol{\Theta}_0,\widehat{\boldsymbol{\beta}})/\partial\boldsymbol{\beta}\|_{op}$, note that $\boldsymbol{\Theta}_0$ has rank r, which satisfies $r \le 2r$, we get

$$\|\boldsymbol{\Delta}_{\boldsymbol{\Theta}}\|_* \le 8\sqrt{r}\|\boldsymbol{\Delta}_{\boldsymbol{\Theta}}\|_F.$$

This proves the result.

Appendix F. Proof of Theorem 2

Proof: Let $\boldsymbol{\Delta}_{\boldsymbol{\Theta}} = \widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_0$, we consider the situation where

$$\mathrm{vec}(\boldsymbol{\Delta}_{\boldsymbol{\Theta}})^{\mathrm{T}}E\{\mathbf{F}_{\boldsymbol{\Theta}}(\mathbf{X},\mathbf{Y}|\boldsymbol{\Theta}^*,\widehat{\boldsymbol{\beta}})\}\mathrm{vec}(\boldsymbol{\Delta}_{\boldsymbol{\Theta}}) > \sigma_{dF}\sqrt{2d\log(d)/(mn)}.$$

First we consider Case I: $\boldsymbol{\Delta}_{\boldsymbol{\Theta}} \notin \mathcal{C}_{\boldsymbol{\Theta}}(\nu)$. Then, with probability at least $1-2\exp\{-d\log(d)\}$,

$$\|\boldsymbol{\Delta}_{\Theta}\|_F^2 \le \|\boldsymbol{\Delta}_{\boldsymbol{\Theta}}\|_{\max}\|\boldsymbol{\Delta}_{\boldsymbol{\Theta}}\|_*\nu\sqrt{\frac{d\log(d)}{mn}} \le 16a\sqrt{r}\|\boldsymbol{\Delta}_{\boldsymbol{\Theta}}\|_F\nu\sqrt{\frac{d\log(d)}{mn}},$$

which implies $\|\boldsymbol{\Delta}_{\boldsymbol{\Theta}}\|_F \le 16a\sqrt{r}\nu\sqrt{d\log(d)/(mn)}$.

Case II: $\boldsymbol{\Delta}_{\boldsymbol{\Theta}} \in \mathcal{C}_{\boldsymbol{\Theta}}(\nu)$. Because $\widehat{\boldsymbol{\Theta}},\widehat{\boldsymbol{\beta}}$ is the minimizer of $\mathcal{L}(\boldsymbol{\Theta},\boldsymbol{\beta}) + \lambda_{\boldsymbol{\Theta}}\|\boldsymbol{\Theta}\|_* + \lambda_{\boldsymbol{\beta}}\|\boldsymbol{\beta}\|_1$, we have

$$\mathcal{L}(\widehat{\boldsymbol{\Theta}},\widehat{\boldsymbol{\beta}}) - \mathcal{L}(\boldsymbol{\Theta}_0,\widehat{\boldsymbol{\beta}}) = \langle\frac{\partial\mathcal{L}(\boldsymbol{\Theta}_0,\widehat{\boldsymbol{\beta}})}{\partial\boldsymbol{\Theta}},\boldsymbol{\Delta}_{\boldsymbol{\Theta}}\rangle + \frac{1}{2}\mathrm{vec}(\widehat{\boldsymbol{\Theta}}-\boldsymbol{\Theta}_0)^{\mathrm{T}}\frac{\partial\mathcal{L}(\boldsymbol{\Theta}^*,\widehat{\boldsymbol{\beta}})}{\partial\mathrm{vec}(\boldsymbol{\Theta})\partial\mathrm{vec}(\boldsymbol{\Theta})^{\mathrm{T}}}\mathrm{vec}(\widehat{\boldsymbol{\Theta}}-\boldsymbol{\Theta}_0)$$

and

$$\begin{aligned}\mathbf{F}_{\boldsymbol{\beta}}(\boldsymbol{\Delta}_{\boldsymbol{\Theta}},\boldsymbol{\Delta}_{\boldsymbol{\Theta}},\mathbf{X},\mathbf{Y}|\boldsymbol{\Theta}^*,\widehat{\boldsymbol{\beta}})/2 &= \mathrm{vec}(\widehat{\boldsymbol{\Theta}}-\boldsymbol{\Theta}_0)^{\mathrm{T}}\frac{\partial^2\mathcal{L}(\boldsymbol{\Theta}^*,\widehat{\boldsymbol{\beta}})}{\partial\mathrm{vec}(\boldsymbol{\Theta})\partial\mathrm{vec}(\boldsymbol{\Theta})^{\mathrm{T}}}\mathrm{vec}(\widehat{\boldsymbol{\Theta}}-\boldsymbol{\Theta}_0)/2\\ &= \mathcal{L}(\widehat{\boldsymbol{\Theta}},\widehat{\boldsymbol{\beta}}) - \mathcal{L}(\boldsymbol{\Theta}_0,\widehat{\boldsymbol{\beta}}) - \langle\frac{\partial\mathcal{L}(\boldsymbol{\Theta}_0,\widehat{\boldsymbol{\beta}})}{\partial\boldsymbol{\Theta}},\boldsymbol{\Delta}_{\boldsymbol{\Theta}}\rangle\\ &\le \langle-\frac{\partial\mathcal{L}(\boldsymbol{\Theta}_0,\widehat{\boldsymbol{\beta}})}{\partial\boldsymbol{\Theta}},\boldsymbol{\Delta}_{\boldsymbol{\Theta}}\rangle + \lambda_{\boldsymbol{\Theta}}\|\boldsymbol{\Theta}_0\|_* - \lambda_{\boldsymbol{\Theta}}\|\widehat{\boldsymbol{\Theta}}\|_*\\ &\le \|\frac{\partial\mathcal{L}(\boldsymbol{\Theta}_0,\widehat{\boldsymbol{\beta}})}{\partial\boldsymbol{\Theta}}\|_{op}\|\boldsymbol{\Delta}_{\boldsymbol{\Theta}}\|_* + \lambda_{\boldsymbol{\Theta}}\|\boldsymbol{\Theta}_0\|_* - \lambda_{\boldsymbol{\Theta}}\|\widehat{\boldsymbol{\Theta}}\|_*\\ &\le \lambda_{\boldsymbol{\Theta}}/2\|\boldsymbol{\Delta}_{\boldsymbol{\Theta}}\|_* + \lambda_{\boldsymbol{\Theta}}\|\boldsymbol{\Theta}_0-\widehat{\boldsymbol{\Theta}}\|_*\\ &\le 12\lambda_{\boldsymbol{\Theta}}\sqrt{r}\|\boldsymbol{\Delta}_{\boldsymbol{\Theta}}\|_F \end{aligned}\tag{27}$$

with probability at least $1-2\exp\{-d\log(d)\}-d^{-1}$. The first inequality holds by the second order mean value theorem for $\mathcal{L}(\widehat{\boldsymbol{\Theta}},\boldsymbol{\beta})$ on $\boldsymbol{\Theta}$. The second inequlity is a known property concerning norms and trace. The third inequality holds by the selection that

$$\begin{aligned}\lambda_{\boldsymbol{\Theta}} \quad \geq \quad & 2c_d\sqrt{d\log(d)/(mn)}+2(2d_{H_2}+2d_{S_2}^2)\max\Big([10\sigma_{1F}/(\alpha_{0\boldsymbol{\beta}}\sqrt{mn})+\{3\lambda_{\boldsymbol{\beta}}\sqrt{s}/\alpha_{0\boldsymbol{\beta}}\}^2]^{1/2} \\ & +3\lambda_{\boldsymbol{\beta}}\sqrt{s}/\alpha_{0\boldsymbol{\beta}}, 8a\sqrt{s}\gamma\sqrt{\log\{\max(p,mn)\}/(mn)}, (4\alpha_{0\boldsymbol{\beta}})^{-1/2}\{2\sigma_{1F}^2\log\{\max(p,mn)\}/(mn)\}^{1/4}\Big),\end{aligned}$$

in the theorem statement which is greater than $2\|\partial\mathcal{L}(\boldsymbol{\Theta}_0,\widehat{\boldsymbol{\beta}})/\partial\boldsymbol{\Theta}\|_{op}$ with probability at least $1-d^{-1}$. The last line holds by Lemma A.20.

Further, by Lemma A.19 we have

$$\begin{aligned}& \mathrm{vec}(\boldsymbol{\Delta}_{\boldsymbol{\Theta}})^{\mathrm{T}}\mathbf{F}_{\boldsymbol{\Theta}}(\mathbf{X},\mathbf{Y}|\boldsymbol{\Theta}^*,\widehat{\boldsymbol{\beta}})\mathrm{vec}(\boldsymbol{\Delta}_{\boldsymbol{\Theta}})/2 \\ \geq \quad & \alpha_{\min}(E[R_{ij}E\{S_2^2(Y_{ij},\mathbf{X}_{ij}|\boldsymbol{\Theta}^*,\widehat{\boldsymbol{\beta}})\mathrm{vec}(\mathbf{z}_{ij})\mathrm{vec}(\mathbf{z}_{ij})^{\mathrm{T}}|Y_{ij}\} \\ & -R_{ij}E\{S_2(Y_{ij},\mathbf{X}_{ij}|\boldsymbol{\Theta}^*,\widehat{\boldsymbol{\beta}})\mathrm{vec}(\mathbf{z}_{ij})|Y_{ij}\}^{\otimes 2}])/4\|\boldsymbol{\Delta}_{\boldsymbol{\Theta}}\|_F^2-137(d_{H_2}+d_{S_2}^2)a^2/\sqrt{mn} \\ = \quad & \alpha_{0\boldsymbol{\Theta}}\|\boldsymbol{\Delta}_{\boldsymbol{\Theta}}\|_F^2-137(d_{H_2}+d_{S_2}^2)a^2/\sqrt{mn}\end{aligned}$$

with probability at least $1-\exp\{-Cd\log(d)\}$. Combine with (27), we have that with probability at least $1-\exp\{-Cd\log(d)\}-2\exp\{-d\log(d)\}$,

$$\alpha_{0\boldsymbol{\Theta}}\|\boldsymbol{\Delta}_{\boldsymbol{\Theta}}\|_F^2 \quad \leq \quad 12\lambda_{\boldsymbol{\Theta}}\sqrt{r}\|\boldsymbol{\Delta}_{\boldsymbol{\Theta}}\|_F+137(d_{H_2}+d_{S_2}^2)a^2/(\sqrt{mn}).$$

Then

$$\{\|\boldsymbol{\Delta}_{\boldsymbol{\Theta}}\|_F-6\lambda_{\boldsymbol{\Theta}}\sqrt{r}/\alpha_{0\boldsymbol{\Theta}}\}^2\leq[137(d_{H_2}+d_{S_2}^2)a^2/(\alpha_{0\boldsymbol{\Theta}}\sqrt{mn})+36\lambda_{\boldsymbol{\Theta}}^2r/\alpha_{0\boldsymbol{\Theta}}^2]$$

Hence

$$\|\boldsymbol{\Delta}_{\boldsymbol{\Theta}}\|_F\leq[137(d_{H_2}+d_{S_2}^2)a^2/(\alpha_{0\boldsymbol{\Theta}}\sqrt{mn})+36\lambda_{\boldsymbol{\Theta}}^2r/\alpha_{0\boldsymbol{\Theta}}^2]^{1/2}+6\lambda_{\boldsymbol{\Theta}}\sqrt{r}/\alpha_{0\boldsymbol{\Theta}}.$$

Combine with the order in Case I and Lemma A.20, we have

$$\begin{aligned}& \|\boldsymbol{\Delta}_{\boldsymbol{\Theta}}\|_F \\ \leq \quad & \max\Big([137(d_{H_2}+d_{S_2}^2)a^2/(\alpha_{0\boldsymbol{\Theta}}\sqrt{mn})+36\lambda_{\boldsymbol{\Theta}}^2r/\alpha_{0\boldsymbol{\Theta}}^2]^{1/2}+6\lambda_{\boldsymbol{\Theta}}\sqrt{r}/\alpha_{0\boldsymbol{\Theta}}, 16a\sqrt{r}\nu\sqrt{d\log(d)/mn}, \\ & (4\alpha_{0\boldsymbol{\Theta}})^{-1/2}\{2\sigma_{dF}^2d\log(d)/(mn)\}^{1/4}\Big),\end{aligned}$$

with probability at least $1-\exp\{-Cd\log(d)\}-2\exp\{-d\log(d)\}-d^{-1}$. This proves the result.

Appendix G. Lemmas for Theorem 3

Lemma A.21 *Assume Conditions (C5) and (C6) hold. Then for unit vectors $\mathbf{v}\in\mathbb{R}^p,\mathbf{u}\in\mathbb{R}^{mn}$ and $\mathbf{w}\in\mathbb{R}^{mn+p}$ such that $\|\mathbf{v}\|_2=1,\|\mathbf{u}\|_2=1$ and $\|\mathbf{w}\|_2=1$, we have*

$$\mathbf{F}_{\boldsymbol{\beta}}(\mathbf{v},\mathbf{v},\mathbf{X},\mathbf{Y}|\boldsymbol{\Theta},\boldsymbol{\beta})\geq E\{\mathbf{F}_{\boldsymbol{\beta}}(\mathbf{v},\mathbf{v},\mathbf{X},\mathbf{Y}|\boldsymbol{\Theta},\boldsymbol{\beta})\}-4c_0\sqrt{(d_{H_2}+d_{S_2}^2)\log\{\max(p,mn)\}/(mn)}, \quad (28)$$

with probability at least $1-2\max(p,mn)^{-1}$,

$$\mathbf{u}^{\mathrm{T}}\mathbf{F}_{\boldsymbol{\Theta}}(\mathbf{X},\mathbf{Y}|\boldsymbol{\Theta},\boldsymbol{\beta})\mathbf{u}\geq\mathbf{u}^{\mathrm{T}}E\{\mathbf{F}_{\boldsymbol{\Theta}}(\mathbf{X},\mathbf{Y}|\boldsymbol{\Theta},\boldsymbol{\beta})\}\mathbf{u}-4\sqrt{(d_{H_2}+d_{S_2}^2)\log(mn)/(mn)}, \quad (29)$$

with probability at least $1-2(mn)^{-1}$, and

$$\mathbf{w}^{\mathrm{T}}\partial^2\mathcal{L}(\boldsymbol{\Theta},\boldsymbol{\beta})/\partial[\{\mathrm{vec}(\boldsymbol{\Theta})^{\mathrm{T}},\boldsymbol{\beta}^{\mathrm{T}}\}^{\mathrm{T}}]^{\otimes 2}\mathbf{w}$$

$$\geq \quad E\left(\mathbf{w}^{\mathrm{T}}\partial^2\mathcal{L}(\boldsymbol{\Theta},\boldsymbol{\beta})/\partial[\{\mathrm{vec}(\boldsymbol{\Theta})^{\mathrm{T}},\boldsymbol{\beta}^{\mathrm{T}}\}^{\mathrm{T}}]^{\otimes 2}\mathbf{w}\right) - 4(c_0+1)\sqrt{(d_{H_2}+d_{S_2}^2)\log(mn+p)/mn}$$

with probability at least $1-2(p+mn)^{-1}$.

Proof: First note that for unit vector $\mathbf{v},\mathbf{v}\in\mathbb{R}^p$, because $\|\mathbf{v}\|_\infty\leq\|\mathbf{v}\|_2=1$, we have

$$\begin{aligned}&|\mathbf{v}^{\mathrm{T}}R_{ij}H_2(Y_{ij},\mathbf{X}_{ij}|\boldsymbol{\Theta},\boldsymbol{\beta})\mathbf{X}_{ij}\mathbf{X}_{ij}^{\mathrm{T}}\mathbf{v}|\\ \leq\quad &\{\|\mathbf{v}\|_\infty\|\mathbf{X}_{ij}\|_1\}^2H_2(Y_{ij},\mathbf{X}_{ij}|\boldsymbol{\Theta},\boldsymbol{\beta})\leq c_0^2d_{H_2},\end{aligned}$$

Similarly

$$|\mathbf{v}^{\mathrm{T}}R_{ij}E\{H_2(Y_{ij},\mathbf{X}_{ij}|\boldsymbol{\Theta},\boldsymbol{\beta})\mathbf{X}_{ij}\mathbf{X}_{ij}^{\mathrm{T}}|Y_{ij}\}\mathbf{v}|\leq c_0^2d_{H_2},$$

and

$$|\mathbf{v}^{\mathrm{T}}R_{ij}E\{S_2^2(Y_{ij},\mathbf{X}_{ij}|\boldsymbol{\Theta},\boldsymbol{\beta})\mathbf{X}_{ij}\mathbf{X}_{ij}^{\mathrm{T}}|Y_{ij}\}\mathbf{v}|\leq c_0^2d_{S_2}^2,$$

and

$$|\mathbf{v}^{\mathrm{T}}R_{ij}E\{S_2(Y_{ij},\mathbf{X}_{ij}|\boldsymbol{\Theta},\boldsymbol{\beta})\mathbf{X}_{ij}|Y_{ij}\}^{\otimes 2}\mathbf{v}|\leq c_0^2d_{S_2}^2.$$

Hence we have each summand in $\mathbf{F}_{\boldsymbol{\beta}}(\mathbf{v},\mathbf{v},\mathbf{X},\mathbf{Y}|\boldsymbol{\Theta},\boldsymbol{\beta})-E\{\mathbf{F}_{\boldsymbol{\beta}}(\mathbf{v},\mathbf{v},\mathbf{X},\mathbf{Y}|\boldsymbol{\Theta},\boldsymbol{\beta})\}$ is in the range of $\{-4c_0^2(d_{H_2}+d_{S_2}^2),4c_0^2(d_{H_2}+d_{S_2}^2)\}$, and in turn is sub-Gaussian with parameter $8c_0^2(d_{H_2}+d_{S_2}^2)$ by Lemma A.1. Here the expectation is taken over $\mathbf{X}$ and $\mathbf{Y}$. Therefore, by Lemma A.1, for any given $\mathbf{v}$

$$\begin{aligned}&\Pr\{|\mathbf{F}_{\boldsymbol{\beta}}(\mathbf{v},\mathbf{v},\mathbf{X},\mathbf{Y}|\boldsymbol{\Theta},\boldsymbol{\beta})-E\{\mathbf{F}_{\boldsymbol{\beta}}(\mathbf{v},\mathbf{v},\mathbf{X},\mathbf{Y}|\boldsymbol{\Theta},\boldsymbol{\beta})\}|>t\}\\ =\quad &\Pr\{|mnF(\mathbf{v},\mathbf{v},\mathbf{X},\mathbf{Y}|\boldsymbol{\Theta},\boldsymbol{\beta})-mnE\{\mathbf{F}_{\boldsymbol{\beta}}(\mathbf{v},\mathbf{v},\mathbf{X},\mathbf{Y}|\boldsymbol{\Theta},\boldsymbol{\beta})\}|>mnt\}\\ \leq\quad &2\exp\left[-\frac{(mn)^2t^2}{2mn\{8c_0^2(d_{H_2}+d_{S_2}^2)\}}\right]\\ =\quad &2\exp\left\{-\frac{mnt^2}{16c_0^2(d_{H_2}+d_{S_2}^2)}\right\}.\end{aligned}$$

Now choose $t=\sqrt{16c_0^2(d_{H_2}+d_{S_2}^2)\log\{\max(p,mn)\}/(mn)}$, we obtain that for any given unit vector $\mathbf{v}$,

$$\begin{aligned}&\Pr\left\{|\mathbf{F}_{\boldsymbol{\beta}}(\mathbf{v},\mathbf{v},\mathbf{X},\mathbf{Y}|\boldsymbol{\Theta},\boldsymbol{\beta})-E\{\mathbf{F}_{\boldsymbol{\beta}}(\mathbf{v},\mathbf{v},\mathbf{X},\mathbf{Y}|\boldsymbol{\Theta},\boldsymbol{\beta})\}|\right.\\ &\left.>\sqrt{16c_0^2(d_{H_2}+d_{S_2}^2)\log\{\max(p,mn)\}/(mn)}\right\}\leq 2\max(p,mn)^{-1}.\end{aligned}\tag{30}$$

Therefore,

$$|\mathbf{F}_{\boldsymbol{\beta}}(\mathbf{v},\mathbf{v},\mathbf{X},\mathbf{Y}|\boldsymbol{\Theta},\boldsymbol{\beta})-E\{\mathbf{F}_{\boldsymbol{\beta}}(\mathbf{v},\mathbf{v},\mathbf{X},\mathbf{Y}|\boldsymbol{\Theta},\boldsymbol{\beta})\}|\leq 4\sqrt{c_0^2(d_{H_2}+d_{S_2}^2)\log\{\max(p,mn)\}/(mn)},$$

with probability at least $1-2\max(p,mn)^{-1}$. This proves (28).

Furthermore, for unit vector $\mathbf{u}\in\mathbb{R}^{mn}$, we have

$$\begin{aligned}&|\mathbf{u}^{\mathrm{T}}R_{ij}H_2(Y_{ij},\mathbf{X}_{ij}|\boldsymbol{\Theta},\boldsymbol{\beta})\mathrm{vec}(\mathbf{z}_{ij})\mathrm{vec}(\mathbf{z}_{ij})^{\mathrm{T}}\mathbf{u}|\\ \leq\quad &\{\|\mathbf{u}\|_\infty\|\mathrm{vec}(\mathbf{z}_{ij})\|_1\}^2|H_2(Y_{ij},\mathbf{X}_{ij}|\boldsymbol{\Theta},\boldsymbol{\beta})|\leq d_{H_2}.\end{aligned}$$

Similarly,

$$\begin{aligned}|\mathbf{u}^{\mathrm{T}}R_{ij}E\{H_2(Y_{ij},\mathbf{X}_{ij}|\boldsymbol{\Theta},\boldsymbol{\beta})\mathrm{vec}(\mathbf{z}_{ij})\mathrm{vec}(\mathbf{z}_{ij})^{\mathrm{T}}|Y_{ij}\}\mathbf{u}| &\leq d_{H_2},\\ |\mathbf{u}^{\mathrm{T}}R_{ij}E\{S_2^2(Y_{ij},\mathbf{X}_{ij}|\boldsymbol{\Theta},\boldsymbol{\beta})\mathrm{vec}(\mathbf{z}_{ij})\mathrm{vec}(\mathbf{z}_{ij})^{\mathrm{T}}|Y_{ij}\}\mathbf{u}| &\leq d_{S_2}^2,\\ |\mathbf{u}^{\mathrm{T}}R_{ij}E\{S_2(Y_{ij},\mathbf{X}_{ij}|\boldsymbol{\Theta},\boldsymbol{\beta})\mathrm{vec}(\mathbf{z}_{ij})|Y_{ij}\}^{\otimes 2}\mathbf{u}| &\leq d_{S_2}^2.\end{aligned}$$

Hence each summand in $\mathbf{u}^{\mathrm{T}}\mathbf{F}_{\boldsymbol{\Theta}}(\mathbf{X},\mathbf{Y}|\boldsymbol{\Theta},\boldsymbol{\beta})\mathbf{u}-\mathbf{u}^{\mathrm{T}}E\{\mathbf{F}_{\boldsymbol{\Theta}}(\mathbf{X},\mathbf{Y}|\boldsymbol{\Theta},\boldsymbol{\beta})\}\mathbf{u}$ is in the range of $[-4(d_{H_2}+d_{S_2}^2),4(d_{H_2}+d_{S_2}^2)]$, and in turn is sub-Gaussian with parameter $8(d_{H_2}+d_{S_2}^2)$ by Lemma A.1. Lemma A.1 further leads to

$$\begin{aligned}&\Pr\left\{|\mathbf{u}^{\mathrm{T}}\mathbf{F}_{\boldsymbol{\Theta}}(\mathbf{X},\mathbf{Y}|\boldsymbol{\Theta},\boldsymbol{\beta})\mathbf{u}-\mathbf{u}^{\mathrm{T}}E\{\mathbf{F}_{\boldsymbol{\Theta}}(X,\mathbf{Y}|\boldsymbol{\Theta},\boldsymbol{\beta})\}\mathbf{u}|>t\right\}\\ =\;&\Pr\left\{|mn\mathbf{u}^{\mathrm{T}}\mathbf{F}_{\boldsymbol{\Theta}}(\mathbf{X},\mathbf{Y}|\boldsymbol{\Theta},t\boldsymbol{\beta})\mathbf{u}-mnE\{\mathbf{u}^{\mathrm{T}}\mathbf{F}_{\boldsymbol{\Theta}}(\mathbf{X},\mathbf{Y}|\boldsymbol{\Theta},\boldsymbol{\beta})\}\mathbf{u}|>mnt\right\}\\ \leq\;&2\exp\left[-\frac{(mn)^2t^2}{2mn\{8(d_{H_2}+d_{S_2}^2)\}}\right]\\ =\;&2\exp\left\{-\frac{mnt^2}{16(d_{H_2}+d_{S_2}^2)}\right\}\end{aligned}$$

for any $t>0$. Let $t=\sqrt{16(d_{H_2}+d_{S_2}^2)\log(mn)/(mn)}$, we have

$$\mathbf{u}^{\mathrm{T}}\mathbf{F}_{\boldsymbol{\Theta}}(\mathbf{X},\mathbf{Y}|\boldsymbol{\Theta},\boldsymbol{\beta})\mathbf{u}\geq\mathbf{u}^{\mathrm{T}}E\{\mathbf{F}_{\boldsymbol{\Theta}}(\mathbf{X},\mathbf{Y}|\boldsymbol{\Theta},\boldsymbol{\beta})\}\mathbf{u}-4\sqrt{(d_{H_2}+d_{S_2}^2)\log(mn)/(mn)} \tag{31}$$

with probability at least $1-2(mn)^{-1}$.

Moreover, using the same arguments as those lead to (30), (31), and the fact that $\|\mathbf{z}_{ij}\|_1=1$, we have for unit vector $\mathbf{w}\in\mathbb{R}^{mn+p}$,

$$\begin{aligned}&\mathbf{w}^{\mathrm{T}}\partial^2\mathcal{L}(\boldsymbol{\Theta},\boldsymbol{\beta})/\partial[\{\mathrm{vec}(\boldsymbol{\Theta})^{\mathrm{T}},\boldsymbol{\beta}^{\mathrm{T}}\}^{\mathrm{T}}]^{\otimes 2}\mathbf{w}\\ \geq\;&E\left(\mathbf{w}^{\mathrm{T}}\partial^2\mathcal{L}(\boldsymbol{\Theta},\boldsymbol{\beta})/\partial[\{\mathrm{vec}(\boldsymbol{\Theta})^{\mathrm{T}},\boldsymbol{\beta}^{\mathrm{T}}\}^{\mathrm{T}}]^{\otimes 2}\mathbf{w}\right)-4\sqrt{(c_0+1)^2(d_{H_2}+d_{S_2}^2)\log(mn+p)/mn},\end{aligned}$$

with probability at least $1-2(p+mn)^{-1}$.

Lemma A.22 *Assume*

$$\boldsymbol{\beta}^{\mathrm{T}}E\{\mathbf{F}_{\boldsymbol{\beta}}(\mathbf{X},\mathbf{Y}|\widetilde{\boldsymbol{\Theta}},\widetilde{\boldsymbol{\beta}})\}\boldsymbol{\beta}>4c_0\sqrt{(d_{H_2}+d_{S_2}^2)\log\{\max(p,mn)\}/(mn)}\|\boldsymbol{\beta}\|_2^2,$$

for all $\boldsymbol{\beta}$ that satisfies $\|\boldsymbol{\beta}\|_\infty\leq 2a$ and $\widetilde{\boldsymbol{\beta}},\widetilde{\boldsymbol{\Theta}}$ that satisfy $\|\widetilde{\boldsymbol{\beta}}\|_\infty\leq a$ and $\|\widetilde{\boldsymbol{\Theta}}\|_{\max}\leq a$. Recall that $\sigma_{\boldsymbol{\beta}}=C_1(2d_{H_2}+2d_{S_2}^2)c_0^2$ for $C_1>1$, $Q^{t-1}=F(\boldsymbol{\Theta}^{t-1},\boldsymbol{\beta}^{t-1})/\lambda_{\boldsymbol{\beta}}$ and

$$g_{\boldsymbol{\beta}}(\boldsymbol{\Theta}^{t-1},\boldsymbol{\beta}^{t-1},Q^{t-1})=\langle\partial\mathcal{L}(\boldsymbol{\Theta}^{t-1},\boldsymbol{\beta}^{t-1})/\partial\boldsymbol{\beta},\boldsymbol{\beta}^{t-1}-\widetilde{\boldsymbol{\beta}}^{t-1}\rangle+\lambda_{\boldsymbol{\beta}}\|\boldsymbol{\beta}^{t-1}-\widetilde{\boldsymbol{\beta}}^{t-1}\|_1,$$

where

$$\widetilde{\boldsymbol{\beta}}^{t-1}\equiv argmin_{\boldsymbol{\beta}}\langle\partial\mathcal{L}(\boldsymbol{\Theta}^{t-1},\boldsymbol{\beta}^{t-1})/\partial\boldsymbol{\beta},\boldsymbol{\beta}\rangle+\lambda_{\boldsymbol{\beta}}\|\boldsymbol{\beta}\|_1,\|\boldsymbol{\beta}\|_1\leq Q^{t-1}.$$

Select

$$0<\eta<1/\sigma_{\boldsymbol{\beta}}.$$

Then we have

$$F(\boldsymbol{\Theta}^{t-1},\boldsymbol{\beta}^t)\leq F(\boldsymbol{\Theta}^{t-1},\boldsymbol{\beta}^{t-1})-\frac{\eta}{2}\frac{g_{\boldsymbol{\beta}}(\boldsymbol{\Theta}^{t-1},\boldsymbol{\beta}^{t-1},Q^{t-1})^2}{(2Q^{t-1})^2},$$

with probability at least $1-2\max(p,mn)^{-1}$.

Proof: Because

$$\boldsymbol{\beta}^{\mathrm{T}} E\{\mathbf{F}_{\boldsymbol{\beta}}(\mathbf{X},\mathbf{Y}|\widetilde{\boldsymbol{\Theta}},\widetilde{\boldsymbol{\beta}})\}\boldsymbol{\beta} \quad > \quad 4c_0\sqrt{(d_{H_2}+d_{S_2}^2)\log\{\max(p,mn)\}/(mn)}\|\boldsymbol{\beta}\|_2^2,$$

the second derivative of $\mathcal{L}(\boldsymbol{\Theta},\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$ is positive definite with probability at least $1-2\max(p,mn)^{-1}$ by Lemma A.21, and hence we majorize $F(\boldsymbol{\Theta}^{t-1},\boldsymbol{\beta})$ at $\boldsymbol{\beta}^{t-1}$ as

$$\begin{aligned} F(\boldsymbol{\Theta}^{t-1},\boldsymbol{\beta}) &\leq F(\boldsymbol{\Theta}^{t-1},\boldsymbol{\beta}^{t-1}) + \langle \boldsymbol{\beta}-\boldsymbol{\beta}^{t-1}, \partial\mathcal{L}(\boldsymbol{\Theta}^{t-1},\boldsymbol{\beta}^{t-1})/\partial\boldsymbol{\beta}\rangle \\ &\quad + \frac{\sigma_{\boldsymbol{\beta}}}{2}\|\boldsymbol{\beta}-\boldsymbol{\beta}^{t-1}\|_2^2 + \lambda_{\boldsymbol{\beta}}(\|\boldsymbol{\beta}\|_1 - \|\boldsymbol{\beta}^{t-1}\|_1) \\ &\leq F(\boldsymbol{\Theta}^{t-1},\boldsymbol{\beta}^{t-1}) + \langle \boldsymbol{\beta}-\boldsymbol{\beta}^{t-1}, \partial\mathcal{L}(\boldsymbol{\Theta}^{t-1},\boldsymbol{\beta}^{t-1})/\partial\boldsymbol{\beta}\rangle \\ &\quad + \frac{1}{2\eta}\|\boldsymbol{\beta}-\boldsymbol{\beta}^{t-1}\|_2^2 + \lambda_{\boldsymbol{\beta}}(\|\boldsymbol{\beta}\|_1 - \|\boldsymbol{\beta}^{t-1}\|_1) \end{aligned} \tag{32}$$

for any $\boldsymbol{\beta}$, where we have used $\sigma_{\boldsymbol{\beta}} \leq 1/\eta$.

Furthermore, for any given $b \geq 0$, let $\boldsymbol{\beta}^t = b\widetilde{\boldsymbol{\beta}}^{t-1} + (1-b)\boldsymbol{\beta}^{t-1}$, where recall that

$$\widetilde{\boldsymbol{\beta}}^{t-1} \equiv \mathrm{argmin}_{\boldsymbol{\beta}}\langle \partial\mathcal{L}(\boldsymbol{\Theta}^{t-1},\boldsymbol{\beta}^{t-1})/\partial\boldsymbol{\beta},\boldsymbol{\beta}\rangle + \lambda_{\boldsymbol{\beta}}\|\boldsymbol{\beta}\|_1, \|\boldsymbol{\beta}\|_1 \leq Q^{t-1} \tag{33}$$

with $Q^{t-1} = F(\boldsymbol{\Theta}^{t-1},\boldsymbol{\beta}^{t-1})/\lambda_{\boldsymbol{\beta}}$, it holds that

$$\begin{aligned} F(\boldsymbol{\Theta}^{t-1},\boldsymbol{\beta}^t) &\leq F(\boldsymbol{\Theta}^{t-1},\boldsymbol{\beta}^{t-1}) + b\langle \partial\mathcal{L}(\boldsymbol{\Theta}^{t-1},\boldsymbol{\beta}^{t-1})/\partial\boldsymbol{\beta}, \widetilde{\boldsymbol{\beta}}^{t-1}-\boldsymbol{\beta}^{t-1}\rangle + \frac{b^2}{2\eta}\|\widetilde{\boldsymbol{\beta}}^{t-1}-\boldsymbol{\beta}^{t-1}\|_2^2 \\ &\quad + \lambda_{\boldsymbol{\beta}}(\|b\widetilde{\boldsymbol{\beta}}^{t-1} + (1-b)\boldsymbol{\beta}^{t-1}\|_1 - \|\boldsymbol{\beta}^{t-1}\|_1) \\ &\leq F(\boldsymbol{\Theta}^{t-1},\boldsymbol{\beta}^{t-1}) + b\langle \partial\mathcal{L}(\boldsymbol{\Theta}^{t-1},\boldsymbol{\beta}^{t-1})/\partial\boldsymbol{\beta}, \widetilde{\boldsymbol{\beta}}^{t-1}-\boldsymbol{\beta}^{t-1}\rangle + \frac{b^2}{2\eta}\|\widetilde{\boldsymbol{\beta}}^{t-1}-\boldsymbol{\beta}^{t-1}\|_2^2 \\ &\quad + b\lambda_{\boldsymbol{\beta}}(\|\widetilde{\boldsymbol{\beta}}^{t-1}-\boldsymbol{\beta}^{t-1}\|_1). \end{aligned} \tag{34}$$

To minimize the right hand side, we set

$$\begin{aligned} b &\equiv \frac{\eta\{-\langle \partial\mathcal{L}(\boldsymbol{\Theta}^{t-1},\boldsymbol{\beta}^{t-1})/\partial\boldsymbol{\beta}, \widetilde{\boldsymbol{\beta}}^{t-1}-\boldsymbol{\beta}^{t-1}\rangle - \lambda_{\boldsymbol{\beta}}(\|\widetilde{\boldsymbol{\beta}}^{t-1}-\boldsymbol{\beta}^{t-1}\|_1)\}}{\|\widetilde{\boldsymbol{\beta}}^{t-1}-\boldsymbol{\beta}^{t-1}\|_2^2} \\ &= \frac{\eta g_{\boldsymbol{\beta}}(\boldsymbol{\Theta}^{t-1},\boldsymbol{\beta}^{t-1},Q^{t-1})}{\|\widetilde{\boldsymbol{\beta}}^{t-1}-\boldsymbol{\beta}^{t-1}\|_2^2}. \end{aligned}$$

Note that $g_{\boldsymbol{\beta}}(\boldsymbol{\Theta}^{t-1},\boldsymbol{\beta}^{t-1},Q^{t-1}) \geq 0$ due to its definition and the definition of $\widetilde{\boldsymbol{\beta}}^{t-1}$, hence $b \geq 0$.

Note that $F(\boldsymbol{\Theta}^{t-1},\boldsymbol{\beta}^{t-1}) = \mathcal{L}(\boldsymbol{\Theta}^{t-1},\boldsymbol{\beta}^{t-1}) - L + \lambda_{\boldsymbol{\Theta}}\|\boldsymbol{\Theta}^{t-1}\|_* + \lambda_{\boldsymbol{\beta}}\|\boldsymbol{\beta}^{t-1}\|_1 \geq \lambda_{\boldsymbol{\beta}}\|\boldsymbol{\beta}^{t-1}\|_1$, hence $\|\boldsymbol{\beta}^{t-1}\|_1 \leq Q^{t-1}$. Since $\widetilde{\boldsymbol{\beta}}^{t-1}$ is the minimizer of (33), we get

$$\begin{aligned} & g_{\boldsymbol{\beta}}(\boldsymbol{\Theta}^{t-1},\boldsymbol{\beta}^{t-1},Q^{t-1}) \\ = \; & \langle \partial\mathcal{L}(\boldsymbol{\Theta}^{t-1},\boldsymbol{\beta}^{t-1})/\partial\boldsymbol{\beta}, \boldsymbol{\beta}^{t-1}-\widetilde{\boldsymbol{\beta}}^{t-1}\rangle + \lambda_{\boldsymbol{\beta}}\|\boldsymbol{\beta}^{t-1}-\widetilde{\boldsymbol{\beta}}^{t-1}\|_1 \\ = \; & \langle \partial\mathcal{L}(\boldsymbol{\Theta}^{t-1},\boldsymbol{\beta}^{t-1})/\partial\boldsymbol{\beta}, \boldsymbol{\beta}^{t-1}-\widetilde{\boldsymbol{\beta}}^{t-1}\rangle + \lambda_{\boldsymbol{\beta}}\|\boldsymbol{\beta}^{t-1}\|_1 - \lambda_{\boldsymbol{\beta}}\|\widetilde{\boldsymbol{\beta}}^{t-1}\|_1 \geq 0. \end{aligned}$$

Plug b in (34), we have

$$F(\boldsymbol{\Theta}^{t-1},\boldsymbol{\beta}^t) - F(\boldsymbol{\Theta}^{t-1},\boldsymbol{\beta}^{t-1}) \leq -\frac{\eta}{2}\frac{g_{\boldsymbol{\beta}}(\boldsymbol{\Theta}^{t-1},\boldsymbol{\beta}^{t-1},Q^{t-1})^2}{\|\widetilde{\boldsymbol{\beta}}^{t-1}-\boldsymbol{\beta}^{t-1}\|_2^2} \leq -\frac{\eta}{2}\frac{g_{\boldsymbol{\beta}}(\boldsymbol{\Theta}^{t-1},\boldsymbol{\beta}^{t-1},Q^{t-1})^2}{(2Q^{t-1})^2}. \tag{35}$$

The last equality hods because $\|\boldsymbol{\beta}^{t-1}\|_1 \leq Q^{t-1}$ and $\|\widetilde{\boldsymbol{\beta}}^{t-1}\|_1 \leq Q^{t-1}$, and hence $\|\widetilde{\boldsymbol{\beta}}^{t-1}-\boldsymbol{\beta}^{t-1}\|_2^2 \leq (2Q^{t-1})^2$. This proves the result.

Lemma A.23 *Assume*

$$\mathrm{vec}(\boldsymbol{\Theta})^{\mathrm{T}} E\{\mathbf{F}_{\boldsymbol{\Theta}}(\mathbf{X},\mathbf{Y}|\widetilde{\boldsymbol{\Theta}},\widetilde{\boldsymbol{\beta}})\}\mathrm{vec}(\boldsymbol{\Theta}) \quad > \quad 4\sqrt{(d_{H_2}+d_{S_2}^2)\log(mn)/(mn)}\|\boldsymbol{\Theta}\|_F^2$$

for all $\boldsymbol{\Theta}$ *that satisfies* $\|\boldsymbol{\Theta}\|_{\max} \le 2a$ *and* $\widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\Theta}}$ *that satisfy* $\|\widetilde{\boldsymbol{\beta}}\|_\infty \le a$ *and* $\|\widetilde{\boldsymbol{\Theta}}\|_{\max} \le a$. *Recall that* $\sigma_{\boldsymbol{\Theta}} = 2d_{H_2} + 2d_{S_2}^2$, $R^{t-1} = F(\boldsymbol{\Theta}^{t-1}, \boldsymbol{\beta}^t)/\lambda_{\boldsymbol{\Theta}}$, *and*

$$g_{\boldsymbol{\Theta}}(\boldsymbol{\Theta}^{t-1}, \boldsymbol{\beta}^t, R^{t-1}) = \langle \mathcal{L}(\boldsymbol{\Theta}^{t-1}, \boldsymbol{\beta}^t)/\partial\boldsymbol{\Theta}, \boldsymbol{\Theta}^{t-1} - \widetilde{\boldsymbol{\Theta}}^{t-1}\rangle + \lambda_{\boldsymbol{\Theta}}\|\boldsymbol{\Theta}^{t-1} - \widetilde{\boldsymbol{\Theta}}^{t-1}\|_*.$$

where

$$\widetilde{\boldsymbol{\Theta}}^{t-1} \equiv \mathit{argmin}_{\boldsymbol{\Theta}}\langle \partial\mathcal{L}(\boldsymbol{\Theta}^{t-1}, \boldsymbol{\beta}^t)/\partial\boldsymbol{\Theta}, \boldsymbol{\Theta}\rangle + \lambda_{\boldsymbol{\Theta}}\|\boldsymbol{\Theta}\|_*, \|\boldsymbol{\Theta}\|_* \le R^{t-1}.$$

Select $0 < \eta_1 < 1/\sigma_{\boldsymbol{\Theta}}$. *Then we have*

$$F(\boldsymbol{\Theta}^t, \boldsymbol{\beta}^t) \le F(\boldsymbol{\Theta}^{t-1}, \boldsymbol{\beta}^t) - \frac{\eta_1}{2}\frac{g_{\boldsymbol{\Theta}}(\boldsymbol{\Theta}^{t-1}, \boldsymbol{\beta}^t, R^{t-1})^2}{(2R^{t-1})^2},$$

with probability at least $1 - 2(mn)^{-1}$.

Proof: Because

$$E\{\mathrm{vec}(\boldsymbol{\Theta})^{\mathrm{T}}\mathbf{F}_{\boldsymbol{\Theta}}(\mathbf{X},\mathbf{Y}|\widetilde{\boldsymbol{\Theta}},\widetilde{\boldsymbol{\beta}})\}\mathrm{vec}(\boldsymbol{\Theta}) \quad > \quad 4\sqrt{(d_{H_2}+d_{S_2}^2)\log(mn)/(mn)}\|\boldsymbol{\Theta}\|_F^2,$$

then the second derivative of $\mathcal{L}(\boldsymbol{\Theta}, \boldsymbol{\beta})$ with respect to $\mathrm{vec}(\boldsymbol{\Theta})$ is positive definite with probability at least $1 - 2(mn)^{-1}$ by Lemma A.21, and hence we majorize $F(\boldsymbol{\Theta}, \boldsymbol{\beta}^t)$ at $\boldsymbol{\Theta}^{t-1}$ as

$$\begin{aligned} F(\boldsymbol{\Theta}, \boldsymbol{\beta}^t) &\le F(\boldsymbol{\Theta}^{t-1}, \boldsymbol{\beta}^t) + \langle \boldsymbol{\Theta} - \boldsymbol{\Theta}^{t-1}, \partial\mathcal{L}(\boldsymbol{\Theta}^{t-1}, \boldsymbol{\beta}^t)/\partial\boldsymbol{\Theta}\rangle \\ &\quad + \frac{\sigma_{\boldsymbol{\Theta}}}{2}\|\boldsymbol{\Theta} - \boldsymbol{\Theta}^{t-1}\|_F^2 + \lambda_{\boldsymbol{\Theta}}(\|\boldsymbol{\Theta}\|_* - \|\boldsymbol{\Theta}^{t-1}\|_*) \\ &\le F(\boldsymbol{\Theta}^{t-1}, \boldsymbol{\beta}^t) + \langle \boldsymbol{\Theta} - \boldsymbol{\Theta}^{t-1}, \partial\mathcal{L}(\boldsymbol{\Theta}^{t-1}, \boldsymbol{\beta}^t)/\partial\boldsymbol{\Theta}\rangle \\ &\quad + \frac{1}{2\eta_1}\|\boldsymbol{\Theta} - \boldsymbol{\Theta}^{t-1}\|_F^2 + \lambda_{\boldsymbol{\Theta}}(\|\boldsymbol{\Theta}\|_* - \|\boldsymbol{\Theta}^{t-1}\|_*) \end{aligned}$$

for any $\boldsymbol{\Theta}$.

Furthermore, for any given $b \ge 0$, let $\boldsymbol{\Theta}^t = b\widetilde{\boldsymbol{\Theta}}^{t-1} + (1-b)\boldsymbol{\Theta}^{t-1}$, where recall that

$$\widetilde{\boldsymbol{\Theta}}^{t-1} \equiv \mathrm{argmin}_{\boldsymbol{\Theta}}\langle \partial\mathcal{L}(\boldsymbol{\Theta}^{t-1}, \boldsymbol{\beta}^t)/\partial\boldsymbol{\Theta}, \boldsymbol{\Theta}\rangle + \lambda_{\boldsymbol{\Theta}}\|\boldsymbol{\Theta}\|_*, \|\boldsymbol{\Theta}\|_* \le R^{t-1} \tag{36}$$

with $R^{t-1} = F(\boldsymbol{\Theta}^{t-1}, \boldsymbol{\beta}^t)/\lambda_{\boldsymbol{\Theta}}$, it holds that

$$\begin{aligned} F(\boldsymbol{\Theta}^t, \boldsymbol{\beta}^t) &\le F(\boldsymbol{\Theta}^{t-1}, \boldsymbol{\beta}^t) + b\langle \partial\mathcal{L}(\boldsymbol{\Theta}^{t-1}, \boldsymbol{\beta}^t)/\partial\boldsymbol{\Theta}, \widetilde{\boldsymbol{\Theta}}^{t-1} - \boldsymbol{\Theta}^{t-1}\rangle + \frac{b^2}{2\eta_1}\|\widetilde{\boldsymbol{\Theta}}^{t-1} - \boldsymbol{\Theta}^{t-1}\|_F^2 \\ &\quad + \lambda_{\boldsymbol{\Theta}}(\|b\widetilde{\boldsymbol{\Theta}}^{t-1} + (1-b)\boldsymbol{\Theta}^{t-1}\|_* - \|\boldsymbol{\Theta}^{t-1}\|_*) \\ &\le F(\boldsymbol{\Theta}^{t-1}, \boldsymbol{\beta}^t) + b\langle \partial\mathcal{L}(\boldsymbol{\Theta}^{t-1}, \boldsymbol{\beta}^t)/\partial\boldsymbol{\Theta}, \widetilde{\boldsymbol{\Theta}}^{t-1} - \boldsymbol{\Theta}^{t-1}\rangle + \frac{b^2}{2\eta_1}\|\widetilde{\boldsymbol{\Theta}}^{t-1} - \boldsymbol{\Theta}^{t-1}\|_F^2 \\ &\quad + b\lambda_{\boldsymbol{\Theta}}(\|\widetilde{\boldsymbol{\Theta}}^{t-1} - \boldsymbol{\Theta}^{t-1}\|_*). \end{aligned} \tag{37}$$

To minimize the right hand side, we set

$$b \equiv \frac{\eta_1\{-\langle \partial\mathcal{L}(\boldsymbol{\Theta}^{t-1}, \boldsymbol{\beta}^t)/\partial\boldsymbol{\Theta}, \widetilde{\boldsymbol{\Theta}}^{t-1} - \boldsymbol{\Theta}^{t-1}\rangle - \lambda_{\boldsymbol{\Theta}}(\|\widetilde{\boldsymbol{\Theta}}^{t-1} - \boldsymbol{\Theta}^{t-1}\|_*)\}}{\|\widetilde{\boldsymbol{\Theta}}^{t-1} - \boldsymbol{\Theta}^{t-1}\|_F^2}$$

$$= \frac{\eta_1 g_{\boldsymbol{\Theta}}(\boldsymbol{\Theta}^{t-1}, \boldsymbol{\beta}^t, R^{t-1})}{\|\widetilde{\boldsymbol{\Theta}}^{t-1} - \boldsymbol{\Theta}^{t-1}\|_F^2}.$$

Note that $g_{\boldsymbol{\Theta}}(\boldsymbol{\Theta}^{t-1}, \boldsymbol{\beta}^t, Q^{t-1}) \geq 0$ due to its definition and the definition of $\widetilde{\boldsymbol{\Theta}}^{t-1}$, hence $b \geq 0$. Note that $F(\boldsymbol{\Theta}^{t-1}, \boldsymbol{\beta}^t) = \mathcal{L}(\boldsymbol{\Theta}^{t-1}, \boldsymbol{\beta}^t) - L + \lambda_{\boldsymbol{\Theta}}\|\boldsymbol{\Theta}^{t-1}\|_* + \lambda_{\boldsymbol{\beta}}\|\boldsymbol{\beta}^t\|_1 \geq \lambda_{\boldsymbol{\Theta}}\|\boldsymbol{\Theta}^{t-1}\|_*$, hence $\|\boldsymbol{\Theta}^{t-1}\|_* \leq R^{t-1}$. Since $\widetilde{\boldsymbol{\Theta}}^{t-1}$ is the minimizer of (36), we get

$$\begin{aligned} & g_{\boldsymbol{\Theta}}(\boldsymbol{\Theta}^{t-1}, \boldsymbol{\beta}^t, R^{t-1}) \\ \equiv\ & \langle \partial \mathcal{L}(\boldsymbol{\Theta}^{t-1}, \boldsymbol{\beta}^t)/\partial \boldsymbol{\Theta}, \boldsymbol{\Theta}^{t-1} - \widetilde{\boldsymbol{\Theta}}^{t-1} \rangle + \lambda_{\boldsymbol{\Theta}}\|\boldsymbol{\Theta}^{t-1} - \widetilde{\boldsymbol{\Theta}}^{t-1}\|_* \\ \geq\ & \langle \partial \mathcal{L}(\boldsymbol{\Theta}^{t-1}, \boldsymbol{\beta}^t)/\partial \boldsymbol{\Theta}, \boldsymbol{\Theta}^{t-1} - \widetilde{\boldsymbol{\Theta}}^{t-1} \rangle + \lambda_{\boldsymbol{\Theta}}\|\boldsymbol{\Theta}^{t-1}\|_* - \lambda_{\boldsymbol{\Theta}}\|\widetilde{\boldsymbol{\Theta}}^{t-1}\|_* \geq 0. \end{aligned}$$

Plug b in (37), we have

$$F(\boldsymbol{\Theta}^t, \boldsymbol{\beta}^t) - F(\boldsymbol{\Theta}^{t-1}, \boldsymbol{\beta}^t) \leq -\frac{\eta_1}{2} \frac{g_{\boldsymbol{\Theta}}(\boldsymbol{\Theta}^{t-1}, \boldsymbol{\beta}^t, R^{t-1})^2}{\|\widetilde{\boldsymbol{\Theta}}^{t-1} - \boldsymbol{\Theta}^{t-1}\|_F^2} \leq -\frac{\eta_1}{2} \frac{g_{\boldsymbol{\Theta}}(\boldsymbol{\Theta}^{t-1}, \boldsymbol{\beta}^t, R^{t-1})^2}{(2R^{t-1})^2}. \tag{38}$$

The last equality hods because $\|\boldsymbol{\Theta}^{t-1}\|_1 \leq R^{t-1}$ and $\|\widetilde{\boldsymbol{\Theta}}^{t-1}\|_1 \leq R^{t-1}$, and hence $\|\widetilde{\boldsymbol{\Theta}}^{t-1} - \boldsymbol{\Theta}^{t-1}\|_2^2 \leq (2R^{t-1})^2$. This proves the result.

Appendix H. Proof of Theorem 3

Proof: Let

$$\widetilde{\boldsymbol{\beta}}^t = \text{argmin} \langle \partial \mathcal{L}(\boldsymbol{\Theta}^{t-1}, \boldsymbol{\beta}^t)/\partial \boldsymbol{\beta}, \boldsymbol{\beta} \rangle + \lambda_{\boldsymbol{\beta}}\|\boldsymbol{\beta}\|_1, \|\boldsymbol{\beta}\|_1 \leq Q^t,$$

and recall that

$$\widetilde{\boldsymbol{\Theta}}^{t-1} \equiv \text{argmin}_{\boldsymbol{\Theta}} \langle \partial \mathcal{L}(\boldsymbol{\Theta}^{t-1}, \boldsymbol{\beta}^t)/\partial \boldsymbol{\Theta}, \boldsymbol{\Theta} \rangle + \lambda_{\boldsymbol{\Theta}}\|\boldsymbol{\Theta}\|_*, \|\boldsymbol{\Theta}\|_* \leq R^{t-1}$$

as defined in (36). Because

$$\begin{aligned} & E\{\{\boldsymbol{\beta}^{\mathrm{T}}, \text{vec}(\boldsymbol{\Theta})^{\mathrm{T}}\}\partial^2 \mathcal{L}(\boldsymbol{\Theta}, \boldsymbol{\beta})/\partial[\{\text{vec}(\boldsymbol{\Theta})^{\mathrm{T}}, \boldsymbol{\beta}^{\mathrm{T}}\}^{\mathrm{T}}]^{\otimes 2}\{\boldsymbol{\beta}^{\mathrm{T}}, \text{vec}(\boldsymbol{\Theta})^{\mathrm{T}}\}^{\mathrm{T}}\} \\ \geq\ & 4(c_0+1)\sqrt{(d_{H_2} + d_{S_2}^2)\log\{\max(p, mn)\}/(mn)}\|\{\boldsymbol{\beta}^{\mathrm{T}}, \text{vec}(\boldsymbol{\Theta})^{\mathrm{T}}\}^{\mathrm{T}}\|_2^2, \end{aligned}$$

we have

$$\begin{aligned} & F(\widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\beta}}) - F(\boldsymbol{\Theta}^{t-1}, \boldsymbol{\beta}^t) \\ \geq\ & \langle \partial \mathcal{L}(\boldsymbol{\Theta}^{t-1}, \boldsymbol{\beta}^t)/\partial \boldsymbol{\Theta}, \widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^{t-1} \rangle + \lambda_{\boldsymbol{\Theta}}(\|\widehat{\boldsymbol{\Theta}}\|_* - \|\boldsymbol{\Theta}^{t-1}\|_*) \\ & + \langle \partial \mathcal{L}(\boldsymbol{\Theta}^{t-1}, \boldsymbol{\beta}^t)/\partial \boldsymbol{\beta}, \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^t \rangle + \lambda_{\boldsymbol{\beta}}(\|\widehat{\boldsymbol{\beta}}\|_1 - \|\boldsymbol{\beta}^t\|_1), \end{aligned}$$

with probability at least $1 - 2(mn+p)^{-1}$ by Lemma A.21. Hence

$$\begin{aligned} F(\boldsymbol{\Theta}^{t-1}, \boldsymbol{\beta}^t) - F(\widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\beta}}) \leq\ & \langle \partial \mathcal{L}(\boldsymbol{\Theta}^{t-1}, \boldsymbol{\beta}^t)/\partial \boldsymbol{\Theta}, \boldsymbol{\Theta}^{t-1} - \widehat{\boldsymbol{\Theta}} \rangle + \lambda_{\boldsymbol{\Theta}}(\|\boldsymbol{\Theta}^{t-1}\|_* - \|\widehat{\boldsymbol{\Theta}}\|_*) \\ & + \langle \partial \mathcal{L}(\boldsymbol{\Theta}^{t-1}, \boldsymbol{\beta}^t)/\partial \boldsymbol{\beta}, \boldsymbol{\beta}^t - \widehat{\boldsymbol{\beta}} \rangle + \lambda_{\boldsymbol{\beta}}(\|\boldsymbol{\beta}^t\|_1 - \|\widehat{\boldsymbol{\beta}}\|_1) \\ \leq\ & \langle \partial \mathcal{L}(\boldsymbol{\Theta}^{t-1}, \boldsymbol{\beta}^t)/\partial \boldsymbol{\Theta}, \boldsymbol{\Theta}^{t-1} - \widetilde{\boldsymbol{\Theta}}^{t-1} \rangle + \lambda_{\boldsymbol{\Theta}}(\|\boldsymbol{\Theta}^{t-1}\|_* - \|\widetilde{\boldsymbol{\Theta}}^{t-1}\|_*) \\ & + \langle \partial \mathcal{L}(\boldsymbol{\Theta}^{t-1}, \boldsymbol{\beta}^t)/\partial \boldsymbol{\beta}, \boldsymbol{\beta}^t - \widetilde{\boldsymbol{\beta}}^t \rangle + \lambda_{\boldsymbol{\beta}}(\|\boldsymbol{\beta}^t\|_1 - \|\widetilde{\boldsymbol{\beta}}^t\|_1) \\ \leq\ & \langle \partial \mathcal{L}(\boldsymbol{\Theta}^{t-1}, \boldsymbol{\beta}^t)/\partial \boldsymbol{\Theta}, \boldsymbol{\Theta}^{t-1} - \widetilde{\boldsymbol{\Theta}}^{t-1} \rangle + \lambda_{\boldsymbol{\Theta}}(\|\boldsymbol{\Theta}^{t-1} - \widetilde{\boldsymbol{\Theta}}^{t-1}\|_*) \end{aligned}$$

$$
\begin{aligned}
& +\langle \partial \mathcal{L}(\boldsymbol{\Theta}^{t-1}, \boldsymbol{\beta}^t)/\partial \boldsymbol{\beta}, \boldsymbol{\beta}^t - \widetilde{\boldsymbol{\beta}}^t \rangle + \lambda_{\boldsymbol{\beta}}(\|\boldsymbol{\beta}^t - \widetilde{\boldsymbol{\beta}}^t\|_1) \\
= \;& g_{\boldsymbol{\Theta}}(\boldsymbol{\Theta}^{t-1}, \boldsymbol{\beta}^t, R^{t-1}) + g_{\boldsymbol{\beta}}(\boldsymbol{\Theta}^{t-1}, \boldsymbol{\beta}^t, Q^t), \qquad (39)
\end{aligned}
$$

where recall that $g_{\boldsymbol{\beta}}(\boldsymbol{\Theta}^{t-1}, \boldsymbol{\beta}^t, Q^t) = \langle \partial \mathcal{L}(\boldsymbol{\Theta}^{t-1}, \boldsymbol{\beta}^t)/\partial \boldsymbol{\beta}, \boldsymbol{\beta}^t - \widetilde{\boldsymbol{\beta}}^t \rangle + \lambda_{\boldsymbol{\beta}}(\|\boldsymbol{\beta}^t - \widetilde{\boldsymbol{\beta}}^t\|_1)$ and $g_{\boldsymbol{\Theta}}(\boldsymbol{\Theta}^{t-1}, \boldsymbol{\beta}^t, R^{t-1}) = \langle \mathcal{L}(\boldsymbol{\Theta}^{t-1}, \boldsymbol{\beta}^t)/\partial \boldsymbol{\Theta}, \boldsymbol{\Theta}^{t-1} - \widetilde{\boldsymbol{\Theta}}^{t-1} \rangle + \lambda_{\boldsymbol{\Theta}} \|\boldsymbol{\Theta}^{t-1} - \widetilde{\boldsymbol{\Theta}}^{t-1}\|_*$. The second inequality holds because $F(\boldsymbol{\Theta}, \boldsymbol{\beta})$ is minimized at $\widehat{\boldsymbol{\Theta}}$ and $\widehat{\boldsymbol{\beta}}$, and hence $\|\widehat{\boldsymbol{\beta}}\|_1 \le F(\widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\beta}})/\lambda_{\boldsymbol{\beta}} \le F(\boldsymbol{\Theta}^t, \boldsymbol{\beta}^t)/\lambda_{\boldsymbol{\beta}} = Q^t$ and $\|\widehat{\boldsymbol{\Theta}}\|_* \le F(\widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\beta}})/\lambda_{\boldsymbol{\Theta}} \le F(\boldsymbol{\Theta}^{t-1}, \boldsymbol{\beta}^t)/\lambda_{\boldsymbol{\beta}} \le R^{t-1}$. Furthermore,

$$
\begin{aligned}
g_{\boldsymbol{\beta}}(\boldsymbol{\Theta}^{t-1}, \boldsymbol{\beta}^t, Q^t) &= \langle \partial \mathcal{L}(\boldsymbol{\Theta}^{t-1}, \boldsymbol{\beta}^t)/\partial \boldsymbol{\beta}, \boldsymbol{\beta}^t - \widetilde{\boldsymbol{\beta}}^t \rangle + \lambda_{\boldsymbol{\beta}}(\|\boldsymbol{\beta}^t - \widetilde{\boldsymbol{\beta}}^t\|_1) \\
&= \langle \partial \mathcal{L}(\boldsymbol{\Theta}^t, \boldsymbol{\beta}^t)/\partial \boldsymbol{\beta}, \boldsymbol{\beta}^t - \widetilde{\boldsymbol{\beta}}^t \rangle + \langle \partial \mathcal{L}(\boldsymbol{\Theta}^{t-1}, \boldsymbol{\beta}^t)/\partial \boldsymbol{\beta} - \partial \mathcal{L}(\boldsymbol{\Theta}^t, \boldsymbol{\beta}^t)/\partial \boldsymbol{\beta}, \boldsymbol{\beta}^t - \widetilde{\boldsymbol{\beta}}^t \rangle \\
&\quad + \lambda_{\boldsymbol{\beta}}(\|\boldsymbol{\beta}^t - \widetilde{\boldsymbol{\beta}}^t\|_1) \\
&\le \langle \partial \mathcal{L}(\boldsymbol{\Theta}^t, \boldsymbol{\beta}^t)/\partial \boldsymbol{\beta}, \boldsymbol{\beta}^t - \widetilde{\boldsymbol{\beta}}^t \rangle + \lambda_{\boldsymbol{\beta}} \|\boldsymbol{\beta}^t - \widetilde{\boldsymbol{\beta}}^t\|_1 \\
&\quad + \|\partial \mathcal{L}(\boldsymbol{\Theta}^{t-1}, \boldsymbol{\beta}^t)/\partial \boldsymbol{\beta} - \partial \mathcal{L}(\boldsymbol{\Theta}^t, \boldsymbol{\beta}^t)/\partial \boldsymbol{\beta}\|_2 \|\boldsymbol{\beta}^t - \widetilde{\boldsymbol{\beta}}^t\|_2 \\
&\le g_{\boldsymbol{\beta}}(\boldsymbol{\Theta}^t, \boldsymbol{\beta}^t, Q^t) + 2\sigma_{\boldsymbol{\beta}\boldsymbol{\Theta}} Q^t \|\boldsymbol{\Theta}^{t-1} - \boldsymbol{\Theta}^t\|_F. \qquad (40)
\end{aligned}
$$

The last inequality holds by Remark 1 and $\|\boldsymbol{\beta}^t - \boldsymbol{\beta}\|_2 \le \|\boldsymbol{\beta}^t - \boldsymbol{\beta}\|_1 \le 2Q^t$. Furthermore, because $\boldsymbol{\Theta}^t$ is the minimizer of $\frac{1}{2}\|\boldsymbol{\Theta} - \boldsymbol{\Theta}^{t-1} + \eta_1 \frac{\partial \mathcal{L}(\boldsymbol{\Theta}^{t-1}, \boldsymbol{\beta}^t)}{\partial \boldsymbol{\Theta}}\|_F^2 + \eta_1 \lambda_{\boldsymbol{\Theta}} \|\boldsymbol{\Theta}\|_*$, we have

$$
\boldsymbol{\Theta}^t - \boldsymbol{\Theta}^{t-1} + \eta_1 \frac{\partial \mathcal{L}(\boldsymbol{\Theta}^{t-1}, \boldsymbol{\beta}^t)}{\partial \boldsymbol{\Theta}} + \eta_1 \lambda_{\boldsymbol{\Theta}} \partial \|\boldsymbol{\Theta}^t\|_*/\partial \boldsymbol{\Theta} = 0.
$$

Therefore,

$$
\begin{aligned}
& F(\boldsymbol{\Theta}^{t-1}, \boldsymbol{\beta}^t) - F(\boldsymbol{\Theta}^t, \boldsymbol{\beta}^t) \\
= \;& \langle \partial \mathcal{L}(\boldsymbol{\Theta}^t, \boldsymbol{\beta}^t)/\partial \boldsymbol{\Theta} + \lambda_{\boldsymbol{\Theta}} \partial \|\boldsymbol{\Theta}^t\|_*/\partial \boldsymbol{\Theta}, \boldsymbol{\Theta}^{t-1} - \boldsymbol{\Theta}^t \rangle \\
& + \{\mathrm{vec}(\boldsymbol{\Theta}^{t-1}) - \mathrm{vec}(\boldsymbol{\Theta}^t)\}^{\mathrm{T}} \frac{\partial^2 \mathcal{L}(\boldsymbol{\Theta}^*, \boldsymbol{\beta}^t)}{2\partial \mathrm{vec}(\boldsymbol{\Theta}) \mathrm{vec}(\boldsymbol{\Theta})^{\mathrm{T}}} \{\mathrm{vec}(\boldsymbol{\Theta}^{t-1}) - \mathrm{vec}(\boldsymbol{\Theta}^t)\} \\
= \;& \langle \partial \mathcal{L}(\boldsymbol{\Theta}^{t-1}, \boldsymbol{\beta}^t)/\partial \boldsymbol{\Theta} + \lambda_{\boldsymbol{\Theta}} \partial \|\boldsymbol{\Theta}^t\|_*/\partial \boldsymbol{\Theta}, \boldsymbol{\Theta}^{t-1} - \boldsymbol{\Theta}^t \rangle \\
& + \langle \partial \mathcal{L}(\boldsymbol{\Theta}^t, \boldsymbol{\beta}^t)/\partial \boldsymbol{\Theta} - \partial \mathcal{L}(\boldsymbol{\Theta}^{t-1}, \boldsymbol{\beta}^t)/\partial \boldsymbol{\Theta}, \boldsymbol{\Theta}^{t-1} - \boldsymbol{\Theta}^t \rangle \\
& + \{\mathrm{vec}(\boldsymbol{\Theta}^{t-1}) - \mathrm{vec}(\boldsymbol{\Theta}^t)\}^{\mathrm{T}} \frac{\partial^2 \mathcal{L}(\boldsymbol{\Theta}^*, \boldsymbol{\beta}^t)}{2\partial \mathrm{vec}(\boldsymbol{\Theta}) \mathrm{vec}(\boldsymbol{\Theta})^{\mathrm{T}}} \{\mathrm{vec}(\boldsymbol{\Theta}^{t-1}) - \mathrm{vec}(\boldsymbol{\Theta}^t)\} \\
= \;& \langle \partial \mathcal{L}(\boldsymbol{\Theta}^{t-1}, \boldsymbol{\beta}^t)/\partial \boldsymbol{\Theta} + \lambda_{\boldsymbol{\Theta}} \partial \|\boldsymbol{\Theta}^t\|_*/\partial \boldsymbol{\Theta}, \boldsymbol{\Theta}^{t-1} - \boldsymbol{\Theta}^t \rangle \\
& - \{\mathrm{vec}(\boldsymbol{\Theta}^{t-1}) - \mathrm{vec}(\boldsymbol{\Theta}^t)\}^{\mathrm{T}} \frac{\partial^2 \mathcal{L}(\boldsymbol{\Theta}^{**}, \boldsymbol{\beta}^t)}{\partial \mathrm{vec}(\boldsymbol{\Theta}) \mathrm{vec}(\boldsymbol{\Theta})^{\mathrm{T}}} \{\mathrm{vec}(\boldsymbol{\Theta}^{t-1}) - \mathrm{vec}(\boldsymbol{\Theta}^t)\} \\
& + \{\mathrm{vec}(\boldsymbol{\Theta}^{t-1}) - \mathrm{vec}(\boldsymbol{\Theta}^t)\}^{\mathrm{T}} \frac{\partial^2 \mathcal{L}(\boldsymbol{\Theta}^*, \boldsymbol{\beta}^t)}{2\partial \mathrm{vec}(\boldsymbol{\Theta}) \mathrm{vec}(\boldsymbol{\Theta})^{\mathrm{T}}} \{\mathrm{vec}(\boldsymbol{\Theta}^{t-1}) - \mathrm{vec}(\boldsymbol{\Theta}^t)\} \\
\ge \;& \langle \partial \mathcal{L}(\boldsymbol{\Theta}^{t-1}, \boldsymbol{\beta}^t)/\partial \boldsymbol{\Theta} + \lambda_{\boldsymbol{\Theta}} \partial \|\boldsymbol{\Theta}^t\|_*/\partial \boldsymbol{\Theta}, \boldsymbol{\Theta}^{t-1} - \boldsymbol{\Theta}^t \rangle \\
& - \{\mathrm{vec}(\boldsymbol{\Theta}^{t-1}) - \mathrm{vec}(\boldsymbol{\Theta}^t)\}^{\mathrm{T}} \frac{\partial^2 \mathcal{L}(\boldsymbol{\Theta}^{**}, \boldsymbol{\beta}^t)}{\partial \mathrm{vec}(\boldsymbol{\Theta}) \mathrm{vec}(\boldsymbol{\Theta})^{\mathrm{T}}} \{\mathrm{vec}(\boldsymbol{\Theta}^{t-1}) - \mathrm{vec}(\boldsymbol{\Theta}^t)\} \\
= \;& 1/\eta_1 \|\boldsymbol{\Theta}^{t-1} - \boldsymbol{\Theta}^t\|_F^2 - \{\mathrm{vec}(\boldsymbol{\Theta}^{t-1}) - \mathrm{vec}(\boldsymbol{\Theta}^t)\}^{\mathrm{T}} \frac{\partial^2 \mathcal{L}(\boldsymbol{\Theta}^{**}, \boldsymbol{\beta}^t)}{\partial \mathrm{vec}(\boldsymbol{\Theta}) \partial \mathrm{vec}(\boldsymbol{\Theta})^{\mathrm{T}}} \{\mathrm{vec}(\boldsymbol{\Theta}^{t-1}) - \mathrm{vec}(\boldsymbol{\Theta}^t)\} \\
\ge \;& (1/\eta_1 - \sigma_{\boldsymbol{\Theta}}) \|\boldsymbol{\Theta}^{t-1} - \boldsymbol{\Theta}^t\|_F^2, \qquad (41)
\end{aligned}
$$

where $\boldsymbol{\Theta}^*$ and $\boldsymbol{\Theta}^{**}$ are points on the line between $\boldsymbol{\Theta}^{t-1}$ and $\boldsymbol{\Theta}^t$. Hence

$$
\|\boldsymbol{\Theta}^{t-1} - \boldsymbol{\Theta}^t\|_F^2 \le \{F(\boldsymbol{\Theta}^{t-1}, \boldsymbol{\beta}^t) - F(\boldsymbol{\Theta}^t, \boldsymbol{\beta}^t)\}/(1/\eta_1 - \sigma_{\boldsymbol{\Theta}}).
$$

The last inequality holds by the fact that $\|\partial^2\mathcal{L}(\mathbf{\Theta}^{**},\boldsymbol{\beta}^t)/\partial\mathrm{vec}(\mathbf{\Theta})\partial\mathrm{vec}(\mathbf{\Theta})^{\mathrm{T}}\|_2 \leq \sigma_{\mathbf{\Theta}}$ as shown in Remark 1. Combining (40), (41), Lemmas A.22 and A.23, we obtain that with probability at least $1-2\max(p,mn)^{-1}-2(mn)^{-1}$,

$$\begin{aligned}
& \{g_{\mathbf{\Theta}}(\mathbf{\Theta}^{t-1},\boldsymbol{\beta}^t,R^{t-1})+g_{\boldsymbol{\beta}}(\mathbf{\Theta}^{t-1},\boldsymbol{\beta}^t,Q^t)\}^2 \\
\leq\ & 4\{g_{\mathbf{\Theta}}(\mathbf{\Theta}^{t-1},\boldsymbol{\beta}^t,R^{t-1})^2+g_{\boldsymbol{\beta}}(\mathbf{\Theta}^t,\boldsymbol{\beta}^t,Q^t)^2+4\sigma^2_{\boldsymbol{\beta}\mathbf{\Theta}}(Q^t)^2\|\mathbf{\Theta}^{t-1}-\mathbf{\Theta}^t\|_F^2\} \\
\leq\ & 4[2(2R^{t-1})^2/\eta_1\{F(\mathbf{\Theta}^{t-1},\boldsymbol{\beta}^t)-F(\mathbf{\Theta}^t,\boldsymbol{\beta}^t)\} \\
& +2(2Q^t)^2/\eta\{F(\mathbf{\Theta}^t,\boldsymbol{\beta}^t)-F(\mathbf{\Theta}^t,\boldsymbol{\beta}^{t+1})\} \\
& +4\sigma^2_{\boldsymbol{\beta}\mathbf{\Theta}}(Q^t)^2\{F(\mathbf{\Theta}^{t-1},\boldsymbol{\beta}^t)-F(\mathbf{\Theta}^t,\boldsymbol{\beta}^t)\}/(1/\eta_1-\sigma_{\mathbf{\Theta}})] \\
\leq\ & C(t)\{F(\mathbf{\Theta}^{t-1},\boldsymbol{\beta}^t)-F(\mathbf{\Theta}^t,\boldsymbol{\beta}^t)\}+C(t)\{F(\mathbf{\Theta}^t,\boldsymbol{\beta}^t)-F(\mathbf{\Theta}^t,\boldsymbol{\beta}^{t+1})\} \\
=\ & C(t)\{F(\mathbf{\Theta}^{t-1},\boldsymbol{\beta}^t)-F(\mathbf{\Theta}^t,\boldsymbol{\beta}^{t+1})\},
\end{aligned}$$

where $C(t)=\max\{32(R^{t-1})^2/\eta_1+16\sigma^2_{\boldsymbol{\beta}\mathbf{\Theta}}(Q^t)^2/(1/\eta_1-\sigma_{\mathbf{\Theta}}),32(Q^t)^2/\eta\}$. Combining with (39), we have

$$\{F(\mathbf{\Theta}^{t-1},\boldsymbol{\beta}^t)-F(\widehat{\mathbf{\Theta}},\widehat{\boldsymbol{\beta}})\}^2 \leq C(t)\{F(\mathbf{\Theta}^{t-1},\boldsymbol{\beta}^t)-F(\widehat{\mathbf{\Theta}},\widehat{\boldsymbol{\beta}})-F(\mathbf{\Theta}^t,\boldsymbol{\beta}^{t+1})+F(\widehat{\mathbf{\Theta}},\widehat{\boldsymbol{\beta}})\}$$

with probability at least $1-2\max(p,mn)^{-1}-2(mn)^{-1}-2(mn+p)^{-1}$. Let $\Delta_t=F(\mathbf{\Theta}^{t-1},\boldsymbol{\beta}^t)-F(\widehat{\mathbf{\Theta}},\widehat{\boldsymbol{\beta}})$, the above inequality can be written as

$$\Delta_{t+1}\leq\Delta_t-\frac{1}{C(t)}(\Delta_t)^2=\Delta_t\left\{1-\frac{\Delta_t}{C(t)}\right\}\leq\Delta_t\left\{1+\frac{\Delta_t}{C(t)}\right\}^{-1}.$$

Taking inverse, we get $\Delta_{t+1}^{-1}\geq\Delta_t^{-1}+C(t)^{-1}$, hence $\Delta_{t+1}^{-1}\geq\Delta_0^{-1}+\sum_{k=0}^t C(t)^{-1}$, which leads to

$$\Delta_{t+1}\leq\frac{1}{1/\{F(\mathbf{\Theta}^0,\boldsymbol{\beta}^0)-F(\widehat{\mathbf{\Theta}},\widehat{\boldsymbol{\beta}})\}+\sum_{k=0}^t 1/C(k)}.$$

Hence when

$$\sum_{t=0}^{T-1}1/C(t)\geq\left\{\frac{1}{\epsilon}-\frac{1}{F(\mathbf{\Theta}^0,\boldsymbol{\beta}^0)-F(\widehat{\mathbf{\Theta}},\widehat{\boldsymbol{\beta}})}\right\},$$

we have

$$F(\mathbf{\Theta}^{\mathrm{T}},\boldsymbol{\beta}^{\mathrm{T}})-F(\widehat{\mathbf{\Theta}},\widehat{\boldsymbol{\beta}})\leq\epsilon,$$

with probability at least $1-2\max(p,mn)^{-1}-2(mn)^{-1}-2(mn+p)^{-1}\geq 1-6(mn+p)^{-1}$. Hence $F(\mathbf{\Theta}^{\mathrm{T}},\boldsymbol{\beta}^{\mathrm{T}})$ is the ϵ-optimal solution for (1) with probability at least $1-6(mn+p)^{-1}$. This proves the result.

Appendix I. Proof of the Consistency of Pseudo-likelihood Estimator in Regression with Finite Dimensional Parameter

Theorem A.1 *Assume Condition (C1)–(C6) hold and $\boldsymbol{\beta}$ is finite dimensional. Furthermore, assume the conditional density of Y_i given $\mathbf{X}_i$ is $f(Y_i,\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{X}_i)$. Let $\widehat{\boldsymbol{\beta}}$ be the maximizer for the pseudo-likelihood*

$$\mathcal{L}(\boldsymbol{\beta})\equiv n^{-1}\sum_{i=1}^n\mathcal{L}_i(\boldsymbol{\beta}),$$

where $\mathcal{L}_i(\boldsymbol{\beta})=R_i\ell_i(\boldsymbol{\beta})$,

$$\ell_i(\boldsymbol{\beta})=\left[\log\{f(Y_i,\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i)\}-\log\left\{\int f(Y_i,\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})g(\mathbf{X})d\mathbf{X}\right\}\right],$$

R_i is the missing indicator for the ith observation. Assume $\mathbf{X}_i$ and R_i are independent given Y_i. We have $\widehat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}_0$.

Proof: Everywhere in this proof, E denotes an expectation taken with respect to the original true distribution that governs the data generation process $f(Y, \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{X})g(\mathbf{X})$, for example, for any function $\mathbf{a}(\mathbf{X}, Y)$, $E\{\mathbf{a}(\mathbf{X}, Y)\} = \int \mathbf{a}(\mathbf{x}, y) f(y, \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x}) g(\mathbf{x}) d\mu(\mathbf{x}, y)$. Let $f_2(Y_i, \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i)$ be the derivative of $f(Y_i, \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i)$ with respect to $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i$, $S_2(Y_i, \mathbf{X}_i, \boldsymbol{\beta})$ be the derivative of $\log f(Y_i, \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i)$ with respect to $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i$, define

$$\begin{aligned}
\mathbf{S}(Y_i, \mathbf{X}_i, \boldsymbol{\beta}) &\equiv \frac{\partial \mathcal{L}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \\
&= n^{-1} \sum_{i=1}^{n} R_i \left\{ \frac{f_2(Y_i, \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i)}{f(Y_i, \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i)} \mathbf{X}_i - \frac{\int f_2(Y_i, \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}) g(\mathbf{X}) \mathbf{X} d\mathbf{X}}{\int f(Y_i, \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}) g(\mathbf{X}) d\mathbf{X}} \right\} \\
&= n^{-1} \sum_{i=1}^{n} R_i S_2(Y_i, \mathbf{X}_i, \boldsymbol{\beta}) \mathbf{X}_i - R_i E_{\boldsymbol{\beta}}\{S_2(Y_i, \mathbf{X}, \boldsymbol{\beta}) \mathbf{X} | Y_i\}.
\end{aligned}$$

Here, the notation $E_{\boldsymbol{\beta}}$ means an expectation is taken with respect to the distribution $f(Y, \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})g(\mathbf{X})$. It is clear that

$$\frac{\partial E\{\mathcal{L}(\boldsymbol{\beta})\}}{\partial \boldsymbol{\beta}}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} = E\left\{ \frac{\partial \mathcal{L}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} \right\} = E\{\mathbf{S}(Y_i, \mathbf{X}_i, \boldsymbol{\beta}_0)\} = 0. \tag{42}$$

Also,

$$\begin{aligned}
\frac{\partial^2 \mathcal{L}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^{\mathrm{T}}} &= n^{-1} \sum_{i=1}^{n} R_i \Big(\mathbf{H}(Y_i, \mathbf{X}_i, \boldsymbol{\beta}) - E_{\boldsymbol{\beta}}\{\mathbf{H}(Y_i, \mathbf{X}_i, \boldsymbol{\beta}) | Y_i\} \\
&\quad - E_{\boldsymbol{\beta}}\left\{ \mathbf{S}(Y_i, \mathbf{X}_i, \boldsymbol{\beta})^{\otimes 2} | Y_i \right\} + \left[E_{\boldsymbol{\beta}}\{\mathbf{S}(Y_i, \mathbf{X}_i, \boldsymbol{\beta}) | Y_i\} \right]^{\otimes 2} \Big),
\end{aligned} \tag{43}$$

where $\mathbf{H}(Y_i, \mathbf{X}_i, \boldsymbol{\beta})$ is the second partial derivative of $\log\{f(Y_i, \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i)\}$ with respect to $\boldsymbol{\beta}$. Now using the independence between $\mathbf{X}_i$ and R_i given Y_i, we have

$$\begin{aligned}
E\left\{ \frac{\partial^2 \mathcal{L}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^{\mathrm{T}}}|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} \right\} &= E\left(E\left[R_i \mathbf{H}(Y_i, \mathbf{X}_i, \boldsymbol{\beta}) | Y_i \right] - E(R_i|Y_i) E_{\boldsymbol{\beta}_0}\{\mathbf{H}(Y_i, \mathbf{X}_i, \boldsymbol{\beta}_0) | Y_i\} \right) \\
&\quad - E\left\{ E(R_i|Y_i) \left(E\left\{ \mathbf{S}(Y_i, \mathbf{X}_i, \boldsymbol{\beta}_0)^{\otimes 2} | Y_i \right\} - \left[E\{\mathbf{S}(Y_i, \mathbf{X}_i, \boldsymbol{\beta}_0) | Y_i\} \right]^{\otimes 2} \right) \right\} \\
&= -E\left\{ E(R_i|Y_i) \left(E\left\{ \mathbf{S}(Y_i, \mathbf{X}_i, \boldsymbol{\beta}_0)^{\otimes 2} | Y_i \right\} - \left[E\{\mathbf{S}(Y_i, \mathbf{X}_i, \boldsymbol{\beta}_0) | Y_i\} \right]^{\otimes 2} \right) \right\} \\
&= -E\left[E(R_i|Y_i) \mathrm{var}\left\{ \mathbf{S}(Y_i, \mathbf{X}_i, \boldsymbol{\beta}_0) | Y_i \right\} \right],
\end{aligned} \tag{44}$$

which is negative definite. Note that $E_{\boldsymbol{\beta}_0} = E$ according to our definition of the two notations. Now by Taylor expansion, let $\mathbf{r}$ be a vector with $\|\mathbf{r}\|_2 = 1$, assume $\mathbf{r}$ times the third derivative of $\mathcal{L}(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$ has bounded L_2 norm, for any $\epsilon \in (0, 1/2)$, we have

$$\begin{aligned}
& \mathcal{L}(\boldsymbol{\beta}_0 + \mathbf{r} n^{-1/2+\epsilon}) \\
= \; & \mathcal{L}(\boldsymbol{\beta}_0) + \frac{\partial \mathcal{L}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} \mathbf{r} n^{-1/2+\epsilon} + \frac{1}{2} \mathbf{r}^{\mathrm{T}} \frac{\partial^2 \mathcal{L}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^{\mathrm{T}}}|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} \mathbf{r} n^{-1+2\epsilon} + O_p(n^{-3/2+3\epsilon}) \\
= \; & \mathcal{L}(\boldsymbol{\beta}_0) + \left[E\left\{ \frac{\partial \mathcal{L}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} \right\} + O_p(n^{-1/2}) \right] \mathbf{r} n^{-1/2+\epsilon} \\
+ \; & \frac{1}{2} \mathbf{r}^{\mathrm{T}} \left[E\left\{ \frac{\partial^2 \mathcal{L}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^{\mathrm{T}}}|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} \right\} + o_p(1) \right] \mathbf{r} n^{-1+2\epsilon} + O_p(n^{-3/2+3\epsilon})
\end{aligned}$$

$$= \mathcal{L}(\boldsymbol{\beta}_0) + \frac{1}{2}\mathbf{r}^{\mathrm{T}} E\left\{\frac{\partial^2 \mathcal{L}(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^{\mathrm{T}}}|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}\right\}\mathbf{r}n^{-1+2\epsilon} + o_p(n^{-1+2\epsilon}), \tag{45}$$

where the second equality holds by the central limit theorem, the third equality holds by (42). Now by (44), $E\left\{\partial^2\mathcal{L}(\boldsymbol{\beta})/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^{\mathrm{T}}|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}\right\}$ is negative definite. Hence, $\mathcal{L}(\boldsymbol{\beta}_0) > \mathcal{L}(\boldsymbol{\beta}_0+\mathbf{r}/\sqrt{n})$ in probability when n is sufficiently large for any $\mathbf{r}$ with $\|\mathbf{r}\|_2 = 1$. Therefore, there is a maximizer for $\mathcal{L}(\boldsymbol{\beta})$ in the ball with center $\boldsymbol{\beta}_0$ and radius $n^{-/2+\epsilon}$. Let the maximizer be $\widehat{\boldsymbol{\beta}}$. Obviously $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| \leq n^{-/2+\epsilon}$ in probability, hence $\widehat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}_0$.

References

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

Afonso S Bandeira, Ramon Van Handel, et al. Sharp nonasymptotic bounds on the norm of random matrices with independent entries. *The Annals of Probability*, 44(4):2479–2506, 2016.

X. Bi, A. Qu, J. Wang, and X. Shen. A group-specific recommender system. *Journal of the American Statistical Association*, 112:1344–1353, 2016.

P. J. Bickel, Y. Ritov, A. B. Tsybakov, et al. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

T. Cai, T. T. Cai, and A. Zhang. Structured matrix completion with applications to genomic data integration. *Journal of the American Statistical Association*, 111:621–633, 2016.

T. T. Cai and W.-X. Zhou. Matrix completion via max-norm constrained optimization. *Electronic Journal of Statistics*, 10:1493–1525, 2016.

E. J. Candés and Y. Plan. Matrix completion with noise. *Proceedings of the IEEE*, 98: 925–936, 2010.

E. J. Candés and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9:717–772, 2009.

Kai-Yang Chiang, Cho-Jui Hsieh, and Inderjit S Dhillon. Matrix completion with noisy side information. In *Advances in Neural Information Processing Systems*, volume 28, page 3447–3455, 2015.

D. M. Christopher, R. Prabhakar, and S. Hinrich. *Introduction to Information Retrieval*. Cambridge University Press, New York, 2008.

J. M. Hernández-Lobato, N. Houlsby, and Z. Ghahramani. Probabilistic matrix factorization with non-random missing data. In *International Conference on Machine Learning*, pages 1512–1520. PMLR, 2014.

D. Hsu, S. M. Kakade, and T. Zhang. Robust matrix decomposition with sparse corruptions. *IEEE Transactions on Information Theory*, 57(11):7221–7234, 2011.

Y. Hu, Y. Koren, and C. Volinsky. Collaborative filtering for implicit feedback datasets. In *IEEE International Conference on Data Mining*, pages 263–272. IEEE, 2008.

O. Klopp. Noisy low-rankmatrix completion with general sampling distribution. *Bernoulli*, 20:282–303, 2014.

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

Michel Ledoux. *The Concentration of Measure Phenomenon*, volume 89. American Mathematical Soc., 2001.

D. Liang, L. Charlin, J. McInerney, and D. M. Blei. Modeling user exposure in recommendation. In *International Conference on World Wide Web*, pages 951–961, 2016.

Po-Ling Loh and Martin J Wainwright. High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. *The Annals of Statistics*, 40(3):1637–1664, 2012.

X. Mao, Wong R. K. W, and Chen S. X. Matrix completion under low-rank missing mechanism. *ArXiv*, abs/1812.07813, 2018.

X. Mao, Chen S. X., and Wong R. K. W. Matrix completion with covariate information. *Journal of the American Statistical Association*, 114:198–210, 2019.

B. M. Marlin and R. S. Zemel. Collaborative prediction and ranking with non-random missing data. In *ACM Conference on Recommender Systems*, pages 5–12, 2009.

R. Mazumder, T. Hastie, and R. Tibshirani. Spectral regularization algorithms for learning large incomplete matrices. *Journal of Machine Learning Research*, 11:2287–2322, 2010.

W. Miao and E. J. Tchetgen Tchetgen. On varieties of doubly robust estimators under missingness not at random with a shadow variable. *Biometrika*, 103(2):475–482, 2016.

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

R. Pan, Y. Zhou, B. Cao, N. N. Liu, R. Lukose, M. Scholz, and Q. Yang. One-class collaborative filtering. In *International Conference on Data Mining*, pages 502–511. IEEE, 2008.

Gilles Pisier. *The Volume of Convex Bodies and Banach Space Geometry*, volume 94. Cambridge University Press, 1999.

B. Recht. A simpler approach to matrix completion. *Journal of Machine Learning Research*, 12:3413–3430, 2011.

G. Robin, H. T. Wai, J. Josse, O. Klopp, and É. Moulines. Low-rank interaction with sparse additive effects model for large data frames. In *Advances in Neural Information Processing Systems*, pages 5501–5511, 2018.

G. Robin, O. Klopp, J. Josse, É. Moulines, and R. Tibshirani. Main effects and interactions in mixed and incomplete data frames. *Journal of the American Statistical Association*, 115(531):1292–1303, 2020.

A. Rohde and A. B. Tsybakov. Estimation of high-dimensional low-rank matrices. *The Annals of Statistics*, 39:887–930, 2011.

Nathan Srebro and Russ R Salakhutdinov. Collaborative filtering in a non-uniform world: Learning with the weighted trace norm. In *Advances in Neural Information Processing Systems*, volume 23, page 2056–2064, 2010.

H. Steck. Training and testing of recommender systems on data missing not at random. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 713–722, 2010.

S. A. Van De Geer, P. Bühlmann, et al. On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics*, 3:1360–1392, 2009.

A. Van Der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes*. Springer, 2000.

Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019. doi: 10.1017/9781108627771.

Miao Xu, Rong Jin, and Zhi-Hua Zhou. Speedup matrix completion with side information: Application to multi-label learning. In *Advances in Neural Information Processing Systems*, volume 26, page 2301–2309, 2013.

J. Zhao and J. Shao. Semiparametric pseudo-likelihoods in generalized linear models with nonignorable missing data. *Journal of the American Statistical Association*, 110(512): 1577–1590, 2015.

Y. Zhu, X. Shen, and C. Ye. Personalized prediction and sparsity pursuit in latent factor models. *Journal of the American Statistical Association*, 111:241–252, 2016.