# Inference for Maximin Effects: A Sampling Approach to Aggregating Heterogenous High-dimensional Regression Models *

Zijian Guo

November 15, 2020

**Abstract**

Heterogeneity is an important feature of modern data analysis and a central task is to extract information from the massive and heterogeneous data. In this paper, we consider aggregation of heterogeneous regression vectors under multiple high-dimensional linear models. We adopt the definition of maximin effect (Meinshausen, Bühlmann, AoS, 43(4), 1801–1830) and further define the maximin effect for a targeted population by allowing for covariate shift. A ridge-type maximin effect is introduced to balance reward optimality and statistical stability. To identify the maximin effect in high dimensions, we estimate the regression covariance matrix by a debiased estimator and construct the optimal weight vector for the maximin effect. The resulted estimator of the maximin effect is not necessarily asymptotic normal since the constructed weight vector might have a mixture distribution. We devise a novel sampling approach to construct confidence intervals for any linear contrast of maximin effects in high dimensions. The coverage and precision properties of the constructed confidence intervals are studied. The proposed method is demonstrated over simulations and a genetic data set on growing yeast colonies under different environments.

**Keywords:** Ridge-type maximin effect; covariate shift; stability; confidence interval; debiasing; adversarial reward.

*Z. Guo is an assistant professor at Department of Statistics, Rutgers University. The research of Z. Guo was supported in part by the NSF DMS 1811857, 2015373 and NIH R01GM140463-01.

# 1 Introduction

Data heterogeneity commonly occurs in the analysis of massive data sets, for example, the data being collected from different sub-populations [27] or the experiments being conducted under different environments [5, 15]. Despite of its importance, there is a lack of statistical inference methods for analyzing heterogeneous data sets. A central task is to effectively extract and summarize information from the massive and potentially heterogeneous data. To model the data heterogeneity, we consider multiple high-dimensional linear models,

$$Y^{(l)}_{n_l\times 1} = X^{(l)}_{n_l\times p} b^{(l)}_{p\times 1} + \epsilon^{(l)}_{n_l\times 1}, \quad \text{for} \quad 1 \leq l \leq L \tag{1}$$

where $L$ is the number of groups and for the group $l$, $b^{(l)} \in \mathbb{R}^p$ denotes the regression vector, $n_l$ denotes the sample size and $Y^{(l)} \in \mathbb{R}^{n_l}$, $X^{(l)} \in \mathbb{R}^{n_l\times p}$ and $\epsilon^{(l)} \in \mathbb{R}^{n_l}$ denote outcome, design matrix and the model error, respectively. The group label $l$ might correspond to an unique data source, for example, a specific sub-population or an experiment environment. Throughout the paper, the group labelling $l$ is assumed to be known a priori and within the group $l$, $\{(X^{(l)}_{i,\cdot}, Y^{(l)}_i)\}_{1\leq i\leq n_l}$ are identically and independently generated from the distribution $\mathcal{P}_l$ with $X^{(l)}_{1,\cdot}$ following the covariate distribution $\mathcal{P}^l_X$ and the model errors $\epsilon^{(l)}$ being independent of $X^{(l)}$. Across the groups $\{1, \ldots, L\}$, the covariate distribution $P^l_X$, the regression vector $b^{(l)} \in \mathbb{R}^p$ and the distribution of $\epsilon^{(l)}$ are allowed to vary.

To identify an effect vector which is representative for all $L$ groups, [29] introduced the maximin effect $\beta^* \in \mathbb{R}^p$ as

$$\beta^* := \arg\max_{\beta\in\mathbb{R}^p} R(\beta) \quad \text{with} \quad R(\beta) = \min_{1\leq l\leq L} \left\{ \mathbf{E}_{\mathcal{P}_l}[Y^{(l)}_1]^2 - \mathbf{E}_{\mathcal{P}_l}[Y^{(l)}_1 - (X^{(l)}_{1,\cdot})^\intercal\beta]^2 \right\}. \tag{2}$$

For a given $\beta \in \mathbb{R}^p$, $R(\beta)$ is referred to as reward since it measures the minimum explained variance across $L$ groups. The maximin effect is defined as the vector maximizing this worst-case reward $R(\beta)$. [29] showed that $\beta^* \in \mathbb{R}^p$ is identified as a weighted average of $\{b^{(l)}\}_{1\leq l\leq L}$ and the aggregation weights can be computed by solving a quadratic optimization problem.

We focus on the model (1) in the high-dimensional setting with $p \gg \max_{1\leq l\leq L}\{n_l\}$. New maximin effect definitions will be introduced to accommodate for possible covariate shift and non-uniquely defined maximin effect in (2). The major goal is to construct confidence intervals for linear contrasts of various maximin effects in high dimensions.

## 1.1 Our results and contribution

In the model (1), we study how to identify a best representation of the heterogeneous regression vectors $\{b^{(l)}\}_{1\leq l\leq L}$ for a targeted population with covariate distribution $\mathcal{Q}$. We consider the covariate shift setting: the targeted covariate distribution $\mathcal{Q}$ is possibly different from any of the covariate distributions $\{\mathcal{P}_X^l\}_{1\leq l\leq L}$. To accommodate for covariate shift, we introduce the covariate shift maximin effect $\beta^*(\mathcal{Q}) \in \mathbb{R}^p$ as a generalization of the maximin effect in (2); see its definition in (5). $\beta^*(\mathcal{Q})$ can be identified as a weighted average of $\{b^{(l)}\}_{1\leq l\leq L}$ and the weights depend on the covariate distribution $\mathcal{Q}$. Specifically, we define $\Sigma^{\mathcal{Q}} = \mathbf{E}X_{1,\cdot}^{\mathcal{Q}}(X_{1,\cdot}^{\mathcal{Q}})^{\intercal}$ with $X_{1,\cdot}^{\mathcal{Q}} \in \mathbb{R}^p$ following the covariate distribution $\mathcal{Q}$ and the regression covariance matrix $\Gamma^{\mathcal{Q}} \in \mathbb{R}^{L\times L}$ with $\Gamma_{l,k}^{\mathcal{Q}} = [b^{(l)}]^{\intercal}\Sigma^{\mathcal{Q}}b^{(k)}$ for $1 \leq l, k \leq L$. The optimal aggregation weights for $\beta^*(\mathcal{Q})$ are determined by $\Gamma^{\mathcal{Q}}$.

The maximin effect in (2) is defined to maximize the reward $R(\beta)$, but, without explicitly stated in (2), an equally important goal is to define an inference target which can be stably estimated. We introduce a ridge-type maximin effect to balance reward optimality and statistical stability. We impose a ridge penalty (with level $\delta > 0$) in construction of the aggregation weights and use these weights to define the ridge-type maximin effect $\beta_\delta^*(\mathcal{Q}) \in \mathbb{R}^p$; see (7). The ridge-type effect is a more stable target to make inference for, especially when the regression covariance matrix $\Gamma^{\mathcal{Q}}$ is nearly singular. In numerical studies, with a suitable penalty level $\delta > 0$, our proposed confidence intervals for $\beta_\delta^*(\mathcal{Q})$ are much shorter than those for the maximin effect (with no penalty); see Figure 4. Although the reward value of $\beta_\delta^*(\mathcal{Q})$ with $\delta > 0$ is typically smaller than the optimal value, we show in Theorem 1 that $\beta_\delta^*(\mathcal{Q})$ with $\delta > 0$ satisfies certain types of reward optimality. Importantly, the reward reduction for any $\delta > 0$ can be estimated in a data-dependent way. This provides a practical method for users to choose $\delta$ such that the reward of $\beta_\delta^*(\mathcal{Q})$ is comparable to the optimal reward but our inference procedures for $\beta_\delta^*(\mathcal{Q})$ are more stable.

We construct the confidence interval (CI) for $x_{\rm new}^{\intercal}\beta_\delta^*(\mathcal{Q})$ for any $x_{\rm new} \in \mathbb{R}^p$ and $\delta \geq 0$. Our proposed method is built on accurately the regression covariance matrix $\Gamma^{\mathcal{Q}}$. In the presence of covariate shift, we either assume that $\Sigma^{\mathcal{Q}}$ is known or estimate $\Sigma^{\mathcal{Q}}$ by $\widehat{\Sigma}^{\mathcal{Q}} = \sum_{i=1}^{N_{\mathcal{Q}}} X_{i,\cdot}^{\mathcal{Q}}(X_{i,\cdot}^{\mathcal{Q}})^{\intercal}/N_{\mathcal{Q}}$, where $\{X_{i,\cdot}^{\mathcal{Q}}\}_{1\leq i\leq N_{\mathcal{Q}}}$ are i.i.d. samples following the distribution $\mathcal{Q}$. For $1 \leq l \leq L$, we use $\widehat{b}_{init}^{(l)}$ to denote a reasonable penalized estimator (e.g. Lasso) of $b^{(l)}$. We propose a debiased estimator $\widehat{\Gamma}^{\mathcal{Q}}$ of $\Gamma^{\mathcal{Q}}$ through conducting entry-wise bias correction of the plug-in estimator $[\widehat{b}_{init}^{(l)}]^{\intercal}\widehat{\Sigma}^{\mathcal{Q}}\widehat{b}_{init}^{(k)}$ for $1 \leq l, k \leq L$. A novel projection direction is proposed to handle the covariate shift. For the covariate shift setting, we establish the asymptotic normality of our proposed matrix estimator $\widehat{\Gamma}^{\mathcal{Q}} \in \mathbb{R}^{L\times L}$; see Theorem 2.

A major challenge of uncertainty quantification for maximin effect lies at the irregularity of the maximin weight vector. With our proposed $\widehat{\Gamma}^{\mathcal{Q}}$, we estimate the optimal aggregation weight $\gamma^* \in \mathbb{R}^L$ by $\widehat{\gamma} \in \mathbb{R}^L$, defined as the maximizer of $\gamma^{\intercal}\widehat{\Gamma}^{\mathcal{Q}}\gamma$ over $L$-dimensional simplex. Although $\widehat{\Gamma}^{\mathcal{Q}} - \Gamma^{\mathcal{Q}}$ is asymptotically normal, the weight vector error $\widehat{\gamma} - \gamma^*$ can be the mixture of an (asymptotically) normal distribution and a point mass. The mixture distribution of the aggregation weights makes it challenging to construct CIs for the maximin effect based on asymptotic normality. To address this, we propose a novel sampling procedure for CI construction. For $1 \leq m \leq M$, we sample $\widehat{\Gamma}^{[m]}$ following the asymptotic distribution of $\widehat{\Gamma}^{\mathcal{Q}}$ and use $\widehat{\Gamma}^{[m]}$ to construct an sampled weight vector $\widehat{\gamma}^{[m]} \in \mathbb{R}^L$ and a sampled interval $\mathrm{Int}_{\alpha}^{[m]}(x_{\mathrm{new}})$. We construct CI for $x_{\mathrm{new}}^{\intercal}\beta^*_{\delta}(\mathcal{Q})$ by taking a union of the intervals $\{\mathrm{Int}_{\alpha}^{[m]}(x_{\mathrm{new}})\}_{1\leq m\leq M}$. This proposed CI is shown to have a valid coverage level. We study its precision properties by comparing its length to an oracle CI with the knowledge of optimal aggregation weights. The length of our proposed CI is at most longer than the oracle CI by a order of $\sqrt{\log M}$.

We will demonstrate the numerical performance of our proposed CI over different targeted distributions $\mathcal{Q}$ and ridge penalty levels $\delta$.

To sum up, the contributions of the current paper are three-folded.

1. We generalize the maximin effect definition [29] to allow for covariate shift and propose ridge-type maximin effects to balance reward optimality and statistical stability.

2. We propose a novel sampling approach to quantify the uncertainty of the maximin weights. To the authors' best knowledge, the proposed CI is the first inference method for the high-dimensional maximin effect.

3. We characterize the dependence of sampling accuracy on the sampling number $M$. The theoretical analysis of Theorem 3 can be useful for other sampling methods.

## 1.2 Existing literature

The most relevant works are [29, 33], where [29] introduced the concept of maximin effect and studied its estimation problem and [33] constructed CI for the maximin effect $\beta^*$ in low dimensions, relying on the asymptotic normality of the magging estimator in [29]. The current paper focuses on CI construction in high dimensions. Our proposed CI relies on a sampling procedure, instead of directly applying asymptotic normality. The proposed CI is effective for two important settings (not covered by [33]): the maximin effect $\beta^*$ is not uniquely defined and the maximin effect estimator is not asymptotically normal. The maximin projection learning was proposed in [35] for individualized treatment selection. We will show that the maximin projection is equal to a normalized maximin effect in (5); see Proposition 4 in Section A.1 in the supplement.

Inference for a single high-dimensional linear model has been actively investigated in the recent decade. Debiasing, desparsifying or Neyman's Orthogonalization [2, 13, 14, 19, 25, 42, 50] were proposed for inference for regression coefficients. Inference problems for linear contrasts and quadratic functionals were studied in [1, 9, 11, 53] and [10, 22, 23, 44], respectively. In contrast to the aforementioned works, the current paper focus on a different task of aggregating heterogeneous regression vectors. The inference target is simultaneously determined by multiple linear models and the targeted covariate distribution. Our proposed debiasing estimator for $\Gamma^{\mathcal{Q}}$ extends the existing methods to the more challenging covariate shift setting. We shall emphasize that the debiased estimator serves as an initial estimator and CI construction for maximin effects relies on the novel sampling approach.

There is an active and vast literature on analysis of heterogenous data for different purposes. Inference for the shared component of regression functions has been considered under multiple high-dimensional linear models [26] and partially linear models [51]. In contrast, the regression vectors $\{b^{(l)}\}_{1\leq l\leq L}$ in our model (2) are not assumed to share any similarity. Invariance principle has been developed to identify causal effect [30, 34] with heterogenous data. Distributional robustness has been studied in [21, 37] by minimizing a worst case over a class of distributions. We refer to [6] as a review of the connection between invariance principle and distributional robustness. Inference in the presence of covariate shift has been studied in [36, 38, 41] with the weighting methods.

Uncertainty quantification with sampling methods has a long history in statistics, such as, bootstrap [16, 17], subsampling [31], generalized fiducial inference [24, 47, 49] and repro sampling [45]. As a major difference, we do not directly sample from the original data but instead sample the estimator of the regression covariance matrix. This makes our proposed sampling approach computationally efficient even if a large number of samples are drawn.

**Notations.** Define $n = \min_{1\leq l\leq L}\{n_l\}$. Let $[p] = \{1, 2, \ldots, p\}$. For a set $S$, $|S|$ denotes the cardinality of $S$ and $S^c$ denotes its complement. For a vector $x \in \mathbb{R}^p$ and a subset $S \subset [p]$, $x_S$ is the sub-vector of $x$ with indices in $S$ and $x_{-S}$ is the sub-vector with indices in $S^c$. The $\ell_q$ norm of a vector $x$ is defined as $\|x\|_q = (\sum_{l=1}^p |x_l|^q)^{\frac{1}{q}}$ for $q \geq 0$ with $\|x\|_0 = |\{1 \leq l \leq p : x_l \neq 0\}|$ and $\|x\|_\infty = \max_{1\leq l\leq p} |x_l|$. For a matrix $X$, $X_{i,\cdot}$ and $X_{\cdot,j}$ are used to denote its $i-$th row and $j$-th column; for a set $S$, $X_{S,\cdot}$ denotes the sub-matrix of $X$ with row indices belonging to $S$. We use $c$ and $C$ to denote generic positive constants that may vary from place to place. For random objects $X_1$ and $X_2$, we use $X_1 \overset{d}{=} X_2$ to denote that they are equal in distribution. We use $X \sim \mathcal{Q}$ to denote that the random vector $X$ follows the distribution $\mathcal{Q}$ and $\{X_i\}_{1\leq i\leq n} \overset{\text{i.i.d.}}{\sim} \mathcal{Q}$ to denote that the sequence random vectors $X_i$ are independent and identically distributed (i.i.d.) following the distribution $\mathcal{Q}$. For a sequence of random variables $X_n$ indexed by $n$, we use $X_n \overset{p}{\to} X$ and $X_n \overset{d}{\to} X$ to represent that $X_n$ converges to $X$ in probability and in distribution, respectively. For two positive sequences $a_n$ and $b_n$, $a_n \lesssim b_n$ means that $\exists C > 0$ such that $a_n \leq C b_n$ for all $n$; $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $b_n \lesssim a_n$, and $a_n \ll b_n$ if $\limsup_{n\to\infty} a_n/b_n = 0$. For a matrix $A$, we use $\|A\|_F$, $\|A\|_2$ and $\|A\|_\infty$ to denote its Frobenius norm, spectral norm and element-wise maximum norm, respectively. For a symmetric matrix $A$, we use $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ to denote its maximum and minimum eigenvalues, respectively; we use $\lambda_j(A)$ to denote its $j$-th largest singular value. For a symmetric matrix $A \in \mathbb{R}^{L\times L}$ and its eigen-decomposition $A = U\Lambda U^\intercal$, we define $A_+ = U\Lambda_+ U^\intercal$ with $(\Lambda_+)_{l,l} = \max\{\Lambda_{l,l}, 0\}$ for $1 \leq l \leq L$.

For a matrix $D \in \mathbb{R}^{L\times L}$, we use $\mathrm{vecl}(D) \in \mathbb{R}^{L(L+1)/2}$ to denote the long vector of stacking the columns of the lower triangle part of $D$. We define the index mapping $\pi$ as

$$\text{for} \quad (l,k) \in \mathcal{I}_L := \{(l,k) : 1 \leq k \leq l \leq L\}, \quad \pi(l,k) = \frac{(2L-k)(k-1)}{2} + l. \tag{3}$$

This one-to-one correspondence $\pi$ maps the index set $\mathcal{I}_L = \{(l,k) : 1 \leq k \leq l \leq L\}$ of the lower triangular part of $D$ to the index set $[L(L+1)/2]$ of $\mathrm{vecl}(D)$. For $(l,k) \in \mathcal{I}_L$, we have $[\mathrm{vecl}(D)]_{\pi(l,k)} = D_{l,k}$.

## 2 Maximin Effect for A Targeted Population

We consider the heterogeneous data $\left\{Y^{(l)}, X^{(l)}\right\}_{1\leq l\leq L}$, where the group labels $\{1,2,\ldots,L\}$ are known a priori and $\{(Y_i^{(l)}, X_{i,\cdot}^{(l)})\}_{1\leq i\leq n_l} \overset{\text{i.i.d.}}{\sim} \mathcal{P}^l$ inside the group $l$. To capture the heterogeneity, we allow the distributions $\{\mathcal{P}^l\}_{1\leq l\leq L}$ to be different from each other. We consider the model (1) and the joint distribution $\mathcal{P}^l$ is determined by the covariate distribution $\mathcal{P}_X^l$, the regression vector $b^{(l)}$ and the error distribution $\epsilon_i^{(l)}$. In biology applications, we may adopt the model (1) by taking $X^{(l)}$ and $Y^{(l)}$ as the genetic variants and phenotype, respectively. When the group label $l$ corresponds to a specific environment, $b^{(l)}$ captures how genetic variants affect phenotype under this environment. With an environmental change, the regression vector $b^{(l)}$ might change across groups $\{1,2,\cdots,L\}$. If the data is collected over a different sub-population, then the covariate distribution $\mathcal{P}_X^l$ is likely to change across $1 \leq l \leq L$. The changes in $b^{(l)}$ and $\mathcal{P}_X^l$ might happen simultaneously.

We use $\mathcal{Q}$ to denote a targeted covariate distribution on the $p$-dimensional covariates and introduce the definition of covariate shift used in the current paper.

**Definition 1.** *If $\mathcal{P}_X^l \neq \mathcal{Q}$ for some $1 \leq l \leq L$, we refer to this as* **covariate shift***; If $\mathcal{P}_X^l = \mathcal{Q}$ for all $1 \leq l \leq L$, we refer to this as* **no covariate shift***.*

By the above definition, the covariate shift happens if any of the covariate distributions $\{\mathcal{P}_X^l\}_{1\leq l\leq L}$ differs from the targeted distribution $\mathcal{Q}$.

We now generalize the maximin effect definition in (2) to aggregate the data heterogeneity for a targeted population $\mathcal{Q}$. We consider the hypothetical data generation mechanism,

$$Y_1^{*,(l)} = [X_{1,\cdot}^{\mathcal{Q}}]^{\intercal} b^{(l)} + \epsilon_1^{(l)} \quad \text{for} \quad 1 \leq l \leq L, \tag{4}$$

where $b^{(l)}$ and $\epsilon_1^{(l)}$ are the same as those in (1) and $X_{1,\cdot}^{\mathcal{Q}} \sim \mathcal{Q}$. We use the super-index $^*$ in $Y_1^{*,(l)}$ to denote that the outcome variable is hypothetical instead of being observed. Under (4), we define the maximin effect with respect to the covariate distribution $\mathcal{Q}$ as

$$\beta^*(\mathcal{Q}) = \arg\max_{\beta \in \mathbb{R}^p} R_{\mathcal{Q}}(\beta) \quad \text{with} \quad R_{\mathcal{Q}}(\beta) = \min_{1 \leq l \leq L} \left[ \mathbf{E}^*(Y_{1,\cdot}^{*(l)})^2 - \mathbf{E}^*(Y_{1,\cdot}^{*(l)} - (X_{1,\cdot}^{\mathcal{Q}})^{\intercal}\beta)^2 \right]$$

where $\mathbf{E}^*$ is the expectation taken over the joint distribution of $(Y_1^{*(l)}, X_{1,\cdot}^{\mathcal{Q}})$ defined in (4). The maximin effect $\beta^*(\mathcal{Q})$ can be interpreted from an adversarial perspective [29]: in a two-side game, we select an effect vector $\beta \in \mathbb{R}^p$ and the counter agent will then choose the most challenging scenario (from one of the $L$ groups) for this selected $\beta$. Our goal is to select the effect vector $\beta \in \mathbb{R}^p$ such that the worst-case reward $R(\beta)$ returned by the counter agent is maximized. With (4), we simplify the expression of $\beta^*(\mathcal{Q})$ as

$$\beta^*(\mathcal{Q}) = \arg\max_{\beta \in \mathbb{R}^p} R_{\mathcal{Q}}(\beta) \quad \text{with} \quad R_{\mathcal{Q}}(\beta) = \min_{b \in \mathbb{B}} \left[ 2b^{\intercal}\Sigma^{\mathcal{Q}}\beta - \beta^{\intercal}\Sigma^{\mathcal{Q}}\beta \right] \tag{5}$$

where $\mathbb{B} = \{b^{(1)}, \ldots, b^{(L)}\}$ and $\Sigma^{\mathcal{Q}} = \mathbf{E} X_{1,\cdot}^{\mathcal{Q}} (X_{1,\cdot}^{\mathcal{Q}})^{\intercal}$.

A few remarks are in order for the definition (5). First, in the no covariate shift setting, the covariate shift maximin effect $\beta^*(\mathcal{Q})$ in (5) is reduced to the regular maximin effect (2). Second, the definition of $\beta^*(\mathcal{Q})$ is useful in aggregating the outcome models for a new or unseen distribution, that is, the target covariate distribution $\mathcal{Q}$ differs from those of $X_{i,\cdot}^{(l)}$. As long as the model (1) is correctly specified, the regression vectors $\{b^{(l)}\}_{1 \leq l \leq L}$ remain the same even in the presence of covariate shift. However, the maximin effect $\beta^*(\mathcal{Q})$ depends on the target distribution $\mathcal{Q}$. Third, the definition in (5) decouples two sources of data heterogeneity: changes in $\{b^{(1)}, \ldots, b^{(L)}\}$ and the covariate shift. With respect to $\mathcal{Q}$, the maximin effect $\beta^*(\mathcal{Q})$ itself is defined as an optimal aggregation of heterogeneous regression vectors $\{b^{(1)}, \ldots, b^{(L)}\}$. In addition, the maximin effect in (5) allows us to explore the heterogeneity in covariate distributions. After calculating $\beta^*(\mathcal{Q}_1), \ldots, \beta^*(\mathcal{Q}_J)$ for different covariate distributions $\{\mathcal{Q}_1, \cdots, \mathcal{Q}_J\}$ with a positive integer $J > 0$, it is possible to further combine the information contained in these maximin effects; see Figure 7 for an example.

The following proposition shows how to identify the maximin effects $\beta^*(\mathcal{Q})$.

**Proposition 1.** *If $\lambda_{\min}(\Sigma^{\mathcal{Q}}) > 0$, then $\beta^*(\mathcal{Q})$ defined in (5) can be expressed as*

$$\beta^*(\mathcal{Q}) = \sum_{l=1}^{L} [\gamma^*(\mathcal{Q})]_l b^{(l)} \quad \textit{with} \quad \gamma^*(\mathcal{Q}) = \arg\min_{\gamma \in \Delta^L} \gamma^{\intercal}\Gamma^{\mathcal{Q}}\gamma \tag{6}$$

*where* $\Gamma^{\mathcal{Q}}_{lk} = (b^{(l)})^\intercal \Sigma^{\mathcal{Q}} b^{(k)}$ *for* $1 \leq l, k \leq L$ *and* $\Delta^L = \{\gamma \in \mathbb{R}^L : \gamma_j \geq 0, \ \sum_{j=1}^{L} \gamma_j = 1\}$ *is the simplex over* $\mathbb{R}^L$. *Furthermore,* $\max_{\beta \in \mathbb{R}^p} \min_{b \in \mathbb{B}} \left[2b^\intercal \Sigma^{\mathcal{Q}} \beta - \beta^\intercal \Sigma^{\mathcal{Q}} \beta\right] = [\beta^*(\mathcal{Q})]^\intercal \Sigma^{\mathcal{Q}} \beta^*(\mathcal{Q})$.

The above proposition shows that the covariate shift maximin effect $\beta^*(\mathcal{Q})$ can be identified as a weighted average of $\{b^{(l)}\}_{1 \leq l \leq L}$ and the aggregation weight vector $\gamma^*(\mathcal{Q})$ depends on the covariate distribution $\mathcal{Q}$. The above proposition provides an explicit way of computing the maximin effect when the regression covariance matrix $\Gamma^{\mathcal{Q}} \in \mathbb{R}^{L \times L}$ is known. Proposition 1 is implied by Theorem 1 in [29] and the definition of $\beta^*(\mathcal{Q})$ in (5).

**Remark 1.** In studying the individualized treatment effect, [35] proposed the maximum projection $\beta^{*,\mathrm{MP}} = \arg\max_{\|\beta\|_2 \leq 1} \min_{1 \leq l \leq L} \beta^\intercal b^{(l)}$. Although the model in [35] is different from our model (1), we can identify the maximin projection $\beta^{*,\mathrm{MP}}$ as a scaled version of the general maximin effect $\beta^*$ in (5) with the targeted covariate distribution $\mathcal{Q}$ as the identity design. See more details in Section A.1 in the supplement.

# 3 Ridge-type Maximin: Optimality and Stability

In this section, we define a ridge-type maximin effect, which strikes the balance between reward optimality and statistical stability. The maximin effect $\beta^*(\mathcal{Q})$ defined in (5) is only targeted at optimizing the reward function, without considering the stability. Since our goal is to effectively aggregate $\{b^{(l)}\}_{1 \leq l \leq L}$, a better strategy is to define a vector $\beta \in \mathbb{R}^p$ such that the corresponding reward $R(\beta)$ is comparable to $R(\beta^*)$ but we are able to estimate this $\beta$ more accurately than the maximin effect $\beta^*$.

One challenging setting for making stable inference for $\beta^*$ is that the weight vector $\gamma^*$ in (6) might not even be uniquely defined. Through numerical explorations, [33] observed that, if the maximin effect is not uniquely defined, confidence intervals are typically wide even in the low-dimensional setting. This suggests a source of instability: $\Gamma^{\mathcal{Q}}$ in (6) is singular or nearly singular.

We introduce the ridge-type maximin effect as a more stable inference target,

$$\beta^*_\delta(\mathcal{Q}) = \sum_{l=1}^{L} [\gamma^*_\delta(\mathcal{Q})]_l \cdot b^{(l)} \quad \text{with} \quad \gamma^*_\delta(\mathcal{Q}) = \arg\min_{\gamma \in \Delta^L} \left[\gamma^\intercal \Gamma^{\mathcal{Q}} \gamma + \delta \|\gamma\|_2^2\right] \tag{7}$$

where the ridge penalty $\delta\|\gamma\|_2^2$ is imposed to compute the aggregation weight and $\beta^*_{\delta=0}$ is reduced to $\beta^*$ in (5). Even if the maximin effect in (5) is not uniquely defined, the ridge-type maximin effect defined in (7) will be uniquely defined for any given $\delta > 0$.

When there is no confusion from the context, we will omit $\mathcal{Q}$ in the corresponding discussion and write $\Gamma^{\mathcal{Q}}, \beta^*(\mathcal{Q}), \gamma^*(\mathcal{Q}), \beta^*_\delta(\mathcal{Q}), \gamma^*_\delta(\mathcal{Q})$ as $\Gamma, \beta^*, \gamma^*, \beta^*_\delta, \gamma^*_\delta$, respectively.

We now demonstrate the stability of the ridge-type maximin effect by considering the special case $L = 2$. The explicit expression $\gamma^*_\delta = ([\gamma^*_\delta]_1, [\gamma^*_\delta]_2)^\intercal \in \Delta^2$ in (7) is obtained with $[\gamma^*_\delta]_1 = \arg\min_{\gamma_1 \in [0,1]} \gamma_1^2(\Gamma_{11} + \delta) + 2\Gamma_{12}\gamma_1(1-\gamma_1) + (\Gamma_{22} + \delta)(1-\gamma_1)^2$. We solve $[\gamma^*_\delta]_1$ as

$$[\gamma^*_\delta]_1 = \min\left\{\max\left\{\frac{\Gamma_{22} + \delta - \Gamma_{12}}{\Gamma_{11} + \Gamma_{22} + 2\delta - 2\Gamma_{12}}, 0\right\}, 1\right\}. \tag{8}$$

In (8) with $\Gamma_{11} + \Gamma_{22} - 2\Gamma_{12}$ being close to zero, if $\delta = 0$, it will be challenging to stably estimate the weight $[\gamma^*_\delta]_1$ as even small estimation errors of $\{\Gamma_{11}, \Gamma_{22}, \Gamma_{12}\}$ might lead to a large estimation error of the weight. In contrast, it is much easier to construct a stable estimator of $[\gamma^*_\delta]_1$ for a positive value of $\delta$. The following proposition characterizes the property of $\beta^*_\delta$ for the setting $\Gamma_{11} + \Gamma_{22} - 2\Gamma_{12} \to 0$, that is, $\lambda_{\min}(\Gamma) \to 0$.

**Proposition 2.** *For the setting $L = 2$, we assume that, with $n, p \to \infty$, $\max\{|\Gamma_{11} - \Gamma_{12}| + |\Gamma_{22} - \Gamma_{12}|\} \to 0$ and $\delta = \delta(n,p) \gg \max\{|\Gamma_{11} - \Gamma_{12}| + |\Gamma_{22} - \Gamma_{12}|\}$. Then $\beta^*_\delta$ and $\gamma^*_\delta$ defined in* (7) *satisfy $[\gamma^*_\delta]_1 \to 1/2$ and $R_{\mathcal{Q}}(\beta^*_\delta) - R_{\mathcal{Q}}(\beta^*) \to 0$, with $n, p \to \infty$.*

Since $\Gamma_{11} + \Gamma_{22} - 2\Gamma_{12} = (b^{(1)} - b^{(2)})^\intercal \Sigma^{\mathcal{Q}} (b^{(1)} - b^{(2)})$, then $\Gamma_{11} + \Gamma_{22} - 2\Gamma_{12} \to 0$ corresponds to $\|b^{(1)} - b^{(2)}\|_2 \to 0$ for a positive definite $\Sigma^{\mathcal{Q}}$. Then Proposition 2 states that, if $\|b^{(1)} - b^{(2)}\|_2$ is close to zero, the ridge-type maximin effect (with $\delta > 0$) automatically aggregates $b^{(1)}$ and $b^{(2)}$ with asymptotically equal weights. Furthermore, the ridge-type maximin effect is asymptotically achieving the optimal reward as the non-penalized maximin effect.

The penalty value $\delta$ affects the stability of our proposed estimator for a general $L \geq 2$. We will characterize the dependence on $\delta$ in Theorem 3.

In Section 7, we consider the simulation setting 1 with $L = 2$, $p = 500$ and $\Gamma_{1,1} = 3.96$, $\Gamma_{1,2} = 3.97$ and $\Gamma_{2,2} = 4.02$ and have $\gamma^*_{\delta=0} = (1,0)^\intercal$ and $\gamma^*_{\delta=2} = (0.53, 0.47)^\intercal$. We observe that the CIs for the ridge-type effect (with $\delta > 0$) are typically arrower than CIs for the non-penalized maximin effect (with $\delta = 0$); for $\delta = 2$, the CI lengths are only half of those for $\delta = 0$; see Table 1 and Figure 4 in Section 7 for details.

In the following, we discuss what is the effect of $\delta$ on the reward value $R_{\mathcal{Q}}(\beta^*_\delta)$. To measure the reward optimality, we introduce the $\nu$-optimal maximin effect:

**Definition 2.** *A vector $\beta \in \mathbb{R}^p$ is defined to be $\nu$-optimal maximin effect if it satisfies $R_{\mathcal{Q}}[\beta] \geq R_{\mathcal{Q}}[\beta^*(\mathcal{Q})] - \nu$ where the reward function $R_{\mathcal{Q}}$ and the maximin effect $\beta^*(\mathcal{Q})$ are defined in* (5) *and $\nu = \nu(n,p) > 0$ is a positive number possibly growing with $n$ and $p$.*

For the $\nu$-optimal effect vector, the corresponding reward $R_{\mathcal{Q}}[\beta]$ might be worse off than the optimal reward $R_{\mathcal{Q}}[\beta^*]$ but the reduction is at most $\nu$. The following theorem characterizes the worse-off amount for the ridge-type maximin effect in (7).

**Theorem 1.** *Suppose $\lambda_L(\mathcal{B}) > 0$ with $\mathcal{B} = \left(b^{(1)}, \ldots, b^{(L)}\right) \in \mathbb{R}^{p \times L}$, then the ridge-type minimizer $\beta^*_\delta$ defined in* (7) *satisfies*

$$R_{\mathcal{Q}}(\beta^*_\delta) \geq R_{\mathcal{Q}}(\beta^*) - 2\delta(\|\gamma^*_\delta\|_\infty - \|\gamma^*_\delta\|_2^2) \geq R_{\mathcal{Q}}(\beta^*) - \frac{\delta}{2} \cdot \left(1 - \frac{1}{L}\right), \tag{9}$$

*where $R_{\mathcal{Q}}$ and $\beta^*$ are defined in* (5) *and $\gamma^*_\delta$ is defined in* (7).

The condition $\lambda_L(\mathcal{B}) > 0$ rules out the exact collinearity among the columns of $\mathcal{B}$; however, the above theorem allows for $\lambda_L(\mathcal{B}) \to 0$ with $n, p \to \infty$, that is, the nearly collinear setting. The penalty level $\delta$ in (7) controls the sub-optimality of the ridge-type maximin effect: $\beta^*_\delta$ is $\nu$-optimal with $\nu = 2\delta(\|\gamma^*_\delta\|_\infty - \|\gamma^*_\delta\|_2^2)$. In Figure 1, we plot the reward values $R_{\mathcal{Q}}[\beta^*_\delta]$ over $\delta \in [0, 5]$ for three simulation settings detailed in Section 7. For settings 1 and 2, for $0 \leq \delta \leq 5$, $R_{\mathcal{Q}}(\beta^*_\delta)$ is above 95% of the optimal value $R_{\mathcal{Q}}(\beta^*)$. For setting 3, $R_{\mathcal{Q}}(\beta^*_\delta)$ is more sensitive to the choice of $\delta$.

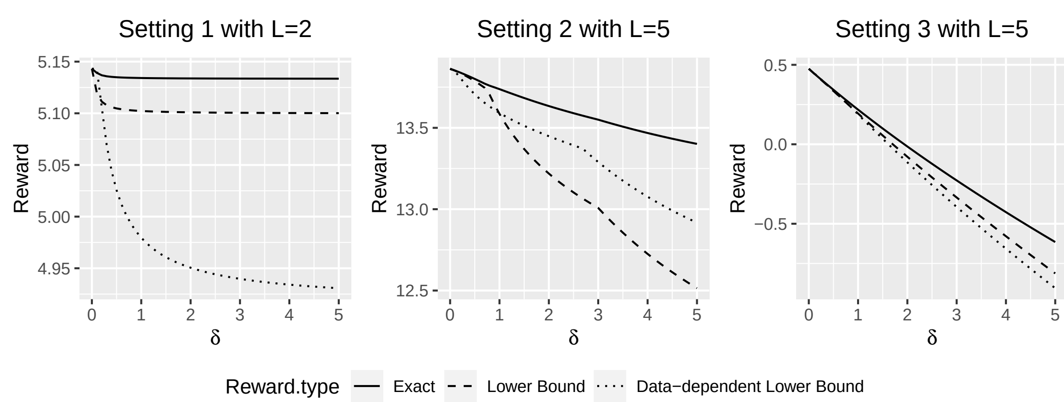



Figure 1: Dependence of reward on $\delta$: the reward type "Exact", "Lower Bound" and "Data-dependent Lower Bound" correspond to $R_{\mathcal{Q}}(\beta^*_\delta)$, $R_{\mathcal{Q}}(\beta^*) - 2\delta(\|\gamma^*_\delta\|_\infty - \|\gamma^*_\delta\|_2^2)$ and $R_{\mathcal{Q}}(\beta^*) - 2\delta(\|\widehat{\gamma}_\delta\|_\infty - \|\widehat{\gamma}_\delta\|_2^2)$, respectively. The exact settings are specified in Section 7.

A useful implication of Theorem 1 is to quantify the reward reduction $R_{\mathcal{Q}}[\beta^*] - R_{\mathcal{Q}}[\beta^*_\delta]$ in a data dependent way. Specifically, we estimate $\gamma^*_\delta$ by a consistent estimator $\widehat{\gamma}_\delta$ (see (11)) and estimate an upper bound for the reward reduction by $2\delta(\|\widehat{\gamma}_\delta\|_\infty - \|\widehat{\gamma}_\delta\|_2^2)$. In Figure 1, we plot lower bounds $R_{\mathcal{Q}}(\beta^*) - 2\delta(\|\gamma^*_\delta\|_\infty - \|\gamma^*_\delta\|_2^2)$ and $R_{\mathcal{Q}}(\beta^*) - 2\delta(\|\widehat{\gamma}_\delta\|_\infty - \|\widehat{\gamma}_\delta\|_2^2)$. We have discussed how to choose $\delta$ in a data-dependent way at the beginning of Section 7; see Figure 3 and the related discussion.

# 4 Challenges of Inference for Maximin Effects

## 4.1 High-dimensional challenge: bias and variance tradeoff

We briefly review the inference procedure proposed in [33] for the low-dimensional setting. Denote $\widetilde{b}^{(l)}$ by the ordinary least square estimator computed based on $(X^{(l)}, Y^{(l)})$. [29] proposed the magging estimator $\widetilde{\beta} = \sum_{l=1}^{L} \widetilde{\gamma}_l \widetilde{b}^{(l)}$ where the weight vector $\widetilde{\gamma} \in \mathbb{R}^L$ is defined as $\widetilde{\gamma} = \arg\min_{\gamma \in \Delta^L} \gamma^\intercal \widetilde{\Gamma} \gamma$ with

$$\widetilde{\Gamma}_{l,k} = [\widetilde{b}^{(l)}]^\intercal \left( \frac{1}{\sum_{l=1}^{L} n_l} \sum_{l=1}^{L} \sum_{i=1}^{n_l} X_{i\cdot}^{(l)} [X_{i\cdot}^{(l)}]^\intercal \right) \widetilde{b}^{(k)} \quad \text{for} \quad 1 \leq l, k \leq L. \tag{10}$$

In the low-dimensional setting with no covariate shift, valid inference procedures have been proposed in [33] by establishing the asymptotic normality of the magging estimator $\widetilde{\beta}$ under certain assumptions; see Theorem 1 of [33].

The estimator $\widetilde{\Gamma}_{l,k}$ in (10) can be viewed as a plug-in estimator of $\Gamma_{l,k} = [b^{(l)}]^\intercal \Sigma b^{(k)}$, which replaces $\Sigma, b^{(l)}$ and $b^{(k)}$ by corresponding reasonable estimators. Despite its effectiveness in low dimensions, the plug-in type estimator of $\Gamma$ as in (10) is in general not effective in high dimensions. We consider two state-of-the-art estimators as examples. For $p \geq n$, we might take $\widetilde{b}^{(l)}$ in (10) as the Lasso estimator [40]. The resulted weight vector $\widetilde{\gamma}$ is not suitable for further statistical inference as the plug-in estimator $\widetilde{\Gamma}_{l,k}$ in (10) inherits the bias from $\widetilde{b}^{(l)}$ and $\widetilde{b}^{(k)}$ [10, 25, 42, 50]. As an alternative, we might take $\widetilde{b}^{(l)}$ as the coordinate bias-corrected estimator proposed by [25, 42, 50]. However, the plug-in debiased Lasso estimator induces both large bias and variance in estimating $\Gamma_{l,k}$ since a main component $[\widetilde{b}^{(l)} - b^{(l)}]^\intercal \Sigma [\widetilde{b}^{(k)} - b^{(k)}]$ of the estimation error $\widetilde{\Gamma}_{l,k} - \Gamma_{l,k}$ is of order $p/n$. These issues are demonstrated in a numerical study in Section D.1 in the supplement.

We address these issues in Section 5.1 through directly estimating the matrix $\Gamma^{\mathcal{Q}} \in \mathbb{R}^{L \times L}$ in the presence of covariate shift. In the rest of the current section, we use $\widehat{\Gamma}^{\mathcal{Q}}$ to denote a data-dependent estimator of $\Gamma^{\mathcal{Q}}$ and define $\widehat{\gamma}_\delta \in \mathbb{R}^L$ as the ridge-type weight vector,

$$\widehat{\gamma}_\delta = \arg\min_{\gamma \in \Delta^L} \left[ \gamma^\intercal \widehat{\Gamma}^{\mathcal{Q}} \gamma + \delta \|\gamma\|_2^2 \right]. \tag{11}$$

## 4.2 Maximin effect challenge: mixture distribution

Beyond high dimensionality, a further challenge is the irregularity of the maximin effect as a function of the model parameters. To illustrate this, we consider the special case with $L = 2$ and $\delta = 0$ and the optimal weight is $(\gamma_1^*, 1 - \gamma_1^*) \in \mathbb{R}^2$ where $\gamma_1^*$ is defined in (8) with

$\delta = 0$. We obtain an explicit solution $(\widehat{\gamma}_1, 1 - \widehat{\gamma}_1) \in \mathbb{R}^2$ of (11) with $\delta = 0$:

$$\widehat{\gamma}_1 = \min\{\max\{\bar{\gamma}_1, 0\}, 1\} \quad \text{with} \quad \bar{\gamma}_1 = \frac{\widehat{\Gamma}_{22} - \widehat{\Gamma}_{12}}{\widehat{\Gamma}_{11} + \widehat{\Gamma}_{22} - 2\widehat{\Gamma}_{12}}. \tag{12}$$

Even if the errors $(\widehat{\Gamma}_{11} - \Gamma_{11}, \widehat{\Gamma}_{12} - \Gamma_{12}, \widehat{\Gamma}_{22} - \Gamma_{22})^\intercal$ are asymptotically normal, $\widehat{\gamma}_1 - \gamma_1^*$ is not necessarily asymptotically normal. This happens due to the irregularity in the definition of $\gamma_1^*$: the truncation at the thresholds 0 and 1. We decompose the error $\widehat{\gamma}_1 - \gamma_1^*$ as

$$(\bar{\gamma}_1 - \gamma_1^*) \cdot \mathbf{1}\{0 < \bar{\gamma}_1 < 1\} + (-\gamma_1^*) \cdot \mathbf{1}\{\bar{\gamma}_1 \leq 0\} + (1 - \gamma_1^*) \cdot \mathbf{1}\{\bar{\gamma}_1 \geq 1\}, \tag{13}$$

with $\bar{\gamma}_1$ defined in (12). By taking $\gamma_1^* = \gamma_1^*(n,p) \asymp 1/\sqrt{n}$, the probability of $\{\bar{\gamma}_1 \leq 0\}$ might not vanish while the probability of $\{\bar{\gamma}_1 \geq 1\}$ is asymptotically zero. In this case, if the term $\sqrt{n}(\bar{\gamma}_1 - \gamma_1^*)$ is asymptotically normal, then $\sqrt{n}(\widehat{\gamma}_1 - \gamma_1^*)$ will be a mixture of an asymptotic normal and a single mass at the point $-\sqrt{n}\gamma_1^*$. For a positive $\Gamma_{11} + \Gamma_{22} - 2\Gamma_{12}$, then $\gamma_1^* \asymp 1/\sqrt{n}$ can be made more explicit with expressing it in the form of $b^{(1)}$ and $b^{(2)}$: $\left|(b^{(1)} - b^{(2)})^\intercal \Sigma^{\mathbb{Q}} b^{(2)}\right| = c/\sqrt{n}$ for a small positive $c > 0$. That is, the irregularity setting is when the difference vector $b^{(1)} - b^{(2)}$ is nearly orthogonal to $\Sigma^{\mathbb{Q}} b^{(2)}$. By symmetry, another irregular setting is $\left|(b^{(2)} - b^{(1)})^\intercal \Sigma^{\mathbb{Q}} b^{(1)}\right| = c/\sqrt{n}$, which results in $|\gamma_1^* - 1| \asymp 1/\sqrt{n}$ and $\sqrt{n}(\widehat{\gamma} - \gamma^*)$ becomes the mixture of an asymptotic normal distribution and a point mass at $\sqrt{n}(1 - \gamma^*)$. In Figure 2, we illustrate the mixture distribution in (13) by reporting the proportions of 0 and 1 for $\widehat{\gamma}_1$ in (12). When $\gamma_1^*$ is close to 0, $\widehat{\gamma}_1 = 0$ for more than 15% of 500 simulations; when $\gamma_1^*$ is close to 1, $\widehat{\gamma}_1 = 1$ for more than 60% of 500 simulations.

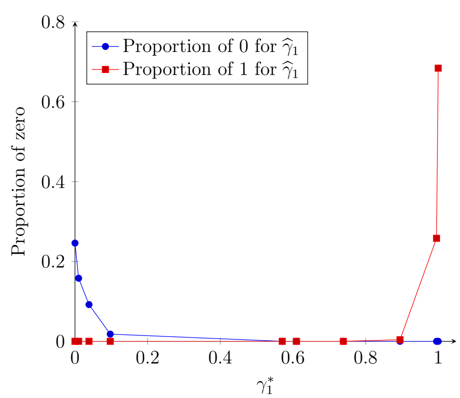



Figure 2: The proportions of 0 and 1 (out of 500 simulations) for $\widehat{\gamma}_1$ with $\gamma_1^*$ changing from 0 to 1. The simulation setting is the same as that of Figure 6 in Section 7.

The mixture distribution for the weight vector in (13) has posed challenges for constructing CIs for the maximin effect relying on asymptotic normal limiting distribution. To address this, we shall devise a novel sampling approach in Section 5.2.

# 5 Inference for Maximin Effects in High Dimensions

## 5.1 Estimation of $\Gamma^{\mathcal{Q}}$ with covariate shift

When there is possible covariate shift, we assume that we either have access to a sample $\{X_{i,\cdot}^{\mathcal{Q}}\}_{1\leq i\leq N_{\mathcal{Q}}} \overset{\text{i.i.d}}{\sim} \mathcal{Q}$ or have a prior knowledge of $\Sigma^{\mathcal{Q}} = \mathbf{E} X_{1,\cdot}^{\mathcal{Q}}(X_{1,\cdot}^{\mathcal{Q}})^{\intercal}$. For $1 \leq l \leq L$, we randomly split the data $(X^{(l)}, Y^{(l)})$ into two approximate equal-size subsamples $(X_{A_l,\cdot}^{(l)}, Y_{A_l}^{(l)})$ and $(X_{B_l,\cdot}^{(l)}, Y_{B_l}^{(l)})$, where the index sets $A_l$ and $B_l$ satisfy $A_l \cap B_l = \varnothing$, $A_l \cup B_l = [n_l]$ and $|A_l| = \lfloor n_l/2 \rfloor$. We randomly split the data $\{X_{i,\cdot}^{\mathcal{Q}}\}_{1\leq i\leq N_{\mathcal{Q}}}$ as $X_{A,\cdot}^{\mathcal{Q}}$ and $X_{B,\cdot}^{\mathcal{Q}}$, where the index sets $A$ and $B$ satisfy $A \cap B = \varnothing$, $A \cup B = [N_{\mathcal{Q}}]$ and $|A| = \lfloor N_{\mathcal{Q}}/2 \rfloor$.

The proposed estimator $\widehat{\Gamma}^{\mathcal{Q}}$ is of two steps. In the first step, we estimate $\{b^{(l)}\}_{1\leq l\leq L}$ by applying Lasso [40] to the sub-sample with the index set $A_l$:

$$\widehat{b}_{init}^{(l)} = \arg\min_{b\in\mathbb{R}^p} \frac{\|Y_{A_l}^{(l)} - X_{A_l,\cdot}^{(l)} b\|_2^2}{2|A_l|} + \lambda_l \sum_{j=1}^{p} \frac{\|X_{A_l,j}^{(l)}\|_2}{\sqrt{|A_l|}} |b_j|, \text{ with } \lambda_l = \sqrt{\frac{(2+c)\log p}{|A_l|}}\sigma_l \tag{14}$$

for some constant $c > 0$. As alternatives, we can replace Lasso with other tuning-free penalized estimators, such as scaled Lasso [39] or square-root Lasso [3]. We further construct an initial estimator of $\Gamma_{l,k}^{\mathcal{Q}}$ as $[\widehat{b}_{init}^{(l)}]^{\intercal}\widehat{\Sigma}^{\mathcal{Q}}\widehat{b}_{init}^{(k)}$ for $1 \leq l, k \leq L$ with $\widehat{\Sigma}^{\mathcal{Q}} = \frac{1}{|B|}\sum_{i\in B} X_{i,\cdot}^{\mathcal{Q}}(X_{i,\cdot}^{\mathcal{Q}})^{\intercal}$.

In the second step, we correct the bias of the initial estimator using the other half data. Specifically, the plug-in estimator $[\widehat{b}_{init}^{(l)}]^{\intercal}\widehat{\Sigma}^{\mathcal{Q}}\widehat{b}_{init}^{(k)}$ has the error decomposition:

$$\begin{aligned}(\widehat{b}_{init}^{(l)})^{\intercal}\widehat{\Sigma}^{\mathcal{Q}}\widehat{b}_{init}^{(k)} - (b^{(l)})^{\intercal}\Sigma^{\mathcal{Q}}b^{(k)} &= (\widehat{b}_{init}^{(k)})^{\intercal}\widehat{\Sigma}^{\mathcal{Q}}(\widehat{b}_{init}^{(l)} - b^{(l)}) + (\widehat{b}_{init}^{(l)})^{\intercal}\widehat{\Sigma}^{\mathcal{Q}}(\widehat{b}_{init}^{(k)} - b^{(k)}) \\ &\quad - (\widehat{b}_{init}^{(l)} - b^{(l)})^{\intercal}\widehat{\Sigma}^{\mathcal{Q}}(\widehat{b}_{init}^{(k)} - b^{(k)}) + (b^{(l)})^{\intercal}(\widehat{\Sigma}^{\mathcal{Q}} - \Sigma^{\mathcal{Q}})b^{(k)}.\end{aligned}$$

The bias correction relies on accurately estimating $(\widehat{b}_{init}^{(k)})^{\intercal}\widehat{\Sigma}^{\mathcal{Q}}(\widehat{b}_{init}^{(l)} - b^{(l)})$ and $(\widehat{b}_{init}^{(l)})^{\intercal}\widehat{\Sigma}^{\mathcal{Q}}(\widehat{b}_{init}^{(k)} - b^{(k)})$, which are dominating terms in the above error decomposition.

We detail how to approximate $(\widehat{b}_{init}^{(k)})^{\intercal}\widehat{\Sigma}^{\mathcal{Q}}(\widehat{b}_{init}^{(l)} - b^{(l)})$ in a data-dependent way and a similar procedure will be proposed to approximate $(\widehat{b}_{init}^{(l)})^{\intercal}\widehat{\Sigma}^{\mathcal{Q}}(\widehat{b}_{init}^{(k)} - b^{(k)})$. With $\widehat{\Sigma}^{(l)} = \frac{1}{|B_l|}\sum_{i\in B_l} X_{i,\cdot}^{(l)}[X_{i,\cdot}^{(l)}]^{\intercal}$ and $\widehat{u}^{(l,k)} \in \mathbb{R}^p$ denoting a projection direction to be constructed, we approximate $(\widehat{b}_{init}^{(k)})^{\intercal}\widehat{\Sigma}^{\mathcal{Q}}(\widehat{b}_{init}^{(l)} - b^{(l)})$ by $-\frac{1}{|B_l|}[\widehat{u}^{(l,k)}]^{\intercal}[X_{B_l,\cdot}^{(l)}]^{\intercal}(Y_{B_l}^{(l)} - X_{B_l,\cdot}^{(l)}\widehat{b}_{init}^{(l)})$ and the approximation error is decomposed as

$$[\widehat{\Sigma}^{(l)}\widehat{u}^{(l,k)} - \widehat{\Sigma}^{\mathcal{Q}}\widehat{b}_{init}^{(k)}]^{\intercal}(\widehat{b}_{init}^{(l)} - b^{(l)}) - \frac{1}{|B_l|}[\widehat{u}^{(l,k)}]^{\intercal}[X_{B_l,\cdot}^{(l)}]^{\intercal}\epsilon_{B_l}^{(l)}. \tag{15}$$

The intuition is to construct $\widehat{u}^{(l,k)} \in \mathbb{R}^p$ satisfying $\widehat{\Sigma}^{(l)}\widehat{u}^{(l,k)} - \widehat{\Sigma}^{\mathcal{Q}}\widehat{b}_{init}^{(k)} \approx \mathbf{0}$, which guarantees the first term of (15) to be small. We construct $\widehat{u}^{(l,k)}$ as follows

$$\widehat{u}^{(l,k)} = \arg\min_{u\in\mathbb{R}^p} u^{\intercal}\widehat{\Sigma}^{(l)}u \quad \text{subject to } \|\widehat{\Sigma}^{(l)}u - \omega^{(k)}\|_\infty \leq \|\omega^{(k)}\|_2\mu_l \tag{16}$$

$$\left|[\omega^{(k)}]^{\intercal}\widehat{\Sigma}^{(l)}u - \|\omega^{(k)}\|_2^2\right| \leq \|\omega^{(k)}\|_2^2\mu_l \tag{17}$$

where $\mu_l \asymp \sqrt{\log p/|B_l|}$ and

$$\omega^{(k)} = \widetilde{\Sigma}^{\mathcal{Q}}\widehat{b}^{(k)}_{init} \in \mathbb{R}^p \quad \text{with} \quad \widetilde{\Sigma}^{\mathcal{Q}} = \frac{1}{|A|}\sum_{i\in A} X^{\mathcal{Q}}_{i,\cdot}(X^{\mathcal{Q}}_{i,\cdot})^{\intercal}. \tag{18}$$

Since the objective $u^{\intercal}\widehat{\Sigma}^{(l)}u$ in (16) is proportional to the variance of the second term in (15), the constraint and the objective value in (16) are used to control the first and the second term in (15), respectively. The additional constraint (17) is seemingly useless to control the approximation error in (15). However, we shall emphasize that this additional constraint ensures that the second term in (15) dominates the first term in (15), which is critical in constructing an asymptotically normal estimator of $\Gamma^{\mathcal{Q}}_{l,k}$. The additional constraint (17) is particularly useful in the presence of covariate shift, that is, $\Sigma^{(l)} \neq \Sigma^{\mathcal{Q}}$ for some $1 \leq l \leq L$.

Finally, we construct the bias-corrected estimator of $\Gamma^{\mathcal{Q}}_{l,k}$ as

$$\widehat{\Gamma}^{\mathcal{Q}}_{l,k} = (\widehat{b}^{(l)}_{init})^{\intercal}\widehat{\Sigma}^{\mathcal{Q}}\widehat{b}^{(k)}_{init} + [\widehat{u}^{(l,k)}]^{\intercal}\frac{1}{|B_l|}[X^{(l)}_{B_l,\cdot}]^{\intercal}(Y^{(l)}_{B_l} - X^{(l)}_{B_l,\cdot}\widehat{b}^{(l)}_{init}) + [\widehat{u}^{(k,l)}]^{\intercal}\frac{1}{|B_k|}[X^{(k)}_{B_k,\cdot}]^{\intercal}(Y^{(k)}_{B_k} - X^{(k)}_{B_k,\cdot}\widehat{b}^{(k)}_{init}) \tag{19}$$

where $\widehat{u}^{(l,k)}$ and $\widehat{u}^{(k,l)}$ are defined in the optimization algorithm (16) and (17). Then we estimate the maximin weight vector by $\widehat{\gamma}_\delta$ defined in (11).

If $\Sigma^{\mathcal{Q}}$ is known, we modify $\widehat{\Gamma}^{\mathcal{Q}}_{l,k}$ in (19) by replacing $\widehat{\Sigma}^{\mathcal{Q}}$ and $\omega^{(k)}$ in (18) by $\Sigma^{\mathcal{Q}}$ and $\omega^{(k)} = \Sigma^{\mathcal{Q}}\widehat{b}^{(k)}_{init}$, respectively. This modified estimator (with known $\Sigma^{\mathcal{Q}}$) is of a smaller variance as there is no uncertainty of estimating $\Sigma^{\mathcal{Q}}$; see Figure 5 for numerical comparisons.

**No covariate shift.** A few simplifications can be made when there is no covariate shift. For $1 \leq l \leq L$, we estimate $b^{(l)}$ by applying Lasso to the whole data set $(X^{(l)}, Y^{(l)})$: $\widehat{b}^{(l)}_{init} = \arg\min_{b\in\mathbb{R}^p} \|Y^{(l)} - X^{(l)}b\|_2^2/(2n_l) + \lambda_l \sum_{j=1}^p \|X^{(l)}_j\|_2/\sqrt{n_l}\cdot|b_j|$ with $\lambda = \sqrt{(2+c)\log p/n_l}\sigma_l$ for some constant $c > 0$. We slightly abuse the notation by using $\widehat{b}^{(l)}_{init}$ to denote the Lasso estimator based on the non-split data set. Since $\Sigma^{(l)} = \Sigma^{\mathcal{Q}}$ for $1 \leq l \leq L$, we define $\widehat{\Sigma} = \frac{1}{\sum_{l=1}^L n_l + N_{\mathcal{Q}}}\left(\sum_{l=1}^L\sum_{i=1}^{n_l} X^{(l)}_{i,\cdot}[X^{(l)}_{i,\cdot}]^{\intercal} + \sum_{i=1}^{N_{\mathcal{Q}}} X^{(l)}_{i,\cdot}[X^{(l)}_{i,\cdot}]^{\intercal}\right)$ and estimate $\Gamma_{l,k}$ as

$$\widehat{\Gamma}^{\mathcal{Q}}_{l,k} = (\widehat{b}^{(l)}_{init})^{\intercal}\widehat{\Sigma}\widehat{b}^{(k)}_{init} + (\widehat{b}^{(l)}_{init})^{\intercal}\frac{1}{n_k}[X^{(k)}]^{\intercal}(Y^{(k)} - X^{(k)}\widehat{b}^{(k)}_{init}) + (\widehat{b}^{(k)}_{init})^{\intercal}\frac{1}{n_l}[X^{(l)}]^{\intercal}(Y^{(l)} - X^{(l)}\widehat{b}^{(l)}_{init}). \tag{20}$$

This estimator can be viewed as a special case of (19): the projection directions $\widehat{u}^{(l,k)}$ and $\widehat{u}^{(k,l)}$ are set as $\widehat{b}^{(k)}_{init}$ and $\widehat{b}^{(l)}_{init}$, respectively. The optimization in (16) and (17) is not even needed for constructing the projection directions in the no covariate shift setting. Sample splitting is not needed for this setting.

## 5.2 Sampling approach for maximin effects

In the following, we propose a sampling procedure to address the irregularity issues of the maximin effects discussed in Section 4.2. For our proposed estimator $\widehat{\Gamma}^{\mathcal{Q}}$ in (19), we will show in Theorem 2 that the stacked long vector $\mathrm{vecl}(\widehat{\Gamma}^{\mathcal{Q}} - \Gamma^{\mathcal{Q}}) \in \mathbb{R}^{L(L+1)/2}$ can be approximated by a multivariate Gaussian random vector $S^* \in \mathbb{R}^{L(L+1)/2}$ with mean $\mathbf{0}$ and covariance matrix $\mathbf{V}$. For the pairs $(l_1, k_1), (l_2, k_2) \in \mathcal{I}_L$, the index mapping $\pi$ in (3) maps the matrix indexes $(l_1, k_1), (l_2, k_2) \in \mathcal{I}_L$ to the corresponding stacked vector index $\pi(l_1, k_1), \pi(l_2, k_2) \in [L(L+1)/2]$; the term $\mathbf{V}_{\pi(l_1,k_1),\pi(l_2,k_2)}$ approximates the covariance between $\widehat{\Gamma}^{\mathcal{Q}}_{l_1,k_1} - \Gamma^{\mathcal{Q}}_{l_1,k_1}$ and $\widehat{\Gamma}^{\mathcal{Q}}_{l_2,k_2} - \Gamma^{\mathcal{Q}}_{l_2,k_2}$. We estimate $\mathbf{V}_{\pi(l_1,k_1),\pi(l_2,k_2)}$ by $\widehat{\mathbf{V}}_{\pi(l_1,k_1),\pi(l_2,k_2)}$ defined as

$$
\begin{aligned}
&\frac{\widehat{\sigma}_{l_1}^2}{|B_{l_1}|}(\widehat{u}^{(l_1,k_1)})^\intercal \widehat{\Sigma}^{(l_1)} \left[\widehat{u}^{(l_2,k_2)}\mathbf{1}(l_2 = l_1) + \widehat{u}^{(k_2,l_2)}\mathbf{1}(k_2 = l_1)\right] \\
&\quad + \frac{\widehat{\sigma}_{k_1}^2}{|B_{k_1}|}(\widehat{u}^{(k_1,l_1)})^\intercal \widehat{\Sigma}^{(k_1)} \left[\widehat{u}^{(l_2,k_2)}\mathbf{1}(l_2 = k_1) + \widehat{u}^{(k_2,l_2)}\mathbf{1}(k_2 = k_1)\right] \\
&+ \frac{1}{|B|N_{\mathcal{Q}}} \sum_{i=1}^{N_{\mathcal{Q}}} \left((\widehat{b}^{(l_1)}_{init})^\intercal X^{\mathcal{Q}}_{i,\cdot}(\widehat{b}^{(k_1)}_{init})^\intercal X^{\mathcal{Q}}_{i,\cdot}(\widehat{b}^{(l_2)}_{init})^\intercal X^{\mathcal{Q}}_{i,\cdot}(\widehat{b}^{(k_2)}_{init})^\intercal X^{\mathcal{Q}}_{i,\cdot} - (\widehat{b}^{(l_1)}_{init})^\intercal \bar{\Sigma}^{\mathcal{Q}}\widehat{b}^{(k_1)}_{init}(\widehat{b}^{(l_2)}_{init})^\intercal \bar{\Sigma}^{\mathcal{Q}}\widehat{b}^{(k_2)}_{init}\right)
\end{aligned}
\tag{21}
$$

with $\bar{\Sigma}^{\mathcal{Q}} = \frac{1}{N_{\mathcal{Q}}}\sum_{i=1}^{N_{\mathcal{Q}}} X^{\mathcal{Q}}_{i,\cdot}(X^{\mathcal{Q}}_{i,\cdot})^\intercal$ and $\widehat{\sigma}_l^2 = \|Y^{(l)} - X^{(l)}\widehat{b}^{(l)}\|_2^2/n_l$ for $1 \le l \le L$.

With the estimated $\widehat{\mathbf{V}}$, we sample $S^{[m]} \in \mathbb{R}^{L(L+1)/2}$ for $1 \le m \le M$ as

$$
S^{[m]} \sim N\left(\mathbf{0}, \widehat{\mathbf{V}} + d_0/n \cdot \mathbf{I}\right), \quad \text{with} \quad d_0 = \max\left\{\max_{(l,k)\in\mathcal{I}_L}\left\{\tau \cdot n \cdot \widehat{\mathbf{V}}_{\pi(l,k),\pi(l,k)}\right\}, 1\right\} \tag{22}
$$

where $\tau > 0$ is a positive tuning parameter (default set as 0.2) and $\mathbf{I}$ is the identity matrix of the same dimension as $\widehat{\mathbf{V}}$. The rescaled diagonal entry $n \cdot \widehat{\mathbf{V}}_{\pi(l,k),\pi(l,k)}$ will be shown in Proposition 3 to be of a constant order under regularity conditions and hence $d_0$ is a constant depending on the maximum of $n \cdot \widehat{\mathbf{V}}$ and a pre-specified constant $\tau$.

The sampling distribution in (22) is slightly noisier than that with the covariance matrix $\widehat{\mathbf{V}}$ and the enlargement in the generating variance is at the scale of $1/n$. We emphasize that this slightly enlarged variance in the sampling distribution is useful for the inference for maximin effects from two different perspectives: firstly, for a nearly singular $\widehat{\mathbf{V}}$, the covariance matrix $\widehat{\mathbf{V}} + d_0/n \cdot \mathbf{I}$ is still positive-definite; secondly, for the setting of nearly null signals (i.e. $\max_{1\le l\le L}\|b^{(l)}\|_2$ being close to zero), the covariance $\widehat{\mathbf{V}}$ might not accurately quantify the uncertainty of $\widehat{\Gamma}^{\mathcal{Q}}$ due to the bias components of $\widehat{\Gamma}^{\mathcal{Q}}$. This noisier sampler provides a stochastic upper bound for the uncertainty of the proposed $\widehat{\Gamma}^{\mathcal{Q}}$ in (19).

For the sampled $S^{[m]}$ in (22), we construct the $m$-th sampled estimator of $\mathrm{vecl}(\Gamma^{\mathcal{Q}})$ as

$$\mathrm{vecl}(\widehat{\Gamma}^{[m]}) = \mathrm{vecl}(\widehat{\Gamma}^{\mathcal{Q}}) - S^{[m]} \quad \text{for } 1 \leq m \leq M. \tag{23}$$

We construct weight vector $\widehat{\gamma}_{\delta}^{[m]}$ corresponding to the $m$-th sample as

$$\widehat{\gamma}_{\delta}^{[m]} = \operatorname*{arg\,min}_{\gamma \in \Delta^L} \left[\gamma^{\intercal} \widehat{\Gamma}_{+}^{[m]} \gamma + \delta \|\gamma\|_2^2\right] \quad \text{for} \quad \delta \geq 0. \tag{24}$$

We provide some intuitions on this sampling idea. By (23), we can write

$$\mathrm{vecl}(\widehat{\Gamma}^{[m]}) \overset{d}{\approx} \mathrm{vecl}(\Gamma^{\mathcal{Q}}) + S^{*} - S^{[m]} \quad \text{for } 1 \leq m \leq M,$$

where $\overset{d}{\approx}$ indicates approximately equal in distribution. With a large enough sampling number $M$, there exists $1 \leq m^* \leq M$ such that $S^{[m^*]}$ is sufficiently close to $S^*$. As a consequence, the estimation error of the resulting weight estimator $\widehat{\gamma}_{\delta}^{[m^*]} - \gamma_{\delta}^{*}$ is almost negligible in comparison to the uncertainty arising from estimating $\{b^{(l)}\}_{1 \leq l \leq L}$.

## 5.3 CI construction for $x_{\text{new}}^{\intercal} \beta_{\delta}^{*}$

For the group $l$ with $1 \leq l \leq L$, we adopt the existing debiased estimator of $x_{\text{new}}^{\intercal} b^{(l)}$ in [11],

$$\widehat{x_{\text{new}}^{\intercal} b^{(l)}} = x_{\text{new}}^{\intercal} \widehat{b}_{init}^{(l)} + [\widehat{v}^{(l)}]^{\intercal} \frac{1}{n_l} (X^{(l)})^{\intercal} (Y^{(l)} - X^{(l)} \widehat{b}_{init}^{(l)}) \tag{25}$$

where $\widehat{b}_{init}^{(l)}$ is the Lasso estimator on the non-split data $(X^{(l)}, Y^{(l)})$ and the projection direction $\widehat{v}^{(l)} \in \mathbb{R}^p$ is defined as

$$\widehat{v}^{(l)} = \operatorname*{arg\,min}_{v \in \mathbb{R}^p} v^{\intercal} \frac{1}{n_l} (X^{(l)})^{\intercal} X^{(l)} v \quad \text{s.t.} \quad \max_{w \in \mathcal{C}(x_{\text{new}})} \left| \langle w, \frac{1}{n_l} (X^{(l)})^{\intercal} X^{(l)} v - x_{\text{new}} \rangle \right| \leq \eta_l \tag{26}$$

with $\mathcal{C}(x_{\text{new}}) = \{e_1, e_2, \cdots, e_p, x_{\text{new}} / \|x_{\text{new}}\|_2\}$ and the positive tuning parameter $\eta_l \asymp \|x_{\text{new}}\|_2 \sqrt{\log p / n_l}$. The main idea of (25) is to correct the bias of $x_{\text{new}}^{\intercal} \widehat{b}_{init}^{(l)}$ and more detailed discussion can be found in [11].

In the current paper, we focus on the inference problem for $x_{\text{new}}^{\intercal} \beta_{\delta}^{*}$ for any $x_{\text{new}} \in \mathbb{R}^p$ and $\delta \geq 0$. We aggregate $\{\widehat{x_{\text{new}}^{\intercal} b^{(l)}}\}_{1 \leq l \leq L}$ with the sampled weight vector $\widehat{\gamma}_{\delta}^{[m]}$ in (24):

$$\widehat{x_{\text{new}}^{\intercal} \beta}^{[m]} = \sum_{l=1}^{L} [\widehat{\gamma}_{\delta}^{[m]}]_l \cdot \widehat{x_{\text{new}}^{\intercal} b^{(l)}} \quad \text{for} \quad 1 \leq m \leq M. \tag{27}$$

For $1 \leq m \leq M$, we construct the sampled interval centering at $\widehat{x_{\rm new}^{\intercal}\beta}^{[m]}$,

$$\mathrm{Int}_{\alpha}^{[m]}(x_{\rm new}) = \left(\widehat{x_{\rm new}^{\intercal}\beta}^{[m]} - z_{1-\alpha/2}\widehat{\mathrm{se}}^{[m]}(x_{\rm new}), \widehat{x_{\rm new}^{\intercal}\beta}^{[m]} + z_{1-\alpha/2}\widehat{\mathrm{se}}^{[m]}(x_{\rm new})\right)$$

where $z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the standard normal distribution and

$$\widehat{\mathrm{se}}^{[m]}(x_{\rm new}) = 1.01\sqrt{\sum_{l=1}^{L}[\widehat{\gamma}_{\delta}^{[m]}]_l^2 \frac{\widehat{\sigma}_l^2}{n_l^2}[\widehat{v}^{(l)}]^{\intercal}(X^{(l)})^{\intercal}X^{(l)}\widehat{v}^{(l)}} \quad \text{with } \widehat{v}^{(l)} \text{ defined in (26)}$$

Here, the estimated standard error is slightly enlarged by a factor of 1.01 to offset the finite sample bias. Then we propose the CI for $x_{\rm new}^{\intercal}\beta_{\delta}^{*}$ as

$$\mathrm{CI}_{\alpha}\left(x_{\rm new}^{\intercal}\beta_{\delta}^{*}\right) = \cup_{m=1}^{M}\mathrm{Int}_{\alpha}^{[m]}(x_{\rm new}). \tag{28}$$

We refer to the above CI construction as Sampling Aggregation for Ridge-type maximin effects, shorthanded as SAR. A few remarks are in order for the above CI construction. Firstly, the intuition of (28) is that there exists $1 \leq m^{*} \leq M$ such that $\widehat{\gamma}_{\delta}^{[m^{*}]}$ is fairly close to the truth $\gamma_{\delta}^{*}$ and the uncertainty of the sampled point estimator $\widehat{x_{\rm new}^{\intercal}\beta}^{[m^{*}]}$ mainly comes out of $\{\widehat{x_{\rm new}^{\intercal}b^{(l)}}\}_{1\leq l\leq L}$. Secondly, although the CI in (28) is a union of $M$ intervals, its length is still well controlled, mainly due to the fact that the centers $\{\widehat{x_{\rm new}^{\intercal}\beta}^{[m]}\}_{1\leq m\leq M}$ are close to the targeted value $x_{\rm new}^{\intercal}\beta_{\delta}^{*}$ and standard errors $\widehat{\mathrm{se}}^{[m]}(x_{\rm new})$ are of rate $\|x_{\rm new}\|_2/\sqrt{n}$; see (37) in Theorem 4 and the "Efficiency Ratio" plot in Figures 1 and 6.

Thirdly, the proposed method is computationally efficient, in the sense that, the samples $S^{[m]}$ are generated after the implementation of high-dimensional optimization problems. After sampling each $S^{[m]}$, we mainly solve a $L$-dimension optimization problem in (24). This has significantly reduced the computation burden in comparison to sampling from the original data and conducting the high-dimensional optimization problems for each sample.

As a next step, we can extend the inference procedure in (28) to test the null hypothesis of $x_{\rm new}^{\intercal}\beta^{*}$ by $\phi_{\alpha} = \mathbf{1}\left(0 \notin \mathrm{CI}_{\alpha}\left(x_{\rm new}^{\intercal}\beta_{\delta}^{*}\right)\right)$. If $x_{\rm new}$ is taken as the $j$-th Euclidean basis, then this is to test the significance of the $j$-th variable.

Algorithm 1 summarizes the construction of $\mathrm{CI}_{\alpha}\left(x_{\rm new}^{\intercal}\beta_{\delta}^{*}\right)$, with the corresponding tuning parameters selection presented at the beginning of Section 7.

**Algorithm 1** Sampling Aggregation for Ridge-type maximin effects (SAR)

**Input:** Data $\{X^{(l)}, Y^{(l)}\}_{1\leq l\leq L}$, $X^{\mathcal{Q}}$; loading $x_{\rm new}\in\mathbb{R}^p$; level $\alpha\in(0,1)$; sampling size $M$; tuning parameters $\delta\geq 0$, $\tau>0$, $\mu_l>0$, $\lambda_l>0$ and $\eta_l>0$ for $1\leq l\leq L$.

**Output:** Confidence interval $\mathrm{CI}_\alpha\,(x_{\rm new}^\intercal\beta^*_\delta)$; point estimator $\widehat{x_{\rm new}^\intercal\beta^*_\delta}$.

**for** $l\leftarrow l$ to $L$ **do**
  Compute $\widehat{b}^{(l)}$ in (14) with $\lambda_l>0$ and $\widehat{\sigma}_l^2=\|Y^{(l)}-X^{(l)}\widehat{b}^{(l)}\|_2^2/n_l$;
  Compute $\widehat{v}^{(l)}$ in (26) with $\eta_l>0$ and $\widehat{x_{\rm new}^\intercal b^{(l)}}$ in (25);
**end for** ▷ Construction of Initial Estimators

**for** $(l,k)\leftarrow\mathcal{I}_L$ **do**
  Compute $\widehat{u}^{(l,k)}$ in (16),(17) with $\mu_l>0$;
  Compute $\widehat{u}^{(k,l)}$ in (16),(17) with $\mu_k>0$;
  Compute $\widehat{\Gamma}^{\mathcal{Q}}_{l,k}$ in (19);
**end for** ▷ Estimation of $\Gamma^{\mathcal{Q}}$

Compute $\widehat{\gamma}_\delta$ in (11) with $\delta\geq 0$;
Compute $\widehat{x_{\rm new}^\intercal\beta^*_\delta}=\sum_{l=1}^L[\widehat{\gamma}_\delta]_l\cdot\widehat{x_{\rm new}^\intercal b^{(l)}}$; ▷ Point estimation

**for** $(l_1,k_1),(l_2,k_2)\leftarrow\mathcal{I}_L$ **do**
  Compute $\widehat{\mathbf{V}}_{\pi(l_1,k_1),\pi(l_2,k_2)}$ in (21);
**end for** ▷ Uncertainty quantification of $\widehat{\Gamma}^{\mathcal{Q}}$

**for** $m\leftarrow 1,2,\ldots,M$ **do**
  Sample $\mathrm{vecl}(\widehat{\Gamma}^{[m]})$ in (22), (23) with $\widehat{\mathbf{V}}$ and $\tau>0$;
  Compute $\widehat{\gamma}^{[m]}_\delta$ in (24) with $\delta\geq 0$;
  Construct the interval $\mathrm{Int}^{[m]}_\alpha(x_{\rm new})$ with $\widehat{\gamma}^{[m]}_\delta$;
**end for** ▷ Sampling for Ridge-type Maximin

Construct $\mathrm{CI}_\alpha\,(x_{\rm new}^\intercal\beta^*_\delta)$ in (28). ▷ Confidence Interval Aggregation

# 6 Theoretical Justification

## 6.1 Inference properties of $\widehat{\Gamma}^{\mathcal{Q}}$

Before stating the main result, we introduce the assumptions for the model (1).

(A1) For $1\leq l\leq L$, the data $\{X^{(l)}_{i,\cdot},\epsilon^{(l)}_i\}_{1\leq i\leq n_l}$ are independently and identically generated, where $\epsilon^{(l)}_1$ is Gaussian with mean 0 and covariance $\sigma_l^2$ and independent of $X^{(l)}_{1,\cdot}$, and

$X^{(l)}_{1,\cdot} \in \mathbb{R}^p$ is sub-gaussian with $\Sigma^{(l)} = \mathbf{E}X^{(l)}_{1,\cdot}[X^{(l)}_{1,\cdot}]^\intercal$ satisfying $c_0 \leq \lambda_{\min}\left(\Sigma^{(l)}\right) \leq \lambda_{\max}\left(\Sigma^{(l)}\right) \leq C_0$ for positive constants $C_0 > c_0 > 0$. The data $\{X^{\mathcal{Q}}_{i,\cdot}\}_{1\leq i\leq N_{\mathcal{Q}}} \overset{i.i.d.}{\sim} \mathcal{Q}$ and $X^{\mathcal{Q}}_{1,\cdot}$ is sub-gaussian with $\Sigma^{\mathcal{Q}} = \mathbf{E}X^{\mathcal{Q}}_{1,\cdot}[X^{\mathcal{Q}}_{1,\cdot}]^\intercal$ satisfying $c_1 \leq \lambda_{\min}\left(\Sigma^{\mathcal{Q}}\right) \leq \lambda_{\max}\left(\Sigma^{\mathcal{Q}}\right) \leq C_1$ for positive constants $C_1 > c_1 > 0$. $L$ is finite and $\max_{1\leq l\leq L}\|b^{(l)}\|_2 \leq C$ for a positive constant $C > 0$.

(A2) Define $s = \max_{1\leq l\leq L}\|b^{(l)}\|_0$ and $n = \min_{1\leq l\leq L} n_l$. $n \asymp \max_{1\leq l\leq L} n_l$ and the model complexity parameters $(s, n, p)$ satisfy $(s\log p)^2/n \to 0$.

Assumption (A1) is commonly assumed for the theoretical analysis of high-dimensional linear models; c.f. [7]. The Gaussian error $\epsilon^{(l)}_1$ assumption can be relaxed to sub-gaussian with a more refined analysis. We impose the positive definiteness assumption on $\Sigma^{(l)}$ to guarantee the restricted eigenvalue condition [4, 52]. The model complexity condition $n \gg (s\log p)^2$ in (A2) is commonly assumed in the inference problem for a single high-dimensional linear regression [25, 42, 50] and has been shown in [9] as the minimum sample size requirement for constructing adaptive confidence intervals for single regression coefficients. The boundedness assumptions on $L$ and $\|b^{(l)}\|_2$ are mainly imposed to simplify the presentation, so is the assumption $n \asymp \max_{1\leq l\leq L} n_l$. The boundedness assumption on $\|b^{(l)}\|_2$ can be implied by a finite second order moment of the outcome variable and a positive definite $\Sigma^{(l)}$. In our analysis, we shall point out the dependence on $L$, $\|b^{(l)}\|_2$ and $n_l$ whenever possible. Define $\mathbf{V} = (\mathbf{V}_{\pi(l_1,k_1),\pi(l_2,k_2)})_{(l_1,k_1)\in\mathcal{I}_L,(l_2,k_2)\in\mathcal{I}_L} \in \mathbb{R}^{L(L+1)/2\times L(L+1)/2}$ as

$$
\begin{aligned}
\mathbf{V}_{\pi(l_1,k_1),\pi(l_2,k_2)} = & \frac{\sigma^2_{l_1}}{|B_{l_1}|}(\widehat{u}^{(l_1,k_1)})^\intercal\widehat{\Sigma}^{(l_1)}\left[\widehat{u}^{(l_2,k_2)}\mathbf{1}(l_2 = l_1) + \widehat{u}^{(k_2,l_2)}\mathbf{1}(k_2 = l_1)\right] \\
& + \frac{\sigma^2_{k_1}}{|B_{k_1}|}(\widehat{u}^{(k_1,l_1)})^\intercal\widehat{\Sigma}^{(k_1)}\left[\widehat{u}^{(l_2,k_2)}\mathbf{1}(l_2 = k_1) + \widehat{u}^{(k_2,l_2)}\mathbf{1}(k_2 = k_1)\right] \\
& \frac{1}{|B|}(\mathbf{E}[b^{(l_1)}]^\intercal X^{\mathcal{Q}}_{i,\cdot}[b^{(k_1)}]^\intercal X^{\mathcal{Q}}_{i,\cdot}[b^{(l_2)}]^\intercal X^{\mathcal{Q}}_{i,\cdot}[b^{(k_2)}]^\intercal X^{\mathcal{Q}}_{i,\cdot} - (b^{(l_1)})^\intercal\Sigma^{\mathcal{Q}}b^{(k_1)}(b^{(l_2)})^\intercal\Sigma^{\mathcal{Q}}b^{(k_2)})
\end{aligned}
\tag{29}
$$

where the index mapping $\pi$ and the matrix index set $\mathcal{I}_L$ are defined in (3).

The following theorem shows that our proposed estimator $\mathrm{vecl}(\widehat{\Gamma}^{\mathcal{Q}})$ can be approximately by a multivariate Gaussian random vector with zero mean and covariance $\mathbf{V}$.

**Theorem 2.** *Consider the model* (1). *Suppose Condition* (A1) *holds and* $\frac{s\log p}{\min\{n, N_{\mathcal{Q}}\}} \to 0$ *with* $n = \min_{1\leq l\leq L} n_l$ *and* $s = \max_{1\leq l\leq L}\|b^{(l)}\|_0$. *Then the proposed estimator* $\widehat{\Gamma}^{\mathcal{Q}} \in \mathbb{R}^{L\times L}$ *in* (19) *satisfies* $\widehat{\Gamma}^{\mathcal{Q}} - \Gamma^{\mathcal{Q}} = D + \mathrm{Rem}$ *where, as* $n, p, N_{\mathcal{Q}} \to \infty$, *there exists* $D^* \in \mathbb{R}^{L\times L}$ *and* $S^* \in \mathbb{R}^{L(L+1)/2}$ *such that* $D^* \overset{d}{=} D$, $\mathrm{vecl}(D^*) - S^* = o(N_{\mathcal{Q}}^{-2/3})$ *almost surely, and*

$$S^* \mid X^{\mathcal{Q}}_{A,\cdot}, \{X^{(l)}, \epsilon^{(l)}_{A_l}\}_{1\leq l\leq L} \sim \mathcal{N}(0, \mathbf{V})$$

*with the covariance matrix* $\mathbf{V}$ *defined in* (29)*; for* $1 \le l, k \le L$*, the reminder term* $\mathrm{Rem}_{l,k}$ *satisfies with probability larger than* $1 - \min\{n,p\}^{-c}$,

$$|\mathrm{Rem}_{l,k}| \lesssim (1 + \|\omega^{(k)}\|_2 + \|\omega^{(l)}\|_2)\frac{s \log p}{n} + (\|b^{(k)}\|_2 + \|b^{(l)}\|_2)\sqrt{\frac{s(\log p)^2}{nN_{\mathcal{Q}}}}, \tag{30}$$

*where* $c > 0$ *is a positive constant and* $\omega^{(k)}$ *and* $\omega^{(l)}$ *are defined in* (18).

The above theorem shows that $\mathrm{vecl}(\widehat{\Sigma}^{\mathcal{Q}}) - \mathrm{vecl}(\Sigma^{\mathcal{Q}})$ can be approximated in distribution by $S^*$ and the approximation error decomposes of two parts: the reminder terms $\mathrm{Rem}_{l,k}$ controlled in (30) and the distribution approximation error $\mathrm{vecl}(D^*) - S^*$ of the order $N_{\mathcal{Q}}^{-2/3}$. We control the diagonal of the covariance matrix $\mathbf{V}$ in the following proposition.

**Proposition 3.** *Suppose that the assumptions of Theorem 2 hold. Then with probability larger than* $1 - \min\{n,p\}^{-c}$*, the diagonal element* $\mathbf{V}_{\pi(l,k),\pi(l,k)}$ *in* (29) *for* $(l,k) \in \mathcal{I}_L$ *satisfies,*

$$\frac{\|\omega^{(l)}\|_2^2}{n_k} + \frac{\|\omega^{(k)}\|_2^2}{n_l} \lesssim \mathbf{V}_{\pi(l,k),\pi(l,k)} \lesssim \frac{\|\omega^{(l)}\|_2^2}{n_k} + \frac{\|\omega^{(k)}\|_2^2}{n_l} + \frac{\|b^{(l)}\|_2^2 \|b^{(k)}\|_2^2}{N_{\mathcal{Q}}} \tag{31}$$

*where* $c > 0$ *is a positive constant and* $\omega^{(l)}$ *and* $\omega^{(k)}$ *are defined in* (18)*. If* $\Sigma^{\mathcal{Q}}$ *is known, then with probability larger than* $1 - \min\{n,p\}^{-c}$*,* $n \cdot \mathbf{V}_{\pi(l,k),\pi(l,k)} \lesssim \lambda^2_{\max}(\Sigma^{\mathcal{Q}})(\|b^{(k)}\|_2^2 + \|b^{(l)}\|_2^2 + s\frac{\log p}{n})$*. If we estimate* $\Sigma^{\mathcal{Q}}$ *by* $\{X^{\mathcal{Q}}_{i,\cdot}\}_{1 \le i \le N_{\mathcal{Q}}}$*, then with probability larger than* $1 - \min\{n,p\}^{-c}$*,*

$$n \cdot \mathbf{V}_{\pi(l,k),\pi(l,k)} \lesssim \lambda^2_{\max}(\Sigma^{\mathcal{Q}}) \left(1 + \sqrt{\frac{p}{N_{\mathcal{Q}}}}\right)^4 \left(\|b^{(k)}\|_2^2 + \|b^{(l)}\|_2^2 + s\frac{\log p}{n}\right) + \frac{n}{N_{\mathcal{Q}}}\|b^{(l)}\|_2^2 \|b^{(k)}\|_2^2. \tag{32}$$

The above proposition controls the variance $\mathbf{V}_{\pi(l,k),\pi(l,k)}$ of $\widehat{\Gamma}^{\mathcal{Q}}_{l,k}$ in the general setting with covariate shift. For known $\Sigma^{\mathcal{Q}}$, we show that the diagonal elements of $\mathbf{V}$ is of order $1/n$ under the assumption (A2). If the matrix $\Sigma^{\mathcal{Q}}$ is estimated from the data, then the control of variance in (32) shows that the diagonal elements of $\mathbf{V}$ is of order $1/\sqrt{n}$ if $N_{\mathcal{Q}} \gtrsim \max\{p, n\}$ and the assumption (A2) holds. This requires a relatively large sample size $N_{\mathcal{Q}}$ of the unlabelled covariate data for the targeted population while the sizes of the labelled data $\{n_l\}_{1 \le l \le L}$ are allowed to be much smaller than $p$. The condition $N_{\mathcal{Q}} \gtrsim \max\{p, n\}$ is mainly imposed for bounding the diagonal entries of $n \cdot \mathbf{V}$ but not for establishing the asymptotic normality in Theorem 2. We have explored the dependence of the finite sample performance on $N$; see Figure 5 for more details.

Now we combine Theorem 2 and Proposition 3 to discuss the necessary conditions when the approximation errors in Theorem 2 are negligible in comparison to the multivariate normal component. Specifically, if (A2) holds and

$$N_{\mathcal{Q}} \gg \max\{s(\log p)^2, n^{3/4}, \sqrt{ns[\log \max\{N_{\mathcal{Q}}, p\}]^3}\}, \tag{33}$$

then the approximation errors satisfy $\sqrt{n}\|\text{Rem}\|_\infty/d_0 \xrightarrow{p} 0$ and $\sqrt{n}\|\text{vecl}(D^*) - S^*\|_\infty/d_0 \xrightarrow{p} 0$, where $d_0$ is defined in (22). That is, the approximation error is negligible in comparison to the sampling uncertainty in (22). As a remark, if $N_{\mathcal{Q}} \gtrsim n$, then (33) is reduced to $N_{\mathcal{Q}} \gg s[\log\max\{N_{\mathcal{Q}}, p\}]^3$.

For the setting with no covariate shift, we can establish theoretical results for the estimator in (21), which are similar to Theorem 2 and Proposition 3. We present the details in Section A.2 in the supplement.

## 6.2 Sampling accuracy

We quantify the accuracy of the sampling step. For $L > 0$ and $0 < \alpha_0 < 1/2$, define

$$C^*(L, \alpha_0) = \frac{\pi^{L(L+1)/4}}{2\sqrt{2\pi}\psi\left(L(L+1)/4 + 1\right)} \cdot \frac{\exp\left(-F^{-1}_{\chi^2_r}(1-\alpha_0)\right)}{\prod_{i=1}^{\frac{L(L+1)}{2}}\left[n \cdot \lambda_i(\mathbf{V}) + 3d_0/2\right]} \tag{34}$$

where $\psi(\cdot)$ denotes the gamma function, $1 \leq r \leq L(L+1)/2$ is the rank of $\mathbf{V}$ defined in (29) and $F^{-1}_{\chi^2_r}(1-\alpha_0)$ denotes the $1-\alpha_0$ quantile of the $\chi^2$ distribution with degree of freedom $r$. For a finite group number $L$, we have $n \cdot \lambda_i(\mathbf{V}) \lesssim n \cdot \|\mathbf{V}\|_\infty \lesssim d_0$. Under Condition (A2) and $N_{\mathcal{Q}} \gtrsim \max\{n, p\}$, we apply Proposition 3 to show that $\|n\mathbf{V}\|_\infty$ and $d_0$ are bounded with a high probability. This implies that $C^*(L, \alpha_0) \geq c$ for a positive constant $c > 0$.

**Theorem 3.** *Consider the model* (1)*. Suppose Conditions* (A1)*,* (A2) *and* (33) *hold. For* $0 < \alpha_0 < 1/2$*, define* $\text{err}_n(M) = \left[\frac{2\log n}{C^*(L,\alpha_0)M}\right]^{\frac{2}{L(L+1)}}$ *with* $C^*(L, \alpha_0)$ *in* (34)*. If* $\text{err}_n(M) \leq c\min\{1, \sqrt{n}(\lambda_{\min}(\Gamma^{\mathcal{Q}}) + \delta)\}$ *for a small positive constant* $c > 0$ *and* $\delta$ *used in the definition* (7)*, then with probability larger than* $1 - \alpha_0 - \min\{N_{\mathcal{Q}}, n, p\}^{-c_1}$ *for some* $c_1 > 0$*, there exists* $1 \leq m^* \leq M$ *such that*

$$\|\widehat{\gamma}_\delta^{[m^*]} - \gamma_\delta^*\|_2 \leq \frac{2\sqrt{2}\text{err}_n(M)}{\lambda_{\min}(\Gamma^{\mathcal{Q}}) + \delta} \cdot \frac{1}{\sqrt{n}} \tag{35}$$

The above theorem characterizes the dependence of the sampling accuracy on the sampling number $M$ and the penalty level $\delta$. We shall choose a large enough sampling number $M$ such that $\text{err}_n(M)$ converges to zero and is much smaller than $\lambda_{\min}(\Gamma^{\mathcal{Q}}) + \delta$. The upper bound in (35) guarantees that at least one sampled maximin weight $\widehat{\gamma}_\delta^{[m^*]}$ converges to the true weight $\gamma_\delta^*$ at a rate faster than $1/\sqrt{n}$. As a consequence, the uncertainty $\|\widehat{\gamma}_\delta^{[m^*]} - \gamma_\delta^*\|_2$ of constructing the maximin weight vector is negligible. The sampling accuracy for estimating $\gamma_\delta^*$ in (35) reveals that a larger value of $\delta$ reduces the uncertainty of our sampled weight estimator $\widehat{\gamma}_{\delta}^{[m^*]}$. If $\Gamma^{\mathcal{Q}}$ is (nearly) singular, we shall choose a positive penalty level $\delta > 0$ such that $\delta + \lambda_{\min}(\Gamma^{\mathcal{Q}}) \gg \mathrm{err}_n(M)$.

When $C^*(L, \alpha_0) \geq c$ for a positive constant $c > 0$ and $\lambda_{\min}(\Gamma^{\mathcal{Q}}) + \delta$ is of a constant order, we can simply choose $M \gg \log n$. In practice, we set $M = 500$ and observe reliable inference results. If our goal is to achieve the sampling accuracy $\mathrm{err}_n(M) = a_*(\lambda_{\min}(\Gamma^{\mathcal{Q}})+\delta)$ for a small positive constant $a_* > 0$, then we shall choose $M \asymp 2\log n \prod_{i=1}^{\frac{L(L+1)}{2}} \frac{n \cdot \lambda_i(\mathbf{V})+3d_0/2}{a_*(\lambda_{\min}(\Gamma^{\mathcal{Q}})+\delta)}$ that is, a larger sampling number $M$ is required for larger values of the group number $L$ and the spectrum of the covariance matrix $\mathbf{V}$.

## 6.3 Statistical inference for maximin effects

The following error decomposition reveals intuition on our proposed inference procedure,

$$\widehat{x_{\mathrm{new}}^{\intercal}\beta}^{[m]} - x_{\mathrm{new}}^{\intercal}\beta = \sum_{l=1}^{L}([\widehat{\gamma}_{\delta}^{[m]}]_l - [\gamma_{\delta}^*]_l) \cdot \widehat{x_{\mathrm{new}}^{\intercal}b^{(l)}} + \sum_{l=1}^{L}[\gamma_{\delta}^*]_l \cdot (\widehat{x_{\mathrm{new}}^{\intercal}b^{(l)}} - x_{\mathrm{new}}^{\intercal}b^{(l)}) \tag{36}$$

With $m = m^*$ in Theorem 3, we only need to quantify the uncertainty of $\sum_{l=1}^{L}[\gamma_{\delta}^*]_l \cdot (\widehat{x_{\mathrm{new}}^{\intercal}b^{(l)}} - x_{\mathrm{new}}^{\intercal}b^{(l)})$ since the uncertainty of $\sum_{l=1}^{L}([\widehat{\gamma}_{\delta}^{[m]}]_l - [\gamma_{\delta}^*]_l) \cdot \widehat{x_{\mathrm{new}}^{\intercal}b^{(l)}}$ is negligible. The following theorem establishes the properties of $\mathrm{CI}(x_{\mathrm{new}}^{\intercal}\beta_{\delta}^*)$ in (28) .

**Theorem 4.** *Consider the model* (1)*. Suppose Conditions* (A1)*,* (A2) *and* (33) *hold. If the sampling number $M$ are chosen such that* $\mathrm{err}_n(M) \ll \min\{1, \lambda_{\min}(\Gamma^{\mathcal{Q}}) + \delta\}$ *with* $\mathrm{err}_n(M)$ *defined in Theorem 2 and $\delta$ in the definition* (7)*, then* $\mathrm{CI}_{\alpha}(x_{\mathrm{new}}^{\intercal}\beta_{\delta}^*)$ *in* (28) *satisfies*

$$\lim_{n,p\to\infty} \mathbf{P}\left(x_{\mathrm{new}}^{\intercal}\beta_{\delta}^* \in \mathrm{CI}\left(x_{\mathrm{new}}^{\intercal}\beta_{\delta}^*\right)\right) \geq 1 - \alpha - \alpha_0,$$

*where $\alpha$ is the pre-specified significance level and $0 < \alpha_0 < 1/2$ is a small positive number. By further assuming* $\lambda_{\min}(\Gamma^{\mathcal{Q}}) + \delta \gg \sqrt{d_0 \log M/n}$ *with $d_0$ in* (22)*, then with probability larger than* $1 - \min\{n,p\}^{-c} - M^{-c}$*,*

$$\mathcal{L}\left(\mathrm{CI}\left(x_{\mathrm{new}}^{\intercal}\beta_{\delta}^*\right)\right) \lesssim \frac{\sqrt{d_0 \log M/n}}{\lambda_{\min}(\Gamma^{\mathcal{Q}}) + \delta} \cdot \left(\frac{\log n \|x_{\mathrm{new}}\|_2}{\sqrt{n}} + \max_{1\leq l\leq L} |x_{\mathrm{new}}^{\intercal}b^{(l)}|\right) + \frac{\|x_{\mathrm{new}}\|_2}{\sqrt{n}} \tag{37}$$

*where* $\mathcal{L}\left(\mathrm{CI}\left(x_{\mathrm{new}}^{\intercal}\beta_{\delta}^*\right)\right)$ *denotes the length of the interval and $c > 0$ is a positive constant.*

The above theorem justifies the validity of the inference procedures. The small positive value $0 < \alpha_0 < 1/2$ is the price to pay for establishing the sampling accuracy in Theorem 3, which guard against the case that $\mathrm{vecl}(\widehat{\Gamma}^{\mathcal{Q}} - \Gamma^{\mathcal{Q}})$ is near the tail of Gaussian distribution.

A major concern for the union-type CI in (28) is its conservativeness. We study the length of this union-type CI and compare it with an oracle CI with knowing the optimal weight $\gamma_{\delta}^{*}$, whose length is of order $\|x_{\text{new}}\|_2/\sqrt{n}$. Then we have

$$\frac{\mathcal{L}\left(\mathrm{CI}\left(x_{\text{new}}^{\intercal}\beta_{\delta}^{*}\right)\right)}{\|x_{\text{new}}\|_2/\sqrt{n}} \asymp 1+\sqrt{d_0 \log M}\cdot\frac{\log n/\sqrt{n}+\max_{1\leq l\leq L}|x_{\text{new}}^{\intercal}b^{(l)}|/\|x_{\text{new}}\|_2}{\lambda_{\min}(\Gamma^{\mathcal{Q}})+\delta}.$$

We further bound this by $1+\sqrt{d_0\log M}/(\lambda_{\min}(\Gamma^{\mathcal{Q}})+\delta)$. Since Proposition 3 implies that $d_0$ is of a constant order under regularity conditions, the union-type CI is enlarged by at most $\sqrt{\log M}/(\lambda_{\min}(\Gamma^{\mathcal{Q}})+\delta)$ in terms of order of magnitude, where the $\log M$ comes out of aggregating $M$ confidence intervals. There exist scenarios where the ratio $\max_{1\leq l\leq L}|x_{\text{new}}^{\intercal}b^{(l)}|/\|x_{\text{new}}\|_2$ is small. For example, if all elements of $x_{\text{new}}$ are of the same order, then $\max_{1\leq l\leq L}|x_{\text{new}}^{\intercal}b^{(l)}|/\|x_{\text{new}}\|_2 \lesssim \sqrt{s/p}$ due to the sparsity of $b^{(l)}$. Then the length of our proposed CI is of the same order as the oracle CI under the mild condition $\lambda_{\min}(\Gamma^{\mathcal{Q}})+\delta \gg \sqrt{d_0\log M}(\log n/\sqrt{n}+\sqrt{s/p})$. In Section 7, we carefully examine the finite sample length of the proposed CI and compare it with some oracle benchmark relying on asymptotic normality; see the plots of " Efficiency Ratio" in Figures 1 and 6.

# 7 Simulation Results

**Algorithm 1 Implementation.** The most important tuning parameter is the ridge penalty $\delta$ in the definition of the maximin effect in (7). We choose the value of $\delta$ through balancing reward optimality and statistical stability. Note that $\lambda_{\min}(\Gamma^{\mathcal{Q}})$ may serve as a stability indicator for the non-penalized maximin effect: if $\lambda_{\min}(\Gamma^{\mathcal{Q}})>c$ for a positive constant $c>0$, the uniquely defined maximin effect can be estimated stably. We then use $\lambda_{\min}(\widehat{\Gamma}^{\mathcal{Q}})$ as a data-dependent indicator for stability of the non-penalized maximin effect.

We illustrate how to choose $\delta$ across three simulation settings, whose details are presented in Section 7. From the leftmost to the rightmost in Figure 3, the corresponding values of $\lambda_{\min}(\widehat{\Gamma}^{\mathcal{Q}})$ are 0.10, 0.09 and 0.63. This indicates that estimation of non-penalized maximin effect is more stable for setting 3 and less stable for settings 1 and 2. In Figure 3, we plot the estimated rewards for $0\leq\delta\leq 5$, where the weight vector $\widehat{\gamma}_{\delta}$ is constructed from a single data set with $n=500$ and $N_{\mathcal{Q}}=2,000$. For settings 1 and 2, even for $\delta=2$, the estimated awards are above 95% of the estimated optimal rewards (corresponding to $\delta=0$). Hence, for settings 1 and 2, we may recommend $\delta=2$ as a good balance between reward optimality and stability. For setting 3, since the estimated reward is more sensitive to $\delta$ and $\lambda_{\min}(\widehat{\Gamma}^{\mathcal{Q}})$ is relatively large, we may recommend the non-penalized maximin effect with $\delta=0$.

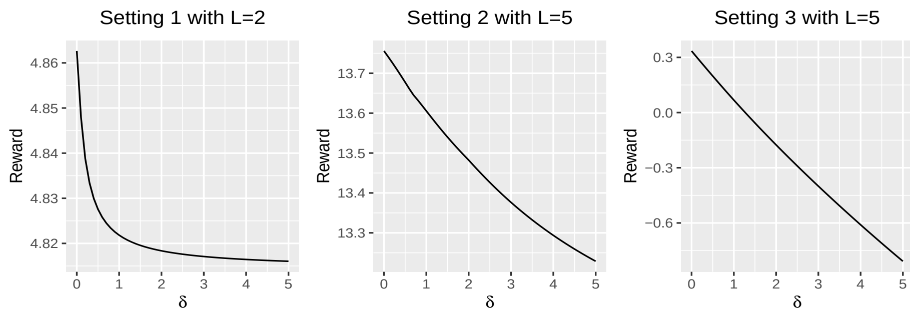



Figure 3: Plot of data dependent rewards in the simulation settings $1, 2$ and $3$ (with covariate shift). From the leftmost to the rightmost, the true optimal rewards $5.14, 13.86$ and $0.48$ (corresponding to $\delta = 0$) are estimated by $4.86, 13.75$ and $0.33$, respectively.

We set the sampling number $M = 500$ and $\tau = 0.2$ in (22). We apply existing methods to compute the initial estimators $\{\widehat{b}^{(l)}_{init}, \widehat{\sigma}^2_l, \widehat{x_{\rm new}^{\intercal} b^{(l)}}\}_{1\leq l\leq L}$: the Lasso estimator $\widehat{b}^{(l)}_{init}$ is implemented by the R-package `glmnet` [20] with the tuning parameters $\{\lambda_l\}_{1\leq l\leq L}$ chosen by cross-validation; the noise level is estimated by $\widehat{\sigma}^2_l = \|Y^{(l)} - X^{(l)}\widehat{b}^{(l)}_{init}\|_2^2/n_l$; the debiased estimator $\widehat{x_{\rm new}^{\intercal} b^{(l)}}$ is implemented using the `FIHR` package with its built-in tuning parameter selection. [1] For the construction of $\widehat{u}^{(l,k)}$ in (16) and (17), we solve its dual problem

$$\widehat{h} = \arg\min_{h\in\mathbb{R}^{p+1}} \frac{1}{4} h^{\intercal} H^{\intercal} \widehat{\Sigma}^{(l)} H h + (\omega^{(k)})^{\intercal} H h / \|\omega^{(k)}\|_2 + \lambda \|h\|_1 \text{ with } H = \left[\frac{\omega^{(k)}}{\|\omega^{(k)}\|_2}, \mathbf{I}_{p\times p}\right],$$

where we adopt the notation $0/0 = 0$. This dual problem is unbounded from below when $H^{\intercal}\widehat{\Sigma}^{(l)} H$ is singular and $\lambda$ is close to zero. We choose the smallest $\lambda > 0$ such that the dual problem is bounded from below. We then construct the direction $\widehat{u}^{(l,k)}$ as $\widehat{u}^{(l,k)} = -\frac{1}{2}(\widehat{h}_{-1} + \widehat{h}_1 \omega^{(k)}/\|\omega^{(k)}\|_2)$. Recall that sample splitting is used for our proposed estimator $\widehat{\Gamma}^{\mathcal{Q}}$ in the covariate-shift setting. In numerical studies, we implement this proposed estimator using the full sample and observe reliable inference properties in the simulation studies.

Throughout the simulation, we generate the data $\{X^{(l)}, Y^{(l)}\}_{1\leq l\leq L}$ following (1), where, for the $l$-th group, the $p$-dimensional covariates $\{X^{(l)}_{i,\cdot}\}_{1\leq i\leq n_l} \overset{\text{i.i.d.}}{\sim} N(\mathbf{0}, \Sigma^{(l)})$ and the errors $\{\epsilon^{(l)}_i\}_{1\leq i\leq n_l} \overset{\text{i.i.d.}}{\sim} N(0, \sigma_l)$. The dimension $p$ is set as 500. In the covariate shift setting, we assume that we either have access to $N_{\mathcal{Q}}$ i.i.d. samples $\{X^{\mathcal{Q}}_{i,\cdot}\}_{1\leq i\leq N_{\mathcal{Q}}} \overset{\text{i.i.d.}}{\sim} N(\mathbf{0}, \Sigma^{\mathcal{Q}})$ or we know $\Sigma^{\mathcal{Q}}$ a priori. For $1 \leq l \leq L$, we take $n_l = n$, $\sigma_l = 1$ and $\Sigma^{(l)} = \Sigma$, where $\Sigma_{j,k} = 0.6^{|j-k|}$ for $1 \leq j, k \leq p$; we conduct 500 simulations and report the average measures.

**Dependence on $n$ and $\delta > 0$.** We consider the following covariate shift settings with targeted sample size $N_{\mathcal{Q}} = 2,000$.

[1] The code of `FIHR` is available at `https://statistics.rutgers.edu/home/zijguo/Software.html`.

Setting 1 (covariate shift with $L = 2$): $b_j^{(1)} = j/40$ for $1 \leq j \leq 10$, $b_j^{(1)} = 1$ for $j = 22, 23$, $b_j^{(1)} = 0.5$ for $j = 498$, $b_j^{(1)} = -0.5$ for $j = 499, 500$ and $b_j^{(1)} = 0$ otherwise; $b_j^{(2)} = b_j^{(1)}$ for $1 \leq j \leq 10, j = 22, 23$, $b_j^{(2)} = 1$ for $j = 500$ and $b_j^{(2)} = 0$ otherwise; $\Sigma_{i,i}^{\mathcal{Q}} = 1.5$ for $1 \leq i \leq 500$, $\Sigma_{i,j}^{\mathcal{Q}} = 0.9$ for $1 \leq i \neq j \leq 5$, $\Sigma_{i,j}^{\mathcal{Q}} = 0.9$ for $499 \leq i \neq j \leq 500$ and $\Sigma_{i,j}^{\mathcal{Q}} = \Sigma_{i,j}$ otherwise; $[x_{\text{new}}]_j = 1$ for $498 \leq j \leq 500$ and $[x_{\text{new}}]_j = 0$ otherwise.

Setting 2 (covariate shift with $L \geq 2$): $b_j^{(1)} = j/10$ for $1 \leq j \leq 10$, $b_j^{(1)} = (10 - j)/10$ for $11 \leq j \leq 20$, $b_j^{(1)} = 0.2$ for $j = 21$, $b_j^{(1)} = 1$ for $j = 22, 23$ and $b_j^{(1)} = 0$ for $24 \leq j \leq 500$; For $2 \leq l \leq L$, $b_j^{(l)} = b_j^{(1)} + 0.1 \cdot (l-1)/\sqrt{300}$ for $1 \leq j \leq 10$, $b_j^{(2)} = -0.3 \cdot (l-1)/\sqrt{300}$ for $11 \leq j \leq 20$, $b_j^{(l)} = 0.5 \cdot (l-1)$ for $j = 21$, $b_j^{(l)} = 0.2 \cdot (j-1)$ for $j = 22, 23$ and $b_j^{(2)} = 0$ for $24 \leq j \leq 500$; $\Sigma_{i,i}^{\mathcal{Q}} = 1.1$ for $1 \leq i \leq 500$, $\Sigma_{i,j}^{\mathcal{Q}} = 0.75$ for $1 \leq i \neq j \leq 6$ and $\Sigma_{i,j}^{\mathcal{Q}} = \Sigma_{i,j}$ otherwise; $[x_{\text{new}}]_j = 1$ for $21 \leq j \leq 23$ and $[x_{\text{new}}]_j = 0$ otherwise.

Setting 3 (covariate shift with $L \geq 2$)): $b^{(1)}$, $b^{(2)}$ and $\Sigma^{\mathcal{Q}}$ are the same as Setting 2; for $l \geq 3$, $\{b_j^{(l)}\}_{1 \leq j \leq 6}$ are independently generated following standard normal and $b_j^{(l)} = 0$ for $7 \leq j \leq 500$; $x_{\text{new}} \sim \mathcal{N}(\mathbf{0}, \Sigma^{\text{new}})$ with $\Sigma_{i,j}^{\text{new}} = 0.5^{1+|i-j|}/25$ for $1 \leq i, j \leq 500$.

For setting 1, the Root Mean Square Error (RMSE) of the point estimator $\widehat{x_{\text{new}}^{\intercal}\beta_{\delta}^{*}}$ in Algorithm 1 is reported in Table 1. The RMSE decreases with a larger sample size or a larger $\delta$ value. The reward plot in Figure 3 suggests $\delta = 2$ as a reasonable choice. For $n = 300, 500$, the RMSE for $\delta = 2$ is almost half of that for $\delta = 0$.

| n | $\delta = 0$ | $\delta = 0.1$ | $\delta = 0.5$ | $\delta = 1$ | $\delta = 2$ | $\delta = 3$ | $\delta = 4$ | $\delta = 5$ |
|---|---|---|---|---|---|---|---|---|
| 100 | 0.158 | 0.153 | 0.139 | 0.132 | 0.128 | 0.127 | 0.127 | 0.126 |
| 200 | 0.121 | 0.114 | 0.095 | 0.088 | 0.085 | 0.084 | 0.084 | 0.083 |
| 300 | 0.116 | 0.104 | 0.079 | 0.070 | 0.067 | 0.066 | 0.066 | 0.065 |
| 500 | 0.106 | 0.091 | 0.063 | 0.057 | 0.055 | 0.055 | 0.055 | 0.054 |

Table 1: Root Mean Square Error of $\widehat{x_{\text{new}}^{\intercal}\beta_{\delta}^{*}}$ for Setting 1

The inference results for setting 1 are reported in Figure 4. We plot the empirical coverage and length of our proposed CIs over $\delta \in \{0, 0.1, 0.5, 1, 2, 3, 4, 5\}$, with corresponding $x_{\text{new}}^{\intercal}\beta_{\delta}^{*}$ as $\{0.1, 0.164, 0.192, 0.196, 0.198, 0.1986, 0.1989, 0.199\}$. In Figure 4, for $n = 300, 500$, with increasing $\delta$ from 0 to 2 (the recommended value), the empirical coverage levels drop from 100% to the desired 95%; the CI lengths are reduced by around 50%. For $\delta = 2$, the empirical coverage is around 85% for $n = 100$ and 92.5% for $n = 200$.

We introduce a measure named "efficiency ratio" to study the conservativeness of the constructed union-type CI. The oracle CI length (relying on asymptotic normality) is calculated as multiplying the estimated standard error of the proposed maximin estimator

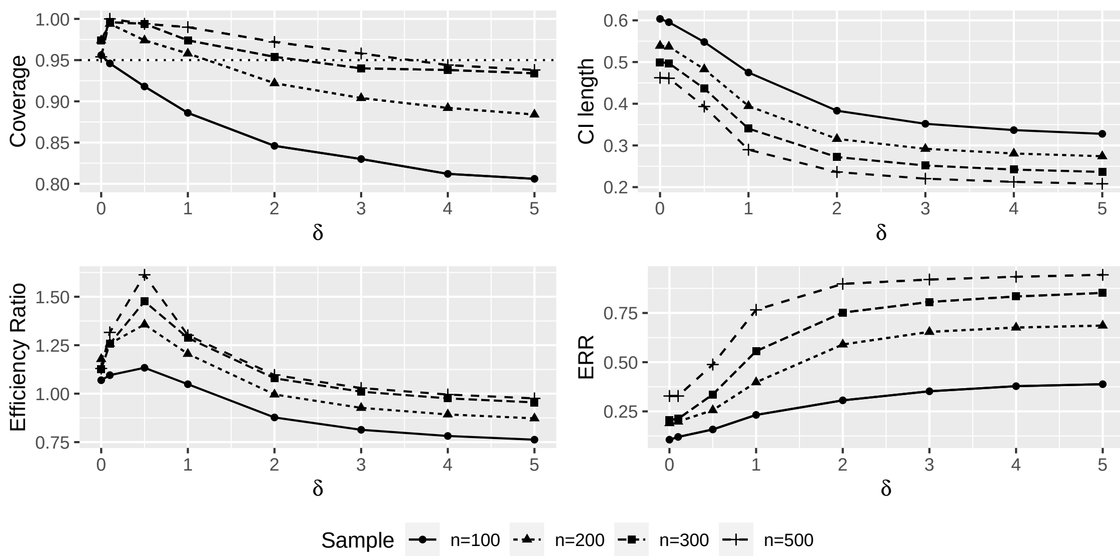


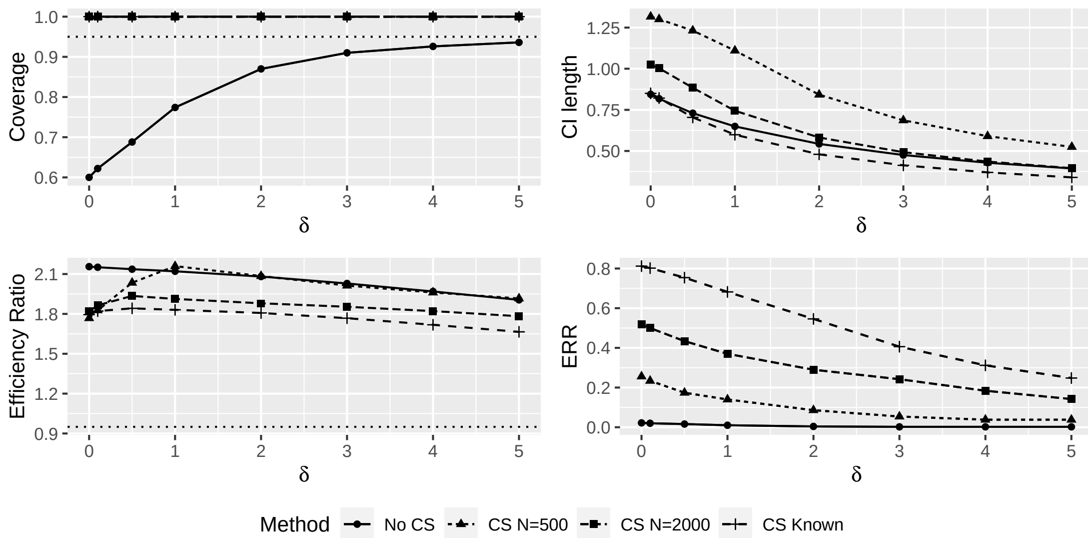

Figure 4: Dependence on $\delta$ and $n$: setting 1 (covariate shift) with unknown $\Sigma^{\mathcal{Q}}$. "Coverage" and "CI Length" stand for the empirical coverage and the average length of our proposed CI; "Efficiency Ratio" represents the ratio of the length of CI in (28) to an oracle CI relying on normality; "ERR" represents the empirical rejection rate out of 500 simulations.

$\widehat{x_{\text{new}}^{\intercal}\beta_{\delta}^{*}}$ by a factor of $2 \cdot 1.96$. Since the estimated standard error is computed over 500 simulations, this oracle CI can only be computed in oracle settings. We define the "efficiency ratio" as the ratio of the length of our proposed CI over that of this oracle CI length. Since the limiting distribution of $\widehat{x_{\text{new}}^{\intercal}\beta^{*}}$ is not necessarily normal, this oracle CI based on asymptotic normality might not even be valid; see Figure 6 for an example. This proposed "efficiency ratio" is mainly used to give some hints about the efficiency of the proposed CI. In Figure 4, the efficiency ratios are below 1.5 in most cases; for $\delta = 2$, the ratio is below 1 for $n = 100$ (corresponding to under-coverage) but is slightly above 1 for $n = 200, 300, 500$.

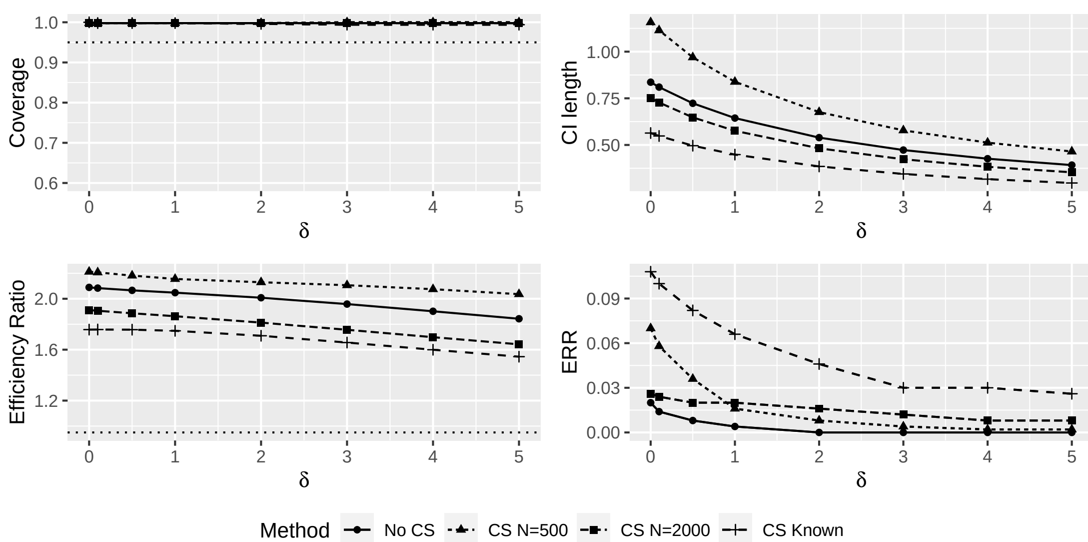
We report the empirical rejection rate (ERR) of the test $\phi_{\alpha} = \mathbf{1}\left(0 \notin \mathrm{CI}_{\alpha}\left(x_{\text{new}}^{\intercal}\beta_{\delta}^{*}\right)\right)$, where ERR is defined as the proportion of null hypothesis $x_{\text{new}}^{\intercal}\beta_{\delta}^{*} = 0$ being rejected out of the 500 simulations. Under the null hypothesis, ERR is an empirical measure of the type I error; under the alternative hypothesis, ERR is an empirical measure of the power. In Figure 4, ERR, as a measure of the empirical power, increases with a larger $n$ and $\delta$.

The results for settings 2 and 3 are similar to those for setting 1 and are presented in Section D.2 in the supplement. These settings illustrate that our proposed method works for a general finite group number $L \geq 2$.

**Comparison of algorithms with or without covariate shift.** We consider covariate




(a) Simulation settings with covariate shift




(b) Simulation settings with no covariate shift

Figure 5: Comparison of covariate shift and no covariate shift algorithms with $n = 500$. The methods "No CS", "CS N=500", "CS N=2000", "CS Known" represent algorithms assuming no covariate shift, Algorithm 1 (covariate shift) with $N_{\mathcal{Q}} = 500$, with $N_{\mathcal{Q}} = 2000$ and known $\Sigma^{\mathcal{Q}}$, respectively. "Coverage" and "CI Length" stand for the empirical coverage and the average length of our proposed CI; "Efficiency Ratio" represents the ratio of the length of CI in (28) to an oracle CI relying on normality; "ERR" represents the empirical rejection rate out of 500 simulations.

shift algorithm (Algorithm 1) with $N_{\mathcal{Q}} = 500, 2,000$ or known $\Sigma^{\mathcal{Q}}$ and compare it with the proposed algorithm assuming no covariate shift, which replaces the general estimator $\widehat{\Gamma}^{\mathcal{Q}}$ in Algorithm 1 by (21). We set $L = 2$; generate $b^{(1)} \in \mathbb{R}^p$ as $b_j^{(1)} = j/40$ for $1 \leq j \leq 10$, $b_j^{(1)} = 1$ for $j = 22, 23$ and $b_j^{(1)} = 0.5$ for $j = 498$, $b_j^{(1)} = -0.5$ for $j = 499, 500$ and $b_j^{(1)} = 0$ otherwise; generate $b^{(2)} \in \mathbb{R}^p$ as $b_j^{(2)} = b_j^{(1)}$ for $1 \leq j \leq 10, j = 22, 23$, $b_j^{(2)} = 1$ for $j = 500$ and $b_j^{(2)} = 0$ otherwise. Set the loading $[x_{\rm new}]_j = 1$ for $j = 499, 500$ and $[x_{\rm new}]_j = 0$ otherwise. For the setting with covariate shift, we generate $\Sigma^{\mathcal{Q}}$ as $\Sigma^{\mathcal{Q}}_{i,i} = 1.5$ for $1 \leq i \leq 500$, $\Sigma^{\mathcal{Q}}_{i,j} = 0.6$ for $1 \leq i \neq j \leq 5$, $\Sigma^{\mathcal{Q}}_{i,j} = -0.9$ for $499 \leq i \neq j \leq 500$ and $\Sigma^{\mathcal{Q}}_{i,j} = \Sigma_{i,j}$ otherwise.

The results for $n = 500$ are reported in Figure 5. The top of Figure 5 corresponds to the simulation settings with covariate shift. The no covariate shift algorithm does not achieve the 95% coverage due to the bias of assuming no covariate shift. In contrast, all three covariate shift algorithms achieve the 95% coverage level and the CI lengths decrease with an increasing $N_{\mathcal{Q}}$, where the setting of known $\Sigma^{\mathcal{Q}}$ can be viewed as $N_{\mathcal{Q}} = \infty$. Correspondingly, the empirical power, measured by ERR, increases with a larger $N_{\mathcal{Q}}$. The bottom of Figure 5 corresponds to the simulation settings with no covariate shift. All algorithm achieves the desired coverage levels. Since the no covariate shift algorithm can be viewed as the setting with $N_{\mathcal{Q}} = 1,000$, the corresponding lengths of the constructed CIs decrease with a larger $N_{\mathcal{Q}}$. The results for sample sizes $\{100, 200, 300\}$ are similar to those for $n = 500$ and are reported in Section D.3 in the supplement.

**Irregularity of maximin effects.** We consider the settings where the distribution of $\widehat{\gamma}$ is likely to be a mixture distribution, as in (13). We set $L = 2$; set the loading as $[x_{\rm new}]_j = j/5$ for $1 \leq j \leq 5$ and $[x_{\rm new}]_j = 0$ for $6 \leq j \leq 500$. generate $b^{(1)} \in \mathbb{R}^p$ as $b_j^{(1)} = j/40$ for $1 \leq j \leq 10$, $b_j^{(1)} = (10 - j)/40$ for $11 \leq j \leq 20$, $b_{21}^{(1)} = 0.2, b_{22}^{(1)} = b_{23}^{(1)} = 1$ and $b_j^{(1)} = 0$ for $24 \leq j \leq 500$ and generate $b^{(2)} \in \mathbb{R}^p$ as $b_j^{(2)} = b_j^{(1)} + \text{perb}/\sqrt{300}$ for $1 \leq j \leq 10$, $b_j^{(2)} = 0$ for $11 \leq j \leq 20$ and $b_{21}^{(2)} = 0.5, b_{22}^{(2)} = b_{23}^{(2)} = 0.2$ and $b_j^{(2)} = 0$ for $24 \leq j \leq 500$. Here, settings 1 to 10 correspond to the values of perb taken as $\{1, 1.125, 1.25, 1.5, 3.75, 4, 5, 7, 10, 12\}$. As reported in Figure 6, the constructed CI achieves the desired coverage level when the sample size reaches 300 and the CI length decreases with a larger sample size. An interesting observation is that, for perturbation settings 9 and 10, although the efficiency ratio is around 1.5, the corresponding coverage levels are just achieving the desired 95% level, which indicates the oracle CI would be under-coverage.

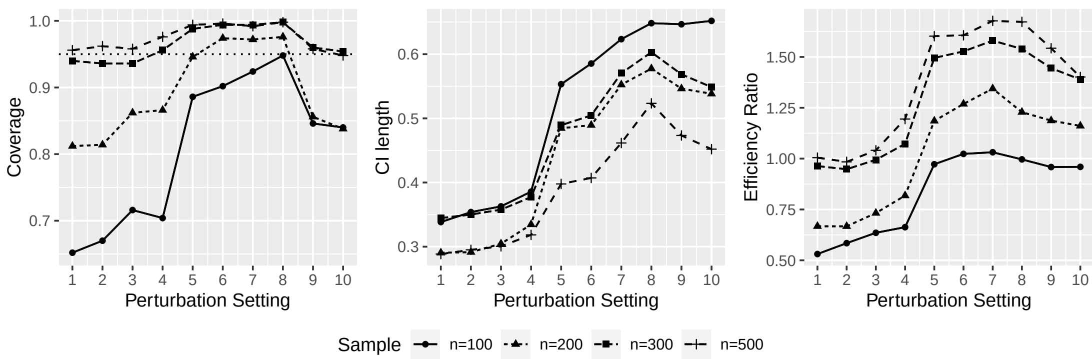



Figure 6: Inference for the maximin effects: ten settings 1 to 10 correspond to perb being taken as $\{1, 1.125, 1.25, 1.5, 3.75, 4, 5, 7, 10, 12\}$. "Coverage" and "CI Length" stand for the empirical coverage and the average length of our proposed CI; "Efficiency Ratio" represents the ratio of the length of CI in (28) to an oracle CI relying on normality.

# 8 Real Data Applications

We apply our proposed method to a genome-wide association study [5], which studies the yeast colony growth under 46 different growth media. The study is based on $n = 1,008$ Saccharomyces cerevisiae segregants crossbred from a laboratory strain and a wine strain. A set of $p = 4,410$ genetic markers has selected out of the total 11,623 genetic markers [5]. The outcome variables of interest are the end-point colony sizes under different growth media. To demonstrate our method, we consider the colony sizes under five growth media (i.e. $L = 5$): "Ethanol", "Lactate", "Lactose", "Sorbitol" and "Trehalose". Our model (1) can be applied here with $L = 5$ and each $1 \leq l \leq 5$ corresponds to one growth media (environment). The goal is to find a vector $\beta^*(\mathcal{Q}) \in \mathbb{R}^p$ as a best representation of the heterogeneous regression vectors under different growth media.

All of these outcome variables are normalized to have variance 1, with corresponding variance explained by the genetic markers $0.60, 0.69, 0.68, 0.51$ and $0.66$. We study the inference for the ridge-type maximin effect for no covariate shift setting and two settings with covariate shift. For covariate shift setting 1, we set $\Sigma^{\mathcal{Q}} = \mathrm{I}$; for setting 2, we set $\Sigma^{\mathcal{Q}}_{j,j} = 1$ for $1 \leq j \leq p$ and $\Sigma^{\mathcal{Q}}_{j,l} = 0.75$ for $j \neq l$ and $j, l \in \mathcal{S}_0$, where $\mathcal{S}_0$ is a set of 14 important genetic markers. We report the estimated reward (as a minimum of the explained variance across $L = 5$ groups) in Table 2, where $\widehat{R}$ denotes replacing the expectation in (5) with its sample average. With a larger ridge penalty level $\delta$, the estimated reward decreases by 10% to 15% when $\delta$ is increased from 0 to 0.5. Since $\lambda_{\min}(\widehat{\Gamma}^{\mathcal{Q}})$ is used as a stability indicator for the estimator $\widehat{\beta^*_{\delta=0}}$, the results in Table 2 suggest the use of some positive $\delta$

values, e.g. $\delta = 0.5$.

| | $\widehat{R}(\widehat{\beta^*_{\delta=0}})$ | $\widehat{R}(\widehat{\beta^*_{\delta=0.1}})$ | $\widehat{R}(\widehat{\beta^*_{\delta=0.5}})$ | $\widehat{R}(\widehat{\beta^*_{\delta=1}})$ | $\lambda_{\min}(\widehat{\Gamma}^{\mathcal{Q}})$ |
|---:|---:|---:|---:|---:|---:|
| No Covariate Shift | 0.470 | 0.462 | 0.421 | 0.398 | 0.032 |
| Covariate Shift Setting 1 | 0.359 | 0.346 | 0.302 | 0.276 | 0.082 |
| Covariate Shift Setting 2 | 0.186 | 0.173 | 0.164 | 0.143 | 0.110 |

Table 2: Real data: dependence of estimated rewards on $\delta$.

In Figure 7, we plot the constructed CI for $[\beta^*_\delta(\mathcal{Q})]_j$ for $j \in \mathcal{S}$, where $\mathcal{S} \subset \mathcal{S}_0$ is a set of ten pre-selected regression indexes. We vary $\delta$ across $\{0, 0.1, 0.5, 1\}$ and $\mathcal{Q}$ across {no covariate shift, covariate shift setting 1 and 2}. We observe that the constructed CIs get shorter with a larger $\delta$. Additionally, for different targeted distribution $\mathcal{Q}$, the inference results vary but we can observe a consistent patten across three different targeted covariate distributions. In particular, the variable corresponding to index 2 is most significant and the corresponding CIs are below zero across different $\delta$; the variable with index 3 becomes significant for $\delta \geq 0.1$; the variables with indexes $\{1, 9, 10\}$ are significant until $\delta$ reaches 1 while all remaining variables are not significant even for $\delta = 1$.

# 9 Conclusion and Discussions

We introduce the covariate shift and ridge-type maximin effects as effective summaries of heterogeneous high-dimensional regression vectors. The sampling approach is effective for quantifying the uncertainty of the constructed aggregation weights of the maximin effect. Beyond linear models in (1), an interesting research direction is the model aggregation of more complex models in high dimensions, for example, the sparse additive models [28, 32]. Another important question is on model aggregation when the linear models in (1) are possibly misspecified [8, 46]. Both questions are left for future research.

# Supplement

We present additional methods and theories, proofs and additional simulation results in the supplement.

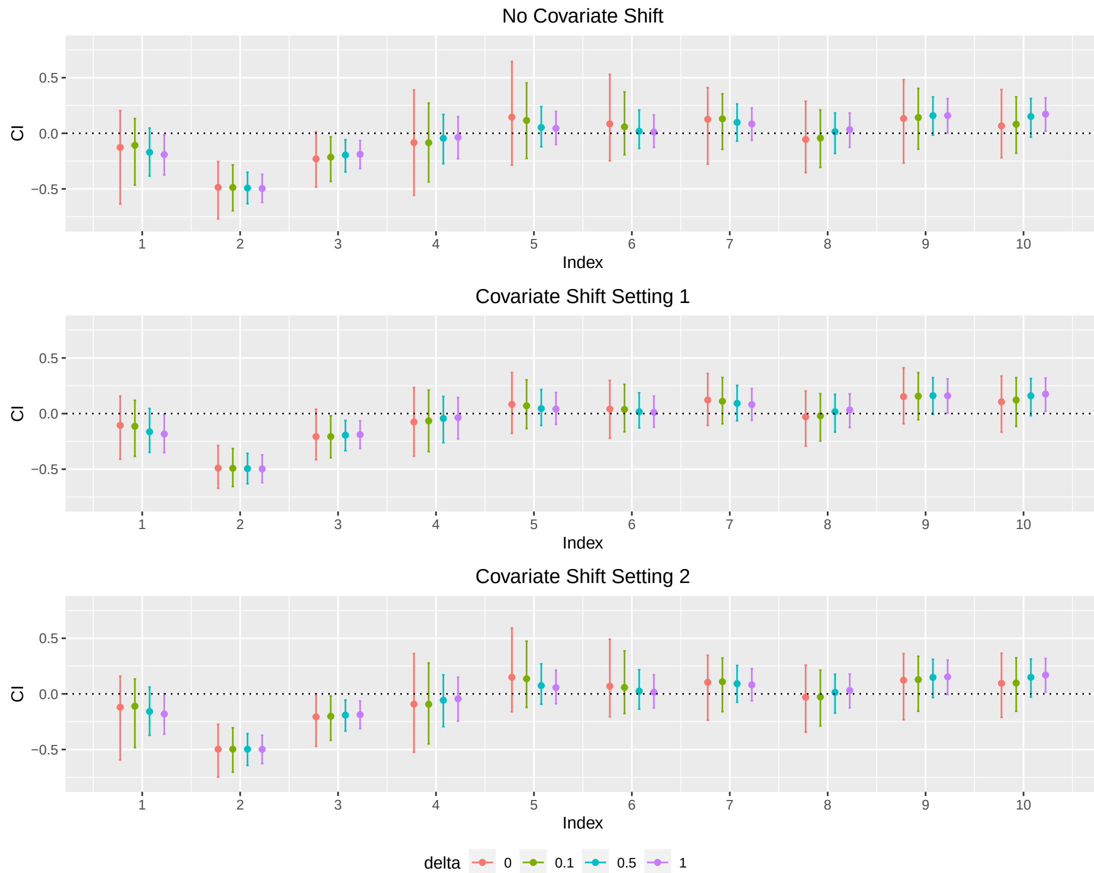



Figure 7: Plot of CI for $[\beta^*_\delta(\mathcal{Q})]_j$ for $j \in \mathcal{S}$, where $\mathcal{S}$ is a set of ten genetic markers. Here, $\delta$ is varied across $\{0, 0.1, 0.5, 1\}$ and $\mathcal{Q}$ is varied across {no covariate shift, covariate shift setting 1 and 2}.

# References


[1] Susan Athey, Guido W Imbens, and Stefan Wager. Approximate residual balancing: debiased inference of average treatment effects in high dimensions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(4):597–623, 2018.

[2] Alexandre Belloni, Victor Chernozhukov, and Christian Hansen. High-dimensional methods and inference on structural and treatment effects. *Journal of Economic Perspectives*, 28(2):29–50, 2014.

[3] Alexandre Belloni, Victor Chernozhukov, and Lie Wang. Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika*, 98(4):791–806, 2011.

[4] Peter J Bickel, Yaacov Ritov, and Alexandre B Tsybakov. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

[5] Joshua S Bloom, Ian M Ehrenreich, Wesley T Loo, Thúy-Lan Võ Lite, and Leonid Kruglyak. Finding the sources of missing heritability in a yeast cross. *Nature*, 494(7436):234–237, 2013.

[6] Peter Bühlmann. Invariance, causality and robustness. *arXiv preprint arXiv:1812.08233*, 2018.

[7] Peter Bühlmann and Sara van de Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.

[8] Peter Bühlmann and Sara van de Geer. High-dimensional inference in misspecified linear models. *Electronic Journal of Statistics*, 9(1):1449–1473, 2015.

[9] T Tony Cai and Zijian Guo. Confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity. *The Annals of Statistics*, 45(2):615–646, 2017.

[10] T Tony Cai and Zijian Guo. Semisupervised inference for explained variance in high dimensional linear regression and its applications. *Journal of the Royal Statistical Society Series B*, 82(2):391–419, 2020.

[11] Tianxi Cai, Tony Cai, and Zijian Guo. Optimal statistical inference for individualized treatment effects in high-dimensional models. *arXiv preprint arXiv:1904.12891*, 2019.

[12] Tony Cai and Weidong Liu. Adaptive thresholding for sparse covariance matrix estimation. *Journal of the American Statistical Association*, 106(494):672–684, 2011.

[13] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters: Double/debiased machine learning. *The Econometrics Journal*, 21(1):C1–C68, 2018.

[14] Victor Chernozhukov, Christian Hansen, and Martin Spindler. Valid post-selection and post-regularization inference: An elementary, general approach. *Annu. Rev. Econ.*, 7(1):649–688, 2015.

[15] Ellen Deusser. Heterogeneity of ribosomal populations in escherichia coli cells grown in different media. *Molecular and General Genetics MGG*, 119(3):249–258, 1972.

[16] B Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26, 1979.

[17] Bradley Efron and Robert J Tibshirani. *An introduction to the bootstrap*. CRC press, 1994.

[18] Uwe Einmahl. Extensions of results of komlós, major, and tusnády to the multivariate case. *Journal of multivariate analysis*, 28(1):20–68, 1989.

[19] Max H Farrell. Robust inference on average treatment effects with possibly more covariates than observations. *Journal of Econometrics*, 189(1):1–23, 2015.

[20] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1, 2010.

[21] Rui Gao, Xi Chen, and Anton J Kleywegt. Wasserstein distributional robustness and regularization in statistical learning. *arXiv preprint arXiv:1712.06050*, 2017.

[22] Zijian Guo, Claude Renaux, Peter Bühlmann, and T Tony Cai. Group inference in high dimensions with applications to hierarchical testing. *arXiv preprint arXiv:1909.01503*, 2019.

[23] Zijian Guo, Wanjie Wang, T Tony Cai, and Hongzhe Li. Optimal estimation of genetic relatedness in high-dimensional linear models. *Journal of the American Statistical Association*, 114(525):358–369, 2019.

[24] J. Hannig, H. Iyer, R. C. S. Lai, and T. C. M. Lee. Generalized fiducial inference: A review and new results. *Journal of American Statistical Association*, 2016. To appear. Accepted in March 2016. doidoi:10.1080/01621459.2016.1165102.

[25] Adel Javanmard and Andrea Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research*, 15(1):2869–2909, 2014.

[26] Molei Liu, Yin Xia, Tianxi Cai, and Kelly Cho. Integrative high dimensional multiple testing with heterogeneity under data sharing constraints. *arXiv preprint arXiv:2004.00816*, 2020.

[27] Gitta H Lubke and Bengt Muthén. Investigating population heterogeneity with factor mixture models. *Psychological methods*, 10(1):21, 2005.

[28] Lukas Meier, Sara van de Geer, and Peter Bühlmann. High-dimensional additive modeling. *The Annals of Statistics*, 37(6B):3779–3821, 2009.

[29] Nicolai Meinshausen and Peter Bühlmann. Maximin effects in inhomogeneous large-scale data. *The Annals of Statistics*, 43(4):1801–1830, 2015.

[30] Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference using invariant prediction: identification and confidence intervals. *arXiv preprint arXiv:1501.01332*, 2015.

[31] Dimitris N Politis, Joseph P Romano, and Michael Wolf. *Subsampling*. Springer Science & Business Media, 1999.

[32] Pradeep Ravikumar, John Lafferty, Han Liu, and Larry Wasserman. Sparse additive models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(5):1009–1030, 2009.

[33] Dominik Rothenhäusler, Nicolai Meinshausen, and Peter Bühlmann. Confidence intervals for maximin effects in inhomogeneous large-scale data. In *Statistical Analysis for High-Dimensional Data*, pages 255–277. Springer, 2016.

[34] Dominik Rothenhäusler, Nicolai Meinshausen, Peter Bühlmann, and Jonas Peters. Anchor regression: heterogeneous data meets causality. *arXiv preprint arXiv:1801.06229*, 2018.

[35] Chengchun Shi, Rui Song, Wenbin Lu, and Bo Fu. Maximin projection learning for optimal treatment decision with heterogeneous individualized treatment effects. *Journal of the Royal Statistical Society. Series B, Statistical methodology*, 80(4):681, 2018.

[36] Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244, 2000.

[37] Aman Sinha, Hongseok Namkoong, Riccardo Volpi, and John Duchi. Certifying some distributional robustness with principled adversarial training. *arXiv preprint arXiv:1710.10571*, 2017.

[38] Masashi Sugiyama, Matthias Krauledat, and Klaus-Robert MÃžller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8(May):985–1005, 2007.

[39] Tingni Sun and Cun-Hui Zhang. Scaled sparse linear regression. *Biometrika*, 101(2):269–284, 2012.

[40] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

[41] Yuta Tsuboi, Hisashi Kashima, Shohei Hido, Steffen Bickel, and Masashi Sugiyama. Direct density ratio estimation for large-scale covariate shift adaptation. *Journal of Information Processing*, 17:138–155, 2009.

[42] Sara van de Geer, Peter Bühlmann, Yaacov Ritov, and Ruben Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202, 2014.

[43] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. In Yonina Eldar and Gitta Kutyniok, editors, *Compressed Sensing: Theory and Applications*, pages 210–268. Cambridge University Press, 2012.

[44] Nicolas Verzelen and Elisabeth Gassiat. Adaptive estimation of high-dimensional signal-to-noise ratios. *Bernoulli*, 24(4B):3683–3710, 2018.

[45] Peng Wang and Minge Xie. Repro sampling method for statistical inference of high dimensional linear models. *Research Manuscript*.

[46] Larry Wasserman. Discussion:" a significance test for the lasso". *The Annals of Statistics*, 42(2):501–508, 2014.

[47] M. Xie and K. Singh. Confidence distribution, the frequentist distribution estimator of a parameter: a review. *International Statistical Review*, 81:3–39, 2013.

[48] Fei Ye and Cun-Hui Zhang. Rate minimaxity of the lasso and dantzig selector for the lq loss in lr balls. *The Journal of Machine Learning Research*, 11:3519–3540, 2010.

[49] Sandy L Zabell. Ra fisher and fiducial argument. *Statistical Science*, 7(3):369–387, 1992.

[50] Cun-Hui Zhang and Stephanie S Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):217–242, 2014.

[51] Tianqi Zhao, Guang Cheng, and Han Liu. A partially linear framework for massive heterogeneous data. *Annals of statistics*, 44(4):1400, 2016.

[52] Shuheng Zhou. Restricted eigenvalue conditions on subgaussian random matrices. *arXiv preprint arXiv:0912.4045*, 2009.

[53] Yinchu Zhu and Jelena Bradic. Linear hypothesis testing in dense high-dimensional linear models. *Journal of the American Statistical Association*, pages 1–18, 2018.

# A Additional Methods and Theories

In Section A.1, we discuss the connection to the maximin projection learning problem in [35]. In Section A.2, we present additional results for the setting assuming no covariate shift.

## A.1 Connection to maximin projection

Follow the model setting in [35]: for the group $l$ with $1 \leq l \leq L$, the outcome $Y_i^{(l)} \in \mathbb{R}$, the binary treatment $A_i^{(l)} \in \{0,1\}$ and the baseline covariates $X_{i,\cdot}^{(l)} \in \mathbb{R}^p$, are modeled by

$$Y_i^{(l)} = h_l(X_{i,\cdot}^{(l)}) + A_i^{(l)} \cdot [(b^{(l)})^\intercal X_{i,\cdot}^{(l)} + c] + e_i^{(l)}$$

where $h_l : \mathbb{R}^p \to \mathbb{R}$, $b^{(l)} \in \mathbb{R}^p$, $c \in \mathbb{R}$ and the error $e_i^{(l)}$ satisfies $\mathbf{E}(e_i^{(l)} \mid X_{i,\cdot}^{(l)}, A_i^{(l)}) = 0$. In studying the individualized treatment effect, [35] has proposed the maximum projection

$$\beta^{*,\mathrm{MP}} = \arg\max_{\|\beta\|_2 \leq 1} \min_{1 \leq l \leq L} \beta^\intercal b^{(l)}. \tag{38}$$

**Proposition 4.** *The maximum projection $\beta^{*,\mathrm{MP}}$ in* (38) *satisfies*

$$\beta^{*,\mathrm{MP}} = \frac{1}{\|\beta^*(\mathrm{I})\|_2} \beta^*(\mathrm{I}) \quad \text{with} \quad \beta^*(\mathrm{I}) = \sum_{l=1}^{L} [\gamma^*(\mathrm{I})]_l b^{(l)}$$

*where $\gamma^*(\mathrm{I}) = \arg\min_{\gamma \in \Delta^L} \gamma^\intercal \Gamma^{\mathrm{I}} \gamma$ and $\Gamma^{\mathrm{I}}_{lk} = (b^{(l)})^\intercal b^{(k)}$ for $1 \leq l, k \leq L$.*

Through comparing the above proposition with Proposition 1, we note that the maximin projection is proportional to the general maximin effect defined in (5) with $\Sigma^Q = \mathrm{I}$ and hence the identification of $\beta^*(\mathrm{I})$ is instrumental in identifying $\beta^{*,\mathrm{MP}}$. We refer to [35] for more details on maximin projection in low-dimensional setting.

## A.2 Inference for $\Gamma^{\mathcal{Q}}$ with No Covariate Shift

We consider the special setting with no covariate shift. The corresponding results are similar to Theorem 2 and Proposition 3 which are derived for the general setting allowing for covariate shift. We define the covariance matrices $\mathbf{V}^{(j)} = (\mathbf{V}^{(j)}_{\pi(l_1,k_1),\pi(l_2,k_2)})_{(l_1,k_1)\in\mathcal{I}_L,(l_2,k_2)\in\mathcal{I}_L} \in \mathbb{R}^{L(L+1)/2 \times L(L+1)/2}$ for $j = 1, 2$ as

$$\begin{aligned}\mathbf{V}^{(1)}_{\pi(l_1,k_1),\pi(l_2,k_2)} = &\frac{\sigma_{l_1}^2}{n_{l_1}^2} [b^{(k_1)}]^\intercal [X^{(l_1)}]^\intercal \left[X^{(l_2)} b^{(k_2)} \mathbf{1}(l_2 = l_1) + X^{(k_2)} b^{(l_2)} \mathbf{1}(k_2 = l_1)\right] \\ &+ \frac{\sigma_{k_1}^2}{n_{l_1}^2} [b^{(l_1)}]^\intercal [X^{(k_1)}]^\intercal \left[X^{(l_2)} b^{(k_2)} \mathbf{1}(l_2 = k_1) + X^{(k_2)} b^{(l_2)} \mathbf{1}(k_2 = k_1)\right]\end{aligned} \tag{39}$$

$$\mathbf{V}^{(1)}_{\pi(l_1,k_1),\pi(l_2,k_2)} = \frac{\mathbf{E}[b^{(l_1)}]^\intercal X^{\mathcal{Q}}_{i,\cdot}[b^{(k_1)}]^\intercal X^{\mathcal{Q}}_{i,\cdot}[b^{(l_2)}]^\intercal X^{\mathcal{Q}}_{i,\cdot}[b^{(k_2)}]^\intercal X^{\mathcal{Q}}_{i,\cdot} - (b^{(l_1)})^\intercal \Sigma^{\mathcal{Q}} b^{(k_1)} (b^{(l_2)})^\intercal \Sigma^{\mathcal{Q}} b^{(k_2)}}{\sum_{l=1}^{L} n_l + N_{\mathcal{Q}}} \tag{40}$$

We establish the properties of the estimator $\widehat{\Gamma}^{\mathcal{Q}}_{l,k}$ in (21) for no covariate shift setting.

**Proposition 5.** *Consider the model* (1). *Suppose Condition* (A1) *holds and* $s \log p/n \to 0$ *with* $n = \min_{1\leq l\leq L} n_l$ *and* $s = \max_{1\leq l\leq L} \|b^{(l)}\|_0$. *If* $\{X^{(l)}_{i,\cdot}\}_{1\leq i\leq n_l} \overset{i.i.d.}{\sim} \mathcal{Q}$ *for* $1 \leq l \leq L$, *then the estimator* $\widehat{\Gamma}^{\mathcal{Q}}_{l,k}$ *defined in* (21) *satisfies*

$$\widehat{\Gamma}^{\mathcal{Q}}_{l,k} - \Gamma^{\mathcal{Q}}_{l,k} = D^{(1)}_{l,k} + D^{(2)}_{l,k} + \mathrm{Rem}_{l,k},$$

*where* $\mathrm{vecl}(D^{(1)}) \mid \{X^{(l)}\}_{1\leq l\leq L} \sim \mathcal{N}(0, \mathbf{V}^{(1)})$ *with* $\mathbf{V}^{(1)}$ *defined in* (39) *and the components of* $D^{(1)}$ *are uncorrelated with the components of* $D^{(2)}$*; there exists* $D^{(2),*} \in \mathbb{R}^{L\times L}$ *and* $T^* \sim \mathcal{N}(0, \mathbf{V}^{(2)})$ *with* $\mathbf{V}^{(2)}$ *defined in* (40) *such that* $D^{(2),*} \overset{d}{=} D^{(2)}$, $\mathrm{vecl}(D^{(2),*}) - T^* = o((\sum_{l=1}^{L} n_l + N_{\mathcal{Q}})^{-2/3})$ *almost surely; for* $1 \leq l, k \leq L$, *the reminder term* $\mathrm{Rem}_{l,k}$ *satisfies with probability larger than* $1 - \min\{n,p\}^{-c}$ *for a constant* $c > 0$,

$$|\mathrm{Rem}_{l,k}| \lesssim (1 + \|b^{(k)}\|_2 + \|b^{(l)}\|_2) \cdot \frac{s \log p}{n}. \tag{41}$$

*With probability larger than* $1 - p^{-c}$ *for a positive constant* $c > 0$,

$$\mathbf{V}^{(1)}_{\pi(l,k),\pi(l,k)} \asymp \frac{\|b^{(k)}\|_2^2 + \|b^{(k)}\|_2^2}{n} \quad \text{and} \quad \mathbf{V}^{(2)}_{\pi(l,k),\pi(l,k)} \lesssim \frac{\|b^{(l)}\|_2^2 \|b^{(k)}\|_2^2}{\sum_{l=1}^{L} n_l + N_{\mathcal{Q}}}. \tag{42}$$

Proposition 5 shows that $\mathrm{vecl}(\widehat{\Gamma}^{\mathcal{Q}}) - \mathrm{vecl}(\Gamma^{\mathcal{Q}})$ is approximated by the summation of two random vectors $\mathrm{vecl}(D^{(1)})$ and $\mathrm{vecl}(D^{(2)})$. We show that $\mathrm{vecl}(D^{(1)})$ and $\mathrm{vecl}(D^{(2)})$ are approximated by (conditionally or unconditionally) multivariate Gaussian separately, which is sufficiently for our analysis of characterizing the sampling accuracy.

In general, the theoretical results results for the general covariate shift setting holds for the more special no covariate shift setting. We simply replace $D$ and $\mathbf{V}$ in Theorem 2 with $D^{(1)} + D^{(2)}$ and $\mathbf{V}^{(1)} + \mathbf{V}^{(2)}$ in Proposition 5, respectively. We estimate $\mathbf{V}^{(1)}_{\pi(l_1,k_1),\pi(l_2,k_2)}$ in (39) and $\mathbf{V}^{(2)}_{\pi(l_1,k_1),\pi(l_2,k_2)}$ in (40) by

$$\begin{aligned}\widehat{\mathbf{V}}^{(1)}_{\pi(l_1,k_1),\pi(l_2,k_2)} = & \frac{\widehat{\sigma}^2_{l_1}}{n^2_{l_1}} [b^{(l_1)}]^\intercal [X^{(l_1)}]^\intercal X^{(l_1)} \left[b^{(l_2)} \mathbf{1}(l_2 = l_1) + b^{(k_2)} \mathbf{1}(k_2 = l_1)\right] \\ & + \frac{\widehat{\sigma}^2_{k_1}}{n^2_{l_1}} [b^{(k_1)}]^\intercal [X^{(k_1)}]^\intercal X^{(k_1)} \left[b^{(l_2)} \mathbf{1}(l_2 = k_1) + b^{(k_2)} \mathbf{1}(k_2 = k_1)\right]\end{aligned} \tag{43}$$

$$\widehat{\mathbf{V}}^{(2)}_{\pi(l_1,k_1),\pi(l_2,k_2)} = \frac{\sum_{i=1}^{N_{\mathcal{Q}}}\left((\widehat{b}^{(l_1)}_{init})^{\intercal}X^{\mathcal{Q}}_{i,\cdot}(\widehat{b}^{(k_1)}_{init})^{\intercal}X^{\mathcal{Q}}_{i,\cdot}(\widehat{b}^{(l_2)}_{init})^{\intercal}X^{\mathcal{Q}}_{i,\cdot}(\widehat{b}^{(k_2)}_{init})^{\intercal}X^{\mathcal{Q}}_{i,\cdot} - (\widehat{b}^{(l_1)}_{init})^{\intercal}\widehat{\Sigma}\widehat{b}^{(k_1)}_{init}(\widehat{b}^{(l_2)}_{init})^{\intercal}\widehat{\Sigma}\widehat{b}^{(k_2)}_{init}\right)}{(N+N_{\mathcal{Q}})^2}$$
$$+\frac{\sum_{l=1}^{L}\sum_{i=1}^{n_l}\left((\widehat{b}^{(l_1)}_{init})^{\intercal}X^{(l)}_{i,\cdot}(\widehat{b}^{(k_1)}_{init})^{\intercal}X^{(l)}_{i,\cdot}(\widehat{b}^{(l_2)}_{init})^{\intercal}X^{(l)}_{i,\cdot}(\widehat{b}^{(k_2)}_{init})^{\intercal}X^{(l)}_{i,\cdot} - (\widehat{b}^{(l_1)}_{init})^{\intercal}\widehat{\Sigma}\widehat{b}^{(k_1)}_{init}(\widehat{b}^{(l_2)}_{init})^{\intercal}\widehat{\Sigma}\widehat{b}^{(k_2)}_{init}\right)}{(N+N_{\mathcal{Q}})^2} \tag{44}$$

with $N=\sum_{l=1}^{L} n_l$ and $\widehat{\Sigma}=\frac{1}{N+N_{\mathcal{Q}}}\left(\sum_{l=1}^{L}\sum_{i=1}^{n_l} X^{(l)}_{i,\cdot}[X^{(l)}_{i,\cdot}]^{\intercal}+\sum_{i=1}^{N_{\mathcal{Q}}} X^{(l)}_{i,\cdot}[X^{(l)}_{i,\cdot}]^{\intercal}\right)$.

The sampling accuracy Theorem 3 also holds for the no covariate shift by replacing the constant $C^*(L,\alpha_0)$ in (34) by

$$C_1^*(L,\alpha_0)=\frac{\pi^{L(L+1)/4}}{2\sqrt{2\pi}\psi\left(L(L+1)/4+1\right)}\cdot\frac{\exp\left(-2F^{-1}_{\chi^2_{r_1}}(1-\frac{\alpha_0}{2})-F^{-1}_{\chi^2_{r_2}}(1-\frac{\alpha_0}{2})\right)}{\prod_{i=1}^{\frac{L(L+1)}{2}}\left[n\cdot\lambda_i(\mathbf{V}^{(1)}+\mathbf{V}^{(2)})+3d_0/2\right]} \tag{45}$$

where $\psi(\cdot)$ denotes the gamma function, $1\le r_1,r_2\le L(L+1)/2$ denote the ranks of $\mathbf{V}^{(1)}$ and $\mathbf{V}^{(2)}$, respectively. It follows from (42) in Proposition 5 and Condition (A2) that $n\cdot(\mathbf{V}^{(1)}+\mathbf{V}^{(2)})$ and $d_0$ are of a constant order with a high probability. Then for a finite $L$, we show that $C_1^*(L,\alpha_0)\ge c$ for a small positive constant $c>0$ .

**Proposition 6.** *Consider the model* (1) *with no covariate shift. Suppose Conditions* (A1) *and* (A2) *hold. For* $0<\alpha_0<1/2$, *we define* $\mathrm{err}_n(M)=\left[\frac{2\log n}{C_1^*(L,\alpha_0)M}\right]^{\frac{2}{L(L+1)}}$ *with the positive constant* $C_1^*(L,\alpha_0)$ *defined in* (34). *If* $\mathrm{err}_n(M)\le c$ *for a small positive constant* $c>0$, *then there exists a positive constant* $c_1>0$ *such that*

$$\mathbf{P}\left(\min_{1\le m\le M}\|\widehat{\Gamma}^{[m]}-\Gamma^{\mathcal{Q}}\|_F/\sqrt{2}\le \mathrm{err}_n(M)/\sqrt{n}\right)\ge 1-\alpha_0-n^{-c_1}-p^{-c_1}. \tag{46}$$

*By further assuming* $\lambda_{\min}(\Gamma^{\mathcal{Q}})+\delta>2\sqrt{2}\mathrm{err}_n(M)/\sqrt{n}$ *with* $\delta$ *defined in* (7), *then with probability larger than* $1-\alpha_0-n^{-c_1}-p^{-c_1}$, *there exists* $1\le m^*\le M$ *such that*

$$\|\widehat{\gamma}^{[m^*]}_{\delta}-\gamma^*_{\delta}\|_2\le\frac{\sqrt{2}\|\widehat{\Gamma}^{[m^*]}-\Gamma^{\mathcal{Q}}\|_2}{\lambda_{\min}(\widehat{\Gamma}^{[m^*]})+\delta}\|\gamma^*_{\delta}\|_2\le\frac{2\sqrt{2}\mathrm{err}_n(M)}{\lambda_{\min}(\Gamma^{\mathcal{Q}})+\delta}\cdot\frac{1}{\sqrt{n}} \tag{47}$$

In comparison to Theorem 3 for the general covariate shift setting, the assumption (33) can be removed for the simpler setting assuming no covariate shift.

# B Proofs

In this section, we present the proofs of Theorems 1 to 4 and Propositions 1 to 6. The proofs of additional lemmas are in Section C.

## B.1 High probability events

We introduce the following events to facilitate the proofs,

$$
\begin{aligned}
\mathcal{G}_0 &= \left\{ \left\| \frac{1}{n_l} [X^{(l)}]^\intercal \epsilon^{(l)} \right\|_\infty \lesssim \sqrt{\frac{\log p}{n_l}} \quad \text{for } 1 \leq l \leq L \right\}; \\
\mathcal{G}_1 &= \left\{ \max\left\{ \|\widehat{b}^{(l)}_{init} - b^{(l)}\|_2, \frac{1}{\sqrt{n_l}} \|X^{(l)}(\widehat{b}^{(l)}_{init} - b^{(l)})\|_2 \right\} \lesssim \sqrt{\|b^{(l)}\|_0 \frac{\log p}{n_l}} \sigma_l \quad \text{for } 1 \leq l \leq L \right\}, \\
\mathcal{G}_2 &= \left\{ \|\widehat{b}^{(l)}_{init} - b^{(l)}\|_1 \lesssim \|b^{(l)}\|_0 \sqrt{\frac{\log p}{n_l}} \sigma_l, \|[\widehat{b}^{(l)}_{init} - b^{(l)}]_{\mathcal{S}_l^c}\|_1 \leq C \|[\widehat{b}^{(l)}_{init} - b^{(l)}]_{\mathcal{S}_l}\|_1 \quad \text{for } 1 \leq l \leq L \right\}, \\
\mathcal{G}_3 &= \left\{ |\widehat{\sigma}_l^2 - \sigma_l^2| \lesssim \|b^{(l)}\|_0 \frac{\log p}{n_l} + \sqrt{\frac{\log p}{n_l}} \quad \text{for } 1 \leq l \leq L \right\},
\end{aligned} \tag{48}
$$

where $\mathcal{S}_l \subset [p]$ denotes the support of $b^{(l)}$ for $1 \leq l \leq L$.

Recall that, for $1 \leq l \leq L$, $\widehat{b}^{(l)}_{init}$ is the Lasso estimator defined in (14) with $\lambda_l = \sqrt{(2+c)\log p/|A_l|}\sigma_l$ for some constant $c > 0$ or the Lasso estimator based on the non-split data; $\widehat{\sigma}_l^2 = \|Y^{(l)} - X^{(l)}\widehat{b}^{(l)}\|_2^2/n_l$ for $1 \leq l \leq L$. Then under Condition (A1), we can establish

$$
\mathbf{P}\left(\cap_{j=0}^3 \mathcal{G}_j\right) \geq 1 - \min\{n, p\}^{-c}. \tag{49}
$$

The above high-probability statement follows from the existing literature results on the theoretical analysis of Lasso estimator. We shall point to the exact literature results. The control of the probability of $\mathcal{G}_0$ follows from Lemma 6.2 of [7]. Regarding the events $\mathcal{G}_1$ and $\mathcal{G}_2$, the control of $\|\widehat{b}^{(l)}_{init} - b^{(l)}\|_1$, $\|\widehat{b}^{(l)}_{init} - b^{(l)}\|_2$ and $\frac{1}{\sqrt{n_l}}\|X^{(l)}(\widehat{b}^{(l)}_{init} - b^{(l)})\|_2$ can be found in Theorem3 of [48], Theorem 7.2 of [4] or Theorem 6.1 of [7]; the control of $\|[\widehat{b}^{(l)}_{init} - b^{(l)}]_{\mathcal{S}_l^c}\|_1 \leq C\|[\widehat{b}^{(l)}_{init} - b^{(l)}]_{\mathcal{S}_l}\|_1$ can be found in Corollary B.2 of [4] or Lemma 6.3 of [7]. For the event $\mathcal{G}_4$, its probability can be controlled as Theorem 2 or (20) in [39].

We further define the following events,

$$
\begin{aligned}
\mathcal{G}_4 &= \left\{ \|\widetilde{\Sigma}^{\mathcal{Q}} - \Sigma^{\mathcal{Q}}\|_2 \lesssim \sqrt{\frac{p}{N}} + \frac{p}{N} \right\} \\
\mathcal{G}_5 &= \left\{ \max_{\mathcal{S} \subset [p], |\mathcal{S}| \leq s} \max_{\|w_{\mathcal{S}^c}\|_1 \leq C \|w_{\mathcal{S}}\|_1} \left| \frac{w^\intercal \left(\frac{1}{N}\sum_{i=1}^N X^{\mathcal{Q}}_{i,\cdot}[X^{\mathcal{Q}}_{i,\cdot}]^\intercal\right) w}{w^\intercal E(X^{\mathcal{Q}}_{i,\cdot}[X^{\mathcal{Q}}_{i,\cdot}])w} - 1 \right| \lesssim \sqrt{\frac{s \log p}{N}} \right\} \\
\mathcal{G}_6(w, v, t) &= \left\{ \left| w^\intercal \left(\widehat{\Sigma}^{\mathcal{Q}} - \Sigma^{\mathcal{Q}}\right) v \right|, \left| w^\intercal \left(\widetilde{\Sigma}^{\mathcal{Q}} - \Sigma^{\mathcal{Q}}\right) v \right| \lesssim t \frac{\|(\Sigma^{\mathcal{Q}})^{1/2} w\|_2 \|(\Sigma^{\mathcal{Q}})^{1/2} v\|_2}{\sqrt{N_{\mathcal{Q}}}} \right\}
\end{aligned} \tag{50}
$$

for any given vectors $w, v \in \mathbb{R}^p$. If $X^{\mathcal{Q}}_{i,\cdot}$ is sub-gaussian, then we have

$$
\mathbf{P}\left(\mathcal{G}_4 \cap \mathcal{G}_5\right) \geq 1 - p^{-c} \tag{51}
$$

$$\mathbf{P}\left(\mathcal{G}_6(w,v,t)\right) \geq 1-2\exp(-ct^2) \tag{52}$$

Since $X^{\mathcal{Q}}_{i,\cdot}$ is sub-gaussian, it follows from equation (5.26) of [43] that the event $\mathcal{G}_4$ holds with probability larger than $1-\exp(-cp)$ for some positive constant $c>0$.; it follows from Theorem 1.6 of [52] that the event $\mathcal{G}_5$ holds with probability larger than $1-p^{-c}$ for some positive constant $c>0$. The proof of (52) follows from Lemma 10 in the supplement of [10].

## B.2 Proofs of Theorem 1 and Proposition 2

For $\mathcal{B}=\begin{pmatrix} b^{(1)} & b^{(2)} & \ldots & b^{(L)}\end{pmatrix} \in \mathbb{R}^{p\times L}$, we define its SVD as

$$\mathcal{B}=U\Lambda V^{\intercal}$$

where $U\in\mathbb{R}^{p\times L}$, $\Lambda\in\mathbb{R}^{L\times L}$ and $V\in\mathbb{R}^{L\times L}$ with $\Lambda_{1,1}\geq\ldots\geq\Lambda_{L,L}>0$. We define $\Delta=\delta\cdot U\Lambda^{-2}U^{\intercal}\in\mathbb{R}^{p\times p}$. As a consequence,

$$[\mathcal{B}]^{\intercal}\Delta\mathcal{B}=\delta\mathrm{I} \quad \text{and} \quad \mathcal{B}^{\intercal}(\Sigma^{\mathcal{Q}}+\Delta)\mathcal{B}=\Gamma^{\mathcal{Q}}+\delta\cdot\mathrm{I}. \tag{53}$$

It follows from Proposition 1 and the definition of $\beta^*_\delta$ in (7) that

$$\beta^*_\delta=\max_{\beta\in\mathbb{R}^p}\min_{b\in\mathbb{B}}\left[2b^{\intercal}(\Sigma^{\mathcal{Q}}+\Delta)\beta-\beta^{\intercal}(\Sigma^{\mathcal{Q}}+\Delta)\beta\right]$$

and

$$\min_{b\in\mathbb{B}}\left[2b^{\intercal}(\Sigma^{\mathcal{Q}}+\Delta)\beta^*_\delta-[\beta^*_\delta]^{\intercal}(\Sigma^{\mathcal{Q}}+\Delta)\beta^*_\delta\right]=[\beta^*_\delta]^{\intercal}(\Sigma^{\mathcal{Q}}+\Delta)\beta^*_\delta=[\gamma^*_\delta]^{\intercal}\left(\Gamma^{\mathcal{Q}}+\delta\cdot\mathrm{I}\right)\gamma^*_\delta \tag{54}$$

Now we compute the lower bound for $R_{\mathcal{Q}}(\beta^*_\delta)=\min_{b\in\mathbb{B}}\left[2b^{\intercal}\Sigma^{\mathcal{Q}}\beta^*_\delta-[\beta^*_\delta]^{\intercal}\Sigma^{\mathcal{Q}}\beta^*_\delta\right].$

We apply (53) and $\beta^*_\delta=\mathcal{B}\gamma^*_\delta$ and establish

$$\begin{aligned}
&\min_{b\in\mathbb{B}}\left[2b^{\intercal}(\Sigma^{\mathcal{Q}}+\Delta)\beta^*_\delta-[\beta^*_\delta]^{\intercal}(\Sigma^{\mathcal{Q}}+\Delta)\beta^*_\delta\right]\\
&=\min_{b\in\mathbb{B}}\left[2b^{\intercal}(\Sigma^{\mathcal{Q}}+\Delta)\beta^*_\delta-[\beta^*_\delta]^{\intercal}\Sigma^{\mathcal{Q}}\beta^*_\delta\right]-\delta\|\gamma^*_\delta\|_2^2\\
&\leq\min_{b\in\mathbb{B}}\left[2b^{\intercal}\Sigma^{\mathcal{Q}}\beta^*_\delta-[\beta^*_\delta]^{\intercal}\Sigma^{\mathcal{Q}}\beta^*_\delta\right]+2\max_{b\in\mathbb{B}}b^{\intercal}\Delta\beta^*_\delta-\delta\|\gamma^*_\delta\|_2^2\\
&=R_{\mathcal{Q}}(\beta^*_\delta)+2\delta\max_{\gamma\in\Delta^L}\gamma^{\intercal}\gamma^*_\delta-\delta\|\gamma^*_\delta\|_2^2
\end{aligned}$$

where the last inequality follows from Proposition 1 and the function $b^{\intercal}\Delta\beta^*_\delta$ is linear in $b$, $\beta^*_\delta=\mathcal{B}\gamma^*_\delta$ and (53). Combined with (54), we establish

$$\begin{aligned}
R_{\mathcal{Q}}(\beta^*_\delta)&\geq[\gamma^*_\delta]^{\intercal}\left(\Gamma^{\mathcal{Q}}+\delta\cdot\mathrm{I}\right)\gamma^*_\delta-2\delta\max_{\gamma\in\Delta^L}\gamma^{\intercal}\gamma^*_\delta+\delta\|\gamma^*_\delta\|_2^2\\
&\geq[\gamma^*]^{\intercal}\Gamma^{\mathcal{Q}}\gamma^*_\delta+2\delta\|\gamma^*_\delta\|_2^2-2\delta\max_{\gamma\in\Delta^L}\gamma^{\intercal}\gamma^*_\delta\\
&=R_{\mathcal{Q}}(\beta^*)+2\delta\left(\|\gamma^*_\delta\|_2^2-\max_{\gamma\in\Delta^L}\gamma^{\intercal}\gamma^*_\delta\right)
\end{aligned} \tag{55}$$

Note that $\max_{\gamma \in \Delta^L} \gamma^\intercal \gamma_\delta^* - \|\gamma_\delta^*\|_2^2 \geq 0$ and

$$\max_{\gamma \in \Delta^L} \gamma^\intercal \gamma_\delta^* - \|\gamma_\delta^*\|_2^2 = \|\gamma_\delta^*\|_\infty - \|\gamma_\delta^*\|_2^2$$

We use $j^* \in [L]$ to denote the index such that $[\gamma_\delta^*]_{j^*} = \|\gamma_\delta^*\|_\infty$. Then we have

$$\begin{aligned}
\|\gamma_\delta^*\|_\infty - \|\gamma_\delta^*\|_2^2 &= [\gamma_\delta^*]_{j^*} - [\gamma_\delta^*]_{j^*}^2 - \sum_{l \neq j^*} [\gamma_\delta^*]_l^2 \\
&\leq [\gamma_\delta^*]_{j^*} - [\gamma_\delta^*]_{j^*}^2 - \frac{1}{L-1} \left( \sum_{l \neq j^*} [\gamma_\delta^*]_l \right)^2 \qquad (56) \\
&= [\gamma_\delta^*]_{j^*} - [\gamma_\delta^*]_{j^*}^2 - \frac{1}{L-1} (1 - [\gamma_\delta^*]_{j^*})^2
\end{aligned}$$

We take the minimum value of the right hand side with respect to $[\gamma_\delta^*]_{j^*}$ over the domain $[1/L, 1]$. Then we obtain

$$\max_{\frac{1}{L} \leq [\gamma_\delta^*]_{j^*} \leq 1} [\gamma_\delta^*]_{j^*} - [\gamma_\delta^*]_{j^*}^2 - \frac{1}{L-1} (1 - [\gamma_\delta^*]_{j^*})^2 = \frac{1}{4} \left( 1 - \frac{1}{L} \right)$$

where the minimum value is achieved at $[\gamma_\delta^*]_{j^*} = \frac{1+\frac{1}{L}}{2}$. Combined with (55) and (56), we establish

$$R_{\mathcal{Q}}(\beta_\delta^*) \geq R_{\mathcal{Q}}(\beta^*) - \frac{\delta}{2} \cdot \left( 1 - \frac{1}{L} \right).$$

## B.3 Proof of Proposition 2

We establish $[\gamma_\delta^*]_1 \to 1/2$ by (8) and the condition $\delta = \delta(n,p) \gg \max\{|\Gamma_{11} - \Gamma_{12}| + |\Gamma_{22} - \Gamma_{12}|\}$. We note that $\lambda_{\min}(\Gamma) \to 0$ implies that $\lambda_{\min}(\mathcal{B}) \to 0$ as $n, p \to \infty$. In the following discussion, we shall use the notation $n \to \infty$ and adopt the conventional notation: $p = p(n)$ and $p \to \infty$ as $n \to \infty$. We consider two separate cases:

1. If $\lambda_{\min}(\mathcal{B}) > 0$ for a given $n$, then we have $\max_{\gamma \in \Delta^L} \gamma^\intercal \gamma_\delta^* - \|\gamma_\delta^*\|_2^2 = \|\gamma_\delta^*\|_\infty - \|\gamma_\delta^*\|_2^2 \to 0$ based on the fact that $[\gamma_\delta^*]_1 \to 1/2$. By (55), for any given small positive $\epsilon_0 > 0$, there exists $N_0$ such that for $n \geq N_0$ and $\lambda_{\min}(\mathcal{B}) > 0$, then $|R_{\mathcal{Q}}(\beta_\delta^*) - R_{\mathcal{Q}}(\beta^*)| \leq \epsilon_0$.

2. If $\lambda_{\min}(\mathcal{B}) = 0$ for a given $n$, then $b^{(1)}$ and $b^{(2)}$ are collinear. By the assumption $\lambda_{\min}(\Gamma^{\mathcal{Q}}) \to 0$, we have $b^{(2)} = (1 + c_n) b^{(1)}$ with $c_n \to 0$. Hence $\beta_\delta^* = b^{(1)}(1 + \gamma_\delta^* c_n)$ and

   $R_{\mathcal{Q}}(\beta_\delta^*) = \min\{(1 + \gamma_\delta^* c_n), (1 + c_n)(1 + \gamma_\delta^* c_n)\}\, 2[b^{(1)}]^\intercal \Sigma^{\mathcal{Q}} b^{(1)} - (1 + \gamma_\delta^* c_n)^2 [b^{(1)}]^\intercal \Sigma^{\mathcal{Q}} b^{(1)}.$

   For any $\gamma_\delta^* \in [0,1]$, we have $R_{\mathcal{Q}}(\beta_\delta^*) \to [b^{(1)}]^\intercal \Sigma^{\mathcal{Q}} b^{(1)}$ for any $\delta \geq 0$. For any given small positive $\epsilon_0 > 0$, there exists $N_1$ such that for $n \geq N_1$ and $\lambda_{\min}(\mathcal{B}) = 0$, then $|R_{\mathcal{Q}}(\beta_\delta^*) - R_{\mathcal{Q}}(\beta^*)| \leq \epsilon_0$.

For any $\delta \geq 0$ and $\epsilon_0 > 0$, we take $N_* = \max\{N_0, N_1\}$ and have shown that for $n \geq N_*$, $|R_{\mathcal{Q}}(\beta^*_\delta) - R_{\mathcal{Q}}(\beta^*)| \leq \epsilon_0$. That is, $R_{\mathcal{Q}}(\beta^*_\delta) - R_{\mathcal{Q}}(\beta^*) \to 0$.

## B.4 Proof of Theorem 2

We decompose the error $\widehat{\Gamma}^{\mathcal{Q}}_{l,k} - \Gamma^{\mathcal{Q}}_{l,k}$ as

$$\begin{aligned}
\widehat{\Gamma}^{\mathcal{Q}}_{l,k} - \Gamma^{\mathcal{Q}}_{l,k} = & \frac{1}{|B_l|}(\widehat{u}^{(l,k)})^\intercal [X^{(l)}_{B_l,\cdot}]^\intercal \epsilon^{(l)}_{B_l} + \frac{1}{|B_k|}(\widehat{u}^{(k,l)})^\intercal [X^{(k)}_{B_k,\cdot}]^\intercal \epsilon^{(k)}_{B_k} \\
& + (b^{(l)})^\intercal (\widehat{\Sigma}^{\mathcal{Q}} - \Sigma^{\mathcal{Q}}) b^{(k)} - (\widehat{b}^{(l)}_{init} - b^{(l)})^\intercal \widehat{\Sigma}^{\mathcal{Q}} (\widehat{b}^{(k)}_{init} - b^{(k)}) \\
& + (\widehat{\Sigma}^{\mathcal{Q}} \widehat{b}^{(k)}_{init} - \widehat{\Sigma}^{(l)} \widehat{u}^{(l,k)})^\intercal (\widehat{b}^{(l)}_{init} - b^{(l)}) + (\widehat{\Sigma}^{\mathcal{Q}} \widehat{b}^{(l)}_{init} - \widehat{\Sigma}^{(l)} \widehat{u}^{(k,l)})^\intercal (\widehat{b}^{(k)}_{init} - b^{(k)})
\end{aligned} \tag{57}$$

We define $D_{l,k} = D^{(a)}_{l,k} + D^{(b)}_{l,k}$ with

$$D^{(a)}_{l,k} = \frac{1}{|B_l|}(\widehat{u}^{(l,k)})^\intercal [X^{(l)}_{B_l,\cdot}]^\intercal \epsilon^{(l)}_{B_l} + \frac{1}{|B_k|}(\widehat{u}^{(k,l)})^\intercal [X^{(k)}_{B_k,\cdot}]^\intercal \epsilon^{(k)}_{B_k}$$

and

$$D^{(b)}_{l,k} = (b^{(l)})^\intercal (\widehat{\Sigma}^{\mathcal{Q}} - \Sigma^{\mathcal{Q}}) b^{(k)}.$$

Note that $D^{(a)}_{l,k}$ is a function of $X^{\mathcal{Q}}_{A,\cdot}$, $\{X^{(l)}\}_{1\leq l\leq L}$ and $\{\epsilon^{(k)}\}_{1\leq l\leq L}$ and $D^{(b)}_{l,k}$ is a function of the sub-sample $X^{\mathcal{Q}}_{B,\cdot}$ and hence $D^{(a)}_{l,k}$ is independent of $D^{(b)}_{l,k}$. We define $\mathrm{Rem}_{l,k}$ as

$$\begin{aligned}
\mathrm{Rem}_{l,k} = & (\widehat{b}^{(l)}_{init} - b^{(l)})^\intercal \widehat{\Sigma}^{\mathcal{Q}} (\widehat{b}^{(k)}_{init} - b^{(k)}) + (\widehat{\Sigma}^{\mathcal{Q}} \widehat{b}^{(k)}_{init} - \widehat{\Sigma}^{(l)} \widehat{u}^{(l,k)})^\intercal (\widehat{b}^{(l)}_{init} - b^{(l)}) \\
& + (\widehat{\Sigma}^{\mathcal{Q}} \widehat{b}^{(l)}_{init} - \widehat{\Sigma}^{(l)} \widehat{u}^{(k,l)})^\intercal (\widehat{b}^{(k)}_{init} - b^{(k)}).
\end{aligned}$$

By (57), we have $\widehat{\Gamma}^{\mathcal{Q}}_{l,k} - \Gamma^{\mathcal{Q}}_{l,k} = D_{l,k} + \mathrm{Rem}_{l,k}$. In the following, we control the decomposition term by term.

**Control of the reminder term** $\mathrm{Rem}_{l,k}$. The following lemma controls the reminder term $\mathrm{Rem}_{l,k}$ in (30).

**Lemma 1.** *With probability larger than* $1 - \min\{n,p\}^{-c}$, *we have*

$$\left|(\widehat{b}^{(l)}_{init} - b^{(l)})^\intercal \widehat{\Sigma}^{\mathcal{Q}} (\widehat{b}^{(k)}_{init} - b^{(k)})\right| \lesssim \sqrt{\frac{\|b^{(l)}\|_0 \|b^{(k)}\|_0 (\log p)^2}{n_l n_k}}. \tag{58}$$

$$\left|(\widehat{\Sigma}^{\mathcal{Q}} \widehat{b}^{(k)}_{init} - \widehat{\Sigma}^{(l)} \widehat{u}^{(l,k)})^\intercal (\widehat{b}^{(l)}_{init} - b^{(l)})\right| \lesssim \|\widetilde{\Sigma}^{\mathcal{Q}} \widehat{b}^{(k)}_{init}\|_2 \frac{\|b^{(l)}\|_0 \log p}{n_l} + \|\widehat{b}^{(k)}_{init}\|_2 \sqrt{\frac{\|b^{(l)}\|_0 (\log p)^2}{n_l N_{\mathcal{Q}}}} \tag{59}$$

$$\left|(\widehat{\Sigma}^{\mathcal{Q}}\widehat{b}^{(l)}_{init} - \widehat{\Sigma}^{(l)}\widehat{u}^{(k,l)})^{\intercal}\widehat{\Sigma}^{\mathcal{Q}}(\widehat{b}^{(k)}_{init} - b^{(k)})\right| \lesssim \|\widetilde{\Sigma}^{\mathcal{Q}}\widehat{b}^{(l)}_{init}\|_2 \frac{\|b^{(k)}\|_0 \log p}{n_k} + \|\widehat{b}^{(l)}_{init}\|_2 \sqrt{\frac{\|b^{(k)}\|_0 (\log p)^2}{n_k N_{\mathcal{Q}}}} \tag{60}$$

The proof of the above lemma is presented in Section C.1.

**Distribution of $D^{(a)}$.** Since $\widehat{b}^{(k)}_{init}$ is a function of $(X^{(l)}_{A_l,\cdot}, \epsilon^{(k)}_{A_l})$, the projection direction $\widehat{u}^{(l,k)}$ is a function of $X^{\mathcal{Q}}_{A,\cdot}$, $X^{(l)}_{B_l,\cdot}$ and $(X^{(l)}_{A_l,\cdot}, \epsilon^{(k)}_{A_l})$. By the Gaussian error assumption, we establish

$$\mathrm{vecl}(D^{(a)}) \mid X^{\mathcal{Q}}_{A,\cdot}, \{X^{(l)}\}_{1\le l\le L}, \{\epsilon^{(k)}_{A_l}\}_{1\le l\le L} \sim N(\mathbf{0}, \mathbf{V}^{(a)}) \tag{61}$$

where

$$\begin{aligned}
\mathbf{V}^{(a)}_{\pi(l_1,k_1),\pi(l_2,k_2)} &= \mathrm{Cov}(D^{(a)}_{l_1,k_1}, D^{(a)}_{l_2,k_2}) \\
&= \frac{\sigma^2_{l_1}}{|B_{l_1}|}(\widehat{u}^{(l_1,k_1)})^{\intercal}\widehat{\Sigma}^{(l_1)}\left[\widehat{u}^{(l_2,k_2)}\mathbf{1}(l_2 = l_1) + \widehat{u}^{(k_2,l_2)}\mathbf{1}(k_2 = l_1)\right] \\
&+ \frac{\sigma^2_{k_1}}{|B_{k_1}|}(\widehat{u}^{(k_1,l_1)})^{\intercal}\widehat{\Sigma}^{(k_1)}\left[\widehat{u}^{(l_2,k_2)}\mathbf{1}(l_2 = k_1) + \widehat{u}^{(k_2,l_2)}\mathbf{1}(k_2 = k_1)\right]
\end{aligned} \tag{62}$$

with the index mapping $\pi$ defined in (3).

**Distribution of $D^{(2)}$.** For $i \in B$ and $(l,k) \in \mathcal{I}_L$, we define $W_{i,\cdot} \in \mathbb{R}^{L(L+1)/2}$ as

$$W_{i,\pi(l,k)} = [b^{(l)}]^{\intercal}(X^{\mathcal{Q}}_{i,\cdot}[X^{\mathcal{Q}}_{i,\cdot}]^{\intercal} - \Sigma^{\mathcal{Q}})b^{(k)}$$

Then we have

$$[\mathrm{vecl}(D^{(b)})]_{\pi(l,k)} = D^{(b)}_{l,k} = \frac{1}{|B|}\sum_{i\in B} W_{i,\pi(l,k)},$$

where the index mapping $\pi$ is defined in (3). The random variable $W_{i,\cdot} \in \mathbb{R}^{L(L+1)/2}$ is of mean zero and covariance matrix $\mathbf{C} \in \mathbb{R}^{L(L+1)/2\times L(L+1)/2}$, defined as,

$$\begin{aligned}
&\mathbf{C}_{\pi(l_1,k_1),\pi(l_2,k_2)} = \mathbf{E}W_{i,\pi(l_1,k_1)}[W_{i,\pi(l_2,k_2)}]^{\intercal} \\
&= \mathbf{E}\left([b^{(l_1)}]^{\intercal}X^{\mathcal{Q}}_{i,\cdot}[b^{(k_1)}]^{\intercal}X^{\mathcal{Q}}_{i,\cdot} - (b^{(l_1)})^{\intercal}\Sigma^{\mathcal{Q}}b^{(k_1)}\right)\left([b^{(l_2)}]^{\intercal}X^{\mathcal{Q}}_{i,\cdot}[b^{(k_2)}]^{\intercal}X^{\mathcal{Q}}_{i,\cdot} - (b^{(l_2)})^{\intercal}\Sigma^{\mathcal{Q}}b^{(k_2)}\right).
\end{aligned} \tag{63}$$

Now we shall show that $[\mathrm{vecl}(D^{(b)})]$ can be approximated by a normal random variable. The following lemma is the multivariate version of Komlós, Major, and Tusnády theory, which restates Theorem 1 of [18] in the current paper's terminology.

**Lemma 2.** *Let $W_{i,\cdot} : \Omega \to \mathbb{R}^d$ be a random vector with mean zero and covariance matrix* $\mathbf{C}$. *Suppose that $\mathbf{E}\|W_{i,\cdot}\|_2^{r_0} < \infty$ for $r_0 \geq 3$. Then one can construct a probability space $(\Omega_0, \mathcal{A}_0, P_0)$ and two sequences of independent random vectors $\{W_{i,\cdot}^0\}$ and $Z_{i,\cdot}^0$ with $W_{i,\cdot}^0$ having the same distribution as $W_{i,\cdot}$ and $Z_{i,\cdot}^0 \sim \mathcal{N}(0, \mathbf{C})$ such that*

$$\sum_{i=1}^{n} W_{i,\cdot}^0 - \sum_{i=1}^{n} Z_{i,\cdot}^0 = o(n^{1/r_0}) \quad a.s. \tag{64}$$

Now we verify conditions of the above lemma. For $i \in B$, $\mathbf{E}W_{i,\cdot}$ is of mean zero and covariance $\mathbf{C}$ defined in (117). For any $r_0 \geq 3$, we have

$$\begin{aligned}
\mathbf{E}\|W_{i,\cdot}\|_2^{r_0} = \mathbf{E}\left(\sum_{(l,k)\in\mathcal{I}_L} W_{i,\pi(l,k)}^2\right)^{\frac{r_0}{2}} &\leq (L(L+1)/2)^{\frac{r_0}{2}}\, \mathbf{E} \max_{(l,k)\in\mathcal{I}_L} |W_{i,\pi(l,k)}|^{r_0} \\
&\leq (L(L+1)/2)^{\frac{r_0}{2}} \sum_{(l,k)\in\mathcal{I}_L} \mathbf{E}|W_{i,\pi(l,k)}|^{r_0} \\
&\leq (L(L+1)/2)^{\frac{r_0}{2}+1} \max_{(l,k)\in\mathcal{I}_L} \mathbf{E}|W_{i,\pi(l,k)}|^{r_0}
\end{aligned}$$

Since $(b^{(l)})^\intercal X_{i,\cdot}^{\mathcal{Q}}$ and $(b^{(k)})^\intercal X_{i,\cdot}^{\mathcal{Q}}$ are sub-gaussian random variables, then $[b^{(l)}]^\intercal (X_{i,\cdot}^{\mathcal{Q}}[X_{i,\cdot}^{\mathcal{Q}}]^\intercal - \Sigma) b^{(k)}$ is sub-exponential, that is, $\mathbf{E}|W_{i,\pi(l,k)}|^{r_0} \lesssim r_0^{r_0}$. Hence

$$\mathbf{E}\|W_{i,\cdot}\|_2^{r} \lesssim (L(L+1)/2)^{\frac{r_0}{2}+1}\, r_0^{r_0}. \tag{65}$$

For any given positive $r_0 \geq 3$, we have shown that $\mathbf{E}\|W_{i,\cdot}\|_2^{r_0} \leq C_0$ for some constant $C_0 > 0$. By applying Lemma 2, we establish that there exist two sequences of independent random vectors $W_{i,\cdot}^0$ and $Z_{i,\cdot}^0 \sim \mathcal{N}(0, \mathbf{C})$ for $i \in B$ such that $W_{i,\cdot}^0$ has the same distribution as $W_{i,\cdot}$ and

$$\left\|\sum_{i\in B} W_{i,\cdot}^0 - \sum_{i\in B} Z_{i,\cdot}^0\right\|_2 \lesssim |N_{\mathcal{Q}}|^{\frac{1}{r_0}} \quad \text{a.s.} \tag{66}$$

We define

$$\text{vecl}(D^{(b),*}) = \frac{1}{|B|}\sum_{i\in B} W_{i,\cdot}^0$$

and as a consequence, the corresponding random matrix $D^{(b),*}$ has the same distribution with $D^{(b)}$. We couple the underlying probability spaces of $D^{(a)}$ and $D^{(b),*}$ and use the product measure of these two spaces as the joint probability measure of $D^{(a)}$ and $D^{(b),*}$.

After coupling, the two random objects $D^{(b),*}$ and $D^{(a)}$ are independent of each other. Since $D^{(a)}$ is independent of $D^{(b)}$, we establish

$$\operatorname{vecl}(D^{(a)}) + \operatorname{vecl}(D^{(b)}) \overset{d}{=} \operatorname{vecl}(D^{(a)}) + \operatorname{vecl}(D^{(b),*}) \tag{67}$$

We define $D^* = D^{(a)} + D^{(b),*}$. A combination of (66) and (67) leads to

$$[\operatorname{vecl}(D^*)] - S^* = o(N_{\mathcal{Q}}^{1/r_0-1}) \quad \text{with} \quad S^* = \operatorname{vecl}(D^{(a)}) + \frac{1}{|B|}\sum_{i\in B} Z^0_{i,\cdot}. \tag{68}$$

Conditioning on $X^{\mathcal{Q}}_{A,\cdot}$ and $\{X^{(l)}, \epsilon^{(k)}_{A_l}\}_{1\leq l\leq L}$, the random variable $S^* \sim \mathcal{N}(0, \mathbf{V})$ with $\mathbf{V}$ defined in (29). We then establish the distribution of $D$ in Theorem 2 since $1 - 1/r_0 \in (2/3, 1)$.

## B.5 Proof of Proposition 3

We have the expression for the diagonal element of $\mathbf{V}$ as

$$\begin{aligned}\mathbf{V}_{\pi(l,k),\pi(l,k)} = {}& \frac{\sigma_l^2}{|B_l|}(\widehat{u}^{(l,k)})^\intercal \widehat{\Sigma}^{(l)}\left[\widehat{u}^{(l,k)} + \widehat{u}^{(k,l)}\mathbf{1}(k=l)\right] + \frac{\sigma_k^2}{|B_k|}(\widehat{u}^{(k,l)})^\intercal \widehat{\Sigma}^{(k)}\left[\widehat{u}^{(l,k)}\mathbf{1}(l=k) + \widehat{u}^{(k,l)}\right] \\ & + \frac{1}{|B|}(\mathbf{E}[b^{(l)}]^\intercal X^{\mathcal{Q}}_{i,\cdot}[b^{(k)}]^\intercal X^{\mathcal{Q}}_{i,\cdot}[b^{(l)}]^\intercal X^{\mathcal{Q}}_{i,\cdot}[b^{(k)}]^\intercal X^{\mathcal{Q}}_{i,\cdot} - (b^{(l)})^\intercal \Sigma^{\mathcal{Q}} b^{(k)} (b^{(l)})^\intercal \Sigma^{\mathcal{Q}} b^{(k)})\end{aligned}$$

We introduce the following lemma, which restates Lemma 1 of [11] in the current paper's terminology.

**Lemma 3.** *Suppose that the condition* (A1) *holds, then with probability larger than* $1-p^{-c}$,

$$c\frac{\|\omega^{(k)}\|_2^2}{n_l} \leq \frac{1}{|B_l|}(\widehat{u}^{(l,k)})^\intercal \widehat{\Sigma}^{(l)} \widehat{u}^{(l,k)} \leq C\frac{\|\omega^{(k)}\|_2^2}{n_l}, \quad \textit{for} \quad 1 \leq l,k \leq L,$$

$$c\frac{\|\omega^{(l)}\|_2^2}{n_k} \leq \frac{1}{|B_k|}(\widehat{u}^{(k,l)})^\intercal \widehat{\Sigma}^{(k)} \widehat{u}^{(l,k)} \leq C\frac{\|\omega^{(l)}\|_2^2}{n_k}, \quad \textit{for} \quad 1 \leq l,k \leq L,$$

*for some positive constants* $C > c > 0$.

We apply the above lemma to establish the lower bound in (31) and also part of the upper bound in (31). Since $X^{\mathcal{Q}}_{i,\cdot}$ is sub-gaussian, we have

$$\left|\mathbf{E}[b^{(l_1)}]^\intercal X^{\mathcal{Q}}_{i,\cdot}[b^{(k_1)}]^\intercal X^{\mathcal{Q}}_{i,\cdot}[b^{(l_2)}]^\intercal X^{\mathcal{Q}}_{i,\cdot}[b^{(k_2)}]^\intercal X^{\mathcal{Q}}_{i,\cdot}\right| \lesssim \|b^{(l_1)}\|_2\|b^{(k_1)}\|_2\|b^{(l_2)}\|_2\|b^{(k_2)}\|_2 \tag{69}$$

and

$$(b^{(l_1)})^\intercal \Sigma^{\mathcal{Q}} b^{(k_1)} (b^{(l_2)})^\intercal \Sigma^{\mathcal{Q}} b^{(k_2)} \lesssim \|b^{(l_1)}\|_2\|b^{(k_1)}\|_2\|b^{(l_2)}\|_2\|b^{(k_2)}\|_2. \tag{70}$$

Then we establish the upper bound of (31) by taking $l_1 = l_2 = l$ and $k_1 = k_2 = k$.

To establish (32), we control $\|\omega^{(k)}\|_2$ as follows,

$$\begin{aligned}\|\omega^{(k)}\|_2 = \|\widetilde{\Sigma}^{\mathcal{Q}}\widehat{b}^{(k)}_{init}\|_2 &\leq \|\Sigma^{\mathcal{Q}}\widehat{b}^{(k)}_{init}\|_2 + \|(\widetilde{\Sigma}^{\mathcal{Q}} - \Sigma^{\mathcal{Q}})\widehat{b}^{(k)}_{init}\|_2 \\ &\leq \lambda_{\max}(\Sigma^{\mathcal{Q}})\|\widehat{b}^{(k)}_{init}\|_2 + \|\widetilde{\Sigma}^{\mathcal{Q}} - \Sigma^{\mathcal{Q}}\|_2\|\widehat{b}^{(k)}_{init}\|_2\end{aligned}$$

On the event $\mathcal{G}_4$, we establish

$$\|\omega^{(k)}\|_2 \lesssim \lambda_{\max}(\Sigma^{\mathcal{Q}})\left(1 + \sqrt{\frac{p}{N_{\mathcal{Q}}}} + \frac{p}{N_{\mathcal{Q}}}\right)\left(\|b^{(k)}\|_2 + \sqrt{s\log p/n}\right)$$

With a similar bound for $\|\omega^{(l)}\|_2$, we establish (32). For the setting of known $\Sigma^{\mathcal{Q}}$, on the event $\mathcal{G}_2$ defined in (48), we establish

$$\|\omega^{(k)}\|_2 = \|\Sigma^{\mathcal{Q}}\widehat{b}^{(k)}_{init}\|_2 \lesssim \lambda_{\max}(\Sigma^{\mathcal{Q}})\|\widehat{b}^{(k)}_{init}\|_2 \lesssim \lambda_{\max}(\Sigma^{\mathcal{Q}})\left(\|b^{(k)}\|_2 + \sqrt{s\log p/n}\right).$$

## B.6 Proof of Theorem 3

We will first prove

$$\mathbf{P}\left(\min_{1\leq m\leq M}\|\widehat{\Gamma}^{[m]} - \Gamma^{\mathcal{Q}}\|_F/\sqrt{2} \leq \mathrm{err}_n(M)/\sqrt{n}\right) \geq 1 - \alpha_0 - n^{-c_1} - p^{-c_1}, \tag{71}$$

and then prove (35).

**Proof of** (71)**.** Denote all data by $\mathcal{O}$, that is,

$$\mathcal{O} = \{X^{(l)}, Y^{(l)}\}_{1\leq l\leq L} \cup \{X^{\mathcal{Q}}\}.$$

Define $n = \min_{1\leq l\leq L} n_l$. We rescale the difference as $\widehat{Z} = \sqrt{n}\left(\mathrm{vecl}(\widehat{\Gamma}) - \mathrm{vecl}(\Gamma)\right)$ and the sampled difference as $Z^{[m]} = \sqrt{n}S^{[m]}$. Hence,

$$\widehat{Z} - Z^{[m]} = \sqrt{n}\left(\mathrm{vecl}(\widehat{\Gamma}^{[m]}) - \mathrm{vecl}(\Gamma)\right) \tag{72}$$

The rescaled covariance matrix $\mathbf{Cov}$ and estimated covariance matrix $\widehat{\mathbf{Cov}}$ are defined as

$$\mathbf{Cov} = n\mathbf{V} \quad \text{and} \quad \widehat{\mathbf{Cov}} = n\widehat{\mathbf{V}}.$$

Let $f(Z \mid \mathcal{O})$ denote the conditional density function of $Z^{[m]}$ given the data $\mathcal{O}$, that is,

$$f(Z \mid \mathcal{O}) = \frac{1}{\sqrt{2\pi\det(\widehat{\mathbf{Cov}} + d_0\mathbf{I})}}\exp\left(-\frac{1}{2}Z^\intercal(\widehat{\mathbf{Cov}} + d_0\mathbf{I})^{-1}Z\right).$$

We define the following function to facilitate the proof,

$$g(Z)=\frac{1}{\sqrt{2\pi\det(\mathbf{Cov}+\frac{3}{2}d_0\mathbf{I})}}\exp\left(-\frac{1}{2}Z^{\intercal}(\mathbf{Cov}+\frac{1}{2}d_0\mathbf{I})^{-1}Z\right). \tag{73}$$

We define the following event for the data $\mathcal{O}$,

$$\begin{aligned}
\mathcal{E}_1&=\left\{\|\widehat{\mathbf{Cov}}-\mathbf{Cov}\|_2<d_0/2\right\}\\
\mathcal{E}_2&=\cup_{1\le l,k\le L}\left\{\mathrm{Rem}_{l,k}\text{ satisfies (30)}\right\}\\
\mathcal{E}_3&=\left\{g(\widehat{Z})\cdot\mathbf{1}_{\mathcal{E}_1\cap\mathcal{E}_2}\ge c^*(1-\alpha_0)\right\}
\end{aligned} \tag{74}$$

where $\widehat{Z}=\sqrt{n}\left(\mathrm{vecl}(\widehat{\Gamma})-\mathrm{vecl}(\Gamma)\right)$ is a function of $\mathcal{O}$ and

$$c^*(1-\alpha_0)=\frac{1}{2\sqrt{2\pi}}\prod_{i=1}^{\frac{L(L+1)}{2}}\left[n\cdot\lambda_i(\mathbf{V})+3d_0/2\right]^{-1}\exp\left(-F^{-1}_{\chi^2_r}(1-\alpha_0)\right), \tag{75}$$

with $r$ denoting the rank of $\mathbf{V}$ and $F^{-1}_{\chi^2_r}(1-\alpha_0)$ denoting $1-\alpha_0$ quantile of the $\chi^2$ distribution with degree of freedom $r$.

The following two lemmas show that the event $\mathcal{E}_1\cap\mathcal{E}_2\cap\mathcal{E}_3$ holds with a high probability.

**Lemma 4.** *Suppose that the conditions of Theorem 3 hold, then we have*

$$\mathbf{P}\left(\mathcal{E}_1\cap\mathcal{E}_2\right)\ge 1-\min\{N_{\mathcal{Q}},n,p\}^{-c} \tag{76}$$

*for some positive constant $c>0$.*

**Lemma 5.** *Suppose that the conditions of Theorem 3, then we have*

$$\mathbf{P}\left(\mathcal{E}_3\right)\ge 1-\alpha_0-\mathbf{P}((\mathcal{E}_1\cap\mathcal{E}_2)^c) \tag{77}$$

The proofs of Lemmas 4 and 5 are presented in Sections C.3 and C.4, respectively.

$$\mathbf{Cov}=n\mathbf{V}\quad\text{and}\quad\widehat{\mathbf{Cov}}=n\widehat{\mathbf{V}}.$$

We lower bound the targeted probability in (71) as

$$\begin{aligned}
&\mathbf{P}\left(\min_{1\le m\le M}\|\widehat{\Gamma}^{[m]}-\Gamma^{\mathcal{Q}}\|_F\le\sqrt{2}\mathrm{err}_n(M)/\sqrt{n}\right)\\
&\ge\mathbf{P}\left(\min_{1\le m\le M}\|\widehat{Z}-Z^{[m]}\|_2\le\mathrm{err}_n(M)\right)\\
&\ge\mathbf{E}_{\mathcal{O}}\mathbf{P}\left(\min_{1\le m\le M}\|\widehat{Z}-Z^{[m]}\|_2\le\mathrm{err}_n(M)\mid\mathcal{O}\right)\cdot\mathbf{1}_{\mathcal{O}\in\mathcal{E}_1\cap\mathcal{E}_2\cap\mathcal{E}_3}
\end{aligned} \tag{78}$$

where $\mathbf{P}(\cdot \mid \mathcal{O})$ denotes the conditional probability with respect to the observed data $\mathcal{O}$ and $\mathbf{E}_{\mathcal{O}}$ denotes the expectation taken with respect to the observed data $\mathcal{O}$. Note that

$$\begin{aligned}
&\mathbf{P}\left(\min_{1\leq m\leq M} \|\widehat{Z} - Z^{[m]}\|_2 \leq \mathrm{err}_n(M) \mid \mathcal{O}\right) \\
&= 1 - \mathbf{P}\left(\min_{1\leq m\leq M} \|\widehat{Z} - Z^{[m]}\|_2 \geq \mathrm{err}_n(M) \mid \mathcal{O}\right) \\
&= 1 - \prod_{m=1}^{M}\left[1 - \mathbf{P}\left(\|\widehat{Z} - Z^{[m]}\|_2 \leq \mathrm{err}_n(M) \mid \mathcal{O}\right)\right] \\
&\geq 1 - \exp\left[-M \cdot \mathbf{P}\left(\|\widehat{Z} - Z^{[m]}\|_2 \leq \mathrm{err}_n(M) \mid \mathcal{O}\right)\right],
\end{aligned}$$

where the second equality follows from the conditional independence of $\{Z^{[m]}\}_{1\leq m\leq M}$ given the data $\mathcal{O}$ and the last inequality follows from $1-x \leq e^{-x}$. Hence, we have

$$\begin{aligned}
&\mathbf{P}\left(\min_{1\leq m\leq M} \|\widehat{Z} - Z^{[m]}\|_2 \leq \mathrm{err}_n(M) \mid \mathcal{O}\right) \cdot \mathbf{1}_{\mathcal{O}\in\mathcal{E}_1\cap\mathcal{E}_2\cap\mathcal{E}_3} \\
&\geq \left(1 - \exp\left[-M \cdot \mathbf{P}\left(\|\widehat{Z} - Z^{[m]}\|_2 \leq \mathrm{err}_n(M) \mid \mathcal{O}\right)\right]\right) \cdot \mathbf{1}_{\mathcal{O}\in\mathcal{E}_1\cap\mathcal{E}_2\cap\mathcal{E}_3} \\
&= 1 - \exp\left[-M \cdot \mathbf{P}\left(\|\widehat{Z} - Z^{[m]}\|_2 \leq \mathrm{err}_n(M) \mid \mathcal{O}\right) \cdot \mathbf{1}_{\mathcal{O}\in\mathcal{E}_1\cap\mathcal{E}_2\cap\mathcal{E}_3}\right]
\end{aligned} \tag{79}$$

and it is sufficient to establish an lower bound for

$$\mathbf{P}\left(\|\widehat{Z} - Z^{[m]}\|_2 \leq \mathrm{err}_n(M) \mid \mathcal{O}\right) \cdot \mathbf{1}_{\mathcal{O}\in\mathcal{E}_1\cap\mathcal{E}_2\cap\mathcal{E}_3}. \tag{80}$$

On the event $\mathcal{O} \in \mathcal{E}_1$, we have

$$\mathbf{Cov} + \frac{3}{2}d_0\mathbf{I} \succ \widehat{\mathbf{Cov}} + d_0\mathbf{I} \succ \mathbf{Cov} + \frac{1}{2}d_0\mathbf{I} \tag{81}$$

and

$$f(Z^{[m]} \mid \mathcal{O})\mathbf{1}_{\mathcal{O}\in\mathcal{E}_1} \geq g(Z^{[m]})\mathbf{1}_{\mathcal{O}\in\mathcal{E}_1}.$$

We apply the above inequality and further lower bound the targeted probability in (80) as

$$\begin{aligned}
&\mathbf{P}\left(\|\widehat{Z} - Z^{[m]}\|_2 \leq \mathrm{err}_n(M) \mid \mathcal{O}\right) \cdot \mathbf{1}_{\mathcal{O}\in\mathcal{E}_1\cap\mathcal{E}_2\cap\mathcal{E}_3} \\
&= \int f(Z^{[m]} \mid \mathcal{O})\mathbf{1}_{\left\{\|\widehat{Z} - Z^{[m]}\|_2 \leq \mathrm{err}_n(M)\right\}} dZ^{[m]} \cdot \mathbf{1}_{\mathcal{O}\in\mathcal{E}_1\cap\mathcal{E}_2\cap\mathcal{E}_3} \\
&\geq \int g(Z^{[m]})\mathbf{1}_{\left\{\|\widehat{Z} - Z^{[m]}\|_2 \leq \mathrm{err}_n(M)\right\}} dZ^{[m]} \cdot \mathbf{1}_{\mathcal{O}\in\mathcal{E}_1\cap\mathcal{E}_2\cap\mathcal{E}_3} \\
&= \int g(\widehat{Z})\mathbf{1}_{\left\{\|\widehat{Z} - Z^{[m]}\|_2 \leq \mathrm{err}_n(M)\right\}} dZ^{[m]} \cdot \mathbf{1}_{\mathcal{O}\in\mathcal{E}_1\cap\mathcal{E}_2\cap\mathcal{E}_3} \\
&\quad + \int [g(Z^{[m]}) - g(\widehat{Z})]\mathbf{1}_{\left\{\|\widehat{Z} - Z^{[m]}\|_2 \leq \mathrm{err}_n(M)\right\}} dZ^{[m]} \cdot \mathbf{1}_{\mathcal{O}\in\mathcal{E}_1\cap\mathcal{E}_2\cap\mathcal{E}_3}
\end{aligned} \tag{82}$$

By the definition of $\mathcal{E}_3$, we establish

$$
\begin{aligned}
&\int g(\widehat{Z})\mathbf{1}_{\left\{\|\widehat{Z}-Z^{[m]}\|_2\leq \mathrm{err}_n(M)\right\}}dZ^{[m]}\cdot \mathbf{1}_{\mathcal{O}\in\mathcal{E}_1\cap\mathcal{E}_2\cap\mathcal{E}_3}\\
&\geq c^*(1-\alpha_0)\cdot\int \mathbf{1}_{\left\{\|\widehat{Z}-Z^{[m]}\|_2\leq \mathrm{err}_n(M)\right\}}dZ^{[m]}\cdot \mathbf{1}_{\mathcal{O}\in\mathcal{E}_1\cap\mathcal{E}_2\cap\mathcal{E}_3}\\
&\geq c^*(1-\alpha_0)\cdot C(L)\cdot \mathrm{err}_n(M)^{L(L+1)/2}\cdot \mathbf{1}_{\mathcal{O}\in\mathcal{E}_1\cap\mathcal{E}_2\cap\mathcal{E}_3}
\end{aligned}
\tag{83}
$$

where $C(L)=\frac{\pi^{L(L+1)/4}}{\psi(L(L+1)/4+1)}$ denotes the volume of the unit ball in $\frac{L(L+1)}{2}$-dimension. Note that there exists $t\in(0,1)$ such that

$$g(Z^{[m]})-g(\widehat{Z})=[\nabla g(\widehat{Z}+t(Z^{[m]}-\widehat{Z}))]^{\intercal}(Z^{[m]}-\widehat{Z})$$

with

$$\nabla g(w)=\frac{1}{\sqrt{2\pi\mathrm{det}(\mathbf{Cov}+\frac{3}{2}d_0\mathbf{I})}}\exp\left(-\frac{1}{2}w^{\intercal}(\mathbf{Cov}+\frac{1}{2}d_0\mathbf{I})^{-1}w\right)(\mathbf{Cov}+\frac{1}{2}d_0\mathbf{I})^{-1}w.$$

Since $\lambda_{\min}(\mathbf{Cov}+\frac{1}{2}d_0\mathbf{I})\geq d_0/2$, then $\nabla g$ is bounded and $\left|g(Z^{[m]})-g(\widehat{Z})\right|\leq C\|Z^{[m]}-\widehat{Z}\|_2$ for a positive constant $C>0$. Then we establish

$$
\begin{aligned}
&\left|\int[g(Z^{[m]})-g(\widehat{Z})]\mathbf{1}_{\left\{\|\widehat{Z}-Z^{[m]}\|_2\leq \mathrm{err}_n(M)\right\}}dZ^{[m]}\cdot \mathbf{1}_{\mathcal{O}\in\mathcal{E}_1\cap\mathcal{E}_2\cap\mathcal{E}_3}\right|\\
&\leq C\mathrm{err}_n(M)\cdot\int \mathbf{1}_{\left\{\|\widehat{Z}-Z^{[m]}\|_2\leq \mathrm{err}_n(M)\right\}}dZ^{[m]}\cdot \mathbf{1}_{\mathcal{O}\in\mathcal{E}_1\cap\mathcal{E}_2\cap\mathcal{E}_3}\\
&= C\mathrm{err}_n(M)\cdot C(L)\cdot \mathrm{err}_n(M)^{L(L+1)/2}\cdot \mathbf{1}_{\mathcal{O}\in\mathcal{E}_1\cap\mathcal{E}_2\cap\mathcal{E}_3}.
\end{aligned}
\tag{84}
$$

By assuming

$$C\mathrm{err}_n(M)\leq \frac{1}{2}c^*(1-\alpha_0),$$

we combine (82), (83) and (84) and obtain

$$\mathbf{P}\left(\|\widehat{Z}-Z^{[m]}\|_2\leq \mathrm{err}_n(M)\mid\mathcal{O}\right)\cdot\mathbf{1}_{\mathcal{O}\in\mathcal{E}_1\cap\mathcal{E}_2\cap\mathcal{E}_3}\geq\frac{1}{2}c^*(1-\alpha_0)C(L)\mathrm{err}_n(M)^{L(L+1)/2}\cdot\mathbf{1}_{\mathcal{O}\in\mathcal{E}_1\cap\mathcal{E}_2\cap\mathcal{E}_3}$$

Together with (79), we establish

$$
\begin{aligned}
&\mathbf{P}\left(\min_{1\leq m\leq M}\|\widehat{Z}-Z^{[m]}\|_2\leq \mathrm{err}_n(M)\mid\mathcal{O}\right)\cdot\mathbf{1}_{\mathcal{O}\in\mathcal{E}_1\cap\mathcal{E}_2\cap\mathcal{E}_3}\\
&\geq 1-\exp\left[-M\cdot\frac{1}{2}c^*(1-\alpha_0)C(L)\mathrm{err}_n(M)^{L(L+1)/2}\cdot\mathbf{1}_{\mathcal{O}\in\mathcal{E}_1\cap\mathcal{E}_2\cap\mathcal{E}_3}\right]\\
&=\left(1-\exp\left[-M\cdot\frac{1}{2}c^*(1-\alpha_0)C(L)\mathrm{err}_n(M)^{L(L+1)/2}\right]\right)\cdot\mathbf{1}_{\mathcal{O}\in\mathcal{E}_1\cap\mathcal{E}_2\cap\mathcal{E}_3}
\end{aligned}
\tag{85}
$$

Together with (78), we have

$$\mathbf{P}\left(\min_{1\leq m\leq M}\|\widehat{Z}-Z^{[m]}\|_2\leq \mathrm{err}_n(M)\right)$$
$$\geq\left(1-\exp\left[-M\cdot\frac{1}{2}c^*(1-\alpha_0)C(L)\mathrm{err}_n(M)^{L(L+1)/2}\right]\right)\mathbf{P}\left(\mathcal{E}_1\cap\mathcal{E}_2\cap\mathcal{E}_3\right)$$

Since $C^*(L,\alpha_0)$ defined in (34) satisfies $C^*(L,\alpha_0)=C(L)c^*(1-\alpha_0)$, we choose

$$\mathrm{err}_n(M)=\left[\frac{2\log n}{C^*(L,\alpha_0)M}\right]^{\frac{2}{L(L+1)}}$$

and establish

$$\mathbf{P}\left(\min_{1\leq m\leq M}\|\widehat{Z}-Z^{[m]}\|_2\leq \mathrm{err}_n(M)\right)\geq(1-n^{-1})\cdot\mathbf{P}\left(\mathcal{E}_1\cap\mathcal{E}_2\cap\mathcal{E}_3\right).$$

We further apply Lemmas 4 and 5 and establish

$$\begin{aligned}\mathbf{P}\left(\min_{1\leq m\leq M}\|\widehat{Z}-Z^{[m]}\|_2\leq \mathrm{err}_n(M)\right)&\geq(1-n^{-1})(1-\alpha_0-\exp(-cn)-p^{-c})\\&\geq 1-\alpha_0-n^{-c_1}-p^{-c_1}.\end{aligned}$$

**Proof of** (35)**.** We shall quantify the error of the ridge-type sampled weights $\widehat{\gamma}_\delta^{[m]}$ by the corresponding error of $\|\widehat{\Gamma}^{[m]}-\Gamma^{\mathcal{Q}}\|_2$.

**Lemma 6.** *If* $\|\widehat{\Gamma}^{[m]}-\Gamma^{\mathcal{Q}}\|_2\leq(\lambda_{\min}(\Gamma^{\mathcal{Q}})+\delta)/2$, *then*

$$\|\widehat{\gamma}_\delta^{[m]}-\gamma_\delta^*\|_2\leq\frac{2\|\widehat{\Gamma}^{[m]}-\Gamma^{\mathcal{Q}}\|_F}{\lambda_{\min}(\Gamma^{\mathcal{Q}})+\delta}\|\gamma_\delta^*\|_2\tag{86}$$

*where* $\widehat{\gamma}_\delta^{[m]}$ *and* $\gamma_\delta^*=\gamma_\delta^*(\mathcal{Q})$ *are defined in* (24) *and* (7), *respectively.*

The above lemma will be proved in Section B.6.1.

It follows from (71) that, with probability larger than $1-\alpha_0-n^{-c_1}-p^{-c_1}$, there exists $1\leq m^*\leq M$ such that $\|\widehat{\Gamma}^{[m^*]}-\Gamma^{\mathcal{Q}}\|_F\leq\sqrt{2}\mathrm{err}(M)/\sqrt{n}$. Under the assumption $\sqrt{2}\mathrm{err}(M)/\sqrt{n}\leq(\lambda_{\min}(\Gamma^{\mathcal{Q}})+\delta)/2$, we apply (86), together with $\|\gamma_\delta^*\|_2\leq\|\gamma_\delta^*\|_1\leq 1$, to establish (35).

### B.6.1 Proof of Lemma 6

Recall that $\gamma^*=\gamma^*_{\delta=0}$ and $\widehat{\gamma}^{[m]}=\widehat{\gamma}^{[m]}_{\delta=0}$ where $\gamma^*_\delta$ is defined in (6) and $\widehat{\gamma}^{[m]}_\delta$ is defined in (24). We first consider the setting with $\delta=0$ and control $\|\gamma^*-\widehat{\gamma}^{[m]}\|_2$; then we extend the proof to control $\|\gamma^*_\delta - \widehat{\gamma}^{[m]}_\delta\|_2$ for $\delta \geq 0$. By the definition of $\gamma^*$ in (6) with $\delta = 0$, for any $t \in (0,1)$, we have

$$(\gamma^*)^\intercal \Gamma^{\mathcal{Q}} \gamma^* \leq \left[\gamma^* + t(\widehat{\gamma}^{[m]} - \gamma^*)\right]^\intercal \Gamma^{\mathcal{Q}} \left[\gamma^* + t(\widehat{\gamma}^{[m]} - \gamma^*)\right]$$

and hence

$$0 \leq 2t(\gamma^*)^\intercal \Gamma^{\mathcal{Q}} (\widehat{\gamma}^{[m]} - \gamma^*) + t^2 (\widehat{\gamma}^{[m]} - \gamma^*)^\intercal \Gamma^{\mathcal{Q}} (\widehat{\gamma}^{[m]} - \gamma^*)$$

By taking $t \to 0+$, we have

$$(\gamma^*)^\intercal \Gamma^{\mathcal{Q}} (\widehat{\gamma}^{[m]} - \gamma^*) \geq 0. \tag{87}$$

By this definition of (24) with $\delta = 0$, for any $t \in (0,1)$, we have

$$(\widehat{\gamma}^{[m]})^\intercal \widehat{\Gamma}^{[m]}_+ \widehat{\gamma}^{[m]} \leq \left[\widehat{\gamma}^{[m]} + t(\gamma^* - \widehat{\gamma}^{[m]})\right]^\intercal \widehat{\Gamma}^{[m]}_+ \left[\widehat{\gamma}^{[m]} + t(\gamma^* - \widehat{\gamma}^{[m]})\right]$$

This gives us

$$2(\gamma^*)^\intercal \widehat{\Gamma}^{[m]}_+ (\gamma^* - \widehat{\gamma}^{[m]}) + (t-2)(\gamma^* - \widehat{\gamma}^{[m]})^\intercal \widehat{\Gamma}^{[m]}_+ (\gamma^* - \widehat{\gamma}^{[m]}) \geq 0 \tag{88}$$

Since $2 - t > 0$, we have

$$(\gamma^* - \widehat{\gamma}^{[m]})^\intercal \widehat{\Gamma}^{[m]}_+ (\gamma^* - \widehat{\gamma}^{[m]}) \leq \frac{2}{2-t} (\gamma^*)^\intercal \widehat{\Gamma}^{[m]}_+ (\gamma^* - \widehat{\gamma}^{[m]}). \tag{89}$$

It follows from (87) that

$$\begin{aligned}(\gamma^*)^\intercal \widehat{\Gamma}^{[m]}_+ (\gamma^* - \widehat{\gamma}^{[m]}) &= (\gamma^*)^\intercal \Gamma^{\mathcal{Q}} (\gamma^* - \widehat{\gamma}^{[m]}) + (\gamma^*)^\intercal (\widehat{\Gamma}^{[m]}_+ - \Gamma^{\mathcal{Q}})(\gamma^* - \widehat{\gamma}^{[m]}) \\ &\leq (\gamma^*)^\intercal (\widehat{\Gamma}^{[m]}_+ - \Gamma^{\mathcal{Q}})(\gamma^* - \widehat{\gamma}^{[m]})\end{aligned}$$

Combined with (89), we have

$$\begin{aligned}(\gamma^* - \widehat{\gamma}^{[m]})^\intercal \widehat{\Gamma}^{[m]}_+ (\gamma^* - \widehat{\gamma}^{[m]}) &\leq \frac{2}{2-t} (\gamma^*)^\intercal (\widehat{\Gamma}^{[m]}_+ - \Gamma^{\mathcal{Q}})(\gamma^* - \widehat{\gamma}^{[m]}) \\ &\leq \frac{2\|\gamma^*\|_2}{2-t} \|\widehat{\Gamma}^{[m]}_+ - \Gamma^{\mathcal{Q}}\|_2 \|\gamma^* - \widehat{\gamma}^{[m]}\|_2 \\ &\leq \frac{2\|\gamma^*\|_2}{2-t} \|\widehat{\Gamma}^{[m]}_+ - \Gamma^{\mathcal{Q}}\|_F \|\gamma^* - \widehat{\gamma}^{[m]}\|_2\end{aligned} \tag{90}$$

By the definition of $(\widehat{\Gamma}^{[m]} + \delta \mathrm{I})_+$ and $\Gamma^{\mathcal{Q}} + \delta \mathrm{I}$ is positive semi-definite, we have

$$\|(\widehat{\Gamma}^{[m]} + \delta \mathrm{I})_+ - (\Gamma^{\mathcal{Q}} + \delta \mathrm{I})\|_F \leq \|(\widehat{\Gamma}^{[m]} + \delta \mathrm{I}) - (\Gamma^{\mathcal{Q}} + \delta \mathrm{I})\|_F = \|\widehat{\Gamma}^{[m]} - \Gamma^{\mathcal{Q}}\|_F. \tag{91}$$

Since $\gamma^*_\delta = \arg\min_{\gamma \in \Delta^L} \gamma^\intercal \left(\Gamma^{\mathcal{Q}} + \delta \mathrm{I}\right) \gamma$ and $\widehat{\gamma}^{[m]}_\delta = \arg\min_{\gamma \in \Delta^L} \gamma^\intercal \left(\widehat{\Gamma}^{\mathcal{Q}} + \delta \mathrm{I}\right) \gamma$, we can apply the same argument for (90) by replacing respectively $\Gamma^{\mathcal{Q}}$ and $\widehat{\Gamma}^{[m]}_+$ by $\Gamma^{\mathcal{Q}} + \delta \mathrm{I}$ and $(\widehat{\Gamma}^{[m]} + \delta \mathrm{I})_+$ and then apply (91) to establish

$$(\widehat{\gamma}^{[m]}_\delta - \gamma^*_\delta)^\intercal \left(\widehat{\Gamma}^{[m]} + \delta \mathrm{I}\right)_+ (\widehat{\gamma}^{[m]}_\delta - \gamma^*_\delta) \leq \frac{2\|\gamma^*_\delta\|_2}{2-t} \|\widehat{\Gamma}^{[m]} - \Gamma^{\mathcal{Q}}\|_F \|\widehat{\gamma}^{[m]}_\delta - \gamma^*_\delta\|_2$$

We apply the above bound by taking $t \to 0+$ and establish

$$\|\widehat{\gamma}_\delta^{[m]} - \gamma_\delta^*\|_2 \leq \frac{\|\widehat{\Gamma}^{[m]} - \Gamma^{\mathcal{Q}}\|_F}{(\lambda_{\min}(\widehat{\Gamma}^{[m]}) + \delta)_+} \|\gamma_\delta^*\|_2$$

Since $\left|\lambda_{\min}(\widehat{\Gamma}^{[m]}) - \lambda_{\min}(\Gamma^{\mathcal{Q}})\right| \leq \|\widehat{\Gamma}^{[m]} - \Gamma^{\mathcal{Q}}\|_2$, if $\lambda_{\min}(\Gamma^{\mathcal{Q}}) + \delta > 2\|\widehat{\Gamma}^{[m]} - \Gamma^{\mathcal{Q}}\|_2$, then we have $\lambda_{\min}(\widehat{\Gamma}^{\mathcal{Q}}) + \delta > \frac{1}{2}(\lambda_{\min}(\Gamma^{\mathcal{Q}}) + \delta)$, and

$$\|\widehat{\gamma}_\delta^{[m]} - \gamma_\delta^*\|_2 \leq \frac{2\|\widehat{\Gamma}^{[m]} - \Gamma^{\mathcal{Q}}\|_F}{\lambda_{\min}(\Gamma^{\mathcal{Q}}) + \delta} \|\gamma_\delta^*\|_2$$

## B.7 Proof of Proposition 6

The proof of Proposition 6 is similar way to that of Theorem 3 in Section B.6. We define similar events as those in (74) by modifying the definition of $\mathcal{E}_2$ as

$$\mathcal{E}_2 = \cup_{1\leq l,k\leq L} \left\{\text{Rem}_{l,k} \text{ satisfies (41)}\right\}$$

and re-defining $\mathbf{Cov}$ and $\widehat{\mathbf{Cov}}$ as

$$\mathbf{Cov} = n(\mathbf{V}^{(1)} + \mathbf{V}^{(2)}) \quad \text{and} \quad \widehat{\mathbf{Cov}} = n(\widehat{\mathbf{V}}^{(1)} + \widehat{\mathbf{V}}^{(2)}).$$

where $\mathbf{V}^{(1)}$ and $\mathbf{V}^{(2)}$ are defined in (39) and (40), respectively and $\widehat{\mathbf{V}}^{(1)}$ and $\widehat{\mathbf{V}}^{(2)}$ are defined in (43) and (44), respectively. The key is to establish the high probability event lemmas 4 and 5. The proof of 4 follows the same argument as that of the general covariate shift setting in Section C.3 and is omitted here. We shall present a separate proof of Lemma 5 for the no covariate shift setting in Section C.4.

## B.8 Proof of Theorem 4

We introduce the following high-probability events to facilitate the discussion.

$$\begin{aligned}
\mathcal{E}_4 &= \left\{\frac{1}{n_l}[\widehat{v}^{(l)}]^\intercal (X^{(l)})^\intercal X^{(l)} \widehat{v}^{(l)} \asymp \|x_{\text{new}}\|_2 \quad \text{for } 1 \leq l \leq L\right\} \\
\mathcal{E}_5 &= \left\{\frac{\left|\sum_{l=1}^L \left([\widehat{\gamma}_\delta^{[m]}]_l - [\gamma_\delta^*]_l\right) \widehat{x_{\text{new}}^\intercal b^{(l)}}\right|}{\sqrt{\sum_{l=1}^L [\gamma_\delta^*]_l^2 \frac{\sigma_l^2}{n_l^2} [\widehat{v}^{(l)}]^\intercal (X^{(l)})^\intercal X^{(l)} \widehat{v}^{(l)}}} \lesssim \sqrt{n} \|\widehat{\gamma}_\delta^{[m]} - \gamma_\delta^*\|_2\right\} \\
\mathcal{E}_6 &= \left\{\|\widehat{\gamma}_\delta^{[m^*]} - \gamma_\delta^*\|_2 \leq \frac{2\sqrt{2}\text{err}_n(M)}{\lambda_{\min}(\Gamma^{\mathcal{Q}}) + \delta} \cdot \frac{1}{\sqrt{n}} \quad \text{for some } 1 \leq m^* \leq M\right\}
\end{aligned} \tag{92}$$

We apply Lemma 1 of [11] and establish that $\mathbf{P}\left(\mathcal{E}_4\right) \geq 1-p^{-c}$. On the event $\mathcal{E}_4$, we have

$$\sqrt{\sum_{l=1}^{L}[\gamma_\delta^*]_l^2 \frac{\sigma_l^2}{n_l^2}[\widehat{v}^{(l)}]^\intercal (X^{(l)})^\intercal X^{(l)}\widehat{v}^{(l)}} \asymp \frac{\|\gamma_\delta^*\|_2 \|x_{\rm new}\|_2}{\sqrt{n}} \geq \frac{1}{\sqrt{L}} \cdot \frac{\|x_{\rm new}\|_2}{\sqrt{n}}. \tag{93}$$

Note that

$$\begin{aligned}\left|\sum_{l=1}^{L}\left([\widehat{\gamma}_\delta^{[m]}]_l - [\gamma_\delta^*]_l\right)\widehat{x_{\rm new}^\intercal b^{(l)}}\right| &\leq \|\widehat{\gamma}_\delta^{[m]} - \gamma_\delta^*\|_2 \sqrt{\sum_{l=1}^{L}[\widehat{x_{\rm new}^\intercal b^{(l)}}]^2} \\ &\lesssim \|\widehat{\gamma}_\delta^{[m]} - \gamma_\delta^*\|_2 \sqrt{\sum_{l=1}^{L}[\widehat{x_{\rm new}^\intercal b^{(l)}} - x_{\rm new}^\intercal b^{(l)}]^2 + [x_{\rm new}^\intercal b^{(l)}]^2}\end{aligned} \tag{94}$$

By Theorem 2 of [11] and (93), with probability larger than $1-\min\{n,p\}^{-c}$ for some positive constant $c>0$,

$$\begin{aligned}&\frac{\left|\sum_{l=1}^{L}\left([\widehat{\gamma}_\delta^{[m]}]_l - [\gamma_\delta^*]_l\right)\widehat{x_{\rm new}^\intercal b^{(l)}}\right|}{\sqrt{\sum_{l=1}^{L}[\gamma_\delta^*]_l^2 \frac{\sigma_l^2}{n_l^2}[\widehat{v}^{(l)}]^\intercal (X^{(l)})^\intercal X^{(l)}\widehat{v}^{(l)}}} \\ &\lesssim \|\widehat{\gamma}_\delta^{[m]} - \gamma_\delta^*\|_2 \cdot \frac{\sqrt{L}\cdot\left(\log n \cdot \frac{\|x_{\rm new}\|_2}{\sqrt{n}} + \max_{1\leq l\leq L}|x_{\rm new}^\intercal b^{(l)}|\right)}{\frac{1}{\sqrt{L}}\cdot\frac{\|x_{\rm new}\|_2}{\sqrt{n}}} \lesssim \sqrt{n}\|\widehat{\gamma}_\delta^{[m]} - \gamma_\delta^*\|_2\end{aligned} \tag{95}$$

where the last inequality follows from the bounded $\|b^{(l)}\|_2$ and finite $L$. This implies that $\mathbf{P}(\mathcal{E}_5) \geq 1-\min\{n,p\}^{-c}$. It follows from (35) of Theorem 3 that $\mathbf{P}(\mathcal{E}_6) \geq 1-\alpha_0-n^{-c_1}-p^{-c_1}$ for some positive constant $c_1>0$.

**Coverage property.** It follows from Theorem 2 of [11] that

$$\frac{\sum_{l=1}^{L}[\gamma_\delta^*]_l[\widehat{x_{\rm new}^\intercal b^{(l)}} - x_{\rm new}^\intercal b^{(l)}]}{\sqrt{\sum_{l=1}^{L}[\gamma_\delta^*]_l^2 \frac{\sigma_l^2}{n_l^2}[\widehat{v}^{(l)}]^\intercal (X^{(l)})^\intercal X^{(l)}\widehat{v}^{(l)}}} \xrightarrow{d} N(0,1). \tag{96}$$

We shall take $m^*$ as in Theorem 3. By the CI definition in (28), we have

$$\begin{aligned}&\mathbf{P}\left(x_{\rm new}^\intercal\beta_\delta^* \not\in {\rm CI}_\alpha\left(x_{\rm new}^\intercal\beta_\delta^*\right)\right) \leq \mathbf{P}\left(x_{\rm new}^\intercal\beta_\delta^* \not\in {\rm Int}_\alpha^{[m^*]}(x_{\rm new})\right) \\ &= \mathbf{P}\left(\frac{|\widehat{x_{\rm new}^\intercal\beta}^{[m^*]} - x_{\rm new}^\intercal\beta_\delta^*|}{\sqrt{\sum_{l=1}^{L}[\gamma_\delta^*]_l^2 \frac{\sigma_l^2}{n_l^2}[\widehat{v}^{(l)}]^\intercal (X^{(l)})^\intercal X^{(l)}\widehat{v}^{(l)}}} \geq 1.01\cdot z_{1-\alpha/2}\sqrt{\frac{\sum_{l=1}^{L}[\widehat{\gamma}_\delta^{[m^*]}]_l^2 \frac{\widehat{\sigma}_l^2}{n_l^2}[\widehat{v}^{(l)}]^\intercal (X^{(l)})^\intercal X^{(l)}\widehat{v}^{(l)}}{\sum_{l=1}^{L}[\gamma_\delta^*]_l^2 \frac{\sigma_l^2}{n_l^2}[\widehat{v}^{(l)}]^\intercal (X^{(l)})^\intercal X^{(l)}\widehat{v}^{(l)}}}\right)\end{aligned}$$

By applying (36) with $m=m^*$, we further upper bound the above inequality as

$$
\begin{aligned}
&\mathbf{P}\left(x_{\text{new}}^{\intercal}\beta_{\delta}^{*} \notin \mathrm{CI}_{\alpha}\left(x_{\text{new}}^{\intercal}\beta_{\delta}^{*}\right)\right) \\
&\leq \mathbf{P}\left(\frac{|\sum_{l=1}^{L}[\gamma_{\delta}^{*}]_l \cdot (\widehat{x_{\text{new}}^{\intercal}b^{(l)}} - x_{\text{new}}^{\intercal}b^{(l)})|}{\sqrt{\sum_{l=1}^{L}[\gamma_{\delta}^{*}]_l^2 \frac{\sigma_l^2}{n_l^2}[\widehat{v}^{(l)}]^{\intercal}(X^{(l)})^{\intercal}X^{(l)}\widehat{v}^{(l)}}} \geq z_{1-\alpha/2}\sqrt{\frac{\sum_{l=1}^{L}[\widehat{\gamma}_{\delta}^{[m^*]}]_l^2 \frac{\widehat{\sigma}_l^2}{n_l^2}[\widehat{v}^{(l)}]^{\intercal}(X^{(l)})^{\intercal}X^{(l)}\widehat{v}^{(l)}}{\sum_{l=1}^{L}[\gamma_{\delta}^{*}]_l^2 \frac{\sigma_l^2}{n_l^2}[\widehat{v}^{(l)}]^{\intercal}(X^{(l)})^{\intercal}X^{(l)}\widehat{v}^{(l)}}}\right) \\
&+ \mathbf{P}\left(\frac{|\sum_{l=1}^{L}([\widehat{\gamma}_{\delta}^{[m^*]}]_l - [\gamma_{\delta}^{*}]_l) \cdot \widehat{x_{\text{new}}^{\intercal}b^{(l)}}|}{\sqrt{\sum_{l=1}^{L}[\gamma_{\delta}^{*}]_l^2 \frac{\sigma_l^2}{n_l^2}[\widehat{v}^{(l)}]^{\intercal}(X^{(l)})^{\intercal}X^{(l)}\widehat{v}^{(l)}}} \geq 0.01 \cdot z_{1-\alpha/2}\sqrt{\frac{\sum_{l=1}^{L}[\widehat{\gamma}_{\delta}^{[m^*]}]_l^2 \frac{\widehat{\sigma}_l^2}{n_l^2}[\widehat{v}^{(l)}]^{\intercal}(X^{(l)})^{\intercal}X^{(l)}\widehat{v}^{(l)}}{\sum_{l=1}^{L}[\gamma_{\delta}^{*}]_l^2 \frac{\sigma_l^2}{n_l^2}[\widehat{v}^{(l)}]^{\intercal}(X^{(l)})^{\intercal}X^{(l)}\widehat{v}^{(l)}}}\right)
\end{aligned} \tag{97}
$$

Since $\mathrm{err}_n(M) \ll \lambda_{\min}(\Gamma^{\mathcal{Q}}) + \delta$, on the event $\mathcal{G}_3 \cap \mathcal{E}_4 \cap \mathcal{E}_5 \cap \mathcal{E}_6$ with $\mathcal{G}_3$ defined in (48) and $\mathcal{E}_4, \mathcal{E}_5$ and $\mathcal{E}_6$ defined in (92), we have

$$
\frac{|\sum_{l=1}^{L}([\widehat{\gamma}_{\delta}^{[m^*]}]_l - [\gamma_{\delta}^{*}]_l) \cdot \widehat{x_{\text{new}}^{\intercal}b^{(l)}}|}{\sqrt{\sum_{l=1}^{L}[\gamma_{\delta}^{*}]_l^2 \frac{\sigma_l^2}{n_l^2}[\widehat{v}^{(l)}]^{\intercal}(X^{(l)})^{\intercal}X^{(l)}\widehat{v}^{(l)}}} \leq 0.01 \cdot z_{1-\alpha/2}\sqrt{\frac{\sum_{l=1}^{L}[\widehat{\gamma}_{\delta}^{[m^*]}]_l^2 \frac{\widehat{\sigma}_l^2}{n_l^2}[\widehat{v}^{(l)}]^{\intercal}(X^{(l)})^{\intercal}X^{(l)}\widehat{v}^{(l)}}{\sum_{l=1}^{L}[\gamma_{\delta}^{*}]_l^2 \frac{\sigma_l^2}{n_l^2}[\widehat{v}^{(l)}]^{\intercal}(X^{(l)})^{\intercal}X^{(l)}\widehat{v}^{(l)}}}.
$$

This implies

$$
\begin{aligned}
&\mathbf{P}\left(\frac{|\sum_{l=1}^{L}([\widehat{\gamma}_{\delta}^{[m^*]}]_l - [\gamma_{\delta}^{*}]_l) \cdot \widehat{x_{\text{new}}^{\intercal}b^{(l)}}|}{\sqrt{\sum_{l=1}^{L}[\gamma_{\delta}^{*}]_l^2 \frac{\sigma_l^2}{n_l^2}[\widehat{v}^{(l)}]^{\intercal}(X^{(l)})^{\intercal}X^{(l)}\widehat{v}^{(l)}}} \leq 0.01 \cdot z_{1-\alpha/2}\sqrt{\frac{\sum_{l=1}^{L}[\widehat{\gamma}_{\delta}^{[m^*]}]_l^2 \frac{\widehat{\sigma}_l^2}{n_l^2}[\widehat{v}^{(l)}]^{\intercal}(X^{(l)})^{\intercal}X^{(l)}\widehat{v}^{(l)}}{\sum_{l=1}^{L}[\gamma_{\delta}^{*}]_l^2 \frac{\sigma_l^2}{n_l^2}[\widehat{v}^{(l)}]^{\intercal}(X^{(l)})^{\intercal}X^{(l)}\widehat{v}^{(l)}}}\right) \\
&\geq \mathbf{P}\left(\mathcal{G}_3 \cap \mathcal{E}_4 \cap \mathcal{E}_5 \cap \mathcal{E}_6\right) \geq 1 - \alpha_0 - n^{-c_1} - p^{-c_1}.
\end{aligned} \tag{98}
$$

Since the events $\mathcal{G}_3$ and $\mathcal{E}_4$ hold with high probability, we have

$$
\sqrt{\frac{\sum_{l=1}^{L}[\widehat{\gamma}_{\delta}^{[m^*]}]_l^2 \frac{\widehat{\sigma}_l^2}{n_l^2}[\widehat{v}^{(l)}]^{\intercal}(X^{(l)})^{\intercal}X^{(l)}\widehat{v}^{(l)}}{\sum_{l=1}^{L}[\gamma_{\delta}^{*}]_l^2 \frac{\sigma_l^2}{n_l^2}[\widehat{v}^{(l)}]^{\intercal}(X^{(l)})^{\intercal}X^{(l)}\widehat{v}^{(l)}}} \xrightarrow{p} 1.
$$

Together with (96), we establish

$$
\mathbf{P}\left(\frac{|\sum_{l=1}^{L}[\gamma_{\delta}^{*}]_l \cdot (\widehat{x_{\text{new}}^{\intercal}b^{(l)}} - x_{\text{new}}^{\intercal}b^{(l)})|}{\sqrt{\sum_{l=1}^{L}[\gamma_{\delta}^{*}]_l^2 \frac{\sigma_l^2}{n_l^2}[\widehat{v}^{(l)}]^{\intercal}(X^{(l)})^{\intercal}X^{(l)}\widehat{v}^{(l)}}} \geq z_{1-\alpha/2}\sqrt{\frac{\sum_{l=1}^{L}[\widehat{\gamma}_{\delta}^{[m^*]}]_l^2 \frac{\widehat{\sigma}_l^2}{n_l^2}[\widehat{v}^{(l)}]^{\intercal}(X^{(l)})^{\intercal}X^{(l)}\widehat{v}^{(l)}}{\sum_{l=1}^{L}[\gamma_{\delta}^{*}]_l^2 \frac{\sigma_l^2}{n_l^2}[\widehat{v}^{(l)}]^{\intercal}(X^{(l)})^{\intercal}X^{(l)}\widehat{v}^{(l)}}}\right) \to \alpha.
$$

Combined with (97) and (98), we have

$$
\lim_{n,p\to\infty} \mathbf{P}\left(x_{\text{new}}^{\intercal}\beta_{\delta}^{*} \notin \mathrm{CI}_{\alpha}\left(x_{\text{new}}^{\intercal}\beta_{\delta}^{*}\right)\right) \leq \alpha + \alpha_0.
$$

**Confidence interval length.** We control the length by the decomposition (36) for $1 \leq m \leq M$. On the event $\mathcal{E}_4$, we have

$$\sqrt{\sum_{l=1}^{L} [\widehat{\gamma}_{\delta}^{[m]}]_l^2 \frac{\sigma_l^2}{n_l^2} [\widehat{v}^{(l)}]^{\intercal} (X^{(l)})^{\intercal} X^{(l)} \widehat{v}^{(l)}} \asymp \frac{\|\widehat{\gamma}_{\delta}^{[m]}\|_2 \|x_{\rm new}\|_2}{\sqrt{n}} \lesssim \frac{\|x_{\rm new}\|_2}{\sqrt{n}}. \tag{99}$$

Note that with probability larger than $1 - M^{-c}$, we have

$$\left|\widehat{\Gamma}_{l,k}^{[m]} - \Gamma_{l,k}^{\mathcal{Q}}\right| \lesssim \sqrt{d_0/n}\sqrt{2\log M}$$

where the high probability bound follows from Theorem 2 and Proposition 3. As a consequence, for a finite $L$, with probability larger than $1 - M^{-c}$,

$$\max_{1\leq m\leq M} \|\widehat{\Gamma}^{[m]} - \Gamma^{\mathcal{Q}}\|_2 \lesssim \sqrt{L \cdot d_0/n} \cdot \sqrt{\log M}$$

By Lemma 6, if $\sqrt{L \cdot d_0/n}\sqrt{\log M} \ll (\lambda_{\min}(\Gamma^{\mathcal{Q}}) + \delta)/2$, then

$$\max_{1\leq m\leq M} \|\widehat{\gamma}_{\delta}^{[m]} - \gamma_{\delta}^{*}\|_2 \lesssim \frac{2\sqrt{L \cdot d_0/n} \cdot \sqrt{\log M}}{\lambda_{\min}(\Gamma^{\mathcal{Q}}) + \delta} \|\gamma_{\delta}^{*}\|_2 \tag{100}$$

Together with (94) and (95), we establish

$$\left|\sum_{l=1}^{L} \left([\widehat{\gamma}_{\delta}^{[m]}]_l - [\gamma_{\delta}^{*}]_l\right) \widehat{x_{\rm new}^{\intercal} b^{(l)}}\right| \lesssim \frac{\sqrt{L \cdot d_0/n} \cdot \sqrt{\log M}}{2(\lambda_{\min}(\Gamma^{\mathcal{Q}}) + \delta)} \cdot \sqrt{L} \cdot (\log n \cdot \frac{\|x_{\rm new}\|_2}{\sqrt{n}} + \max_{1\leq l\leq L} |x_{\rm new}^{\intercal} b^{(l)}|) \tag{101}$$

By the decomposition (36), we establish (37) by combining (99) and (101).

## B.9 Proof of Proposition 1

We now supply a proof of Proposition 1, whose original proof can be found in [29]. Let $\mathbb{H}$ denotes the convex hull of the set $\mathbb{B} = \{b^{(1)}, \ldots, b^{(L)}\}$. By the linear form of $b$, we have

$$\begin{aligned}
\beta^* &= \arg\max_{\beta \in \mathbb{R}^p} \min_{b \in \mathbb{B}} \left[2b^{\intercal}\Sigma^{\mathcal{Q}}\beta - \beta^{\intercal}\Sigma^{\mathcal{Q}}\beta\right] \\
&= \arg\max_{\beta \in \mathbb{R}^p} \min_{b \in \mathbb{H}} \left[2b^{\intercal}\Sigma^{\mathcal{Q}}\beta - \beta^{\intercal}\Sigma^{\mathcal{Q}}\beta\right]
\end{aligned}$$

We decompose $\Sigma^{\mathcal{Q}} = C^{\intercal}C$ such that $C$ is invertible. Define $\widetilde{\mathbb{H}} = C^{-1}\mathbb{H}$. Then we have $\beta^* = C^{-1}\xi^*$ with

$$\xi^* = \arg\max_{\xi \in \mathbb{R}^p} \min_{u \in \widetilde{\mathbb{H}}} [2u^{\intercal}\xi - \xi^{\intercal}\xi] \tag{102}$$

If we interchange min and max in the above equation, then we have

$$\xi^* = \arg\min_{\xi \in \widetilde{\mathbb{H}}} \xi^\intercal \xi \tag{103}$$

We will justify this inter-change by showing that the solution $\xi^*$ defined in (103) is the solution to (102). For any $\nu \in [0,1]$ and $\mu \in \widetilde{\mathbb{H}}$, we use the fact $\xi^* + \nu(\mu - \xi^*) \in \widetilde{\mathbb{H}}$ and obtain

$$\|\xi^* + \nu(\mu - \xi^*)\|_2^2 \geq \|\xi^*\|_2^2$$

This leads to

$$(\xi^*)^\intercal \mu - (\xi^*)^\intercal \xi^* \geq 0$$

and hence

$$2(\xi^*)^\intercal \mu - (\xi^*)^\intercal \xi^* \geq (\xi^*)^\intercal \xi^*, \quad \text{for any} \quad \mu \in \widetilde{\mathbb{H}}$$

By taking $\xi$ as $\xi^*$ in the optimization problem (102), we have

$$\max_{\xi \in \mathbb{R}^p} \min_{u \in \widetilde{\mathbb{H}}} [2u^\intercal \xi - \xi^\intercal \xi] \geq \min_{u \in \widetilde{\mathbb{H}}} [2u^\intercal \xi^* - [\xi^*]^\intercal \xi^*] \geq (\xi^*)^\intercal \xi^*.$$

In (102), we take $u = \xi^*$, then we have

$$\max_{\xi \in \mathbb{R}^p} \min_{u \in \widetilde{\mathbb{H}}} [2u^\intercal \xi - \xi^\intercal \xi] \leq \max_{\xi \in \mathbb{R}^p} [2[\xi^*]^\intercal \xi - \xi^\intercal \xi] = (\xi^*)^\intercal \xi^*$$

By matching the above two bounds, $\xi^*$ is the optimal solution to (102) and

$$\max_{\xi \in \mathbb{R}^p} \min_{u \in \widetilde{\mathbb{H}}} [2u^\intercal \xi - \xi^\intercal \xi] = [\xi^*]^\intercal \xi^*.$$

Since $\beta^* = C^{-1} \xi^*$ and $\Sigma^{\mathcal{Q}} = C^\intercal C$, we have

$$\beta^* = \arg\min_{\beta \in \mathbb{H}} \beta^\intercal \Sigma^{\mathcal{Q}} \beta \tag{104}$$

and

$$\max_{\beta \in \mathbb{R}^p} \min_{b \in \mathbb{H}} \left[2b^\intercal \Sigma^{\mathcal{Q}} \beta - \beta^\intercal \Sigma^{\mathcal{Q}} \beta\right] = [\beta^*]^\intercal \Sigma^{\mathcal{Q}} \beta^*.$$

We establish (6) by combining (104) and the fact that $\beta \in \mathbb{H}$ can be expressed as $\beta = \mathcal{B}\gamma$ for $\gamma \in \Delta^L$.

## B.10 Proof of Proposition 4

We can write the maximin definition in the following form

$$\beta^{*,\mathrm{MP}} = \arg\max_{\|\beta\|_2 \leq 1} \min_{b \in \mathbb{B}} \beta^\intercal b \tag{105}$$

where $\mathbb{B} = \{b^{(1)}, \ldots, b^{(L)}\}$. Since $b^\intercal \beta$ is linear in $b$, we can replace $\mathbb{B}$ with its convex hull $\mathbb{H}$,

$$\beta^{*,\mathrm{MP}} = \arg\max_{\|\beta\|_2 \leq 1} \min_{b \in \mathbb{H}} b^\intercal \beta$$

We exchange the max and min in the above equation and have

$$\min_{b \in \mathbb{H}} \max_{\|\beta\|_2 \leq 1} b^\intercal \beta = \min_{b \in \mathbb{H}} \|b\|_2$$

We define

$$\xi = \arg\min_{b \in \mathbb{H}} \|b\|_2 \quad \text{and} \quad \xi^* = \xi / \|\xi\|$$

We claim that $\xi^* = \xi / \|\xi\|$ is the optimal solution of (105). For any $\mu \in \mathbb{H}$, we have $\xi + \nu(\mu - \xi) \in \mathbb{H}$ for $\nu \in [0, 1]$ and have

$$\|\xi + \nu(\mu - \xi)\|_2^2 \geq \|\xi\|_2^2 \quad \text{for any} \quad \nu \in [0, 1] \tag{106}$$

By taking $\nu \to 0$, we have

$$\mu^\intercal \xi - \|\xi\|_2^2 \geq 0$$

By dividing both sides by $\|\xi\|_2$, we have

$$\mu^\intercal \xi^* \geq \|\xi\|_2 \quad \text{for any} \quad \mu \in \mathbb{H}. \tag{107}$$

In the definition of (105), we take $\beta = \xi^*$ and have

$$\max_{\|\beta\|_2 \leq 1} \min_{b \in \mathbb{H}} b^\intercal \beta \geq \min_{b \in \mathbb{H}} b^\intercal \xi^* \geq \|\xi\|_2 \tag{108}$$

where the last inequality follows from (107) . Additionally, we take $b = \xi$ in the definition of (105) and have

$$\max_{\|\beta\|_2 \leq 1} \min_{b \in \mathbb{H}} b^\intercal \beta \leq \max_{\|\beta\|_2 \leq 1} \xi^\intercal \beta = \|\xi\|_2$$

Combined with (108), we have shown that

$$\xi^* = \arg\max_{\|\beta\|_2 \leq 1} \min_{b \in \mathbb{H}} b^\intercal \beta$$

that is, $\beta^{*,\mathrm{MP}} = \xi^*$.

## B.11 Proof of Proposition 5

The difference between the proposed estimator $\widehat{\Gamma}_{l,k}$ and $\Gamma_{l,k}$ is

$$\widehat{\Gamma}_{l,k} - \Gamma_{l,k} = D^{(1)}_{l,k} + D^{(2)}_{l,k} + \mathrm{Rem}_{l,k},$$

where $D^{(1)}_{l,k} = \frac{1}{n_k}(b^{(l)})^\intercal [X^{(k)}]^\intercal \epsilon^{(k)} + \frac{1}{n_l}(b^{(k)})^\intercal [X^{(l)}]^\intercal \epsilon^{(l)}$, $D^{(2)}_{l,k} = (b^{(l)})^\intercal(\widehat{\Sigma} - \Sigma)b^{(k)}$, and

$$\begin{aligned}\mathrm{Rem}_{l,k} &= \frac{1}{n_l}(\widehat{b}^{(k)}_{init} - b^{(k)})^\intercal [X^{(l)}]^\intercal \epsilon^{(l)} + \frac{1}{n_k}(\widehat{b}^{(l)}_{init} - b^{(l)})^\intercal [X^{(k)}]^\intercal \epsilon^{(k)} \\ &- (\widehat{b}^{(l)}_{init} - b^{(l)})^\intercal \widehat{\Sigma}(\widehat{b}^{(k)}_{init} - b^{(k)}) + [\widehat{b}^{(l)}_{init}]^\intercal(\widehat{\Sigma} - \widetilde{\Sigma}^{(k)})(\widehat{b}^{(k)}_{init} - b^{(k)}) + [\widehat{b}^{(k)}_{init}]^\intercal(\widehat{\Sigma} - \widetilde{\Sigma}^{(l)})(\widehat{b}^{(l)}_{init} - b^{(l)})\end{aligned} \tag{109}$$

with $\widetilde{\Sigma}^{(l)} = \frac{1}{n_l}\sum_{i=1}^{n_l} X^{(l)}_{i,\cdot}[X^{(l)}_{i,\cdot}]^\intercal$ and

$$\widehat{\Sigma} = \frac{1}{\sum_{l=1}^{L} n_l + N_{\mathcal{Q}}}\left(\sum_{l=1}^{L}\sum_{i=1}^{n_l} X^{(l)}_{i,\cdot}[X^{(l)}_{i,\cdot}]^\intercal + \sum_{i=1}^{N_{\mathcal{Q}}} X^{(l)}_{i,\cdot}[X^{(l)}_{i,\cdot}]^\intercal\right).$$

In the following, we shall control $D^{(1)}, D^{(2)}$ and $\mathrm{Rem}_{l,k}$ separately.

**Control of $D^{(1)}$ and $D^{(2)}$.** By the Gaussian error assumption, we show that $\mathrm{vecl}(D^{(1)}) \mid \{X^{(l)}\}_{1\le l\le L} \sim \mathcal{N}(0, \mathbf{V}^{(1)})$ with $\mathbf{V}^{(1)}$ defined in (39).

For $1 \le j \le L$ and $(l,k) \in \mathcal{I}_L$, we define $W_{i,\cdot} \in \mathbb{R}^{L(L+1)/2}$ as

$$W^{(j)}_{i,\pi(l,k)} = [b^{(l)}]^\intercal(X^{(l)}_{i,\cdot}[X^{(l)}_{i,\cdot}]^\intercal - \Sigma^{\mathcal{Q}})b^{(k)} \quad \text{for} \quad 1 \le i \le n_j$$

and

$$W^{\mathcal{Q}}_{i,\pi(l,k)} = [b^{(l)}]^\intercal(X^{\mathcal{Q}}_{i,\cdot}[X^{\mathcal{Q}}_{i,\cdot}]^\intercal - \Sigma^{\mathcal{Q}})b^{(k)} \quad \text{for} \quad 1 \le i \le n_j$$

where the index mapping $\pi$ is defined in (3). Then we have

$$\{W^{(j)}_{i,\pi(l,k)}\}_{1\le j\le L, 1\le i\le n_j} \quad \text{and} \quad \{W^{\mathcal{Q}}_{i,\pi(l,k)}\}_{1\le i\le N_{\mathcal{Q}}} \quad \text{are i.i.d. random variables.}$$

Regarding $D^{(2)}$, we have

$$[\mathrm{vecl}(D^{(2)})]_{\pi(l,k)} = D^{(2)}_{l,k} = \frac{1}{\sum_{l=1}^{L} n_l + N_{\mathcal{Q}}}\left(\sum_{j=1}^{L}\sum_{i=1}^{n_j} W^{(j)}_{i,\pi(l,k)} + \sum_{i=1}^{N_{\mathcal{Q}}} W^{\mathcal{Q}}_{i,\pi(l,k)}\right).$$

The random variable $W^{(1)}_{1,\cdot}$ is of mean zero and covariance matrix $\mathbf{C}$ defined in (117). By (65), we have shown that $\mathbf{E}\|W^{(1)}_{1,\cdot}\|_2^{r_0} \le C_0$ for some positive constants $r_0 \ge 3$ and $C_0 > 0$. By applying Lemma 2, we establish that there exist two sequences of independent random vectors $\{W^{(j),0}_{i,\cdot}\}_{1\leq i\leq n_j}$ and $\{W^{\mathcal{Q},0}_{i,\cdot}\}_{1\leq i\leq N_{\mathcal{Q}}}$ for $1\leq j\leq L$ and $\{Z^{(j),0}_{i,\cdot}\}_{1\leq i\leq n_j}$ and $\{Z^{\mathcal{Q},0}_{i,\cdot}\}_{1\leq i\leq N_{\mathcal{Q}}}$ for $1\leq j\leq L$ such that

$$Z^{(j),0}_{i,\cdot}\sim\mathcal{N}(0,\mathbf{C})\quad\text{and}\quad Z^{\mathcal{Q},0}_{i,\cdot}\sim\mathcal{N}(0,\mathbf{C}),$$

$$\{W^{(j),0}_{i,\cdot}\}_{1\leq i\leq n_j}\overset{d}{=}\{W^{(j)}_{i,\cdot}\}_{1\leq i\leq n_j}\quad\text{and}\quad\{W^{\mathcal{Q},0}_{i,\cdot}\}_{1\leq i\leq N_{\mathcal{Q}}}\overset{d}{=}\{W^{\mathcal{Q}}_{i,\cdot}\}_{1\leq i\leq N_{\mathcal{Q}}}\tag{110}$$

and

$$\left\|\sum_{j=1}^{L}\sum_{i=1}^{n_j}W^{(j),0}_{i,\pi(l,k)}+\sum_{i=1}^{N_{\mathcal{Q}}}W^{\mathcal{Q},0}_{i,\pi(l,k)}-\sum_{j=1}^{L}\sum_{i=1}^{n_j}Z^{(j),0}_{i,\pi(l,k)}-\sum_{i=1}^{N_{\mathcal{Q}}}Z^{\mathcal{Q},0}_{i,\pi(l,k)}\right\|\lesssim(\sum_{l=1}^{L}n_l+N_{\mathcal{Q}})^{\frac{1}{r_0}}\quad\text{a.s.}\tag{111}$$

We define

$$\mathrm{vecl}(D^{(2),*})=\frac{1}{\sum_{l=1}^{L}n_l+N_{\mathcal{Q}}}\left(\sum_{j=1}^{L}\sum_{i=1}^{n_j}W^{(j),0}_{i,\cdot}+\sum_{i=1}^{N_{\mathcal{Q}}}W^{\mathcal{Q},0}_{i,\cdot}\right)$$

and

$$T^{*}=\frac{1}{\sum_{l=1}^{L}n_l+N_{\mathcal{Q}}}\left(\sum_{j=1}^{L}\sum_{i=1}^{n_j}Z^{(j),0}_{i,\cdot}+\sum_{i=1}^{N_{\mathcal{Q}}}Z^{\mathcal{Q},0}_{i,\cdot}\right)\sim\mathcal{N}(0,\mathbf{V}^{(2)})$$

with $\mathbf{V}^{(2)}$ defined in (40). As a consequence, the corresponding random matrix $D^{(2),*}$ has the same distribution with $D^{(2)}$ and $\|\mathrm{vecl}(D^{(2),*})-T^{*}\|\ll(\sum_{l=1}^{L}n_l+N_{\mathcal{Q}})^{1-\frac{1}{r_0}}\leq(\sum_{l=1}^{L}n_l+N_{\mathcal{Q}})^{-2/3}$ almost surely.

**Control of** $\mathrm{Rem}_{l,k}$**.** We control $\mathrm{Rem}_{l,k}$ by the definition of $\mathrm{Rem}_{l,k}$ in (109) and the following Lemma.

**Lemma 7.** *With probability larger than* $1-\min\{n,p\}^{-c}$*, we have*

$$\begin{aligned}
&\left|\frac{1}{n_l}(\widehat{b}^{(k)}_{init}-b^{(k)})^{\intercal}[X^{(l)}]^{\intercal}\epsilon^{(l)}\right|\lesssim\frac{s\log p}{n};\\
&\left|\frac{1}{n_k}(\widehat{b}^{(l)}_{init}-b^{(l)})^{\intercal}[X^{(k)}]^{\intercal}\epsilon^{(k)}\right|\lesssim\frac{s\log p}{n};\\
&\left|(\widehat{b}^{(l)}_{init}-b^{(l)})^{\intercal}\widehat{\Sigma}(\widehat{b}^{(k)}_{init}-b^{(k)})\right|\lesssim\frac{s\log p}{n}.\\
&\left|[\widehat{b}^{(l)}_{init}]^{\intercal}(\widehat{\Sigma}-\widetilde{\Sigma}^{(k)})(\widehat{b}^{(k)}_{init}-b^{(k)})\right|\lesssim\left(\|b^{(l)}\|_2+\sqrt{\frac{s\log p}{n}}\right)\frac{s\log p}{n}\\
&\left|[\widehat{b}^{(k)}_{init}]^{\intercal}(\widehat{\Sigma}-\widetilde{\Sigma}^{(l)})(\widehat{b}^{(l)}_{init}-b^{(l)})\right|\lesssim\left(\|b^{(k)}\|_2+\sqrt{\frac{s\log p}{n}}\right)\frac{s\log p}{n}
\end{aligned}$$

**Proof of** (42)**.** The control of $\mathbf{V}^{(1)}_{\pi(l,k),\pi(l,k)}$ follows from the definition (39) and the concentration result, for any $1 \leq l \leq L$,

$$\mathbf{P}\left([b^{(l)}]^{\intercal}\left(\frac{1}{n_l}(X^{(l)})^{\intercal}X^{(l)} - \Sigma_l\right)b^{(l)} \leq C\sqrt{\frac{\log p}{n_l}}\|b^{(l)}\|_2^2\right) \geq 1 - p^{-c}$$

for some positive constant $C > 0$ and $c > 0$. The control of $\mathbf{V}^{(2)}_{\pi(l,k),\pi(l,k)}$ follows from (69) and (70)

# C Additional Proofs

## C.1 Proof of Lemma 1

On the event $\mathcal{G}_1 \cap \mathcal{G}_6(\widehat{b}^{(l)}_{init} - b^{(l)}, \widehat{b}^{(l)}_{init} - b^{(l)}, \sqrt{\log p})$, we have

$$\frac{1}{|B|}\sum_{i\in B}[(X^{\mathcal{Q}}_{i,\cdot})^{\intercal}(\widehat{b}^{(l)}_{init} - b^{(l)})]^2 \lesssim \frac{\|b^{(l)}\|_0 \log p}{n}\sigma_l^2.$$

Then we have

$$\begin{aligned}\left|(\widehat{b}^{(l)}_{init} - b^{(l)})^{\intercal}\widehat{\Sigma}^{\mathcal{Q}}(\widehat{b}^{(k)}_{init} - b^{(k)})\right| &\leq \frac{1}{|B|}\|X^{\mathcal{Q}}_{B,\cdot}(\widehat{b}^{(l)}_{init} - b^{(l)})\|_2\|X^{\mathcal{Q}}_{B,\cdot}(\widehat{b}^{(k)}_{init} - b^{(k)})\|_2 \\ &\lesssim \sqrt{\frac{\|b^{(l)}\|_0\|b^{(k)}\|_0(\log p)^2}{n_l n_k}}\end{aligned}$$

and establish (58). We decompose

$$\begin{aligned}&(\widehat{\Sigma}^{\mathcal{Q}}\widehat{b}^{(k)}_{init} - \widehat{\Sigma}^{(l)}\widehat{u}^{(l,k)})^{\intercal}(\widehat{b}^{(l)}_{init} - b^{(l)}) \\ &= (\widetilde{\Sigma}^{\mathcal{Q}}\widehat{b}^{(k)}_{init} - \widehat{\Sigma}^{(l)}\widehat{u}^{(l,k)})^{\intercal}(\widehat{b}^{(l)}_{init} - b^{(l)}) + [\widehat{b}^{(k)}_{init}]^{\intercal}(\widehat{\Sigma}^{\mathcal{Q}} - \widetilde{\Sigma}^{\mathcal{Q}})^{\intercal}(\widehat{b}^{(l)}_{init} - b^{(l)})\end{aligned} \tag{112}$$

Regarding the first term of (112), we apply Hölder's inequality and establish

$$\left|(\widetilde{\Sigma}^{\mathcal{Q}}\widehat{b}^{(k)}_{init} - \widehat{\Sigma}^{(l)}\widehat{u}^{(l,k)})^{\intercal}(\widehat{b}^{(l)}_{init} - b^{(l)})\right| \leq \|\widetilde{\Sigma}^{\mathcal{Q}}\widehat{b}^{(k)}_{init} - \widehat{\Sigma}^{(l)}\widehat{u}^{(l,k)}\|_\infty\|\widehat{b}^{(l)}_{init} - b^{(l)}\|_1$$

By the optimization constraint (16), on the event $\mathcal{G}_2$, we have

$$\left|(\widetilde{\Sigma}^{\mathcal{Q}}\widehat{b}^{(k)}_{init} - \widehat{\Sigma}^{(l)}\widehat{u}^{(l,k)})^{\intercal}(\widehat{b}^{(l)}_{init} - b^{(l)})\right| \lesssim \|\widetilde{\Sigma}^{\mathcal{Q}}\widehat{b}^{(k)}_{init}\|_2\sqrt{\frac{\log p}{n_l}} \cdot \|b^{(l)}\|_0\sqrt{\frac{\log p}{n_l}} \tag{113}$$

Regarding the second term of (112), conditioning on $\widehat{b}^{(k)}_{init}$ and $\widehat{b}^{(l)}_{init}$, on the event $\mathcal{G}_6(\widehat{b}^{(k)}_{init}, \widehat{b}^{(l)}_{init} - b^{(l)}, \sqrt{\log p})$,

$$\left|[\widehat{b}^{(k)}_{init}]^{\intercal}(\widehat{\Sigma}^{\mathcal{Q}} - \widetilde{\Sigma}^{\mathcal{Q}})^{\intercal}(\widehat{b}^{(l)}_{init} - b^{(l)})\right| \lesssim \frac{\sqrt{\log p}}{\sqrt{N_{\mathcal{Q}}}}\|\widehat{b}^{(k)}_{init}\|_2\|\widehat{b}^{(l)}_{init} - b^{(l)}\|_2.$$

On the event $\mathcal{G}_1$, we further have

$$\left|[\widehat{b}^{(k)}_{init}]^{\intercal}(\widehat{\Sigma}^{\mathcal{Q}}-\widetilde{\Sigma}^{\mathcal{Q}})^{\intercal}(\widehat{b}^{(l)}_{init}-b^{(l)})\right|\lesssim\|\widehat{b}^{(k)}_{init}\|_2\sqrt{\frac{\|b^{(l)}\|_0(\log p)^2}{n_l N_{\mathcal{Q}}}}$$

Combined with (113), we establish (59). We establish (60) through applying the similar argument for (59) by exchanging the role of $l$ and $k$. Together with (49), (51) and (52) with $t=\sqrt{\log p}$, we establish the lemma.

## C.2 Proof of Lemmas 7

On the event $\mathcal{G}_0\cap\mathcal{G}_2$, we apply the Hölder's inequality and establish

$$\left|\frac{1}{n_l}(\widehat{b}^{(k)}_{init}-b^{(k)})^{\intercal}[X^{(l)}]^{\intercal}\epsilon^{(l)}\right|\leq\|\widehat{b}^{(k)}_{init}-b^{(k)}\|_1\left\|\frac{1}{n_l}[X^{(l)}]^{\intercal}\epsilon^{(l)}\right\|_{\infty}\lesssim\frac{s\log p}{n}.$$

Similarly, we establish $\left|\frac{1}{n_k}(\widehat{b}^{(l)}_{init}-b^{(l)})^{\intercal}[X^{(k)}]^{\intercal}\epsilon^{(k)}\right|\lesssim s\log p/n$. Note that

$$\left|(\widehat{b}^{(l)}_{init}-b^{(l)})^{\intercal}\widehat{\Sigma}(\widehat{b}^{(k)}_{init}-b^{(k)})\right|\leq\left|(\widehat{b}^{(l)}_{init}-b^{(l)})^{\intercal}\Sigma(\widehat{b}^{(k)}_{init}-b^{(k)})\right|+\left|(\widehat{b}^{(l)}_{init}-b^{(l)})^{\intercal}(\widehat{\Sigma}-\Sigma)(\widehat{b}^{(k)}_{init}-b^{(k)})\right|.$$

On the event $\mathcal{G}_1\cap\mathcal{G}_2$, we have

$$\left|(\widehat{b}^{(l)}_{init}-b^{(l)})^{\intercal}\Sigma(\widehat{b}^{(k)}_{init}-b^{(k)})\right|\lesssim\frac{s\log p}{n}.$$

Note that

$$\left|(\widehat{b}^{(l)}_{init}-b^{(l)})^{\intercal}(\widehat{\Sigma}-\Sigma)(\widehat{b}^{(k)}_{init}-b^{(k)})\right|\leq\sqrt{\left|(\widehat{b}^{(l)}_{init}-b^{(l)})^{\intercal}(\widehat{\Sigma}-\Sigma)(\widehat{b}^{(l)}_{init}-b^{(l)})\right|\left|(\widehat{b}^{(k)}_{init}-b^{(k)})^{\intercal}(\widehat{\Sigma}-\Sigma)(\widehat{b}^{(k)}_{init}-b^{(k)})\right|}$$

On the event $\mathcal{G}_1\cap\mathcal{G}_5$ with $N=\sum_{l=1}^{L}n_l+N_{\mathcal{Q}}$, we have

$$\left|(\widehat{b}^{(l)}_{init}-b^{(l)})^{\intercal}(\widehat{\Sigma}-\Sigma)(\widehat{b}^{(k)}_{init}-b^{(k)})\right|\lesssim\frac{s\log p}{n}\cdot\sqrt{\frac{s\log p}{\sum_{l=1}^{L}n_l+N_{\mathcal{Q}}}}. \tag{114}$$

Combined with $\left|(\widehat{b}^{(l)}_{init}-b^{(l)})^{\intercal}\Sigma(\widehat{b}^{(k)}_{init}-b^{(k)})\right|\lesssim\frac{s\log p}{n}$, we establish

$$\left|(\widehat{b}^{(l)}_{init}-b^{(l)})^{\intercal}\widehat{\Sigma}(\widehat{b}^{(k)}_{init}-b^{(k)})\right|\lesssim\frac{s\log p}{n}.$$

Note that

$$\begin{aligned}
&[\widehat{b}^{(l)}_{init}]^{\intercal}(\widehat{\Sigma}-\widetilde{\Sigma}^{(k)})(\widehat{b}^{(k)}_{init}-b^{(k)})\\
&=[\widehat{b}^{(l)}_{init}-b^{(l)}]^{\intercal}(\widehat{\Sigma}-\widetilde{\Sigma}^{(k)})(\widehat{b}^{(k)}_{init}-b^{(k)})+[b^{(l)}]^{\intercal}(\widehat{\Sigma}-\widetilde{\Sigma}^{(k)})(\widehat{b}^{(k)}_{init}-b^{(k)})\\
&=[\widehat{b}^{(l)}_{init}-b^{(l)}]^{\intercal}(\widehat{\Sigma}-\Sigma)(\widehat{b}^{(k)}_{init}-b^{(k)})+[b^{(l)}]^{\intercal}(\widehat{\Sigma}-\Sigma)(\widehat{b}^{(k)}_{init}-b^{(k)})\\
&-[\widehat{b}^{(l)}_{init}-b^{(l)}]^{\intercal}(\widetilde{\Sigma}^{(k)}-\Sigma)(\widehat{b}^{(k)}_{init}-b^{(k)})-[b^{(l)}]^{\intercal}(\widetilde{\Sigma}^{(k)}-\Sigma)(\widehat{b}^{(k)}_{init}-b^{(k)})
\end{aligned} \tag{115}$$

With a similar proof for (114), we show that, on the event $\mathcal{G}_2 \cap \mathcal{G}_5$,

$$\left|[b^{(l)}]^\intercal(\widehat{\Sigma} - \Sigma)(\widehat{b}^{(k)}_{init} - b^{(k)})\right| \lesssim \|b^{(l)}\|_2 \sqrt{\frac{s \log p}{n}} \sqrt{\frac{s \log p}{\sum_{l=1}^{L} n_l + N_{\mathcal{Q}}}}$$

$$\left|[\widehat{b}^{(l)}_{init} - b^{(l)}]^\intercal(\widetilde{\Sigma}^{(k)} - \Sigma)(\widehat{b}^{(k)}_{init} - b^{(k)})\right| \lesssim \frac{s \log p}{n} \sqrt{\frac{s \log p}{n}}$$

$$\left|[b^{(l)}]^\intercal(\widetilde{\Sigma}^{(k)} - \Sigma)(\widehat{b}^{(k)}_{init} - b^{(k)})\right| \lesssim \|b^{(l)}\|_2 \frac{s \log p}{n}$$

By the above inequalities, (114) and (115), we establish

$$\left|[\widehat{b}^{(l)}_{init}]^\intercal(\widehat{\Sigma} - \widetilde{\Sigma}^{(k)})(\widehat{b}^{(k)}_{init} - b^{(k)})\right| \lesssim \left(\|b^{(l)}\|_2 + \sqrt{\frac{s \log p}{n}}\right) \frac{s \log p}{n}.$$

We can apply a similar argument to control $\left|[\widehat{b}^{(k)}_{init}]^\intercal(\widehat{\Sigma} - \widetilde{\Sigma}^{(l)})(\widehat{b}^{(l)}_{init} - b^{(l)})\right|$. Together with (49) and (51), we establish the lemma.

## C.3 Proof of Lemma 4

The control of the event $\mathcal{E}_2$ follows from the proof of (30). In the following, we shall control the event $\mathcal{E}_1$ by establish the following two high probability inequalities.

For $\mathbf{V}_{\pi(l_1,k_1),\pi(l_2,k_2)}$ defined in (29), we express it as

$$\mathbf{V}_{\pi(l_1,k_1),\pi(l_2,k_2)} = \mathbf{V}^{(a)}_{\pi(l_1,k_1),\pi(l_2,k_2)} + \mathbf{V}^{(b)}_{\pi(l_1,k_1),\pi(l_2,k_2)} \tag{116}$$

where $\mathbf{V}^{(a)}_{\pi(l_1,k_1),\pi(l_2,k_2)}$ defined in (62) and

$$\mathbf{V}^{(b)}_{\pi(l_1,k_1),\pi(l_2,k_2)} = \frac{1}{|B|}(\mathbf{E}[b^{(l_1)}]^\intercal X^{\mathcal{Q}}_{i,\cdot}[b^{(k_1)}]^\intercal X^{\mathcal{Q}}_{i,\cdot}[b^{(l_2)}]^\intercal X^{\mathcal{Q}}_{i,\cdot}[b^{(k_2)}]^\intercal X^{\mathcal{Q}}_{i,\cdot} - (b^{(l_1)})^\intercal \Sigma^{\mathcal{Q}} b^{(k_1)} (b^{(l_2)})^\intercal \Sigma^{\mathcal{Q}} b^{(k_2)}) \tag{117}$$

For $\widehat{\mathbf{V}}_{\pi(l_1,k_1),\pi(l_2,k_2)}$ defined in (21), we express it as

$$\widehat{\mathbf{V}}_{\pi(l_1,k_1),\pi(l_2,k_2)} = \widehat{\mathbf{V}}^{(a)}_{\pi(l_1,k_1),\pi(l_2,k_2)} + \widehat{\mathbf{V}}^{(b)}_{\pi(l_1,k_1),\pi(l_2,k_2)} \tag{118}$$

with

$$\begin{aligned}\widehat{\mathbf{V}}^{(a)}_{\pi(l_1,k_1),\pi(l_2,k_2)} =& \frac{\widehat{\sigma}^2_{l_1}}{|B_{l_1}|}(\widehat{u}^{(l_1,k_1)})^\intercal \widehat{\Sigma}^{(l_1)} \left[\widehat{u}^{(l_2,k_2)} \mathbf{1}(l_2 = l_1) + \widehat{u}^{(k_2,l_2)} \mathbf{1}(k_2 = l_1)\right] \\ &+ \frac{\widehat{\sigma}^2_{k_1}}{|B_{k_1}|}(\widehat{u}^{(k_1,l_1)})^\intercal \widehat{\Sigma}^{(k_1)} \left[\widehat{u}^{(l_2,k_2)} \mathbf{1}(l_2 = k_1) + \widehat{u}^{(k_2,l_2)} \mathbf{1}(k_2 = k_1)\right]\end{aligned} \tag{119}$$

$$\widehat{\mathbf{V}}^{(b)}_{\pi(l_1,k_1),\pi(l_2,k_2)} = \frac{\sum_{i=1}^{N_{\mathcal{Q}}}\left((\widehat{b}^{(l_1)}_{init})^{\intercal}X^{\mathcal{Q}}_{i,\cdot}(\widehat{b}^{(k_1)}_{init})^{\intercal}X^{\mathcal{Q}}_{i,\cdot}(\widehat{b}^{(l_2)}_{init})^{\intercal}X^{\mathcal{Q}}_{i,\cdot}(\widehat{b}^{(k_2)}_{init})^{\intercal}X^{\mathcal{Q}}_{i,\cdot} - (\widehat{b}^{(l_1)}_{init})^{\intercal}\bar{\Sigma}^{\mathcal{Q}}\widehat{b}^{(k_1)}_{init}(\widehat{b}^{(l_2)}_{init})^{\intercal}\bar{\Sigma}^{\mathcal{Q}}\widehat{b}^{(k_2)}_{init}\right)}{|B|N_{\mathcal{Q}}} \tag{120}$$

where $\bar{\Sigma}^{\mathcal{Q}} = \frac{1}{N_{\mathcal{Q}}}\sum_{i=1}^{N_{\mathcal{Q}}} X^{\mathcal{Q}}_{i,\cdot}(X^{\mathcal{Q}}_{i,\cdot})^{\intercal}$ and $\widehat{\sigma}^2_l = \|Y^{(l)} - X^{(l)}\widehat{b}^{(l)}\|_2^2/n_l$ for $1 \le l \le L$.

With probability larger than $1-\exp(-cn) - \min\{N_{\mathcal{Q}},p\}^{-c}$ for some positive constant $c>0$, we have

$$n \cdot \left|\widehat{\mathbf{V}}^{(a)}_{\pi(l_1,k_1),\pi(l_2,k_2)} - \mathbf{V}^{(a)}_{\pi(l_1,k_1),\pi(l_2,k_2)}\right| \le Cd_0\left(\frac{s\log p}{n} + \sqrt{\frac{\log p}{n}}\right) \le \frac{d_0}{4}. \tag{121}$$

$$N_{\mathcal{Q}} \cdot \left|\widehat{\mathbf{V}}^{(b)}_{\pi(l_1,k_1),\pi(l_2,k_2)} - \mathbf{V}^{(b)}_{\pi(l_1,k_1),\pi(l_2,k_2)}\right| \lesssim \log\max\{N_{\mathcal{Q}},p\}\sqrt{\frac{s\log p\log N_{\mathcal{Q}}}{n}} + \frac{(\log N_{\mathcal{Q}})^{5/2}}{\sqrt{N_{\mathcal{Q}}}}. \tag{122}$$

We combine (121) and (122) and establish

$$\begin{aligned}
&\|\widehat{\mathbf{Cov}} - \mathbf{Cov}\|_2 \\
&\le n \cdot \left|\widehat{\mathbf{V}}^{(a)}_{\pi(l_1,k_1),\pi(l_2,k_2)} - \mathbf{V}^{(a)}_{\pi(l_1,k_1),\pi(l_2,k_2)}\right| + n \cdot \left|\widehat{\mathbf{V}}^{(b)}_{\pi(l_1,k_1),\pi(l_2,k_2)} - \mathbf{V}^{(b)}_{\pi(l_1,k_1),\pi(l_2,k_2)}\right| \\
&\le \frac{d_0}{4} + \frac{\sqrt{n\cdot s[\log\max\{N_{\mathcal{Q}},p\}]^3}}{N_{\mathcal{Q}}} + \frac{n\cdot(\log N_{\mathcal{Q}})^{5/2}}{N_{\mathcal{Q}}^{3/2}} \le d_0/2,
\end{aligned}$$

where the last inequality follows from $n < N_{\mathcal{Q}}^{4/3}$ and $\sqrt{s[\log\max\{N_{\mathcal{Q}},p\}]^3} \le cN_{\mathcal{Q}}/\sqrt{n}$.

### C.3.1 Proof of (121)

$$\begin{aligned}
&n \cdot \left|\widehat{\mathbf{V}}^{(a)}_{\pi(l_1,k_1),\pi(l_2,k_2)} - \mathbf{V}^{(a)}_{\pi(l_1,k_1),\pi(l_2,k_2)}\right| \\
&\lesssim \left|\widehat{\sigma}^2_{l_1} - \sigma^2_{l_1}\right| (\widehat{u}^{(l_1,k_1)})^{\intercal}\widehat{\Sigma}^{(l_1)}\left[\widehat{u}^{(l_2,k_2)}\mathbf{1}(l_2=l_1) + \widehat{u}^{(k_2,l_2)}\mathbf{1}(k_2=l_1)\right] \\
&+ \left|\widehat{\sigma}^2_{k_1} - \sigma^2_{k_1}\right| (\widehat{u}^{(k_1,l_1)})^{\intercal}\widehat{\Sigma}^{(k_1)}\left[\widehat{u}^{(l_2,k_2)}\mathbf{1}(l_2=k_1) + \widehat{u}^{(k_2,l_2)}\mathbf{1}(k_2=k_1)\right]
\end{aligned} \tag{123}$$

Since

$$\begin{aligned}
&\left|(\widehat{u}^{(l_1,k_1)})^{\intercal}\widehat{\Sigma}^{(l_1)}\left[\widehat{u}^{(l_2,k_2)}\mathbf{1}(l_2=l_1) + \widehat{u}^{(k_2,l_2)}\mathbf{1}(k_2=l_1)\right]\right| \\
&\le \sqrt{(\widehat{u}^{(l_1,k_1)})^{\intercal}\widehat{\Sigma}^{(l_1)}\widehat{u}^{(l_1,k_1)} \cdot (\widehat{u}^{(l_1,k_2)})^{\intercal}\widehat{\Sigma}^{(l_1)}\widehat{u}^{(l_1,k_2)}} \\
&+ \sqrt{(\widehat{u}^{(l_1,k_1)})^{\intercal}\widehat{\Sigma}^{(l_1)}\widehat{u}^{(l_1,k_1)} \cdot (\widehat{u}^{(l_1,l_2)})^{\intercal}\widehat{\Sigma}^{(l_1)}\widehat{u}^{(l_1,l_2)}}
\end{aligned} \tag{124}$$

we have

$$\left|(\widehat{u}^{(l_1,k_1)})^{\intercal}\widehat{\Sigma}^{(l_1)}\left[\widehat{u}^{(l_2,k_2)}\mathbf{1}(l_2=l_1) + \widehat{u}^{(k_2,l_2)}\mathbf{1}(k_2=l_1)\right]\right| \lesssim n\max_{(l,k)\in\mathcal{I}_L}\mathbf{V}^{(a)}_{\pi(l,k),\pi(l,k)} \le d_0$$

Similarly, we have

$$\left|(\widehat{u}^{(k_1,l_1)})^\intercal \widehat{\Sigma}^{(k_1)}\left[\widehat{u}^{(l_2,k_2)}\mathbf{1}(l_2=k_1)+\widehat{u}^{(k_2,l_2)}\mathbf{1}(k_2=k_1)\right]\right| \lesssim d_0$$

Hence, on the event $\mathcal{G}_3$, we establish (121).

### C.3.2 Proof of (122)

Define

$$W_{i,1}=[b^{(l_1)}]^\intercal X^{\mathcal{Q}}_{i,\cdot},\ W_{i,2}=[b^{(k_1)}]^\intercal X^{\mathcal{Q}}_{i,\cdot},\ W_{i,3}=[b^{(l_2)}]^\intercal X^{\mathcal{Q}}_{i,\cdot},\ W_{i,4}=[b^{(k_2)}]^\intercal X^{\mathcal{Q}}_{i,\cdot};$$

and

$$\widehat{W}_{i,1}=(\widehat{b}^{(l_1)}_{init})^\intercal X^{\mathcal{Q}}_{i,\cdot},\ \widehat{W}_{i,2}=(\widehat{b}^{(k_1)}_{init})^\intercal X^{\mathcal{Q}}_{i,\cdot},\ \widehat{W}_{i,3}=(\widehat{b}^{(l_2)}_{init})^\intercal X^{\mathcal{Q}}_{i,\cdot},\ \widehat{W}_{i,4}=(\widehat{b}^{(k_2)}_{init})^\intercal X^{\mathcal{Q}}_{i,\cdot}.$$

With the above definitions, we have

$$\begin{aligned}&\mathbf{E}[b^{(l_1)}]^\intercal X^{\mathcal{Q}}_{i,\cdot}[b^{(k_1)}]^\intercal X^{\mathcal{Q}}_{i,\cdot}[b^{(l_2)}]^\intercal X^{\mathcal{Q}}_{i,\cdot}[b^{(k_2)}]^\intercal X^{\mathcal{Q}}_{i,\cdot}-(b^{(l_1)})^\intercal\Sigma^{\mathcal{Q}}b^{(k_1)}(b^{(l_2)})^\intercal\Sigma^{\mathcal{Q}}b^{(k_2)}\\&=\mathbf{E}\prod_{t=1}^{4}W_{i,t}-\mathbf{E}W_{i,1}W_{i,2}\cdot\mathbf{E}W_{i,3}W_{i,4}\end{aligned}\tag{125}$$

and

$$\begin{aligned}&\frac{1}{N_{\mathcal{Q}}}\sum_{i=1}^{N_{\mathcal{Q}}}\left((\widehat{b}^{(l_1)}_{init})^\intercal X^{\mathcal{Q}}_{i,\cdot}(\widehat{b}^{(k_1)}_{init})^\intercal X^{\mathcal{Q}}_{i,\cdot}(\widehat{b}^{(l_2)}_{init})^\intercal X^{\mathcal{Q}}_{i,\cdot}(\widehat{b}^{(k_2)}_{init})^\intercal X^{\mathcal{Q}}_{i,\cdot}-(\widehat{b}^{(l_1)}_{init})^\intercal\widehat{\Sigma}\widehat{b}^{(k_1)}_{init}(\widehat{b}^{(l_2)}_{init})^\intercal\widehat{\Sigma}\widehat{b}^{(k_2)}_{init}\right)\\&=\frac{1}{N_{\mathcal{Q}}}\sum_{i=1}^{N_{\mathcal{Q}}}\prod_{t=1}^{4}\widehat{W}_{i,t}-\frac{1}{N_{\mathcal{Q}}}\sum_{i=1}^{N_{\mathcal{Q}}}\widehat{W}_{i,1}\widehat{W}_{i,2}\cdot\frac{1}{N_{\mathcal{Q}}}\sum_{i=1}^{N_{\mathcal{Q}}}\widehat{W}_{i,3}\widehat{W}_{i,4}\end{aligned}\tag{126}$$

Hence, it is sufficient to control the following terms.

$$\frac{1}{N_{\mathcal{Q}}}\sum_{i=1}^{N_{\mathcal{Q}}}\prod_{t=1}^{4}\widehat{W}_{i,t}-\mathbf{E}\prod_{t=1}^{4}W_{i,t}=\frac{1}{N_{\mathcal{Q}}}\sum_{i=1}^{N_{\mathcal{Q}}}\prod_{t=1}^{4}\widehat{W}_{i,t}-\frac{1}{N_{\mathcal{Q}}}\sum_{i=1}^{N_{\mathcal{Q}}}\prod_{t=1}^{4}W_{i,t}+\frac{1}{N_{\mathcal{Q}}}\sum_{i=1}^{N_{\mathcal{Q}}}\prod_{t=1}^{4}W_{i,t}-\mathbf{E}\prod_{t=1}^{4}W_{i,t}$$

$$\frac{1}{N_{\mathcal{Q}}}\sum_{i=1}^{N_{\mathcal{Q}}}\widehat{W}_{i,1}\widehat{W}_{i,2}-\mathbf{E}W_{i,1}W_{i,2}=\frac{1}{N_{\mathcal{Q}}}\sum_{i=1}^{N_{\mathcal{Q}}}\widehat{W}_{i,1}\widehat{W}_{i,2}-\frac{1}{N_{\mathcal{Q}}}\sum_{i=1}^{N_{\mathcal{Q}}}W_{i,1}W_{i,2}+\frac{1}{N_{\mathcal{Q}}}\sum_{i=1}^{N_{\mathcal{Q}}}W_{i,1}W_{i,2}-\mathbf{E}W_{i,1}W_{i,2}$$

$$\frac{1}{N_{\mathcal{Q}}}\sum_{i=1}^{N_{\mathcal{Q}}}\widehat{W}_{i,3}\widehat{W}_{i,4}-\mathbf{E}W_{i,3}W_{i,4}=\frac{1}{N_{\mathcal{Q}}}\sum_{i=1}^{N_{\mathcal{Q}}}\widehat{W}_{i,3}\widehat{W}_{i,4}-\frac{1}{N_{\mathcal{Q}}}\sum_{i=1}^{N_{\mathcal{Q}}}W_{i,3}W_{i,4}+\frac{1}{N_{\mathcal{Q}}}\sum_{i=1}^{N_{\mathcal{Q}}}W_{i,3}W_{i,4}-\mathbf{E}W_{i,3}W_{i,4}$$

Specifically, we will show that, with probability larger than $1-\min\{N_{\mathcal{Q}},p\}^{-c}$,

$$\left|\frac{1}{N_{\mathcal{Q}}}\sum_{i=1}^{N_{\mathcal{Q}}}W_{i,1}W_{i,2}-\mathbf{E}W_{i,1}W_{i,2}\right|\lesssim\|b^{(l_1)}\|_2\|b^{(k_1)}\|_2\sqrt{\frac{\log N_{\mathcal{Q}}}{N_{\mathcal{Q}}}},\tag{127}$$

$$\left|\frac{1}{N_{\mathcal{Q}}}\sum_{i=1}^{N_{\mathcal{Q}}} W_{i,3}W_{i,4}-\mathbf{E}W_{i,3}W_{i,4}\right| \lesssim \|b^{(l_2)}\|_2\|b^{(k_2)}\|_2\sqrt{\frac{\log N_{\mathcal{Q}}}{N_{\mathcal{Q}}}}, \tag{128}$$

$$\frac{1}{N_{\mathcal{Q}}}\sum_{i=1}^{N_{\mathcal{Q}}}\left(\prod_{t=1}^{4} W_{i,t}-\mathbf{E}\prod_{t=1}^{4} W_{i,t}\right) \lesssim \|b^{(l_1)}\|_2\|b^{(k_1)}\|_2\|b^{(l_2)}\|_2\|b^{(k_2)}\|_2\frac{(\log N_{\mathcal{Q}})^{5/2}}{\sqrt{N_{\mathcal{Q}}}}, \tag{129}$$

$$\left|\frac{1}{N_{\mathcal{Q}}}\sum_{i=1}^{N_{\mathcal{Q}}} \widehat{W}_{i,1}\widehat{W}_{i,2}-\frac{1}{N_{\mathcal{Q}}}\sum_{i=1}^{N_{\mathcal{Q}}} W_{i,1}W_{i,2}\right| \lesssim \sqrt{\frac{s\log p}{n}}\left(\|b^{(l_1)}\|_2+\|b^{(k_1)}\|_2+\sqrt{\frac{s\log p}{n}}\right), \tag{130}$$

$$\left|\frac{1}{N_{\mathcal{Q}}}\sum_{i=1}^{N_{\mathcal{Q}}} \widehat{W}_{i,3}\widehat{W}_{i,4}-\frac{1}{N_{\mathcal{Q}}}\sum_{i=1}^{N_{\mathcal{Q}}} W_{i,3}W_{i,4}\right| \lesssim \sqrt{\frac{s\log p}{n}}\left(\|b^{(l_2)}\|_2+\|b^{(k_2)}\|_2+\sqrt{\frac{s\log p}{n}}\right). \tag{131}$$

If we further assume that $\|b^{(l)}\|_2 \leq C$ for $1 \leq l \leq L$ and $s^2(\log p)^2/n \leq c$ for some positive constants $C > 0$ and $c > 0$, then with probability larger than $1-\min\{N_{\mathcal{Q}}, p\}^{-c}$,

$$\left|\frac{1}{N_{\mathcal{Q}}}\sum_{i=1}^{N_{\mathcal{Q}}}\prod_{t=1}^{4} \widehat{W}_{i,t}-\frac{1}{N_{\mathcal{Q}}}\sum_{i=1}^{N_{\mathcal{Q}}}\prod_{t=1}^{4} W_{i,t}\right| \lesssim \log\max\{N_{\mathcal{Q}}, p\}\sqrt{\frac{s\log p\log N_{\mathcal{Q}}}{n}}. \tag{132}$$

By the expression (125) and (126), we establish (122) by applying (127), (128), (129), (130), (131), (132). In the following, we prove (127), (128) and (129). Then we will present the proofs of (130), (131), (132).

**Proofs of** (127)**,** (128) **and** (129)**.** We shall apply the following lemma to control the above terms, which re-states the Lemma 1 in [12].

**Lemma 8.** *Let $\xi_1, \cdots, \xi_n$ be independent random variables with mean 0. Suppose that there exists some $c > 0$ and $U_n$ such that $\sum_{i=1}^n \mathbf{E}\xi_i^2 \exp\left(c|\xi_i|\right) \leq U_n^2$. Then for $0 < t \leq U_n$,*

$$\mathbf{P}\left(\sum_{i=1}^{n} \xi_i \geq C U_n t\right) \leq \exp(-t^2), \tag{133}$$

*where $C = c + c^{-1}$.*

Define

$$W_{i,1}^0 = \frac{[b^{(l_1)}]^{\intercal} X_{i,\cdot}^{\mathcal{Q}}}{\sqrt{[b^{(l_1)}]^{\intercal}\Sigma^{\mathcal{Q}} b^{(l_1)}}},\ W_{i,2}^0 = \frac{[b^{(k_1)}]^{\intercal} X_{i,\cdot}^{\mathcal{Q}}}{\sqrt{[b^{(k_1)}]^{\intercal}\Sigma^{\mathcal{Q}} b^{(k_1)}}}$$

and

$$W_{i,3}^0 = \frac{[b^{(l_2)}]^{\intercal} X_{i,\cdot}^{\mathcal{Q}}}{\sqrt{[b^{(l_2)}]^{\intercal}\Sigma^{\mathcal{Q}} b^{(l_2)}}}\ W_{i,4}^0 = \frac{[b^{(k_2)}]^{\intercal} X_{i,\cdot}^{\mathcal{Q}}}{\sqrt{[b^{(k_2)}]^{\intercal}\Sigma^{\mathcal{Q}} b^{(k_2)}}}$$

Since $X_{i,\cdot}^{\mathcal{Q}}$ is sub-gaussian, $W_{i,t}^0$ is sub-gaussian and both $W_{i,1}^0 W_{i,2}^0$ and $W_{i,3}^0 W_{i,4}^0$ are sub-exponetial random variables, which follows from Remark 5.18 in [43]. By Corollary 5.17 in [43], we have

$$\mathbf{P}\left(\left|\frac{1}{N_{\mathcal{Q}}}\sum_{i=1}^{N_{\mathcal{Q}}}\left(W_{i,1}^0 W_{i,2}^0 - \mathbf{E}W_{i,1}^0 W_{i,2}^0\right)\right| \geq C\sqrt{\frac{\log N_{\mathcal{Q}}}{N_{\mathcal{Q}}}}\right) \leq 2N_{\mathcal{Q}}^{-c}$$

and

$$\mathbf{P}\left(\left|\frac{1}{N_{\mathcal{Q}}}\sum_{i=1}^{N_{\mathcal{Q}}}\left(W_{i,3}^0 W_{i,4}^0 - \mathbf{E}W_{i,3}^0 W_{i,4}^0\right)\right| \geq C\sqrt{\frac{\log N_{\mathcal{Q}}}{N_{\mathcal{Q}}}}\right) \leq 2N_{\mathcal{Q}}^{-c}$$

where $c$ and $C$ are positive constants. The above inequalities imply (127) and (128) after rescaling.

For $1 \leq t \leq 4$, since $W_{i,t}^0$ is a sub-gaussian random variable, there exist positive constants $C_1 > 0$ and $c > 2$ such that the following concentration inequality holds,

$$\sum_{i=1}^{N_{\mathcal{Q}}}\mathbf{P}\left(\max_{1\leq t\leq 4}|W_{i,t}^0| \geq C_1\sqrt{\log N_{\mathcal{Q}}}\right) \leq N_{\mathcal{Q}}\max_{1\leq i\leq N_{\mathcal{Q}}}\mathbf{P}\left(\max_{1\leq t\leq 4}|W_{i,t}^0| \geq C_1\sqrt{\log N_{\mathcal{Q}}}\right) \lesssim N_{\mathcal{Q}}^{-c} \tag{134}$$

Define

$$H_{i,a} = \prod_{t=1}^{4} W_{i,t}^0 \cdot \mathbf{1}\left(\max_{1\leq t\leq 4}|W_{i,t}^0| \leq C_1\sqrt{\log N_{\mathcal{Q}}}\right) \quad \text{for} \quad 1 \leq t \leq 4,$$

and

$$H_{i,b} = \prod_{t=1}^{4} W_{i,t}^0 \cdot \mathbf{1}\left(\max_{1\leq t\leq 4}|W_{i,t}^0| \geq C_1\sqrt{\log N_{\mathcal{Q}}}\right) \quad \text{for} \quad 1 \leq t \leq 4.$$

Then we have

$$\frac{1}{N_{\mathcal{Q}}}\sum_{i=1}^{N_{\mathcal{Q}}}\prod_{t=1}^{4} W_{i,t}^0 - \mathbf{E}\prod_{t=1}^{4} W_{i,t}^0 = \frac{1}{N_{\mathcal{Q}}}\sum_{i=1}^{N_{\mathcal{Q}}}\left(H_{i,a} - \mathbf{E}H_{i,a}\right) + \frac{1}{N_{\mathcal{Q}}}\sum_{i=1}^{N_{\mathcal{Q}}}\left(H_{i,b} - \mathbf{E}H_{i,b}\right) \tag{135}$$

By applying the Cauchy-Schwarz inequality, we bound $\mathbf{E}H_{i,b}$ as

$$\begin{aligned}|\mathbf{E}H_{i,b}| &\leq \sqrt{\mathbf{E}\left(\prod_{t=1}^{4} W_{i,t}^0\right)^2 \mathbf{P}\left(\max_{1\leq t\leq 4}|W_{i,t}^0| \geq C_1\sqrt{\log N_{\mathcal{Q}}}\right)} \\ &\lesssim \mathbf{P}\left(|W_{i,t}^0| \geq C_1\sqrt{\log N_{\mathcal{Q}}}\right)^{1/2} \lesssim N_{\mathcal{Q}}^{-1/2},\end{aligned} \tag{136}$$

where the last inequality follows from the fact that $W_{i,t}^0$ is a sub-gaussian random variable. Now we apply Lemma 8 to bound $\frac{1}{N_{\mathcal{Q}}}\sum_{i=1}^{N_{\mathcal{Q}}}\left(H_{i,a} - \mathbf{E}H_{i,a}\right)$. By taking $c = c_1/(C_1^2 \log N_{\mathcal{Q}})^2$

for some small positive constant $c_1 > 0$, we have

$$\sum_{i=1}^{N_{\mathcal{Q}}} \mathbf{E}\left(H_{i,a} - \mathbf{E}H_{i,a}\right)^2 \exp\left(c\left|H_{i,a} - \mathbf{E}H_{i,a}\right|\right) \leq C \sum_{i=1}^{N_{\mathcal{Q}}} \mathbf{E}\left(H_{i,a} - \mathbf{E}H_{i,a}\right)^2 \leq C_2 N_{\mathcal{Q}}.$$

By applying Lemma 8 with $U_n = \sqrt{C_2 N_{\mathcal{Q}}}$, $c = c_1/(C_1^2 \log N_{\mathcal{Q}})^2$ and $t = \sqrt{\log N_{\mathcal{Q}}}$, then we have

$$\mathbf{P}\left(\frac{1}{N_{\mathcal{Q}}}\sum_{i=1}^{N_{\mathcal{Q}}}\left(H_{i,a} - \mathbf{E}H_{i,a}\right) \geq C\frac{(\log N_{\mathcal{Q}})^{5/2}}{\sqrt{N_{\mathcal{Q}}}}\right) \lesssim N_{\mathcal{Q}}^{-c}. \tag{137}$$

Note that

$$\begin{aligned}
\mathbf{P}\left(\left|\frac{1}{N_{\mathcal{Q}}}\sum_{i=1}^{N_{\mathcal{Q}}} H_{i,b}\right| \geq C\frac{(\log N_{\mathcal{Q}})^{5/2}}{\sqrt{N_{\mathcal{Q}}}}\right) &\leq \mathbf{P}\left(\frac{1}{N_{\mathcal{Q}}}\sum_{i=1}^{N_{\mathcal{Q}}} |H_{i,b}| \geq C\frac{(\log N_{\mathcal{Q}})^{5/2}}{\sqrt{N_{\mathcal{Q}}}}\right) \\
&\leq \sum_{i=1}^{N_{\mathcal{Q}}} \mathbf{P}\left(|H_{i,b}| \geq C\frac{(\log N_{\mathcal{Q}})^{5/2}}{\sqrt{N_{\mathcal{Q}}}}\right) \\
&\leq \sum_{i=1}^{N_{\mathcal{Q}}} \mathbf{P}\left(\max_{1\leq t\leq 4} |W_{i,t}^0| \geq C_1\sqrt{\log N_{\mathcal{Q}}}\right) \lesssim N_{\mathcal{Q}}^{-c}
\end{aligned} \tag{138}$$

where the last inequality follows from (134).

By the decomposition (135), we have

$$\begin{aligned}
&\mathbf{P}\left(\left|\frac{1}{N_{\mathcal{Q}}}\sum_{i=1}^{N_{\mathcal{Q}}}\left(\prod_{t=1}^{4} W_{i,t}^0 - \mathbf{E}\prod_{t=1}^{4} W_{i,t}^0\right)\right| \geq 3C\frac{(\log N_{\mathcal{Q}})^{5/2}}{\sqrt{N_{\mathcal{Q}}}}\right) \\
&\leq \mathbf{P}\left(\left|\frac{1}{N_{\mathcal{Q}}}\sum_{i=1}^{N_{\mathcal{Q}}}\left(H_{i,a} - \mathbf{E}H_{i,a}\right)\right| \geq C\frac{(\log N_{\mathcal{Q}})^{5/2}}{\sqrt{N_{\mathcal{Q}}}}\right) \\
&+ \mathbf{P}\left(|\mathbf{E}H_{i,b}| \geq C\frac{(\log N_{\mathcal{Q}})^{5/2}}{\sqrt{N_{\mathcal{Q}}}}\right) + \mathbf{P}\left(\left|\frac{1}{N_{\mathcal{Q}}}\sum_{i=1}^{N_{\mathcal{Q}}} H_{i,b}\right| \geq C\frac{(\log N_{\mathcal{Q}})^{5/2}}{\sqrt{N_{\mathcal{Q}}}}\right) \lesssim N_{\mathcal{Q}}^{-c}.
\end{aligned}$$

where the final upper bound follows from (136), (137) and (138). Hence, we establish that (129) holds with probability larger than $1 - N_{\mathcal{Q}}^{-c}$.

**Proofs** (130)**,** (131) **and** (132)**.** It follows from the definitions of $\widehat{W}_{i,t}$ and $W_{i,t}$ that

$$\begin{aligned}
\frac{1}{N_{\mathcal{Q}}}\sum_{i=1}^{N_{\mathcal{Q}}} \widehat{W}_{i,1}\widehat{W}_{i,2} - \frac{1}{N_{\mathcal{Q}}}\sum_{i=1}^{N_{\mathcal{Q}}} W_{i,1}W_{i,2} &= [\widehat{b}_{init}^{(l_1)} - b^{(l_1)}]^{\intercal}\frac{1}{N_{\mathcal{Q}}}\sum_{i=1}^{N_{\mathcal{Q}}} X_{i,\cdot}^{\mathcal{Q}}[X_{i,\cdot}^{\mathcal{Q}}]^{\intercal}[\widehat{b}_{init}^{(k_1)} - b^{(k_1)}] \\
+ [b^{(l_1)}]^{\intercal}\frac{1}{N_{\mathcal{Q}}}\sum_{i=1}^{N_{\mathcal{Q}}} X_{i,\cdot}^{\mathcal{Q}}[X_{i,\cdot}^{\mathcal{Q}}]^{\intercal}[\widehat{b}_{init}^{(k_1)} - b^{(k_1)}] &+ [\widehat{b}_{init}^{(l_1)} - b^{(l_1)}]^{\intercal}\frac{1}{N_{\mathcal{Q}}}\sum_{i=1}^{N_{\mathcal{Q}}} X_{i,\cdot}^{\mathcal{Q}}[X_{i,\cdot}^{\mathcal{Q}}]^{\intercal} b^{(k_1)}
\end{aligned} \tag{139}$$

On the event $\mathcal{G}_2 \cap \mathcal{G}_5$ with $\mathcal{G}_2$ defined in (48) and $\mathcal{G}_5$ defined in (50), we establish (130). By a similar argument, we establish (131).

Furthermore, we define the event

$$
\begin{aligned}
\mathcal{G}_7 &= \left\{ \max_{1 \leq l \leq L} \max_{1 \leq i \leq N_{\mathcal{Q}}} \left| X_{i,\cdot}^{\mathcal{Q}} b^{(l)} \right| \lesssim (\sqrt{C_0} + \sqrt{\log N_{\mathcal{Q}}}) \|b^{(l)}\|_2 \right\} \\
\mathcal{G}_8 &= \left\{ \max_{1 \leq i \leq N_{\mathcal{Q}}} \|X_{i,\cdot}^{\mathcal{Q}}\|_\infty \lesssim (\sqrt{C_0} + \sqrt{\log N_{\mathcal{Q}} + \log p}) \right\}
\end{aligned}
\tag{140}
$$

It follows from the assumption (A1) that $\mathbf{P}(\mathcal{G}_7) \lesssim N_{\mathcal{Q}}^{-c}$ and $\mathbf{P}(\mathcal{G}_8) \lesssim \min\{N_{\mathcal{Q}}, p\}^{-c}$ for some positive constant $c > 0$. Note that

$$
\begin{aligned}
\frac{1}{N_{\mathcal{Q}}} \sum_{i=1}^{N_{\mathcal{Q}}} \left| \widehat{W}_{i,1} \widehat{W}_{i,2} - W_{i,1} W_{i,2} \right| &\leq \frac{1}{N_{\mathcal{Q}}} \sum_{i=1}^{N_{\mathcal{Q}}} \left| [\widehat{b}_{init}^{(l_1)} - b^{(l_1)}]^\intercal X_{i,\cdot}^{\mathcal{Q}} [X_{i,\cdot}^{\mathcal{Q}}]^\intercal [\widehat{b}_{init}^{(k_1)} - b^{(k_1)}] \right| \\
+ \frac{1}{N_{\mathcal{Q}}} \sum_{i=1}^{N_{\mathcal{Q}}} \left| [b^{(l_1)}]^\intercal X_{i,\cdot}^{\mathcal{Q}} [X_{i,\cdot}^{\mathcal{Q}}]^\intercal [\widehat{b}_{init}^{(k_1)} - b^{(k_1)}] \right| &+ \frac{1}{N_{\mathcal{Q}}} \sum_{i=1}^{N_{\mathcal{Q}}} \left| [\widehat{b}_{init}^{(l_1)} - b^{(l_1)}]^\intercal X_{i,\cdot}^{\mathcal{Q}} [X_{i,\cdot}^{\mathcal{Q}}]^\intercal b^{(k_1)} \right|
\end{aligned}
\tag{141}
$$

By the Cauchy-Schwarz inequality, we have

$$
\begin{aligned}
&\frac{1}{N_{\mathcal{Q}}} \sum_{i=1}^{N_{\mathcal{Q}}} \left| [\widehat{b}_{init}^{(l_1)} - b^{(l_1)}]^\intercal X_{i,\cdot}^{\mathcal{Q}} [X_{i,\cdot}^{\mathcal{Q}}]^\intercal [\widehat{b}_{init}^{(k_1)} - b^{(k_1)}] \right| \\
&\leq \frac{1}{N_{\mathcal{Q}}} \sqrt{ \sum_{i=1}^{N_{\mathcal{Q}}} \left( [\widehat{b}_{init}^{(l_1)} - b^{(l_1)}]^\intercal X_{i,\cdot}^{\mathcal{Q}} \right)^2 \sum_{i=1}^{N_{\mathcal{Q}}} \left( [X_{i,\cdot}^{\mathcal{Q}}]^\intercal [\widehat{b}_{init}^{(k_1)} - b^{(k_1)}] \right)^2 }
\end{aligned}
$$

Hence, on the event $\mathcal{G}_1 \cap \mathcal{G}_6(\widehat{b}_{init}^{(k_1)} - b^{(k_1)}, \widehat{b}_{init}^{(k_1)} - b^{(k_1)}, \sqrt{\log p}) \cap \mathcal{G}_6(\widehat{b}_{init}^{(l_1)} - b^{(l_1)}, \widehat{b}_{init}^{(l_1)} - b^{(l_1)}, \sqrt{\log p})$, we have

$$
\frac{1}{N_{\mathcal{Q}}} \sum_{i=1}^{N_{\mathcal{Q}}} \left| [\widehat{b}_{init}^{(l_1)} - b^{(l_1)}]^\intercal X_{i,\cdot}^{\mathcal{Q}} [X_{i,\cdot}^{\mathcal{Q}}]^\intercal [\widehat{b}_{init}^{(k_1)} - b^{(k_1)}] \right| \lesssim \frac{s \log p}{n}
\tag{142}
$$

On the event $\mathcal{G}_7$, we have

$$
\begin{aligned}
&\frac{1}{N_{\mathcal{Q}}} \sum_{i=1}^{N_{\mathcal{Q}}} \left| [b^{(l_1)}]^\intercal X_{i,\cdot}^{\mathcal{Q}} [X_{i,\cdot}^{\mathcal{Q}}]^\intercal [\widehat{b}_{init}^{(k_1)} - b^{(k_1)}] \right| \\
&\lesssim (\sqrt{C_0} + \sqrt{\log N_{\mathcal{Q}}}) \|b^{(l_1)}\|_2 \frac{1}{N_{\mathcal{Q}}} \sum_{i=1}^{N_{\mathcal{Q}}} \left| [X_{i,\cdot}^{\mathcal{Q}}]^\intercal [\widehat{b}_{init}^{(k_1)} - b^{(k_1)}] \right| \\
&\leq (\sqrt{C_0} + \sqrt{\log N_{\mathcal{Q}}}) \|b^{(l_1)}\|_2 \frac{1}{\sqrt{N_{\mathcal{Q}}}} \sqrt{ \sum_{i=1}^{N_{\mathcal{Q}}} \left( [X_{i,\cdot}^{\mathcal{Q}}]^\intercal [\widehat{b}_{init}^{(k_1)} - b^{(k_1)}] \right)^2 }
\end{aligned}
$$

where the last inequality follows from the Cauchy-Schwarz inequality. Hence, on the event $\mathcal{G}_1 \cap \mathcal{G}_7 \cap \mathcal{G}_6(\widehat{b}_{init}^{(k_1)} - b^{(k_1)}, \widehat{b}_{init}^{(k_1)} - b^{(k_1)}, \sqrt{\log p})$, we establish

$$\frac{1}{N_{\mathcal{Q}}} \sum_{i=1}^{N_{\mathcal{Q}}} \left| [b^{(l_1)}]^{\intercal} X_{i,\cdot}^{\mathcal{Q}} [X_{i,\cdot}^{\mathcal{Q}}]^{\intercal} [\widehat{b}_{init}^{(k_1)} - b^{(k_1)}] \right| \lesssim (\sqrt{C_0} + \sqrt{\log N_{\mathcal{Q}}}) \|b^{(l_1)}\|_2 \sqrt{\frac{s \log p}{n}}. \tag{143}$$

Similarly, we establish

$$\frac{1}{N_{\mathcal{Q}}} \sum_{i=1}^{N_{\mathcal{Q}}} \left| [\widehat{b}_{init}^{(l_1)} - b^{(l_1)}]^{\intercal} X_{i,\cdot}^{\mathcal{Q}} [X_{i,\cdot}^{\mathcal{Q}}]^{\intercal} b^{(k_1)} \right| \lesssim (\sqrt{C_0} + \sqrt{\log N_{\mathcal{Q}}}) \|b^{(k_1)}\|_2 \sqrt{\frac{s \log p}{n}}.$$

Combined with (141), (142) and (143), we establish

$$\frac{1}{N_{\mathcal{Q}}} \sum_{i=1}^{N_{\mathcal{Q}}} \left| \widehat{W}_{i,1} \widehat{W}_{i,2} - W_{i,1} W_{i,2} \right| \lesssim \left( \sqrt{\log N_{\mathcal{Q}}} (\|b^{(l_1)}\|_2 + \|b^{(k_1)}\|_2) + \sqrt{\frac{s \log p}{n}} \right) \sqrt{\frac{s \log p}{n}} \tag{144}$$

Similarly, we establish

$$\frac{1}{N_{\mathcal{Q}}} \sum_{i=1}^{N_{\mathcal{Q}}} \left| \widehat{W}_{i,3} \widehat{W}_{i,4} - W_{i,3} W_{i,4} \right| \lesssim \left( \sqrt{\log N_{\mathcal{Q}}} (\|b^{(l_2)}\|_2 + \|b^{(k_2)}\|_2) + \sqrt{\frac{s \log p}{n}} \right) \sqrt{\frac{s \log p}{n}} \tag{145}$$

Define

$$\begin{aligned} H_{i,1} &= W_{i,1} W_{i,2} \quad \text{and} \quad H_{i,2} = W_{i,3} W_{i,4} \\ \widehat{H}_{i,1} &= \widehat{W}_{i,1} \widehat{W}_{i,2} \quad \text{and} \quad \widehat{H}_{i,2} = \widehat{W}_{i,3} \widehat{W}_{i,4} \end{aligned}$$

Then we have

$$\begin{aligned} &\frac{1}{N_{\mathcal{Q}}} \sum_{i=1}^{N_{\mathcal{Q}}} \prod_{t=1}^{4} \widehat{W}_{i,t} - \frac{1}{N_{\mathcal{Q}}} \sum_{i=1}^{N_{\mathcal{Q}}} \prod_{t=1}^{4} W_{i,t} \\ &= \frac{1}{N_{\mathcal{Q}}} \sum_{i=1}^{N_{\mathcal{Q}}} \widehat{H}_{i,1} \widehat{H}_{i,2} - \frac{1}{N_{\mathcal{Q}}} \sum_{i=1}^{N_{\mathcal{Q}}} H_{i,1} H_{i,2} \\ &= \frac{1}{N_{\mathcal{Q}}} \sum_{i=1}^{N_{\mathcal{Q}}} \left( \widehat{H}_{i,1} - H_{i,1} \right) H_{i,2} + \frac{1}{N_{\mathcal{Q}}} \sum_{i=1}^{N_{\mathcal{Q}}} \left( \widehat{H}_{i,2} - H_{i,2} \right) H_{i,1} + \frac{1}{N_{\mathcal{Q}}} \sum_{i=1}^{N_{\mathcal{Q}}} \left( \widehat{H}_{i,1} - H_{i,1} \right) \left( \widehat{H}_{i,2} - H_{i,2} \right) \end{aligned} \tag{146}$$

On the event $\mathcal{G}_7$, we have

$$|H_{i,1}| \lesssim (C_0 + \log N_{\mathcal{Q}}) \|b^{(l_1)}\|_2 \|b^{(k_1)}\|_2 \quad \text{and} \quad |H_{i,2}| \lesssim (C_0 + \log N_{\mathcal{Q}}) \|b^{(l_2)}\|_2 \|b^{(k_2)}\|_2 \tag{147}$$

On the event $\mathcal{G}_7 \cap \mathcal{G}_8$, we have

$$
\begin{aligned}
&\left|\widehat{H}_{i,2} - H_{i,2}\right| \leq \left|[\widehat{b}^{(l_1)}_{init} - b^{(l_1)}]^{\intercal} X^{\mathcal{Q}}_{i,\cdot}[X^{\mathcal{Q}}_{i,\cdot}]^{\intercal}[\widehat{b}^{(k_1)}_{init} - b^{(k_1)}]\right| \\
&+ \left|[b^{(l_1)}]^{\intercal} X^{\mathcal{Q}}_{i,\cdot}[X^{\mathcal{Q}}_{i,\cdot}]^{\intercal}[\widehat{b}^{(k_1)}_{init} - b^{(k_1)}]\right| + \left|[\widehat{b}^{(l_1)}_{init} - b^{(l_1)}]^{\intercal} X^{\mathcal{Q}}_{i,\cdot}[X^{\mathcal{Q}}_{i,\cdot}]^{\intercal} b^{(k_1)}\right| \\
&\lesssim (C_0 + \log N_{\mathcal{Q}} + \log p)\left(s^2\frac{\log p}{n} + s\sqrt{\frac{\log p}{n}}\|b^{(k_1)}\|_2 + s\sqrt{\frac{\log p}{n}}\|b^{(l_1)}\|_2\right)
\end{aligned}
\tag{148}
$$

By the decomposition (146), we combine (147), (148), (144) and (145) and establish

$$
\begin{aligned}
&\left|\frac{1}{N_{\mathcal{Q}}}\sum_{i=1}^{N_{\mathcal{Q}}}\prod_{t=1}^{4}\widehat{W}_{i,t} - \frac{1}{N_{\mathcal{Q}}}\sum_{i=1}^{N_{\mathcal{Q}}}\prod_{t=1}^{4}W_{i,t}\right| \\
&\leq (C_0 + \log N_{\mathcal{Q}})\|b^{(l_2)}\|_2\|b^{(k_2)}\|_2\left(\sqrt{\log N_{\mathcal{Q}}}(\|b^{(l_1)}\|_2 + \|b^{(k_1)}\|_2) + \sqrt{\frac{s\log p}{n}}\right)\sqrt{\frac{s\log p}{n}} \\
&+ \left(\sqrt{\log N_{\mathcal{Q}}}(\|b^{(l_2)}\|_2 + \|b^{(k_2)}\|_2) + \sqrt{\frac{s\log p}{n}}\right)\sqrt{\frac{s\log p}{n}} \\
&\cdot (C_0 + \log N_{\mathcal{Q}} + \log p)\left(s^2\frac{\log p}{n} + s\sqrt{\frac{\log p}{n}}\|b^{(k)}\|_2 + s\sqrt{\frac{\log p}{n}}\|b^{(l)}\|_2 + \|b^{(l_1)}\|_2\|b^{(k_1)}\|_2\right)
\end{aligned}
\tag{149}
$$

If we further assume that $\|b^{(l)}\|_2 \leq C$ for $1 \leq l \leq L$ and $s^2(\log p)^2/n \leq c$ for some positive constants $C > 0$ and $c > 0$, then we establish (132).

### C.4 Proof of Lemma 5

We shall divide the proof into two parts based on whether there is possible covariate shift or not. The proofs are slightly different but the main idea remains the same.

#### C.4.1 The setting with possible covariate shift.

Note that, for any vectors $u_1$ and $u_2$ and any positive constant $c > 0$, we have

$$
\|u_1 + u_2\|_2^2 \leq (1+c)\|u_1\|_2^2 + \left(1 + \frac{1}{c}\right)\|u_2\|_2^2 \tag{150}
$$

We take $u_1 = \sqrt{n}\text{vecl}(D)$, $u_2 = \sqrt{n}\text{vecl}(\text{Rem})$ and $u_1 + u_2 = \widehat{Z}$ and establish the lower bound,

$$\exp\left(-\frac{1}{2}\widehat{Z}^{\intercal}(\mathbf{Cov} + \frac{1}{2}d_0\mathbf{I})^{-1}\widehat{Z}\right) \geq \exp\left(-(1+c)\frac{1}{2}n[\text{vecl}(D)]^{\intercal}(\mathbf{Cov} + \frac{1}{2}d_0\mathbf{I})^{-1}[\text{vecl}(D)]\right)$$
$$\cdot \exp\left(-\left(1+\frac{1}{c}\right)\frac{1}{2}n[\text{vecl}(\text{Rem})]^{\intercal}(\mathbf{Cov} + \frac{1}{2}d_0\mathbf{I})^{-1}[\text{vecl}(\text{Rem})]\right). \tag{151}$$

where $c > 0$ is a small positive constant. Note that the approximation error in (151) can be controlled as

$$\begin{aligned} &n[\text{vecl}(\text{Rem})]^{\intercal}(\mathbf{Cov} + \frac{1}{2}d_0\mathbf{I})^{-1}[\text{vecl}(\text{Rem})]\mathbf{1}_{\mathcal{O}\in\mathcal{E}_2} \\ &\lesssim \frac{n}{d_0}\|\text{Rem}\|_F^2 \lesssim L^2 \cdot \left(\frac{s(\log p)^2}{N_{\mathcal{Q}}} + \frac{(s\log p)^2}{n}\right) \end{aligned} \tag{152}$$

where the last inequality follows from (30), Proposition 3 and the definition of $d_0$ in (22).

Additionally, we take $\text{vecl}(D^*)$ and $S^*$ as in Theorem 2 and then have

$$\begin{aligned} &\exp\left(-(1+c)\frac{1}{2}n[\text{vecl}(D^*)]^{\intercal}(\mathbf{Cov} + \frac{1}{2}d_0\mathbf{I})^{-1}\text{vecl}(D^*)\right) \\ &\geq \exp\left(-(1+c)^2\frac{1}{2}n[S^*]^{\intercal}(\mathbf{Cov} + \frac{1}{2}d_0\mathbf{I})^{-1}[S^*]\right) \\ &\cdot \exp\left(-(1+c)\left(1+\frac{1}{c}\right)\frac{1}{2}n[\text{vecl}(D)^* - S^*]^{\intercal}(\mathbf{Cov} + \frac{1}{2}d_0\mathbf{I})^{-1}(\text{vecl}(D^*) - S^*)\right) \end{aligned} \tag{153}$$

where the last inequality follows from (150). Note that the approximation error in (153) can be controlled as

$$n[\text{vecl}(D^*) - S^*]^{\intercal}(\mathbf{Cov} + \frac{1}{2}d_0\mathbf{I})^{-1}(\text{vecl}(D^*) - S^*) \lesssim \frac{n}{N_{\mathcal{Q}}^{4/3}} \quad \text{a.s.} \tag{154}$$

With $r$ denoting the rank of $\mathbf{Cov}$, we conduct the eigen-decomposition

$$\mathbf{Cov} = \begin{pmatrix} U & U_c \end{pmatrix} \begin{pmatrix} \Lambda & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \begin{pmatrix} U & U_c \end{pmatrix}^{\intercal}$$

where $\Lambda = \text{diag}(\Lambda_{1,1}, \cdots, \Lambda_{r,r})$ with $\Lambda_{1,1} \geq \cdots \geq \Lambda_{r,r} > 0$ and $U \in \mathbb{R}^{L(L+1)/2 \times r}$ denotes the matrix of eigenvectors corresponding to $\Lambda_{1,1} \geq \cdots \geq \Lambda_{r,r} > 0$, and $U_c \in \mathbb{R}^{L(L+1)/2 \times (L(L+1)/2 - r)}$ denotes the matrix of eigenvectors corresponding to zero eigenvalues. Hence, it follows from Theorem 2 that, conditioning on $X^{\mathcal{Q}}_{A,\cdot}, \{X^{(l)}, \epsilon^{(l)}_{A_l}\}_{1\leq l\leq L}$, the transformed random vector $\sqrt{n}\Lambda^{-1/2}U^{\intercal}S^* \in \mathbb{R}^m$ is Gaussian with zero mean and the diagonal covariance matrix $\mathrm{I} \in \mathbb{R}^{r\times r}$ and $\sqrt{n}U_c^{\intercal}S^*$ is Gaussian with zero mean and covariance. As a consequence, their marginal distributions remain the same, that is, $\sqrt{n}\Lambda^{-1/2}U^\intercal S^* \in \mathbb{R}^m$ is Gaussian with zero mean and the diagonal covariance matrix $\mathrm{I} \in \mathbb{R}^{r\times r}$ and $\sqrt{n}U_c^\intercal S^*$ is zero almost surely. For any given $0 < \alpha_0 < 1/2$, we have

$$\mathbf{P}\left(-n\frac{1}{2}[U^\intercal S^*]^\intercal \Lambda^{-1} U^\intercal S^* \geq -\frac{1}{2}F_{\chi_r^2}^{-1}(1-\alpha_0)\right) = 1-\alpha_0, \tag{155}$$

where $F_{\chi_r^2}^{-1}(1-\alpha_0)$ denotes the $1-\alpha_0$ quantile of the $\chi^2$ distribution with degree of freedom $r$. We define the positive constant

$$c_{\alpha_0} = \exp\left(-(1+c)^2\frac{1}{2}F_{\chi_r^2}^{-1}\left(1-\alpha_0\right)\right). \tag{156}$$

We now apply (151), (152) and the complexity condition (A2) and establish that, there exists a small positive constant $c_1 > 0$,

$$\begin{aligned}
&\mathbf{P}\left(\exp\left(-\frac{1}{2}\widehat{Z}^\intercal(\mathbf{Cov}+\frac{1}{2}d_0\mathbf{I})^{-1}\widehat{Z}\right)\cdot \mathbf{1}_{\mathcal{O}\in\mathcal{E}_1\cap\mathcal{E}_2} \geq (1+c_1)\cdot\frac{c_{\alpha_0}}{2}\right)\\
&\geq \mathbf{P}\left(\exp\left(-(1+c)\frac{1}{2}n[\mathrm{vecl}(D)]^\intercal(\mathbf{Cov}+\frac{1}{2}d_0\mathbf{I})^{-1}[\mathrm{vecl}(D)]\right)\cdot \mathbf{1}_{\mathcal{O}\in\mathcal{E}_1\cap\mathcal{E}_2} \geq \left(1+\frac{c_1}{2}\right)\cdot\frac{c_{\alpha_0}}{2}\right)\\
&\geq \mathbf{P}\left(\exp\left(-(1+c)\frac{1}{2}n[\mathrm{vecl}(D)]^\intercal(\mathbf{Cov}+\frac{1}{2}d_0\mathbf{I})^{-1}[\mathrm{vecl}(D)]\right) \geq \left(1+\frac{c_1}{2}\right)\cdot c_{\alpha_0}\right)\\
&-\mathbf{P}\left(\exp\left(-(1+c)\frac{1}{2}n[\mathrm{vecl}(D)]^\intercal(\mathbf{Cov}+\frac{1}{2}d_0\mathbf{I})^{-1}[\mathrm{vecl}(D)]\right)\mathbf{1}_{\mathcal{O}\notin\mathcal{E}_1\cap\mathcal{E}_2} \geq \left(1+\frac{c_1}{2}\right)\cdot\frac{c_{\alpha_0}}{2}\right)
\end{aligned}$$

where the second inequality follows from the union bound. By applying the above inequality and

$$\mathbf{P}\left(\exp\left(-(1+c)\frac{1}{2}n[\mathrm{vecl}(D)]^\intercal(\mathbf{Cov}+\frac{1}{2}d_0\mathbf{I})^{-1}[\mathrm{vecl}(D)]\right)\mathbf{1}_{\mathcal{O}\notin\mathcal{E}_1\cap\mathcal{E}_2} \geq \left(1+\frac{c_1}{2}\right)\cdot\frac{c_{\alpha_0}}{2}\right) \leq \mathbf{P}((\mathcal{E}_1\cap\mathcal{E}_2)^c),$$

we establish

$$\begin{aligned}
&\mathbf{P}\left(\exp\left(-\frac{1}{2}\widehat{Z}^\intercal(\mathbf{Cov}+\frac{1}{2}d_0\mathbf{I})^{-1}\widehat{Z}\right)\cdot \mathbf{1}_{\mathcal{O}\in\mathcal{E}_1\cap\mathcal{E}_2} \geq (1+c_1)\cdot\frac{c_{\alpha_0}}{2}\right)\\
&\geq \mathbf{P}\left(\exp\left(-(1+c)\frac{1}{2}n[\mathrm{vecl}(D)]^\intercal(\mathbf{Cov}+\frac{1}{2}d_0\mathbf{I})^{-1}[\mathrm{vecl}(D)]\right) \geq \left(1+\frac{c_1}{2}\right)\cdot c_{\alpha_0}\right) - \mathbf{P}((\mathcal{E}_1\cap\mathcal{E}_2)^c)\\
&= \mathbf{P}\left(\exp\left(-(1+c)\frac{1}{2}n[\mathrm{vecl}(D^*)]^\intercal(\mathbf{Cov}+\frac{1}{2}d_0\mathbf{I})^{-1}\mathrm{vecl}(D^*)\right) \geq \left(1+\frac{c_1}{2}\right)\cdot c_{\alpha_0}\right) - \mathbf{P}((\mathcal{E}_1\cap\mathcal{E}_2)^c)\\
&\geq \mathbf{P}\left(\exp\left(-(1+c)^2\frac{1}{2}n[S^*]^\intercal(\mathbf{Cov}+\frac{1}{2}d_0\mathbf{I})^{-1}S^*\right) \geq c_{\alpha_0}\right) - \mathbf{P}((\mathcal{E}_1\cap\mathcal{E}_2)^c)
\end{aligned} \tag{157}$$

where the equality follows from $[\mathrm{vecl}(D)] \overset{d}{=} \mathrm{vecl}(D^*)$ and the last inequality follows from (153), (154) and the complexity condition (A2). Hence, it is sufficient to control the lower bound in (157). Note that

$$\begin{aligned}
&\mathbf{P}\left(\exp\left(-(1+c)^2 n\frac{1}{2}[S^*]^\intercal(\mathbf{Cov}+\frac{1}{2}d_0\mathbf{I})^{-1}[S^*]\right)\geq c_{\alpha_0}\right)\\
&=\mathbf{P}\left(\exp\left(-(1+c)^2 n\frac{1}{2}[S^*]^\intercal\left(\begin{pmatrix}U & U_c\end{pmatrix}\begin{pmatrix}\Lambda & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\begin{pmatrix}U & U_c\end{pmatrix}^\intercal+\frac{1}{2}d_0\mathbf{I}\right)^{-1}S^*\right)\geq c_{\alpha_0}\right)\\
&=\mathbf{P}\left(\exp\left(-(1+c)^2 n\frac{1}{2}[U^\intercal S^*]^\intercal(\Lambda+\frac{1}{2}d_0\mathbf{I})^{-1}U^\intercal S^*\right)\geq c_{\alpha_0}\right)\\
&\geq\mathbf{P}\left(\exp\left(-(1+c)^2 n\frac{1}{2}[U^\intercal S^*]^\intercal\Lambda^{-1}U^\intercal S^*\right)\geq c_{\alpha_0}\right)=1-\frac{\alpha_0}{2},
\end{aligned}\tag{158}$$

where the first equality follows from the eigen-decomposition of $\mathbf{Cov}$, the second equality follows from the fact that $\sqrt{n}U_c^\intercal S^*$ is zero almost surely and the last equality follows from the definition of $c_{\alpha_0}$ in (156). We establish (77) by combining (157), (156) and (158) with $c=\sqrt{2}-1$.

#### C.4.2 The setting with no covariate shift.

The proof is similar to the setting with possible covariate shift. We mainly highlight the difference in the proofs. We recall $\mathbf{V}=\mathbf{V}^{(1)}+\mathbf{V}^{(2)}$ and define $\mathbf{Cov}=n\cdot\mathbf{V}$ and

$$\mathbf{Cov}^{(1)}=n\cdot\mathbf{V}^{(1)}\quad\text{and}\quad\mathbf{Cov}^{(2)}=n\cdot\mathbf{V}^{(2)}.$$

Similarly to (151), we establish

$$\begin{aligned}
&\exp\left(-\frac{1}{2}\widehat{Z}^\intercal(\mathbf{Cov}+\frac{1}{2}d_0\mathbf{I})^{-1}\widehat{Z}\right)\\
&\geq\exp\left(-(1+c)n\frac{1}{2}[\mathrm{vecl}(D]^\intercal(\mathbf{Cov}+\frac{1}{2}d_0\mathbf{I})^{-1}[\mathrm{vecl}(D)]\right)\\
&\cdot\exp\left(-\left(1+\frac{1}{c}\right)\frac{1}{2}n[\mathrm{vecl}(\mathrm{Rem})]^\intercal(\mathbf{Cov}+\frac{1}{2}d_0\mathbf{I})^{-1}[\mathrm{vecl}(\mathrm{Rem})]\right).
\end{aligned}\tag{159}$$

where $c>0$ is a small positive constant. Similarly to (160), we control the last term on the right hand side of (159) by

$$n[\mathrm{vecl}(\mathrm{Rem})]^\intercal(\mathbf{Cov}+\frac{1}{2}d_0\mathbf{I})^{-1}[\mathrm{vecl}(\mathrm{Rem})]\mathbf{1}_{\mathcal{O}\in\mathcal{E}_2}\lesssim\frac{n}{d_0}\|\mathrm{Rem}\|_F^2\lesssim L^2\cdot\frac{(s\log p)^2}{n}\tag{160}$$

where the inequality follows from Proposition 5. We now apply (159), (160) and the complexity condition (A2) and establish that, for any constant $c_2 > 0$, there exists a small positive constant $c_1 > 0$,

$$
\begin{aligned}
&\mathbf{P}\left(\exp\left(-\frac{1}{2}\widehat{Z}^\intercal(\mathbf{Cov}+\frac{1}{2}d_0\mathbf{I})^{-1}\widehat{Z}\right)\cdot \mathbf{1}_{\mathcal{O}\in\mathcal{E}_1\cap\mathcal{E}_2} \geq (1+c_1)\cdot\frac{c_2}{2}\right)\\
&\geq \mathbf{P}\left(\exp\left(-(1+c)n[\mathrm{vecl}(D)]^\intercal(\mathbf{Cov}+\frac{1}{2}d_0\mathbf{I})^{-1}[\mathrm{vecl}(D)]\right)\cdot \mathbf{1}_{\mathcal{O}\in\mathcal{E}_1\cap\mathcal{E}_2} \geq \left(1+\frac{c_1}{2}\right)\cdot\frac{c_2}{2}\right)\\
&\geq \mathbf{P}\left(\exp\left(-(1+c)n[\mathrm{vecl}(D)]^\intercal(\mathbf{Cov}+\frac{1}{2}d_0\mathbf{I})^{-1}[\mathrm{vecl}(D)]\right) \geq \left(1+\frac{c_1}{2}\right)\cdot c_2\right)\\
&- \mathbf{P}\left(\exp\left(-(1+c)n[\mathrm{vecl}(D)]^\intercal(\mathbf{Cov}+\frac{1}{2}d_0\mathbf{I})^{-1}[\mathrm{vecl}(D)]\right)\mathbf{1}_{\mathcal{O}\notin\mathcal{E}_1\cap\mathcal{E}_2} \geq \left(1+\frac{c_1}{2}\right)\cdot\frac{c_2}{2}\right)\\
&\geq \mathbf{P}\left(\exp\left(-(1+c)n[\mathrm{vecl}(D)]^\intercal(\mathbf{Cov}+\frac{1}{2}d_0\mathbf{I})^{-1}[\mathrm{vecl}(D)]\right) \geq \left(1+\frac{c_1}{2}\right)\cdot c_2\right) - \mathbf{P}((\mathcal{E}_1\cap\mathcal{E}_2)^c)
\end{aligned}
\tag{161}
$$

where the second inequality follows from the union bound.

We now control the term $\exp\left(-(1+c)n[\mathrm{vecl}(D)]^\intercal(\mathbf{Cov}+\frac{1}{2}d_0\mathbf{I})^{-1}[\mathrm{vecl}(D)]\right)$ by the following inequality,

$$
\begin{aligned}
&\exp\left(-(1+c)n\frac{1}{2}[\mathrm{vecl}(D)]^\intercal(\mathbf{Cov}+\frac{1}{2}d_0\mathbf{I})^{-1}[\mathrm{vecl}(D)]\right)\\
&\geq \exp\left(-(1+c)n[\mathrm{vecl}(D^{(1)})]^\intercal(\mathbf{Cov}+\frac{1}{2}d_0\mathbf{I})^{-1}[\mathrm{vecl}(D^{(1)})]\right)\\
&\cdot\exp\left(-(1+c)n[\mathrm{vecl}(D^{(2)})]^\intercal(\mathbf{Cov}+\frac{1}{2}d_0\mathbf{I})^{-1}[\mathrm{vecl}(D^{(2)})]\right)
\end{aligned}
\tag{162}
$$

Additionally, we take $\mathrm{vecl}(D^{(2),*})$ and $T^*$ as in Proposition 5 and then have

$$
\begin{aligned}
&\exp\left(-(1+c)n[\mathrm{vecl}(D^{(2)})]^\intercal(\mathbf{Cov}+\frac{1}{2}d_0\mathbf{I})^{-1}[\mathrm{vecl}(D^{(2)})]\right)\\
&\overset{d}{=}\exp\left(-(1+c)n[\mathrm{vecl}(D^{(2),*})]^\intercal(\mathbf{Cov}+\frac{1}{2}d_0\mathbf{I})^{-1}\mathrm{vecl}(D^{(2),*})\right)\\
&\geq \exp\left(-(1+c)^2n[T^*]^\intercal(\mathbf{Cov}+\frac{1}{2}d_0\mathbf{I})^{-1}[T^*]\right)\\
&\cdot\exp\left(-(1+c)\left(1+\frac{1}{c}\right)\frac{1}{2}n[\mathrm{vecl}(D^{(2),*})-T^*]^\intercal(\mathbf{Cov}+\frac{1}{2}d_0\mathbf{I})^{-1}(\mathrm{vecl}(D^{(2),*})-T^*)\right)
\end{aligned}
\tag{163}
$$

where the last inequality follows from (150). Note that the approximation error in (163) can be controlled as

$$
n[\mathrm{vecl}(D^{(2),*})-T^*]^\intercal(\mathbf{Cov}+\frac{1}{2}d_0\mathbf{I})^{-1}(\mathrm{vecl}(D^{(2),*})-T^*)\lesssim\frac{n}{(\sum_{l=1}^n n_l+N_{\mathcal{Q}})^{4/3}}\quad \text{a.s.}
\tag{164}
$$

For $j = 1, 2$, with $r_j$ denoting the rank of $\mathbf{Cov}^{(j)}$, we conduct the eigen-decomposition

$$\mathbf{Cov}^{(j)} = \begin{pmatrix} U^{(j)} & U_c^{(j)} \end{pmatrix} \begin{pmatrix} \Lambda^{(j)} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \begin{pmatrix} U^{(j)} & U_c^{(j)} \end{pmatrix}^\intercal$$

where $\Lambda^{(j)} = \mathrm{diag}(\Lambda_{1,1}^{(j)}, \cdots, \Lambda_{r_j,r_j}^{(j)})$ with $\Lambda_{1,1}^{(j)} \geq \cdots \geq \Lambda_{r_j,r_j}^{(j)} > 0$ and $U^{(j)} \in \mathbb{R}^{L(L+1)/2 \times r_j}$ denotes the matrix of eigenvectors corresponding to $\Lambda_{1,1}^{(j)} \geq \cdots \geq \Lambda_{r_j,r_j}^{(j)} > 0$, and $U_c^{(j)} \in \mathbb{R}^{L(L+1)/2 \times (L(L+1)/2 - r_j)}$ denotes its orthogonal complement matrix. We apply the same argument for (155) and establish that

$$\mathbf{P}\left(-n[\mathrm{vecl}(D^{(1)})]^\intercal U^{(1)}[\Lambda^{(1)}]^{-1}[U^{(1)}]^\intercal \mathrm{vecl}(D^{(1)}) \leq -F^{-1}_{\chi^2_{r_1}}\left(1 - \frac{\alpha_0}{2}\right)\right) \leq \frac{\alpha_0}{2}, \tag{165}$$

$$\mathbf{P}\left(-n(T^*)^\intercal U^{(2)}[\Lambda^{(2)}]^{-1}[U^{(2)}]^\intercal T^* \leq -F^{-1}_{\chi^2_{r_2}}\left(1 - \frac{\alpha_0}{2}\right)\right) \leq \frac{\alpha_0}{2}, \tag{166}$$

and $[U_c^{(1)}]^\intercal \mathrm{vecl}(D^{(1)})$ and $[U_c^{(2)}]^\intercal T^*$is zero almost surely where $F^{-1}_{\chi^2_r}\left(1 - \frac{\alpha_0}{2}\right)$ denote the $1 - \frac{\alpha_0}{2}$ quantile of the $\chi^2$ distribution with degree of freedom $r$.

We define the positive constant

$$c_{\alpha_0}^{(1)} = \exp\left(-(1+c)F^{-1}_{\chi^2_{r_1}}\left(1 - \frac{\alpha_0}{2}\right)\right) \quad \text{and} \quad c_{\alpha_0}^{(2)} = \exp\left(-(1+c)^2 F^{-1}_{\chi^2_{r_2}}\left(1 - \frac{\alpha_0}{2}\right)\right). \tag{167}$$

By applying (161) with $c_2 = c_{\alpha_0}^{(1)} c_{\alpha_0}^{(2)}$, we have

$$\begin{aligned}
&\mathbf{P}\left(\exp\left(-\frac{1}{2}\widehat{Z}^\intercal(\mathbf{Cov} + \frac{1}{2}d_0\mathbf{I})^{-1}\widehat{Z}\right) \cdot \mathbf{1}_{\mathcal{O} \in \mathcal{E}_1 \cap \mathcal{E}_2} \geq (1 + c_1) \cdot \frac{c_{\alpha_0}^{(1)} c_{\alpha_0}^{(2)}}{2}\right) \\
&\geq \mathbf{P}\left(\exp\left(-(1+c)n[\mathrm{vecl}(D)]^\intercal(\mathbf{Cov} + \frac{1}{2}d_0\mathbf{I})^{-1}[\mathrm{vecl}(D)]\right) \geq \left(1 + \frac{c_1}{2}\right) \cdot c_{\alpha_0}^{(1)} c_{\alpha_0}^{(2)}\right) - \mathbf{P}((\mathcal{E}_1 \cap \mathcal{E}_2)^c)
\end{aligned} \tag{168}$$

By (162), we know that

$$\begin{aligned}
&\mathbf{P}\left(\exp\left(-(1+c)n[\mathrm{vecl}(D)]^\intercal(\mathbf{Cov} + \frac{1}{2}d_0\mathbf{I})^{-1}[\mathrm{vecl}(D)]\right) \geq \left(1 + \frac{c_1}{2}\right) \cdot c_{\alpha_0}^{(1)} c_{\alpha_0}^{(2)}\right) \\
&\geq 1 - \mathbf{P}\left(\exp\left(-(1+c)n[\mathrm{vecl}(D^{(1)})]^\intercal(\mathbf{Cov} + \frac{1}{2}d_0\mathbf{I})^{-1}\mathrm{vecl}(D^{(1)})\right) \leq c_{\alpha_0}^{(1)}\right) \\
&- \mathbf{P}\left(\exp\left(-(1+c)n[\mathrm{vecl}(D^{(2),*})]^\intercal(\mathbf{Cov} + \frac{1}{2}d_0\mathbf{I})^{-1}\mathrm{vecl}(D^{(2),*})\right) \leq \left(1 + \frac{c_1}{2}\right) \cdot c_{\alpha_0}^{(2)}\right)
\end{aligned} \tag{169}$$

Note that

$$
\begin{aligned}
&\mathbf{P}\left(\exp\left(-(1+c)n[\mathrm{vecl}(D^{(1)})]^{\intercal}(\mathbf{Cov}+\frac{1}{2}d_0\mathbf{I})^{-1}\mathrm{vecl}(D^{(1)})\right)\leq c^{(1)}_{\alpha_0}\right)\\
&\leq \mathbf{P}\left(\exp\left(-(1+c)n[\mathrm{vecl}(D^{(1)})]^{\intercal}(\mathbf{Cov}^{(1)}+\frac{1}{2}d_0\mathbf{I})^{-1}\mathrm{vecl}(D^{(1)})\right)\leq c^{(1)}_{\alpha_0}\right)\\
&= \mathbf{P}\left(\exp\left(-(1+c)n[\mathrm{vecl}(D^{(1)})]^{\intercal}\left(\begin{pmatrix}U^{(1)} & U_c^{(1)}\end{pmatrix}\begin{pmatrix}\Lambda^{(1)} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\begin{pmatrix}U^{(1)} & U_c^{(1)}\end{pmatrix}^{\intercal}+\frac{1}{2}d_0\mathbf{I}\right)^{-1}\mathrm{vecl}(D^{(1)})\right)\leq c^{(1)}_{\alpha_0}\right)\\
&\leq \mathbf{P}\left(-n[\mathrm{vecl}(D^{(1)})]^{\intercal}U^{(1)}[\Lambda^{(1)}]^{-1}[U^{(1)}]^{\intercal}\mathrm{vecl}(D^{(1)})\leq -F^{-1}_{\chi^2_{r_1}}\left(1-\frac{\alpha_0}{2}\right)\right)\leq\frac{\alpha_0}{2}
\end{aligned}
\tag{170}
$$

where the first inequality follows from the fact that $\mathbf{Cov}-\mathbf{Cov}^{(1)}$ is positive definite, the third inequality follows from $d_0>0$ and the final upper bound follows from (165). Similarly, we can show that

$$
\mathbf{P}\left(\exp\left(-(1+c)n[\mathrm{vecl}(D^{(2),*})]^{\intercal}(\mathbf{Cov}+\frac{1}{2}d_0\mathbf{I})^{-1}\mathrm{vecl}(D^{(2),*})\right)\leq\left(1+\frac{c_1}{2}\right)\cdot c^{(2)}_{\alpha_0}\right)\leq\frac{\alpha_0}{2}.
$$

Hence, combined with (168),(169) and (170), we establish the lemma for the setting with no covariate shift.

# D Additional Simulation

## D.1 Bias-variance tradeoff in high dimensions

We consider the no covariate shift setting and compare the proposed estimator in (21) with the plug-in estimators in (10) with $\widetilde{b}^{(l)}$ taken as Lasso estimator [40] or the debiased Lasso estimator [25]. We set $L=2$, generate $b^{(1)}\in\mathbb{R}^p$ as $b^{(1)}_j=j/40$ for $1\leq j\leq 10$, $b^{(1)}_j=(10-j)/40$ for $11\leq j\leq 20$ and $b^{(1)}_j=0$ for $21\leq j\leq 500$ and generate $b^{(2)}\in\mathbb{R}^p$ as $b^{(2)}_j=b^{(2)}_j+0.3$ for $1\leq j\leq 10$, $b^{(2)}_j=0.3$ for $11\leq j\leq 20$ and $b^{(2)}_j=0$ for $21\leq j\leq 500$. In Figure 8, We report average absolute bias, average standard error and average proportion of variance out of the total mean squared error. Since our goal is to estimate the lower triangular part $(\Gamma_{1,1},\Gamma_{2,1},\Gamma_{2,2})$ of the matrix $\Gamma$, we average the corresponding accuracy measures of estimating these three entries. The plug-in Lasso estimator has a larger bias than our proposed estimator while the plug-in debiased Lasso estimator has a large bias and variance. The variance proportion of our proposed estimator is much higher than those for the plug-in estimators, which indicates the success of bias correction and the reliable inference performance by quantifying the uncertainty of the variance component.

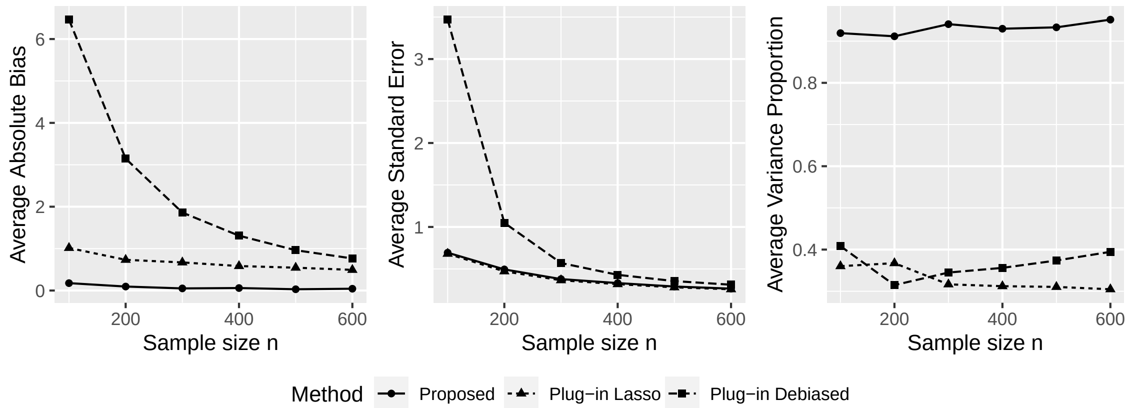


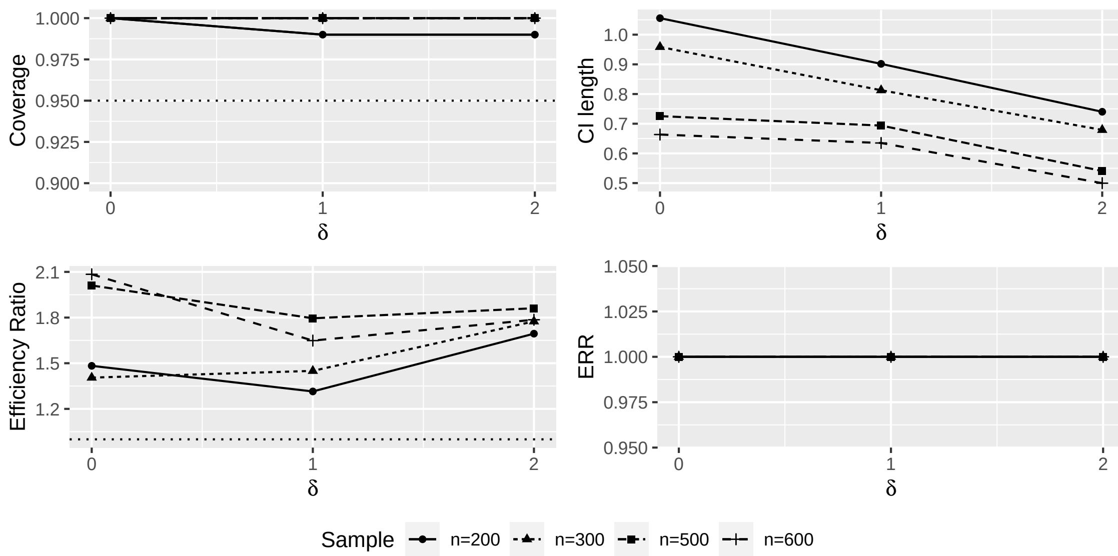

Figure 8: Comparison of our proposed estimator $\widehat{\Gamma}^{\mathcal{Q}}$ of $\Gamma$, the plug-in Lasso estimator and the Plug-in Debiased Lasso estimator.

## D.2 Additional simulation for covariate shift

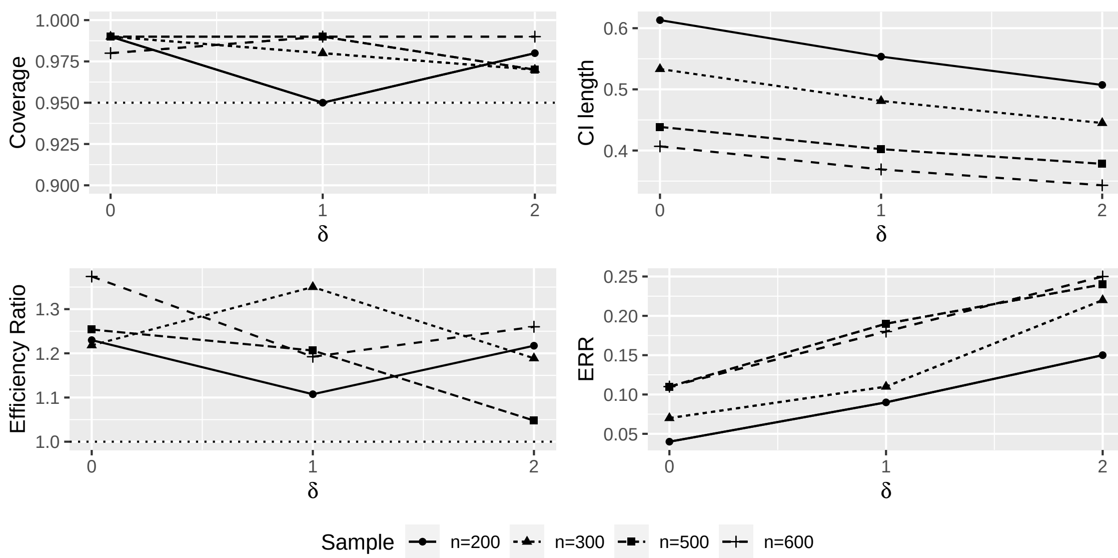
We report the simulation results corresponding to settings 2 and 3 in Section 7. We vary $\delta$ across $\{0,1,2\}$, where the corresponding values $x_{\rm new}^{\intercal}\beta^*_{\delta}$ for setting 2 are $\{0.977, 1.124, 1.287\}$ and for setting 3 are $\{-0.115, -0.136, -0.147\}$. The main observations are similar to those in setting 1, where the CI lengths decrease with a larger value of $\delta$ and a larger sample size. We shall point out some differences here. First, from Figure 3, the recommended choices of $\delta$ are 2 and 0 for settings 2 and 3, respectively. For setting 2, with increasing $\delta$ from 0 to 2, the corresponding confidence intervals drop by 25%, which indicates that the ridge-type maximin effect is more stable. Second, in settings 2 and 3, the constructed confidence intervals are over-coverage; for $n = 500$, the relative efficiency is around 1.8 for setting 2 and around 1.2 for setting 3.

## D.3 Additional method comparison

We present additional simulation results for the simulation settings considered in Figure 5. We simply replace the sample size $n = 500$ with $n \in \{100, 200, 300\}$ and present the corresponding results in Figures 10, 11 and 12, respectively.

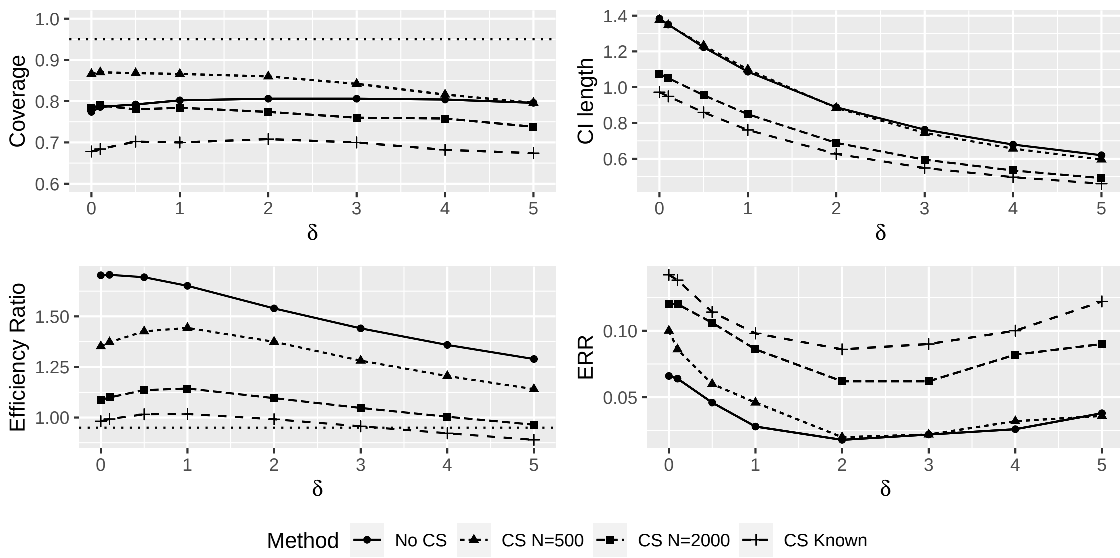



(a) Simulation setting 2 (covariate shift, L=5) with unknown $\Sigma^{\mathcal{Q}}$.

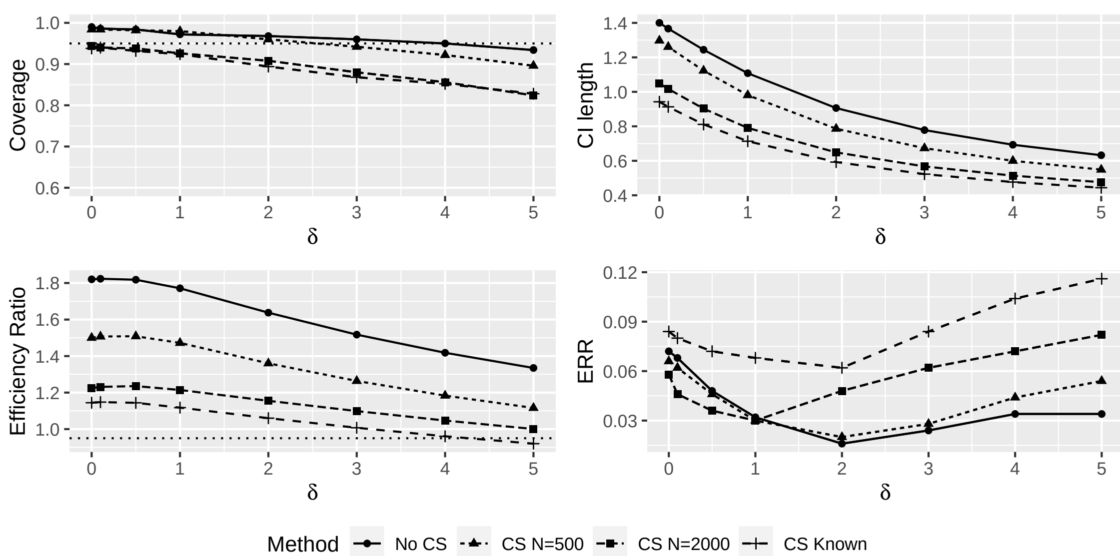



(b) Simulation setting 3 (covariate shift, L=5) with unknown $\Sigma^{\mathcal{Q}}$.

Figure 9: Dependence on $\delta$ and $n$. "Coverage" and "CI Length" stand for the empirical coverage and the average length of our proposed CI; "Efficiency Ratio" represents the ratio of the length of CI in (28) to an oracle CI relying on normality; "ERR" represents the empirical rejection rate out of 500 simulations.

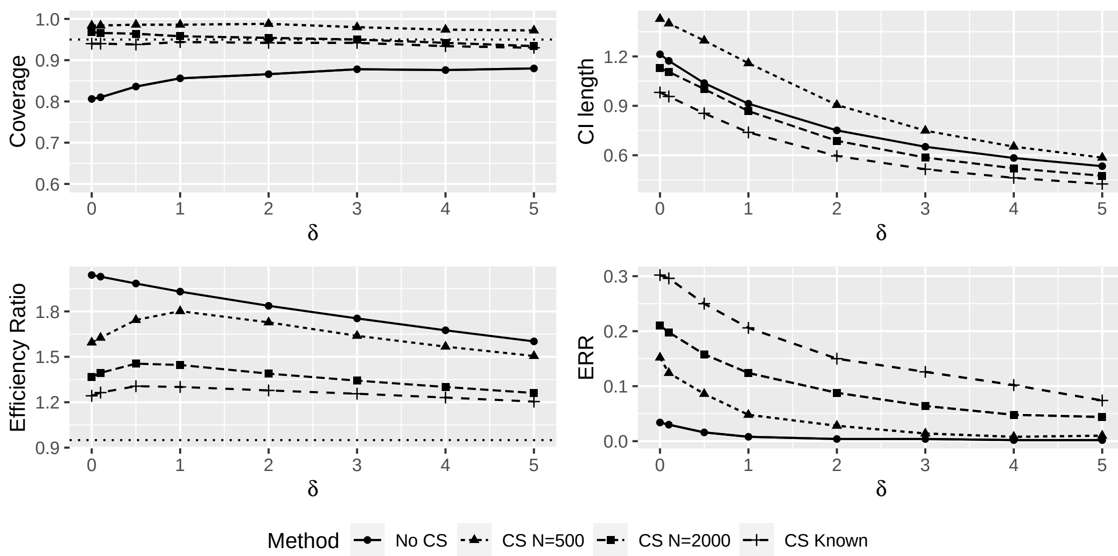



(a) Simulation settings with covariate shift

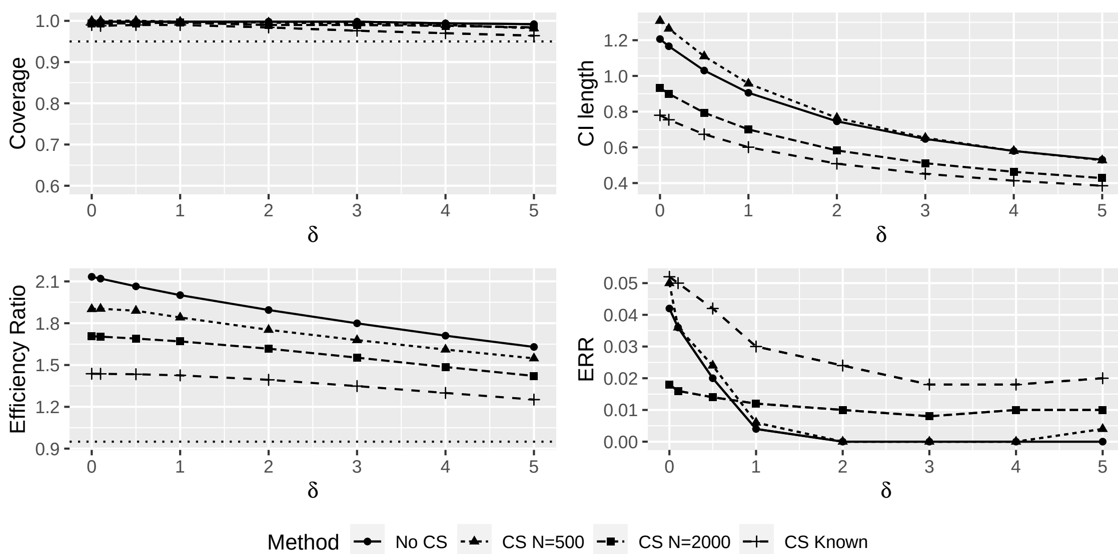



(b) Simulation settings with no covariate shift

Figure 10: Comparison of covariate shift and no covariate shift algorithms with $n = 100$. The methods "No CS", "CS $N = 500$", "CS $N = 2000$", "CS Known" represent algorithms assuming no covariate shift, Algorithm 1 with $N_{\mathcal{Q}} = 500$, with $N_{\mathcal{Q}} = 2000$ and known $\Sigma^{\mathcal{Q}}$, respectively. "Coverage" and "CI Length" stand for the empirical coverage and the average length of our proposed CI; "Efficiency Ratio" represents the ratio of the length of CI in (28) to an oracle CI relying on normality; "ERR" represents the empirical rejection rate out of 500 simulations.

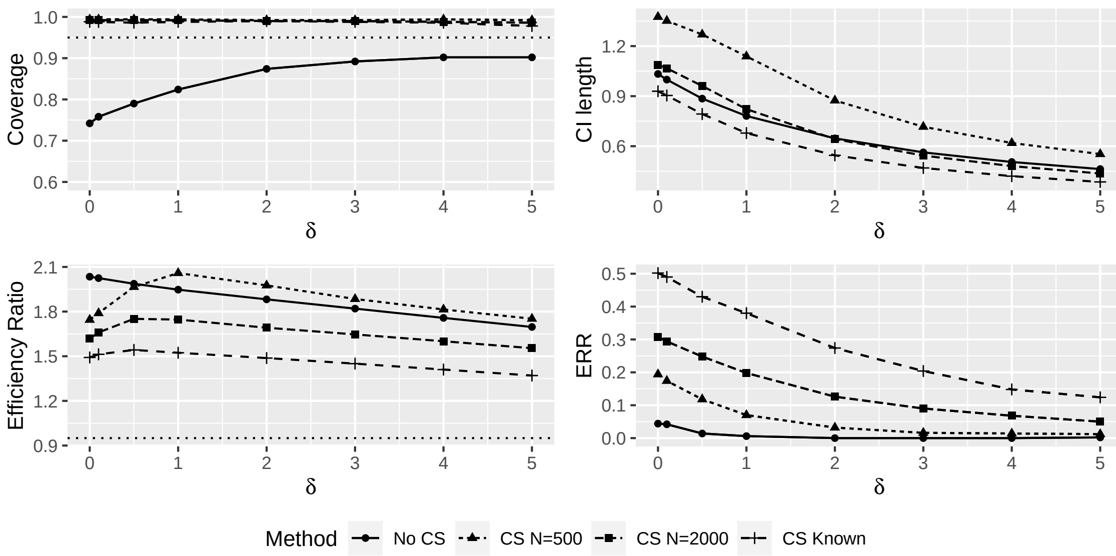



(a) Simulation settings with covariate shift

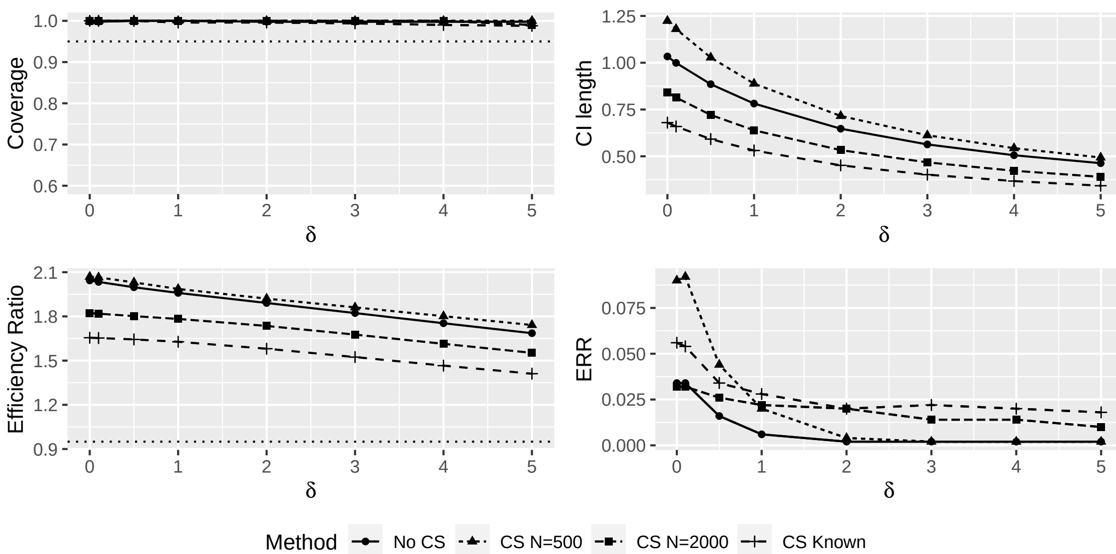



(b) Simulation settings with no covariate shift

Figure 11: Comparison of covariate shift and no covariate shift algorithms with $n = 200$. The methods "No CS", "CS $N = 500$", "CS $N = 2000$", "CS Known" represent algorithms assuming no covariate shift, Algorithm 1 with $N_{\mathcal{Q}} = 500$, with $N_{\mathcal{Q}} = 2000$ and known $\Sigma^{\mathcal{Q}}$, respectively. "Coverage" and "CI Length" stand for the empirical coverage and the average length of our proposed CI; "Efficiency Ratio" represents the ratio of the length of CI in (28) to an oracle CI relying on normality; "ERR" represents the empirical rejection rate out of 500 simulations.




(a) Simulation settings with covariate shift




(b) Simulation settings with no covariate shift

Figure 12: Comparison of covariate shift and no covariate shift algorithms with $n = 300$. The methods "No CS", "CS $N = 500$", "CS $N = 2000$", "CS Known" represent algorithms assuming no covariate shift, Algorithm 1 with $N_{\mathcal{Q}} = 500$, with $N_{\mathcal{Q}} = 2000$ and known $\Sigma^{\mathcal{Q}}$, respectively. "Coverage" and "CI Length" stand for the empirical coverage and the average length of our proposed CI; "Efficiency Ratio" represents the ratio of the length of CI in (28) to an oracle CI relying on normality; "ERR" represents the empirical rejection rate out of 500 simulations.