Journal of Machine Learning Research 25 (2024) 1-83 Submitted 4/23; Revised 3/24; Published 8/24

Uncertainty Quantification of MLE for Entity Ranking with Covariates

Jianqing Fan JQFAN@PRINCETON.EDU
Jikai Hou JIKAIH@PRINCETON.EDU
Department of Operations Research and Financial Engineering
Princeton University
Princeton, NJ, United States
Mengxin Yu MENGXINY@WHARTON.UPENN.EDU
Department of Statistics and Data Science, the Wharton School
University of Pennsylvania
Philadelphia, PA, United States

Editor: Ji Zhu

Abstract

We study statistical estimation and inference for the ranking problems based on pairwise comparisons with additional covariate information. In specific, in this paper, we study a Covariate-Assisted Ranking Estimation (CARE) model in a systematic way, that extends the well-known Bradley-Terry-Luce (BTL) model by incorporating the covariate information. We impose natural identifiability conditions, derive the statistical rates for the MLE under a sparse comparison graph, and obtain its asymptotic distribution. Moreover, we validate our theoretical results through large-scale numerical studies.

Keywords: High-Dimensional Inference, Entity ranking, Ranking with covariates, Uncertainty quantification, Maximum likelihood estimator.

1 Introduction

Ranking plays an essential role in many real-world applications. For example, it is crucial in individual choice (Luce, 2012), psychology (Thurstone, 1927, 2017), recommendation systems (Baltrunas et al., 2010; Li et al., 2019), and many others. The ranked items such as sports teams (Massey, 1997; Turner and Firth, 2012), scientific journals (Stigler, 1994), web pages (Dwork et al., 2001), election candidates (Plackett, 1975), or even movies (Harper and Konstan, 2015) will not only illustrate their qualities but also affect people's future choices. Thus, the ranking problem has been extensively studied in statistics, machine learning, operations research, etc.; see, for example, (Hunter, 2004; Richardson et al., 2006; Jang et al., 2018; Chen et al., 2019, 2022b,a; Liu et al., 2022) for more details.

Among various models for the ranking problem, the most well-known one is the Bradley-Terry-Luce (BTL) model (Bradley and Terry, 1952; Luce, 2012), which assumes the existence of scores $\{\theta_i^*\}_{i=1}^n$ of n compared items such that the preference between item i and

item j is given by

$$\mathbb{P}(\text{item i is preferred over j}) = e^{\theta_i^*}/(e^{\theta_i^*} + e^{\theta_j^*}), \qquad \text{for } (i,j) \in [n] \times [n].$$

The underlying assumption of this BTL model is the scores of compared items are fixed and do not explicitly use their attributes. However, in many real-world applications, covariate information often exists and this heterogeneity needs to be incorporated. For example, US News and Times Higher Education consider many characteristics of universities, such as international research reputation, teaching quality, the ratio between students and professors, and citations to conduct global university rankings. In addition, in NBA basketball competitions, the final rank of a team is also affected by its underlying attributes, such as the ability to defend, make a three-point shot, etc.

Thus, a crucial question still remains open:

> "*Can one design a provably efficient mechanism for ranking by incorporating features of compared items and conduct associated high-dimensional statistical inference?*"

To this end, we follow the idea from related literature Turner and Firth (2012); Li et al. (2022), by incorporating feature information of items into the BTL model, and call the model as the Covariate-Assisted Ranking Estimation (CARE) model. Specifically, we address covariate heterogeneity by assuming the underlying score (ability) of the i-th item is given by $\alpha_i^* + \boldsymbol{x}_i^\top \boldsymbol{\beta}^*$, where $\boldsymbol{x}_i^\top \boldsymbol{\beta}^*$ captures the covariate effect and α_i^* is the intrinsic score that cannot be explained by the covariate. In this case, the outcome of pairwise comparison is modeled as

$$\mathbb{P}(\text{item i is preferred over j}) = \frac{e^{\alpha_i^* + \boldsymbol{x}_i^\top \boldsymbol{\beta}^*}}{e^{\alpha_i^* + \boldsymbol{x}_i^\top \boldsymbol{\beta}^*} + e^{\alpha_j^* + \boldsymbol{x}_j^\top \boldsymbol{\beta}^*}}.$$

We do not assume that all pairs are compared. Rather, each pair is selected at random for comparison. In specific, we let the underlying comparison graph be the Erdős-Rényi random graph with edge probability p. In addition, once a pair is selected, they are compared L times. In this work, we consider the fixed design in the sense that the randomness only comes from results of comparisons.

There are several challenges in studying statistical inference for our CARE model. First, our model incorporates feature information into the original BTL model, not only the underlying scores $\{\alpha_i^*\}_{i=1}^n$ but also $\boldsymbol{\beta}^*$ shall be estimated and analyzed in a novel way. This also gives rise to the issue of identifiability. Second, given consistent estimators, it remains open to quantify these key components' uncertainty. Most existing work focuses more on deriving statistical rates of covergence for those underlying scores via various estimators in the BTL model to achieve specific rank recoveries such as top-K and partial recovery (Chen et al., 2019, 2022b). There are few results established for the inference of the BTL model (Simons and Yao, 1999; Han et al., 2020; Gao et al., 2021; Liu et al., 2022), letting alone the uncertainty quantification for the more general BTL model with covariates (CARE model).

In our work, we resolve the first challenge by designing a novel constrained maximum likelihood estimator (MLE) $(\widehat{\boldsymbol{\alpha}}_M, \widehat{\boldsymbol{\beta}}_M)$ which efficiently estimates the underlying scores $\{\alpha_i^*\}_{i=1}^n$ and $\boldsymbol{\beta}^*$. With some proper initialization, the MLE can be solved by simply running the projected gradient descent algorithm. By leveraging the 'leave-one-out' technique (Chen et al., 2019), we prove that the statistical rate of convergence of the MLE of the intrinsic scores $\{\alpha_i^*\}_{i=1}^n$ and overall scores $\{\alpha_i^* + \mathbf{x}_i^\top \boldsymbol{\beta}^*\}_{i=1}^n$ in ℓ_∞-norm are of order $\mathcal{O}(\sqrt{\log n/npL})$, in which d is the dimension of the observed covariates. These statistical rates reduce to the standard minimax rates for estimating the BTL model when no covariate exists, (Chen et al., 2019).

To take on the second challenge, namely, depicting the asymptotic distribution of the MLE, we first approximate the MLE by the minimizer of the quadratic approximation of our joint likelihood function, whose uncertainty is easier to depict. The critical difficulty lies in quantifying this approximation error. To tackle this issue, we then utilize the 'leave-one-out' technique and derive *novel proofs*, which is valid under the minimal sample complexity up to logarithm terms. In a more specific BTL model, the seminal works by Gao et al. (2021) and Fan et al. (2022) (when considering pairwise comparison) leverage the minimizers of the more restricted diagnonal quadratic approximations of their marginal likelihoods to approximate the MLE. They capture the approximation errors based on a 'leave-two-out' technique. In contrast, in this work, we utilize the minimizer of the quadratic approximation of the joint likelihood to approximate the MLE, and achieve a tighter approximation error than Gao et al. (2021); Fan et al. (2022).

Finally, we conduct numerical experiments to corroborate our theory. The performance of the model is also convincingly illustrated by analysis of the pokemon competition data. From the perspective of stock selection and return prediction, our proposed covariate-assisted BTL model (CARE) outperforms the original BTL model in many aspects.

To summarize, the contributions of this work are of multiple folds. First, we study a Covariate-Assisted Ranking Estimation (CARE) model in a systematic way that extends the well-known Bradley-Terry-Luce (BTL) model by incorporating the covariate information. Specifically, we derive ℓ_∞- and ℓ_2- statistical rates for the MLE of $\{\alpha_i^*\}_{i=1}^n$ and $\boldsymbol{\beta}^*$, respectively. Moreover, we also conduct uncertainty quantification for our MLE, where we improve the approximation errors given in existing works and derive more general asymptotic results. Furthermore, our results hold even on the sparse comparison graph, i.e. the probability of pairwise comparison $p \asymp 1/n$ up to logarithm terms, with minimal sample complexity. Finally, we illustrate our methods via large-scale numerical studies on synthetic and real data. Numerical results lends further support of our proposed CARE model over the original BTL model.

1.1 Prior Arts

Ranking problems based on pairwise comparison for parametric and non-parametric models have received much attention. For the BTL model, Hunter (2004) studies its variants and

establishes theoretical properties using a minorization-maximization algorithm. Chen and Suh (2015) use a two-step method to study the BTL model which is provably optimal in terms of sample complexity. Jang et al. (2016) leverage the spectral method to recover the top-K items with only minimal samples. In addition, Negahban et al. (2012) propose an iterative rank aggregation algorithm named Rank Centrality to recover the underlying scores of the BTL model in optimal ℓ_2- statistical rate. In the sequel, Chen et al. (2019) derive both ℓ_2- and ℓ_∞- optimal statistical rates of those underlying scores and prove that the regularized MLE and spectral method are both optimal for recovering top-K items when the conditional number is a constant. Furthermore, Chen et al. (2022b) prove that for partial recovery, MLE is optimal but the spectral method is not when we have a general conditional number. It is worth noting that the aforementioned prior arts mainly focus on studying the parametric BTL model. There is also a series of works that studies specific non-parametric variants of the BTL model. For instance, Shah and Wainwright (2017) develop a counting-based algorithm to recover top-K ranked items under the nonparametric stochastically transitive model. For more details on the non-parametric comparison models, see Shah et al. (2016); Shah and Wainwright (2017); Chen et al. (2017); Pananjady et al. (2017) and the references therein.

Going beyond the pairwise comparison, there also exist other works which study ranking problems using M-way comparisons ($M \geq 2$). The first well-known model is the Plackett-Luce model and its variants (Plackett, 1975; Guiver and Snelson, 2009; Cheng et al., 2010; Hajek et al., 2014; Maystre and Grossglauser, 2015; Szörényi et al., 2015; Jang et al., 2018; Fan et al., 2022). For instance, a closely related work is Jang et al. (2018), who study the Plackett-Luce model under a uniform hyper-graph. They divide M-way compared data into pairs and utilize the spectral method to derive the ℓ_∞- statistical rate of underlying scores. They further provide a lower bound for sample complexity to recover top-K items in the Plackett-Luce model. Another well-known model is the Thurstone model (Thurstone, 1927), which admits the Plackett-Luce model as a particular case; see Thurstone (1927); Guiver and Snelson (2009); Hajek et al. (2014); Vojnovic and Yun (2016); Jin et al. (2020) for more details.

The aforementioned literature mainly focuses on non-asymptotic statistical consistency results for the underlying scores of compared items through various ranking frameworks. However, the limiting distributional results for ranking models still remain highly under-explored. There are several results on the asymptotic distributions for the ranking scores in the BTL model. For instance, Simons and Yao (1999) derive the asymptotic normality of the MLE of the BTL model in the scenario where all pairs of comparison are fully conducted (i.e., $p = 1$). Han et al. (2020) further extend the results to the regime where the comparison graph (Erdős-Rényi random graph) is dense but not fully connected, i.e., $p \gtrsim n^{-1/10}$. In addition, recently, Liu et al. (2022) propose a Lagrangian debiasing method to derive asymptotic distribution for ranking scores, where they allow sparse comparison graph $p \asymp 1/n$ but require comparison times L to be larger than n^2. Moreover, Gao et al. (2021) utilize a 'leave-two-out' trick to derive asymptotic distributions for ranking scores

with optimal sample complexity in the regime where the comparison graph is sparse, i.e., $p \asymp 1/n$ up to logarithm terms.

All aforementioned models and methods mainly study the estimation and uncertainty quantification for ranking models without considering any individual feature information. Yet, covariate data exist in most applications, and this results in additional challenges in technical derivations and computation. Although there exists some other literature that also considers incorporating covariate information into the ranking problems (Guo et al., 2018; Schäfer and Hüllermeier, 2018; Zhao et al., 2022; Chau et al., 2023; Finch, 2022), their motivations, model settings, methodology, and theoretical contributions are different from us. In specific, most of them assume the underlying scores of all compared items are fully explained by covariates without studying the effects of individual intrinsic scores (i.e., no α_i^*). In addition, we allow the comparisons to be realized through (sparse) comparison graphs, which take on extra challenges. Moreover, in terms of the theoretical contribution, most of them only establish the ℓ_2- statistical rates for estimating β^* whereas we not only obtain ℓ_∞-and ℓ_2- statistical rates for estimators of $\{\alpha_i\}_{i=1}^n$ and β^* but quantify their uncertainty as well. In addition, even though our CARE model is related to Turner and Firth (2012); Li et al. (2022), we propose a systematic model estimation and inference framework. In contrast, these previous works only formulate ranking with covariates intuitively and do not discuss methodological implementations or theoretical guarantees.

Therefore, this paper takes up this challenge by presenting a systematic framework for model estimation and uncertainty quantification of our CARE model over a random comparison graph. Notably, this framework admits all previous advancements made on the BTL model, which do not incorporate covariate information as special cases.

1.2 Notation

We introduce some useful notations before proceeding. We denote by $[M] = \{1, 2, \ldots, M\}$ for any positive integer M. For any vector $\mathbf{u}$ and $q \geq 0$, we use $\|\mathbf{u}\|_{\ell_q}$ to represent the vector ℓ_q norm of $\mathbf{u}$. In addition, the inner product $\langle \mathbf{u}, \mathbf{v} \rangle$ between any pair of vectors $\mathbf{u}$ and $\mathbf{v}$ is defined as the Euclidean inner product $\mathbf{u}^\top \mathbf{v}$. For vector $\mathbf{u} \in \mathbb{R}^m$ and index $i \in [m]$, we denote by $\mathbf{u}_{-i}$ the vector we get by deleting the i-th element in $\mathbf{u}$. For any given matrix $\mathbf{X} \in \mathbb{R}^{d_1 \times d_2}$, we use $\|\mathbf{X}\|$, $\|\mathbf{X}\|_F$, $\|\mathbf{X}\|_*$ and $\|\mathbf{X}\|_{2,\infty}$ to represent the operator norm, Frobenius norm, nuclear norm and two-to-infinity norm of matrix $\mathbf{X}$ respectively. Moreover, we use $\mathbf{X} \succcurlyeq 0$ or $\mathbf{X} \preccurlyeq 0$ to denote positive semidefinite or negative semidefinite of matrix $\mathbf{X}$. Moreover, we use the notation $a_n \lesssim b_n$ or $a_n = O(b_n)$ for non-negative sequences $\{a_n\}$ and $\{b_n\}$ if there exists a constant ν_1 such that $a_n \leq \nu_1 b_n$. We use the notation $a_n \gtrsim b_n$ for non-negative sequences $\{a_n\}$ and $\{b_n\}$ if there is a constant ν_2 such that $a_n \geq \nu_2 b_n$. For simplicity, we define function $\phi(t) = e^t/(e^t + 1)$. We write $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $b_n \lesssim a_n$. For matrix $\boldsymbol{A}$, we denote by $\boldsymbol{A}^+$ its pseudoinverse (Banerjee, 1973).

1.3 Roadmap

The remaining paper is organized as follows. We describe the problem formulation for our BTL model with covariate and derive the corresponding maximum likelihood estimators for its involved parameters in §2. In §3, we establish the statistical estimation results for our MLE. In §4, we further conduct uncertainty quantification for the obtained MLE. In §5, we corroborate our theoretical results by conducting large-scale numerical studies via both synthetic and real data set.

2 Model Formulation

In this section, we introduce the Covariate-Assisted Ranking Estimation (CARE) model which incorporates covariate information into the BTL model. In the traditional BTL model (Bradley and Terry, 1952; Luce, 2012), it is assumed that each item $i \in [n]$ has a latent score θ_i^* and the outcomes of comparisons are modeled as the realizations of the Bernoulli trials:

$$\mathbb{P}\{\text{item } j \text{ is preferred over item } i\} = \frac{e^{\theta_j^*}}{e^{\theta_j^*} + e^{\theta_i^*}}, \quad \forall 1 \le i \ne j \le n. \tag{1}$$

It is worth mentioning that the function $\exp(\cdot)$ in (1) can be replaced by any increasing differentiable functions.

In many applications, one observes individual features $\boldsymbol{x}_i \in \mathbb{R}^d$ and would like to incorporate them for conducting more accurate ranking. As an extension of the parameterization $\exp(\theta_i^*)$ (Chen et al., 2019, 2022b), we model the scores $\exp(\theta_i^*)$ as $\exp\left(\alpha_i^* + \boldsymbol{x}_i^\top \boldsymbol{\beta}^*\right)$ for $1 \le i \le n$. The linear term $\boldsymbol{x}_i^\top \boldsymbol{\beta}^*$ captures the part of the scores explained by the variables $\boldsymbol{x}_i$ and α_i^* represents the intrinsic score that can not be explained by the covariate $\boldsymbol{x}_i$. This leads to modeling the outcomes of comparisons as the Bernoulli trials with probabilities

$$\mathbb{P}\{\text{item } j \text{ is preferred over item } i\} = \frac{e^{\alpha_j^* + \boldsymbol{x}_j^\top \boldsymbol{\beta}^*}}{e^{\alpha_i^* + \boldsymbol{x}_i^\top \boldsymbol{\beta}^*} + e^{\alpha_j^* + \boldsymbol{x}_j^\top \boldsymbol{\beta}^*}}, \quad \forall 1 \le i \ne j \le n. \tag{2}$$

We call this model Covariate Assisted Ranking Estimation (CARE) model.

We do not assume that all pairs are compared, but only those in the comparison graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. Here $\mathcal{V} := \{1, 2, \ldots, n\}$ and $\mathcal{E}$ represent the collections of vertexes (n items) and edges, respectively. More specifically, $(i, j) \in \mathcal{E}$ if and only if item i and item j are compared. Throughout our paper, the comparison graph is assumed to follow the Erdős-Rényi random graph $\mathcal{G}_{n,p}$ (Erdos et al., 1960) where each edge appears independently with probability p. In short, items i and j with $(i, j) \in [n] \times [n]$ are compared at random with probability p.

In addition, for any $(i, j) \in \mathcal{E}$, we observe L independent and identically distributed realizations from the Bernoulli random variables

$$P(y_{i,j}^{(l)} = 1) = \frac{e^{\alpha_j^* + \boldsymbol{x}_j^\top \boldsymbol{\beta}^*}}{e^{\alpha_i^* + \boldsymbol{x}_i^\top \boldsymbol{\beta}^*} + e^{\alpha_j^* + \boldsymbol{x}_j^\top \boldsymbol{\beta}^*}}.$$

Let $\widetilde{\boldsymbol{x}}_i = \left(\boldsymbol{e}_i^\top, \boldsymbol{x}_i^\top\right)^\top$ and $\widetilde{\boldsymbol{\beta}} = \left(\boldsymbol{\alpha}^\top, \boldsymbol{\beta}^\top\right)^\top$, where $\{\boldsymbol{e}_i\}_{i=1}^n$ stand for the canonical basis vectors in $\mathbb{R}^n$ and $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_n)^\top \in \mathbb{R}^n$. Then, the log-likelihood function conditioned on $\mathcal{G}$ is given by

$$L \cdot \sum_{(i,j)\in\mathcal{E}, i>j} \left\{ y_{j,i} \log \frac{e^{\widetilde{\boldsymbol{x}}_i^\top \widetilde{\boldsymbol{\beta}}}}{e^{\widetilde{\boldsymbol{x}}_i^\top \widetilde{\boldsymbol{\beta}}} + e^{\widetilde{\boldsymbol{x}}_j^\top \widetilde{\boldsymbol{\beta}}}} + (1 - y_{j,i}) \log \frac{e^{\widetilde{\boldsymbol{x}}_j^\top \widetilde{\boldsymbol{\beta}}}}{e^{\widetilde{\boldsymbol{x}}_i^\top \widetilde{\boldsymbol{\beta}}} + e^{\widetilde{\boldsymbol{x}}_j^\top \widetilde{\boldsymbol{\beta}}}} \right\}.$$

where $y_{i,j} = \frac{1}{L}\sum_{l=1}^L y_{i,j}^{(l)}$ is a sufficient statistic. The identifiability question arises naturally since we over-parametrized the problem. To remedy this issue, we restrict the parameter space of $\widetilde{\boldsymbol{\beta}}$ onto some constrained set Θ with a natural interpretation. In specific, we denote $\bar{\boldsymbol{x}}_i = [1, \boldsymbol{x}_i^\top]^\top, \forall i \in [n]$, let $\bar{\boldsymbol{X}} = [\bar{\boldsymbol{x}}_1, \cdots, \bar{\boldsymbol{x}}_n]^\top \in \mathbb{R}^{n\times(d+1)}$ and consider the constrained set $\Theta = \{(\boldsymbol{\alpha}, \boldsymbol{\beta}) : \bar{\boldsymbol{X}}^\top \boldsymbol{\alpha} = \mathbf{0}\}$. Throughout the paper, we assume that $\bar{\boldsymbol{X}}$ has rank $d+1$. Under these identifiability constraints, if the true parameter vector $\widetilde{\boldsymbol{\beta}}^* = \left(\alpha_1^*, \alpha_2^*, \ldots, \alpha_n^*, \boldsymbol{\beta}^{*\top}\right)^\top = \left(\boldsymbol{\alpha}^{*\top}, \boldsymbol{\beta}^{*\top}\right)^\top \in \Theta$, the identifiability condition implies that $\sum_{i=1}^n \alpha_i^* = 0$ and $\boldsymbol{X}^\top \boldsymbol{\alpha}^* = 0$ with $\mathbf{X} = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n]^\top \in \mathbb{R}^{n\times d}$. It admits clear interpretation: $\boldsymbol{X}\boldsymbol{\beta}^*$ represents the scores that be captured by the covariates, whereas the $\boldsymbol{\alpha}^* \in \mathbb{R}^n$ represents the residual scores (or equivalently intrinsic scores) that can not be explained by the involved features (i.e., it does not fall into the linear space spanned by the columns of covariates). Next, we prove the identifiable property of Θ rigoriously in Proposition 1.

Proposition 1. *CARE model Eq.* (2) *with parameter space* $\Theta = \{(\boldsymbol{\alpha}, \boldsymbol{\beta}) : \bar{\boldsymbol{X}}^\top \boldsymbol{\alpha} = \mathbf{0}\}$ *is identifiable.*

We denote by $\boldsymbol{Z} \in \mathbb{R}^{(n+d)\times(d+1)}$ a matrix by padding $\mathbf{0} \in \mathbb{R}^{d\times(d+1)}$ matrix to $\bar{\boldsymbol{X}}$, i.e. $\boldsymbol{Z}_{1:n,\cdot} = \bar{\boldsymbol{X}}$ and $\boldsymbol{Z}_{n+1:n+d,\cdot} = \mathbf{0}$. As a result, Θ can be also written as $\{\widetilde{\boldsymbol{\beta}} \in \mathbb{R}^{n+d} : \boldsymbol{Z}^\top \widetilde{\boldsymbol{\beta}} = \mathbf{0}\}$. Denote by $\mathcal{P} = \boldsymbol{I} - \boldsymbol{Z}(\boldsymbol{Z}^\top \boldsymbol{Z})^{-1}\boldsymbol{Z}^\top$ the projection matrix onto space Θ.

Given the aforementioned identifiable condition, we consider the following constrained maximum likelihood estimator (MLE)

$$\widetilde{\boldsymbol{\beta}}_M = \operatorname*{argmin}_{\widetilde{\boldsymbol{\beta}}\in\Theta} \mathcal{L}(\widetilde{\boldsymbol{\beta}}), \tag{3}$$

where

$$\mathcal{L}(\widetilde{\boldsymbol{\beta}}) := \sum_{(i,j)\in\mathcal{E}, i>j} \left\{ -y_{j,i} \left(\widetilde{\boldsymbol{x}}_i^\top \widetilde{\boldsymbol{\beta}} - \widetilde{\boldsymbol{x}}_j^\top \widetilde{\boldsymbol{\beta}} \right) + \log\left(1 + e^{\widetilde{\boldsymbol{x}}_i^\top \widetilde{\boldsymbol{\beta}} - \widetilde{\boldsymbol{x}}_j^\top \widetilde{\boldsymbol{\beta}}}\right) \right\}. \tag{4}$$

Note that when there is no covariate $\{\boldsymbol{x}_i\}_{i=1}^n$, we have $\Theta = \{\boldsymbol{\alpha} \in \mathbb{R}^n, \mathbf{1}^\top \boldsymbol{\alpha} = \mathbf{0}\}$ and the scores are identifiable up to a constant shift. Therefore, our formulation includes those studies of the BTL model without covariate information as special cases (Chen et al., 2019).

The inference question arises naturally if some covariates can explain the underlying scores, namely if some or all components of $\boldsymbol{\beta}^*$ are statistically significant. Similarly, one might ask if the covariates are adequate for determining the underlying scores by testing whether some or all components of $\boldsymbol{\alpha}^*$ are zero. In general, we would expect the variations

among the components of $\boldsymbol{\alpha}^*$ to be smaller than the original scores $\{\theta_i^*\}_{i=1}^n$. This enables us to improve data predictions by shrinking or regularizing the estimate of $\boldsymbol{\alpha}^*$.

In the following context, we rescale $\boldsymbol{x}_i$ to $\boldsymbol{x}_i/K$, where $K > 0$ is a positive number such that $\|\boldsymbol{x}_i\|_2 \leq \sqrt{(d+1)/n}$ for all $\boldsymbol{x}_i$ after the transformation. The likelihood function, prediction and the column space spanned by $\bar{\boldsymbol{X}}$ are not affected by the scaling but this normalization facilitates us with scaling issues in the technical derivations. Therefore, the content that follows will be based on the scaled data and parameters.

3 Rate of Convergence of Maximum Likelihood Estimator

In this section, we show the statistical consistency results for the maximum likelihood estimator $\widetilde{\boldsymbol{\beta}}_M$ in (3). Before proceeding to the main results, we begin by introducing several key assumptions on the design matrix. First, we assume the projected matrix $\mathcal{P}_{\bar{\boldsymbol{X}}} := \bar{\boldsymbol{X}}(\bar{\boldsymbol{X}}^\top \bar{\boldsymbol{X}})^{-1}\bar{\boldsymbol{X}}^\top$ satisfies the following incoherence condition.

Assumption 2. *[Incoherence Condition] Assume that there exists a positive constant c_0 such that*

$$\|\mathcal{P}_{\bar{\boldsymbol{X}}}\|_{2,\infty} = \|\bar{\boldsymbol{X}}(\bar{\boldsymbol{X}}^\top \bar{\boldsymbol{X}})^{-1}\bar{\boldsymbol{X}}^\top\|_{2,\infty} \leq c_0\sqrt{(d+1)/n}.$$

To demonstrate the rationality behind Assumption 2, we first note that the $\|\mathcal{P}_{\bar{\boldsymbol{X}}}\|_F^2 \leq d+1$. Therefore, a sufficient condition for this assumption to hold is when the rows of $\mathcal{P}_{\bar{\boldsymbol{X}}}$ are nearly balanced, with row sum of squares all of the order $(d+1)/n$ or smaller. When there does not exist the covariate (i.e. $\bar{\boldsymbol{X}} = \mathbf{1}$), we have $\mathcal{P}_{\bar{\boldsymbol{X}}} = \mathbf{1}\mathbf{1}^\top/n$. In this scenario, this assumption holds automatically with $c_0 = 1$. The following results of this paper are established under this incoherence condition.

We next introduce a key assumption on the covariates $\boldsymbol{x}_i$ which guarantees a well-behaved landscape of the loss function as well as good statistical properties of the MLE estimator. In specific, we put the following assumption on $\boldsymbol{\Sigma} = \sum_{i>j}(\widetilde{\boldsymbol{x}}_i - \widetilde{\boldsymbol{x}}_j)(\widetilde{\boldsymbol{x}}_i - \widetilde{\boldsymbol{x}}_j)^\top$.

Assumption 3. *Assume that there exist positive constants c_1 and c_2 such that*

$$c_2 n \leq \lambda_{min,\perp}(\boldsymbol{\Sigma}) \leq \|\boldsymbol{\Sigma}\| \leq c_1 n,$$

where $\|\boldsymbol{\Sigma}\|$ is the operator norm of $\boldsymbol{\Sigma}$ and

$$\lambda_{min,\perp}(\boldsymbol{\Sigma}) := \min\left\{\mu | \boldsymbol{z}^\top \boldsymbol{\Sigma}\boldsymbol{z} \geq \mu\|\boldsymbol{z}\|_2^2 \text{ for all } \boldsymbol{z} \in \Theta\right\}.$$

In this Assumption 3, we assume that $\boldsymbol{\Sigma}$ is well-behaved in all directions inside our parameter space Θ, namely, both of its largest and smallest eigenvalues are of order n. This assumption follows directly after we rescale the $\|\boldsymbol{x}_i\|_2$ such that $\|\boldsymbol{x}_i\|_2 \leq \sqrt{(d+1)/n}$ for all $\boldsymbol{x}_i, i \in [n]$. When there is no covariate (i.e., $d = 0$), then $\boldsymbol{\Sigma} = \sum_{i>j}(\mathbf{e}_i - \mathbf{e}_j)(\mathbf{e}_i - \mathbf{e}_j)^\top$ and Assumption 3 holds naturally with $c_1 = c_2 = 1$ (Chen et al., 2019).

We next introduce the condition number of this problem as

$$\kappa_1 := \frac{\max_i w_i^*}{\min_i w_i^*} = \exp\left(\max_{i,j\in[n]}\left(\alpha_i^* + \boldsymbol{x}_i^\top\boldsymbol{\beta}^* - \alpha_j^* - \boldsymbol{x}_j^\top\boldsymbol{\beta}^*\right)\right),$$

where $w_i^* = \exp(\alpha_i^* + \boldsymbol{x}_i^\top\boldsymbol{\beta}^*)$, which extends the condition number in Chen et al. (2019) when there does not exist covariate. With all aforementioned assumptions in hand, we next present our first main theorem on the statistical rates of convergence of the MLE $\widetilde{\boldsymbol{\beta}}_M$. Recall that we assume that the true parameter vector $\widetilde{\boldsymbol{\beta}}^* = \left(\alpha_1^*, \alpha_2^*, \ldots, \alpha_n^*, \boldsymbol{\beta}^{*\top}\right)^\top = \left(\boldsymbol{\alpha}^{*\top}, \boldsymbol{\beta}^{*\top}\right)^\top \in \Theta$, without loss of generality.

Theorem 4. *(Rate of convergence) Suppose $np > c_p \log n$ for some $c_p > 0$ and $d + 1 < n, (d+1)\log n \lesssim np$. Consider $L \le c_4 \cdot n^{c_5}$ for any absolute constants $c_4, c_5 > 0$. Let $\widetilde{\boldsymbol{\beta}}_M = (\widehat{\boldsymbol{\alpha}}_M^\top, \widehat{\boldsymbol{\beta}}_M^\top)^\top$ be the solution of the MLE given in (3). Then with probability at least $1 - O(n^{-6})$, we have*

$$\|\widehat{\boldsymbol{\alpha}}_M - \boldsymbol{\alpha}^*\|_\infty \lesssim \kappa_1^2\sqrt{\frac{(d+1)\log n}{npL}}; \qquad \left\|\widehat{\boldsymbol{\beta}}_M - \boldsymbol{\beta}^*\right\|_2 \lesssim \kappa_1\sqrt{\frac{\log n}{pL}};$$

$$\left\|\widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{\beta}}_M - \widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{\beta}}^*\right\|_\infty \lesssim \kappa_1^2\sqrt{\frac{(d+1)\log n}{npL}}; \qquad \frac{\left\|e^{\widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{\beta}}_M} - e^{\widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{\beta}}^*}\right\|_\infty}{\left\|e^{\widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{\beta}}^*}\right\|_\infty} \lesssim \kappa_1^2\sqrt{\frac{(d+1)\log n}{npL}},$$

where $\widetilde{\boldsymbol{X}} = [\widetilde{\boldsymbol{x}}_1, \widetilde{\boldsymbol{x}}_2, \ldots, \widetilde{\boldsymbol{x}}_n]^\top$.

Recall that we rescaled the covariates such that $\max_{i\in[n]}\|\boldsymbol{x}_i\|_2 \le \sqrt{(d+1)/n}$. This scaling has an impact on the definition of $\boldsymbol{\beta}^*$ and influences its rate. This explains why $\widehat{\boldsymbol{\beta}}_M$ converges slower and does not depends on d. However, when we view $\boldsymbol{x}_i^\top\boldsymbol{\beta}$ as an whole, the estimation rate is $\sqrt{(d+1)\log n/npL}$, and this does not impact on the estimation of the individual score, as shown in Theorem 4.

Remark 5. *Following a similar proof, it holds that*

$$\frac{\|\widehat{\boldsymbol{\alpha}}_M - \boldsymbol{\alpha}^*\|_2}{\|\boldsymbol{\alpha}^*\|_2} \lesssim \kappa_1^2\sqrt{\frac{(d+1)\log n}{npL}}, \qquad \frac{\left\|e^{\widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{\beta}}_M} - e^{\widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{\beta}}^*}\right\|_2}{\left\|e^{\widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{\beta}}^*}\right\|_2} \lesssim \kappa_1^2\sqrt{\frac{(d+1)\log n}{npL}},$$

which are the relative ℓ_2-statistical rates of the intrinsic scores $\alpha_i^*, i \in [n]$ and overall scores $\alpha_i^* + \boldsymbol{x}_i^\top\boldsymbol{\beta}^*, i \in [n]$, respectively. Combining this relative statistical rate in ℓ_2-norm with that in ℓ_∞-norm mentioned in Theorem 4, we conclude that the estimation errors of latent scores and overall scores spread out across all items.

Remark 6. *We note that we are further able to get rid of the factor $\sqrt{d+1}$ in the statistical rates of $\|\widehat{\boldsymbol{\alpha}}_M - \widehat{\boldsymbol{\alpha}}^*\|_\infty$ ($\|\widehat{\boldsymbol{\alpha}}_M - \widehat{\boldsymbol{\alpha}}^*\|_2$ follows directly). This involves analyzing the non-asymptotic expansion of the $\widehat{\boldsymbol{\alpha}}_M$, and the details will be discussed in the following section.*

4 Uncertainty Quantification of MLE

Most existing works on ranking mainly study the first-order asymptotic behavior of their estimators (Hunter, 2004; Chen and Suh, 2015; Jang et al., 2016; Shah and Wainwright, 2017; Chen et al., 2019). Deriving limiting distributional results in ranking models is important for uncertainty quantification, especially when covariate information is incorporated into the ranking problem, are not studied in detail.

This section devotes to understanding the sampling variability of the MLE $\widetilde{\boldsymbol{\beta}}_M$ under the CARE model. Directly studying the asymptotic behavior of $\widetilde{\boldsymbol{\beta}}_M$ is very challenging. To address this issue, we approximate $\widetilde{\boldsymbol{\beta}}_M$ by considering

$$\overline{\boldsymbol{\beta}} := \underset{\widetilde{\boldsymbol{\beta}}\in\Theta}{\operatorname{argmin}}\, \overline{\mathcal{L}}(\widetilde{\boldsymbol{\beta}}), \tag{5}$$

here $\overline{\mathcal{L}}(\widetilde{\boldsymbol{\beta}})$ is the quadratic expansion of the loss function $\mathcal{L}(\widetilde{\boldsymbol{\beta}})$ around $\widetilde{\boldsymbol{\beta}}^*$ given by

$$\overline{\mathcal{L}}(\widetilde{\boldsymbol{\beta}}) = \mathcal{L}(\widetilde{\boldsymbol{\beta}}^*) + \left(\widetilde{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}^*\right)^\top \nabla\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*) + \frac{1}{2}\left(\widetilde{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}^*\right)^\top \nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\left(\widetilde{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}^*\right). \tag{6}$$

According to this definition, $\overline{\boldsymbol{\beta}}$ can also be given by the following linear equations

$$\begin{cases} \mathcal{P}\nabla\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*) + \mathcal{P}\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\left(\overline{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}^*\right) = \mathbf{0}; \\ \mathcal{P}\overline{\boldsymbol{\beta}} = \overline{\boldsymbol{\beta}}. \end{cases}$$

Here $\mathcal{P}$ represents the linear projection onto space Θ. Observe that $\overline{\boldsymbol{\beta}}$ serves as a candidate approximator of $\widetilde{\boldsymbol{\beta}}_M$ whose uncertainty is easier to quantify according to Berry-Esseen theorem (Berry, 1941; Esseen, 1942). It is not a statistical estimator but an auxiliary random variable that we used for the technical proof. It is worth mentioning that when there is no covariate, our linear expansion reduces to $\overline{\boldsymbol{\beta}} \in \mathbb{R}^n = [\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)]^+\nabla\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)$, which is equal to the expansion in Gao et al. (2021) up to the off-diagonal terms in $\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)$.

The critical difficulty falls in proving that the difference $\Delta\widetilde{\boldsymbol{\beta}} := \widetilde{\boldsymbol{\beta}}_M - \overline{\boldsymbol{\beta}}$ is negligible compared to $\overline{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}^*$ under certain conditions. To accommodate this, we derive novel proofs by leveraging the 'leave-one-out' (Chen et al., 2019, 2021) technique to control the approximation error $\Delta\widetilde{\boldsymbol{\beta}} := \widetilde{\boldsymbol{\beta}}_M - \overline{\boldsymbol{\beta}}$ in ℓ_2-norm and $\Delta\boldsymbol{\alpha} := \Delta\widetilde{\boldsymbol{\beta}}_{1:n}$ in ℓ_∞-norm. The upper bounds are summarized in the following Theorem 7.

Theorem 7. *(Approximation error) Under the assumptions of Theorem 4, if* $\kappa_1^2\sqrt{(d+1)\log n/npL} \le c$ *and* $\kappa_1^2\left(\sqrt{(d+1)/np} + \log n/np\right) \le c$ *for some fixed constant* $c > 0$*, we have*

$$\widehat{\boldsymbol{\alpha}}_M = \overline{\boldsymbol{\alpha}} + \Delta\boldsymbol{\alpha} \text{ and } \widetilde{\boldsymbol{\beta}}_M = \overline{\boldsymbol{\beta}} + \Delta\widetilde{\boldsymbol{\beta}},$$

and with probability at least $1 - O(n^{-5})$*, it holds that*

$$\left\|\Delta\widetilde{\boldsymbol{\beta}}\right\|_2 \lesssim \kappa_1^4\frac{(d+1)^{0.5}\log n}{\sqrt{n}pL},$$
$$\left\|\Delta\boldsymbol{\alpha}\right\|_\infty \lesssim \kappa_1^6\frac{(d+1)\log n}{npL} + \frac{\kappa_1^4}{np}\sqrt{\frac{(d+1)\log n}{L}}\left(\sqrt{d+1} + \frac{\log n}{\sqrt{np}}\right).$$

Remark 8. *The assumptions* $\kappa_1^2\sqrt{(d+1)\log n/npL} = \mathcal{O}(1)$ *and* $\kappa_1^2\left(\sqrt{(d+1)/np} + \log n/np\right) \leq c$ *for some fixed constant* $c > 0$ *are mild. When* d *and* κ_1 *are bounded, they hold when* $p \gtrsim \log n/n$. *This matches the lower bound of the sampling probability* p *to ensure the connectivity of the Erdős-Rényi random graph, which is a necessary requirement for item ranking. Besides,* $\kappa_1^2\sqrt{(d+1)\log n/npL} = \mathcal{O}(1)$ *is also required by the consistency results of our estimator according to Theorem 4.*

Remark 9. *Given the non-asymptotic expansion, and approximation error* $\Delta\boldsymbol{\alpha} := \widehat{\boldsymbol{\alpha}}_M - \bar{\boldsymbol{\beta}}_{1:n}$ *presented in Theorem 7, we are able to achieve a tighter* ℓ_∞ *statistical error bound for* $\|\widehat{\boldsymbol{\alpha}}_M - \boldsymbol{\alpha}^*\|_\infty$. *Specifically, Under the assumptions of Theorem 4 and 7, as long as*

$$\kappa_1^5(d+1)\sqrt{\frac{\log n}{npL}} + \kappa_1^3\sqrt{\frac{(d+1)}{np}}\left(\sqrt{d+1} + \frac{\log n}{\sqrt{np}}\right) \lesssim 1.$$

Then with probability at least $1 - O(n^{-5})$ *we have*

$$\|\widehat{\boldsymbol{\alpha}}_M - \boldsymbol{\alpha}^*\|_\infty \lesssim \kappa_1\sqrt{\frac{\log n}{npL}}.$$

Next, we utilize Berry-Essen theorem (Berry, 1941; Esseen, 1942) to derive the asymptotic distribution of the linear conbinaitons of $\widetilde{\boldsymbol{\beta}}_M$, respectively. Since it holds for any linear combinations, the result applies to any finite dimensional distribution of $\widetilde{\boldsymbol{\beta}}_M$.

Theorem 10. *(Asymptotic normality of MLE) Given* $\boldsymbol{c} \in \mathbb{R}^{n+d}$, *let* $\overline{\boldsymbol{c}} = \mathcal{P}\boldsymbol{c}$ *be the projection of* $\boldsymbol{c}$ *onto linear space* Θ. *Under the assumptions of Theorem 7, we have the following decomposition*

$$\sqrt{L}\left(\boldsymbol{c}^\top\widetilde{\boldsymbol{\beta}}_M - \boldsymbol{c}^\top\widetilde{\boldsymbol{\beta}}^*\right) = \sqrt{L}\left(\boldsymbol{c}^\top\widetilde{\boldsymbol{\beta}}_M - \boldsymbol{c}^\top\overline{\boldsymbol{\beta}}\right) + \sqrt{L}\left(\boldsymbol{c}^\top\overline{\boldsymbol{\beta}} - \boldsymbol{c}^\top\widetilde{\boldsymbol{\beta}}^*\right),$$

where

$$\left|\frac{\sqrt{L}\left(\boldsymbol{c}^\top\widetilde{\boldsymbol{\beta}}_M - \boldsymbol{c}^\top\overline{\boldsymbol{\beta}}\right)}{\sqrt{\overline{\boldsymbol{c}}^\top\left[\mathcal{P}\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\mathcal{P}\right]^+\overline{\boldsymbol{c}}}}\right| \lesssim \left[\kappa_1^6\frac{(d+1)\log n}{\sqrt{npL}} + \kappa_1^4\sqrt{\frac{(d+1)\log n}{np}}\left(\sqrt{d+1} + \frac{\log n}{\sqrt{np}}\right)\right]\frac{\|\boldsymbol{c}_{1:n}\|_1}{\|\overline{\boldsymbol{c}}\|_2}$$
$$+ \kappa_1^4\frac{(d+1)^{0.5}\log n}{\sqrt{pL}}\frac{\|\boldsymbol{c}_{n+1:n+d}\|_2}{\|\overline{\boldsymbol{c}}\|_2},$$

with probability exceeding $1 - O(n^{-5})$ *(randomness comes from* $\mathcal{G}$ *and* $y_{i,j}^{(l)}$*) and*

$$\sup_{x\in\mathbb{R}}\left|\mathbb{P}\left(\frac{\sqrt{L}\left(\boldsymbol{c}^\top\overline{\boldsymbol{\beta}} - \boldsymbol{c}^\top\widetilde{\boldsymbol{\beta}}^*\right)}{\sqrt{\overline{\boldsymbol{c}}^\top\left[\mathcal{P}\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\mathcal{P}\right]^+\overline{\boldsymbol{c}}}} \leq x\middle|\mathcal{G}\right) - \mathbb{P}\left(\mathcal{N}(0,1) \leq x\right)\right| \lesssim \frac{\kappa_1}{\sqrt{npL}},$$

with probability exceeding $1 - O(n^{-10})$ *(randomness comes from* $\mathcal{G}$*). Combining the approximation error and asymptotic distribution together, and by taking all randomness into consideration, we further obtain*

$$\begin{aligned}
&\sup_{x\in\mathbb{R}}\left|\mathbb{P}\left(\frac{\sqrt{L}\left(\boldsymbol{c}^\top\widetilde{\boldsymbol{\beta}}_M-\boldsymbol{c}^\top\widetilde{\boldsymbol{\beta}}^*\right)}{\sqrt{\overline{\boldsymbol{c}}^\top\left[\mathcal{P}\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\mathcal{P}\right]^+\overline{\boldsymbol{c}}}}\le x\right)-\mathbb{P}\left(\mathcal{N}(0,1)\le x\right)\right| \\
\lesssim&\frac{\kappa_1}{\sqrt{npL}}+\left[\kappa_1^6\frac{(d+1)\log n}{\sqrt{npL}}+\kappa_1^4\sqrt{\frac{(d+1)\log n}{np}}\left(\sqrt{d+1}+\frac{\log n}{\sqrt{np}}\right)\right]\frac{\|\boldsymbol{c}_{1:n}\|_1}{\|\overline{\boldsymbol{c}}\|_2} \\
&+\kappa_1^4\frac{(d+1)^{0.5}\log n}{\sqrt{pL}}\frac{\|\boldsymbol{c}_{n+1:n+d}\|_2}{\|\overline{\boldsymbol{c}}\|_2}+\frac{1}{n^5}.
\end{aligned}$$

In Theorem 10, we first obtain the distributional guarantee of $\frac{\sqrt{L}(\boldsymbol{c}^\top\overline{\boldsymbol{\beta}}-\boldsymbol{c}^\top\widetilde{\boldsymbol{\beta}}^*)}{\sqrt{\overline{\boldsymbol{c}}^\top\left[\mathcal{P}\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\mathcal{P}\right]^+\overline{\boldsymbol{c}}}}$, conditioned on the comparison graph $\mathcal{G}$, in the sense that we only take the randomness of $y_{i,j}^{(l)}$ into consideration. These results are stronger than the distributional guranttees which make use of all the randomness from $\mathcal{G}$ and $y_{i,j}^{(l)}$, by the dominated convergence theorem. Besides, combining with the approximation results, we further derive distributional guarantees for linear combinations of $\widetilde{\boldsymbol{\beta}}_M$ by taking all the randomness of into consideration. The asymptotic variance of $\mathbf{c}^\top\widetilde{\boldsymbol{\beta}}_M$ is $\overline{\boldsymbol{c}}^\top[\mathcal{P}\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\mathcal{P}]^+\overline{\boldsymbol{c}}$ as presented in Theorem 10, where $[\mathcal{P}\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\mathcal{P}]^+$ is exactly the inverse of the fisher information (conditioned on graph $\mathcal{G}$) that is projected into the space Θ.

Moreover, recall that we have scaled the covariates to satisfy $\max_{i\in[n]}\|\boldsymbol{x}_i\|_2\le\sqrt{(d+1)/n}$ in the data preprocessing step. Then, if κ_1, $\|\boldsymbol{c}_{1:n}\|_1/\|\overline{\boldsymbol{c}}\|_2$ and $\sqrt{n}\|\boldsymbol{c}_{n+1:n+d}\|_2/(\sqrt{d+1}\|\overline{\boldsymbol{c}}\|_2)$ are bounded, the asymptotic normality holds when

$$\max\left\{\frac{(d+1)\log n}{\sqrt{npL}},\frac{(d+1)\sqrt{\log n}}{\sqrt{np}},\frac{\log n\sqrt{(d+1)\log n}}{np}\right\}=o(1). \tag{7}$$

in the sense that it allows the comparison graph to be sparse (when $p\asymp 1/n$ up to logarithmic terms) when the covariate dimension is bounded. This admits all existing uncertainty quantification results for the BTL model without covariates as special cases.

Finally, we comment on the condition $\max\{\|\boldsymbol{c}_{1:n}\|_1,\sqrt{n/(d+1)}\,\|\boldsymbol{c}_{n+1:n+d}\|_2\}\lesssim\|\overline{\boldsymbol{c}}\|_2$. This is only a mild requirement. For instance, when $d=0$ and $\mathbf{c}\in\mathbb{R}^n$ is sparse (the original BTL model), this inequality holds naturally. In addition, the comparison of preference ratings is another significant illustration that meets the requirement. In specific, for testing $H_0:\widetilde{\boldsymbol{x}}_i^\top\widetilde{\boldsymbol{\beta}}^*\le\widetilde{\boldsymbol{x}}_j^\top\widetilde{\boldsymbol{\beta}}^*$ v.s. $H_a:\widetilde{\boldsymbol{x}}_i^\top\widetilde{\boldsymbol{\beta}}^*>\widetilde{\boldsymbol{x}}_j^\top\widetilde{\boldsymbol{\beta}}^*$, we choose $\boldsymbol{c}=\widetilde{\boldsymbol{x}}_i-\widetilde{\boldsymbol{x}}_j\in\mathbb{R}^{n+d}$. In this scenario, the the condition is met since $\max\{\|\boldsymbol{c}_{1:n}\|_1,\sqrt{n/d+1}\|\boldsymbol{c}_{n+1:n+d}\|_2\}=2$, and $\|\overline{\boldsymbol{c}}\|_2=\|\boldsymbol{c}\|_2\asymp\sqrt{2+2(d+1)/n}$.

An important corollary of Theorem 10 is the limiting distribution of $\widehat{\alpha}_{M,S_k}$, where S_k is any subset over $[n]$ with size $k<\infty$. The corresponding theoretical property is summarized in the following Corollary 11.

Corollary 11. *Assume the assumptions of Theorem 7 hold. Then for any fixed $k \in [n]$, as long as*

$$\kappa_1^6 \frac{\sqrt{k}(d+1)\log n}{\sqrt{npL}} + \kappa_1^4 \sqrt{\frac{k(d+1)\log n}{np}} \left(\sqrt{d+1} + \frac{\log n}{\sqrt{np}}\right) = o(1),$$

we have

$$\sqrt{L}\left(\left[\mathcal{P}\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\mathcal{P}\right]^+_{S_k,S_k}\right)^{-1/2}(\widehat{\boldsymbol{\alpha}}_{M,S_k} - \boldsymbol{\alpha}^*_{S_k}) \rightsquigarrow \mathcal{N}\left(\mathbf{0}, \mathbf{I}_k\right),$$

where S_k is any subset over $[n]$ with size k.

Although our results are established under a more general setting than the existing literature on the BTL model, which does not involve the covariates, our specific result in Corollary 11 with $d = 0$ can still compare favorably. Compared with Liu et al. (2022), we need a much smaller sample complexity to establish the asymptotic normality. Specifically, they require $n\log n/\sqrt{L} + \log n/\sqrt{pL} = o(1)$ to derive the asymptotic normality. This condition requires $L \gg n^2$. In contrast, we allow $L = \mathcal{O}(1)$ and our requirement for the sample complexity is minimax optimal up to logarithm terms (Negahban et al., 2012; Chen et al., 2019). Moreover, Liu et al. (2022) use the Lagrangian debiasing method to derive the estimators, which involves an additional tuning parameter.

Compared with Han et al. (2020), we allow sparse compare graphs ($p \gtrsim 1/n$ by ignoring logarithm terms), whereas they require a much denser comparison graph ($p \gtrsim 1/n^{1/10}$) than ours.

We now compare our results with Gao et al. (2021) and Fan et al. (2022) (under the pairwise comparison regime). As the analysis in these two works does not incorporate the condition number κ_1 or covariate, we will consider the regime $\kappa_1 = O(1)$ and $d = 0$ in our theorems for comparison. First of all, both papers show that the asymptotic normality holds even for the sparse regime $p \asymp 1/n$, up to a logarithmic order. However, the choice of approximators and approximation errors are very different. Instead of using the Taylor expansion $\overline{\mathcal{L}}(\cdot)$ given by Eq. (6), Gao et al. (2021); Fan et al. (2022) consider the following quadratic approximation:

$$\overline{\mathcal{L}}_{\text{diag}}(\widetilde{\boldsymbol{\beta}}) = \mathcal{L}(\widetilde{\boldsymbol{\beta}}^*) + \left(\widetilde{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}^*\right)^\top \nabla\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*) + \frac{1}{2}\left(\widetilde{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}^*\right)^\top \left(\mathbf{diag}\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\right)\left(\widetilde{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}^*\right),$$

which only keeps the diagonal part of $\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)$, and define the approximator $\overline{\boldsymbol{\beta}}_{\text{diag}} = \operatorname{argmin}\overline{\mathcal{L}}_{\text{diag}}(\widetilde{\boldsymbol{\beta}})$. In addition, it is worth noting that their approach, which only keeps the diagonal of the Hessian matrix to handle the approximation error, cannot be applied directly to our setting (with covariates) due to several reasons. Firstly, results in Gao et al. (2021) without involving covariates imply that for any pair of indices i, j within the sample space $[n]$, the estimators $\widehat{\alpha}_{M,i}$ and $\widehat{\alpha}_{M,j}$ (corresponds to $\widehat{\theta}_i, \widehat{\theta}_j$ in their paper) are asymptotically independent, which justifies their utilization of a quadratic approximation for a

single variable while keeping others fixed. In our model that incorporates covariates, the corresponding estimators $\widehat{\alpha}_{M,i}$ and $\widehat{\alpha}_{M,j}$ do not exhibit asymptotic independence. Secondly, the approach in Gao et al. (2021) overlooks the off-diagonal elements of the Hessian matrix $\nabla^2\mathcal{L}(\widetilde{\beta}^*)$, which is an acceptable approximation when no covariates are involved since these off-diagonal entries are negligible compared with asymptotic variances. However, in our study, after we take the covariates into consideration, this approximation no longer holds. As a result, an expansion similar to [(2.10), (Gao et al., 2021)] would not give a valid expansion in our case. Therefore, this motivates us to drive our own method presented above.

In terms of the approximation errors, they show that with high probability

$$\left|\frac{\sqrt{L}\left(\widehat{\alpha}_{M,k}-\overline{\boldsymbol{\beta}}_{\mathrm{diag},k}\right)}{\sqrt{\left(\left[\mathcal{P}\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\mathcal{P}\right]^+\right)_{k,k}}}\right| \lesssim \sqrt{\frac{\log n}{np}}+\frac{(\log n)^3}{(np)^2}+\left(\frac{\log n}{npL}\right)^{1/4}+\frac{(\log n)^{7/4}}{(np)^{5/4}L^{1/4}}, \tag{8}$$

while we prove that for our approximation $\overline{\boldsymbol{\beta}}$, with high probability for $k\in[n]$ it holds that

$$\left|\frac{\sqrt{L}\left(\widehat{\alpha}_{M,k}-\overline{\alpha}_{k}\right)}{\sqrt{\left(\left[\mathcal{P}\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\mathcal{P}\right]^+\right)_{k,k}}}\right| \lesssim \sqrt{\frac{\log n}{np}}+\frac{(\log n)^{1.5}}{np}. \tag{9}$$

When $L=o(np/\log n)$, the term $(\log n/npL)^{1/4}$ dominates the right-hand side of (8) and hence also dominates the error rate given by (9) as long as $np\gtrsim(\log n)^2$. In other words, our approximation error is an order of magnitude smaller than that in Gao et al. (2021) and Fan et al. (2022) (when considering the pairwise comparison regime). This holds true for the common case where $L\asymp 1$. We next explain the underlying rationale. In scenarios where the likelihood function does not contain covariates, the off-diagonal elements of the Hessian matrix, though in smaller order compared to the asymptotic variance, still contribute to bias in the remainder term beyond the non-asymptotic expansion. This inherent bias, which emerges when examining marginal likelihoods, is more pronounced than using joint likelihood. Our approach, which employs joint likelihood, effectively accounts for these off-diagonal elements, thereby mitigating their impact and resulting in a more accurate approximation.

Besides investigating the asymptotic behavior of $\widehat{\alpha}_i, i\in[n]$, studying the asymptotic property for $\widehat{\beta}_j, j\in[d]$ is another crucial task as it depicts whether some covariates have any power for explaining latent scores. We deduce these from the Theorem 7 and Theorem 10, and summarize them together with a refined ℓ_2-upper bound of $\|\widehat{\boldsymbol{\beta}}_M-\boldsymbol{\beta}^*\|_2$ in Corollary 12.

Corollary 12. *Under the assumptions of Theorem 7 and Theorem 10, we first obtain a refined upper bound for* $\|\widehat{\boldsymbol{\beta}}_M - \boldsymbol{\beta}^*\|_2$*, namely*

$$\left\|\widehat{\boldsymbol{\beta}}_M - \boldsymbol{\beta}^*\right\|_2 \lesssim \sqrt{\frac{(d+1)}{n}} \cdot \max\left\{\kappa_1^4 \frac{\log n}{pL}, \kappa_1 \sqrt{\frac{\log n}{pL}}\right\}. \tag{10}$$

In addition, we also achieve the distributional results for every $\widehat{\beta}_j, j \in [d]$*,*

$$\begin{aligned}
&\sup_{x\in\mathbb{R}}\left|\mathbb{P}\left(\frac{\sqrt{L}\left(\boldsymbol{e}_{n+j}^\top\widetilde{\boldsymbol{\beta}}_M - \boldsymbol{e}_{n+j}^\top\widetilde{\boldsymbol{\beta}}^*\right)}{\sqrt{\bar{\boldsymbol{e}}_{n+j}^\top\left[\mathcal{P}\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}_M)\mathcal{P}\right]^+\bar{\boldsymbol{e}}_{n+j}}} \le x\right) - \mathbb{P}\left(\mathcal{N}(0,1)\le x\right)\right| \\
\lesssim&\frac{\kappa_1}{\sqrt{npL}} + \kappa_1^4\frac{(d+1)^{0.5}\log n}{\sqrt{pL}} + \frac{1}{n^5}.
\end{aligned}$$

Therefore, when $(d+1)^{0.5}\log n/(pL) \to 0$*, for any* $j \in [d]$*, we have*

$$\mathbb{P}\left(\beta_j^* \in [C_L(\widetilde{\boldsymbol{\beta}}_M), C_U(\widetilde{\boldsymbol{\beta}}_M)]\right) = 1-\alpha,$$

where $[C_L(\widetilde{\boldsymbol{\beta}}_M), C_U(\widetilde{\boldsymbol{\beta}}_M)] = [\boldsymbol{e}_{n+j}^\top\widetilde{\boldsymbol{\beta}}_M - \frac{\sqrt{\bar{\boldsymbol{e}}_{n+j}^\top[\mathcal{P}\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}_M)\mathcal{P}]^+\bar{\boldsymbol{e}}_{n+j}}z_{\alpha/2}}{\sqrt{L}}, \boldsymbol{e}_{n+j}^\top\widetilde{\boldsymbol{\beta}}_M + \frac{\sqrt{\bar{\boldsymbol{e}}_{n+j}^\top[\mathcal{P}\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}_M)\mathcal{P}]^+\bar{\boldsymbol{e}}_{n+j}}z_{\alpha/2}}{\sqrt{L}}]$ *with* $z_{\alpha/2}$ *being the upper* $\alpha/2$*-quantile of the standard Gaussian distribution.*

In Corollary 12, we first obtain a refined upper bound of $\|\widehat{\boldsymbol{\beta}}_M - \boldsymbol{\beta}^*\|_2$, and we next explain the rationality behind this. In Theorem 4, a rough upper bound for $\|\widehat{\boldsymbol{\beta}}_M - \boldsymbol{\beta}^*\|_2$ is obtained via concentration since no precise distributional results are involved in that stage. Given the statistical rates derived in Theorem 4, we then analyze the non-asymptotic approximation and distribution of $\widehat{\boldsymbol{\beta}}_M$, in Theorems 7 and 10, respectively. Finally, based on these distributional results, a refined upper bound for $\|\widehat{\boldsymbol{\beta}}_M - \boldsymbol{\beta}^*\|_2$ is achieved.

If we let $\kappa_1, d = \mathcal{O}(1)$, one observes that the final upper bound of $\|\widehat{\boldsymbol{\beta}}_M - \boldsymbol{\beta}^*\|_2$ involves both rates $\frac{1}{\sqrt{npL}}$ and $\frac{1}{\sqrt{npL}}$, by ignoring logarithm terms. We conjecture the term $\frac{1}{\sqrt{npL}}$, which comes from the approximation error ($\|\widehat{\boldsymbol{\beta}}_M - \overline{\boldsymbol{\beta}}_{n+1:n+d}\|_2$) in Theorem 7, can be improved to $\frac{1}{npL}$. In this case, $\|\widehat{\boldsymbol{\beta}}_M - \boldsymbol{\beta}^*\|_2$ should be bounded by the rate of $\mathcal{O}(\sqrt{\frac{1}{npL}})$ (the same order as the variance of $\widehat{\boldsymbol{\beta}}_M$). The numerical studies in §5.1 validates this conjecture by showing that the rates of $\|\widehat{\boldsymbol{\beta}}_M - \boldsymbol{\beta}^*\|_2$ is proporation to $\frac{1}{\sqrt{pL}}$ after we fix n and d. However, improving the approximation error is highly non-trivial and needs more complex theoretical analysis. Therefore, we will leave this as our future work. It is worth mentioning that the assumption $(d+1)^{0.5}\log n/(pL) \to 0$ is only required while doing inference for $\beta_j^*, j \in [d]$. In all the previously mentioned theorems and corollaries, we do not need this condition and allow $L = 1$.

Besides the refined ℓ_2-bounds, the asymptotic distribution for each $\widehat{\beta}_j, j \in [d]$ and the $(1-\alpha)$- confidence interval for β_j^* are also derived in Corollary 12. This will enable us to determine each covariate's significance in real data studies.

5 Numerical Results

In this section, we conduct numerical experiments using synthetic and real data to validate our theories. In §5.1 and §5.2, we leverage synthetic data to corroborate the statistical rates given in §3 and distributional results given in §4, respectively. In addition, in §5.5, we illustrate further our model and methods by using the mutual funds holding data.

5.1 Rate of Convergence

We begin with the data generation process. Throughout the synthetic data experiments, we set n to be 200 and d to be 5. The covariates are generated independently with $(\boldsymbol{x}_i)_j \sim \text{Uniform}[-0.5, 0.5]$ for all $i \in [n], j \in [d]$. For matrix $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n]^\top \in \mathbb{R}^{n\times d}$, its columns are then normalized such that they have mean 0 and standard deviation 1. Next, we scale $\boldsymbol{x}_i$ by $\boldsymbol{x}_i/K$ so that $\max_{i\in[n]} \|\boldsymbol{x}_i\|_2/K = \sqrt{(d+1)/n}$. We generate $\breve{\boldsymbol{\alpha}} \in \mathbb{R}^n$ by sampling its entries independently from $\text{Uniform}[0.5, \log(5) - 0.5]$. Also, a $\breve{\boldsymbol{\beta}} \in \mathbb{R}^d$ is generated uniformly from the hypersphere $\{\boldsymbol{\beta} : \|\boldsymbol{\beta}\|_2 = 0.5\sqrt{n/(d+1)}\}$. Then we project $(\breve{\boldsymbol{\alpha}}^\top, \breve{\boldsymbol{\beta}}^\top)^\top$ onto linear space Θ and let it be $\widetilde{\boldsymbol{\beta}}^*$. In this way, we ensure $\kappa_1 \le 5$.

To validate the statistical rates in Theorem 4, we use the above method to generate the covariates $\boldsymbol{x}_i$ and $\widetilde{\boldsymbol{\beta}}^*$ for three times. This gives us three different instances of the covariates $\boldsymbol{x}_i$ and the parameter $\widetilde{\boldsymbol{\beta}}^*$. For each given instance, we consider 6 different (p, L) pairs, which are listed below.

p	1	0.5	0.222	0.625	0.4	0.278
L	50	25	25	5	5	5

For each (p, L) pair, comparion graph $\mathcal{E}$, $\{y_{i,j}^{(\ell)}, l \in [L], (i,j) \in \mathcal{E}\}$ is generated and the MLE $\widetilde{\boldsymbol{\beta}}_M$ is calculated based on the available data. This process is repeated 200 times and the averaged $\|\widehat{\boldsymbol{\alpha}}_M - \boldsymbol{\alpha}^*\|_\infty$, $\|\widehat{\boldsymbol{\beta}}_M - \boldsymbol{\beta}^*\|_2/\|\boldsymbol{\beta}^*\|_2$ as well as their associated standard deviations are recorded. The results are depicted in Figure 1 for each of the three instances. Note that $\|\widehat{\boldsymbol{\alpha}}_M - \boldsymbol{\alpha}^*\|_\infty$ and $\|\widehat{\boldsymbol{\beta}}_M - \boldsymbol{\beta}^*\|_2/\|\boldsymbol{\beta}^*\|_2$ are nearly proportional to $1/\sqrt{pL}$, lending further support of the results in Theorem 4. The results are insensitive to three different instances, as expected.

5.2 Distributional Results

We employ the same method given in §5.1 to generate the covariates $\boldsymbol{x}_i$ and $\widetilde{\boldsymbol{\beta}}^*$ once and fix them throughout the simulation. Letting the effective sample size $n_a = n/[(d+1)\log n]$, we choose the 6 pairs (p, L) with $p = 1.25/n_a$ or $p = 2/n_a$ and $L \in \{2, 6, 20\}$. For each (p, L) pair, the graph $\mathcal{E}$ and data $\{y_{i,j}^{(\ell)}, l \in [L], (i,j) \in \mathcal{E}\}$ are generated 250 times and the MLEs $\widetilde{\boldsymbol{\beta}}_M$ for all simulations are recorded. Figure 2 presents the Q-Q plots for checking the normality of $(\widehat{\boldsymbol{\alpha}}_M)_1$, the first component of $\widehat{\boldsymbol{\alpha}}_M$. The results show that $(\widehat{\boldsymbol{\alpha}}_M)_1$ follows closely the normal distribution.

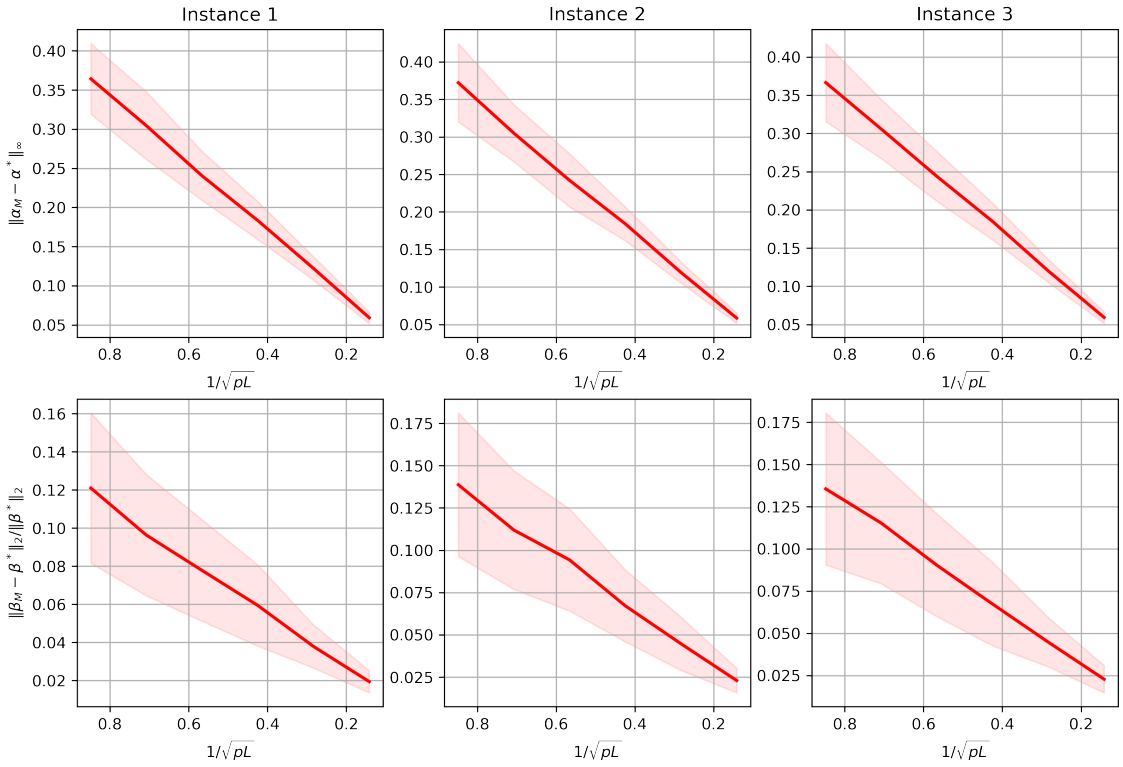

Figure 1: Statistical rates of $\|\widehat{\boldsymbol{\alpha}}_M - \boldsymbol{\alpha}^*\|_\infty$ and $\|\widehat{\boldsymbol{\beta}}_M - \boldsymbol{\beta}^*\|_2/\|\boldsymbol{\beta}^*\|_2$ for three simulated instances (realization of simulated models). The solid red lines and light areas represent the averaged $\|\widehat{\boldsymbol{\alpha}}_M - \boldsymbol{\alpha}^*\|_\infty$, $\|\widehat{\boldsymbol{\beta}}_M - \boldsymbol{\beta}^*\|_2/\|\boldsymbol{\beta}^*\|_2$ and their associated standard errors based on 200 Monte Carlo simulations.

In addition to checking the asymptotic normality, we now verify the asymptotic variance of our estimator. As an illustration, we consider the linear combination $\boldsymbol{c}^\top \widetilde{\boldsymbol{\beta}}_M$, where $\boldsymbol{c} = \boldsymbol{e}_1 + \boldsymbol{e}_{201}$ and $\boldsymbol{e}_i$ is the i-th vector from the standard basis of $\mathbb{R}^{205}$. Based on 250 simulations with $(p, L) = (1.25/n_a, 2)$ and $(p, L) = (2/n_a, 20)$, the histograms of the following two standardized random variables are plotted:

$$A = \frac{\sqrt{L}\left(\boldsymbol{c}^\top \widetilde{\boldsymbol{\beta}}_M - \boldsymbol{c}^\top \widetilde{\boldsymbol{\beta}}^*\right)}{\sqrt{\overline{\boldsymbol{c}}^\top \left[\mathcal{P}\nabla^2 \mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\mathcal{P}\right]^+ \overline{\boldsymbol{c}}}} \text{ and } B = \frac{\sqrt{L}\left(\boldsymbol{c}^\top \widetilde{\boldsymbol{\beta}}_M - \boldsymbol{c}^\top \widetilde{\boldsymbol{\beta}}^*\right)}{\sqrt{\overline{\boldsymbol{c}}^\top \left[\mathcal{P}\nabla^2 \mathcal{L}(\widetilde{\boldsymbol{\beta}}_M)\mathcal{P}\right]^+ \overline{\boldsymbol{c}}}}. \tag{11}$$

This uses the asymptotic theory with plug-in asymptotic variance using the true and estimated parameters, where $\overline{\boldsymbol{c}} = P_\Theta(\boldsymbol{c})$ is the projection of $\boldsymbol{c}$ onto linear space Θ. Figure 3 shows that the histograms follow closely the standard Gaussian density. The first row of Figure 3 is presented in the regime with $(p, L) = (1.25/n_a, 2)$. It holds that, even when the sample size is very small, the two density plots are still very close to the standard Gaussian

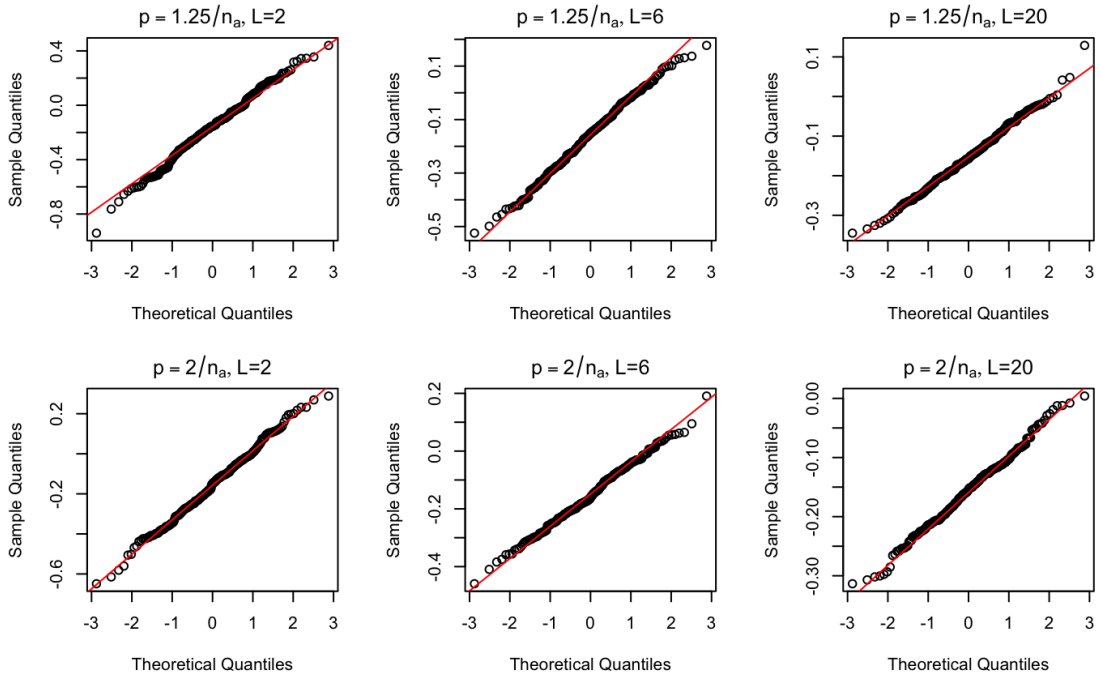

Figure 2: Q-Q Plots for checking the normality of $(\widehat{\boldsymbol{\alpha}}_M)_1$ based on 250 simulations.

density. The second row of Figure 3 is drawn based on the setting where $(p, L) = (2/n_a, 20)$. In this case, the density plots of A and B are more stable and close to the standard Gaussian curve. These results in turn support our theoretical results in Theorem 10.

5.3 Comparison with BTL Model Without Covariates

In this section, we conduct a series of simulations to make comparisons (proportion of information being explained, prediction accuracy, and sensitivity analysis) and present the detailed results below.

- We show the first merit of our proposed method, in terms of the proportion of information being explained.

  We record the following quantity $1 - \|\widehat{\boldsymbol{\alpha}}_M\|_2^2 / \|\widehat{\boldsymbol{\alpha}}_M + \boldsymbol{X}\widehat{\boldsymbol{\beta}}_M\|_2^2$, which quantifies the proportion of information being explained by the covariates. We consider the same (p, L) pair presented in Section 5.1. For each (p, L) pair, we generate the comparison graph $\mathcal{E}$, the comparison results $\{y_{i,j}^{(\ell)}, l \in [L], (i, j) \in \mathcal{E}\}$ and solve the MLE $\widetilde{\boldsymbol{\beta}}_M$ for 200 times. We report the mean and standard deviation of the concerned quantity for each (p, L) pair based on these 200 repetitions. The results are presented in the following Table 1.

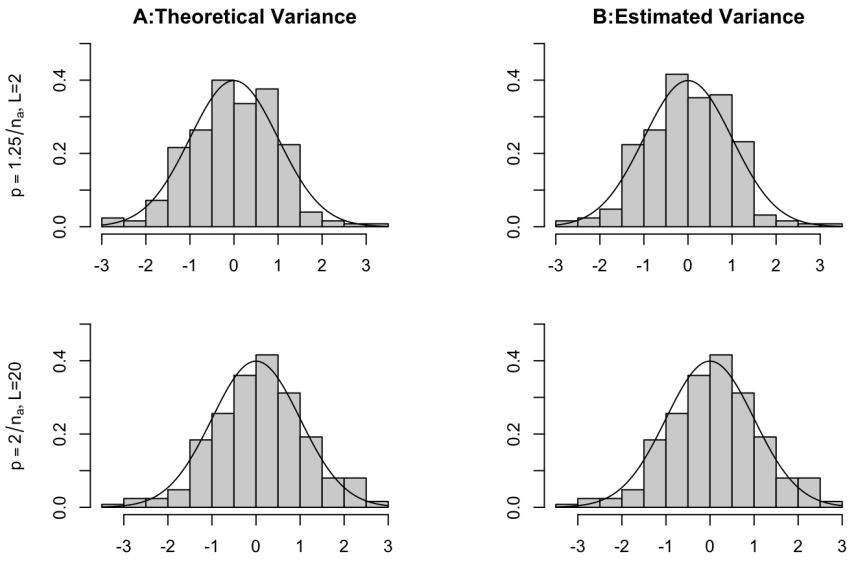

Figure 3: Histograms of standardized random variables A and B in (11) along with the density of the standardized Gaussian random variable. The first row is based on $(p, L) = (1.25/n_a, 2)$ and the second row is based on $(p, L) = (2/n_a, 20)$.

(p, L)	$1 - \|\widehat{\boldsymbol{\alpha}}_M\|_2^2 / \|\widehat{\boldsymbol{\alpha}}_M + \boldsymbol{X}\widehat{\boldsymbol{\beta}}_M\|_2^2$
$(1, 50)$	0.370 ± 0.008
$(0.5, 25)$	0.390 ± 0.016
$(0.222, 25)$	0.415 ± 0.023
$(0.625, 5)$	0.449 ± 0.028
$(0.4, 5)$	0.486 ± 0.030
$(0.278, 5)$	0.531 ± 0.032

Table 1: Mean and standard deviation of $1 - \|\widehat{\boldsymbol{\alpha}}_M\|_2^2 / \|\widehat{\boldsymbol{\alpha}}_M + \boldsymbol{X}\widehat{\boldsymbol{\beta}}_M\|_2^2$ for each (p, L) pair based on 200 repetitions.

We conclude from Table 1, involving the covariates helps reduce the magnitude of unexplained information for all settings. Therefore, this further helps making out-of-sample predictions, and we will discuss this point in the following step.

- Second, we present the prediction performance of our model and compare it with BTL model without covariates. We generate a new set of covariates $[\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_n]^\top$ where $\boldsymbol{z}_i \in \mathbb{R}^d$ falls in the same column space of $\mathbf{X}$. (the detailed simulation setting can be found in the main text). With these new covariates, for any $i \neq j$, the out-of-sample

probability $\mathbb{P}\{\text{item } j \text{ is preferred over item } i\}$ becomes

$$p^*_{i,j} = \frac{e^{\alpha^*_j + \boldsymbol{z}_j^\top \boldsymbol{\beta}^*}}{e^{\alpha^*_i + \boldsymbol{z}_i^\top \boldsymbol{\beta}^*} + e^{\alpha^*_j + \boldsymbol{z}_j^\top \boldsymbol{\beta}^*}}.$$

Using the covariate information, the probability predicted by our model is

$$\widehat{p}^c_{i,j} = \frac{e^{\widehat{\alpha}_{M,j} + \boldsymbol{z}_j^\top \widehat{\boldsymbol{\beta}}_M}}{e^{\widehat{\alpha}_{M,i} + \boldsymbol{z}_i^\top \widehat{\boldsymbol{\beta}}_M} + e^{\widehat{\alpha}_{M,j} + \boldsymbol{z}_j^\top \widehat{\boldsymbol{\beta}}_M}}.$$

However, when one does not take the covariates information into consideration, the best one can do is to use the estimated score $\widehat{\theta}_i$ under the original BTL model, as Chen et al. (2019); Gao et al. (2021) did. In this case, the predicted probability is

$$\widehat{p}^{nc}_{i,j} = \frac{e^{\widehat{\theta}_j}}{e^{\widehat{\theta}_i} + e^{\widehat{\theta}_j}}.$$

In Table 2 we present the mean square error $\sum_{i<j}(\widehat{p}^c_{i,j} - p^*_{i,j})^2$ and $\sum_{i<j}(\widehat{p}^{nc}_{i,j} - p^*_{i,j})^2$ for the six (p, L) pairs we mentioned above. The results show then mean and standard deviation of the mean square error calculated over 200 repetitions. As we can see from Table 2, the estimator $\widehat{p}_{i,j}$ which takes the covariate information into considers performs much better than the one without covariate information.

(p, L)	With covariates	Without covariates
$(1, 50)$	0.970 ± 0.094	110.325 ± 1.377
$(0.5, 25)$	3.909 ± 0.394	113.453 ± 3.086
$(0.222, 25)$	8.975 ± 0.940	118.796 ± 4.293
$(0.625, 5)$	15.537 ± 1.572	125.149 ± 5.441
$(0.4, 5)$	24.874 ± 2.718	134.796 ± 7.432
$(0.278, 5)$	36.487 ± 4.243	146.286 ± 9.961

Table 2: Mean and standard deviation of mean square error $\sum_{i<j}(\widehat{p}^c_{i,j} - p^*_{i,j})^2$ and $\sum_{i<j}(\widehat{p}^{nc}_{i,j} - p^*_{i,j})^2$ under each setting based on 200 simulations.

In the next step, we will also conduct a sensitivity analysis to compare our prediction results with a scenario where no covariates are considered.

- Lastly, we test the sensitivity of our model by violating the linearity assumption of our model. And subsequently, we also compare its performance to that of the BTL model without covariates.

  We modify the underlying model to be

$$p^*_{i,j} = \frac{e^{\alpha^*_j + \boldsymbol{w}_j^\top \boldsymbol{\beta}^* + g(\boldsymbol{w}_j)}}{e^{\alpha^*_i + \boldsymbol{w}_i^\top \boldsymbol{\beta}^* + g(\boldsymbol{w}_i)} + e^{\alpha^*_j + \boldsymbol{w}_j^\top \boldsymbol{\beta}^* + g(\boldsymbol{w}_j)}},$$

where $\boldsymbol{w}_i = \boldsymbol{x}_i$ when we fit the model and $\boldsymbol{w}_i = \boldsymbol{z}_i$ when we make prediction. Here $g(\cdot)$ is a nonlinear function. In our experiment, we fix $d = 5$ and let

$$g(\boldsymbol{w}_i) = c(0.1(\boldsymbol{w}_i)_1(\boldsymbol{w}_i)_2 + 0.2(\boldsymbol{w}_i)_2(\boldsymbol{w}_i)_3 + 0.3(\boldsymbol{w}_i)_4(\boldsymbol{w}_i)_5), \tag{12}$$

and c is chosen to be 20 and 100 to accommodate different levels of non-linearity. With covariate information, the winning probability is still predicted as

$$\widehat{p}^{c}_{i,j} = \frac{e^{\widehat{\alpha}_{M,j} + \boldsymbol{z}_j^\top \widehat{\boldsymbol{\beta}}_M}}{e^{\widehat{\alpha}_{M,i} + \boldsymbol{z}_i^\top \widehat{\boldsymbol{\beta}}_M} + e^{\widehat{\alpha}_{M,j} + \boldsymbol{z}_j^\top \widehat{\boldsymbol{\beta}}_M}}.$$

while the predicted probability of the original BTL model is

$$\widehat{p}^{nc}_{i,j} = \frac{e^{\widehat{\theta}_j}}{e^{\widehat{\theta}_i} + e^{\widehat{\theta}_j}}.$$

In Table 3 and 4 we present the results when c is chosen to be 20 and 100. We again consider the mean square error $\sum_{i<j}(\widehat{p}^{c}_{i,j} - p^*_{i,j})^2$ and $\sum_{i<j}(\widehat{p}^{nc}_{i,j} - p^*_{i,j})^2$ as the metric. The experiments are repeated 200 times for each (p, L) pair and we report the mean and standard deviation.

(p, L)	With covariates	Without covariates
$(1, 50)$	4.160 ± 0.271	201.799 ± 2.056
$(0.5, 25)$	7.173 ± 0.625	204.962 ± 3.729
$(0.222, 25)$	12.411 ± 1.203	209.903 ± 6.010
$(0.625, 5)$	18.871 ± 1.885	216.580 ± 7.871
$(0.4, 5)$	27.797 ± 2.896	225.813 ± 11.310
$(0.278, 5)$	39.372 ± 4.024	237.060 ± 11.710

Table 3: Mean and standard deviation of mean square error $\sum_{i<j}(\widehat{p}^{c}_{i,j} - p^*_{i,j})^2$ and $\sum_{i<j}(\widehat{p}^{nc}_{i,j} - p^*_{i,j})^2$ under each setting based on 200 simulations. The level of non-linearity in (12) is chosen to be $c = 20$.

From Table 3 and 4 we can see that our model consistently performs better than the original BTL model when different levels of non-linearity exist.

5.4 Application to Pokemon Challenge Data Set

We apply the proposed method to study the Pokemon challenge data set. The original data set can be found at `https://www.kaggle.com/c/intelygenz-pokemon-challenge/data`. This data set records the pairwise competition records among 800 pokemon, whose covariate information is also recorded. This data set contains 50000 competition results, each competition takes place between two pokemons and has one winner.

(p, L)	With covariates	Without covariates
$(1, 50)$	49.074 ± 0.960	102.936 ± 1.438
$(0.5, 25)$	51.679 ± 1.986	105.463 ± 2.870
$(0.222, 25)$	56.958 ± 3.026	111.269 ± 4.317
$(0.625, 5)$	63.858 ± 4.076	117.639 ± 5.670
$(0.4, 5)$	72.375 ± 5.574	125.958 ± 7.556
$(0.278, 5)$	84.260 ± 7.122	138.484 ± 9.488

Table 4: Mean and standard deviation of mean square error $\sum_{i<j}(\widehat{p}^{c}_{i,j} - p^*_{i,j})^2$ and $\sum_{i<j}(\widehat{p}^{nc}_{i,j} - p^*_{i,j})^2$ under each setting based on 200 simulations. The level of non-linearity in (12) is chosen to be $c = 100$.

Our experiments mainly focus on predicting the ability of the mega evolved pokemons. We think that a mega evolved pokemon has the same intrinsic ability (same α_i) as pre-evolutionary pokemon, and the mega evolution may only change the covariates $\boldsymbol{x}$. We have 48 mega evolved pokemons in this data set, and we randomly select 28 of them to test our predictive performance. Among these remaining $800 - 28 = 772$ pokemons for training purpose, we select the largest connected component of their comparison graph. Eventually we have 757 pokemons left for training. For each pokemon, we select $\log(\textit{Attack})$, $\log(\textit{HP})$, *Mega or not* as their covariates. Here *Attack* and *HP* denote the ability to attack and durability, respectively. The variable *Mega or not* takes binary value and represents whether this pokemon is mega evolved or not. We optimize the likelihood of our CARE model in (4) using training data and record $\widehat{\boldsymbol{\alpha}}_M$ and $\widehat{\boldsymbol{\beta}}_M$.

We first investigate the statistical significance of these 3 variables we are interested in. This amounts to testing the following hypothesis testing problems for each feature:

$$H_0 : \beta^*_i = 0 \quad \text{v.s.} \quad H_a : \beta^*_i \neq 0,\ i \in [3].$$

The test statistics are given by $\beta_{M,i}/\sqrt{([\mathcal{P}\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}_M)\mathcal{P}]^+)_{n+i,n+i}}$ for all $i \in [3]$ and the corresponding p-values are calculated via the asymptotic normality results in §4. The results are depicted in Table 5.4, and these three variables are all statistically significant.

	Estimate	p-value
Attack	2.743	$< 1e-5$
HP	3.759	$< 1e-5$
Mega or not	1.603	$< 1e-5$

We then evaluate the competitions performance of the 28 mega evolved pokemons in the test sample, whose pre-evolutionary versions are the training data. We predict the score of an evolved pokemon as $\widehat{\theta}^{\text{predicted}} := \mathsf{SOFT}(\widehat{\alpha}_{M,\text{pe}}, \tau_{\text{pe}}) + \boldsymbol{z}^\top\widehat{\boldsymbol{\beta}}_M$, where $\widehat{\alpha}_{M,\text{pe}}$ is

the estimated intrinsic score of the pre-evolutionary version and $\boldsymbol{z}$ is the new covariates of this mega evolved pokemon. Here we apply a soft-thresholding to the $\boldsymbol{\alpha}$ part with $\mathsf{SOFT}(x,\tau) = \mathrm{sign}(x)\cdot\max\{|x|-\tau,0\}$ and $\tau_i = \Phi^{-1}(1-0.025/757)\cdot\sqrt{([\mathcal{P}\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}_M)\mathcal{P}]^+)_{i,i}}$ for each pokemon i in the training data set. This corresponds to set those estimates $\widehat{\alpha}_i$ that are statistically indifferent from zero to zero. The use of significant level $0.025/757$ is to control the familywise false positive rates at a level of 0.05.

In order to formalize the metric we used to test our procedure, we introduce the following notation first. Let $\mathcal{T}_2$ be set of the pokemons we want to predict and let $\mathcal{T}_1$ be set of the pokemons in the training data who have competitions with pokemons in $\mathcal{T}_2$. Given any two pokemons i and j, we let $D_{i,j}$ be the competitions between i and j. For each pokemon i from $\mathcal{T}_1$, we further define the estimated score as $\widehat{\theta}_i^{\text{estimated}} := \mathsf{SOFT}(\widehat{\alpha}_{M,i},\tau_i) + \boldsymbol{x}_i^\top\widehat{\boldsymbol{\beta}}_M$. We use the following quantity $Loss_c$ as a measure of the prediction loss for pokemons in $\mathcal{T}_2$ when we have covariate information

$$Loss_c = \sum_{i\in\mathcal{T}_1, j\in\mathcal{T}_2}\sum_{y\in D_{i,j}}\left(\frac{e^{\widehat{\theta}_j^{\text{predicted}}}}{e^{\widehat{\theta}_i^{\text{estimated}}}+e^{\widehat{\theta}_j^{\text{predicted}}}}-y\right)^2 + \sum_{i,j\in\mathcal{T}_2}\sum_{y\in D_{i,j}}\left(\frac{e^{\widehat{\theta}_j^{\text{predicted}}}}{e^{\widehat{\theta}_i^{\text{predicted}}}+e^{\widehat{\theta}_j^{\text{predicted}}}}-y\right)^2.$$

As a comparison, for the original BTL model, we let $\widehat{\theta}^{\text{BTL}}$ be the estimated score under the original BTL model. Since there no covariate information is involved, for pokemons from $\mathcal{T}_2$, their $\widehat{\theta}^{\text{predicted}}$ are set to be the estimated scores of their pre-evolutionary versions. In addition, for pokemons from $\mathcal{T}_1$, their scores $\widehat{\theta}^{\text{estimates}}$ are set to be $\widehat{\theta}^{\text{BTL}}$ estimated via the training data. The loss for our method and original BTL model are reported in Table 5.4. As we can see, our model achieves a significant improvement over the original BTL model.

With covariates	Without covariates
283.23	316.04

5.5 An Application to Ranking of Stocks

In this subsection, we apply our methods to mutual fund holding data collected from the CRSP Mutual Funds database and the stock prices from Yahoo Finance in 2021 and 2022. Most mutual funds have a variety of stocks and derivatives in their portfolios. The percentage of total net assets allocated to the stocks in a portfolio shows the fund manager's views on their expected future returns. If the percentage of asset A is higher than asset B in a portfolio, it is an indication that the fund manager ranks asset A higher than asset B. As a result, the holding information of the mutual funds provides us with pairwise comparisons between the two assets. Although there are a lot of financial assets such as stocks and derivatives on the market, we concentrate on the stocks in the S&P500 list.

Since the returns of the portfolios reflect the quality of the comparisons, we only consider those portfolios that outperform their peers. At any time t, we look at the returns of all

the funds over the past two years. Those funds with the top 5% returns among all the funds are selected. This selection reflects fund managers actually have the stock-picking ability (we selected the top 25% and got similar results). Then we focus on the holding information of the portfolios corresponding to these funds (approximately, 1400 funds). In other words, we only consider the portfolios that have performed well over the last two years since they are more likely to produce more accurate comparison results. We then collect the pairwise comparison results for the S&P500 stocks in these selected portfolios. In specific, if the percentage of stock A is higher than stock B in a portfolio, stock A is preferred, and the comparison result is discarded if their percentages are the same. Moreover, among all constituents in the S&P500, only stocks that are compared for at least 5 times are kept. We consider the following three covariates: the log returns over the past month and the log returns over the past year, which quantify the short-term and long-term performances of this stock, respectively. The third covariate is the weighted percentage of holdings of the stock, calculated from all the selected portfolios that contain this stock. In specific, letting Portfolios(i) be the set of selected portfolios that contain stock i, the third covariate of stock i is calculated as the weighted percentage defined by

$$\frac{1}{\sum\limits_{\text{q}\in\text{Portfolios(i)}} |\text{q}|} \sum_{\text{q}\in\text{Portfolios(i)}} |\text{q}| \times \text{percentage of total net assets of stock } i \text{ in q},$$

where |q| is the total number of assets in portfolio q.

Since each portfolio's holdings do not change much in a short period (such as one or two months), we select three time points in the past two years as representatives to analyze, and the intervals between these time points are roughly half a year apart. Specifically, we focus on the portfolio holdings recorded on May 10, 2021, October 21, 2021, and April 7, 2022, with 1415, 1417, and 1422 funds respectively chosen for these three time points. Based on the information in these funds, there are 332, 334, and 334 S&P500 stocks, respectively, that were compared at least 5 times. At each time point given above, we calculate the MLE estimator $\widetilde{\boldsymbol{\beta}}_M$ in (3), and the implementation details are deferred to Appendix E.

We first investigate the statistical significance of these 3 variables at each time point. This amounts to testing the following hypothesis testing problems for each feature:

$$H_0 : \beta_i^* = 0 \quad \text{v.s.} \quad H_a : \beta_i^* \neq 0,\ i \in [3].$$

The test statistics are given by $\widehat{\beta}_{M,i}/\sqrt{([\mathcal{P}\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}_M)\mathcal{P}]^+)_{n+i,n+i}}$ for all $i \in [3]$ and the corresponding p-values are calculated via the asymptotic normality results in §4. The results are depicted in Table 5, where most of these three variables are statistically significant at each given time point.

We next turn to compare our model with the original BTL model in terms of predicting future returns. We consider the following two estimators as the representatives of ranking

	one month return	one year return	weighted percentage
May 10, 2021	<1e-5	<1e-5	<1e-5
Oct. 21, 2021	<1e-5	<1e-5	4.49e-3
April 7, 2022	3.76e-3	2.49e-3	5.82e-1

Table 5: p-values for the testing significance of three defined variables on the scores of ranking

scores derived from our model:

$$\theta_i^{\mathrm{CARE},1} = \widetilde{\boldsymbol{x}}_i^\top \widehat{\boldsymbol{\beta}}_M;$$
$$\theta_i^{\mathrm{CARE},2} = \mathtt{SOFT}(\widehat{\alpha}_{M,i}, \tau_i) + \widetilde{\boldsymbol{x}}_i^\top \widehat{\boldsymbol{\beta}}_M.$$

In specific, in $\theta_i^{\mathrm{CARE},1}$ we simply set the $\boldsymbol{\alpha}$ part to 0 as if the scores were completely captured by the covariates, which corresponds to $\theta_i^{\mathrm{CARE},2}$ with $\tau_i = \infty$. In $\theta_i^{\mathrm{CARE},2}$ we apply a soft-thresholding to the $\boldsymbol{\alpha}$ part with $\mathtt{SOFT}(x, \tau) = \mathrm{sign}(x) \cdot \max\{|x| - \tau, 0\}$ and $\tau_i = \Phi^{-1}(0.995) \cdot \sqrt{([\mathcal{P}\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}_M)\mathcal{P}]^+)_{i,i}}$ for each item i. This corresponds to set those estimates $\widehat{\alpha}_i$ that are statistically indifferent from zero to zero. The use of significant level 0.005 is to control the familywise false positive rates for hundreds of stocks. We then generate the ranking results $R_i^{\mathrm{CARE},1}$ and $R_i^{\mathrm{CARE},2}$ for the stocks according to $\theta_i^{\mathrm{CARE},1}$ and $\theta_i^{\mathrm{CARE},2}$. In addition, we also let R_i^{BTL} be the ranking result given by the ranking scores of the original BTL model.

To see if the preference ranking of stocks has better performance, we compute the average log-returns of the top k stocks and bottom k stocks for the subsequent month for each ranking method. The average log-returns of the top k stocks and bottom k stocks for different k ranging from 30 to n is presented in Figure 4. It is observed that the ranking results $R_i^{\mathrm{CARE},1}$ and $R_i^{\mathrm{CARE},2}$ given by our method achieve higher log-returns for the top k stocks and lower log-returns for the bottom k stocks. This implies that our model predicts future returns better than the original BTL model.

6 Conclusion and Discussion

In this paper, we study a Covariate-Assisted Ranking Estimation (CARE) model systematically. This allows us to incorporate the covariate information of compared items into the ranking framework, which includes the standard BTL model as a particular case. We derive the minimal sample complexity required for statistical consistency and uncertainty quantification for MLE based on novel proof techniques and illustrate the theory and methods using the mutual fund holding data set. The empirical results lend further support to the CARE model over the classical BTL model.

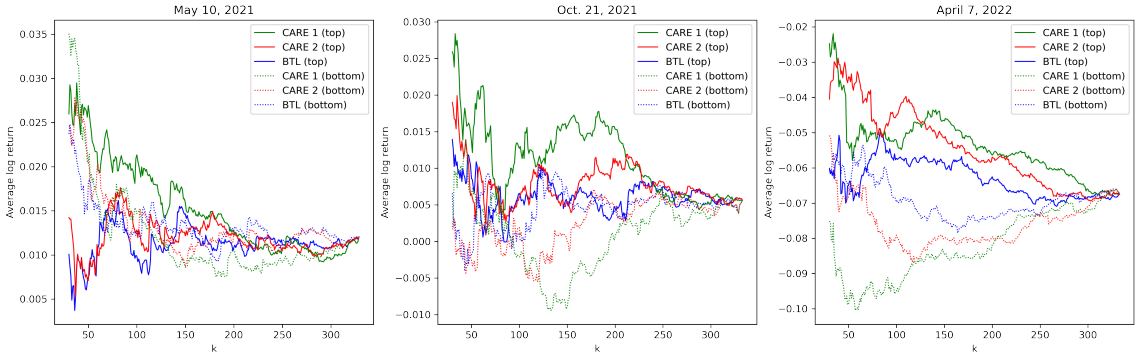

Figure 4: Average log return of the top k stocks and bottom k stocks given by the ranking results $R_i^{\mathrm{CARE},1}$, $R_i^{\mathrm{CARE},2}$ and R_i^{BTL}.

There are a few future directions worth exploring. First, it is worth extending the idea of incorporating covariates into a more general ranking framework, such as the Plackett-Luce or nonparametric models, under a more general comparison graph. In contrast, our work only studies the BTL model with the Erdős-Rényi comparison graph. Second, it would be interesting if some structure assumptions exist on the parameters $\{\alpha_i^*\}_{i=1}^n$ and $\boldsymbol{\beta}^*$, such as sparsity. In this scenario, one shall leverage certain regularizers on $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ in the likelihood function to achieve a solution that generalizes well. Third, except for the covariate, one may also incorporate time information into the ranking framework as in many real applications, the underlying scores of compared items change over time. Lastly, in our paper, we consider the scenario where the underlying score of the i-th item is given by $\alpha_i^* + \mathbf{x}_i^\top \boldsymbol{\beta}^*, i \in [n]$, in the sense that the overall score of the i-th item is the summation of its intrinsic score $\{\alpha_i^*\}_{i=1}^n$ and its covariate times one specific evaluation criterion $\boldsymbol{\beta}$. It would be interesting if we do a ranking based on data evaluated from multiple sources. In specific, suppose that we have L users and n items and the score of the i-th item, $i \in [n]$ evaluated by the ℓ-th person, $\ell \in [L]$, is $\alpha_i^* + \mathbf{x}_i^\top \boldsymbol{\beta}_\ell$. It would be interesting to derive novel statistical estimation and uncertainty quantification principles for ranking models under this setting. We will leave these open problems for future research.

Acknowledgments and Disclosure of Funding

Research supported by the NSF grants DMS-2052926, DMS-2053832, DMS-2210833, and ONR N00014-22-1-2340.

Supplementary materials to "Uncertainty Quantification of MLE for Entity Ranking with Covariates"

Appendix A. Proof Outline of Estimation Results

In this section, we first present the proof outline for Theorem 4. The detailed proof of Theorem 4 is presented in §A.5.

To understand statistical error of $\widetilde{\boldsymbol{\beta}}_M$, we begin with analyzing the regularized MLE $\widetilde{\boldsymbol{\beta}}_\lambda$

$$\widetilde{\boldsymbol{\beta}}_\lambda = \underset{\widetilde{\boldsymbol{\beta}}\in\Theta}{\operatorname{argmin}}\, \mathcal{L}_\lambda(\widetilde{\boldsymbol{\beta}}), \tag{13}$$

where $\mathcal{L}_\lambda(\widetilde{\boldsymbol{\beta}}) := \mathcal{L}(\widetilde{\boldsymbol{\beta}}) + \frac{\lambda}{2}\|\widetilde{\boldsymbol{\beta}}\|_2^2$ for $\lambda > 0$, and then make connections with $\widetilde{\boldsymbol{\beta}}_M$ by a properly chosen λ. The introduction of ℓ_2- regularization as a intermediate step is essential for ensuring the MLE fall in a bounded area around the ground truth through this regularized MLE. Therefore, strong convexity of the loss holds in this bounded area. Prevailing methodologies for examining the BTL model (Chen et al. (2019, Theorem 6) and Chen et al. (2022b, Lemma 8.5)) also relies on this ℓ_2 regularization to ensure strong convexity of the loss.

Before proceeding, we introduce the following two quantities κ_2 and κ_3 indicating the difficulty of recovering $\widetilde{\boldsymbol{\beta}}^*$

$$\kappa_2 := \max_{i\in[n]} |\alpha_i^*|, \quad \kappa_3 := \frac{\left\|\widetilde{\boldsymbol{\beta}}^*\right\|_2}{\sqrt{n(d+1)}}.$$

The regularized MLE solves a strong convex problem whose estimation error bounds are derived in Theorem 13 below.

Theorem 13. *Suppose $np > c_p \log n$ for some $c_p > 0$ and $d < n, d\log n \lesssim np$. We consider $L \le c_4 \cdot n^{c_5}$ for any absolute constants $c_4, c_5 > 0$ and*

$$\lambda = c_\lambda \min\left\{\frac{\kappa_1}{\kappa_2}, \frac{1}{\kappa_3\sqrt{d+1}}\right\}\sqrt{\frac{np\log n}{L}}$$

for some $c_\lambda > 0$. Let $\widetilde{\boldsymbol{\beta}}_\lambda = (\widehat{\boldsymbol{\alpha}}_\lambda^\top, \widehat{\boldsymbol{\beta}}_\lambda^\top)^\top$ be the solution of the regularized MLE Eq. (13). Then with probability at least $1 - O(n^{-6})$, we have

$$\|\widehat{\boldsymbol{\alpha}}_\lambda - \boldsymbol{\alpha}^*\|_\infty \lesssim \kappa_1^2\sqrt{\frac{(d+1)\log n}{npL}}; \quad \left\|\widehat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}^*\right\|_2 \lesssim \kappa_1\sqrt{\frac{\log n}{pL}};$$

$$\left\|\widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{\beta}}_\lambda - \widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{\beta}}^*\right\|_\infty \lesssim \kappa_1^2\sqrt{\frac{(d+1)\log n}{npL}}; \quad \frac{\left\|e^{\widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{\beta}}_\lambda} - e^{\widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{\beta}}^*}\right\|_\infty}{\left\|e^{\widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{\beta}}^*}\right\|_\infty} \lesssim \kappa_1^2\sqrt{\frac{(d+1)\log n}{npL}},$$

where $\widetilde{\boldsymbol{X}} = [\widetilde{\boldsymbol{x}}_1, \widetilde{\boldsymbol{x}}_2, \ldots, \widetilde{\boldsymbol{x}}_n]^\top$.

Theorem 13 presents the statistical rates of the regularized MLE in (13). Before presenting the formal proof for Theorem 13, in the following several subsections A.1, A.2 and A.3 , we focus on establishing its associated building blocks.

Given that our final objective is to establish the statistical rates of the Maximum Likelihood Estimator (MLE) applied to the unregularized loss function (3), we present a formal proof of Theorem 4 in §A.5, leveraging the insights provided by Theorem 13.

A.1 Preliminaries and Basic Results

In this subsection, we study the theoretical properties of the gradient and Hessian of the loss function in (13). Their expressions are given by

$$\nabla \mathcal{L}(\widetilde{\boldsymbol{\beta}}) = \sum_{(i,j)\in\mathcal{E}, i>j} \left\{ -y_{j,i} + \frac{e^{\widetilde{\boldsymbol{x}}_i^\top \widetilde{\boldsymbol{\beta}}}}{e^{\widetilde{\boldsymbol{x}}_i^\top \widetilde{\boldsymbol{\beta}}} + e^{\widetilde{\boldsymbol{x}}_j^\top \widetilde{\boldsymbol{\beta}}}} \right\} (\widetilde{\boldsymbol{x}}_i - \widetilde{\boldsymbol{x}}_j), \tag{14}$$

$$\nabla^2 \mathcal{L}(\widetilde{\boldsymbol{\beta}}) = \sum_{(i,j)\in\mathcal{E}, i>j} \frac{e^{\widetilde{\boldsymbol{x}}_i^\top \widetilde{\boldsymbol{\beta}}} e^{\widetilde{\boldsymbol{x}}_j^\top \widetilde{\boldsymbol{\beta}}}}{\left(e^{\widetilde{\boldsymbol{x}}_i^\top \widetilde{\boldsymbol{\beta}}} + e^{\widetilde{\boldsymbol{x}}_j^\top \widetilde{\boldsymbol{\beta}}}\right)^2} (\widetilde{\boldsymbol{x}}_i - \widetilde{\boldsymbol{x}}_j)(\widetilde{\boldsymbol{x}}_i - \widetilde{\boldsymbol{x}}_j)^\top. \tag{15}$$

The gradient of $\mathcal{L}(\widetilde{\boldsymbol{\beta}})$ at $\widetilde{\boldsymbol{\beta}}^*$ is controlled by the following lemma.

Lemma 14. *With λ given by Theorem 13, the following event*

$$\mathcal{A}_1 = \left\{ \left\| \nabla \mathcal{L}_\lambda \left(\widetilde{\boldsymbol{\beta}}^* \right) \right\|_2 \leq C_0 \sqrt{\frac{n^2 p \log n}{L}} \right\}$$

happens with probability exceeding $1 - O(n^{-11})$ for some $C_0 > 0$ which only depend on c_λ.

Proof For the proof of Lemma 14, we refer to §C.2 for more details. ■

Next, we proceed to analyzing the Hessian matrix $\nabla^2 \mathcal{L}_\lambda(\widetilde{\boldsymbol{\beta}})$. First, we consider $\boldsymbol{L}_\mathcal{G} = \sum_{(i,j)\in\mathcal{E}, i>j} (\widetilde{\boldsymbol{x}}_i - \widetilde{\boldsymbol{x}}_j)(\widetilde{\boldsymbol{x}}_i - \widetilde{\boldsymbol{x}}_j)^\top$ and study its eigenvalues in Lemma 15.

Lemma 15. *Suppose $pn > c_p \log n$ for some $c_p > 0$. The following event*

$$\mathcal{A}_2 = \left\{ \frac{1}{2} c_2 pn \leq \lambda_{min,\perp}(\boldsymbol{L}_\mathcal{G}) \leq \|\boldsymbol{L}_\mathcal{G}\| \leq 2 c_1 pn \right\}$$

happens with probability exceeding $1 - O(n^{-11})$ when n is large enough.

Proof See §C.3 for a detailed proof. ■

In the rest of the content, without loss of generality, we assume the conditions stated in Lemma 15 hold. Moreover, with the help of Lemma 15, we next analyze the Hessian $\nabla^2 \mathcal{L}_\lambda(\widetilde{\boldsymbol{\beta}})$ and summarize its theoretical properties in Lemma 16 and Lemma 17, respectively.

Lemma 16. *Suppose event $\mathcal{A}_2$ holds, we obtain*

$$\lambda_{max}(\nabla^2 \mathcal{L}_\lambda(\widetilde{\boldsymbol{\beta}})) \leq \lambda + \frac{1}{2} c_1 pn, \quad \forall \widetilde{\boldsymbol{\beta}} \in \mathbb{R}^{n+d}.$$

Proof Since $\frac{e^{\widetilde{\boldsymbol{x}}_i^\top \widetilde{\boldsymbol{\beta}}} e^{\widetilde{\boldsymbol{x}}_j^\top \widetilde{\boldsymbol{\beta}}}}{\left(e^{\widetilde{\boldsymbol{x}}_i^\top \widetilde{\boldsymbol{\beta}}} + e^{\widetilde{\boldsymbol{x}}_j^\top \widetilde{\boldsymbol{\beta}}}\right)^2} \leq \frac{1}{4}$, we have

$$\lambda_{\max}(\nabla^2 \mathcal{L}_\lambda(\widetilde{\boldsymbol{\beta}})) \leq \lambda + \frac{1}{4}\|\boldsymbol{L}_{\mathcal{G}}\| \leq \lambda + \frac{1}{2}c_1 pn, \quad \forall \widetilde{\boldsymbol{\beta}} \in \mathbb{R}^{n+d}.$$

■

Lemma 17. *Suppose event $\mathcal{A}_2$ happens. Then for all $\widetilde{\boldsymbol{\beta}}$ such that $\|\boldsymbol{\alpha} - \boldsymbol{\alpha}^*\|_\infty \leq C_1$, $\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 \leq C_2$, we have*

$$\lambda_{min,\perp}(\nabla^2 \mathcal{L}_\lambda(\widetilde{\boldsymbol{\beta}})) \geq \lambda + \frac{c_2 pn}{8\kappa_1 e^C},$$

where $C = 2C_1 + 2\sqrt{\frac{c_3(d+1)}{n}}C_2$.

Proof The formal proof of this Lemma 17 can be found in §C.4. ■

In the following two subsections §A.2 and §A.3, we understand the statistical rates of the regularized estimator $\widetilde{\boldsymbol{\beta}}_\lambda$ via analyzing the iterates in our gradient method.

A.2 Convergence of Projected Gradient Descent

In this subsection, we consider a sequence of iterates $\{\widetilde{\boldsymbol{\beta}}^t\}_{t=0,1,\ldots}$ which is generated by the following projected gradient descent algorithm with stepsize $\eta = \frac{2}{2\lambda + c_1 np}$ and the number of iterations $T = n^5$.

Algorithm 1 Gradient descent for regularized MLE.

Initialize $\widetilde{\boldsymbol{\beta}}^0 = \widetilde{\boldsymbol{\beta}}^*$
for $t = 0, 1, \ldots, T-1$ **do**
  $\zeta = \widetilde{\boldsymbol{\beta}}^t - \eta \nabla \mathcal{L}_\lambda(\widetilde{\boldsymbol{\beta}}^t)$
  $\widetilde{\boldsymbol{\beta}}^{t+1} = \mathcal{P}\zeta$
end for

We initialize at $\widetilde{\boldsymbol{\beta}}^*$ and the target loss given in (13) is strongly convex. The projected gradient descent is employed (since the likelihood has a linear constraint) to ensure the iterates $\widetilde{\boldsymbol{\beta}}^t$ converge to the ℓ_2-regularized MLE $\widetilde{\boldsymbol{\beta}}_\lambda$ exponentially. In the following section, using the leave-one-out analysis, we also show $\widetilde{\boldsymbol{\beta}}^t$ at the same time stays close to $\widetilde{\boldsymbol{\beta}}^*$ for all $t \leq \mathbf{poly}(n)$. Therefore, combine these two parts together, we are able to conclude that $\widetilde{\boldsymbol{\beta}}_\lambda$ is also close to $\widetilde{\boldsymbol{\beta}}^*$.

We summarize the theoretical findings in the following Lemma 18, Lemma 19 and Lemma 20, respectively.

Lemma 18. *Under event $\mathcal{A}_2$, we have*

$$\left\|\widetilde{\boldsymbol{\beta}}^t - \widetilde{\boldsymbol{\beta}}_\lambda\right\|_2 \le \rho^t \left\|\widetilde{\boldsymbol{\beta}}^0 - \widetilde{\boldsymbol{\beta}}_\lambda\right\|_2,$$

where $\rho = 1 - \dfrac{2\lambda}{2\lambda + c_1 np}$.

Next, we prove that the initial point is not far from $\widetilde{\boldsymbol{\beta}}_\lambda$.

Lemma 19. *On the event $\mathcal{A}_1$ happens, it follows that*

$$\left\|\widetilde{\boldsymbol{\beta}}^0 - \widetilde{\boldsymbol{\beta}}_\lambda\right\|_2 = \left\|\widetilde{\boldsymbol{\beta}}_\lambda - \widetilde{\boldsymbol{\beta}}^*\right\|_2 \le \frac{2C_0}{c_\lambda} \max\left\{\frac{\kappa_2}{\kappa_1}, \kappa_3\sqrt{d+1}\right\}\sqrt{n}.$$

Proof See §C.6 for a detailed proof. ■

Combining Lemma 18 and Lemma 19, we obtain the following result on the optimization error.

Lemma 20. *On event $\mathcal{A}_1 \cap \mathcal{A}_2$, there exists some constant C_7 such that*

$$\left\|\widetilde{\boldsymbol{\beta}}^T - \widetilde{\boldsymbol{\beta}}_\lambda\right\|_2 \le C_7\kappa_1\sqrt{\frac{(d+1)\log n}{npL}}.$$

Proof See §C.7 for a detailed proof. ■

In this subsection, we prove that the iterate $\widetilde{\boldsymbol{\beta}}^T$ converges to $\widetilde{\boldsymbol{\beta}}_\lambda$ geometrically and enjoys a good optimization error after $T = n^5$ iterations. In order to prove the distance between $\widetilde{\boldsymbol{\beta}}_\lambda$ and $\widetilde{\boldsymbol{\beta}}^*$, i.e. the statistical error of $\widetilde{\boldsymbol{\beta}}_\lambda$, in the next subsection, we leverage the leave-one-out technique and use induction to prove that the iterate $\widetilde{\boldsymbol{\beta}}^T$ stays close to our initial point $\widetilde{\boldsymbol{\beta}}^0 = \widetilde{\boldsymbol{\beta}}^*$, even after $T = n^5$ iterations.

A.3 Analysis of Leave-one-out Sequences

In this section, we construct the leave-one-out sequences (Ma et al., 2018; Chen et al., 2019, 2020) and bound the statistical error by induction. To construct the leave-one-out sequence, we consider the following loss function for any $m \in [n]$.

$$\begin{aligned}
\mathcal{L}^{(m)}(\widetilde{\boldsymbol{\beta}}) = &\sum_{(i,j)\in\mathcal{E}, i>j, i\neq m, j\neq m} \left\{-y_{j,i}\left(\widetilde{\boldsymbol{x}}_i^\top\widetilde{\boldsymbol{\beta}} - \widetilde{\boldsymbol{x}}_j^\top\widetilde{\boldsymbol{\beta}}\right) + \log\left(1 + e^{\widetilde{\boldsymbol{x}}_i^\top\widetilde{\boldsymbol{\beta}} - \widetilde{\boldsymbol{x}}_j^\top\widetilde{\boldsymbol{\beta}}}\right)\right\} \\
&+ p\sum_{i\neq m}\left\{-\frac{e^{\widetilde{\boldsymbol{x}}_i^\top\widetilde{\boldsymbol{\beta}}^*}}{e^{\widetilde{\boldsymbol{x}}_i^\top\widetilde{\boldsymbol{\beta}}^*} + e^{\widetilde{\boldsymbol{x}}_m^\top\widetilde{\boldsymbol{\beta}}^*}}(\widetilde{\boldsymbol{x}}_i^\top\widetilde{\boldsymbol{\beta}} - \widetilde{\boldsymbol{x}}_m^\top\widetilde{\boldsymbol{\beta}}) + \log\left(1 + e^{\widetilde{\boldsymbol{x}}_i^\top\widetilde{\boldsymbol{\beta}} - \widetilde{\boldsymbol{x}}_m^\top\widetilde{\boldsymbol{\beta}}}\right)\right\}; \\
\mathcal{L}_\lambda^{(m)}(\widetilde{\boldsymbol{\beta}}) = &\mathcal{L}^{(m)}(\widetilde{\boldsymbol{\beta}}) + \frac{\lambda}{2}\|\widetilde{\boldsymbol{\beta}}\|_2^2.
\end{aligned}$$

Then for any $m \in [n]$, we construct the leave-one-out sequence $\left\{\widetilde{\boldsymbol{\beta}}^{t,(m)}\right\}_{t=0,1,\ldots}$ in the way of Algorithm 2.

Algorithm 2 Construction of leave-one-out sequences.

1: **Initialize** $\widetilde{\boldsymbol{\beta}}^{0,(m)} = \widetilde{\boldsymbol{\beta}}^*$
2: **for** $t = 0, 1, \ldots, T-1$ **do**
3: $\quad \zeta = \widetilde{\boldsymbol{\beta}}^{t,(m)} - \eta \nabla \mathcal{L}_\lambda^{(m)} \left(\widetilde{\boldsymbol{\beta}}^{t,(m)} \right)$
4: $\quad \widetilde{\boldsymbol{\beta}}^{t+1,(m)} = \mathcal{P}\zeta$
5: **end for**

With the help of the leave-one-out sequences, we do induction to demonstrate that the iterate $\widetilde{\boldsymbol{\beta}}^T$ will not be far away from $\widetilde{\boldsymbol{\beta}}^*$ when $T = n^5$. In specific, we take again $\eta = \frac{2}{2\lambda + c_1 np}$ and $T = n^5$. With the leave-one-out sequences in hand, we prove the following bounds by induction for $t \le T$.

$$\left\| \widetilde{\boldsymbol{\beta}}^t - \widetilde{\boldsymbol{\beta}}^* \right\|_2 \le C_3 \kappa_1 \sqrt{\frac{\log n}{pL}}; \tag{A}$$

$$\max_{1 \le m \le n} \left\| \widetilde{\boldsymbol{\beta}}^t - \widetilde{\boldsymbol{\beta}}^{t,(m)} \right\|_2 \le C_4 \kappa_1 \sqrt{\frac{(d+1)\log n}{npL}} \le C_4 \kappa_1 \sqrt{\frac{\log n}{pL}}; \tag{B}$$

$$\max_{1 \le m \le n} |\alpha_m^{t,(m)} - \alpha_m^*| \le C_5 \kappa_1^2 \sqrt{\frac{(d+1)\log n}{npL}}; \tag{C}$$

$$\|\boldsymbol{\alpha}^t - \boldsymbol{\alpha}^*\|_\infty \le C_6 \kappa_1^2 \sqrt{\frac{(d+1)\log n}{npL}}. \tag{D}$$

For $t = 0$, since $\widetilde{\boldsymbol{\beta}}^0 = \widetilde{\boldsymbol{\beta}}^{0,(1)} = \widetilde{\boldsymbol{\beta}}^{0,(2)} = \cdots = \widetilde{\boldsymbol{\beta}}^{0,(n)} = \widetilde{\boldsymbol{\beta}}^*$, the (A)$\sim$ (D) hold automatically. In the following lemmas, we prove the conclusions of (A)-(D) for the $(t+1)$-th iteration are true when the results hold for the t-th iteration.

Lemma 21. *Suppose bounds* (A)$\sim$ (D) *hold for the t-th iteration. With probability exceeding $1 - O(n^{-11})$ we have*

$$\left\| \widetilde{\boldsymbol{\beta}}^{t+1} - \widetilde{\boldsymbol{\beta}}^* \right\|_2 \le C_3 \kappa_1 \sqrt{\frac{\log n}{pL}},$$

as long as $0 < \eta \le \dfrac{2}{2\lambda + c_1 np}$, $C_3 \ge \dfrac{20 C_0}{c_2}$ and n is large enough.

Proof See §C.8 for a detailed proof. ■

Lemma 22. *Suppose bounds* (A)$\sim$ (D) *hold for the t-th iteration. With probability exceeding $1 - O(n^{-11})$ we have*

$$\max_{1 \le m \le n} \left\| \widetilde{\boldsymbol{\beta}}^{t+1} - \widetilde{\boldsymbol{\beta}}^{t+1,(m)} \right\|_2 \le C_4 \kappa_1 \sqrt{\frac{(d+1)\log n}{npL}},$$

as long as $0 < \eta \le \dfrac{2}{2\lambda + c_1 np}$, $C_4 \gtrsim \dfrac{1}{c_2}$ *and* $np \gtrsim (d+1)\log n$.

Proof See §C.9 for a detailed proof. ■

Lemma 23. *Suppose bounds* (A)∼ (D) *hold for the* t*-th iteration. With probability exceeding* $1 - O(n^{-11})$ *we have*

$$\max_{1\le m\le n} |\alpha_m^{t+1,(m)} - \alpha_m^*| \le C_5 \kappa_1^2 \sqrt{\frac{(d+1)\log n}{npL}},$$

as long as $C_5 \ge 30c_0(C_0 + c_1C_3 + c_1C_4)$, $C_5 \ge 7.5(1+2\sqrt{c_3})(C_3+C_4)$, $C_5 \ge 30c_\lambda/\sqrt{d+1}$ *and* n *is large enough.*

Proof See §C.10 for a detailed proof. ■

Lemma 24. *Suppose bounds* (A)∼ (D) *hold for the* t*-th iteration. With probability exceeding* $1 - O(n^{-11})$ *we have*

$$\|\boldsymbol{\alpha}^{t+1} - \boldsymbol{\alpha}^*\|_\infty \le C_6 \kappa_1^2 \sqrt{\frac{(d+1)\log n}{npL}},$$

as long as $C_6 \ge C_4 + C_5$ *and* n *is large enough.*

Proof See §C.11 for a detailed proof. ■

With these necessary building blocks at hand, we next prove Theorem 13 by providing statistical rates of the regularized estimator in (13).

A.4 Proof of Theorem 13

In this subsection, we aim at providing proof for Theorem 13 by combining the results established above.

For any $m \in [n]$, by Taylor expansion, one obtains

$$\begin{aligned}\left|e^{\widetilde{\boldsymbol{x}}_m^\top \widetilde{\boldsymbol{\beta}}_\lambda} - e^{\widetilde{\boldsymbol{x}}_m^\top \widetilde{\boldsymbol{\beta}}^*}\right| &\le e^{\omega_m^*} \left|\widetilde{\boldsymbol{x}}_m^\top \widetilde{\boldsymbol{\beta}}_\lambda - \widetilde{\boldsymbol{x}}_m^\top \widetilde{\boldsymbol{\beta}}^*\right|. \\ &\le e^{\widetilde{\boldsymbol{x}}_m^\top \widetilde{\boldsymbol{\beta}}^* + \left|\widetilde{\boldsymbol{x}}_m^\top \widetilde{\boldsymbol{\beta}}_\lambda - \widetilde{\boldsymbol{x}}_m^\top \widetilde{\boldsymbol{\beta}}^*\right|} \left|\widetilde{\boldsymbol{x}}_m^\top \widetilde{\boldsymbol{\beta}}_\lambda - \widetilde{\boldsymbol{x}}_m^\top \widetilde{\boldsymbol{\beta}}^*\right|\end{aligned}$$

where ω_m^* is some real number between $\widetilde{\boldsymbol{x}}_m^\top \widetilde{\boldsymbol{\beta}}_\lambda$ and $\widetilde{\boldsymbol{x}}_m^\top \widetilde{\boldsymbol{\beta}}^*$. As a result, it holds that

$$\begin{aligned}\frac{\left\|e^{\widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{\beta}}_\lambda} - e^{\widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{\beta}}^*}\right\|_\infty}{\left\|e^{\widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{\beta}}^*}\right\|_\infty} &\le \frac{\max_{1\le m\le n} e^{\widetilde{\boldsymbol{x}}_m^\top \widetilde{\boldsymbol{\beta}}^*} e^{\max_{1\le m\le n}\left|\widetilde{\boldsymbol{x}}_m^\top \widetilde{\boldsymbol{\beta}}_\lambda - \widetilde{\boldsymbol{x}}_m^\top \widetilde{\boldsymbol{\beta}}^*\right|} \max_{1\le m\le n}\left|\widetilde{\boldsymbol{x}}_m^\top \widetilde{\boldsymbol{\beta}}_\lambda - \widetilde{\boldsymbol{x}}_m^\top \widetilde{\boldsymbol{\beta}}^*\right|}{\max_{1\le m\le n} e^{\widetilde{\boldsymbol{x}}_m^\top \widetilde{\boldsymbol{\beta}}^*}} \\ &\le e^{\max_{1\le m\le n}\left|\widetilde{\boldsymbol{x}}_m^\top \widetilde{\boldsymbol{\beta}}_\lambda - \widetilde{\boldsymbol{x}}_m^\top \widetilde{\boldsymbol{\beta}}^*\right|} \max_{1\le m\le n}\left|\widetilde{\boldsymbol{x}}_m^\top \widetilde{\boldsymbol{\beta}}_\lambda - \widetilde{\boldsymbol{x}}_m^\top \widetilde{\boldsymbol{\beta}}^*\right|.\end{aligned}$$

By the Cauchy-Schwartz inequality, we have

$$\begin{aligned}\left|\widetilde{\boldsymbol{x}}_m^\top\widetilde{\boldsymbol{\beta}}_\lambda-\widetilde{\boldsymbol{x}}_m^\top\widetilde{\boldsymbol{\beta}}^*\right| &\le |(\widehat{\alpha}_\lambda)_m-\alpha_m^*|+\left|\boldsymbol{x}_m^\top\widehat{\boldsymbol{\beta}}_\lambda-\boldsymbol{x}_m^\top\boldsymbol{\beta}^*\right|\\ &\le \|\widehat{\boldsymbol{\alpha}}_\lambda-\boldsymbol{\alpha}^*\|_\infty+\|\boldsymbol{x}_m\|_2\|\widehat{\boldsymbol{\beta}}_\lambda-\boldsymbol{\beta}^*\|_2\\ &\le \|\boldsymbol{\alpha}^T-\boldsymbol{\alpha}^*\|_\infty+\|\widehat{\boldsymbol{\alpha}}_\lambda-\boldsymbol{\alpha}^T\|_\infty+\|\boldsymbol{x}_m\|_2\|\boldsymbol{\beta}^T-\boldsymbol{\beta}^*\|_2+\|\boldsymbol{x}_m\|_2\|\widehat{\boldsymbol{\beta}}_\lambda-\boldsymbol{\beta}^T\|_2\\ &\le \|\boldsymbol{\alpha}^T-\boldsymbol{\alpha}^*\|_\infty+\left\|\widetilde{\boldsymbol{\beta}}_\lambda-\widetilde{\boldsymbol{\beta}}^T\right\|_2+\|\boldsymbol{x}_m\|_2\|\boldsymbol{\beta}^T-\boldsymbol{\beta}^*\|_2+\|\boldsymbol{x}_m\|_2\left\|\widetilde{\boldsymbol{\beta}}_\lambda-\widetilde{\boldsymbol{\beta}}^T\right\|_2\\ &\le C_8\kappa_1^2\sqrt{\frac{(d+1)\log n}{npL}},\end{aligned}$$

where $C_8\ge C_6+(1+\sqrt{c_3})\,C_7+\sqrt{c_3}C_3$ and n is large enough. The last inequlity holds from the results derived in §A.2 and §A.3. Then for n and L such that $C_8\kappa_1^2\sqrt{\frac{(d+1)\log n}{npL}}\le 0.6$, we have

$$\frac{\left\|e^{\widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{\beta}}_\lambda}-e^{\widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{\beta}}^*}\right\|_\infty}{\left\|e^{\widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{\beta}}^*}\right\|_\infty}\le 2C_8\kappa_1^2\sqrt{\frac{(d+1)\log n}{npL}}.$$

Similarly, we also obtain

$$\begin{aligned}\|\widehat{\boldsymbol{\alpha}}_\lambda-\boldsymbol{\alpha}^*\|_\infty &\le \|\boldsymbol{\alpha}^T-\boldsymbol{\alpha}^*\|_\infty+\|\boldsymbol{\alpha}_\lambda-\boldsymbol{\alpha}^T\|_\infty\\ &\le C_6\kappa_1^2\sqrt{\frac{(d+1)\log n}{npL}}+C_7\kappa_1\sqrt{\frac{(d+1)\log n}{npL}}\\ &\lesssim \kappa_1^2\sqrt{\frac{(d+1)\log n}{npL}};\\ \|\widehat{\boldsymbol{\beta}}_\lambda-\boldsymbol{\beta}^*\|_2 &\le \|\boldsymbol{\beta}^T-\boldsymbol{\beta}^*\|_2+\|\boldsymbol{\beta}_\lambda-\boldsymbol{\beta}^T\|_2\\ &\le C_3\kappa_1\sqrt{\frac{\log n}{pL}}+C_7\kappa_1\sqrt{\frac{(d+1)\log n}{npL}}\lesssim \kappa_1\sqrt{\frac{\log n}{pL}}.\end{aligned}$$

Next, we use the proof ideas and conclusions from Theorem 13 to prove the statistical rate of a non-regularized MLE defined in (3).

A.5 Proof of Theorem 4

With all necessary building blocks at hand, in this subsection, we provide the formal proof for Theorem 4. In specific, we assume $L=O(n^2)$ in the following proof and it is easy to extend the proof to the regime $L\le c_4\cdot n^{c_5}$. The way to solve this is changing the power 11 in Lemma 14 and Lemma 15 to a larger number.

Proof Consider the following MLE

$$\widetilde{\boldsymbol{\beta}}_{\text{con}}:=\underset{\widetilde{\boldsymbol{\beta}}\in\Theta,\|\boldsymbol{\alpha}-\boldsymbol{\alpha}^*\|_\infty\le 0.025,\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2\le 0.025\sqrt{n/(2c_3d+2c_3)}}{\operatorname{argmin}}\mathcal{L}(\widetilde{\boldsymbol{\beta}}).\tag{16}$$

We choose c_λ in the definition of λ such that

$$\lambda n^3 \geq 1 \text{ and } \left(\kappa_2 n + \sqrt{n(d+1)}\kappa_3 + (C_3+C_7)\kappa_1\sqrt{\frac{\log n}{pL}}\right)\lambda \leq \frac{c_2 pn}{20} C_9\sqrt{\frac{\log n}{n^2 pL}}$$

for some $C_9 > 0$. As long as $L \leq c_4 \cdot n^{c_5}$ for some absolute constants $c_4, c_5 > 0$, the proof of Lemma 20 is still valid and Theorem 13 still holds for this λ. For n large enough such that $\widetilde{\boldsymbol{\beta}}_\lambda$ satisfies the constraints in Eq. (16), by the optimality of $\widetilde{\boldsymbol{\beta}}_{\text{con}}$ we know that $\mathcal{L}(\widetilde{\boldsymbol{\beta}}_\lambda) \geq \mathcal{L}(\widetilde{\boldsymbol{\beta}}_{\text{con}})$. On the other hand, by Taylor's expansion

$$\mathcal{L}(\widetilde{\boldsymbol{\beta}}_{\text{con}}) = \mathcal{L}(\widetilde{\boldsymbol{\beta}}_\lambda) + \nabla\mathcal{L}(\widetilde{\boldsymbol{\beta}}_\lambda)^\top\left(\widetilde{\boldsymbol{\beta}}_{\text{con}} - \widetilde{\boldsymbol{\beta}}_\lambda\right) + \frac{1}{2}\left(\widetilde{\boldsymbol{\beta}}_{\text{con}} - \widetilde{\boldsymbol{\beta}}_\lambda\right)^\top \nabla^2\mathcal{L}(\boldsymbol{c})\left(\widetilde{\boldsymbol{\beta}}_{\text{con}} - \widetilde{\boldsymbol{\beta}}_\lambda\right),$$

where $\boldsymbol{c} = \xi\widetilde{\boldsymbol{\beta}}_{\text{con}} + (1-\xi)\widetilde{\boldsymbol{\beta}}_\lambda$ for some $\xi \in [0,1]$.

This leads to

$$\nabla\mathcal{L}(\widetilde{\boldsymbol{\beta}}_\lambda)^\top\left(\widetilde{\boldsymbol{\beta}}_{\text{con}} - \widetilde{\boldsymbol{\beta}}_\lambda\right) + \frac{1}{2}\left(\widetilde{\boldsymbol{\beta}}_{\text{con}} - \widetilde{\boldsymbol{\beta}}_\lambda\right)^\top \nabla^2\mathcal{L}(\boldsymbol{c})\left(\widetilde{\boldsymbol{\beta}}_{\text{con}} - \widetilde{\boldsymbol{\beta}}_\lambda\right) \leq 0. \tag{17}$$

Next, we first define the norm $\|\cdot\|_c$ as

$$\left\|\widetilde{\boldsymbol{\beta}}\right\|_c := \max_{i\in[n]}\left|\widetilde{\boldsymbol{\beta}}_i\right| + \sqrt{\frac{c_3(d+1)}{n}}\sqrt{\sum_{j=n+1}^{n+d}\widetilde{\boldsymbol{\beta}}_j^2}, \qquad \forall\widetilde{\boldsymbol{\beta}} \in \mathbb{R}^{n+d}.$$

Therefore, we have

$$\|\boldsymbol{c} - \widetilde{\boldsymbol{\beta}}^*\|_c \leq \max\left\{\|\widetilde{\boldsymbol{\beta}}_\lambda - \widetilde{\boldsymbol{\beta}}^*\|_c, \|\widetilde{\boldsymbol{\beta}}_{\text{con}} - \widetilde{\boldsymbol{\beta}}^*\|_c\right\} \leq 0.1$$

as long as

$$\left[2(C_6+C_7)\kappa_1^2 + 2\sqrt{2c_3}(C_3+C_7)\kappa_1\right]\sqrt{\frac{(d+1)\log n}{npL}} \leq 0.1.$$

As a result, by Lemma 17 we have

$$\lambda_{\min,\perp}(\nabla^2\mathcal{L}(\boldsymbol{c})) \geq \lambda + \frac{c_2 pn}{8\kappa_1 e^C} \geq \frac{c_2 pn}{10\kappa_1}. \tag{18}$$

Combine Eq. (17) and Eq. (18) we have

$$\begin{aligned}\frac{c_2 pn}{20\kappa_1}\left\|\widetilde{\boldsymbol{\beta}}_{\text{con}} - \widetilde{\boldsymbol{\beta}}_\lambda\right\|_2^2 &\leq \frac{1}{2}\left(\widetilde{\boldsymbol{\beta}}_{\text{con}} - \widetilde{\boldsymbol{\beta}}_\lambda\right)^\top \nabla^2\mathcal{L}(\boldsymbol{c})\left(\widetilde{\boldsymbol{\beta}}_{\text{con}} - \widetilde{\boldsymbol{\beta}}_\lambda\right)\\ &\leq -\nabla\mathcal{L}(\widetilde{\boldsymbol{\beta}}_\lambda)^\top\left(\widetilde{\boldsymbol{\beta}}_{\text{con}} - \widetilde{\boldsymbol{\beta}}_\lambda\right)\\ &\leq \left\|\nabla\mathcal{L}(\widetilde{\boldsymbol{\beta}}_\lambda)\right\|_2\left\|\widetilde{\boldsymbol{\beta}}_{\text{con}} - \widetilde{\boldsymbol{\beta}}_\lambda\right\|_2.\end{aligned}$$

Therefore, after some simple calculations, it holds that

$$\left\|\widetilde{\boldsymbol{\beta}}_{\text{con}} - \widetilde{\boldsymbol{\beta}}_\lambda\right\|_2 \leq \frac{20\kappa_1}{c_2 pn}\left\|\nabla\mathcal{L}(\widetilde{\boldsymbol{\beta}}_\lambda)\right\|_2 \leq \frac{20\kappa_1\lambda}{c_2 pn}\left\|\widetilde{\boldsymbol{\beta}}_\lambda\right\|_2 \leq C_9\kappa_1\sqrt{\frac{\log n}{n^2 pL}}.$$

And, when n is large enough, we have

$$\|\widehat{\boldsymbol{\alpha}}_{\text{con}} - \boldsymbol{\alpha}^*\|_\infty \le \|\widehat{\boldsymbol{\alpha}}_{\text{con}} - \widehat{\boldsymbol{\alpha}}_\lambda\|_\infty + \|\widehat{\boldsymbol{\alpha}}_\lambda - \boldsymbol{\alpha}^*\|_\infty \le C_9\kappa_1\sqrt{\frac{\log n}{n^2pL}} + (C_6+C_7)\kappa_1^2\sqrt{\frac{(d+1)\log n}{npL}} \le 0.01, \tag{19}$$

$$\|\widehat{\boldsymbol{\beta}}_{\text{con}} - \boldsymbol{\beta}^*\|_2 \le \|\widehat{\boldsymbol{\beta}}_{\text{con}} - \widehat{\boldsymbol{\beta}}_\lambda\|_2 + \|\widehat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}^*\|_2 \le C_9\kappa_1\sqrt{\frac{\log n}{n^2pL}} + (C_3+C_7)\kappa_1\sqrt{\frac{\log n}{pL}} \le 0.01\sqrt{\frac{n}{2c_3(d+1)}}, \tag{20}$$

where $\widehat{\boldsymbol{\alpha}}_{\text{con}} := (\widetilde{\boldsymbol{\beta}}_{\text{con}})_{1:n}$ and $\widehat{\boldsymbol{\beta}}_{\text{con}} := (\widetilde{\boldsymbol{\beta}}_{\text{con}})_{n+1:n+d}$. As a result, $\widetilde{\boldsymbol{\beta}}_{\text{con}}$ falls in the interior of the inequality constraints in Eq. (16). By the convexity of $\mathcal{L}$ and its strong convexity in Θ, we have $\widetilde{\boldsymbol{\beta}}_{\text{con}} = \widetilde{\boldsymbol{\beta}}_M$. Therefore, by Eq. (19) and Eq. (20) we have

$$\|\widehat{\boldsymbol{\alpha}}_M - \boldsymbol{\alpha}^*\|_\infty = \|\widehat{\boldsymbol{\alpha}}_{\text{con}} - \boldsymbol{\alpha}^*\|_\infty \le C_{10}\kappa_1^2\sqrt{\frac{(d+1)\log n}{npL}}, \tag{21}$$

$$\|\widehat{\boldsymbol{\beta}}_M - \boldsymbol{\beta}^*\|_2 = \|\widehat{\boldsymbol{\beta}}_{\text{con}} - \boldsymbol{\beta}^*\|_2 \le C_{11}\kappa_1\sqrt{\frac{\log n}{pL}}. \tag{22}$$

The result of $\left\|\widetilde{\boldsymbol{X}}^\top\widetilde{\boldsymbol{\beta}}_M - \widetilde{\boldsymbol{X}}^\top\widetilde{\boldsymbol{\beta}}^*\right\|_\infty$ and $\frac{\left\|e^{\widetilde{\boldsymbol{X}}^\top\widetilde{\boldsymbol{\beta}}_M} - e^{\widetilde{\boldsymbol{X}}^\top\widetilde{\boldsymbol{\beta}}^*}\right\|_\infty}{\left\|e^{\widetilde{\boldsymbol{X}}^\top\widetilde{\boldsymbol{\beta}}^*}\right\|_\infty}$ hold based on the same derivations in Section A.4 and we omit the corresponding details. ■

Appendix B. Proof of Inference Results in Section 4

In this section we first introduce our extension to top-K testing and provide the proof outlines for Theorem 7. Then, we prove Theorem 10 and Corollaries 11 and 12, based on the results of Theorem 7. The details of proofs in this section is deferred to Section C. We next introduce some building blocks for the proof of main theorems.

B.1 Extension to Top-K set hypothesis testing

In this section, we extend the method to conduct the top-K test using estimators for overall scores $\theta_i^* := \alpha_i^* + \mathbf{x}_i^\top \beta^*, i \in [n]$ (for simplicity, here we define the overall scores as θ_i^* and its estimators as $\widehat{\theta}_i := \widehat{\alpha}_i + \mathbf{x}_i^\top \widehat{\beta}, i \in [n]$).

We use the following statistics to construct top-K test simultaneously for all elements in $\mathcal{M}$.

$$\mathcal{T} := \max_{m \in \mathcal{M}} \max_{k \neq m} \frac{\widehat{\theta}_k - \widehat{\theta}_m - (\theta_k^* - \theta_m^*)}{\widehat{\sigma}_{m,k}}$$

It's distribution can be approximated by the bootstrap counterpart

$$\mathcal{G} := \max_{m \in \mathcal{M}} \max_{k \neq m} \sum_{l=1}^{L} \sum_{(i,j) \in \mathcal{E}, i>j} \frac{(\mathbf{e}_m - \mathbf{e}_k)^\top (\nabla^2 \mathcal{L}(\widehat{\theta})^+)(\mathbf{e}_i - \mathbf{e}_j)}{\widehat{\sigma}_{m,k} L} (\phi(\widehat{\theta}_i - \widehat{\theta}_j) - y_{j,i}^{(l)}) \omega_{j,i}^{(l)}.$$

We are able to achieve similar theoretical results by following the similar proof procedure of Theorem 5 in Fan et al. (2023) for $\mathcal{T}$ and $\mathcal{G}$. Let $c_{1-\alpha}$ be the $(1-\alpha)$-th quantile of $\mathcal{G}$, we have the following theorem for the test statistics $\mathcal{T}$.

Theorem 25. *Under the conditions of Theorem 1, we have*

$$|P(\mathcal{T} > c_{1-\alpha}) - \alpha| \to 0.$$

Next, we introduce some applications on constructing (simultaneous) one-sided confidence intervals for out-of-sample ranks via the distribution of $\mathcal{T}$ in the following two examples.

Example 1. *For an item m of interest, and let K be the targeted rank threshold, we are interested in the following testing problem*

$$H_0 : r(m) \leq K \quad \text{versus} \quad H_1 : r(m) > K. \tag{23}$$

Let $\widehat{c}_{1-\alpha}$ be the estimated $(1-\alpha)$-th quantile of $\mathcal{T}$ from the bootstrap samples. As a result, we have

$$P\left(\theta_k^* - \theta_m^* \geq \widehat{\theta}_k - \widehat{\theta}_m - \widehat{c}_{1-\alpha} \widehat{\sigma}_{m,k}\right) \geq 1 - \alpha.$$

Similarly, this implies

$$P\left(r(m) \geq 1 + \sum_{k \neq m} \mathbf{1}(\widehat{\theta}_k - \widehat{\theta}_m > \widehat{c}_{1-\alpha}\widehat{\sigma}_{m,k})\right) \geq 1 - \alpha.$$

This yields a critical region at a significance level of alpha for the test (23)

$$\left\{1 + \sum_{k \neq m} \mathbf{1}(\widehat{\theta}_k - \widehat{\theta}_m > \widehat{c}_{1-\alpha}\widehat{\sigma}_{m,k}) > K\right\}.$$

B.2 Proof Outline

Lemma 26. *For $i \in [n]$, with probability at least $1 - O(n^{-10})$ we have*

- $\left|\left(\nabla \mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\right)_i\right| \lesssim \sqrt{\frac{np \log n}{L}}$;
- $\sum_{\substack{j \neq i \\ np.}} \left(\nabla^2 \mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\right)_{i,j}^2 \lesssim np(1+dp), \quad \sum_{k>n} \left(\nabla^2 \mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\right)_{i,k}^2 \lesssim ndp^2, \quad \sum_{j \in [n], j \neq i} \left|\left(\nabla^2 \mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\right)_{i,j}\right| \lesssim$
- $|y_{j,i} - \mathbb{E}y_{j,i}| \lesssim \sqrt{\frac{\log n}{L}}$, *for any $i, j \in [n], i \neq j$.*

Proof See §D.1 for a detailed proof. ■

Recall that we define the norm $\|\cdot\|_c$ as

$$\left\|\widetilde{\boldsymbol{\beta}}\right\|_c := \max_{i \in [n]} \left|\widetilde{\boldsymbol{\beta}}_i\right| + \sqrt{\frac{c_3(d+1)}{n}} \sqrt{\sum_{j=n+1}^{n+d} \widetilde{\boldsymbol{\beta}}_j^2}, \qquad \forall \widetilde{\boldsymbol{\beta}} \in \mathbb{R}^{n+d}.$$

Then for any $i \in [n]$ and $\widetilde{\boldsymbol{\beta}} \in \mathbb{R}^{n+d}$, we have $\left|\widetilde{\boldsymbol{x}}_i^\top \widetilde{\boldsymbol{\beta}}\right| \leq \left\|\widetilde{\boldsymbol{\beta}}\right\|_c$.

B.3 Proof Outline of Theorem 7

In this subsection, we provide the proof outline for Theorem 7. The following lemma gives a bound for $\left\|\Delta\widetilde{\boldsymbol{\beta}}\right\|_2 := \|\widetilde{\boldsymbol{\beta}}_M - \overline{\boldsymbol{\beta}}\|_2$, which validates the first part of Theorem 7.

Theorem 27. *Under the assumptions of Theorem 4, with probability at least $1 - O(n^{-6})$ we have*

$$\left\|\widetilde{\boldsymbol{\beta}}_M - \overline{\boldsymbol{\beta}}\right\|_2 \lesssim \kappa_1^4 \frac{(d+1)^{0.5} \log n}{\sqrt{n} pL}.$$

Proof See §D.2 for a detailed proof. ■

Next, we consider controlling the magnitude of $\Delta\alpha_i = \widehat{\alpha}_{M,i} - \overline{\alpha}_i$ for $i \in [n]$ in order to prove the second part of Theorem 7.

Recall that we define the quadratic approximation $\overline{\mathcal{L}}(\cdot)$ to $\mathcal{L}(\cdot)$ in (6), which is also given as below:

$$\overline{\mathcal{L}}(\widetilde{\boldsymbol{\beta}}) = \mathcal{L}(\widetilde{\boldsymbol{\beta}}) + \left(\widetilde{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}^*\right)^\top \nabla \mathcal{L}(\widetilde{\boldsymbol{\beta}}^*) + \frac{1}{2}\left(\widetilde{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}^*\right)^\top \nabla^2 \mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\left(\widetilde{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}^*\right).$$

We adopt the following notation for a given vector $\mathbf{x} \in \mathbb{R}^{n+d}$,

$$\overline{\mathcal{L}}\Big|_{\mathbf{x}_{-i}}(x_i) = \overline{\mathcal{L}}(\check{\boldsymbol{\beta}})\Big|_{\check{\boldsymbol{\beta}}_i = x_i, \check{\boldsymbol{\beta}}_{-i} = \mathbf{x}_{-i}},$$

which acts as a marginal likelihood function for the i-th coordinate x_i given the other coordinates $\mathbf{x}_{-i}$ fixed. Accroding to this definition, we have the following proposition. The proof of Proposition 28 is included in §D.3.

Proposition 28. *For $i \in [n]$, $\overline{\alpha}_i$ is the minimizer of the univariate function $\overline{\mathcal{L}}|_{\overline{\boldsymbol{\beta}}_{-i}}$.*

By changing the $n + d - 1$ coordinates that we fix, we define $\overline{\alpha}_i'$ as the minimizer of $\overline{\mathcal{L}}|_{\widetilde{\boldsymbol{\beta}}_{M,-i}}(x_i)$. Here we fix $\widetilde{\boldsymbol{\beta}}_{M,-i}$ and optimize $\overline{\mathcal{L}}$ based on this fixed parameter. Explicitly, the minimizer is calculated as

$$\overline{\alpha}_i' = \alpha_i^* - \frac{\left(\nabla \mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\right)_i + \sum_{j \neq i}\left(\widetilde{\boldsymbol{\beta}}_{M,j} - \widetilde{\boldsymbol{\beta}}_j^*\right)\left(\nabla^2 \mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\right)_{i,j}}{\left(\nabla^2 \mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\right)_{i,i}}.$$

In order to bound $|\widehat{\alpha}_{M,i} - \overline{\alpha}_i|$, we bound $|\overline{\alpha}_i' - \overline{\alpha}_i|$ and $|\widehat{\alpha}_{M,i} - \overline{\alpha}_i'|$ separately.

In terms of $|\overline{\alpha}_i' - \overline{\alpha}_i|$, we provide an upper bound for this quantity in Lemma 29.

Lemma 29. *Under the assumptions of Theorem 4, as long as $\kappa_1^2\sqrt{(d+1)\log n/npL} = \mathcal{O}(1)$, for $i \in [n]$, with probability at least $1 - O(n^{-6})$ we have*

$$\left|\overline{\alpha}_i' - \overline{\alpha}_i\right| \lesssim \kappa_1^5 \frac{(d+1)\log n}{npL} + \frac{\kappa_1^4}{np}\sqrt{\frac{(d+1)\log n}{L}}\left(\sqrt{d+1} + \frac{\log n}{\sqrt{np}}\right) + \kappa_1^2\left(\sqrt{\frac{d+1}{np}} + \frac{\log n}{np}\right)\left\|\widehat{\boldsymbol{\alpha}}_M - \overline{\boldsymbol{\alpha}}\right\|_\infty.$$

Proof The detailed proof of Lemma 29 is given in §D.5. ■

On the other hand, for a given $\mathbf{x} \in \mathbb{R}^{n+d}$, we let

$$\mathcal{L}\Big|_{\mathbf{x}_{-i}}(x_i) = \mathcal{L}(\check{\boldsymbol{\beta}})\Big|_{\check{\boldsymbol{\beta}}_i = x, \check{\boldsymbol{\beta}}_{-i} = \mathbf{x}_{-i}}. \tag{24}$$

Here we consider the marginal loss of $\mathcal{L}(\cdot)$ in the i-th coordinate given other coordinates fixed. Similar to Proposition 28, we have the following proposition for $\mathcal{L}|_{\widetilde{\boldsymbol{\beta}}_{M,-i}}(x)$. The proof of Proposition 30 is also included in §D.3.

Proposition 30. *For $i \in [n]$, $\widehat{\alpha}_{M,i}$ is the minimizer of the univariate function $\mathcal{L}|_{\widetilde{\boldsymbol{\beta}}_{M,-i}}(x)$.*

We now describe the intuition of bounding $|\widehat{\alpha}_{M,i} - \overline{\alpha}'_i|$. Since $\alpha_{M,i}$ is the minimizer of $\mathcal{L}|_{\widetilde{\boldsymbol{\beta}}_{M,-i}}(x)$ and $\overline{\alpha}'_i$ is the minimizer of $\overline{\mathcal{L}}|_{\widetilde{\boldsymbol{\beta}}_{M,-i}}(x)$. Therefore, as long as $\overline{\mathcal{L}}|_{\widetilde{\boldsymbol{\beta}}_{M,-i}}(\cdot)$ and $\mathcal{L}|_{\widetilde{\boldsymbol{\beta}}_{M,-i}}(\cdot)$ are close enough, the difference between their minimizers $\widehat{\alpha}_{M,i}$ and $\overline{\alpha}'_i$ is small. We summarize this finding in the following lemma 31.

Lemma 31. *Under the assumptions of Theorem 4, for $i \in [n]$, with probability at least $1 - O(n^{-6})$ we have*

$$\left|\widehat{\alpha}_{M,i} - \overline{\alpha}'_i\right| \lesssim \kappa_1^6 \frac{(d+1)\log n}{npL}.$$

Proof The proof of Lemma 31 is presented in §D.6. ■

Finally, combining the conclusions of Lemma 29 and Lemma 31 we get the following theorem for $|\widehat{\alpha}_{M,i} - \overline{\alpha}_i|$.

Theorem 32. *Under the assumptions of Theorem 4, as long as $\kappa_1^2\sqrt{(d+1)\log n/npL} = \mathcal{O}(1)$ and $\kappa_1^2\left(\sqrt{(d+1)/np} + \log n/np\right) \le c$ for some fixed constant $c > 0$, with probability at least $1 - O(n^{-5})$, for $i \in [n]$ we have*

$$|\widehat{\alpha}_{M,i} - \overline{\alpha}_i| \lesssim \kappa_1^6 \frac{(d+1)\log n}{npL} + \frac{\kappa_1^4}{np}\sqrt{\frac{(d+1)\log n}{L}}\left(\sqrt{d+1} + \frac{\log n}{\sqrt{np}}\right).$$

If we further assume $np \gtrsim (\log n)^2$, then with probability at least $1 - O(n^{-5})$, for $i \in [n]$ we have

$$|\widehat{\alpha}_{M,i} - \overline{\alpha}_i| \lesssim \kappa_1^6 \frac{(d+1)\log n}{npL} + \kappa_1^4 \frac{d+1}{np}\sqrt{\frac{\log n}{L}}.$$

Proof [Proof of Theorem 32] Combining Lemma 29 and Lemma 31, we know that

$$\begin{aligned}\left\|\widehat{\boldsymbol{\alpha}}_M - \overline{\boldsymbol{\alpha}}\right\|_\infty \lesssim & \kappa_1^6 \frac{(d+1)\log n}{npL} + \frac{\kappa_1^4}{np}\sqrt{\frac{(d+1)\log n}{L}}\left(\sqrt{d+1} + \frac{\log n}{\sqrt{np}}\right) \\ & + \kappa_1^2\left(\sqrt{\frac{d+1}{np}} + \frac{\log n}{np}\right)\left\|\widehat{\boldsymbol{\alpha}}_M - \overline{\boldsymbol{\alpha}}\right\|_\infty\end{aligned}$$

with probability at least $1 - O(n^{-5})$. To reveal the constant hidden in the above inequality, we write it as

$$\begin{aligned}\left\|\widehat{\boldsymbol{\alpha}}_M - \overline{\boldsymbol{\alpha}}\right\|_\infty \le & C_{\text{Hidden}}\left(\kappa_1^6 \frac{(d+1)\log n}{npL} + \frac{\kappa_1^4}{np}\sqrt{\frac{(d+1)\log n}{L}}\left(\sqrt{d+1} + \frac{\log n}{\sqrt{np}}\right)\right) \\ & + C_{\text{Hidden}}\kappa_1^2\left(\sqrt{\frac{d+1}{np}} + \frac{\log n}{np}\right)\left\|\widehat{\boldsymbol{\alpha}}_M - \overline{\boldsymbol{\alpha}}\right\|_\infty\end{aligned}$$

with probability at least $1 - O(n^{-5})$. We choose $c = 1/(2C_{\text{Hidden}})$ in Theorem 32. As a result, as long as $\kappa_1^2 \left(\sqrt{(d+1)/np} + \log n/np\right) \leq c$, we have

$$\begin{aligned}\|\widehat{\boldsymbol{\alpha}}_M - \overline{\boldsymbol{\alpha}}\|_\infty &\leq \frac{C_{\text{Hidden}}}{1-0.5}\left(\kappa_1^6 \frac{(d+1)\log n}{npL} + \frac{\kappa_1^4}{np}\sqrt{\frac{(d+1)\log n}{L}}\left(\sqrt{d+1} + \frac{\log n}{\sqrt{np}}\right)\right) \\ &\lesssim \kappa_1^6 \frac{(d+1)\log n}{npL} + \frac{\kappa_1^4}{np}\sqrt{\frac{(d+1)\log n}{L}}\left(\sqrt{d+1} + \frac{\log n}{\sqrt{np}}\right)\end{aligned}$$

with probability at least $1 - O(n^{-5})$. ■

B.4 Proof of Theorem 7

Proof [Proof of Theorem 7] The conclusion of Theorem 7 follows directly from conclusions of Theorem 27 and Theorem 32. ■

We next prove Theorem 10 based on the results of Theorem 7.

B.5 Proof of Theorem 10

This subsection aims at deriving theoretical proof for Theorem 10.

Proof The following content is conditioned on the event $\mathcal{A}_2$. Recall that $\overline{\boldsymbol{c}}$ is the projection of $\boldsymbol{c}$ onto linear space Θ. Therefore, we obtain

$$\boldsymbol{c}^\top \overline{\boldsymbol{\beta}} - \boldsymbol{c}^\top \widetilde{\boldsymbol{\beta}}^* = \overline{\boldsymbol{c}}^\top \overline{\boldsymbol{\beta}} - \overline{\boldsymbol{c}}^\top \widetilde{\boldsymbol{\beta}}^*.$$

By Proposition 35 we have $\mathcal{P}\nabla\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*) + \mathcal{P}\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)(\overline{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}^*) = \mathbf{0}$. Since $\overline{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\beta}}^* \in \Theta$, we also have $\mathcal{P}\nabla\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*) + \mathcal{P}\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\mathcal{P}(\overline{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}^*) = \mathbf{0}$. Let $\boldsymbol{v} = \left[\mathcal{P}\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\mathcal{P}\right]^+ \overline{\boldsymbol{c}}$. Then we have

$$\begin{aligned}0 &= \boldsymbol{v}^\top \left(\mathcal{P}\nabla\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*) + \mathcal{P}\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\mathcal{P}(\overline{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}^*)\right) \\ &= \boldsymbol{v}^\top \mathcal{P}\nabla\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*) + \overline{\boldsymbol{c}}^\top \overline{\boldsymbol{\beta}} - \overline{\boldsymbol{c}}^\top \widetilde{\boldsymbol{\beta}}^*.\end{aligned}$$

As a result, we have

$$\begin{aligned}\overline{\boldsymbol{c}}^\top \overline{\boldsymbol{\beta}} - \overline{\boldsymbol{c}}^\top \widetilde{\boldsymbol{\beta}}^* &= -\boldsymbol{v}^\top \mathcal{P}\nabla\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*) \\ &= \sum_{l=1}^{L} \sum_{(i,j)\in\mathcal{E}, i>j} \frac{1}{L}\left\{y_{j,i}^{(l)} - \frac{e^{\widetilde{\boldsymbol{x}}_i^\top \widetilde{\boldsymbol{\beta}}}}{e^{\widetilde{\boldsymbol{x}}_i^\top \widetilde{\boldsymbol{\beta}}} + e^{\widetilde{\boldsymbol{x}}_j^\top \widetilde{\boldsymbol{\beta}}}}\right\} \boldsymbol{v}^\top \mathcal{P}(\widetilde{\boldsymbol{x}}_i - \widetilde{\boldsymbol{x}}_j).\end{aligned}$$

For (i,j) such that $(i,j) \in \mathcal{E}, i > j$ and $l \in [L]$, we define

$$X_{i,j}^{(l)} = \frac{1}{L}\left\{y_{j,i}^{(l)} - \frac{e^{\widetilde{\boldsymbol{x}}_i^\top \widetilde{\boldsymbol{\beta}}}}{e^{\widetilde{\boldsymbol{x}}_i^\top \widetilde{\boldsymbol{\beta}}} + e^{\widetilde{\boldsymbol{x}}_j^\top \widetilde{\boldsymbol{\beta}}}}\right\} \boldsymbol{v}^\top \mathcal{P}(\widetilde{\boldsymbol{x}}_i - \widetilde{\boldsymbol{x}}_j).$$

Then we have

$$\frac{\mathbb{E}\left|X_{i,j}^{(l)}\right|^3}{\mathbb{E}\left(X_{i,j}^{(l)}\right)^2} \leq = \frac{\left|\boldsymbol{v}^\top \mathcal{P}(\widetilde{\boldsymbol{x}}_i - \widetilde{\boldsymbol{x}}_j)\right|}{L} \cdot \left(\phi(\widetilde{\boldsymbol{x}}_i^\top \widetilde{\boldsymbol{\beta}}^* - \widetilde{\boldsymbol{x}}_i^\top \widetilde{\boldsymbol{\beta}}^*)^2 + (1-\phi(\widetilde{\boldsymbol{x}}_i^\top \widetilde{\boldsymbol{\beta}}^* - \widetilde{\boldsymbol{x}}_i^\top \widetilde{\boldsymbol{\beta}}^*))^2\right)$$
$$\leq \frac{\|\overline{\boldsymbol{c}}\|_2 \left\|\left[\mathcal{P}\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\mathcal{P}\right]^+\right\| \|\mathcal{P}(\widetilde{\boldsymbol{x}}_i - \widetilde{\boldsymbol{x}}_j)\|_2}{L} \lesssim \frac{\|\overline{\boldsymbol{c}}\|_2}{\lambda_{\min,\perp}(\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*))L} \lesssim \frac{\kappa_1}{npL}\|\overline{\boldsymbol{c}}\|_2$$

with probability exceeding $1 - O(n^{-10})$ (randomness comes from $\mathcal{G}$). As a result, by Berry (1941) we have

$$\sup_{x\in\mathbb{R}}\left|\mathbb{P}\left(\frac{\overline{\boldsymbol{c}}^\top\overline{\boldsymbol{\beta}} - \overline{\boldsymbol{c}}^\top\widetilde{\boldsymbol{\beta}}^*}{\sqrt{\mathrm{Var}\left[\overline{\boldsymbol{c}}^\top\overline{\boldsymbol{\beta}} \mid \mathcal{G}\right]}} \leq x \middle| \mathcal{G}\right) - \mathbb{P}\left(\mathcal{N}(0,1) \leq x\right)\right| \leq \frac{\max_{(i,j)\in\mathcal{E}, i>j, l\in[L]} \mathbb{E}\left|X_{i,j}^{(l)}\right|^3 / \mathbb{E}\left(X_{i,j}^{(l)}\right)^2}{\sqrt{\mathrm{Var}\left[\overline{\boldsymbol{c}}^\top\overline{\boldsymbol{\beta}} \mid \mathcal{G}\right]}}$$
$$\lesssim \frac{\kappa_1}{npL\sqrt{\mathrm{Var}\left[\overline{\boldsymbol{c}}^\top\overline{\boldsymbol{\beta}} \mid \mathcal{G}\right]}}\|\overline{\boldsymbol{c}}\|_2$$

with probability exceeding $1 - O(n^{-10})$ (randomness comes from $\mathcal{G}$). And, we know that $\mathrm{Var}\left[\overline{\boldsymbol{c}}^\top\overline{\boldsymbol{\beta}} \mid \mathcal{G}\right] = \frac{1}{L}\overline{\boldsymbol{c}}^\top\left[\mathcal{P}\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\mathcal{P}\right]^+\overline{\boldsymbol{c}}$ and

$$\overline{\boldsymbol{c}}^\top\left[\mathcal{P}\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\mathcal{P}\right]^+\overline{\boldsymbol{c}} \gtrsim \frac{1}{pn}\|\overline{\boldsymbol{c}}\|_2^2$$

with probability exceeding $1 - O(n^{-10})$. Therefore,

$$\sup_{x\in\mathbb{R}}\left|\mathbb{P}\left(\frac{\sqrt{L}\left(\overline{\boldsymbol{c}}^\top\overline{\boldsymbol{\beta}} - \overline{\boldsymbol{c}}^\top\widetilde{\boldsymbol{\beta}}^*\right)}{\sqrt{\overline{\boldsymbol{c}}^\top\left[\mathcal{P}\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\mathcal{P}\right]^+\overline{\boldsymbol{c}}}} \leq x \middle| \mathcal{G}\right) - \mathbb{P}\left(\mathcal{N}(0,1) \leq x\right)\right| \lesssim \frac{\kappa_1}{\sqrt{npL}}$$

with probability exceeding $1 - O(n^{-10})$ (randomness comes from $\mathcal{G}$). On the other hand, by Theorem 7 we have

$$\left|\frac{\sqrt{L}\left(\boldsymbol{c}^\top\widetilde{\boldsymbol{\beta}}_M - \boldsymbol{c}^\top\overline{\boldsymbol{\beta}}\right)}{\sqrt{\overline{\boldsymbol{c}}^\top\left[\mathcal{P}\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\mathcal{P}\right]^+\overline{\boldsymbol{c}}}}\right| \lesssim \left[\kappa_1^6\frac{(d+1)\log n}{\sqrt{npL}} + \kappa_1^4\sqrt{\frac{(d+1)\log n}{np}}\left(\sqrt{d+1} + \frac{\log n}{\sqrt{np}}\right)\right]\frac{\|\boldsymbol{c}_{1:n}\|_1}{\|\overline{\boldsymbol{c}}\|_2}$$
$$+ \kappa_1^4\frac{(d+1)^{0.5}\log n}{\sqrt{pL}}\frac{\|\boldsymbol{c}_{n+1:n+d}\|_2}{\|\overline{\boldsymbol{c}}\|_2}$$

with probability at least $1 - O(n^{-5})$. Therefore, we conclude the first part of Theorem 10.

We next take all randomness into consideration. For simplicity we denote by

$$\Gamma = \left[\kappa_1^6\frac{(d+1)\log n}{\sqrt{npL}} + \kappa_1^4\sqrt{\frac{(d+1)\log n}{np}}\left(\sqrt{d+1} + \frac{\log n}{\sqrt{np}}\right)\right]\frac{\|\boldsymbol{c}_{1:n}\|_1}{\|\overline{\boldsymbol{c}}\|_2} + \kappa_1^4\frac{(d+1)^{0.5}\log n}{\sqrt{pL}}\frac{\|\boldsymbol{c}_{n+1:n+d}\|_2}{\|\overline{\boldsymbol{c}}\|_2}.$$

To begin with, for fixed x we have

$$\begin{aligned}
&\left|\mathbb{P}\left(\frac{\sqrt{L}\left(\overline{\boldsymbol{c}}^\top\overline{\boldsymbol{\beta}}-\overline{\boldsymbol{c}}^\top\widetilde{\boldsymbol{\beta}}^*\right)}{\sqrt{\overline{\boldsymbol{c}}^\top\left[\mathcal{P}\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\mathcal{P}\right]^+\overline{\boldsymbol{c}}}}\le x\right)-\mathbb{P}\left(\mathcal{N}(0,1)\le x\right)\right| \\
=&\left|\mathbb{E}_{\mathcal{G}}\left[\mathbb{P}\left(\frac{\sqrt{L}\left(\overline{\boldsymbol{c}}^\top\overline{\boldsymbol{\beta}}-\overline{\boldsymbol{c}}^\top\widetilde{\boldsymbol{\beta}}^*\right)}{\sqrt{\overline{\boldsymbol{c}}^\top\left[\mathcal{P}\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\mathcal{P}\right]^+\overline{\boldsymbol{c}}}}\le x\Big|\mathcal{G}\right)\right]-\mathbb{P}\left(\mathcal{N}(0,1)\le x\right)\right| \\
\le&\mathbb{E}_{\mathcal{G}}\left|\mathbb{P}\left(\frac{\sqrt{L}\left(\overline{\boldsymbol{c}}^\top\overline{\boldsymbol{\beta}}-\overline{\boldsymbol{c}}^\top\widetilde{\boldsymbol{\beta}}^*\right)}{\sqrt{\overline{\boldsymbol{c}}^\top\left[\mathcal{P}\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\mathcal{P}\right]^+\overline{\boldsymbol{c}}}}\le x\Big|\mathcal{G}\right)-\mathbb{P}\left(\mathcal{N}(0,1)\le x\right)\right| \\
\lesssim&\frac{\kappa_1}{\sqrt{npL}}+\frac{1}{n^{10}}.
\end{aligned}$$

As a result, we have

$$\sup_{x\in\mathbb{R}}\left|\mathbb{P}\left(\frac{\sqrt{L}\left(\overline{\boldsymbol{c}}^\top\overline{\boldsymbol{\beta}}-\overline{\boldsymbol{c}}^\top\widetilde{\boldsymbol{\beta}}^*\right)}{\sqrt{\overline{\boldsymbol{c}}^\top\left[\mathcal{P}\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\mathcal{P}\right]^+\overline{\boldsymbol{c}}}}\le x\right)-\mathbb{P}\left(\mathcal{N}(0,1)\le x\right)\right|\lesssim\frac{\kappa_1}{\sqrt{npL}}+\frac{1}{n^{10}}. \tag{25}$$

Consider event $A=\left\{\left|\frac{\sqrt{L}\left(\overline{\boldsymbol{c}}^\top\widetilde{\boldsymbol{\beta}}_M-\overline{\boldsymbol{c}}^\top\overline{\boldsymbol{\beta}}\right)}{\sqrt{\overline{\boldsymbol{c}}^\top\left[\mathcal{P}\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\mathcal{P}\right]^+\overline{\boldsymbol{c}}}}\right|\le\Lambda\Gamma\right\}$, where $\Lambda>0$ is some constant such that $\mathbb{P}(A^c)=O(n^{-5})$. Then we consider the following three events

$$\begin{aligned}
B_1=&\left\{\frac{\sqrt{L}\left(\overline{\boldsymbol{c}}^\top\widetilde{\boldsymbol{\beta}}_M-\overline{\boldsymbol{c}}^\top\widetilde{\boldsymbol{\beta}}^*\right)}{\sqrt{\overline{\boldsymbol{c}}^\top\left[\mathcal{P}\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\mathcal{P}\right]^+\overline{\boldsymbol{c}}}}\le x\right\},B_2=\left\{\frac{\sqrt{L}\left(\overline{\boldsymbol{c}}^\top\overline{\boldsymbol{\beta}}-\overline{\boldsymbol{c}}^\top\widetilde{\boldsymbol{\beta}}^*\right)}{\sqrt{\overline{\boldsymbol{c}}^\top\left[\mathcal{P}\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\mathcal{P}\right]^+\overline{\boldsymbol{c}}}}\le x-\Lambda\Gamma\right\}, \\
B_3=&\left\{\frac{\sqrt{L}\left(\overline{\boldsymbol{c}}^\top\overline{\boldsymbol{\beta}}-\overline{\boldsymbol{c}}^\top\widetilde{\boldsymbol{\beta}}^*\right)}{\sqrt{\overline{\boldsymbol{c}}^\top\left[\mathcal{P}\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\mathcal{P}\right]^+\overline{\boldsymbol{c}}}}\le x+\Lambda\Gamma\right\}.
\end{aligned}$$

Then we have

$$\left|\mathbb{P}\left(\frac{\sqrt{L}\left(\overline{\boldsymbol{c}}^\top\widetilde{\boldsymbol{\beta}}_M-\overline{\boldsymbol{c}}^\top\widetilde{\boldsymbol{\beta}}^*\right)}{\sqrt{\overline{\boldsymbol{c}}^\top\left[\mathcal{P}\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\mathcal{P}\right]^+\overline{\boldsymbol{c}}}}\le x\right)-\mathbb{P}\left(\mathcal{N}(0,1)\le x\right)\right|=\left|\mathbb{P}(B_1\cap A)+\mathbb{P}(B_1\cap A^c)-\mathbb{P}\left(\mathcal{N}(0,1)\le x\right)\right| \tag{26}$$

$$\lesssim\left|\mathbb{P}(B_1\cap A)-\mathbb{P}\left(\mathcal{N}(0,1)\le x\right)\right|+\frac{1}{n^5}. \tag{27}$$

On the other hand, for $B_1\cap A$ we have

$$B_2\cap A\subset B_1\cap A\subset B_3\cap A.$$

As a result, we know that

$$\mathbb{P}(B_1\cap A)\le\mathbb{P}\left(B_3\cap A\right)\le\mathbb{P}(B_3). \tag{28}$$

By Eq. (25) we have

$$\left|\mathbb{P}(B_3)-\mathbb{P}\left(\mathcal{N}(0,1)\le x+\Lambda\Gamma\right)\right|\lesssim\frac{\kappa_1}{\sqrt{npL}}+\frac{1}{n^{10}}.$$

On the other hand, we have

$$\left|\mathbb{P}\left(\mathcal{N}(0,1)\le x+\Lambda\Gamma\right)-\mathbb{P}(\mathcal{N}(0,1)\le x)\right|\le\Lambda\Gamma.$$

Therefore, we have

$$\left|\mathbb{P}(B_3)-\mathbb{P}\left(\mathcal{N}(0,1)\le x\right)\right|\lesssim\frac{\kappa_1}{\sqrt{npL}}+\Gamma+\frac{1}{n^{10}}. \tag{29}$$

For $B_1\cap A$ we also have

$$\mathbb{P}(B_1\cap A)\ge\mathbb{P}(B_2\cap A). \tag{30}$$

By definition we have

$$\left|\mathbb{P}(B_2\cap A)-\mathbb{P}(B_2)\right|\le\mathbb{P}(A^c)\lesssim\frac{1}{n^5}.$$

By Eq. (25) we have

$$\left|\mathbb{P}(B_2)-\mathbb{P}\left(\mathcal{N}(0,1)\le x-\Lambda\Gamma\right)\right|\lesssim\frac{\kappa_1}{\sqrt{npL}}+\frac{1}{n^{10}}.$$

On the other hand, we have

$$\left|\mathbb{P}\left(\mathcal{N}(0,1)\le x-\Lambda\Gamma\right)-\mathbb{P}(\mathcal{N}(0,1)\le x)\right|\le\Lambda\Gamma.$$

Therefore, we have

$$|\mathbb{P}(B_2 \cap A) - \mathbb{P}\left(\mathcal{N}(0,1) \le x\right)| \lesssim \frac{\kappa_1}{\sqrt{npL}} + \Gamma + \frac{1}{n^5}. \tag{31}$$

Combine Eq. (27), Eq. (28) and Eq. (30) we know that

$$\begin{aligned}
&\left| \mathbb{P}\left(\frac{\sqrt{L}\left(\overline{\boldsymbol{c}}^\top \widetilde{\boldsymbol{\beta}}_M - \overline{\boldsymbol{c}}^\top \widetilde{\boldsymbol{\beta}}^*\right)}{\sqrt{\overline{\boldsymbol{c}}^\top \left[\mathcal{P}\nabla^2 \mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\mathcal{P}\right]^+ \overline{\boldsymbol{c}}}} \le x \right) - \mathbb{P}\left(\mathcal{N}(0,1) \le x\right) \right| \\
\lesssim &\max\{|\mathbb{P}(B_3) - \mathbb{P}\left(\mathcal{N}(0,1) \le x\right)|, |\mathbb{P}(B_2 \cap A) - \mathbb{P}\left(\mathcal{N}(0,1) \le x\right)|\} + \frac{1}{n^5}.
\end{aligned}$$

Therefore, by Eq. (29), Eq. (31) we have

$$\begin{aligned}
&\left| \mathbb{P}\left(\frac{\sqrt{L}\left(\overline{\boldsymbol{c}}^\top \widetilde{\boldsymbol{\beta}}_M - \overline{\boldsymbol{c}}^\top \widetilde{\boldsymbol{\beta}}^*\right)}{\sqrt{\overline{\boldsymbol{c}}^\top \left[\mathcal{P}\nabla^2 \mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\mathcal{P}\right]^+ \overline{\boldsymbol{c}}}} \le x \right) - \mathbb{P}\left(\mathcal{N}(0,1) \le x\right) \right| \\
\lesssim &\frac{\kappa_1}{\sqrt{npL}} + \Gamma + \frac{1}{n^5}.
\end{aligned}$$

Since the above inequality holds for every $x \in \mathbb{R}$, we prove the desired result.

Thus, we finally conclude our proof of Theorem 10.

■

B.6 Proof of Corollary 11

Proof [Proof of Corollary 11] Let $\{\widetilde{\boldsymbol{e}}_i\}_{i=1}^{n+d}$ be canonical basis vectors of $\mathbb{R}^{n+d}$. By taking $\boldsymbol{c} = \widetilde{\boldsymbol{e}}_k$ for $k \in [n]$ in Theorem 10, we only have to show that $\frac{\|\boldsymbol{c}_{1:n}\|_1}{\|\overline{\boldsymbol{c}}\|_2} \lesssim 1$, and this is also equivalent to $\|\overline{\boldsymbol{c}}\|_2 \gtrsim 1$. By the definition of $\mathcal{P}$ we have

$$\begin{aligned}
\|\overline{\boldsymbol{c}}\|_2^2 &= \left(1 - \left(\boldsymbol{Z}\left(\boldsymbol{Z}^\top \boldsymbol{Z}\right)^{-1} \boldsymbol{Z}^\top\right)_{k,k}\right)^2 + \sum_{i \in [n], i \ne k} \left(\boldsymbol{Z}\left(\boldsymbol{Z}^\top \boldsymbol{Z}\right)^{-1} \boldsymbol{Z}^\top\right)_{i,k}^2 \\
&\ge 1 - 2\left(\boldsymbol{Z}\left(\boldsymbol{Z}^\top \boldsymbol{Z}\right)^{-1} \boldsymbol{Z}^\top\right)_{k,k} \ge 1 - 2\left\|\boldsymbol{Z}\left(\boldsymbol{Z}^\top \boldsymbol{Z}\right)^{-1} \boldsymbol{Z}^\top\right\|_{2,\infty} \\
&\ge 1 - 2c_0\sqrt{\frac{d+1}{n}} \ge 0.1.
\end{aligned}$$

As a result, the first part of Corollary 11 is proved by Theorem 10.

For simplicity we denote by $\Gamma = \kappa_1^6 \frac{(d+1)\log n}{\sqrt{npL}} + \kappa_1^4 \sqrt{\frac{(d+1)\log n}{np}} \left(\sqrt{d+1} + \frac{\log n}{\sqrt{np}} \right)$. To begin with, for fixed x we have

$$\begin{aligned}
&\left| \mathbb{P}\left(\frac{\sqrt{L}\left(\overline{\alpha}_k - \alpha_k^*\right)}{\sqrt{\left(\left[\mathcal{P}\nabla^2 \mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\mathcal{P} \right]^+ \right)_{k,k}}} \le x \right) - \mathbb{P}\left(\mathcal{N}(0,1) \le x\right) \right| \\
=&\left| \mathbb{E}_{\mathcal{G}} \left[\mathbb{P}\left(\frac{\sqrt{L}\left(\overline{\alpha}_k - \alpha_k^*\right)}{\sqrt{\left(\left[\mathcal{P}\nabla^2 \mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\mathcal{P} \right]^+ \right)_{k,k}}} \le x \middle| \mathcal{G} \right) \right] - \mathbb{P}\left(\mathcal{N}(0,1) \le x\right) \right| \\
\le&\mathbb{E}_{\mathcal{G}} \left| \mathbb{P}\left(\frac{\sqrt{L}\left(\overline{\alpha}_k - \alpha_k^*\right)}{\sqrt{\left(\left[\mathcal{P}\nabla^2 \mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\mathcal{P} \right]^+ \right)_{k,k}}} \le x \middle| \mathcal{G} \right) - \mathbb{P}\left(\mathcal{N}(0,1) \le x\right) \right| \\
\lesssim&\frac{\kappa_1}{\sqrt{npL}} + \frac{1}{n^{10}}.
\end{aligned}$$

As a result, we have

$$\sup_{x \in \mathbb{R}} \left| \mathbb{P}\left(\frac{\sqrt{L}\left(\overline{\alpha}_k - \alpha_k^*\right)}{\sqrt{\left(\left[\mathcal{P}\nabla^2 \mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\mathcal{P} \right]^+ \right)_{k,k}}} \le x \right) - \mathbb{P}\left(\mathcal{N}(0,1) \le x\right) \right| \lesssim \frac{\kappa_1}{\sqrt{npL}} + \frac{1}{n^{10}}. \tag{32}$$

Consider event $A = \left\{ \left| \frac{\sqrt{L}\left(\widehat{\alpha}_{M,k} - \overline{\alpha}_k\right)}{\sqrt{\left(\left[\mathcal{P}\nabla^2 \mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\mathcal{P} \right]^+ \right)_{k,k}}} \right| \le \Lambda\Gamma \right\}$, where $\Lambda > 0$ is some constant such that $\mathbb{P}(A^c) = O(n^{-5})$. Then we consider the following three events

$$B_1 = \left\{ \frac{\sqrt{L}\left(\widehat{\alpha}_{M,k} - \alpha_k^*\right)}{\sqrt{\left(\left[\mathcal{P}\nabla^2 \mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\mathcal{P} \right]^+ \right)_{k,k}}} \le x \right\}, B_2 = \left\{ \frac{\sqrt{L}\left(\overline{\alpha}_k - \alpha_k^*\right)}{\sqrt{\left(\left[\mathcal{P}\nabla^2 \mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\mathcal{P} \right]^+ \right)_{k,k}}} \le x - \Lambda\Gamma \right\},$$

$$B_3 = \left\{ \frac{\sqrt{L}\left(\overline{\alpha}_k - \alpha_k^*\right)}{\sqrt{\left(\left[\mathcal{P}\nabla^2 \mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\mathcal{P} \right]^+ \right)_{k,k}}} \le x + \Lambda\Gamma \right\}.$$

Then we have

$$\left|\mathbb{P}\left(\frac{\sqrt{L}\left(\widehat{\alpha}_{M,k}-\alpha_k^*\right)}{\sqrt{\left(\left[\mathcal{P}\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\mathcal{P}\right]^+\right)_{k,k}}}\le x\right)-\mathbb{P}\left(\mathcal{N}(0,1)\le x\right)\right|=\left|\mathbb{P}(B_1\cap A)+\mathbb{P}(B_1\cap A^c)-\mathbb{P}\left(\mathcal{N}(0,1)\le x\right)\right| \tag{33}$$

$$\lesssim \left|\mathbb{P}(B_1\cap A)-\mathbb{P}\left(\mathcal{N}(0,1)\le x\right)\right|+\frac{1}{n^5}. \tag{34}$$

On the other hand, for $B_1\cap A$ we have

$$B_2\cap A\subset B_1\cap A\subset B_3\cap A.$$

As a result, we know that

$$\mathbb{P}(B_1\cap A)\le\mathbb{P}\left(B_3\cap A\right)\le\mathbb{P}(B_3). \tag{35}$$

By Eq. (32) we have

$$\left|\mathbb{P}(B_3)-\mathbb{P}\left(\mathcal{N}(0,1)\le x+\Lambda\Gamma\right)\right|\lesssim\frac{\kappa_1}{\sqrt{npL}}+\frac{1}{n^{10}}.$$

On the other hand, we have

$$\left|\mathbb{P}\left(\mathcal{N}(0,1)\le x+\Lambda\Gamma\right)-\mathbb{P}(\mathcal{N}(0,1)\le x)\right|\le\Lambda\Gamma.$$

Therefore, we have

$$\left|\mathbb{P}(B_3)-\mathbb{P}\left(\mathcal{N}(0,1)\le x\right)\right|\lesssim\frac{\kappa_1}{\sqrt{npL}}+\Gamma+\frac{1}{n^{10}}. \tag{36}$$

For $B_1\cap A$ we also have

$$\mathbb{P}(B_1\cap A)\ge\mathbb{P}(B_2\cap A). \tag{37}$$

By definition we have

$$\left|\mathbb{P}(B_2\cap A)-\mathbb{P}(B_2)\right|\le\mathbb{P}(A^c)\lesssim\frac{1}{n^5}.$$

By Eq. (32) we have

$$\left|\mathbb{P}(B_2)-\mathbb{P}\left(\mathcal{N}(0,1)\le x-\Lambda\Gamma\right)\right|\lesssim\frac{\kappa_1}{\sqrt{npL}}+\frac{1}{n^{10}}.$$

On the other hand, we have

$$\left|\mathbb{P}\left(\mathcal{N}(0,1)\le x-\Lambda\Gamma\right)-\mathbb{P}(\mathcal{N}(0,1)\le x)\right|\le\Lambda\Gamma.$$

Therefore, we have

$$|\mathbb{P}(B_2 \cap A) - \mathbb{P}\left(\mathcal{N}(0,1) \le x\right)| \lesssim \frac{\kappa_1}{\sqrt{npL}} + \Gamma + \frac{1}{n^5}. \tag{38}$$

Combine Eq. (34), Eq. (35) and Eq. (37) we know that

$$\begin{aligned}
&\left| \mathbb{P}\left(\frac{\sqrt{L}\left(\widehat{\alpha}_{M,k} - \alpha_k^*\right)}{\sqrt{\left(\left[\mathcal{P}\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\mathcal{P}\right]^+\right)_{k,k}}} \le x \right) - \mathbb{P}\left(\mathcal{N}(0,1) \le x\right) \right| \\
\lesssim &\max\{|\mathbb{P}(B_3) - \mathbb{P}\left(\mathcal{N}(0,1) \le x\right)|, |\mathbb{P}(B_2 \cap A) - \mathbb{P}\left(\mathcal{N}(0,1) \le x\right)|\} + \frac{1}{n^5}.
\end{aligned}$$

Therefore, by Eq. (36), Eq. (38) we have

$$\begin{aligned}
&\left| \mathbb{P}\left(\frac{\sqrt{L}\left(\widehat{\alpha}_{M,k} - \alpha_k^*\right)}{\sqrt{\left(\left[\mathcal{P}\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\mathcal{P}\right]^+\right)_{k,k}}} \le x \right) - \mathbb{P}\left(\mathcal{N}(0,1) \le x\right) \right| \\
\lesssim &\frac{\kappa_1}{\sqrt{npL}} + \Gamma + \frac{1}{n^5}.
\end{aligned}$$

Since the above inequality holds for every $x \in \mathbb{R}$, we prove the desired result.

■

B.7 Proof of Corollary 12

The proof of Corollary 12 except the refined bound Eq. (10) follows directly from the proof of Theorem 10. Therefore, we omit the details and only prove Eq. (10) here. We prove the following lemma first.

Lemma 33. *Consider some fixed constants $a_{i,j}^{(l)}$ for $(i,j) \in \mathcal{E}, l \in [L]$, and random variable*

$$X = \sum_{l=1}^{L} \sum_{(i,j)\in\mathcal{E}} a_{i,j}^{(l)} \left\{ y_{j,i}^{(l)} - \frac{e^{\widetilde{\boldsymbol{x}}_i^\top \widetilde{\boldsymbol{\beta}}^*}}{e^{\widetilde{\boldsymbol{x}}_i^\top \widetilde{\boldsymbol{\beta}}^*} + e^{\widetilde{\boldsymbol{x}}_j^\top \widetilde{\boldsymbol{\beta}}^*}} \right\}. \tag{39}$$

Conditioned on the comparison graph $\mathcal{G}$, with probability exceeding $1 - O(n^{-11})$ we have

$$|X| \lesssim \sqrt{\kappa_1 \, Var[X \mid \mathcal{G}] \cdot \log n}.$$

Proof Let $X_{i,j}^{(l)} = a_{i,j}^{(l)} \left(y_{j,i}^{(l)} - e^{\widetilde{\boldsymbol{x}}_i^\top \widetilde{\boldsymbol{\beta}}^*} / \left(e^{\widetilde{\boldsymbol{x}}_i^\top \widetilde{\boldsymbol{\beta}}^*} + e^{\widetilde{\boldsymbol{x}}_j^\top \widetilde{\boldsymbol{\beta}}^*} \right) \right)$. Then we know that $\mathbb{E}X_{i,j}^{(l)} = 0$ and $|X_{i,j}^{(l)}| \le |a_{i,j}^{(l)}|$. As a result, by Hoeffding inequality, with probability at least $1 - O(n^{-11})$,

we have

$$\left|\sum_{l=1}^{L}\sum_{(i,j)\in\mathcal{E},i>j} X_{i,j}^{(l)}\right| \lesssim \sqrt{\log n\cdot\sum_{l=1}^{L}\sum_{(i,j)\in\mathcal{E},i>j}\left(a_{i,j}^{(l)}\right)^2}. \tag{40}$$

On the other hand, since $y_{i,j}^{(l)}$ are independent random variables, we know that

$$\begin{aligned}\mathrm{Var}[X\mid\mathcal{G}] &= \sum_{l=1}^{L}\sum_{(i,j)\in\mathcal{E},i>j}\mathrm{Var}[X_{i,j}^{(l)}\mid\mathcal{G}] = \sum_{l=1}^{L}\sum_{(i,j)\in\mathcal{E},i>j}\left(a_{i,j}^{(l)}\right)^2\phi'(\widetilde{\boldsymbol{x}}_i^\top\widetilde{\boldsymbol{\beta}}^*-\widetilde{\boldsymbol{x}}_j^\top\widetilde{\boldsymbol{\beta}}^*)\\ &\gtrsim \frac{1}{\kappa_1}\sum_{l=1}^{L}\sum_{(i,j)\in\mathcal{E},i>j}\left(a_{i,j}^{(l)}\right)^2.\end{aligned}$$

As a result, we have $\sum_{l=1}^{L}\sum_{(i,j)\in\mathcal{E},i>j}\left(a_{i,j}^{(l)}\right)^2\lesssim\kappa_1\mathrm{Var}[X\mid\mathcal{G}]$. Therefore, by Eq. (40) we know that

$$|X| = \left|\sum_{l=1}^{L}\sum_{(i,j)\in\mathcal{E},i>j}X_{i,j}^{(l)}\right|\lesssim\sqrt{\kappa_1\mathrm{Var}[X\mid\mathcal{G}]\log n}$$

with probability exceeding $1-O(n^{-11})$. ■

Proof [Proof of Eq. (10)] Conditioned on the comparison graph $\mathcal{G}$, the entries of $\overline{\boldsymbol{\beta}}$ can be written as the form Eq. (39). By Lemma 33 and union bound, conditioned on the comparison graph $\mathcal{G}$, we know that

$$\begin{aligned}\left\|\overline{\boldsymbol{\beta}}_{n+1:n+d}-\boldsymbol{\beta}^*\right\|_2 = \sqrt{\sum_{j=1}^{d}\left(\overline{\beta}_{n+j}-\beta_j^*\right)^2} &\lesssim \sqrt{\kappa_1\log n\cdot\sum_{j=1}^{d}\mathrm{Var}\left[\overline{\beta}_{n+j}-\beta_j^*\mid\mathcal{G}\right]}\\ &= \sqrt{\kappa_1\log n\cdot\mathbf{tr}\left[\mathrm{Var}\left[\,\overline{\boldsymbol{\beta}}_{n+1:n+d}-\boldsymbol{\beta}^*\mid\mathcal{G}\right]\right]}\end{aligned}\tag{41}$$

with probability at least $1-O(n^{-10})$. By Proposition 35 we know that

$$\mathrm{Var}\left[\,\overline{\boldsymbol{\beta}}_{n+1:n+d}-\boldsymbol{\beta}^*\mid\mathcal{G}\right] = \frac{1}{L}\left(\left[\mathcal{P}\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\mathcal{P}\right]^+\right)_{n+1:n+d,n+1:n+d}.$$

By the definition of Θ we know that

$$\begin{aligned}\lambda_{\max}\left(\left(\left[\mathcal{P}\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\mathcal{P}\right]^+\right)_{n+1:n+d,n+1:n+d}\right) &\le \lambda_{\max}\left(\left[\mathcal{P}\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\mathcal{P}\right]^+\right)\\ &\le \frac{1}{\lambda_{\min,\perp}(\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*))}\lesssim\frac{\kappa_1}{np}\end{aligned}$$

under event $\mathcal{A}_2$. As a result, we have

$$\mathbf{tr}\left[\mathrm{Var}\left[\,\overline{\boldsymbol{\beta}}_{n+1:n+d}-\boldsymbol{\beta}^*\mid\mathcal{G}\right]\right]\leq\frac{1}{L}d\lambda_{\max}\left(\left(\left[\mathcal{P}\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\mathcal{P}\right]^+\right)_{n+1:n+d,n+1:n+d}\right) \tag{42}$$

$$\lesssim\frac{\kappa_1(d+1)}{npL} \tag{43}$$

with probability at least $1-O(n^{-10})$. Combine Eq. (41) and Eq. (43) we know that

$$\left\|\overline{\boldsymbol{\beta}}_{n+1:n+d}-\boldsymbol{\beta}^*\right\|_2\lesssim\kappa_1\sqrt{\frac{(d+1)\log n}{npL}}$$

with probability at least $1-O(n^{-10})$. On the other hand, by Theorem 7, we have

$$\left\|\overline{\boldsymbol{\beta}}_{n+1:n+d}-\widehat{\boldsymbol{\beta}}_M\right\|_2\leq\left\|\Delta\widetilde{\boldsymbol{\beta}}\right\|_2\lesssim\kappa_1^4\frac{(d+1)^{0.5}\log n}{\sqrt{n}pL}$$

with probability at least $1-O(n^{-5})$. We know that

$$\left\|\widehat{\boldsymbol{\beta}}_M-\boldsymbol{\beta}^*\right\|_2\leq\left\|\overline{\boldsymbol{\beta}}_{n+1:n+d}-\boldsymbol{\beta}^*\right\|_2+\left\|\overline{\boldsymbol{\beta}}_{n+1:n+d}-\widehat{\boldsymbol{\beta}}_M\right\|_2\lesssim\sqrt{\frac{d+1}{n}}\max\left\{\kappa_1^4\frac{\log n}{pL},\kappa_1\sqrt{\frac{\log n}{pL}}\right\}$$

with probability at least $1-O(n^{-5})$. ■

Appendix C. Proof of Auxiliary Lemmas in Section A

In this section, we prove detailed proof of aforementioned building blocks.

C.1 Proof of Proposition 1

Proof Assume that there are two parameter vectors $\widetilde{\boldsymbol{\beta}} = (\boldsymbol{\alpha}^\top, \boldsymbol{\beta}^\top)^\top \in \Theta$ and $\widetilde{\boldsymbol{\beta}}' = (\boldsymbol{\alpha}'^\top, \boldsymbol{\beta}'^\top)^\top \in \Theta$ such that

$$\frac{e^{\alpha_j^* + \boldsymbol{x}_j^\top \boldsymbol{\beta}^*}}{e^{\alpha_i^* + \boldsymbol{x}_i^\top \boldsymbol{\beta}^*} + e^{\alpha_j^* + \boldsymbol{x}_j^\top \boldsymbol{\beta}^*}} = \frac{e^{\alpha_j'^* + \boldsymbol{x}_j^\top \boldsymbol{\beta}'^*}}{e^{\alpha_i'^* + \boldsymbol{x}_i^\top \boldsymbol{\beta}'^*} + e^{\alpha_j'^* + \boldsymbol{x}_j^\top \boldsymbol{\beta}'^*}}, \quad \forall 1 \le i \ne j \le n.$$

Since $e^b/(e^a + e^b) = 1/(e^{a-b} + 1)$, we know that

$$\alpha_i^* + \boldsymbol{x}_i^\top \boldsymbol{\beta}^* - (\alpha_j^* + \boldsymbol{x}_j^\top \boldsymbol{\beta}^*) = \alpha_i'^* + \boldsymbol{x}_i^\top \boldsymbol{\beta}'^* - (\alpha_j'^* + \boldsymbol{x}_j^\top \boldsymbol{\beta}'^*), \quad \forall 1 \le i \ne j \le n.$$

Let $c = \alpha_1^* + \boldsymbol{x}_1^\top \boldsymbol{\beta}^* - (\alpha_1'^* + \boldsymbol{x}_1^\top \boldsymbol{\beta}'^*)$ and $\mathbf{1}_n$ be a n-dimensional vertoc whose entries are all one, then we know that

$$\boldsymbol{\alpha}^* + \boldsymbol{X}\boldsymbol{\beta}^* = \boldsymbol{\alpha}'^* + \boldsymbol{X}\boldsymbol{\beta}'^* + c\mathbf{1}_n.$$

Since $\bar{\boldsymbol{X}}^\top \boldsymbol{\alpha}^* = \bar{\boldsymbol{X}}^\top \boldsymbol{\alpha}'^* = \mathbf{0}$, we know that

$$\left\|\boldsymbol{\alpha}^* - \boldsymbol{\alpha}'^*\right\|_2^2 = (\boldsymbol{\alpha}^* - \boldsymbol{\alpha}'^*)^\top (\boldsymbol{\alpha}^* - \boldsymbol{\alpha}'^*) = (\boldsymbol{\alpha}^* - \boldsymbol{\alpha}'^*)^\top (\boldsymbol{X}\boldsymbol{\beta}'^* + c\mathbf{1}_n - \boldsymbol{X}\boldsymbol{\beta}^*) = 0.$$

This implies $\boldsymbol{\alpha}^* = \boldsymbol{\alpha}'^*$. Therefore, we have $\boldsymbol{X}\boldsymbol{\beta}^* = \boldsymbol{X}\boldsymbol{\beta}'^* + c\mathbf{1}_n$. This is equivalent to

$$\bar{\boldsymbol{X}} \begin{bmatrix} c \\ \boldsymbol{\beta}^* - \boldsymbol{\beta}'^* \end{bmatrix} = \mathbf{0}.$$

Since $\bar{\boldsymbol{X}}$ has full column rank, we know that $[c, (\boldsymbol{\beta}^* - \boldsymbol{\beta}'^*)^\top] = \mathbf{0}$. As a result, we must have $\widetilde{\boldsymbol{\beta}} = \widetilde{\boldsymbol{\beta}}'$.

■

C.2 Proof of Lemma 14 and Its Corollary

In this subsection, we provide the proof of Lemma 14 in the sense that we provide an upper bound for the gradient vector in ℓ_2-norm.

Proof The gradient is calculated as

$$\nabla \mathcal{L}_\lambda(\widetilde{\boldsymbol{\beta}}^*) = \lambda \widetilde{\boldsymbol{\beta}}^* + \frac{1}{L} \sum_{(i,j) \in \mathcal{E}, i>j} \sum_{l=1}^{L} \underbrace{\left\{ -y_{j,i}^{(l)} + \frac{e^{\widetilde{\boldsymbol{x}}_i^\top \widetilde{\boldsymbol{\beta}}^*}}{e^{\widetilde{\boldsymbol{x}}_i^\top \widetilde{\boldsymbol{\beta}}^*} + e^{\widetilde{\boldsymbol{x}}_j^\top \widetilde{\boldsymbol{\beta}}^*}} \right\}}_{:= z_{i,j}^{(l)}} (\widetilde{\boldsymbol{x}}_i - \widetilde{\boldsymbol{x}}_j).$$

Since $\mathbb{E}[z_{i,j}^{(l)}] = 0, \|z_{i,j}^{(l)}\|_2 \le \|\widetilde{\boldsymbol{x}}_i - \widetilde{\boldsymbol{x}}_j\|_2 \le \sqrt{6}$, we have

$$\mathbb{E}[z_{i,j}^{(l)} z_{i,j}^{(l)\top}] = \mathrm{Var}[y_{j,i}^{(l)}](\widetilde{\boldsymbol{x}}_i - \widetilde{\boldsymbol{x}}_j)(\widetilde{\boldsymbol{x}}_i - \widetilde{\boldsymbol{x}}_j)^\top \prec (\widetilde{\boldsymbol{x}}_i - \widetilde{\boldsymbol{x}}_j)(\widetilde{\boldsymbol{x}}_i - \widetilde{\boldsymbol{x}}_j)^\top$$
$$\text{and} \qquad \mathbb{E}[z_{i,j}^{(l)\top} z_{i,j}^{(l)}] \le 6.$$

as long as n/d is large enough. Thus, with high probability (with respect to the randomness of $\mathcal{G}$), we have

$$\left\| \sum_{(i,j)\in\mathcal{E}, i>j} \sum_{l=1}^{L} \mathbb{E}\left[z_{i,j}^{(l)} z_{i,j}^{(l)\top} \right] \right\| \le L \left\| \sum_{(i,j)\in\mathcal{E}, i>j} (\widetilde{\boldsymbol{x}}_i - \widetilde{\boldsymbol{x}}_j)(\widetilde{\boldsymbol{x}}_i - \widetilde{\boldsymbol{x}}_j)^\top \right\| = L \|\boldsymbol{L}_{\mathcal{G}}\| \lesssim Lnp$$

and

$$\left| \sum_{(i,j)\in\mathcal{E}, i>j} \sum_{l=1}^{L} \mathbb{E}\left[\boldsymbol{z}_{i,j}^{(l)\top} \boldsymbol{z}_{i,j}^{(l)} \right] \right| \le 6L \left| \sum_{(i,j)\in\mathcal{E}, i>j} 1 \right| \lesssim Ln^2 p.$$

Let $V := \frac{1}{L^2} \max\left\{ \left\| \sum_{(i,j)\in\mathcal{E}} \sum_{l=1}^{L} \mathbb{E}\left[\boldsymbol{z}_{i,j}^{(l)} \boldsymbol{z}_{i,j}^{(l)\top} \right] \right\|, \left| \sum_{(i,j)\in\mathcal{E}} \sum_{l=1}^{L} \mathbb{E}\left[\boldsymbol{z}_{i,j}^{(l)\top} \boldsymbol{z}_{i,j}^{(l)} \right] \right| \right\}$ and $B := \max_{i,j,l} \|z_{i,j}^{(l)}\|/L$. By matrix Bernstein inequality (Tropp, 2015), we have

$$\left\| \nabla \mathcal{L}_\lambda(\widetilde{\boldsymbol{\beta}}^*) - \mathbb{E}\left[\nabla \mathcal{L}_\lambda(\widetilde{\boldsymbol{\beta}}^*) \mid \mathcal{G} \right] \right\|_2 \lesssim \sqrt{V \log(n+d+1)} + B \log(n+d+1) \lesssim \sqrt{\frac{n^2 p \log n}{L}} + \frac{\log n}{L}$$

with probability exceeding $1 - O(n^{-11})$ as long as $d < n$ and $npL \gtrsim \log n$. On the other hand, we have

$$\left\| \mathbb{E}\left[\nabla \mathcal{L}_\lambda(\widetilde{\boldsymbol{\beta}}^*) \mid \mathcal{G} \right] \right\|_2 = \lambda \left\| \widetilde{\boldsymbol{\beta}}^* \right\|_2 \lesssim \frac{1}{\kappa_3 \sqrt{d+1}} \sqrt{\frac{np \log n}{L}} \kappa_3 \sqrt{(d+1)n} \lesssim \sqrt{\frac{n^2 p \log n}{L}}.$$

To summarize, we have

$$\left\| \nabla \mathcal{L}_\lambda(\widetilde{\boldsymbol{\beta}}^*) \right\|_2 \lesssim \sqrt{\frac{n^2 p \log n}{L}}$$

with probability exceeding $1 - O(n^{-11})$. ■

Once Lemma 14 is established, we have the following lemma which can be viewed as a direct corollary of Lemma 14.

Lemma 34. *With λ given by Theorem 13, the following event*

$$\mathcal{A}_3 = \left\{ \left\| \nabla \mathcal{L}_\lambda\left(\widetilde{\boldsymbol{\beta}}\right) \right\|_2 \le C_0 \sqrt{\frac{n^2 p \log n}{L}} + \left(\lambda + \frac{1}{2} c_1 pn \right) r,\ \forall \widetilde{\boldsymbol{\beta}} \text{ s.t. } \left\| \widetilde{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}^* \right\|_2 \le r \right\}$$

is contained by the event $\mathcal{A}_1 \cap \mathcal{A}_2$. As a result, $\mathcal{A}_3$ happens with probability exceeding $1 - O(n^{-11})$.

Proof By the fundamental theorem of calculus we know that

$$\nabla \mathcal{L}_\lambda(\widetilde{\boldsymbol{\beta}}) = \nabla \mathcal{L}_\lambda(\widetilde{\boldsymbol{\beta}}^*) + \int_0^1 \nabla^2 \mathcal{L}_\lambda(\widetilde{\boldsymbol{\beta}}(\tau))(\widetilde{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}^*) d\tau,$$

where $\widetilde{\boldsymbol{\beta}}(\tau) = \widetilde{\boldsymbol{\beta}}^* + \tau(\widetilde{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}^*)$. So for all $\widetilde{\boldsymbol{\beta}}$ such that $\|\widetilde{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}^*\|_2 \le r$, we have

$$\begin{aligned}
\left\|\nabla \mathcal{L}_\lambda(\widetilde{\boldsymbol{\beta}})\right\|_2 &\le \left\|\nabla \mathcal{L}_\lambda(\widetilde{\boldsymbol{\beta}}^*)\right\|_2 + \int_0^1 \left\|\nabla^2 \mathcal{L}_\lambda(\widetilde{\boldsymbol{\beta}}(\tau))(\widetilde{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}^*)\right\|_2 d\tau \\
&\le \left\|\nabla \mathcal{L}_\lambda(\widetilde{\boldsymbol{\beta}}^*)\right\|_2 + \left\|\widetilde{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}^*\right\|_2 \int_0^1 \left\|\nabla^2 \mathcal{L}_\lambda(\widetilde{\boldsymbol{\beta}}(\tau))\right\| d\tau \\
&\le \left\|\nabla \mathcal{L}_\lambda(\widetilde{\boldsymbol{\beta}}^*)\right\|_2 + r \int_0^1 \left\|\nabla^2 \mathcal{L}_\lambda(\widetilde{\boldsymbol{\beta}}(\tau))\right\| d\tau.
\end{aligned}$$

Under event $\mathcal{A}_1$, we have

$$\left\|\nabla \mathcal{L}_\lambda\left(\widetilde{\boldsymbol{\beta}}^*\right)\right\|_2 \le C_0 \sqrt{\frac{n^2 p \log n}{L}}.$$

And, under event $\mathcal{A}_2$, we have $\|\nabla^2 \mathcal{L}_\lambda(\widetilde{\boldsymbol{\beta}}(\tau))\| \le \lambda + c_1 pn/2$. As a result, $\mathcal{A}_3$ is contained by $\mathcal{A}_1 \cap \mathcal{A}_2$. ■

C.3 Proof of Lemma 15

In this subsection, we prove Lemma 15 by demonstrating

$$\mathcal{A}_2 = \left\{ \frac{1}{2} c_2 pn \le \lambda_{\min,\perp}(\boldsymbol{L}_{\mathcal{G}}) \le \|\boldsymbol{L}_{\mathcal{G}}\| \le 2 c_1 pn \right\}$$

holds with high probability.

Proof Let $\boldsymbol{O}$ be any $r \times (n+d)$ matrix with orthonormal rows such that the row space of $\boldsymbol{O}$ is Θ, where r is the dimension of Θ. Then, it holds that

$$\|\boldsymbol{L}_{\mathcal{G}}\| = \|\boldsymbol{O}\boldsymbol{L}_{\mathcal{G}}\boldsymbol{O}^\top\|; \qquad \lambda_{\min,\perp}(\boldsymbol{L}_{\mathcal{G}}) = \lambda_{\min}(\boldsymbol{O}\boldsymbol{L}_{\mathcal{G}}\boldsymbol{O}^\top).$$

Let $\boldsymbol{X}_{i,j} = \boldsymbol{O}(\widetilde{\boldsymbol{x}}_i - \widetilde{\boldsymbol{x}}_j)(\widetilde{\boldsymbol{x}}_i - \widetilde{\boldsymbol{x}}_j)^\top \boldsymbol{O}^\top \mathbf{1}((i,j) \in \mathcal{E})$ for $i > j$. Then we have $\boldsymbol{O}\boldsymbol{L}_{\mathcal{G}}\boldsymbol{O}^\top = \sum_{i>j} \boldsymbol{X}_{i,j}$ and

$$\boldsymbol{X}_{i,j} \succeq \mathbf{0}, \quad \|\boldsymbol{X}_{i,j}\| \le \|\widetilde{\boldsymbol{x}}_i - \widetilde{\boldsymbol{x}}_j\|_2^2 \le 6,$$

as long as n/d is large enough. Furthermore,

$$\begin{aligned}
\lambda_{\min}\left(\mathbb{E}\sum_{i>j} \boldsymbol{X}_{i,j}\right) &= \lambda_{\min}\left(p\boldsymbol{O}\boldsymbol{\Sigma}\boldsymbol{O}^\top\right) = p\lambda_{\min,\perp}(\boldsymbol{\Sigma}) \ge c_2 pn; \\
\lambda_{\max}\left(\mathbb{E}\sum_{i>j} \boldsymbol{X}_{i,j}\right) &= \lambda_{\max}\left(p\boldsymbol{O}\boldsymbol{\Sigma}\boldsymbol{O}^\top\right) = p\lambda_{\max}(\boldsymbol{\Sigma}) \le c_1 pn.
\end{aligned}$$

By the matrix Chernoff inequality (Tropp, 2012), we have

$$\mathbb{P}\left(\lambda_{\min}\left(\sum_{i>j}\boldsymbol{X}_{i,j}\right)\leq\frac{1}{2}\lambda_{\min}\left(\mathbb{E}\sum_{i>j}\boldsymbol{X}_{i,j}\right)\right)\leq(n+d)\cdot 0.8^{c_2pn/6};$$
$$\mathbb{P}\left(\lambda_{\max}\left(\sum_{i>j}\boldsymbol{X}_{i,j}\right)\geq\frac{3}{2}\lambda_{\max}\left(\mathbb{E}\sum_{i>j}\boldsymbol{X}_{i,j}\right)\right)\leq(n+d)\cdot 0.8^{c_2pn/6}.$$

As a result, if $pn>c_p\log n$ for some $c_p>0$, we have

$$\mathbb{P}(\mathcal{A}_2)\geq\mathbb{P}\left(\frac{1}{2}c_2pn\leq\lambda_{\min}\left(\sum_{i>j}\boldsymbol{X}_{i,j}\right)\leq\lambda_{\max}\left(\sum_{i>j}\boldsymbol{X}_{i,j}\right)\leq\frac{3}{2}c_1pn\right)\geq 1-O(n^{-11}).$$

This concludes the proof of Lemma 15. ■

C.4 Proof of Lemma 17

In this subsection, we provide the proof of Lemma 17 by providing a lower bound for $\lambda_{\min,\perp}(\nabla^2\mathcal{L}_\lambda(\widetilde{\boldsymbol{\beta}}))$.

Proof For pair (i,j), without loss of generality we assume $\widetilde{\boldsymbol{x}}_i^\top\widetilde{\boldsymbol{\beta}}\leq\widetilde{\boldsymbol{x}}_j^\top\widetilde{\boldsymbol{\beta}}$, we then obtain

$$\frac{e^{\widetilde{\boldsymbol{x}}_i^\top\widetilde{\boldsymbol{\beta}}}e^{\widetilde{\boldsymbol{x}}_j^\top\widetilde{\boldsymbol{\beta}}}}{\left(e^{\widetilde{\boldsymbol{x}}_i^\top\widetilde{\boldsymbol{\beta}}}+e^{\widetilde{\boldsymbol{x}}_j^\top\widetilde{\boldsymbol{\beta}}}\right)^2}=\frac{e^{\widetilde{\boldsymbol{x}}_i^\top\widetilde{\boldsymbol{\beta}}-\widetilde{\boldsymbol{x}}_j^\top\widetilde{\boldsymbol{\beta}}}}{\left(1+e^{\widetilde{\boldsymbol{x}}_i^\top\widetilde{\boldsymbol{\beta}}-\widetilde{\boldsymbol{x}}_j^\top\widetilde{\boldsymbol{\beta}}}\right)^2}=\frac{e^{-|\widetilde{\boldsymbol{x}}_i^\top\widetilde{\boldsymbol{\beta}}-\widetilde{\boldsymbol{x}}_j^\top\widetilde{\boldsymbol{\beta}}|}}{\left(1+e^{-|\widetilde{\boldsymbol{x}}_i^\top\widetilde{\boldsymbol{\beta}}-\widetilde{\boldsymbol{x}}_j^\top\widetilde{\boldsymbol{\beta}}|}\right)^2}\geq\frac{1}{4}e^{-|\widetilde{\boldsymbol{x}}_i^\top\widetilde{\boldsymbol{\beta}}-\widetilde{\boldsymbol{x}}_j^\top\widetilde{\boldsymbol{\beta}}|}.$$

One the other hand, it holds that

$$\begin{aligned}|\widetilde{\boldsymbol{x}}_i^\top\widetilde{\boldsymbol{\beta}}-\widetilde{\boldsymbol{x}}_j^\top\widetilde{\boldsymbol{\beta}}|&\leq|\widetilde{\boldsymbol{x}}_i^\top\widetilde{\boldsymbol{\beta}}^*-\widetilde{\boldsymbol{x}}_j^\top\widetilde{\boldsymbol{\beta}}^*|+|\widetilde{\boldsymbol{x}}_i^\top\widetilde{\boldsymbol{\beta}}^*-\widetilde{\boldsymbol{x}}_i^\top\widetilde{\boldsymbol{\beta}}|+|\widetilde{\boldsymbol{x}}_j^\top\widetilde{\boldsymbol{\beta}}^*-\widetilde{\boldsymbol{x}}_j^\top\widetilde{\boldsymbol{\beta}}|.\\&\leq\log(\kappa_1)+|\alpha_i-\alpha_i^*|+|\boldsymbol{x}_i^\top\boldsymbol{\beta}^*-\boldsymbol{x}_i^\top\boldsymbol{\beta}|+|\alpha_j-\alpha_j^*|+|\boldsymbol{x}_j^\top\boldsymbol{\beta}^*-\boldsymbol{x}_j^\top\boldsymbol{\beta}|\\&\leq\log(\kappa_1)+2\|\boldsymbol{\alpha}-\boldsymbol{\alpha}^*\|_\infty+2\sqrt{\frac{c_3(d+1)}{n}}\|\boldsymbol{\beta}^*-\boldsymbol{\beta}\|_2\\&\leq\log(\kappa_1)+2C_1+2\sqrt{\frac{c_3(d+1)}{n}}C_2.\end{aligned}$$

Therefore, we obtain

$$\frac{e^{\widetilde{\boldsymbol{x}}_i^\top\widetilde{\boldsymbol{\beta}}}e^{\widetilde{\boldsymbol{x}}_j^\top\widetilde{\boldsymbol{\beta}}}}{\left(e^{\widetilde{\boldsymbol{x}}_i^\top\widetilde{\boldsymbol{\beta}}}+e^{\widetilde{\boldsymbol{x}}_j^\top\widetilde{\boldsymbol{\beta}}}\right)^2}\geq\frac{1}{4\kappa_1e^C},$$

where $C=2C_1+2\sqrt{\frac{c_3(d+1)}{n}}C_2$. As a result, for the Hessian $\nabla^2\mathcal{L}_\lambda(\widetilde{\boldsymbol{\beta}})$ we have

$$\lambda_{\min,\perp}(\nabla^2\mathcal{L}_\lambda(\widetilde{\boldsymbol{\beta}}))\geq\lambda+\frac{1}{4\kappa_1e^C}\lambda_{\min,\perp}(\boldsymbol{L}_\mathcal{G})\geq\lambda+\frac{c_2pn}{8\kappa_1e^C}.$$

This completes the proof of Lemma 17. ■

C.5 Proof of Lemma 18

Proof Since $\mathcal{L}_\lambda(\cdot)$ is λ-strongly convex and $\lambda + \frac{1}{2}c_1 np$-smooth on the event $\mathcal{A}_2$, we know that

$$\left\|\widetilde{\boldsymbol{\beta}}^t - \eta\nabla_\lambda(\widetilde{\boldsymbol{\beta}}_\lambda) - \widetilde{\boldsymbol{\beta}}^*\right\|_2 \leq \rho\left\|\widetilde{\boldsymbol{\beta}}^t - \widetilde{\boldsymbol{\beta}}_\lambda\right\|_2.$$

As a result, when event $\mathcal{A}_2$ happens, we have

$$\left\|\widetilde{\boldsymbol{\beta}}^{t+1} - \widetilde{\boldsymbol{\beta}}^*\right\|_2 = \left\|\mathcal{P}\left(\widetilde{\boldsymbol{\beta}}^t - \eta\nabla_\lambda(\widetilde{\boldsymbol{\beta}}_\lambda) - \widetilde{\boldsymbol{\beta}}^*\right)\right\|_2 \leq \left\|\widetilde{\boldsymbol{\beta}}^t - \eta\nabla_\lambda(\widetilde{\boldsymbol{\beta}}_\lambda) - \widetilde{\boldsymbol{\beta}}^*\right\|_2 \leq \rho\left\|\widetilde{\boldsymbol{\beta}}^t - \widetilde{\boldsymbol{\beta}}_\lambda\right\|_2.$$

Therefore, under event $\mathcal{A}_2$, we have

$$\left\|\widetilde{\boldsymbol{\beta}}^t - \widetilde{\boldsymbol{\beta}}^*\right\|_2 \leq \rho^t\left\|\widetilde{\boldsymbol{\beta}}^0 - \widetilde{\boldsymbol{\beta}}^*\right\|_2.$$

■

C.6 Proof of Lemma 19

In this section, we prove $\widetilde{\boldsymbol{\beta}}_\lambda$ is not far from the initial point $\widetilde{\boldsymbol{\beta}}^*$.

Proof Since $\widetilde{\boldsymbol{\beta}}_\lambda$ is the minimizer, we have that $\mathcal{L}_\lambda(\widetilde{\boldsymbol{\beta}}^*) \geq \mathcal{L}_\lambda(\widetilde{\boldsymbol{\beta}}_\lambda)$. By the mean value theorem, for some $\widetilde{\boldsymbol{\beta}}'$ between $\widetilde{\boldsymbol{\beta}}^*$ and $\widetilde{\boldsymbol{\beta}}_\lambda$, we have

$$\mathcal{L}_\lambda(\widetilde{\boldsymbol{\beta}}_\lambda) = \mathcal{L}_\lambda(\widetilde{\boldsymbol{\beta}}^*) + \nabla\mathcal{L}_\lambda(\widetilde{\boldsymbol{\beta}}^*)^\top(\widetilde{\boldsymbol{\beta}}_\lambda - \widetilde{\boldsymbol{\beta}}^*) + \frac{1}{2}(\widetilde{\boldsymbol{\beta}}_\lambda - \widetilde{\boldsymbol{\beta}}^*)^\top\nabla^2\mathcal{L}_\lambda(\widetilde{\boldsymbol{\beta}}')(\widetilde{\boldsymbol{\beta}}_\lambda - \widetilde{\boldsymbol{\beta}}^*).$$

As a result, we have

$$\begin{aligned}\mathcal{L}_\lambda(\widetilde{\boldsymbol{\beta}}^*) &\geq \mathcal{L}_\lambda(\widetilde{\boldsymbol{\beta}}^*) + \nabla\mathcal{L}_\lambda(\widetilde{\boldsymbol{\beta}}^*)^\top(\widetilde{\boldsymbol{\beta}}_\lambda - \widetilde{\boldsymbol{\beta}}^*) + \frac{1}{2}(\widetilde{\boldsymbol{\beta}}_\lambda - \widetilde{\boldsymbol{\beta}}^*)^\top\nabla^2\mathcal{L}_\lambda(\widetilde{\boldsymbol{\beta}}^*)(\widetilde{\boldsymbol{\beta}}_\lambda - \widetilde{\boldsymbol{\beta}}^*) \\ &\geq \mathcal{L}_\lambda(\widetilde{\boldsymbol{\beta}}^*) + \nabla\mathcal{L}_\lambda(\widetilde{\boldsymbol{\beta}}^*)^\top(\widetilde{\boldsymbol{\beta}}_\lambda - \widetilde{\boldsymbol{\beta}}^*) + \frac{\lambda}{2}\left\|\widetilde{\boldsymbol{\beta}}_\lambda - \widetilde{\boldsymbol{\beta}}^*\right\|_2^2.\end{aligned}$$

Therefore, we get

$$\begin{aligned}\frac{\lambda}{2}\left\|\widetilde{\boldsymbol{\beta}}_\lambda - \widetilde{\boldsymbol{\beta}}^*\right\|_2^2 &\leq -\nabla\mathcal{L}_\lambda(\widetilde{\boldsymbol{\beta}}^*)^\top(\widetilde{\boldsymbol{\beta}}_\lambda - \widetilde{\boldsymbol{\beta}}^*) \\ &\leq \left\|\nabla\mathcal{L}_\lambda(\widetilde{\boldsymbol{\beta}}^*)\right\|_2\left\|\widetilde{\boldsymbol{\beta}}_\lambda - \widetilde{\boldsymbol{\beta}}^*\right\|_2.\end{aligned}$$

As a result, on event $\mathcal{A}_1$ we have

$$\left\|\widetilde{\boldsymbol{\beta}}_\lambda - \widetilde{\boldsymbol{\beta}}^*\right\|_2 \leq \frac{2\left\|\nabla\mathcal{L}_\lambda(\widetilde{\boldsymbol{\beta}}^*)\right\|_2}{\lambda} \leq \frac{2C_0}{c_\lambda}\max\left\{\frac{\kappa_2}{\kappa_1}, \kappa_3\sqrt{d+1}\right\}\sqrt{n}.$$

We conclude the proof of Lemma 19. ■

C.7 Proof of Lemma 20

Proof Combine Lemma 18 and Lemma 19 we have

$$\begin{aligned}\left\|\widetilde{\boldsymbol{\beta}}^T-\widetilde{\boldsymbol{\beta}}_\lambda\right\|_2 &\le \rho^T\left\|\widetilde{\boldsymbol{\beta}}^0-\widetilde{\boldsymbol{\beta}}_\lambda\right\|_2\\ &\le \left(1-\frac{2\lambda}{2\lambda+c_1np}\right)^{n^5}\frac{2C_0}{c_\lambda}\max\left\{\frac{\kappa_2}{\kappa_1},\kappa_3\sqrt{d+1}\right\}\sqrt{n}\\ &\le \frac{2C_0}{c_\lambda}\max\left\{\frac{\kappa_2}{\kappa_1},\kappa_3\sqrt{d+1}\right\}\sqrt{n}\exp\left(-\frac{2\lambda n^5}{2\lambda+c_1np}\right)\\ &\le C_7\kappa_1\sqrt{\frac{(d+1)\log n}{npL}}\end{aligned}$$

for $L\le c_4\cdot n^{c_5}$ and n which is large enough. ■

C.8 Proof of Lemma 21

Proof By definition we know that

$$\widetilde{\boldsymbol{\beta}}^{t+1}-\widetilde{\boldsymbol{\beta}}^*=\mathcal{P}\left(\widetilde{\boldsymbol{\beta}}^t-\eta\nabla\mathcal{L}_\lambda(\widetilde{\boldsymbol{\beta}}^t)\right)-\widetilde{\boldsymbol{\beta}}^*=\mathcal{P}\left(\widetilde{\boldsymbol{\beta}}^t-\eta\nabla\mathcal{L}_\lambda(\widetilde{\boldsymbol{\beta}}^t)-\widetilde{\boldsymbol{\beta}}^*\right).$$

Consider $\widetilde{\boldsymbol{\beta}}(\tau)=\widetilde{\boldsymbol{\beta}}^*+\tau\left(\widetilde{\boldsymbol{\beta}}^t-\widetilde{\boldsymbol{\beta}}^*\right)$. By the fundamental theorem of calculus we have

$$\begin{aligned}\widetilde{\boldsymbol{\beta}}^t-\eta\nabla\mathcal{L}_\lambda(\widetilde{\boldsymbol{\beta}}^t)-\widetilde{\boldsymbol{\beta}}^* &= \widetilde{\boldsymbol{\beta}}^t-\eta\nabla\mathcal{L}_\lambda(\widetilde{\boldsymbol{\beta}}^t)-\left[\widetilde{\boldsymbol{\beta}}^*-\eta\nabla\mathcal{L}_\lambda(\widetilde{\boldsymbol{\beta}}^*)\right]-\eta\nabla\mathcal{L}_\lambda(\widetilde{\boldsymbol{\beta}}^*)\\ &= \left\{\boldsymbol{I}_{n+d}-\eta\int_0^1\nabla^2\mathcal{L}_\lambda(\widetilde{\boldsymbol{\beta}}(\tau))d\tau\right\}\left(\widetilde{\boldsymbol{\beta}}^t-\widetilde{\boldsymbol{\beta}}^*\right)-\eta\nabla\mathcal{L}_\lambda(\widetilde{\boldsymbol{\beta}}^*).\end{aligned}$$

Let npL be large enough such that

$$2C_6\kappa_1^2\sqrt{\frac{(d+1)\log n}{npL}}\le 0.1,\quad 2C_3\kappa_1\sqrt{c_3}\sqrt{\frac{(d+1)\log n}{npL}}\le 0.1.$$

In addition, by the assumption of induction, we have

$$\|\boldsymbol{\alpha}(\tau)-\boldsymbol{\alpha}^*\|_\infty\le 0.05,\quad \|\boldsymbol{\beta}(\tau)-\boldsymbol{\beta}^*\|_2\le 0.05\sqrt{\frac{n}{c_3(d+1)}}.$$

Then by Lemma 17, we have

$$\lambda_{\min,\perp}\left(\nabla^2\mathcal{L}_\lambda(\widetilde{\boldsymbol{\beta}}(\tau))\right)\ge\lambda+\frac{c_2pn}{8\kappa_1e^{0.2}}\ge\lambda+\frac{c_2pn}{10\kappa_1},\quad\forall 0\le\tau\le 1.$$

On the other hand, by Lemma 16, we have

$$\lambda_{\max}\left(\nabla^2\mathcal{L}_\lambda(\widetilde{\boldsymbol{\beta}}(\tau))\right)\le\lambda+\frac{1}{2}c_1pn.$$

Let $\boldsymbol{A} = \int_0^1 \nabla^2 \mathcal{L}_\lambda(\widetilde{\boldsymbol{\beta}}(\tau))d\tau$, then we have

$$\lambda + \frac{c_2 pn}{10\kappa_1} \le \lambda_{\min,\perp}(A) \le \lambda_{\max}(A) \le \lambda + \frac{1}{2}c_1 pn.$$

Since $\widetilde{\boldsymbol{\beta}}^t - \widetilde{\boldsymbol{\beta}}^* \in \Theta$, it holds that

$$\begin{aligned}\left\|\mathcal{P}(\boldsymbol{I}_{n+d} - \eta\boldsymbol{A})(\widetilde{\boldsymbol{\beta}}^t - \widetilde{\boldsymbol{\beta}}^*)\right\|_2 &\le \max\{|1-\eta\lambda_{\min,\perp}(A), 1-\eta\lambda_{\max}(A)|\}\left\|\widetilde{\boldsymbol{\beta}}^t - \widetilde{\boldsymbol{\beta}}^*\right\|_2 \\ &\le \left(1 - \frac{c_2}{20\kappa_1}\eta pn\right)\left\|\widetilde{\boldsymbol{\beta}}^t - \widetilde{\boldsymbol{\beta}}^*\right\|_2.\end{aligned}$$

Therefore, on the event $\mathcal{A}_1$, we have

$$\begin{aligned}\left\|\widetilde{\boldsymbol{\beta}}^{t+1} - \widetilde{\boldsymbol{\beta}}^*\right\|_2 &\le \left\|\mathcal{P}\left(\boldsymbol{I}_{n+d} - \eta\boldsymbol{A}\right)(\widetilde{\boldsymbol{\beta}}^t - \widetilde{\boldsymbol{\beta}}^*) - \eta\mathcal{P}\nabla\mathcal{L}_\lambda(\widetilde{\boldsymbol{\beta}}^*)\right\|_2 \\ &\le \left\|\mathcal{P}\left(\boldsymbol{I}_{n+d} - \eta\boldsymbol{A}\right)(\widetilde{\boldsymbol{\beta}}^t - \widetilde{\boldsymbol{\beta}}^*)\right\|_2 + \eta\left\|\nabla\mathcal{L}_\lambda(\widetilde{\boldsymbol{\beta}}^*)\right\|_2 \\ &\le \left(1 - \frac{c_2}{20\kappa_1}\eta pn\right)\left\|\widetilde{\boldsymbol{\beta}}^t - \widetilde{\boldsymbol{\beta}}^*\right\|_2 + C_0\eta\sqrt{\frac{n^2 p\log n}{L}} \\ &\le \left(1 - \frac{c_2}{20\kappa_1}\eta pn\right)C_3\kappa_1\sqrt{\frac{\log n}{pL}} + C_0\eta\sqrt{\frac{n^2 p\log n}{L}} \\ &\le C_3\kappa_1\sqrt{\frac{\log n}{pL}},\end{aligned}$$

as long as $C_3 \ge \frac{20C_0}{c_2}$. ■

C.9 Proof of Lemma 22

Proof For any $m \in [n]$, by definition we have

$$\widetilde{\boldsymbol{\beta}}^{t+1} - \widetilde{\boldsymbol{\beta}}^{t+1,(m)} = \mathcal{P}\left(\widetilde{\boldsymbol{\beta}}^t - \eta\nabla\mathcal{L}_\lambda(\widetilde{\boldsymbol{\beta}}^t) - \left[\widetilde{\boldsymbol{\beta}}^{t,(m)} - \eta\nabla\mathcal{L}_\lambda^{(m)}(\widetilde{\boldsymbol{\beta}}^{t,(m)})\right]\right).$$

We consider $\widetilde{\boldsymbol{\beta}}(\tau) = \widetilde{\boldsymbol{\beta}}^{t,(m)} + \tau\left(\widetilde{\boldsymbol{\beta}}^t - \widetilde{\boldsymbol{\beta}}^{t,(m)}\right)$. By the fundamental theorem of calculus we have

$$\begin{aligned}&\widetilde{\boldsymbol{\beta}}^t - \eta\nabla\mathcal{L}_\lambda(\widetilde{\boldsymbol{\beta}}^t) - \left[\widetilde{\boldsymbol{\beta}}^{t,(m)} - \eta\nabla\mathcal{L}_\lambda^{(m)}(\widetilde{\boldsymbol{\beta}}^{t,(m)})\right] &(44)\\ =&\widetilde{\boldsymbol{\beta}}^t - \eta\nabla\mathcal{L}_\lambda(\widetilde{\boldsymbol{\beta}}^t) - \left[\widetilde{\boldsymbol{\beta}}^{t,(m)} - \eta\nabla\mathcal{L}_\lambda(\widetilde{\boldsymbol{\beta}}^{t,(m)})\right] - \eta\left(\nabla\mathcal{L}_\lambda(\widetilde{\boldsymbol{\beta}}^{t,(m)}) - \nabla\mathcal{L}_\lambda^{(m)}(\widetilde{\boldsymbol{\beta}}^{t,(m)})\right) &(45)\\ =&\left\{\boldsymbol{I}_{n+d} - \eta\int_0^1 \nabla^2\mathcal{L}_\lambda(\widetilde{\boldsymbol{\beta}}(\tau))d\tau\right\}\left(\widetilde{\boldsymbol{\beta}}^t - \widetilde{\boldsymbol{\beta}}^{t,(m)}\right) - \eta\left(\nabla\mathcal{L}_\lambda(\widetilde{\boldsymbol{\beta}}^{t,(m)}) - \nabla\mathcal{L}_\lambda^{(m)}(\widetilde{\boldsymbol{\beta}}^{t,(m)})\right). &(46)\end{aligned}$$

From (A)$\sim$ (D) we know that

$$\|\boldsymbol{\alpha}^{t,(m)}-\boldsymbol{\alpha}^*\|_\infty \le \|\boldsymbol{\alpha}^{t}-\boldsymbol{\alpha}^*\|_\infty + \max_{1\le m\le n}\left\|\widetilde{\boldsymbol{\beta}}^{t,(m)}-\widetilde{\boldsymbol{\beta}}^{t}\right\|_2 \le (C_4+C_6)\kappa_1^2\sqrt{\frac{(d+1)\log n}{npL}};$$

$$\|\boldsymbol{\beta}^{t,(m)}-\boldsymbol{\beta}^*\|_2 \le \left\|\widetilde{\boldsymbol{\beta}}^{t}-\widetilde{\boldsymbol{\beta}}^{*}\right\|_2 + \max_{1\le m\le n}\left\|\widetilde{\boldsymbol{\beta}}^{t,(m)}-\widetilde{\boldsymbol{\beta}}^{t}\right\|_2 \le (C_3+C_4)\kappa_1\sqrt{\frac{\log n}{pL}};$$

$$\|\boldsymbol{\alpha}^{t}-\boldsymbol{\alpha}^*\|_\infty \le C_6\kappa_1^2\sqrt{\frac{(d+1)\log n}{npL}};$$

$$\|\boldsymbol{\beta}^{t}-\boldsymbol{\beta}^*\|_2 \le \left\|\widetilde{\boldsymbol{\beta}}^{t}-\widetilde{\boldsymbol{\beta}}^{*}\right\|_2 \le C_3\kappa_1\sqrt{\frac{\log n}{pL}}.$$

Consider npL which is large enough such that

$$2(C_4+C_6)\kappa_1^2\sqrt{\frac{(d+1)\log n}{npL}},\ 2(C_3+C_4)\kappa_1\sqrt{c_3}\sqrt{\frac{(d+1)\log n}{npL}} \le 0.1.$$

Then we also have

$$2C_6\kappa_1^2\sqrt{\frac{(d+1)\log n}{npL}},\ 2C_3\kappa_1\sqrt{c_3}\sqrt{\frac{(d+1)\log n}{npL}} \le 0.1.$$

Use the same approach derived in §C.8, we have

$$\left\|\mathcal{P}\left\{\boldsymbol{I}_{n+d}-\eta\int_0^1\nabla^2\mathcal{L}_\lambda(\widetilde{\boldsymbol{\beta}}(\tau))d\tau\right\}\left(\widetilde{\boldsymbol{\beta}}^{t}-\widetilde{\boldsymbol{\beta}}^{t,(m)}\right)\right\|_2 \le \left(1-\frac{c_2}{20\kappa_1}\eta pn\right)\left\|\widetilde{\boldsymbol{\beta}}^{t}-\widetilde{\boldsymbol{\beta}}^{t,(m)}\right\|_2, \tag{47}$$

as long as $0<\eta\le\dfrac{2}{2\lambda+c_1np}$.

On the other hand, since $\left\|\mathcal{P}(\nabla\mathcal{L}_\lambda(\widetilde{\boldsymbol{\beta}}^{t,(m)})-\nabla\mathcal{L}_\lambda^{(m)}(\widetilde{\boldsymbol{\beta}}^{t,(m)}))\right\|_2 \le \left\|\nabla\mathcal{L}_\lambda(\widetilde{\boldsymbol{\beta}}^{t,(m)})-\nabla\mathcal{L}_\lambda^{(m)}(\widetilde{\boldsymbol{\beta}}^{t,(m)})\right\|_2$, it remains to bound $\left\|\nabla\mathcal{L}_\lambda(\widetilde{\boldsymbol{\beta}}^{t,(m)})-\nabla\mathcal{L}_\lambda^{(m)}(\widetilde{\boldsymbol{\beta}}^{t,(m)})\right\|_2$. By definition, we have

$$\begin{aligned}
&\nabla\mathcal{L}_\lambda(\widetilde{\boldsymbol{\beta}}^{t,(m)})-\nabla\mathcal{L}_\lambda^{(m)}(\widetilde{\boldsymbol{\beta}}^{t,(m)})\\
=&\sum_{i\neq m}\left\{\left(-y_{m,i}+\frac{e^{\widetilde{\boldsymbol{x}}_i^\top\widetilde{\boldsymbol{\beta}}^{t,(m)}}}{e^{\widetilde{\boldsymbol{x}}_i^\top\widetilde{\boldsymbol{\beta}}^{t,(m)}}+e^{\widetilde{\boldsymbol{x}}_m^\top\widetilde{\boldsymbol{\beta}}^{t,(m)}}}\right)\mathbf{1}((i,m)\in\mathcal{E})-p\left(-y_{m,i}^*+\frac{e^{\widetilde{\boldsymbol{x}}_i^\top\widetilde{\boldsymbol{\beta}}^{t,(m)}}}{e^{\widetilde{\boldsymbol{x}}_i^\top\widetilde{\boldsymbol{\beta}}^{t,(m)}}+e^{\widetilde{\boldsymbol{x}}_m^\top\widetilde{\boldsymbol{\beta}}^{t,(m)}}}\right)\right\}(\widetilde{\boldsymbol{x}}_i-\widetilde{\boldsymbol{x}}_m)\\
=&\underbrace{\sum_{i\neq m}\left\{\left(-\frac{e^{\widetilde{\boldsymbol{x}}_i^\top\widetilde{\boldsymbol{\beta}}^{*}}}{e^{\widetilde{\boldsymbol{x}}_i^\top\widetilde{\boldsymbol{\beta}}^{*}}+e^{\widetilde{\boldsymbol{x}}_m^\top\widetilde{\boldsymbol{\beta}}^{*}}}+\frac{e^{\widetilde{\boldsymbol{x}}_i^\top\widetilde{\boldsymbol{\beta}}^{t,(m)}}}{e^{\widetilde{\boldsymbol{x}}_i^\top\widetilde{\boldsymbol{\beta}}^{t,(m)}}+e^{\widetilde{\boldsymbol{x}}_m^\top\widetilde{\boldsymbol{\beta}}^{t,(m)}}}\right)(\mathbf{1}((i,m)\in\mathcal{E})-p)\right\}(\widetilde{\boldsymbol{x}}_i-\widetilde{\boldsymbol{x}}_m)}_{:=\boldsymbol{u}^m}\\
&+\underbrace{\frac{1}{L}\sum_{(i,m)\in\mathcal{E}}\sum_{l=1}^{L}\left(-y_{m,i}^{(l)}+\frac{e^{\widetilde{\boldsymbol{x}}_i^\top\widetilde{\boldsymbol{\beta}}^{*}}}{e^{\widetilde{\boldsymbol{x}}_i^\top\widetilde{\boldsymbol{\beta}}^{*}}+e^{\widetilde{\boldsymbol{x}}_m^\top\widetilde{\boldsymbol{\beta}}^{*}}}\right)(\widetilde{\boldsymbol{x}}_i-\widetilde{\boldsymbol{x}}_m)}_{:=\boldsymbol{v}^m}.
\end{aligned}$$

By definition, we also have

$$v_j^m = \begin{cases} \frac{1}{L}\sum_{l=1}^{L}\left(-y_{m,j}^{(l)} + \frac{e^{\widetilde{\boldsymbol{x}}_j^\top \widetilde{\boldsymbol{\beta}}^*}}{e^{\widetilde{\boldsymbol{x}}_j^\top \widetilde{\boldsymbol{\beta}}^*}+e^{\widetilde{\boldsymbol{x}}_m^\top \widetilde{\boldsymbol{\beta}}^*}}\right), & \text{if } (j,m)\in\mathcal{E} \\ \frac{1}{L}\sum_{i:(i,m)\in\mathcal{E}}\sum_{l=1}^{L}\left(y_{m,i}^{(l)} - \frac{e^{\widetilde{\boldsymbol{x}}_i^\top \widetilde{\boldsymbol{\beta}}^*}}{e^{\widetilde{\boldsymbol{x}}_i^\top \widetilde{\boldsymbol{\beta}}^*}+e^{\widetilde{\boldsymbol{x}}_m^\top \widetilde{\boldsymbol{\beta}}^*}}\right), & \text{if } j=m; \\ \frac{1}{L}\sum_{i:(i,m)\in\mathcal{E}}\sum_{l=1}^{L}\left(-y_{m,i}^{(l)} + \frac{e^{\widetilde{\boldsymbol{x}}_i^\top \widetilde{\boldsymbol{\beta}}^*}}{e^{\widetilde{\boldsymbol{x}}_i^\top \widetilde{\boldsymbol{\beta}}^*}+e^{\widetilde{\boldsymbol{x}}_m^\top \widetilde{\boldsymbol{\beta}}^*}}\right)((\widetilde{\boldsymbol{x}}_i)_j - (\widetilde{\boldsymbol{x}}_m)_j), & \text{if } j>n; \\ 0, & \text{else.} \end{cases}$$

Consider random variable $M = |\{i:(i,m)\in\mathcal{E}\}|$. By Chernoff bound (Tropp, 2012), we know that

$$\mathbb{P}(M \geq 2pn) \leq (e/4)^{pn} \leq O(n^{-11}),$$

as long as $np > c_p \log n$ for some $c_p > 0$. As long as $\|\boldsymbol{x}_i - \boldsymbol{x}_m\|_2 \leq 2\sqrt{c_3(d+1)/n} \leq 1$, we have $|(\widetilde{\boldsymbol{x}}_i)_j - (\widetilde{\boldsymbol{x}}_m)_j)| \leq 1$ for $j > n$. Since $\left|-y_{m,i}^{(l)} + \frac{e^{\widetilde{\boldsymbol{x}}_i^\top \widetilde{\boldsymbol{\beta}}^*}}{e^{\widetilde{\boldsymbol{x}}_i^\top \widetilde{\boldsymbol{\beta}}^*}+e^{\widetilde{\boldsymbol{x}}_m^\top \widetilde{\boldsymbol{\beta}}^*}}\right| \leq 1$, by Hoeffding's inequality and union bound, we get

$$|v_j^m| \lesssim \sqrt{\frac{M\log n}{L}}, \text{ if } j=m \text{ or } j>n;$$
$$|v_j^m| \lesssim \sqrt{\frac{\log n}{L}}, \text{ if } (j,m)\in\mathcal{E}.$$

with probability exceeding $1 - O(n^{-11})$ conditioning on $\mathcal{E}$ as long as $d < n$. On the other hand, since $M \leq 2pn$ with probability exceeding $1 - O(n^{-11})$, we have

$$\|\boldsymbol{v}^m\|_2^2 \lesssim (d+1)\frac{2pn\log n}{L} + 2pn\frac{\log n}{L} \lesssim \frac{pn(d+1)\log n}{L}$$

with probability exceeding $1 - O(n^{-11})$.

On the other hand, for $\boldsymbol{u}^m$ we have

$$u_j^m = \begin{cases} \xi_j(1-p), & \text{if } (j,m)\in\mathcal{E} \\ -\sum_{i:(i,m)\in\mathcal{E}}\xi_i\left(\mathbf{1}((i,m)\in\mathcal{E}) - p\right), & \text{if } j=m; \\ \sum_{i:(i,m)\in\mathcal{E}}\xi_i\left(\mathbf{1}((i,m)\in\mathcal{E}) - p\right)((\widetilde{\boldsymbol{x}}_i)_j - (\widetilde{\boldsymbol{x}}_m)_j), & \text{if } j>n; \\ -\xi_j p, & \text{else,} \end{cases}$$

where

$$\xi_j = -\frac{e^{\widetilde{\boldsymbol{x}}_j^\top \widetilde{\boldsymbol{\beta}}^*}}{e^{\widetilde{\boldsymbol{x}}_j^\top \widetilde{\boldsymbol{\beta}}^*}+e^{\widetilde{\boldsymbol{x}}_m^\top \widetilde{\boldsymbol{\beta}}^*}} + \frac{e^{\widetilde{\boldsymbol{x}}_j^\top \widetilde{\boldsymbol{\beta}}^{t,(m)}}}{e^{\widetilde{\boldsymbol{x}}_j^\top \widetilde{\boldsymbol{\beta}}^{t,(m)}}+e^{\widetilde{\boldsymbol{x}}_m^\top \widetilde{\boldsymbol{\beta}}^{t,(m)}}} = -\frac{1}{1+e^{\widetilde{\boldsymbol{x}}_m^\top \widetilde{\boldsymbol{\beta}}^* - \widetilde{\boldsymbol{x}}_j^\top \widetilde{\boldsymbol{\beta}}^*}} + \frac{1}{1+e^{\widetilde{\boldsymbol{x}}_m^\top \widetilde{\boldsymbol{\beta}}^{t,(m)} - \widetilde{\boldsymbol{x}}_j^\top \widetilde{\boldsymbol{\beta}}^{t,(m)}}}.$$

Consider $g(x) = \dfrac{1}{1+e^x}$. Since $\left|g'(x)\right| \le 1$, we have that

$$\begin{aligned}
|\xi_j| &= \left|g(\widetilde{\boldsymbol{x}}_m^\top \widetilde{\boldsymbol{\beta}}^{t,(m)} - \widetilde{\boldsymbol{x}}_j^\top \widetilde{\boldsymbol{\beta}}^{t,(m)}) - g(\widetilde{\boldsymbol{x}}_m^\top \widetilde{\boldsymbol{\beta}}^{*} - \widetilde{\boldsymbol{x}}_j^\top \widetilde{\boldsymbol{\beta}}^{*})\right| \\
&\le \left|(\widetilde{\boldsymbol{x}}_m^\top \widetilde{\boldsymbol{\beta}}^{t,(m)} - \widetilde{\boldsymbol{x}}_j^\top \widetilde{\boldsymbol{\beta}}^{t,(m)}) - (\widetilde{\boldsymbol{x}}_m^\top \widetilde{\boldsymbol{\beta}}^{*} - \widetilde{\boldsymbol{x}}_j^\top \widetilde{\boldsymbol{\beta}}^{*})\right| \\
&\le \left|\widetilde{\boldsymbol{x}}_m^\top \widetilde{\boldsymbol{\beta}}^{t,(m)} - \widetilde{\boldsymbol{x}}_m^\top \widetilde{\boldsymbol{\beta}}^{*}\right| + \left|\widetilde{\boldsymbol{x}}_j^\top \widetilde{\boldsymbol{\beta}}^{t,(m)} - \widetilde{\boldsymbol{x}}_j^\top \widetilde{\boldsymbol{\beta}}^{*}\right| \\
&\le \left|\alpha_m^{t,(m)} - \alpha_m^*\right| + \left|\boldsymbol{x}_m^\top \boldsymbol{\beta}^{t,(m)} - \boldsymbol{x}_m^\top \boldsymbol{\beta}^{*}\right| + \left|\alpha_j^{t,(m)} - \alpha_j^*\right| + \left|\boldsymbol{x}_j^\top \boldsymbol{\beta}^{t,(m)} - \boldsymbol{x}_j^\top \boldsymbol{\beta}^{*}\right| \\
&\le 2\left\|\boldsymbol{\alpha}^{t,(m)} - \boldsymbol{\alpha}^{*}\right\|_\infty + 2\sqrt{c_3(d+1)/n}\left\|\boldsymbol{\beta}^{t,(m)} - \boldsymbol{\beta}^{*}\right\|_2 \\
&\le \left[2(C_4+C_6)\kappa_1^2 + 2(C_3+C_4)\kappa_1\sqrt{c_3}\right]\sqrt{\frac{(d+1)\log n}{npL}} := \widetilde{C}_1\sqrt{\frac{(d+1)\log n}{npL}}.
\end{aligned}$$

By Bernstein inequality we know that

$$\begin{aligned}
|u_j^m| &\lesssim \sqrt{\left(p\sum_{i=1}^n \xi_i^2\right)\log n} + \max_{1\le i\le n}|\xi_i|\log n \\
&\le \left(\sqrt{np\log n} + \log n\right)\widetilde{C}_1\sqrt{\frac{(d+1)\log n}{npL}},\ \text{if } j = m \text{ or } j > n.
\end{aligned}$$

As a result, for $\boldsymbol{u}^m$ we have

$$\begin{aligned}
\|\boldsymbol{u}^m\|_2^2 &= (u_m^m)^2 + \sum_{j>n}(u_j^m)^2 + \sum_{j:(j,m)\in\mathcal{E}}(u_j^m)^2 + \sum_{j:(j,m)\notin\mathcal{E}, j\ne m, j\le n}(u_j^m)^2 \\
&\lesssim (d+1)\left(\sqrt{np\log n} + \log n\right)^2 \widetilde{C}_1^2\frac{(d+1)\log n}{npL} + np\widetilde{C}_1^2\frac{(d+1)\log n}{npL} + p^2 n\widetilde{C}_1^2\frac{(d+1)\log n}{npL} \\
&\lesssim pn(d+1)\log n\widetilde{C}_1^2\frac{(d+1)\log n}{npL}.
\end{aligned}$$

In summary, there exists constants D_1, D_2 which are independent of $C_i, i \ge 0$ such that

$$\|\boldsymbol{v}^m\|_2 \le D_1\sqrt{\frac{pn(d+1)\log n}{L}},\quad \|\boldsymbol{u}^m\|_2 \le D_2\widetilde{C}_1(d+1)\log n\sqrt{\frac{1}{L}} \tag{48}$$

with probability exceeding $1-O(n^{-11})$. Combining Eq. (46), Eq. (47) and Eq. (48) we have

$$\begin{aligned}\left\|\widetilde{\boldsymbol{\beta}}^{t+1}-\widetilde{\boldsymbol{\beta}}^{t+1,(m)}\right\|_2 \leq & \left(1-\frac{c_2}{20\kappa_1}\eta pn\right)\left\|\widetilde{\boldsymbol{\beta}}^{t}-\widetilde{\boldsymbol{\beta}}^{t,(m)}\right\|_2 \\ & +\eta\left(D_1\sqrt{\frac{pn(d+1)\log n}{L}}+D_2\widetilde{C}_1(d+1)\log n\sqrt{\frac{1}{L}}\right) \\ \leq & \left(1-\frac{c_2}{20\kappa_1}\eta pn\right)C_4\kappa_1\sqrt{\frac{(d+1)\log n}{npL}} \\ & +\eta\left(D_1\sqrt{\frac{pn(d+1)\log n}{L}}+D_2\widetilde{C}_1(d+1)\log n\sqrt{\frac{1}{L}}\right) \\ \leq & C_4\kappa_1\sqrt{\frac{(d+1)\log n}{npL}},\end{aligned}$$

as long as $C_4 \geq \frac{40D_1}{c_2}$ and n is large enough such that $C_4 \geq \frac{40D_2}{c_2}\widetilde{C}_1\sqrt{\frac{(d+1)\log n}{np}}$. ■

C.10 Proof of Lemma 23

Proof For $m \in [n]$, we have

$$\begin{aligned}\alpha_m^{t+1,(m)}-\alpha_m^* &= \left[\mathcal{P}\left(\widetilde{\boldsymbol{\beta}}^{t,(m)}-\eta\nabla\mathcal{L}_\lambda^{(m)}\left(\widetilde{\boldsymbol{\beta}}^{t,(m)}\right)-\widetilde{\boldsymbol{\beta}}^*\right)\right]_m \\ &= \alpha_m^{t,(m)}-\eta\left[\mathcal{P}\nabla\mathcal{L}_\lambda^{(m)}\left(\widetilde{\boldsymbol{\beta}}^{t,(m)}\right)\right]_m-\alpha_m^* \\ &= \underbrace{\alpha_m^{t,(m)}-\eta\left[\nabla\mathcal{L}_\lambda^{(m)}\left(\widetilde{\boldsymbol{\beta}}^{t,(m)}\right)\right]_m-\alpha_m^*}_{\mu_1}+\underbrace{\eta\left[(I-\mathcal{P})\nabla\mathcal{L}_\lambda^{(m)}\left(\widetilde{\boldsymbol{\beta}}^{t,(m)}\right)\right]_m}_{\mu_2}\end{aligned}$$

For μ_2, we have

$$\begin{aligned}|\mu_2| &\leq \eta\left\|(I-\mathcal{P})\nabla\mathcal{L}_\lambda^{(m)}\left(\widetilde{\boldsymbol{\beta}}^{t,(m)}\right)\right\|_\infty \\ &\leq \eta\left\|I-\mathcal{P}\right\|_{2,\infty}\left\|\nabla\mathcal{L}_\lambda^{(m)}\left(\widetilde{\boldsymbol{\beta}}^{t,(m)}\right)\right\|_2 \\ &\leq c_0\eta\sqrt{\frac{d+1}{n}}\left\|\nabla\mathcal{L}_\lambda^{(m)}\left(\widetilde{\boldsymbol{\beta}}^{t,(m)}\right)\right\|_2.\end{aligned}$$

By Lemma 34, with probability at least $1-O(n^{-11})$ we have

$$\left\|\nabla\mathcal{L}_\lambda^{(m)}\left(\widetilde{\boldsymbol{\beta}}^{t,(m)}\right)\right\|_2 \leq (C_0+c_1C_3+c_1C_4)\kappa_1\sqrt{\frac{n^2p\log n}{L}},$$

as long as

$$npL \geq \frac{4c_\lambda^2}{c_1^2}\min\left\{\frac{\kappa_1^2}{\kappa_2^2},\frac{1}{(d+1)\kappa_3^2}\right\}\log n.$$

In this case, for μ_2 we have

$$|\mu_2| \le c_0(C_0 + c_1C_3 + c_1C_4)\eta\kappa_1\sqrt{\frac{(d+1)np\log n}{L}} := \widetilde{C}_2\eta\kappa_1\sqrt{\frac{(d+1)np\log n}{L}}. \tag{49}$$

On the other hand, for μ_1 we have

$$\mu_1 = \alpha_m^{t,(m)} - \eta\left[\nabla\mathcal{L}_\lambda^{(m)}\left(\widetilde{\boldsymbol{\beta}}^{t,(m)}\right)\right]_m - \alpha_m^* \tag{50}$$

$$= \alpha_m^{t,(m)} - \alpha_m^* - \eta p\sum_{i\neq m}\left\{\frac{e^{\widetilde{\boldsymbol{x}}_i^\top\widetilde{\boldsymbol{\beta}}^*}}{e^{\widetilde{\boldsymbol{x}}_i^\top\widetilde{\boldsymbol{\beta}}^*} + e^{\widetilde{\boldsymbol{x}}_m^\top\widetilde{\boldsymbol{\beta}}^*}} - \frac{e^{\widetilde{\boldsymbol{x}}_i^\top\widetilde{\boldsymbol{\beta}}^{t,(m)}}}{e^{\widetilde{\boldsymbol{x}}_i^\top\widetilde{\boldsymbol{\beta}}^{t,(m)}} + e^{\widetilde{\boldsymbol{x}}_m^\top\widetilde{\boldsymbol{\beta}}^{t,(m)}}}\right\} - \eta\lambda\alpha_m^{t,(m)}. \tag{51}$$

Normalizing the numerators below to 1 and by the mean value theorem, there exists some c_i between $\widetilde{\boldsymbol{x}}_m^\top\widetilde{\boldsymbol{\beta}}^* - \widetilde{\boldsymbol{x}}_i^\top\widetilde{\boldsymbol{\beta}}^*$ and $\widetilde{\boldsymbol{x}}_m^\top\widetilde{\boldsymbol{\beta}}^{t,(m)} - \widetilde{\boldsymbol{x}}_i^\top\widetilde{\boldsymbol{\beta}}^{t,(m)}$ such that

$$\begin{aligned}\frac{e^{\widetilde{\boldsymbol{x}}_i^\top\widetilde{\boldsymbol{\beta}}^*}}{e^{\widetilde{\boldsymbol{x}}_i^\top\widetilde{\boldsymbol{\beta}}^*} + e^{\widetilde{\boldsymbol{x}}_m^\top\widetilde{\boldsymbol{\beta}}^*}} - \frac{e^{\widetilde{\boldsymbol{x}}_i^\top\widetilde{\boldsymbol{\beta}}^{t,(m)}}}{e^{\widetilde{\boldsymbol{x}}_i^\top\widetilde{\boldsymbol{\beta}}^{t,(m)}} + e^{\widetilde{\boldsymbol{x}}_m^\top\widetilde{\boldsymbol{\beta}}^{t,(m)}}} &= -\frac{e^{c_i}}{(1+e^{c_i})^2}\left[\widetilde{\boldsymbol{x}}_m^\top\widetilde{\boldsymbol{\beta}}^* - \widetilde{\boldsymbol{x}}_i^\top\widetilde{\boldsymbol{\beta}}^* - \widetilde{\boldsymbol{x}}_m^\top\widetilde{\boldsymbol{\beta}}^{t,(m)} + \widetilde{\boldsymbol{x}}_i^\top\widetilde{\boldsymbol{\beta}}^{t,(m)}\right] \\ &= -\frac{e^{c_i}}{(1+e^{c_i})^2}\left[\alpha_m^* - \alpha_i^* - \alpha_m^{t,(m)} + \alpha_i^{t,(m)}\right] \\ &\quad -\frac{e^{c_i}}{(1+e^{c_i})^2}\left[\boldsymbol{x}_m^\top\boldsymbol{\beta}^* - \boldsymbol{x}_i^\top\boldsymbol{\beta}^* - \boldsymbol{x}_m^\top\boldsymbol{\beta}^{t,(m)} + \boldsymbol{x}_i^\top\boldsymbol{\beta}^{t,(m)}\right].\end{aligned} \tag{52}$$

Combining Eq. (51) and Eq. (52), we have

$$\begin{aligned}\mu_1 &= \left(1 - \eta p\sum_{i\neq m}\frac{e^{c_i}}{(1+e^{c_i})^2}\right)\left(\alpha_m^{t,(m)} - \alpha_m^*\right) \\ &\quad + \eta p\sum_{i\neq m}\frac{e^{c_i}}{(1+e^{c_i})^2}\left[\alpha_i^{t,(m)} - \alpha_i^* + (\boldsymbol{x}_m - \boldsymbol{x}_i)^\top\left(\boldsymbol{\beta}^* - \boldsymbol{\beta}^{t,(m)}\right)\right] - \eta\lambda\alpha_m^{t,(m)} \\ &= \left(1 - \eta\lambda - \eta p\sum_{i\neq m}\frac{e^{c_i}}{(1+e^{c_i})^2}\right)\left(\alpha_m^{t,(m)} - \alpha_m^*\right) \\ &\quad + \eta p\sum_{i\neq m}\frac{e^{c_i}}{(1+e^{c_i})^2}\left[\alpha_i^{t,(m)} - \alpha_i^* + (\boldsymbol{x}_m - \boldsymbol{x}_i)^\top\left(\boldsymbol{\beta}^* - \boldsymbol{\beta}^{t,(m)}\right)\right] - \eta\lambda\alpha_m^*.\end{aligned}$$

By taking absolute value on both side, we get

$$\begin{aligned}
|\mu_1| \le & \left|1-\eta\lambda-\eta p\sum_{i\neq m}\frac{e^{c_i}}{(1+e^{c_i})^2}\right||\alpha_m^{t,(m)}-\alpha_m^*| \\
& +\frac{\eta p}{4}\sum_{i\neq m}\left[|\alpha_i^{t,(m)}-\alpha_i^*|+\|\boldsymbol{x}_m-\boldsymbol{x}_i\|_2\|\boldsymbol{\beta}^*-\boldsymbol{\beta}^{t,(m)}\|_2\right]+\eta\lambda|\alpha_m^*| \\
\le & \left|1-\eta\lambda-\eta p\sum_{i\neq m}\frac{e^{c_i}}{(1+e^{c_i})^2}\right||\alpha_m^{t,(m)}-\alpha_m^*| \\
& +\frac{\eta p}{4}\left[\sqrt{n}\|\boldsymbol{\alpha}^{t,(m)}-\boldsymbol{\alpha}^*\|_2+n\cdot 2\sqrt{\frac{c_3(d+1)}{n}}\|\boldsymbol{\beta}^*-\boldsymbol{\beta}^{t,(m)}\|_2\right]+\eta\lambda\|\boldsymbol{\alpha}^*\|_\infty \\
\le & \left|1-\eta\lambda-\eta p\sum_{i\neq m}\frac{e^{c_i}}{(1+e^{c_i})^2}\right||\alpha_m^{t,(m)}-\alpha_m^*| \\
& +\frac{\eta p}{4}\sqrt{n}\left(1+2\sqrt{c_3(d+1)}\right)\left\|\widetilde{\boldsymbol{\beta}}^*-\widetilde{\boldsymbol{\beta}}^{t,(m)}\right\|_2+\eta\lambda\|\boldsymbol{\alpha}^*\|_\infty.
\end{aligned}$$

Since $1-\eta\lambda-\eta p\sum_{i\neq m}\frac{e^{c_1}}{(1+e^{c_i})^2}\ge 1-\eta\lambda-\eta p\frac{n}{4}\ge 0$, we have

$$\begin{aligned}
\left|1-\eta\lambda-\eta p\sum_{i\neq m}\frac{e^{c_i}}{(1+e^{c_i})^2}\right| &= 1-\eta\lambda-\eta p\sum_{i\neq m}\frac{e^{c_i}}{(1+e^{c_i})^2} \\
&\le 1-\eta p(n-1)\min_{i\neq m}\frac{e^{c_i}}{(1+e^{c_i})^2}.
\end{aligned}$$

By the defintion of c_i, we have

$$\begin{aligned}
\max_{i\neq m}|c_i| &\le \max_{i\neq m}|\widetilde{\boldsymbol{x}}_m^\top\widetilde{\boldsymbol{\beta}}^*-\widetilde{\boldsymbol{x}}_i^\top\widetilde{\boldsymbol{\beta}}^*|+\max_{i\neq m}\left|\widetilde{\boldsymbol{x}}_m^\top\widetilde{\boldsymbol{\beta}}^*-\widetilde{\boldsymbol{x}}_i^\top\widetilde{\boldsymbol{\beta}}^*-\widetilde{\boldsymbol{x}}_m^\top\widetilde{\boldsymbol{\beta}}^{t,(m)}+\widetilde{\boldsymbol{x}}_i^\top\widetilde{\boldsymbol{\beta}}^{t,(m)}\right| \\
&\le \log\kappa_1+\max_{i\neq m}|\alpha_m^*-\alpha_i^*-\alpha_m^{t,(m)}+\alpha_i^{t,(m)}|+\max_{i\neq m}\left|(\boldsymbol{x}_m-\boldsymbol{x}_i)^\top\left(\boldsymbol{\beta}^*-\boldsymbol{\beta}^{t,(m)}\right)\right| \\
&\le \log\kappa_1+2\|\boldsymbol{\alpha}^{t,(m)}-\boldsymbol{\alpha}^*\|_\infty+2\sqrt{\frac{c_3(d+1)}{n}}\|\boldsymbol{\beta}^*-\boldsymbol{\beta}^{t,(m)}\|_2.
\end{aligned}$$

Consider npL which is large enough such that

$$2(C_4+C_6)\kappa_1^2\sqrt{\frac{(d+1)\log n}{npL}}\le 0.1,\quad 2(C_3+C_4)\kappa_1\sqrt{c_3}\sqrt{\frac{(d+1)\log n}{npL}}\le 0.1.$$

Then we have

$$\begin{aligned}\max_{i\neq m}|c_i| \leq& \log\kappa_1 + 2\|\boldsymbol{\alpha}^{t,(m)} - \boldsymbol{\alpha}^*\|_\infty + 2\sqrt{\frac{c_3(d+1)}{n}}\|\boldsymbol{\beta}^* - \boldsymbol{\beta}^{t,(m)}\|_2\\ \leq& \log\kappa_1 + 2\left(\|\boldsymbol{\alpha}^t - \boldsymbol{\alpha}^*\|_\infty + \max_{1\leq m\leq n}\left\|\widetilde{\boldsymbol{\beta}}^t - \widetilde{\boldsymbol{\beta}}^{t,(m)}\right\|_2\right)\\ &+ 2\sqrt{\frac{c_3(d+1)}{n}}\left(\left\|\widetilde{\boldsymbol{\beta}}^t - \widetilde{\boldsymbol{\beta}}^*\right\|_2 + \max_{1\leq m\leq n}\left\|\widetilde{\boldsymbol{\beta}}^t - \widetilde{\boldsymbol{\beta}}^{t,(m)}\right\|_2\right)\\ \leq& \log\kappa_1 + 0.2.\end{aligned}$$

Then it holds that,

$$\min_{i\neq m}\frac{e^{c_i}}{(1+e^{c_i})^2} = \min_{i\neq m}\frac{e^{-|c_i|}}{(1+e^{-|c_i|})^2} \geq \min_{i\neq m}\frac{e^{-|c_i|}}{4} = \frac{e^{-\max_{i\neq m}|c_i|}}{4} \geq \frac{1}{4\kappa_1 e^{0.2}} \geq \frac{1}{5\kappa_1}.$$

Using $n-1 \geq \frac{n}{2}$ for $n \geq 2$, we have

$$\begin{aligned}|\mu_1| \leq& \left(1 - \frac{1}{10\kappa_1}\eta p n\right)|\alpha_m^{t,(m)} - \alpha_m^*|\\ &+ \frac{1+2\sqrt{c_3(d+1)}}{4}\eta p\sqrt{n}\left(\left\|\widetilde{\boldsymbol{\beta}}^t - \widetilde{\boldsymbol{\beta}}^*\right\|_2 + \max_{1\leq m\leq n}\left\|\widetilde{\boldsymbol{\beta}}^t - \widetilde{\boldsymbol{\beta}}^{t,(m)}\right\|_2\right) + \eta\lambda\kappa_2\\ \leq& \left(1 - \frac{1}{10\kappa_1}\eta p n\right)C_5\kappa_1^2\sqrt{\frac{(d+1)\log n}{npL}}\\ &+ \frac{1+2\sqrt{c_3(d+1)}}{4}\eta p\sqrt{n}(C_3+C_4)\kappa_1\sqrt{\frac{\log n}{pL}} + \eta\lambda\kappa_2.\end{aligned}$$

Combine this result with Eq. (49), we get

$$\begin{aligned}\left|\alpha_m^{t+1,(m)} - \alpha_m^*\right| \leq& |\mu_1| + |\mu_2|\\ \leq& \widetilde{C}_2\eta\kappa_1\sqrt{\frac{(d+1)np\log n}{L}} + \left(1 - \frac{1}{10\kappa_1}\eta p n\right)C_5\kappa_1^2\sqrt{\frac{(d+1)\log n}{npL}}\\ &+ \frac{1+2\sqrt{c_3(d+1)}}{4}\eta p\sqrt{n}(C_3+C_4)\kappa_1\sqrt{\frac{\log n}{pL}} + \eta\lambda\kappa_2\\ \leq& C_5\kappa_1^2\sqrt{\frac{(d+1)\log n}{npL}}.\end{aligned}$$

as long as $C_5 \geq 30\widetilde{C}_2$, $C_5 \geq 7.5(1+2\sqrt{c_3})(C_3+C_4)$ and $C_5 \geq 30c_\lambda/\sqrt{d+1}$. This concludes our proof for Lemma 23. ■

C.11 Proof of Lemma 24

Proof For any $m \in [n]$, we have

$$\begin{aligned}
|\alpha_m^{t+1} - \alpha_m^*| &\le |\alpha_m^{t+1} - \alpha_m^{t+1,(m)}| + |\alpha_m^{t+1,(m)} - \alpha_m^*| \\
&\le \left\|\widetilde{\boldsymbol{\beta}}_m^{t+1} - \widetilde{\boldsymbol{\beta}}_m^{t+1,(m)}\right\|_2 + |\alpha_m^{t+1,(m)} - \alpha_m^*| \\
&\le C_4\kappa_1\sqrt{\frac{(d+1)\log n}{npL}} + C_5\kappa_1^2\sqrt{\frac{(d+1)\log n}{npL}} \\
&\le (C_4 + C_5)\kappa_1^2\sqrt{\frac{(d+1)\log n}{npL}}.
\end{aligned}$$

As a result, we have

$$\left\|\boldsymbol{\alpha}^{t+1} - \boldsymbol{\alpha}^*\right\|_\infty \le C_6\kappa_1^2\sqrt{\frac{(d+1)\log n}{npL}},$$

as long as $C_6 \ge C_4 + C_5$. ■

Appendix D. Proof of Auxiliary Lemmas in Section B

D.1 Proof of Lemma 26 and Two Propositions (Propositions 35 and 36)

Proof [Proof of Lemma 26] (1) By definition for $i \in [n]$ we have

$$\begin{aligned}\left(\nabla \mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\right)_i &= \sum_{j\neq i,(i,j)\in\mathcal{E}} \left\{-y_{j,i} + \phi(\widetilde{\boldsymbol{x}}_i^\top \widetilde{\boldsymbol{\beta}}^* - \widetilde{\boldsymbol{x}}_j^\top \widetilde{\boldsymbol{\beta}}^*)\right\} \\ &= \frac{1}{L}\sum_{j\neq i,(i,j)\in\mathcal{E}} \sum_{l=1}^{L} \left\{-y_{j,i}^{(l)} + \phi(\widetilde{\boldsymbol{x}}_i^\top \widetilde{\boldsymbol{\beta}}^* - \widetilde{\boldsymbol{x}}_j^\top \widetilde{\boldsymbol{\beta}}^*)\right\}.\end{aligned}$$

Since $\left|-y_{j,i}^{(l)} + \phi(\widetilde{\boldsymbol{x}}_i^\top \widetilde{\boldsymbol{\beta}}^* - \widetilde{\boldsymbol{x}}_j^\top \widetilde{\boldsymbol{\beta}}^*)\right| \le 1$, by Bernstein inequality we have

$$\begin{aligned}\left|\left(\nabla \mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\right)_i - \mathbb{E}\left[\left(\nabla \mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\right)_i \middle| \mathcal{G}\right]\right| &\lesssim \frac{1}{L}\left(\sqrt{\log n \left(\sum_{j\neq i,(i,j)\in\mathcal{E}} 1\right) L} + \log n\right) \\ &\lesssim \sqrt{\frac{np\log n}{L}}\end{aligned}$$

with probability exceeding $1 - O(n^{-10})$, as long as $npL \gtrsim \log n$. On the other hand, since $\mathbb{E}\left[-y_{j,i}^{(l)} + \phi(\widetilde{\boldsymbol{x}}_i^\top \widetilde{\boldsymbol{\beta}}^* - \widetilde{\boldsymbol{x}}_j^\top \widetilde{\boldsymbol{\beta}}^*)\right] = 0$, we know that $\mathbb{E}\left[\left(\nabla \mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\right)_i \middle| \mathcal{G}\right] = 0$. As a result, we have

$$\left|\left(\nabla \mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\right)_i\right| \lesssim \sqrt{\frac{np\log n}{L}}.$$

(2) By definition we have

$$\begin{aligned}\sum_{j\neq i}\left(\nabla^2 \mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\right)_{i,j}^2 &= \left\|\left(\sum_{j\neq i,(i,j)\in\mathcal{E}} \phi'(\widetilde{\boldsymbol{x}}_i^\top \widetilde{\boldsymbol{\beta}}^* - \widetilde{\boldsymbol{x}}_j^\top \widetilde{\boldsymbol{\beta}}^*)\,(\widetilde{\boldsymbol{x}}_i - \widetilde{\boldsymbol{x}}_j)\right)_{-i}\right\|_2^2 \\ &= \sum_{j\neq i,(i,j)\in\mathcal{E}} 1 + \left\|\sum_{j\neq i,(i,j)\in\mathcal{E}} \phi'(\widetilde{\boldsymbol{x}}_i^\top \widetilde{\boldsymbol{\beta}}^* - \widetilde{\boldsymbol{x}}_j^\top \widetilde{\boldsymbol{\beta}}^*)\,(\boldsymbol{x}_i - \boldsymbol{x}_j)\right\|_2^2 \\ &\le \sum_{j\neq i,(i,j)\in\mathcal{E}} 1 + \left(\sum_{j\neq i,(i,j)\in\mathcal{E}} 1\right) \sum_{j\neq i,(i,j)\in\mathcal{E}} \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2 \\ &\lesssim np + np\cdot dp\end{aligned}$$

with probability at least $1 - O(n^{-10})$. Similarly, we have

$$\begin{aligned}\sum_{k>n}\left(\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\right)_{i,k}^2 &= \left\|\sum_{j\neq i,(i,j)\in\mathcal{E}}\phi'(\widetilde{\boldsymbol{x}}_i^\top\widetilde{\boldsymbol{\beta}}^* - \widetilde{\boldsymbol{x}}_j^\top\widetilde{\boldsymbol{\beta}}^*)\,(\boldsymbol{x}_i-\boldsymbol{x}_j)\right\|_2^2\\ &\le \left(\sum_{j\neq i,(i,j)\in\mathcal{E}}1\right)\sum_{j\neq i,(i,j)\in\mathcal{E}}\|\boldsymbol{x}_i-\boldsymbol{x}_j\|_2^2\\ &\lesssim np\cdot dp\end{aligned}$$

and

$$\sum_{j\in[n],j\neq i}\left|\left(\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\right)_{i,j}\right| = \sum_{j\in[n],j\neq i}\left|\phi'(\widetilde{\boldsymbol{x}}_i^\top\widetilde{\boldsymbol{\beta}}^* - \widetilde{\boldsymbol{x}}_j^\top\widetilde{\boldsymbol{\beta}}^*)\right|\mathbf{1}((i,j)\in\mathcal{E}) \le \sum_{j\in[n],j\neq i}\mathbf{1}((i,j)\in\mathcal{E}) \lesssim np$$

with probability at least $1 - O(n^{-10})$.

(3) For $i,j\in[n], i\neq j$, by definition we know that $y_{j,i} = \dfrac{1}{L}\sum_{l=1}^{L} y_{j,i}^{(l)}$ is the average of L independent Bernoulli random variables. By Hoeffding's inequality we know that

$$|y_{j,i} - \mathbb{E}y_{j,i}| \lesssim \sqrt{\frac{\log n}{L}}$$

with probability at least $1 - O(n^{-12})$. As a result, by union bound we know that

$$|y_{j,i} - \mathbb{E}y_{j,i}| \lesssim \sqrt{\frac{\log n}{L}}$$

holds for all $i,j\in[n], i\neq j$ with probability at least $1 - O(n^{-10})$. ■

We also include here two propositions which are also involved in the later proofs.

Proposition 35. *$\widetilde{\boldsymbol{\beta}}$ is the solution of the following linear equations*

$$\begin{cases}\mathcal{P}\nabla\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*) + \mathcal{P}\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\left(\overline{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}^*\right) = \mathbf{0};\\ \mathcal{P}\overline{\boldsymbol{\beta}} = \overline{\boldsymbol{\beta}}.\end{cases}$$

Proposition 35 follows from Eq. (5), which gives the definition of $\overline{\boldsymbol{\beta}}$.

Proposition 36. *Under event $\mathcal{A}_2$, we have*

$$Var\left[\,\overline{\boldsymbol{\beta}}\mid\mathcal{G}\right] = \frac{1}{L}\left[\mathcal{P}\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\mathcal{P}\right]^+.$$

We next provide the proof of Proposition 36 here.

Proof [Proof of Proposition 36] Since $\mathcal{P}\nabla\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*) + \mathcal{P}\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)(\overline{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}^*) = \mathbf{0}$, by taking variance (conditioned on $\mathcal{G}$) on the both sides we have

$$\mathcal{P}\mathrm{Var}\left[\nabla\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*) \mid \mathcal{G}\right]\mathcal{P} + \mathcal{P}\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\mathrm{Var}\left[\overline{\boldsymbol{\beta}} \mid \mathcal{G}\right]\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\mathcal{P} + 2\mathcal{P}\mathrm{Cov}\left(\nabla\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*), \overline{\boldsymbol{\beta}} \mid \mathcal{G}\right)\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\mathcal{P} = \mathbf{0}. \tag{53}$$

On the other hand, by considering the covariance (conditioned on $\mathcal{G}$) of $\nabla\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)$ and $\mathcal{P}\nabla\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*) + \mathcal{P}\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)(\overline{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}^*)$ we get

$$\mathrm{Var}\left[\nabla\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*) \mid \mathcal{G}\right]\mathcal{P} + \mathrm{Cov}\left(\nabla\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*), \overline{\boldsymbol{\beta}} \mid \mathcal{G}\right)\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\mathcal{P} = \mathbf{0}.$$

As a result, we know that

$$\mathcal{P}\mathrm{Cov}\left(\nabla\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*), \overline{\boldsymbol{\beta}} \mid \mathcal{G}\right)\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\mathcal{P} = -\mathcal{P}\mathrm{Var}\left[\nabla\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*) \mid \mathcal{G}\right]\mathcal{P}. \tag{54}$$

Combine Eq. (53) and Eq. (54) we get

$$\mathcal{P}\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\mathrm{Var}\left[\overline{\boldsymbol{\beta}} \mid \mathcal{G}\right]\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\mathcal{P} = \mathcal{P}\mathrm{Var}\left[\nabla\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*) \mid \mathcal{G}\right]\mathcal{P} = \frac{1}{L}\mathcal{P}\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\mathcal{P}. \tag{55}$$

By taking variance on the both sides of $\mathcal{P}\overline{\boldsymbol{\beta}} = \overline{\boldsymbol{\beta}}$, we have $\mathcal{P}\mathrm{Var}\left[\overline{\boldsymbol{\beta}} \mid \mathcal{G}\right]\mathcal{P} = \mathrm{Var}\left[\overline{\boldsymbol{\beta}} \mid \mathcal{G}\right]$. This also implies $(\boldsymbol{I} - \mathcal{P})\mathrm{Var}\left[\overline{\boldsymbol{\beta}} \mid \mathcal{G}\right] = 0$. As a result, Eq. (55) can be also written as

$$\mathcal{P}\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\mathcal{P}\mathrm{Var}\left[\overline{\boldsymbol{\beta}} \mid \mathcal{G}\right]\mathcal{P}\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\mathcal{P} = \frac{1}{L}\mathcal{P}\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\mathcal{P}.$$

This immediately implies $\mathcal{P}\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\mathcal{P}(\boldsymbol{I} - \mathrm{Var}\left[\overline{\boldsymbol{\beta}} \mid \mathcal{G}\right]\mathcal{P}\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\mathcal{P}) = \mathbf{0}$. Under event $\mathcal{A}_2$, we have $\lambda_{\min,\perp}(\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)) > 0$. As a result, for any $\widetilde{\boldsymbol{\beta}} \in \Theta$, we have

$$\begin{aligned}
&\lambda_{\min,\perp}(\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*))\left\|\left(\boldsymbol{I} - \mathrm{Var}\left[\overline{\boldsymbol{\beta}} \mid \mathcal{G}\right]\mathcal{P}\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\mathcal{P}\right)\widetilde{\boldsymbol{\beta}}\right\|_2^2 \\
\leq &\widetilde{\boldsymbol{\beta}}^\top\left(\boldsymbol{I} - \mathrm{Var}\left[\overline{\boldsymbol{\beta}} \mid \mathcal{G}\right]\mathcal{P}\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)^\top\mathcal{P}\right)^\top\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\left(\boldsymbol{I} - \mathrm{Var}\left[\overline{\boldsymbol{\beta}} \mid \mathcal{G}\right]\mathcal{P}\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\mathcal{P}\right)\widetilde{\boldsymbol{\beta}} \\
= &\widetilde{\boldsymbol{\beta}}^\top\left(\boldsymbol{I} - \mathrm{Var}\left[\overline{\boldsymbol{\beta}} \mid \mathcal{G}\right]\mathcal{P}\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)^\top\mathcal{P}\right)^\top\mathcal{P}\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\mathcal{P}\left(\boldsymbol{I} - \mathrm{Var}\left[\overline{\boldsymbol{\beta}} \mid \mathcal{G}\right]\mathcal{P}\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\mathcal{P}\right)\widetilde{\boldsymbol{\beta}} = 0.
\end{aligned}$$

As a result, we have $(\boldsymbol{I} - \mathrm{Var}\left[\overline{\boldsymbol{\beta}} \mid \mathcal{G}\right]\mathcal{P}\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\mathcal{P})\widetilde{\boldsymbol{\beta}} = \mathbf{0}$ for all $\widetilde{\boldsymbol{\beta}} \in \Theta$. Combine this fact with $(\boldsymbol{I} - \mathcal{P})\mathrm{Var}\left[\overline{\boldsymbol{\beta}} \mid \mathcal{G}\right] = \mathbf{0}$, we know that

$$\mathrm{Var}\left[\overline{\boldsymbol{\beta}} \mid \mathcal{G}\right] = \left[\mathcal{P}\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\mathcal{P}\right]^+.$$

■

D.2 Proof of Theorem 27

Proof We know that

$$\begin{aligned}\mathbf{0} = \mathcal{P}\nabla\mathcal{L}(\widetilde{\boldsymbol{\beta}}_M) &= \mathcal{P}\nabla\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*) + \mathcal{P}\int_0^1 \nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^* + t(\widetilde{\boldsymbol{\beta}}_M - \widetilde{\boldsymbol{\beta}}^*))\left(\widetilde{\boldsymbol{\beta}}_M - \widetilde{\boldsymbol{\beta}}^*\right) dt \\ &= \mathcal{P}\nabla\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*) + \mathcal{P}\left\{\int_0^1 \nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^* + t(\widetilde{\boldsymbol{\beta}}_M - \widetilde{\boldsymbol{\beta}}^*))dt\right\}\left(\widetilde{\boldsymbol{\beta}}_M - \widetilde{\boldsymbol{\beta}}^*\right).\end{aligned}$$

Let

$$\boldsymbol{R} = \left\{\int_0^1 \nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^* + t(\widetilde{\boldsymbol{\beta}}_M - \widetilde{\boldsymbol{\beta}}^*))dt\right\}\left(\widetilde{\boldsymbol{\beta}}_M - \widetilde{\boldsymbol{\beta}}^*\right) - \nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\left(\widetilde{\boldsymbol{\beta}}_M - \widetilde{\boldsymbol{\beta}}^*\right),$$

we have

$$\mathbf{0} = \mathcal{P}\nabla\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*) + \mathcal{P}\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\left(\widetilde{\boldsymbol{\beta}}_M - \widetilde{\boldsymbol{\beta}}^*\right) + \mathcal{P}\boldsymbol{R}. \tag{56}$$

On the other hand, we know that

$$\mathbf{0} = \mathcal{P}\nabla\overline{\mathcal{L}}(\overline{\boldsymbol{\beta}}) = \mathcal{P}\nabla\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*) + \mathcal{P}\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\left(\overline{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}^*\right). \tag{57}$$

Combine Eq. (56) and Eq. (57) we have

$$\mathcal{P}\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\left(\overline{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}_M\right) = \mathcal{P}\boldsymbol{R}. \tag{58}$$

For $\boldsymbol{R}$ we have

$$\begin{aligned}\|\boldsymbol{R}\|_2 &\le \left\|\int_0^1 \nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^* + t(\widetilde{\boldsymbol{\beta}}_M - \widetilde{\boldsymbol{\beta}}^*)) - \nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)dt\right\| \left\|\widetilde{\boldsymbol{\beta}}_M - \widetilde{\boldsymbol{\beta}}^*\right\|_2 \\ &\le \int_0^1 \left\|\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^* + t(\widetilde{\boldsymbol{\beta}}_M - \widetilde{\boldsymbol{\beta}}^*)) - \nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\right\| dt \left\|\widetilde{\boldsymbol{\beta}}_M - \widetilde{\boldsymbol{\beta}}^*\right\|_2.\end{aligned}$$

By definition we have

$$\begin{aligned}&\left\|\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^* + t(\widetilde{\boldsymbol{\beta}}_M - \widetilde{\boldsymbol{\beta}}^*)) - \nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\right\| \\ =& \left\|\sum_{(i,j)\in\mathcal{E}, i>j}\left(\phi'(t(\widetilde{\boldsymbol{x}}_i^\top\widetilde{\boldsymbol{\beta}}_M - \widetilde{\boldsymbol{x}}_j^\top\widetilde{\boldsymbol{\beta}}_M) + (1-t)(\widetilde{\boldsymbol{x}}_i^\top\widetilde{\boldsymbol{\beta}}^* - \widetilde{\boldsymbol{x}}_j^\top\widetilde{\boldsymbol{\beta}}^*)) - \phi'(\widetilde{\boldsymbol{x}}_i^\top\widetilde{\boldsymbol{\beta}}^* - \widetilde{\boldsymbol{x}}_j^\top\widetilde{\boldsymbol{\beta}}^*)\right)(\widetilde{\boldsymbol{x}}_i - \widetilde{\boldsymbol{x}}_j)(\widetilde{\boldsymbol{x}}_i - \widetilde{\boldsymbol{x}}_j)^\top\right\| \\ \le& \left\|\sum_{(i,j)\in\mathcal{E}, i>j}\left|\phi'(t(\widetilde{\boldsymbol{x}}_i^\top\widetilde{\boldsymbol{\beta}}_M - \widetilde{\boldsymbol{x}}_j^\top\widetilde{\boldsymbol{\beta}}_M) + (1-t)(\widetilde{\boldsymbol{x}}_i^\top\widetilde{\boldsymbol{\beta}}^* - \widetilde{\boldsymbol{x}}_j^\top\widetilde{\boldsymbol{\beta}}^*)) - \phi'(\widetilde{\boldsymbol{x}}_i^\top\widetilde{\boldsymbol{\beta}}^* - \widetilde{\boldsymbol{x}}_j^\top\widetilde{\boldsymbol{\beta}}^*)\right|(\widetilde{\boldsymbol{x}}_i - \widetilde{\boldsymbol{x}}_j)(\widetilde{\boldsymbol{x}}_i - \widetilde{\boldsymbol{x}}_j)^\top\right\| \\ \lesssim& \left\|\sum_{(i,j)\in\mathcal{E}, i>j}\left|t(\widetilde{\boldsymbol{x}}_i^\top\widetilde{\boldsymbol{\beta}}_M - \widetilde{\boldsymbol{x}}_j^\top\widetilde{\boldsymbol{\beta}}_M) - t(\widetilde{\boldsymbol{x}}_i^\top\widetilde{\boldsymbol{\beta}}^* - \widetilde{\boldsymbol{x}}_j^\top\widetilde{\boldsymbol{\beta}}^*)\right|(\widetilde{\boldsymbol{x}}_i - \widetilde{\boldsymbol{x}}_j)(\widetilde{\boldsymbol{x}}_i - \widetilde{\boldsymbol{x}}_j)^\top\right\| \\ \lesssim& \left\|\sum_{(i,j)\in\mathcal{E}, i>j}(\widetilde{\boldsymbol{x}}_i - \widetilde{\boldsymbol{x}}_j)(\widetilde{\boldsymbol{x}}_i - \widetilde{\boldsymbol{x}}_j)^\top\right\| \left\|\widetilde{\boldsymbol{\beta}}_M - \widetilde{\boldsymbol{\beta}}^*\right\|_c \lesssim \|\boldsymbol{L}_{\mathcal{G}}\| \left\|\widetilde{\boldsymbol{\beta}}_M - \widetilde{\boldsymbol{\beta}}^*\right\|_c, \quad \forall t \in [0,1].\end{aligned}$$

By Lemma 15 we know that

$$\left\|\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*+t(\widetilde{\boldsymbol{\beta}}_M-\widetilde{\boldsymbol{\beta}}^*))-\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\right\|\lesssim np\left\|\widetilde{\boldsymbol{\beta}}_M-\widetilde{\boldsymbol{\beta}}^*\right\|_c$$

holds uniformly for all $t\in[0,1]$ with probability at least $1-O(n^{-11})$. As a result, we get

$$\|\boldsymbol{R}\|_2\lesssim\int_0^1 np\left\|\widetilde{\boldsymbol{\beta}}_M-\widetilde{\boldsymbol{\beta}}^*\right\|_c dt\left\|\widetilde{\boldsymbol{\beta}}_M-\widetilde{\boldsymbol{\beta}}^*\right\|_2\lesssim np\left\|\widetilde{\boldsymbol{\beta}}_M-\widetilde{\boldsymbol{\beta}}^*\right\|_c\left\|\widetilde{\boldsymbol{\beta}}_M-\widetilde{\boldsymbol{\beta}}^*\right\|_2 \tag{59}$$

with probability exceeding $1-O(n^{-11})$. Since $\overline{\boldsymbol{\beta}},\widetilde{\boldsymbol{\beta}}_M\in\Theta$, we know that

$$\begin{aligned}
&\lambda_{\min,\perp}\left(\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\right)\left\|\overline{\boldsymbol{\beta}}-\widetilde{\boldsymbol{\beta}}_M\right\|_2^2\leq\left(\overline{\boldsymbol{\beta}}-\widetilde{\boldsymbol{\beta}}_M\right)^\top\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\left(\overline{\boldsymbol{\beta}}-\widetilde{\boldsymbol{\beta}}_M\right)\\
=&\left(\overline{\boldsymbol{\beta}}-\widetilde{\boldsymbol{\beta}}_M\right)^\top\mathcal{P}\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\left(\overline{\boldsymbol{\beta}}-\widetilde{\boldsymbol{\beta}}_M\right)=\left(\overline{\boldsymbol{\beta}}-\widetilde{\boldsymbol{\beta}}_M\right)^\top\mathcal{P}\boldsymbol{R}\\
\leq&\left\|\overline{\boldsymbol{\beta}}-\widetilde{\boldsymbol{\beta}}_M\right\|_2\|\mathcal{P}\boldsymbol{R}\|_2\leq\left\|\overline{\boldsymbol{\beta}}-\widetilde{\boldsymbol{\beta}}_M\right\|_2\|\boldsymbol{R}\|_2.
\end{aligned}$$

As a result, we get

$$\left\|\overline{\boldsymbol{\beta}}-\widetilde{\boldsymbol{\beta}}_M\right\|_2\leq\frac{\|\boldsymbol{R}\|_2}{\lambda_{\min,\perp}\left(\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\right)}. \tag{60}$$

By Lemma 17 we know that

$$\lambda_{\min,\perp}\left(\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\right)\gtrsim\frac{pn}{\kappa_1}. \tag{61}$$

with probability exceeding $1-O(n^{-11})$. Therefore, combine Eq. (59), Eq. (60) and Eq. (61) we get

$$\left\|\overline{\boldsymbol{\beta}}-\widetilde{\boldsymbol{\beta}}_M\right\|_2\leq\frac{\|\boldsymbol{R}\|_2}{\lambda_{\min,\perp}\left(\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\right)}\lesssim\kappa_1^4\frac{(d+1)^{0.5}\log n}{\sqrt{n}pL}$$

with probability exceeding $1-O(n^{-6})$. ■

D.3 Proof of Proposition 28 and Proposition 30

We denote by $\Psi=\{\widetilde{\boldsymbol{x}}_i-\widetilde{\boldsymbol{x}}_j:i,j\in[n]\}$. By the definition of $\mathcal{L}(\cdot)$ and $\overline{\mathcal{L}}(\cdot)$ we know that for any $\widetilde{\boldsymbol{\beta}}\in\mathbb{R}^{n+d}$ and $\boldsymbol{v}\in\Psi^\perp$, we have

$$\mathcal{L}(\widetilde{\boldsymbol{\beta}})=\mathcal{L}(\widetilde{\boldsymbol{\beta}}+\boldsymbol{v})\text{ and }\overline{\mathcal{L}}(\widetilde{\boldsymbol{\beta}})=\overline{\mathcal{L}}(\widetilde{\boldsymbol{\beta}}+\boldsymbol{v}).$$

On the other hand, under Assumption 3 we have the following lemma.

Lemma 37. *For any $\widetilde{\boldsymbol{\beta}}\in\mathbb{R}^{n+d}$, there exists a $\boldsymbol{v}\in\Psi^\perp$ such that $\widetilde{\boldsymbol{\beta}}+\boldsymbol{v}\in\Theta$.*

Proof We only have to show that for any $\widetilde{\boldsymbol{\beta}} \in \mathbb{R}^{n+d}$, $\widetilde{\boldsymbol{\beta}} + \Psi^{\perp} \bigcap \Theta \neq \varnothing$. Assume there exists a $\widetilde{\boldsymbol{\beta}} \in \mathbb{R}^{n+d}$ such that $\widetilde{\boldsymbol{\beta}} + \Psi^{\perp} \bigcap \Theta = \varnothing$. First of all, we must have $\boldsymbol{Z}^{\top}\widetilde{\boldsymbol{\beta}} \notin \boldsymbol{Z}^{\top}\Psi^{\perp}$. Since $\boldsymbol{Z}^{\top}\Psi^{\perp} \subset \mathbb{R}^{d+1}$, we must have $\dim(\boldsymbol{Z}^{\top}\Psi^{\perp}) \leq d$ as $\boldsymbol{Z}^{\top}\Psi^{\perp} \neq \mathbb{R}^{d+1}$. Since $\dim(\Psi^{\perp}) = d+1$, we know that there exists a non-zero vector $\boldsymbol{v} \in \Psi^{\perp}$ such that $\boldsymbol{Z}^{\top}\boldsymbol{v} = \mathbf{0}$. By the definition of Θ, we know that $\boldsymbol{v} \in \Theta$. Recall the definition of $\boldsymbol{\Sigma}$ and Assumption 3 in §3. Since $\boldsymbol{v} \in \Psi^{\perp}$, we know that $\boldsymbol{\Sigma}\boldsymbol{v} = 0$, so we must have $\lambda_{\min,\perp}(\boldsymbol{\Sigma}) = 0$. As a result, this contradicts to Assumption 3 since $c_2 = 0$. ■

With this lemma, we then turn to prove Proposition 28 and 30.

Proof [Proof of Proposition 28 and 30] Assume there exists a z such that $\overline{\mathcal{L}}|_{\overline{\boldsymbol{\beta}}_{-i}}(z) < \overline{\mathcal{L}}|_{\overline{\boldsymbol{\beta}}_{-i}}(\overline{\alpha}_i)$. Then we let $\boldsymbol{w} \in \mathbb{R}^{n+d}$ be the vector such that $\boldsymbol{w}_{-i} = \overline{\boldsymbol{\beta}}_{-i}$ and $w_i = z$. And, let $\boldsymbol{v}$ be the vector in $\Psi^{\perp}$ such that $\boldsymbol{w} + \boldsymbol{v} \in \Theta$. Then we have

$$\overline{\mathcal{L}}(\boldsymbol{w}+\boldsymbol{v}) = \overline{\mathcal{L}}(\boldsymbol{w}) = \overline{\mathcal{L}}|_{\overline{\boldsymbol{\beta}}_{-i}}(z) < \overline{\mathcal{L}}|_{\overline{\boldsymbol{\beta}}_{-i}}(\overline{\alpha}_i) = \overline{\mathcal{L}}(\overline{\boldsymbol{\beta}}).$$

This contradicts to the definition of $\overline{\boldsymbol{\beta}}$ Eq. (5).

Similarly, if we assume that there exists a z such that $\mathcal{L}|_{\widetilde{\boldsymbol{\beta}}_{M,-i}}(z) < \mathcal{L}|_{\widetilde{\boldsymbol{\beta}}_{M,-i}}(x)(\widehat{\alpha}_{M,i})$. Then we let $\boldsymbol{w} \in \mathbb{R}^{n+d}$ be the vector such that $\boldsymbol{w}_{-i} = \widetilde{\boldsymbol{\beta}}_{M,-i}$ and $w_i = z$. And, let $\boldsymbol{v}$ be the vector in $\Psi^{\perp}$ such that $\boldsymbol{w} + \boldsymbol{v} \in \Theta$. Then we have

$$\mathcal{L}(\boldsymbol{w}+\boldsymbol{v}) = \mathcal{L}(\boldsymbol{w}) = \mathcal{L}|_{\widetilde{\boldsymbol{\beta}}_{M,-i}}(z) < \mathcal{L}|_{\widetilde{\boldsymbol{\beta}}_{M,-i}}(\widehat{\alpha}_{M,i}) = \mathcal{L}(\widetilde{\boldsymbol{\beta}}_M).$$

This contradicts to the definition of $\widetilde{\boldsymbol{\beta}}_M$ Eq. (3). ■

D.4 Auxiliary Results for Proving Lemma 29

In this section we include two results which are helpful to the proof of Lemma 29 in §D.5. These two results are analogies of Theorem 4 and Theorem 27 which we have proven before. The main difference is we replace $\mathcal{L}(\cdot)$ with $\mathcal{L}^{(i)}(\cdot)$. As a result, the following two results can be viewed as the leave-one-out version of Theorem 4 and Theorem 27. To be more specific, we define $\widetilde{\boldsymbol{\beta}}_M^{(i)}$ as

$$\widetilde{\boldsymbol{\beta}}_M^{(i)} = \operatorname*{argmin}_{\widetilde{\boldsymbol{\beta}} \in \Theta} \mathcal{L}^{(i)}(\widetilde{\boldsymbol{\beta}})$$

and let $\overline{\boldsymbol{\beta}}^{(i)}$ be the solution of the following equations

$$\begin{cases} \mathcal{P}\nabla\mathcal{L}^{(i)}(\widetilde{\boldsymbol{\beta}}^*) + \mathcal{P}\nabla^2\mathcal{L}^{(i)}(\widetilde{\boldsymbol{\beta}}^*)\left(\overline{\boldsymbol{\beta}}^{(i)} - \widetilde{\boldsymbol{\beta}}^*\right) = \mathbf{0}; \\ \mathcal{P}\overline{\boldsymbol{\beta}}^{(i)} = \overline{\boldsymbol{\beta}}^{(i)}. \end{cases}$$

Lemma 38. *Suppose $np > c_p \log n$ for some $c_p > 0$ and $d+1 < n, (d+1)\log n \lesssim np$. We consider $L \leq c_4 \cdot n^{c_5}$ for any absolute constants $c_4, c_5 > 0$. Then for every $i \in [n]$ and*

$\widetilde{\boldsymbol{\beta}}_M^{(i)} = (\widehat{\boldsymbol{\alpha}}_M^{(i)\top}, \widehat{\boldsymbol{\beta}}_M^{(i)\top})^\top$, *with probability at least* $1 - O(n^{-6})$ *we have*

$$\|\widehat{\boldsymbol{\alpha}}_M^{(i)} - \boldsymbol{\alpha}^*\|_\infty \lesssim \kappa_1^2 \sqrt{\frac{(d+1)\log n}{npL}}, \left\|\widehat{\boldsymbol{\beta}}_M^{(i)} - \boldsymbol{\beta}^*\right\|_2 \lesssim \kappa_1 \sqrt{\frac{\log n}{pL}} \text{ and } \left\|\widetilde{\boldsymbol{\beta}}_M^{(i)} - \widetilde{\boldsymbol{\beta}}_M\right\|_2 \lesssim \kappa_1 \sqrt{\frac{(d+1)\log n}{npL}}.$$

Lemma 39. *Under the assumptions of Theorem 38, for every* $i \in [n]$*, with probability at least* $1 - O(n^{-6})$ *we have*

$$\left\|\widetilde{\boldsymbol{\beta}}_M^{(i)} - \overline{\boldsymbol{\beta}}^{(i)}\right\|_2 \lesssim \kappa_1^4 \frac{(d+1)^{0.5}\log n}{\sqrt{n}pL}.$$

The proof of Lemma 38 and Lemma 39 are almost the same as the previous results so we omit the proof details here. One can show Lemma 38 by mimicing the proof in §A.5, §C.7 and the results in Lemma 22. In addition, Lemma 39 can be proved by mimicing the proof in §D.2.

D.5 Proof of Lemma 29

Proof Since $\overline{\alpha}_i$ can be expressed as

$$\overline{\alpha}_i = \alpha_i^* - \frac{\left(\nabla \mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\right)_i + \sum_{j \neq i}\left(\overline{\boldsymbol{\beta}}_j - \widetilde{\boldsymbol{\beta}}_j^*\right)\left(\nabla^2 \mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\right)_{i,j}}{\left(\nabla^2 \mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\right)_{i,i}},$$

we have

$$\overline{\alpha}_i' - \overline{\alpha}_i = \frac{\sum_{j \neq i}\left(\widetilde{\boldsymbol{\beta}}_{M,j} - \overline{\boldsymbol{\beta}}_j\right)\left(\nabla^2 \mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\right)_{i,j}}{\left(\nabla^2 \mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\right)_{i,i}}. \tag{62}$$

We decompose Eq. (62) as

$$\overline{\alpha}_i' - \overline{\alpha}_i = \underbrace{\frac{\sum_{j \neq i}\left(\widetilde{\boldsymbol{\beta}}_{M,j} - \widetilde{\boldsymbol{\beta}}_{M,j}^{(i)}\right)\left(\nabla^2 \mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\right)_{i,j}}{\left(\nabla^2 \mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\right)_{i,i}}}_{A_1} + \underbrace{\frac{\sum_{j \neq i}\left(\widetilde{\boldsymbol{\beta}}_{M,j}^{(i)} - \overline{\boldsymbol{\beta}}_j^{(i)}\right)\left(\nabla^2 \mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\right)_{i,j}}{\left(\nabla^2 \mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\right)_{i,i}}}_{A_2}$$
$$+ \underbrace{\frac{\sum_{j \neq i}\left(\overline{\boldsymbol{\beta}}_j^{(i)} - \overline{\boldsymbol{\beta}}_j\right)\left(\nabla^2 \mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\right)_{i,j}}{\left(\nabla^2 \mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\right)_{i,i}}}_{A_3}.$$

Next, we bound A_1-A_3 one by one. Before proceeding, the denominator $\left(\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\right)_{i,i}$ can be bounded as

$$\begin{aligned}\left(\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\right)_{i,i} &= \sum_{j\neq i}\phi'(\widetilde{\boldsymbol{x}}_i^\top\widetilde{\boldsymbol{\beta}}^* - \widetilde{\boldsymbol{x}}_j^\top\widetilde{\boldsymbol{\beta}}^*)\mathbf{1}((i,j)\in\mathcal{E}) \\ &\gtrsim \frac{1}{\kappa_1}\sum_{j\neq i}\mathbf{1}((i,j)\in\mathcal{E}) \gtrsim \frac{np}{\kappa_1}\end{aligned}$$

with probability at least $1-O(n^{-11})$.

For A_1, by Lemma 38 we have

$$\left\|\widetilde{\boldsymbol{\beta}}_M - \widetilde{\boldsymbol{\beta}}_M^{(i)}\right\|_2 \lesssim \kappa_1\sqrt{\frac{(d+1)\log n}{npL}}.$$

So the numerator of A_1 can be bounded as

$$\begin{aligned}\left|\sum_{j\neq i}\left(\widetilde{\boldsymbol{\beta}}_{M,j} - \widetilde{\boldsymbol{\beta}}_{M,j}^{(i)}\right)\left(\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\right)_{i,j}\right| &\leq \left\|\widetilde{\boldsymbol{\beta}}_M - \widetilde{\boldsymbol{\beta}}_M^{(i)}\right\|_2\sqrt{\sum_{j\neq i}\left(\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\right)_{i,j}^2} \\ &\lesssim \kappa_1\sqrt{\frac{(d+1)\log n}{npL}}\sqrt{(d+1)np} \lesssim \kappa_1(d+1)\sqrt{\frac{\log n}{L}}\end{aligned}$$

with probability at least $1-O(n^{-10})$. As a result, A_1 can be bounded as

$$|A_1| \lesssim \frac{\kappa_1(d+1)\sqrt{\frac{\log n}{L}}}{np/\kappa_1} \leq \kappa_1^2\frac{d+1}{np}\sqrt{\frac{\log n}{L}} \tag{63}$$

with probability at least $1-O(n^{-10})$.

When it comes to A_2, by definition we know that $\widetilde{\boldsymbol{\beta}}_{M,j}^{(i)} - \overline{\boldsymbol{\beta}}_j^{(i)}$ is independent with $\left(\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\right)_{i,j}$ for all $j\in[n+d]$. As a result, by Bernstein's inequality and Lemma 39 we know that

$$\left|\sum_{j\in[n],j\neq i}\left(\widetilde{\boldsymbol{\beta}}_{M,j}^{(i)} - \overline{\boldsymbol{\beta}}_j^{(i)}\right)\left(\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\right)_{i,j} - \mathbb{E}\left[\sum_{j\in[n],j\neq i}\left(\widetilde{\boldsymbol{\beta}}_{M,j}^{(i)} - \overline{\boldsymbol{\beta}}_j^{(i)}\right)\left(\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\right)_{i,j}\middle|\widetilde{\boldsymbol{\beta}}_M^{(i)},\overline{\boldsymbol{\beta}}^{(i)}\right]\right| \tag{64}$$

$$\lesssim\sqrt{(\log n)\sum_{j\in[n],j\neq i}\left(\widetilde{\boldsymbol{\beta}}_{M,j}^{(i)} - \overline{\boldsymbol{\beta}}_j^{(i)}\right)^2\mathbb{E}\left(\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\right)_{i,j}^2} + (\log n)\left\|\widehat{\boldsymbol{\alpha}}_M^{(i)} - \overline{\boldsymbol{\alpha}}^{(i)}\right\|_\infty \tag{65}$$

$$\lesssim\sqrt{p\log n}\left\|\widehat{\boldsymbol{\alpha}}_M^{(i)} - \overline{\boldsymbol{\alpha}}^{(i)}\right\|_2 + (\log n)\left\|\widehat{\boldsymbol{\alpha}}_M^{(i)} - \overline{\boldsymbol{\alpha}}^{(i)}\right\|_\infty \leq \sqrt{p\log n}\left\|\widetilde{\boldsymbol{\beta}}_M^{(i)} - \overline{\boldsymbol{\beta}}^{(i)}\right\|_2 + (\log n)\left\|\widehat{\boldsymbol{\alpha}}_M^{(i)} - \overline{\boldsymbol{\alpha}}^{(i)}\right\|_\infty \tag{66}$$

$$\lesssim\kappa_1^4\frac{(d+1)^{0.5}(\log n)^{1.5}}{\sqrt{np}L} + (\log n)\left\|\widehat{\boldsymbol{\alpha}}_M^{(i)} - \overline{\boldsymbol{\alpha}}^{(i)}\right\|_\infty \tag{67}$$

with probability at least $1 - O(n^{-6})$. On the other hand, by Lemma 39 we have

$$\left| \mathbb{E}\left[\sum_{j\in[n],j\neq i} \left(\widetilde{\boldsymbol{\beta}}_{M,j}^{(i)} - \overline{\boldsymbol{\beta}}_j^{(i)}\right) \left(\nabla^2 \mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\right)_{i,j} \middle| \widetilde{\boldsymbol{\beta}}_M^{(i)}, \overline{\boldsymbol{\beta}}^{(i)} \right] \right| \tag{68}$$

$$\leq \sqrt{\sum_{j\in[n],j\neq i} \left(\widetilde{\boldsymbol{\beta}}_{M,j}^{(i)} - \overline{\boldsymbol{\beta}}_j^{(i)}\right)^2} \sqrt{\sum_{j\in[n],j\neq i} \left(\mathbb{E}\left(\nabla^2 \mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\right)_{i,j}\right)^2} \tag{69}$$

$$\lesssim p\sqrt{n} \left\| \widetilde{\boldsymbol{\beta}}_M^{(i)} - \overline{\boldsymbol{\beta}}^{(i)} \right\|_2 \lesssim \kappa_1^4 \frac{(d+1)^{0.5} \log n}{L} \tag{70}$$

and

$$\left| \sum_{k>i} \left(\widetilde{\boldsymbol{\beta}}_{M,k}^{(i)} - \overline{\boldsymbol{\beta}}_k^{(i)}\right) \left(\nabla^2 \mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\right)_{i,k} \right| \leq \left\| \widetilde{\boldsymbol{\beta}}_M^{(i)} - \overline{\boldsymbol{\beta}}^{(i)} \right\|_2 \sqrt{\sum_{k>i} \left(\nabla^2 \mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\right)_{i,k}^2} \lesssim \kappa_1^4 \frac{(d+1)\log n}{L} \tag{71}$$

with probability exceeding $1 - O(n^{-6})$. Combine Eq. (67), Eq. (70) and Eq. (71) we finally bound the numerator of A_2 as

$$\left| \sum_{j\neq i} \left(\widetilde{\boldsymbol{\beta}}_{M,j}^{(i)} - \overline{\boldsymbol{\beta}}_j^{(i)}\right) \left(\nabla^2 \mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\right)_{i,j} \right| \lesssim \kappa_1^4 \frac{(d+1)\log n}{L} + (\log n) \left\| \widehat{\boldsymbol{\alpha}}_M^{(i)} - \overline{\boldsymbol{\alpha}}^{(i)} \right\|_\infty$$

with probability at least $1 - O(n^{-6})$. As a result, we have

$$|A_2| \lesssim \kappa_1^5 \frac{(d+1)\log n}{npL} + \kappa_1 \frac{\log n}{np} \left\| \widehat{\boldsymbol{\alpha}}_M^{(i)} - \overline{\boldsymbol{\alpha}}^{(i)} \right\|_\infty \tag{72}$$

with probability at least $1 - O(n^{-6})$.

We finally consider bounding A_3. By definition, we know that

$$\begin{cases} \mathcal{P}\nabla \mathcal{L}(\widetilde{\boldsymbol{\beta}}^*) + \mathcal{P}\nabla^2 \mathcal{L}(\widetilde{\boldsymbol{\beta}}^*) \left(\overline{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}^*\right) = \mathbf{0}; \\ \mathcal{P}\nabla \mathcal{L}^{(i)}(\widetilde{\boldsymbol{\beta}}^*) + \mathcal{P}\nabla^2 \mathcal{L}^{(i)}(\widetilde{\boldsymbol{\beta}}^*) \left(\overline{\boldsymbol{\beta}}^{(i)} - \widetilde{\boldsymbol{\beta}}^*\right) = \mathbf{0}. \end{cases}$$

Combine the two equations we get

$$\underbrace{\mathcal{P}\left(\nabla \mathcal{L}(\widetilde{\boldsymbol{\beta}}^*) - \nabla \mathcal{L}^{(i)}(\widetilde{\boldsymbol{\beta}}^*)\right)}_{\boldsymbol{w}_1} + \underbrace{\mathcal{P}\left(\nabla^2 \mathcal{L}(\widetilde{\boldsymbol{\beta}}^*) - \nabla^2 \mathcal{L}^{(i)}(\widetilde{\boldsymbol{\beta}}^*)\right)\left(\overline{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}^*\right)}_{\boldsymbol{w}_2} + \mathcal{P}\nabla^2 \mathcal{L}^{(i)}(\widetilde{\boldsymbol{\beta}}^*)\left(\overline{\boldsymbol{\beta}} - \overline{\boldsymbol{\beta}}^{(i)}\right) = \mathbf{0}. \tag{73}$$

For $\boldsymbol{w}_1$, it is easy to see $\|\mathcal{P}(\nabla \mathcal{L}(\widetilde{\boldsymbol{\beta}}^*) - \nabla \mathcal{L}^{(i)}(\widetilde{\boldsymbol{\beta}}^*))\|_2 \leq \|\nabla \mathcal{L}(\widetilde{\boldsymbol{\beta}}^*) - \nabla \mathcal{L}^{(i)}(\widetilde{\boldsymbol{\beta}}^*)\|_2$. By definition we have

$$\nabla \mathcal{L}(\widetilde{\boldsymbol{\beta}}^*) - \nabla \mathcal{L}^{(i)}(\widetilde{\boldsymbol{\beta}}^*) = \sum_{j\neq i} (-y_{j,i} + \phi(\widetilde{\boldsymbol{x}}_i^\top \widetilde{\boldsymbol{\beta}}^* - \widetilde{\boldsymbol{x}}_j^\top \widetilde{\boldsymbol{\beta}}^*)) \mathbf{1}((i,j) \in \mathcal{E})(\widetilde{\boldsymbol{x}}_i - \widetilde{\boldsymbol{x}}_j).$$

For $(\nabla\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*) - \nabla\mathcal{L}^{(i)}(\widetilde{\boldsymbol{\beta}}^*))_i$, by Hoeffding inequality, we have

$$\begin{aligned}\left|(\nabla\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*) - \nabla\mathcal{L}^{(i)}(\widetilde{\boldsymbol{\beta}}^*))_i\right| &= \frac{1}{L}\left|\sum_{(i,j)\in\mathcal{E}}\sum_{l=1}^{L}(-y_{j,i}^{(l)} + \phi(\widetilde{\boldsymbol{x}}_i^\top\widetilde{\boldsymbol{\beta}}^* - \widetilde{\boldsymbol{x}}_j^\top\widetilde{\boldsymbol{\beta}}^*))\mathbf{1}((i,j)\in\mathcal{E})\right| \\ \lesssim &\frac{1}{L}\sqrt{L\sum_{j\neq i}\mathbf{1}((i,j)\in\mathcal{E})\log n}\end{aligned}$$

conditioned on $\mathcal{G}$ with probability at least $1-O(n^{-10})$. On the other hand, since $\sum_{j\neq i,j\in[n]}\mathbf{1}((i,j)\in\mathcal{E}) \lesssim np$ with probability at least $1-O(n^{-10})$, we have

$$\left|(\nabla\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*) - \nabla\mathcal{L}^{(i)}(\widetilde{\boldsymbol{\beta}}^*))_i\right| \lesssim \sqrt{\frac{np\log n}{L}}$$

with probability at least $1-O(n^{-10})$. Furthermore, by Lemma 26 we have

$$\begin{aligned}\sum_{j\neq i,j\in[n]}\left(\nabla\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*) - \nabla\mathcal{L}^{(i)}(\widetilde{\boldsymbol{\beta}}^*)\right)_j^2 &= \sum_{j\neq i,j\in[n]}\left((-y_{j,i} + \phi(\widetilde{\boldsymbol{x}}_i^\top\widetilde{\boldsymbol{\beta}}^* - \widetilde{\boldsymbol{x}}_j^\top\widetilde{\boldsymbol{\beta}}^*))\mathbf{1}((i,j)\in\mathcal{E})\right)^2 \\ &\lesssim \frac{\log n}{L}\sum_{j\neq i}\mathbf{1}((i,j)\in\mathcal{E}) \lesssim \frac{np\log n}{L}\end{aligned}$$

with probability at least $1-O(n^{-10})$. For $(\nabla\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*) - \nabla\mathcal{L}^{(i)}(\widetilde{\boldsymbol{\beta}}^*))_{n+1:n+d}$, we have

$$\begin{aligned}\left\|\left(\nabla\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*) - \nabla\mathcal{L}^{(i)}(\widetilde{\boldsymbol{\beta}}^*)\right)_{n+1:n+d}\right\|_2 &= \left\|\sum_{j\neq i}(-y_{j,i} + \phi(\widetilde{\boldsymbol{x}}_i^\top\widetilde{\boldsymbol{\beta}}^* - \widetilde{\boldsymbol{x}}_j^\top\widetilde{\boldsymbol{\beta}}^*))\mathbf{1}((i,j)\in\mathcal{E})(\boldsymbol{x}_i - \boldsymbol{x}_j)\right\|_2 \\ &\leq \sum_{j\neq i}\mathbf{1}((i,j)\in\mathcal{E})\left|-y_{j,i} + \phi(\widetilde{\boldsymbol{x}}_i^\top\widetilde{\boldsymbol{\beta}}^* - \widetilde{\boldsymbol{x}}_j^\top\widetilde{\boldsymbol{\beta}}^*)\right|\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2 \\ &\lesssim np\sqrt{\frac{\log n}{L}}\sqrt{\frac{d+1}{n}} \leq \sqrt{\frac{(d+1)np\log n}{L}}\end{aligned}$$

with probability at least $1-O(n^{-10})$. Therefore, for $\boldsymbol{w}_1$ we have

$$\|\boldsymbol{w}_1\|_2^2 \leq \left\|\nabla\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*) - \nabla\mathcal{L}^{(i)}(\widetilde{\boldsymbol{\beta}}^*)\right\|_2^2 \leq \left(\nabla\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*) - \nabla\mathcal{L}^{(i)}(\widetilde{\boldsymbol{\beta}}^*)\right)_i^2 + \sum_{j\neq i,j\in[n]}\left(\nabla\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*) - \nabla\mathcal{L}^{(i)}(\widetilde{\boldsymbol{\beta}}^*)\right)_j^2 \tag{74}$$

$$+ \left\|\left(\nabla\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*) - \nabla\mathcal{L}^{(i)}(\widetilde{\boldsymbol{\beta}}^*)\right)_{n+1:n+d}\right\|_2^2 \lesssim \frac{(d+1)np\log n}{L} \tag{75}$$

with probability exceeding $1-O(n^{-10})$.

For $\boldsymbol{w}_2$, since $\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*) - \nabla^2\mathcal{L}^{(i)}(\widetilde{\boldsymbol{\beta}}^*) = \sum_{j\neq i}(\mathbf{1}((i,j)\in\mathcal{E}) - p)\phi'(\widetilde{\boldsymbol{x}}_i^\top\widetilde{\boldsymbol{\beta}}^* - \widetilde{\boldsymbol{x}}_j^\top\widetilde{\boldsymbol{\beta}}^*)(\widetilde{\boldsymbol{x}}_i - \widetilde{\boldsymbol{x}}_j)(\widetilde{\boldsymbol{x}}_i - \widetilde{\boldsymbol{x}}_j)^\top$, it holds that

$$\begin{aligned}&\left(\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*) - \nabla^2\mathcal{L}^{(i)}(\widetilde{\boldsymbol{\beta}}^*)\right)\left(\overline{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}^*\right) \\ =&\sum_{j\neq i}(\mathbf{1}((i,j)\in\mathcal{E}) - p)\phi'(\widetilde{\boldsymbol{x}}_i^\top\widetilde{\boldsymbol{\beta}}^* - \widetilde{\boldsymbol{x}}_j^\top\widetilde{\boldsymbol{\beta}}^*)\left((\widetilde{\boldsymbol{x}}_i - \widetilde{\boldsymbol{x}}_j)^\top\left(\overline{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}^*\right)\right)(\widetilde{\boldsymbol{x}}_i - \widetilde{\boldsymbol{x}}_j).\end{aligned}$$

As a result, it follows that

$$\begin{aligned}
&\left\|\mathcal{P}\left(\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)-\nabla^2\mathcal{L}^{(i)}(\widetilde{\boldsymbol{\beta}}^*)\right)\left(\overline{\boldsymbol{\beta}}-\widetilde{\boldsymbol{\beta}}^*\right)\right\|_2 \le \left\|\left(\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)-\nabla^2\mathcal{L}^{(i)}(\widetilde{\boldsymbol{\beta}}^*)\right)\left(\overline{\boldsymbol{\beta}}-\widetilde{\boldsymbol{\beta}}^*\right)\right\|_2 \\
\le&\sum_{j\neq i}\left\|(p-\mathbf{1}((i,j)\in\mathcal{E}))\phi'(\widetilde{\boldsymbol{x}}_i^\top\widetilde{\boldsymbol{\beta}}^*-\widetilde{\boldsymbol{x}}_j^\top\widetilde{\boldsymbol{\beta}}^*)\left((\widetilde{\boldsymbol{x}}_i-\widetilde{\boldsymbol{x}}_j)^\top\left(\overline{\boldsymbol{\beta}}-\widetilde{\boldsymbol{\beta}}^*\right)\right)(\widetilde{\boldsymbol{x}}_i-\widetilde{\boldsymbol{x}}_j)\right\|_2 \\
\le&\sum_{j\neq i}|p-\mathbf{1}((i,j)\in\mathcal{E})|\left|(\widetilde{\boldsymbol{x}}_i-\widetilde{\boldsymbol{x}}_j)^\top\left(\overline{\boldsymbol{\beta}}-\widetilde{\boldsymbol{\beta}}^*\right)\right|\|\widetilde{\boldsymbol{x}}_i-\widetilde{\boldsymbol{x}}_j\|_2 \\
\lesssim&\sum_{j\neq i}|p-\mathbf{1}((i,j)\in\mathcal{E})|\left\|\overline{\boldsymbol{\beta}}-\widetilde{\boldsymbol{\beta}}^*\right\|_c \lesssim np\left\|\overline{\boldsymbol{\beta}}-\widetilde{\boldsymbol{\beta}}^*\right\|_c
\end{aligned}$$

with probability exceeding $1-O(n^{-10})$. On the other hand, we obtain

$$\begin{aligned}
\left\|\widetilde{\boldsymbol{\beta}}^*-\overline{\boldsymbol{\beta}}\right\|_c &\le \left\|\widetilde{\boldsymbol{\beta}}^*-\widetilde{\boldsymbol{\beta}}_M\right\|_c+\left\|\widetilde{\boldsymbol{\beta}}_M-\overline{\boldsymbol{\beta}}\right\|_c \\
&\lesssim \kappa_1^2\sqrt{\frac{(d+1)\log n}{npL}}+\|\widehat{\boldsymbol{\alpha}}_M-\overline{\boldsymbol{\alpha}}\|_\infty+\sqrt{\frac{d+1}{n}}\kappa_1^4\frac{(d+1)^{0.5}\log n}{\sqrt{n}pL} \\
&\lesssim \kappa_1^2\sqrt{\frac{(d+1)\log n}{npL}}+\|\widehat{\boldsymbol{\alpha}}_M-\overline{\boldsymbol{\alpha}}\|_\infty
\end{aligned}$$

with probability at least $1-O(n^{-6})$. That is to say, we have

$$\|\boldsymbol{w}_2\|_2 \lesssim \kappa_1^2\sqrt{\frac{(d+1)np\log n}{L}}+np\,\|\widehat{\boldsymbol{\alpha}}_M-\overline{\boldsymbol{\alpha}}\|_\infty \tag{76}$$

with probability at least $1-O(n^{-6})$.

Combine Eq. (75) and Eq. (76) with Eq. (73), we know that

$$\left\|\mathcal{P}\nabla^2\mathcal{L}^{(i)}(\widetilde{\boldsymbol{\beta}}^*)\left(\overline{\boldsymbol{\beta}}-\overline{\boldsymbol{\beta}}^{(i)}\right)\right\|_2=\|\boldsymbol{w}_1+\boldsymbol{w}_2\|_2\le\|\boldsymbol{w}_1\|_2+\|\boldsymbol{w}_2\|_2\lesssim\kappa_1^2\sqrt{\frac{(d+1)np\log n}{L}}+np\,\|\widehat{\boldsymbol{\alpha}}_M-\overline{\boldsymbol{\alpha}}\|_\infty$$

with probability at least $1-O(n^{-6})$. Since $\overline{\boldsymbol{\beta}}-\overline{\boldsymbol{\beta}}^{(i)}\in\Theta$, we have

$$\begin{aligned}
&\lambda_{\min,\perp}\left(\nabla^2\mathcal{L}^{(i)}(\widetilde{\boldsymbol{\beta}}^*)\right)\left\|\overline{\boldsymbol{\beta}}-\overline{\boldsymbol{\beta}}^{(i)}\right\|_2^2\le\left(\overline{\boldsymbol{\beta}}-\overline{\boldsymbol{\beta}}^{(i)}\right)^\top\nabla^2\mathcal{L}^{(i)}(\widetilde{\boldsymbol{\beta}}^*)\left(\overline{\boldsymbol{\beta}}-\overline{\boldsymbol{\beta}}^{(i)}\right) \\
=&\left(\overline{\boldsymbol{\beta}}-\overline{\boldsymbol{\beta}}^{(i)}\right)^\top\mathcal{P}\nabla^2\mathcal{L}^{(i)}(\widetilde{\boldsymbol{\beta}}^*)\left(\overline{\boldsymbol{\beta}}-\overline{\boldsymbol{\beta}}^{(i)}\right)\le\left\|\overline{\boldsymbol{\beta}}-\overline{\boldsymbol{\beta}}^{(i)}\right\|_2\left\|\mathcal{P}\nabla^2\mathcal{L}^{(i)}(\widetilde{\boldsymbol{\beta}}^*)\left(\overline{\boldsymbol{\beta}}-\overline{\boldsymbol{\beta}}^{(i)}\right)\right\|_2.
\end{aligned}$$

As a result, we know that

$$\left\|\overline{\boldsymbol{\beta}}-\overline{\boldsymbol{\beta}}^{(i)}\right\|_2\le\frac{\left\|\mathcal{P}\nabla^2\mathcal{L}^{(i)}(\widetilde{\boldsymbol{\beta}}^*)\left(\overline{\boldsymbol{\beta}}-\overline{\boldsymbol{\beta}}^{(i)}\right)\right\|_2}{\lambda_{\min,\perp}\left(\nabla^2\mathcal{L}^{(i)}(\widetilde{\boldsymbol{\beta}}^*)\right)}\lesssim\kappa_1^3\sqrt{\frac{(d+1)\log n}{npL}}+\kappa_1\,\|\widehat{\boldsymbol{\alpha}}_M-\overline{\boldsymbol{\alpha}}\|_\infty \tag{77}$$

with probability at least $1-O(n^{-6})$. Therefore, for A_3 we finally achieve

$$|A_3| \lesssim \left| \frac{\sum_{j\neq i}\left(\overline{\boldsymbol{\beta}}_j^{(i)} - \overline{\boldsymbol{\beta}}_j\right)\left(\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\right)_{i,j}}{np/\kappa_1} \right| \leq \frac{\left\|\overline{\boldsymbol{\beta}}^{(i)} - \overline{\boldsymbol{\beta}}\right\|_2 \sqrt{\sum_{j\neq i}\left(\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\right)_{i,j}^2}}{np/\kappa_1} \tag{78}$$

$$\lesssim \kappa_1^4 \frac{d+1}{np}\sqrt{\frac{\log n}{L}} + \kappa_1^2\sqrt{\frac{d+1}{np}}\left\|\widehat{\boldsymbol{\alpha}}_M - \overline{\boldsymbol{\alpha}}\right\|_\infty \tag{79}$$

with probability at least $1-O(n^{-6})$.

And, by Eq. (77) we know that

$$\begin{aligned}\left\|\widehat{\boldsymbol{\alpha}}_M^{(i)} - \overline{\boldsymbol{\alpha}}^{(i)}\right\|_\infty &\leq \left\|\widetilde{\boldsymbol{\beta}}_M^{(i)} - \widetilde{\boldsymbol{\beta}}_M\right\|_2 + \left\|\widehat{\boldsymbol{\alpha}}_M - \overline{\boldsymbol{\alpha}}\right\|_\infty + \left\|\overline{\boldsymbol{\beta}} - \overline{\boldsymbol{\beta}}^{(i)}\right\|_2 \\ &\lesssim \kappa_1^3\sqrt{\frac{(d+1)\log n}{npL}} + \kappa_1\left\|\widehat{\boldsymbol{\alpha}}_M - \overline{\boldsymbol{\alpha}}\right\|_\infty\end{aligned}$$

with probability exceeding $1-O(n^{-6})$. Combine with Eq. (72) we have

$$|A_2| \lesssim \kappa_1^5\frac{(d+1)\log n}{npL} + \kappa_1^4\frac{\log n}{np}\sqrt{\frac{(d+1)\log n}{npL}} + \kappa_1^2\frac{\log n}{np}\left\|\widehat{\boldsymbol{\alpha}}_M - \overline{\boldsymbol{\alpha}}\right\|_\infty \tag{80}$$

with probability exceeding $1-O(n^{-6})$. Combine Eq. (63), Eq. (80) and Eq. (79) we know that

$$|\overline{\alpha}_i' - \overline{\alpha}_i| \lesssim \kappa_1^5\frac{(d+1)\log n}{npL} + \frac{\kappa_1^4}{np}\sqrt{\frac{(d+1)\log n}{L}}\left(\sqrt{d+1} + \frac{\log n}{\sqrt{np}}\right) + \kappa_1^2\left(\sqrt{\frac{d+1}{np}} + \frac{\log n}{np}\right)\left\|\widehat{\boldsymbol{\alpha}}_M - \overline{\boldsymbol{\alpha}}\right\|_\infty$$

with probability at least $1-O(n^{-6})$.

■

D.6 Proof of Lemma 31

Proof Since $\widehat{\alpha}_{M,i}$ is the minimizer of $\mathcal{L}\Big|_{\widetilde{\boldsymbol{\beta}}_{M,-i}}(\cdot)$, we know that $\left(\mathcal{L}\Big|_{\widetilde{\boldsymbol{\beta}}_{M,-i}}\right)'(\widehat{\alpha}_{M,i}) = 0$. By the mean value theorem we know that

$$\left(\mathcal{L}\Big|_{\widetilde{\boldsymbol{\beta}}_{M,-i}}\right)'(\widehat{\alpha}_{M,i}) = \left(\mathcal{L}\Big|_{\widetilde{\boldsymbol{\beta}}_{M,-i}}\right)'(\alpha_i^*) + \left(\mathcal{L}\Big|_{\widetilde{\boldsymbol{\beta}}_{M,-i}}\right)''(b_1)(\widehat{\alpha}_{M,i} - \alpha_i^*),$$

where b_1 is some real number between α_i^* and $\widehat{\alpha}_{M,i}$. As a result, we have

$$\widehat{\alpha}_{M,i} = \alpha_i^* - \frac{\left(\mathcal{L}\Big|_{\widetilde{\boldsymbol{\beta}}_{M,-i}}\right)'(\alpha_i^*)}{\left(\mathcal{L}\Big|_{\widetilde{\boldsymbol{\beta}}_{M,-i}}\right)''(b_1)}.$$

By the definition Eq. (4) and Eq. (24), we have

$$\left(\mathcal{L}\Big|_{\widetilde{\boldsymbol{\beta}}_{M,-i}}\right)'(x) = \sum_{j\neq i,(i,j)\in\mathcal{E}} \left\{-y_{j,i} + \phi(\widetilde{\boldsymbol{x}}_i^\top \widetilde{\boldsymbol{\beta}}_M - \widetilde{\boldsymbol{x}}_j^\top \widetilde{\boldsymbol{\beta}}_M + x - \widehat{\alpha}_{M,i})\right\}$$
$$\left(\mathcal{L}\Big|_{\widetilde{\boldsymbol{\beta}}_{M,-i}}\right)''(x) = \sum_{j\neq i,(i,j)\in\mathcal{E}} \phi'(\widetilde{\boldsymbol{x}}_i^\top \widetilde{\boldsymbol{\beta}}_M - \widetilde{\boldsymbol{x}}_j^\top \widetilde{\boldsymbol{\beta}}_M + x - \widehat{\alpha}_{M,i}).$$

We first estimate the difference $\left(\mathcal{L}\Big|_{\widetilde{\boldsymbol{\beta}}_{M,-i}}\right)''(b_1) - \left(\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\right)_{i,i}$. We have

$$\begin{aligned}
&\left|\left(\mathcal{L}\Big|_{\widetilde{\boldsymbol{\beta}}_{M,-i}}\right)''(b_1) - \left(\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\right)_{i,i}\right| \\
=&\left|\sum_{j\neq i,(i,j)\in\mathcal{E}} \phi'(\widetilde{\boldsymbol{x}}_i^\top \widetilde{\boldsymbol{\beta}}_M - \widetilde{\boldsymbol{x}}_j^\top \widetilde{\boldsymbol{\beta}}_M + b_1 - \widehat{\alpha}_{M,i}) - \phi'(\widetilde{\boldsymbol{x}}_i^\top \widetilde{\boldsymbol{\beta}}^* - \widetilde{\boldsymbol{x}}_j^\top \widetilde{\boldsymbol{\beta}}^*)\right| \\
\leq& \sum_{j\neq i,(i,j)\in\mathcal{E}} \left|(\widetilde{\boldsymbol{x}}_i^\top \widetilde{\boldsymbol{\beta}}_M - \widetilde{\boldsymbol{x}}_j^\top \widetilde{\boldsymbol{\beta}}_M + b_1 - \widehat{\alpha}_{M,i}) - (\widetilde{\boldsymbol{x}}_i^\top \widetilde{\boldsymbol{\beta}}^* - \widetilde{\boldsymbol{x}}_j^\top \widetilde{\boldsymbol{\beta}}^*)\right| \\
\lesssim& \sum_{j\neq i,(i,j)\in\mathcal{E}} \left(\left\|\widetilde{\boldsymbol{\beta}}_M - \widetilde{\boldsymbol{\beta}}^*\right\|_c + |b_1 - \widehat{\alpha}_{M,i}|\right) \\
\lesssim& np\left\|\widetilde{\boldsymbol{\beta}}_M - \widetilde{\boldsymbol{\beta}}^*\right\|_c
\end{aligned}$$

with probability at least $1 - O(n^{-11})$. On the other hand, we have

$$\left(\mathcal{L}\Big|_{\widetilde{\boldsymbol{\beta}}_{M,-i}}\right)'(\widetilde{\boldsymbol{\beta}}_i^*) - \left(\left(\nabla\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\right)_i + \sum_{j\neq i}\left(\widetilde{\boldsymbol{\beta}}_{M,j} - \widetilde{\boldsymbol{\beta}}_j^*\right)\left(\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\right)_{i,j}\right) \tag{81}$$
$$= \sum_{j\neq i,(i,j)\in\mathcal{E}} \left\{-y_{j,i} + \phi(\widetilde{\boldsymbol{x}}_i^\top \widetilde{\boldsymbol{\beta}}_M - \widetilde{\boldsymbol{x}}_j^\top \widetilde{\boldsymbol{\beta}}_M + \alpha_i^* - \widehat{\alpha}_{M,i})\right\} - \sum_{j\neq i,(i,j)\in\mathcal{E}} \left\{-y_{j,i} + \phi(\widetilde{\boldsymbol{x}}_i^\top \widetilde{\boldsymbol{\beta}}^* - \widetilde{\boldsymbol{x}}_j^\top \widetilde{\boldsymbol{\beta}}^*)\right\} \tag{82}$$
$$- \sum_{j\neq i, j\in[n+d]} \left(\widetilde{\boldsymbol{\beta}}_{M,j} - \widetilde{\boldsymbol{\beta}}_j^*\right)\left(\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\right)_{i,j} \tag{83}$$
$$= \sum_{j\neq i,(i,j)\in\mathcal{E}} \left\{\phi(\widetilde{\boldsymbol{x}}_i^\top \widetilde{\boldsymbol{\beta}}_M - \widetilde{\boldsymbol{x}}_j^\top \widetilde{\boldsymbol{\beta}}_M + \alpha_i^* - \widehat{\alpha}_{M,i}) - \phi(\widetilde{\boldsymbol{x}}_i^\top \widetilde{\boldsymbol{\beta}}^* - \widetilde{\boldsymbol{x}}_j^\top \widetilde{\boldsymbol{\beta}}^*) - \left(\alpha_j^* - \widehat{\alpha}_{M,j}\right)\phi'(\widetilde{\boldsymbol{x}}_i^\top \widetilde{\boldsymbol{\beta}}^* - \widetilde{\boldsymbol{x}}_j^\top \widetilde{\boldsymbol{\beta}}^*)\right\} \tag{84}$$
$$- \sum_{k\in[d]} \left(\widetilde{\boldsymbol{\beta}}_{M,n+k} - \widetilde{\boldsymbol{\beta}}_{n+k}^*\right)\left(\sum_{j\neq i,(i,j)\in\mathcal{E}} \phi'(\widetilde{\boldsymbol{x}}_i^\top \widetilde{\boldsymbol{\beta}}^* - \widetilde{\boldsymbol{x}}_j^\top \widetilde{\boldsymbol{\beta}}^*)\left(\boldsymbol{x}_i - \boldsymbol{x}_j\right)_k\right) \tag{85}$$
$$= \sum_{j\neq i,(i,j)\in\mathcal{E}} r_j, \tag{86}$$

where

$$r_j = \phi(\widetilde{\boldsymbol{x}}_i^\top \widetilde{\boldsymbol{\beta}}_M - \widetilde{\boldsymbol{x}}_j^\top \widetilde{\boldsymbol{\beta}}_M + \alpha_i^* - \widehat{\alpha}_{M,i}) - \phi(\widetilde{\boldsymbol{x}}_i^\top \widetilde{\boldsymbol{\beta}}^* - \widetilde{\boldsymbol{x}}_j^\top \widetilde{\boldsymbol{\beta}}^*) - \left(\alpha_j^* - \widehat{\alpha}_{M,j}\right) \phi'(\widetilde{\boldsymbol{x}}_i^\top \widetilde{\boldsymbol{\beta}}^* - \widetilde{\boldsymbol{x}}_j^\top \widetilde{\boldsymbol{\beta}}^*)$$
$$- (\boldsymbol{x}_i - \boldsymbol{x}_j)^\top (\widehat{\boldsymbol{\beta}}_M - \boldsymbol{\beta}^*) \phi'(\widetilde{\boldsymbol{x}}_i^\top \widetilde{\boldsymbol{\beta}}^* - \widetilde{\boldsymbol{x}}_j^\top \widetilde{\boldsymbol{\beta}}^*).$$

By Taylor expansion we know that

$$\begin{aligned}\phi(\widetilde{\boldsymbol{x}}_i^\top \widetilde{\boldsymbol{\beta}}_M - \widetilde{\boldsymbol{x}}_j^\top \widetilde{\boldsymbol{\beta}}_M + \alpha_i^* - \widehat{\alpha}_{M,i}) =& \phi(\widetilde{\boldsymbol{x}}_i^\top \widetilde{\boldsymbol{\beta}}^* - \widetilde{\boldsymbol{x}}_j^\top \widetilde{\boldsymbol{\beta}}^*) \\ &+ \phi'(\widetilde{\boldsymbol{x}}_i^\top \widetilde{\boldsymbol{\beta}}^* - \widetilde{\boldsymbol{x}}_j^\top \widetilde{\boldsymbol{\beta}}^*) \left((\boldsymbol{x}_i - \boldsymbol{x}_j)^\top (\widehat{\boldsymbol{\beta}}_M - \boldsymbol{\beta}^*) + \alpha_j^* - \widehat{\alpha}_{M,j}\right) \\ &+ \phi''(b_2) \left((\boldsymbol{x}_i - \boldsymbol{x}_j)^\top (\widehat{\boldsymbol{\beta}}_M - \boldsymbol{\beta}^*) + \alpha_j^* - \widehat{\alpha}_{M,j}\right)^2,\end{aligned}$$

where b_2 is some real number between $\widetilde{\boldsymbol{x}}_i^\top \widetilde{\boldsymbol{\beta}}^* - \widetilde{\boldsymbol{x}}_j^\top \widetilde{\boldsymbol{\beta}}^*$ and $\widetilde{\boldsymbol{x}}_i^\top \widetilde{\boldsymbol{\beta}}_M - \widetilde{\boldsymbol{x}}_j^\top \widetilde{\boldsymbol{\beta}}_M + \alpha_i^* - \widehat{\alpha}_{M,i}$. As a result, we have

$$|r_j| \le |\phi''(b_2)| \left((\boldsymbol{x}_i - \boldsymbol{x}_j)^\top (\widehat{\boldsymbol{\beta}}_M - \boldsymbol{\beta}^*) + \alpha_j^* - \widehat{\alpha}_{M,j}\right)^2 \lesssim \left\|\widetilde{\boldsymbol{\beta}}_M - \widetilde{\boldsymbol{\beta}}^*\right\|_c^2. \tag{87}$$

Combine Eq. (86) and Eq. (87) we have

$$\left| \left(\mathcal{L}\Big|_{\widetilde{\boldsymbol{\beta}}_{M,-i}} \right)' (\widetilde{\boldsymbol{\beta}}_i^*) - \left(\left(\nabla \mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\right)_i + \sum_{j \neq i} \left(\widetilde{\boldsymbol{\beta}}_{M,j} - \widetilde{\boldsymbol{\beta}}_j^*\right) \left(\nabla^2 \mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\right)_{i,j} \right) \right| \lesssim np \left\|\widetilde{\boldsymbol{\beta}}_M - \widetilde{\boldsymbol{\beta}}^*\right\|_c^2$$

with probability exceeding $1 - O(n^{-11})$. As a result, we have

$$\begin{aligned}\widehat{\alpha}_{M,i} - \overline{\alpha}_i' =& \frac{\left(\nabla \mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\right)_i + \sum_{j \neq i} \left(\widetilde{\boldsymbol{\beta}}_{M,j} - \widetilde{\boldsymbol{\beta}}_j^*\right) \left(\nabla^2 \mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\right)_{i,j}}{\left(\nabla^2 \mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\right)_{i,i}} - \frac{\left(\mathcal{L}\Big|_{\widetilde{\boldsymbol{\beta}}_{M,-i}} \right)' (\alpha_i^*)}{\left(\mathcal{L}\Big|_{\widetilde{\boldsymbol{\beta}}_{M,-i}} \right)'' (b_1)} \\ =& \underbrace{\frac{\left(\mathcal{L}\Big|_{\widetilde{\boldsymbol{\beta}}_{M,-i}} \right)'' (b_1) - \left(\nabla^2 \mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\right)_{i,i}}{\left(\mathcal{L}\Big|_{\widetilde{\boldsymbol{\beta}}_{M,-i}} \right)'' (b_1) \cdot \left(\nabla^2 \mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\right)_{i,i}} \left(\left(\nabla \mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\right)_i + \sum_{j \neq i} \left(\widetilde{\boldsymbol{\beta}}_{M,j} - \widetilde{\boldsymbol{\beta}}_j^*\right) \left(\nabla^2 \mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\right)_{i,j} \right)}_{A_1} \\ &+ \underbrace{\frac{\left(\mathcal{L}\Big|_{\widetilde{\boldsymbol{\beta}}_{M,-i}} \right)' (\alpha_i^*) - \left(\left(\nabla \mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\right)_i + \sum_{j \neq i} \left(\widetilde{\boldsymbol{\beta}}_{M,j} - \widetilde{\boldsymbol{\beta}}_j^*\right) \left(\nabla^2 \mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\right)_{i,j} \right)}{\left(\mathcal{L}\Big|_{\widetilde{\boldsymbol{\beta}}_{M,-i}} \right)'' (b_1)}}_{A_2},\end{aligned}$$

where

$$|A_1| \le \frac{np\left\|\widetilde{\boldsymbol{\beta}}_M - \widetilde{\boldsymbol{\beta}}^*\right\|_c}{\frac{np}{\kappa_1}\left(\frac{np}{\kappa_1} - \left\|\widetilde{\boldsymbol{\beta}}_M - \widetilde{\boldsymbol{\beta}}^*\right\|_c\right)}\left|\left(\nabla\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\right)_i + \sum_{j\neq i}\left(\widetilde{\boldsymbol{\beta}}_{M,j} - \widetilde{\boldsymbol{\beta}}^*_j\right)\left(\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\right)_{i,j}\right|$$
$$\lesssim \frac{\kappa_1^4}{np}\sqrt{\frac{(d+1)\log n}{npL}}\left|\left(\nabla\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\right)_i + \sum_{j\neq i}\left(\widetilde{\boldsymbol{\beta}}_{M,j} - \widetilde{\boldsymbol{\beta}}^*_j\right)\left(\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\right)_{i,j}\right|$$

and

$$|A_2| \lesssim \frac{1}{\frac{np}{\kappa_1} - \left\|\widetilde{\boldsymbol{\beta}}_M - \widetilde{\boldsymbol{\beta}}^*\right\|_c} np\left\|\widetilde{\boldsymbol{\beta}}_M - \widetilde{\boldsymbol{\beta}}^*\right\|_c^2 \lesssim \kappa_1^5\frac{(d+1)\log n}{npL}$$

with probability exceeding $1 - O(n^{-6})$. On the other hand, we know that

$$\begin{aligned}
&\left|\left(\nabla\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\right)_i + \sum_{j\neq i}\left(\widetilde{\boldsymbol{\beta}}_{M,j} - \widetilde{\boldsymbol{\beta}}^*_j\right)\left(\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\right)_{i,j}\right| \\
\le&\left|\left(\nabla\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\right)_i\right| + \left|\sum_{j\neq i}\left(\widetilde{\boldsymbol{\beta}}_{M,j} - \widetilde{\boldsymbol{\beta}}^*_j\right)\left(\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\right)_{i,j}\right| \\
\le&\left|\left(\nabla\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\right)_i\right| + \|\widehat{\boldsymbol{\alpha}}_M - \boldsymbol{\alpha}^*\|_\infty\sum_{j\in[n],j\neq i}\left|\left(\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\right)_{i,j}\right| + \|\widehat{\boldsymbol{\beta}}_M - \boldsymbol{\beta}^*\|_2\sqrt{\sum_{k>n}\left(\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}^*)\right)_{i,k}^2} \\
\lesssim&\sqrt{\frac{np\log n}{L}} + \kappa_1^2\sqrt{\frac{(d+1)\log n}{npL}}np + \kappa_1\sqrt{\frac{\log n}{pL}}\sqrt{dnp^2} \lesssim \kappa_1^2\sqrt{\frac{(d+1)np\log n}{L}}
\end{aligned}$$

with probability at least $1 - O(n^{-6})$. To sum up, we have

$$\begin{aligned}
\left|\alpha_{M,i} - \overline{\alpha}'_i\right| \le |A_1| + |A_2| &\lesssim \frac{\kappa_1^4}{np}\sqrt{\frac{(d+1)\log n}{npL}}\kappa_1^2\sqrt{\frac{(d+1)np\log n}{L}} + \kappa_1^5\frac{(d+1)\log n}{npL} \\
&\lesssim \kappa_1^6\frac{(d+1)\log n}{npL}
\end{aligned}$$

with probability at least $1 - O(n^{-6})$. ∎

Appendix E. Details of Real Data Experiments

In this section, we provide the details in computing the maximum likelihood estimator $(\widehat{\boldsymbol{\alpha}}_M, \widehat{\boldsymbol{\beta}}_M)$. We first generated the variables $\boldsymbol{X}$ and comparisons as described in §5.4. We standardized each column to make sure they have mean 0 and standard deviation 1 and then multiplied by 2/27 (scale to the order of $\sqrt{d + 1/n}$ as mentioned in the main text). In real-world data the numbers of comparisons between each compared pair are not the same, so we

denote by $L_{i,j}$ the number of comparisons between pair (i,j). Let $\{y_{i,j}^{(l)} : (i,j) \in \mathcal{E}, l \in [L_{i,j}]\}$ be all the comparisons we have, then the negative log-likelihood can be written as

$$\mathcal{L}(\widetilde{\boldsymbol{\beta}}) := \sum_{(i,j)\in\mathcal{E}, i>j} \sum_{l=1}^{L_{i,j}} \left\{ -y_{j,i}^{(l)} \left(\widetilde{\boldsymbol{x}}_i^\top \widetilde{\boldsymbol{\beta}} - \widetilde{\boldsymbol{x}}_j^\top \widetilde{\boldsymbol{\beta}} \right) + \log\left(1 + e^{\widetilde{\boldsymbol{x}}_i^\top \widetilde{\boldsymbol{\beta}} - \widetilde{\boldsymbol{x}}_j^\top \widetilde{\boldsymbol{\beta}}} \right) \right\}.$$

References

Linas Baltrunas, Tadas Makcinskas, and Francesco Ricci. Group recommendations with rank aggregation and collaborative filtering. In *Proceedings of the fourth ACM conference on Recommender systems*, pages 119–126, 2010.

KS Banerjee. Generalized inverse of matrices and its applications, 1973.

Andrew C Berry. The accuracy of the gaussian approximation to the sum of independent variates. *Transactions of the american mathematical society*, 49(1):122–136, 1941.

Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

Siu Lun Chau, Mihai Cucuringu, and Dino Sejdinovic. Spectral ranking with covariates. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2022, Grenoble, France, September 19–23, 2022, Proceedings, Part V*, pages 70–86. Springer, 2023.

Pinhan Chen, Chao Gao, and Anderson Y Zhang. Optimal full ranking from pairwise comparisons. *The Annals of Statistics*, 50(3):1775–1805, 2022a.

Pinhan Chen, Chao Gao, and Anderson Y Zhang. Partial recovery for top-k ranking: Optimality of mle and suboptimality of the spectral method. *The Annals of Statistics*, 50(3):1618–1652, 2022b.

Xi Chen, Sivakanth Gopi, Jieming Mao, and Jon Schneider. Competitive analysis of the top-k ranking problem. In *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1245–1264. SIAM, 2017.

Yuxin Chen and Changho Suh. Spectral mle: Top-k rank aggregation from pairwise comparisons. In *International Conference on Machine Learning*, pages 371–380. PMLR, 2015.

Yuxin Chen, Jianqing Fan, Cong Ma, and Kaizheng Wang. Spectral method and regularized mle are both optimal for top-k ranking. *Annals of statistics*, 47(4):2204, 2019.

Yuxin Chen, Yuejie Chi, Jianqing Fan, Cong Ma, and Yuling Yan. Noisy matrix completion: Understanding statistical guarantees for convex relaxation via nonconvex optimization. *SIAM journal on optimization*, 30(4):3098–3121, 2020.

Yuxin Chen, Yuejie Chi, Jianqing Fan, Cong Ma, et al. Spectral methods for data science: A statistical perspective. *Foundations and Trends® in Machine Learning*, 14(5):566–806, 2021.

Weiwei Cheng, Krzysztof Dembczynski, and Eyke Hüllermeier. Label ranking methods based on the plackett-luce model. In *ICML*, 2010.

Cynthia Dwork, Ravi Kumar, Moni Naor, and Dandapani Sivakumar. Rank aggregation methods for the web. In *Proceedings of the 10th international conference on World Wide Web*, pages 613–622, 2001.

Paul Erdos, Alfréd Rényi, et al. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci*, 5(1):17–60, 1960.

Carl-Gustav Esseen. On the liapunov limit error in the theory of probability. *Ark. Mat. Astr. Fys.*, 28:1–19, 1942.

Jianqing Fan, Zhipeng Lou, Weichen Wang, and Mengxin Yu. Ranking inferences based on the top choice of multiway comparisons. *arXiv preprint arXiv:2211.11957*, 2022.

Jianqing Fan, Zhipeng Lou, Weichen Wang, and Mengxin Yu. Spectral ranking inferences based on general multiway comparisons. *arXiv preprint arXiv:2308.02918*, 2023.

Holmes Finch. An introduction to the analysis of ranked response data. *Practical Assessment, Research, and Evaluation*, 27(1):7, 2022.

Chao Gao, Yandi Shen, and Anderson Y Zhang. Uncertainty quantification in the bradley-terry-luce model. *arXiv preprint arXiv:2110.03874*, 2021.

John Guiver and Edward Snelson. Bayesian inference for plackett-luce ranking models. In *proceedings of the 26th annual international conference on machine learning*, pages 377–384, 2009.

Yuan Guo, Peng Tian, Jayashree Kalpathy-Cramer, Susan Ostmo, J Peter Campbell, Michael F Chiang, Deniz Erdogmus, Jennifer G Dy, and Stratis Ioannidis. Experimental design under the bradley-terry model. In *IJCAI*, pages 2198–2204, 2018.

Bruce Hajek, Sewoong Oh, and Jiaming Xu. Minimax-optimal inference from partial rankings. *Advances in Neural Information Processing Systems*, 27, 2014.

Ruijian Han, Rougang Ye, Chunxi Tan, and Kani Chen. Asymptotic theory of sparse bradley–terry model. *The Annals of Applied Probability*, 30(5):2491–2515, 2020.

F Maxwell Harper and Joseph A Konstan. The movielens datasets: History and context. *Acm transactions on interactive intelligent systems (tiis)*, 5(4):1–19, 2015.

David R Hunter. Mm algorithms for generalized bradley-terry models. *The annals of statistics*, 32(1):384–406, 2004.

Minje Jang, Sunghyun Kim, Changho Suh, and Sewoong Oh. Top-k ranking from pairwise comparisons: When spectral ranking is optimal. *arXiv preprint arXiv:1603.04153*, 2016.

Minje Jang, Sunghyun Kim, and Changho Suh. Top-k rank aggregation from m-wise comparisons. *IEEE Journal of Selected Topics in Signal Processing*, 12(5):989–1004, 2018.

Tao Jin, Pan Xu, Quanquan Gu, and Farzad Farnoud. Rank aggregation via heterogeneous thurstone preference models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 4353–4360, 2020.

Hanwei Li, David Simchi-Levi, Michelle Xiao Wu, and Weiming Zhu. Estimating and exploiting the impact of photo layout: A structural approach. *Available at SSRN 3470877*, 2019.

Xinran Li, Dingdong Yi, and Jun S Liu. Bayesian analysis of rank data with covariates and heterogeneous rankers. *Statistical Science*, 37(1):1–23, 2022.

Yue Liu, Ethan X Fang, and Junwei Lu. Lagrangian inference for ranking problems. *Operations Research*, 2022.

R Duncan Luce. *Individual choice behavior: A theoretical analysis*. Courier Corporation, 2012.

Cong Ma, Kaizheng Wang, Yuejie Chi, and Yuxin Chen. Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval and matrix completion. In *International Conference on Machine Learning*, pages 3345–3354. PMLR, 2018.

Kenneth Massey. Statistical models applied to the rating of sports teams. *Bluefield College*, 1997.

Lucas Maystre and Matthias Grossglauser. Fast and accurate inference of plackett–luce models. *Advances in neural information processing systems*, 28, 2015.

Sahand Negahban, Sewoong Oh, and Devavrat Shah. Iterative ranking from pair-wise comparisons. *Advances in neural information processing systems*, 25, 2012.

Ashwin Pananjady, Cheng Mao, Vidya Muthukumar, Martin J Wainwright, and Thomas A Courtade. Worst-case vs average-case design for estimation from fixed pairwise comparisons. *arXiv preprint arXiv:1707.06217*, 2017.

Robin L Plackett. The analysis of permutations. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 24(2):193–202, 1975.

Matthew Richardson, Amit Prakash, and Eric Brill. Beyond pagerank: machine learning for static ranking. In *Proceedings of the 15th international conference on World Wide Web*, pages 707–715, 2006.

Dirk Schäfer and Eyke Hüllermeier. Dyad ranking using plackett–luce models based on joint feature representations. *Machine Learning*, 107:903–941, 2018.

Nihar Shah, Sivaraman Balakrishnan, Aditya Guntuboyina, and Martin Wainwright. Stochastically transitive models for pairwise comparisons: Statistical and computational issues. In *International Conference on Machine Learning*, pages 11–20. PMLR, 2016.

Nihar B Shah and Martin J Wainwright. Simple, robust and optimal ranking from pairwise comparisons. *The Journal of Machine Learning Research*, 18(1):7246–7283, 2017.

Gordon Simons and Yi-Ching Yao. Asymptotics when the number of parameters tends to infinity in the bradley-terry model for paired comparisons. *The Annals of Statistics*, 27 (3):1041–1060, 1999.

Stephen M Stigler. Citation patterns in the journals of statistics and probability. *Statistical Science*, pages 94–108, 1994.

Balázs Szörényi, Róbert Busa-Fekete, Adil Paul, and Eyke Hüllermeier. Online rank elicitation for plackett-luce: A dueling bandits approach. *Advances in Neural Information Processing Systems*, 28, 2015.

Louis L Thurstone. The method of paired comparisons for social values. *Journal of Abnormal Psychology*, 21(4), 1927.

Louis L Thurstone. A law of comparative judgment. In *Scaling*, pages 81–92. Routledge, 2017.

Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12(4):389–434, 2012.

Joel A Tropp. An introduction to matrix concentration inequalities. *arXiv preprint arXiv:1501.01571*, 2015.

Heather Turner and David Firth. Bradley-terry models in r: the bradleyterry2 package. *Journal of Statistical Software*, 48:1–21, 2012.

Milan Vojnovic and Seyoung Yun. Parameter estimation for generalized thurstone choice models. In *International Conference on Machine Learning*, pages 498–506. PMLR, 2016.

Zhibing Zhao, Ao Liu, and Lirong Xia. Learning mixtures of random utility models with features from incomplete preferences. *arXiv preprint arXiv:2006.03869*, 2022.