Blocking, rerandomization, and regression adjustment in randomized experiments with high-dimensional covariates

Ke Zhu

Center for Statistical Science, Department of Industrial Engineering, Tsinghua University, Beijing, 100084, China.

Hanzhong Liu

Center for Statistical Science, Department of Industrial Engineering, Tsinghua University, Beijing, 100084, China.

Yuehan Yang†

School of Statistics and Mathematics, Central University of Finance and Economics, Beijing, 102206, China.

Summary. Blocking, a special case of rerandomization, is routinely implemented in the design stage of randomized experiments to balance baseline covariates. Regression adjustment is highly encouraged in the analysis stage to adjust for the remaining covariate imbalances. Researchers have recommended combining these techniques; however, the research on this combination in a randomization-based inference framework with a large number of covariates is limited. This paper proposes several methods that combine the blocking, rerandomization, and regression adjustment techniques in randomized experiments with high-dimensional covariates. In the design stage, we suggest the implementation of blocking or rerandomization or both techniques to balance a fixed number of covariates most relevant to the outcomes. For the analysis stage, we propose regression adjustment methods based on the Lasso to adjust for the remaining imbalances in the additional high-dimensional covariates. Moreover, we establish the asymptotic properties of the proposed Lasso-adjusted average treatment effect estimators and outline conditions under which these estimators are more efficient than the unadjusted estimators. In addition, we provide conservative variance estimators to facilitate valid inferences. Our analysis is randomization-based, allowing the outcome data generating models to be misspecified. Simulation studies and two real data analyses demonstrate the advantages of the proposed methods.

Keywords: Lasso, Randomization inference, Regression adjustment, Rerandomization, Stratified randomization

1. Introduction

Randomized experiments are the basis for evaluating the effect of a treatment on an outcome and are widely used in the fields of industry, social sciences, and biomedical sciences (see, e.g., Fisher, 1935; Box et al., 2005; Imbens and Rubin, 2015; Rosenberger

†Corresponding author: yyh@cufe.edu.cn

and Lachin, 2015). In randomized experiments, complete randomization of the treatment assignments can balance the baseline covariates on average. However, scholars recognized that covariate imbalances often occur in a particular treatment assignment (see, e.g., Fisher, 1926; Rosenberger and Sverdlov, 2008; Morgan and Rubin, 2012; Athey and Imbens, 2017). To increase the estimation efficiency of the treatment effect, certain researchers have recommended balancing the key covariates in the design stage (Fisher, 1926; Efron, 1971; Zelen, 1974; Morgan and Rubin, 2012; Krieger et al., 2019), whereas other researchers have recommended implementing adjustments for the covariate imbalances in the analysis stage (Fisher, 1935; Yang and Tsiatis, 2001; Miratrix et al., 2013; Lin, 2013; Bloniarz et al., 2016).

Fisher (1926) was the first to recommend the use of blocking or stratification in the design stage to balance a few discrete covariates that were the most relevant to the outcomes. Since this research, blocking as an experimental design method has been widely used and investigated (see, e.g., Wilk, 1955; Green and Byar, 1978; Imai et al., 2008; Imbens, 2011; Higgins et al., 2015; Schochet, 2016; Pashley and Miratrix, 2020). While blocking can effectively balance discrete covariates, the balancing of continuous covariates through this approach is less intuitive. Rerandomization is a more general approach to balance both discrete and continuous covariates (Morgan and Rubin, 2012, 2015; Branson et al., 2016; Zhou et al., 2018; Li et al., 2018, 2020). Recently, scholars have recommended combining blocking and rerandomization techniques (Schultzberg and Johansson, 2019; Wang et al., 2020). D. B. Rubin summarized this design strategy as "Block what you can and rerandomize what you cannot".

Notably, blocking or rerandomization or their combination can balance only a fixed number of covariates. However, in modern randomized experiments, a large number of baseline covariates are often collected, and the number of covariates can be larger than the sample size. For example, in a clinical trial, the researcher may record the demographic and genetic information of each patient. Bloniarz et al. (2016) highlighted that in such high-dimensional settings, most of the covariates may not be related to the outcomes, and thus, the important covariates must be selected to realize efficient treatment effect estimation. In the design stage, in the event in which pre-experimental data (outcomes and covariates) were available, Johansson and Schultzberg (2020) used the Lasso (Tibshirani, 1996) for variable selection and rerandomization to balance those selected covariates. However, when pre-experimental outcome information is not available, it is difficult to perform covariate selection in the design stage. A more realistic approach is to use the Lasso in the analysis stage to simultaneously perform covariate selection and regression adjustment. Regression adjustment has been widely used to analyze randomized experiments and increase the associated efficiency (see, e.g., Fisher, 1935; Miratrix et al., 2013; Lin, 2013; Bloniarz et al., 2016; Liu and Yang, 2020). Li and Ding (2020) showed that regression adjustment is equivalent to rerandomization (with a threshold tending to zero) for a certain set of covariates.

In contrast to regression adjustment based on the Lasso, blocking and rerandomization techniques do not use the outcome data and thus can avoid bias due to the specification search of the outcome model. Many scholars have indicated that it is preferable to avoid the occurrence of covariate imbalances before treatment is administered, rather than performing post-treatment regression adjustment (Cox, 2007; Freedman, 2008a,b;

Rosenbaum, 2010). In this context, a trade-off method is to combine blocking, rerandomization, and high-dimensional regression adjustment. In such a method, blocking or rerandomization or both are implemented in the design stage to balance a few covariates that most significantly affect the outcomes. Next, we perform covariate selection and regression adjustment by using, for example, the Lasso in the analysis stage to adjust for the remaining imbalances in the additional high-dimensional covariates and further increase the efficiency.

Recently, researchers have proposed methods to combine blocking, rerandomization, and regression adjustment in randomized experiments with low-dimensional covariates. Bugni et al. (2018) showed that the regression of the outcome on the treatment and stratum indicators can increase the estimation efficiency of the average treatment effect in covariate-adaptive randomized experiments, including the randomized block (stratified randomized) experiments as special cases. When the propensity scores (proportion of treated units in each block) differ across blocks, this method yields an inconsistent estimator. To address this issue, Bugni et al. (2019) proposed a weighted regression-adjusted estimator. Moreover, several scholars have discussed regression adjustment based on additional covariates beyond the stratum indicators Ma et al. (2020a); Wang et al. (2019); Ye et al. (2021); Ma et al. (2020b); Ye et al. (2020); Liu et al. (2020). Notably, these studies considered a supper-population framework and assumed that the number of blocks was fixed with their sizes tending to infinity. Considering a finite population framework, Liu and Yang (2020) proposed a weighted regression adjustment method for randomized block experiments and demonstrated that this approach could increase the efficiency even when the number of blocks tended to infinity with the block sizes being fixed. However, this method assumes homogeneous propensity scores across blocks, can manage only low-dimensional covariates, and cannot consider rerandomization in the design stage. Li and Ding (2020) established a unified theory for rerandomization followed by regression adjustment. However, the authors did not consider high-dimensional covariates or the combination of blocking and rerandomization in the design stage.

Overall, the research on the simultaneous implementation of blocking, rerandomization, and regression adjustment and consideration of high-dimensional covariates is limited. To address this aspect, in this study, we develop approaches and theoretical guarantees for the abovementioned combination in general settings with both homogeneous and heterogeneous block sizes, propensity scores, and treatment effects. Our asymptotic results are obtained under a finite population and randomization-based inference framework. The potential outcomes and covariates are fixed quantities, and the sole source of randomness is the treatment assignment. Thus, we allow the outcome data generating models to be mis-specified. Our contributions are three-fold and can be summarized as follows.

First, we propose two approaches to combine blocking and high-dimensional regression adjustment, specifically, two Lasso-adjusted average treatment effect estimators in randomized block (stratified randomized) experiments. We show that under mild conditions, both estimators are consistent and asymptotically normal. Moreover, the asymptotic variance of the first estimator is no greater than that of the unadjusted estimator when the propensity scores are the same across blocks. The second estimator is obtained from a projection perspective and is asymptotically more efficient than, or at least as efficient as, the unadjusted estimator, even when the propensity scores differ across blocks. In addition, we use Neyman-type conservative variance estimators to construct confidence intervals for the average treatment effect. Although Bloniarz et al. (2016) also established an asymptotic theory for a Lasso-adjusted estimator, they did not consider blocking. The technical difficulty lies in establishing novel concentration inequalities for the weighted sample mean and sample covariance under stratified randomization.

Second, we propose a Lasso-adjusted estimator to combine rerandomization and regression adjustment, thereby extending the results of Li and Ding (2020) to high-dimensional settings. We show that in the case of rerandomization, the proposed Lasso-adjusted estimator improves, or at least does not degrade, the precision compared with that obtained using the unadjusted estimator.

Third, we propose two Lasso-adjusted estimators in randomized experiments when both blocking and rerandomization are used in the design stage. We investigate the asymptotic properties of the estimators and outline conditions under which they are more efficient than the unadjusted estimator. Moreover, we propose conservative variance estimators to facilitate valid inferences.

The remaining paper is organized as follows: Section 2 introduces the framework and notation. Section 3 describes the two Lasso adjustment methods in randomized block or stratified randomized experiments and the examination of their asymptotic properties. Section 4 describes the method combining rerandomization and Lasso adjustment. Section 5 describes the two novel "blocking + rerandomization + Lasso adjustment" methods and the establishment of their asymptotic theory. Details of the simulation studies and two real data analyses are presented in Sections 6 and 7, respectively. Section 8 presents the concluding remarks. All the proofs are presented in the Supplementary Material.

2. Framework and notation

Consider a randomized experiment with n units and a binary treatment. For each unit i, $i = 1, \dots, n$, Z_i is an indicator of treatment assignment. Specifically, $Z_i = 1$ if unit i is assigned to the treatment group and $Z_i = 0$ otherwise. The treatment assignment is randomized. For example, in a completely randomized experiment, the probability distribution of the treatment assignment vector $\boldsymbol{Z} = (Z_1, \dots, Z_n)^{\mathrm{T}}$ is

$$P(\boldsymbol{Z} = \boldsymbol{z}) = \frac{n_1! n_0!}{n!}, \quad \sum_{i=1}^{n} I(z_i = 1) = n_1, \quad z_i = 0, 1,$$

where $n_1 = \sum_{i=1}^{n} Z_i$ is the number of treated units, $n_0 = \sum_{i=1}^{n}(1 - Z_i)$ is the number of control units, and $I(\cdot)$ is an indicator function. We define the treatment effects by using the Neyman–Rubin potential outcomes framework (Spława-Neyman et al., 1990; Rubin, 1974). For unit i, $Y_i(1)$ and $Y_i(0)$ are the potential outcomes under the treatment and the control, respectively. We define the unit-level treatment effect as $\tau_i = Y_i(1) - Y_i(0)$. As each unit is assigned to either the treatment group or the control group, but not both, we cannot simultaneously observe $Y_i(1)$ and $Y_i(0)$. Thus, τ_i cannot be identified without strong modelling assumptions pertaining to the potential outcomes. Under the stable unit treatment value assumption (SUTVA) (Rubin, 1980), the average treatment effect, defined as follows, is identifiable:

$$\tau = \frac{1}{n}\sum_{i=1}^{n}\tau_i = \frac{1}{n}\sum_{i=1}^{n}\{Y_i(1) - Y_i(0)\}.$$

The observed outcome is $Y_i = Z_i Y_i(1) + (1 - Z_i) Y_i(0)$. For each unit i, we observe a p^*-dimensional baseline/pre-treatment covariate vector $\boldsymbol{x}_i^* = (x_{i1}^*, \dots, x_{ip^*}^*)^{\mathrm{T}} \in \mathbb{R}^{p^*}$, where p^* is comparable to or even larger than n. According to prior experiments or domain knowledge, certain covariates are most relevant to the potential outcomes and may thus be preferentially balanced in the design stage. Without loss of generality, we define these covariates as the first k elements of $\boldsymbol{x}_i^*$, denoted by $\boldsymbol{w}_i^* = (x_{i1}^*, \dots, x_{ik^*}^*)$. Although the remaining covariates may also be relevant, the designer does not have prior information regarding their importance. These covariates are candidates for performing regression adjustment in the analysis stage. Throughout the paper, we assume that k^* is fixed, and p^* diverges with n. For simplicity, we do not index p^* with n. The objective is to make valid and efficient inferences on the average treatment effect τ by using the observed data $\{Y_i, Z_i, \boldsymbol{x}_i^*\}_{i=1}^n$.

Notation. For an L-dimensional column vector $\boldsymbol{u} = (u_1, \dots, u_L)^{\mathrm{T}}$, let $\|\boldsymbol{u}\|_0$, $\|\boldsymbol{u}\|_1$, $\|\boldsymbol{u}\|_2$, and $\|\boldsymbol{u}\|_\infty$ denote the ℓ_0, ℓ_1, ℓ_2 and ℓ_∞ norms, respectively. For a subset $S \subset \{1, \dots, L\}$, S^c is the complementary set of S and $\boldsymbol{u}_S = (u_j, j \in S)^{\mathrm{T}}$ is the restriction of $\boldsymbol{u}$ on S. Let $|S|$ be the cardinality of S. For a matrix Σ, $\Lambda_{\max}(\Sigma)$ indicates the largest eigenvalue of A. $\xrightarrow{d}$ and $\xrightarrow{p}$ denote convergence in distribution and in probability, respectively. $\dot{\sim}$ denotes the asymptotic equivalence. We use c, C, $\dots$, to denote universal constants that do not change with n but whose precise value may change from line to line.

3. Blocking and high-dimensional regression adjustment

3.1. *Randomized block experiments and unadjusted treatment effect estimator*

Blocking is a traditional approach to balance discrete covariates in experimental design. Experiments in which blocking is implemented are usually known as stratified randomized experiments or randomized block experiments. Blocking can increase the estimation efficiency of the average treatment effect when the blocking variables are relevant to the outcomes (Fisher, 1926; Imai, 2008; Imbens and Rubin, 2015). Moreover, Liu and Yang (2020) showed that a weighted regression adjustment can further improve the estimation efficiency when a fixed number of covariates is available. In this section, we describe two methods that combine blocking and high-dimensional regression adjustment to manage homogeneous and heterogeneous propensity scores across blocks. Specifically, we study the properties of randomized block experiments by using covariates $\boldsymbol{w}_i^*$ in the design stage and performing regression adjustment using the Lasso in the analysis stage to adjust for the remaining imbalances in covariates $\boldsymbol{x}_i^*$.

First, we introduce the randomized block experiments. Before the physical implementation of the experiment, we stratify the units into M blocks based on the covariates $\boldsymbol{w}_i^*$ by using a function $B : \mathrm{support}(\boldsymbol{w}_i^*) \to \{1, \dots, M\}$, where $\mathrm{support}(\boldsymbol{w}_i^*)$ is the support of

$\boldsymbol{w}_i^*$. For example, we stratify the units according to their gender, location, and discretized age. Let $B_i = B(\boldsymbol{w}_i^*)$ denote the block indicator for unit i. We remove the covariates that are linearly dependent on B_i from $\boldsymbol{x}_i^*$ and $\boldsymbol{w}_i^*$, and denote the remaining covariates as $\boldsymbol{x}_i$ and $\boldsymbol{w}_i$ with dimensions of p and k, respectively. Let $n_{[m]} = \sum_{i=1}^n I(B_i = m)$ denote the number of units in block m ($m = 1, \ldots, M$). Hereafter, subscript "$[m]$" indicates block-specific quantities. Within block m, $n_{[m]1}$ units are randomly assigned to the treatment group, and the remaining $n_{[m]0}$ units are assigned to the control group. The total number of treated units is $n_1 = \sum_{m=1}^M n_{[m]1}$. We assume that $2 \leqslant n_{[m]1} \leqslant n_{[m]} - 2$. The treatment assignments are independent across blocks, and thus, the probability distribution of $\boldsymbol{Z} = (Z_1, \ldots, Z_n)^{\mathrm{T}}$ in randomized block experiments is

$$P(\boldsymbol{Z} = \boldsymbol{z}) = \prod_{m=1}^M \frac{n_{[m]1}! n_{[m]0}!}{n_{[m]}!}, \quad \sum_{i \in [m]} I(z_i = 1) = n_{[m]1}, \quad z_i = 0, 1,$$

where $i \in [m]$ indexes the unit i in block m.

Certain additional notations are defined as follows: $\pi_{[m]} = n_{[m]}/n$ is the proportion of block size for block m, and $e_{[m]} = n_{[m]1}/n_{[m]}$ is the propensity score. For potential outcomes or transformed potential outcomes $R_i(z)$ ($z = 0, 1$), the block-specific average and sample mean are defined as

$$\bar{R}_{[m]}(z) = \frac{1}{n_{[m]}} \sum_{i \in [m]} R_i(z), \quad \bar{R}_{[m]z} = \frac{1}{n_{[m]z}} \sum_{i \in [m]} I(Z_i = z) R_i(z),$$

respectively. The overall average and overall (weighted) sample mean are denoted as

$$\bar{R}(z) = \frac{1}{n} \sum_{i=1}^n R_i(z) = \sum_{m=1}^M \pi_{[m]} \bar{R}_{[m]}(z), \quad \bar{R}_z = \sum_{m=1}^M \pi_{[m]} \bar{R}_{[m]z},$$

respectively. For finite population quantities

$$\boldsymbol{H} = (\boldsymbol{H}_1, \ldots, \boldsymbol{H}_n)^{\mathrm{T}}, \quad \boldsymbol{Q} = (\boldsymbol{Q}_1, \ldots, \boldsymbol{Q}_n)^{\mathrm{T}},$$

where $\boldsymbol{H}_i$ and $\boldsymbol{Q}_i$ are the column vectors. The block-specific covariance and overall covariance are denoted as

$$S_{[m]\boldsymbol{HQ}} = \frac{1}{n_{[m]} - 1} \sum_{i \in [m]} (\boldsymbol{H}_i - \bar{\boldsymbol{H}}_{[m]})(\boldsymbol{Q}_i - \bar{\boldsymbol{Q}}_{[m]})^{\mathrm{T}}, \quad \Sigma_{\boldsymbol{HQ}} = \sum_{m=1}^M \pi_{[m]} S_{[m]\boldsymbol{HQ}}.$$

When $\boldsymbol{H} = \boldsymbol{Q}$, we occasionally simplify the subscript as $S^2_{[m]\boldsymbol{H}} = S_{[m]\boldsymbol{HH}}$. The corresponding sample quantities are denoted as $s_{[m]\boldsymbol{HQ}}$, $\hat{\Sigma}_{\boldsymbol{HQ}}$, and $s^2_{[m]\boldsymbol{H}}$. These quantities depend on n; however, they are not indexed with n to ensure simplicity of notation.

Thus, the block-specific average treatment effect is

$$\tau_{[m]} = \frac{1}{n_{[m]}} \sum_{i \in [m]} \{Y_i(1) - Y_i(0)\} = \bar{Y}_{[m]}(1) - \bar{Y}_{[m]}(0),$$

with the overall average treatment effect being

$$\tau = \frac{1}{n}\sum_{i=1}^{n}\{Y_i(1) - Y_i(0)\} = \sum_{m=1}^{M}\pi_{[m]}\tau_{[m]}.$$

The difference-in-means of the outcomes within block m is an unbiased estimator of $\tau_{[m]}$:

$$\hat{\tau}_{[m],\mathrm{unadj}} = \frac{1}{n_{[m]1}}\sum_{i\in[m]} Z_i Y_i(1) - \frac{1}{n_{[m]0}}\sum_{i\in[m]}(1 - Z_i)Y_i(0) = \bar{Y}_{[m]1} - \bar{Y}_{[m]0}.$$

Thus, a plug-in estimator of τ is the following weighted difference-in-means:

$$\hat{\tau}_{\mathrm{unadj}} = \sum_{m=1}^{M}\pi_{[m]}\hat{\tau}_{[m],\mathrm{unadj}} = \sum_{m=1}^{M}\pi_{[m]}(\bar{Y}_{[m]1} - \bar{Y}_{[m]0}).$$

Under mild conditions, $\hat{\tau}_{\mathrm{unadj}}$ is unbiased and $\sqrt{n}(\hat{\tau}_{\mathrm{unadj}} - \tau)$ is asymptotically normal with a mean of zero and variance of

$$\sigma^2_{\mathrm{unadj}} = \lim_{n\to\infty}\sum_{m=1}^{M}\pi_{[m]}\left\{\frac{S^2_{[m]Y(1)}}{e_{[m]}} + \frac{S^2_{[m]Y(0)}}{1 - e_{[m]}} - S^2_{[m]\{Y(1)-Y(0)\}}\right\}.$$

The asymptotic variance can be conservatively estimated by

$$\hat{\sigma}^2_{\mathrm{unadj}} = \sum_{m=1}^{M}\pi_{[m]}\left\{\frac{s^2_{[m]Y(1)}}{e_{[m]}} + \frac{s^2_{[m]Y(0)}}{1 - e_{[m]}}\right\}.$$

More detailed discussions of this aspect can be found in the work of Liu and Yang (2020) and the references therein.

Notably, $\hat{\tau}_{\mathrm{unadj}}$ does not incorporate covariate information. In the following two sections, we introduce two Lasso-adjusted average treatment effect estimators and study their asymptotic properties. The first and second approaches are applicable for cases involving equal and unequal propensity scores across blocks, respectively.

3.2. *Lasso adjustment for equal propensity scores*

For equal propensity scores, $e_{[m]} = e^*$ for all $m = 1, \dots, M$, we define weights $\omega_i = Z_i\{n_{[m]}/(n_{[m]1} - 1)\}^{1/2} + (1 - Z_i)\{n_{[m]}/(n_{[m]0} - 1)\}^{1/2}$, for $i \in [m]$. Liu and Yang (2020) proposed the following weighted linear regression with low-dimensional covariates:

$$\begin{aligned}\omega_i Y_i \quad &\sim \quad \alpha\omega_i + \tau\omega_i Z_i + \sum_{m=2}^{M}\zeta_{[m]0}\omega_i I(B_i = m) + \sum_{m=2}^{M}\zeta_{[m]1}\omega_i Z_i\{I(B_i = m) - \pi_{[m]}\} \\ &\quad + \omega_i \boldsymbol{x}_i^{\mathrm{T}}\theta_0 + \omega_i Z_i\left\{\boldsymbol{x}_i - \sum_{m=1}^{M} I(B_i = m)\bar{\boldsymbol{x}}_{[m]}\right\}^{\mathrm{T}}\theta_1. \qquad (1)\end{aligned}$$

The regression-adjusted average treatment effect estimator $\hat{\tau}_{\mathrm{ols}}$ is defined as the ordinary least squares (OLS) estimator of τ (coefficient of $\omega_i Z_i$). This regression adjustment is equivalent to performing two weighted regressions in which we derive the weighted least squares-adjusted vectors by using the treated and control units, respectively,

$$\hat{\boldsymbol{\beta}}_{\text{ols},z} = \arg\min_{\boldsymbol{\beta}} \frac{1}{2} \sum_{m=1}^{M} \frac{\pi_{[m]}}{n_{[m]z} - 1} \sum_{i \in [m]} I(Z_i = z) \Big[Y_i(z) - \bar{Y}_{[m]z} - \big\{ \boldsymbol{x}_i - \bar{\boldsymbol{x}}_{[m]z} \big\}^{\mathrm{T}} \boldsymbol{\beta} \Big]^2,$$

where $z = 0, 1$. In this case,

$$\hat{\tau}_{\text{ols}} = \sum_{m=1}^{M} \pi_{[m]} \Big[\big\{ \bar{Y}_{[m]1} - (\bar{\boldsymbol{x}}_{[m]1} - \bar{\boldsymbol{x}}_{[m]})^{\mathrm{T}} \hat{\boldsymbol{\beta}}_{\text{ols},1} \big\} - \big\{ \bar{Y}_{[m]0} - (\bar{\boldsymbol{x}}_{[m]0} - \bar{\boldsymbol{x}}_{[m]})^{\mathrm{T}} \hat{\boldsymbol{\beta}}_{\text{ols},0} \big\} \Big],$$

where $(\bar{\boldsymbol{x}}_{[m]1} - \bar{\boldsymbol{x}}_{[m]})^{\mathrm{T}} \hat{\boldsymbol{\beta}}_{\text{ols},1}$ and $(\bar{\boldsymbol{x}}_{[m]0} - \bar{\boldsymbol{x}}_{[m]})^{\mathrm{T}} \hat{\boldsymbol{\beta}}_{\text{ols},0}$ adjust for the imbalances of the covariate means between the treatment and the control groups in block m.

REMARK 1. *When both the block sizes and propensity scores are the same across blocks, i.e., $n_{[m]} = n/M$ and $e_{[m]} = e^* \in (0, 1)$, the weighted least squares-adjusted vectors become*

$$\hat{\boldsymbol{\beta}}_{\text{ols},z} = \arg\min_{\boldsymbol{\beta}} \frac{1}{2n} \sum_{m=1}^{M} \sum_{i \in [m]} I(Z_i = z) \Big[Y_i(z) - \bar{Y}_{[m]z} - \big\{ \boldsymbol{x}_i - \bar{\boldsymbol{x}}_{[m]z} \big\}^{\mathrm{T}} \boldsymbol{\beta} \Big]^2.$$

The weighted regression (1) is equivalent to the following unweighted regression:

$$Y_i \sim \alpha + \tau Z_i + \sum_{m=2}^{M} \zeta_{[m]0} I(B_i = m) + \sum_{m=2}^{M} \zeta_{[m]1} Z_i \{ I(B_i = m) - \pi_{[m]} \} + \boldsymbol{x}_i^{\mathrm{T}} \theta_0 + Z_i \{ \boldsymbol{x}_i - \bar{\boldsymbol{x}} \}^{\mathrm{T}} \theta_1,$$

which is the regression with covariates $I(B_i = m)$ and $\boldsymbol{x}_i$, as suggested by Lin (2013). The resulting estimator $\hat{\tau}_{\text{ols}}$ ensures efficiency gains when two conditions are satisfied, namely, (1) the propensity scores are equal, and (2) the $n_{[m]}$ values are identical or $n_{[m]} \to \infty$ for all $m = 1, \ldots, M$. Otherwise, the efficiency may be degraded.

In a high-dimensional setting, when $p > n$, the OLS framework is ill-posed. Variable selection or regularization is necessary for an effective regression adjustment. In this study, we use the Lasso (Tibshirani, 1996) to simultaneously perform variable selection and regression adjustment. We add l_1 penalties on the regression coefficients θ_0 and θ_1 in the weighted regression (1), or equivalently, we derive the Lasso-adjusted vectors as

$$\hat{\boldsymbol{\beta}}_{\text{lasso},z} = \arg\min_{\boldsymbol{\beta}} \frac{1}{2} \sum_{m=1}^{M} \frac{\pi_{[m]}}{n_{[m]z} - 1} \sum_{i \in [m]} I(Z_i = z) \Big[Y_i(z) - \bar{Y}_{[m]z} - \big\{ \boldsymbol{x}_i - \bar{\boldsymbol{x}}_{[m]z} \big\}^{\mathrm{T}} \boldsymbol{\beta} \Big]^2 + \lambda_z ||\boldsymbol{\beta}||_1,$$

where $z = 0, 1$, λ_1 and λ_0 are tuning parameters. In practice, we can choose λ_z by cross-validation. We define the Lasso-adjusted average treatment effect estimator as

$$\hat{\tau}_{\text{lasso}} = \sum_{m=1}^{M} \pi_{[m]} \Big[\big\{ \bar{Y}_{[m]1} - (\bar{\boldsymbol{x}}_{[m]1} - \bar{\boldsymbol{x}}_{[m]})^{\mathrm{T}} \hat{\boldsymbol{\beta}}_{\text{lasso},1} \big\} - \big\{ \bar{Y}_{[m]0} - (\bar{\boldsymbol{x}}_{[m]0} - \bar{\boldsymbol{x}}_{[m]})^{\mathrm{T}} \hat{\boldsymbol{\beta}}_{\text{lasso},0} \big\} \Big].$$

REMARK 2. *When only one block ($M = 1$) is present, $\hat{\tau}_{\text{lasso}}$ is the same as the Lasso-adjusted estimator proposed by Bloniarz et al. (2016). Thus, $\hat{\tau}_{\text{lasso}}$ generalizes the estimator proposed by Bloniarz et al. from completely randomized experiments to randomized block experiments.*

To investigate the asymptotic properties of $\hat{\tau}_{\text{lasso}}$, we decompose the potential outcomes into a linear combination of the relevant covariates and an error term. Let $\boldsymbol{X}_j = (x_{1j}, \dots, x_{nj})^{\mathrm{T}}$ denote the observed values of the jth covariate, $j = 1, \dots, p$. Let S be the set of relevant covariates and $s = |S|$. Let $\boldsymbol{X}_S = (\boldsymbol{X}_j, j \in S)$ be the sub-matrix consisting of the covariates in S. We define the projection coefficient $\boldsymbol{\beta}(z)$ as

$$\{\boldsymbol{\beta}(z)\}_S = \Sigma^{-1}_{\boldsymbol{X}_S \boldsymbol{X}_S} \Sigma_{\boldsymbol{X}_S Y(z)}, \quad \{\boldsymbol{\beta}(z)\}_{S^c} = \mathbf{0}.$$

Next, we define the projection errors $\varepsilon_i(z)$ through the following equation:

$$Y_i(z) = \bar{Y}_{[m]}(z) + (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_{[m]})^{\mathrm{T}} \boldsymbol{\beta}(z) + \varepsilon_i(z), \quad i \in [m]. \tag{2}$$

We consider a finite population and randomization-based inference, i.e., all of the quantities in the above equation are fixed and deterministic real numbers, and the randomness is solely a result of the treatment assignment $\boldsymbol{Z}$. The following conditions must be satisfied to ensure the consistency and asymptotic normality of $\hat{\tau}_{\text{lasso}}$.

CONDITION 1. *There exists a constant $c \in (0, 0.5)$ independent of n, such that $c \leqslant e_{[m]} \leqslant 1 - c$ for $m = 1, \dots, M$.*

CONDITION 2. *There exists a constant $L < \infty$ independent of n, such that, for $z = 0, 1$ and $j = 1, \dots, p$, $n^{-1} \sum_{m=1}^{M} \sum_{i \in [m]} \{Y_i(z) - \bar{Y}_{[m]}(z)\}^2 \leqslant L^{1/2}$, $n^{-1} \sum_{m=1}^{M} \sum_{i \in [m]} (x_{ij} - \bar{x}_{[m],j})^4 \leqslant L$, and $n^{-1} \sum_{m=1}^{M} \sum_{i \in [m]} \{\varepsilon_i(z) - \bar{\varepsilon}_{[m]}(z)\}^4 \leqslant L$.*

CONDITION 3. *The weighted variances $\sum_{m=1}^{M} \pi_{[m]} S^2_{[m]\varepsilon(1)} / e_{[m]}$, $\sum_{m=1}^{M} \pi_{[m]} S^2_{[m]\varepsilon(0)} / (1 - e_{[m]})$, and $\sum_{m=1}^{M} \pi_{[m]} S^2_{[m]\{\varepsilon(1) - \varepsilon(0)\}}$ tend to finite limits, with positive values for the first two terms. The limit of $\sum_{m=1}^{M} \pi_{[m]} S^2_{[m]\varepsilon(1)} / e_{[m]} + \sum_{m=1}^{M} \pi_{[m]} S^2_{[m]\varepsilon(0)} / (1 - e_{[m]}) - \sum_{m=1}^{M} \pi_{[m]} S^2_{[m]\{\varepsilon(1) - \varepsilon(0)\}}$ is strictly positive.*

REMARK 3. *To satisfy Condition 1, the propensity scores must be bounded away from zero and one. In Condition 2, the bounded fourth moments are used to manage the high-dimensional covariates. This condition was also used by Bloniarz et al. (2016), Fogarty (2018a,b), Freedman (2008a), and Lin (2013) to study the asymptotic properties of various average treatment effect estimators. When p is fixed, we can weaken it to the following condition: for $z = 0, 1$ and $j = 1, \dots, p$,*

$$\frac{1}{n} \max_{m=1,\dots,M} \max_{i \in [m]} \{\varepsilon_i(z) - \bar{\varepsilon}_{[m]}(z)\}^2 \to 0, \quad \frac{1}{n} \max_{m=1,\dots,M} \max_{i \in [m]} (x_{ij} - \bar{x}_{[m],j})^2 \to 0.$$

Condition 3 ensures that the asymptotic variance of $\sqrt{n} \hat{\tau}_{\text{lasso}}$ has a finite and positive limit.

Because we consider high-dimensional settings, additional conditions must be set to ensure that the Lasso can consistently estimate the projection coefficient vector $\boldsymbol{\beta}(z)$.

CONDITION 4. *There exist constants $C > 0$ and $\xi > 1$ independent of n, such that* $||\boldsymbol{h}_S||_1 \leqslant Cs||\Sigma_{\boldsymbol{XX}}\boldsymbol{h}||_\infty$, $\forall \boldsymbol{h} \in \{\boldsymbol{h} : ||\boldsymbol{h}_{S^c}||_1 \leqslant \xi||\boldsymbol{h}_S||_1\}$.

CONDITION 5. *For constants $0 < \eta < (\xi - 1)/(\xi + 1)$, the tuning parameters of the Lasso satisfy*

$$s\sqrt{\log p}\lambda_z \to 0 \quad \text{and} \quad \lambda_z \geqslant \frac{1}{\eta}\Big\{\sqrt{\frac{120L}{c^3}}\sqrt{\frac{\log p}{n}} + \delta_n\Big\}, \quad z = 0, 1,$$

where $\delta_n = \max_{z=0,1} ||\sum_{m=1}^{M} \pi_{[m]} S_{[m]\boldsymbol{X}\varepsilon(z)}||_\infty$.

REMARK 4. *Bloniarz et al. (2016) proposed similar conditions to determine the l_1 convergence rate of the Lasso estimator in completely randomized experiments, without blocking. We can weaken the strict sparsity condition to a weak sparsity condition. Specifically, our theorems hold if we define* $s(z) = \sum_{j=1}^{p} \min\{|\beta_j(z)|/\lambda_z, 1\}$ *and* $s = \max\{s(1), s(0)\}$. *In this definition, we allow $\boldsymbol{\beta}(1)$ and $\boldsymbol{\beta}(0)$ to have many small but non-zero entries. Moreover, Condition 5 implies that* $s\log p/\sqrt{n} \to 0$.

CONDITION 6. *There exists a constant $e^* \in (0, 1)$ such that*

$$\max_{m=1,\ldots,M} |e_{[m]} - e^*| \to 0.$$

REMARK 5. *To satisfy Condition 6, the propensity scores must be asymptotically equal rather than exactly equal ($e_{[m]} = e^*$ for all $m = 1, \ldots, M$). Exactly equal propensity scores are unrealistic in certain cases due to practical restrictions.*

The main results are presented in the following text.

THEOREM 1. *If Conditions 1–5 hold, then*

$$\sqrt{n}(\hat{\tau}_{\text{lasso}} - \tau) \xrightarrow{d} N(0, \sigma^2_{\text{lasso}}),$$

where

$$\sigma^2_{\text{lasso}} = \lim_{n\to\infty} \sum_{m=1}^{M} \pi_{[m]} \Big\{ \frac{S^2_{[m]\varepsilon(1)}}{e_{[m]}} + \frac{S^2_{[m]\varepsilon(0)}}{1 - e_{[m]}} - S^2_{[m]\{\varepsilon(1)-\varepsilon(0)\}} \Big\}.$$

Furthermore, if Condition 6 holds, then

$$\sigma^2_{\text{lasso}} - \sigma^2_{\text{unadj}} = -\sum_{m=1}^{M} \pi_{[m]} \Delta^2_{[m]} \leqslant 0,$$

where $\Delta^2_{[m]} = \boldsymbol{\beta}^{\mathrm{T}}_{[m]} S^2_{[m]\boldsymbol{X}} \boldsymbol{\beta}_{[m]} / \{e_{[m]}(1 - e_{[m]})\}$ *with* $\boldsymbol{\beta}_{[m]} = (1 - e_{[m]})\boldsymbol{\beta}(1) + e_{[m]}\boldsymbol{\beta}(0)$.

REMARK 6. *The proof of Theorem 1 essentially relies on novel concentration inequalities for the weighted sample mean and weighted sample covariance in the case of stratified randomization. These entities are crucial for deriving the l_1-convergence rate of the Lasso estimator in a finite population and randomization-based inference framework. We obtain these inequalities in general asymptotic regimes, including the cases of (1) M tending to infinity with fixed $n_{[m]}$, and (2) $n_{[m]}$ tending to infinity with fixed M. These inequalities are of independent interest in other fields in which stratified sampling without replacement is performed. This aspect is extensively discussed in the Supplementary Material.*

Theorem 1 implies that the Lasso-adjusted estimator $\hat{\tau}_{\mathrm{lasso}}$ is consistent and asymptotically normal. Moreover, compared to the unadjusted estimator, $\hat{\tau}_{\mathrm{lasso}}$ enhances or at least does not degrade the precision when the propensity scores are asymptotically equal across blocks. In the case of unequal propensity scores, the efficiency may be degraded. A novel estimator to manage unequal propensity scores is described in the next section.

Next, we define a Neyman-type conservative estimator for the asymptotic variance. For $z = 0, 1$, let $\hat{s}(z) = ||\hat{\boldsymbol{\beta}}_{\mathrm{lasso},z}||_0$ be the number of selected covariates by the Lasso. We estimate the block-specific variance $S^2_{[m]\varepsilon(z)}$ by the corresponding block-specific sample variance

$$\hat{s}^2_{[m]\varepsilon(z)} = \frac{1}{n_{[m]z} - 1} \sum_{i \in [m]} I(Z_i = z) \Big\{ Y_i(z) - \bar{Y}_{[m]z} - (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_{[m]z})^{\mathrm{T}} \hat{\boldsymbol{\beta}}_{\mathrm{lasso},z} \Big\}^2 .$$

Subsequently, we estimate $\sigma^2_{\mathrm{lasso}}$ as

$$\hat{\sigma}^2_{\mathrm{lasso}} = \sum_{m=1}^{M} \pi_{[m]} \Big\{ \frac{n_1}{n_1 - \hat{s}(1) - 1} \frac{\hat{s}^2_{[m]\varepsilon(1)}}{e_{[m]}} + \frac{n_0}{n_0 - \hat{s}(0) - 1} \frac{\hat{s}^2_{[m]\varepsilon(0)}}{1 - e_{[m]}} \Big\},$$

where the factor $n_z / \{n_z - \hat{s}(z) - 1\}$ adjusts for the degrees of freedom.

CONDITION 7. *There exists a constant $C < \infty$ such that $\Lambda_{\max}(\Sigma_{\boldsymbol{XX}}) \leqslant C$.*

REMARK 7. *Condition 7 was also used by Bloniarz et al. (2016) to ensure that $\max\{\hat{s}(0), \hat{s}(1)\} = O_p(s)$, according to which $\hat{\sigma}^2_{\mathrm{lasso}}$ is asymptotically equivalent to the variance estimator without any adjustment for the degrees of freedom. However, our simulation experience shows that the unadjusted estimator may underestimate the variance in finite samples.*

THEOREM 2. *If Conditions 1–7 hold, $\hat{\sigma}^2_{\mathrm{lasso}}$ converges in probability to*

$$\lim_{n \to \infty} \sum_{m=1}^{M} \pi_{[m]} \Big\{ \frac{S^2_{[m]\varepsilon(1)}}{e_{[m]}} + \frac{S^2_{[m]\varepsilon(0)}}{1 - e_{[m]}} \Big\},$$

which is no less than $\sigma^2_{\mathrm{lasso}}$ and no greater than the probability limit of $\hat{\sigma}^2_{\mathrm{unadj}}$.

According to Theorem 2, $\hat{\sigma}^2_{\mathrm{lasso}}$ is a conservative variance estimator. If the unit-level treatment effect τ_i is constant within each block, i.e., $\tau_i = \tau_{[m]}$ for all $i \in [m]$, then,

$\boldsymbol{\beta}(1) = \boldsymbol{\beta}(0)$ and $S^2_{[m]\{\varepsilon(1)-\varepsilon(0)\}} = 0$. In this case, $\hat{\sigma}^2_{\text{lasso}}$ is a consistent estimator. Given $0 < \alpha < 1$, let $q_{\alpha/2}$ denote the upper $\alpha/2$ quantile of a standard normal distribution. According to Theorems 1 and 2, a $1-\alpha$ confidence interval for τ is

$$\left[\hat{\tau}_{\text{lasso}} - q_{\alpha/2}\hat{\sigma}_{\text{lasso}}/\sqrt{n}, \hat{\tau}_{\text{lasso}} + q_{\alpha/2}\hat{\sigma}_{\text{lasso}}/\sqrt{n}\right],$$

the asymptotic coverage rate of which is greater than or equal to $1-\alpha$. Moreover, the length of the confidence interval is less than or equal to that based on $(\hat{\tau}_{\text{unadj}}, \hat{\sigma}^2_{\text{unadj}})$. Thus, $\hat{\tau}_{\text{lasso}}$ increases, or, at least, does not degrade both the estimation and inference efficiencies for equal propensity scores.

3.3. *Lasso adjustment for unequal propensity scores*

When the propensity scores differ across blocks, $\hat{\tau}_{\text{lasso}}$ may deteriorate the efficiency compared with that achieved using $\hat{\tau}_{\text{unadj}}$. To address this problem, we propose another Lasso-adjusted estimator based on a projection perspective. The new estimator can ensure efficiency gains even when the propensity scores differ across blocks.

We consider the covariates as potential outcomes that are not affected by the treatment assignment, i.e., $\boldsymbol{x}_i(1) = \boldsymbol{x}_i(0) = \boldsymbol{x}_i$. The average treatment effect of the covariates is $\tau_{\boldsymbol{x}} = \sum_{m=1}^{M} \pi_{[m]}\{\bar{\boldsymbol{x}}_{[m]}(1) - \bar{\boldsymbol{x}}_{[m]}(0)\} = \mathbf{0}$. $\hat{\tau}_{\boldsymbol{x}} = \sum_{m=1}^{M} \pi_{[m]}(\bar{\boldsymbol{x}}_{[m]1} - \bar{\boldsymbol{x}}_{[m]0})$ denotes the weighted difference-in-means estimator of the covariates. To decrease the variance of $\hat{\tau}_{\text{unadj}}$, we project it onto $\hat{\tau}_{\boldsymbol{x}_S}$, which is the weighted difference-in-means estimator of the relevant covariates. We define the projection coefficient vector $\boldsymbol{\gamma}_{\text{proj}}$ as

$$\begin{aligned}(\boldsymbol{\gamma}_{\text{proj}})_S &= \arg\min_{\boldsymbol{\gamma}_S} E(\hat{\tau}_{\text{unadj}} - \tau - \hat{\tau}_{\boldsymbol{x}_S}^{\mathrm{T}}\boldsymbol{\gamma}_S)^2 = \operatorname{cov}(\hat{\tau}_{\boldsymbol{x}_S})^{-1}\operatorname{cov}(\hat{\tau}_{\boldsymbol{x}_S}, \hat{\tau}_{\text{unadj}}) \\ &= \left\{\sum_{m=1}^{M} \pi_{[m]} \frac{S^2_{[m]\boldsymbol{X}_S}}{e_{[m]}(1-e_{[m]})}\right\}^{-1} \left\{\sum_{m=1}^{M} \pi_{[m]} \frac{S_{[m]\boldsymbol{X}_S Y(1)}}{e_{[m]}} + \sum_{m=1}^{M} \pi_{[m]} \frac{S_{[m]\boldsymbol{X}_S Y(0)}}{1-e_{[m]}}\right\},\end{aligned}$$

with $(\boldsymbol{\gamma}_{\text{proj}})_{S^c} = \mathbf{0}$. The oracle projection estimator $\tilde{\tau}_{\text{proj}} = \hat{\tau}_{\text{unadj}} - \hat{\tau}_{\boldsymbol{x}_S}^{\mathrm{T}}(\boldsymbol{\gamma}_{\text{proj}})_S = \hat{\tau}_{\text{unadj}} - \hat{\tau}_{\boldsymbol{x}}^{\mathrm{T}}\boldsymbol{\gamma}_{\text{proj}}$ is consistent, asymptotically normal, and has the smallest asymptotic variance among the estimators of the form $\hat{\tau}_{\text{unadj}} - \hat{\tau}_{\boldsymbol{x}}^{\mathrm{T}}\boldsymbol{\gamma}$ for adjusted vector $\boldsymbol{\gamma}$ with $\boldsymbol{\gamma}_{S^c} = \mathbf{0}$.

REMARK 8. *In the case of equal propensity scores, i.e., $e_{[m]} = e^*$, we have*

(i) $\bar{\boldsymbol{x}}_{[m]1} - \bar{\boldsymbol{x}}_{[m]} = (1-e^*)\{\bar{\boldsymbol{x}}_{[m]1} - \bar{\boldsymbol{x}}_{[m]0}\}$, $\bar{\boldsymbol{x}}_{[m]0} - \bar{\boldsymbol{x}}_{[m]} = -e^*\{\bar{\boldsymbol{x}}_{[m]1} - \bar{\boldsymbol{x}}_{[m]0}\}$,

(ii) $\boldsymbol{\gamma}_{\text{proj}} = (1-e^*)\boldsymbol{\beta}(1) + e^*\boldsymbol{\beta}(0)$,

owing to which

$$\hat{\tau}_{\text{lasso}} \dot{\sim} \hat{\tau}_{\text{unadj}} - \hat{\tau}_{\boldsymbol{x}}^{\mathrm{T}}\{(1-e^*)\boldsymbol{\beta}(1) + e^*\boldsymbol{\beta}(0)\} = \tilde{\tau}_{\text{proj}}.$$

Thus, $\hat{\tau}_{\text{lasso}}$ has the same asymptotic distribution as $\tilde{\tau}_{\text{proj}}$. Therefore, when the propensity scores are equal, $\hat{\tau}_{\text{lasso}}$ is the best estimator among the class of estimators $\{\hat{\tau}_{\text{unadj}} - \hat{\tau}_{\boldsymbol{x}}^{\mathrm{T}}\boldsymbol{\gamma} : \boldsymbol{\gamma} \in \mathbb{R}^p,\ \boldsymbol{\gamma}_{S^c} = 0\}$.

However, $\tilde{\tau}_{\text{proj}}$ is not feasible in practice because it depends on the unknown vector $\boldsymbol{\gamma}_{\text{proj}}$. To consistently estimate $\boldsymbol{\gamma}_{\text{proj}}$, we decompose it into two terms. We denote $e_{[m]z} = ze_{[m]} + (1-z)(1-e_{[m]})$ to simplify the notation. In this case, $\boldsymbol{\gamma}_{\text{proj}} = \boldsymbol{\gamma}(0) + \boldsymbol{\gamma}(1)$, where

$$\{\boldsymbol{\gamma}(z)\}_S = \left\{ \sum_{m=1}^{M} \pi_{[m]} \frac{S^2_{[m]\boldsymbol{X}_S}}{e_{[m]}(1-e_{[m]})} \right\}^{-1} \left\{ \sum_{m=1}^{M} \pi_{[m]} \frac{S_{[m]\boldsymbol{X}_S Y(z)}}{e_{[m]z}} \right\}, \quad \{\boldsymbol{\gamma}(z)\}_{S^c} = \mathbf{0}.$$

Intuitively, if the set S is known, we can estimate $\{\boldsymbol{\gamma}(z)\}_S$ by replacing the block-specific covariances with the corresponding sample covariances:

$$(\hat{\boldsymbol{\gamma}}_{\text{ols},z})_S = \left\{ \sum_{m=1}^{M} \pi_{[m]} \frac{s^2_{[m]\boldsymbol{X}_S(z)}}{e_{[m]z}(1-e_{[m]z})} \right\}^{-1} \left\{ \sum_{m=1}^{M} \pi_{[m]} \frac{s_{[m]\boldsymbol{X}_S Y(z)}}{e_{[m]z}} \right\},$$

where $s^2_{[m]\boldsymbol{X}_S(z)}$ stands for the sample covariance of $\boldsymbol{X}_S$ under treatment arm z. To formulate $(\hat{\boldsymbol{\gamma}}_{\text{ols},z})_S$ as a weighted least squares problem, we introduce the following weights:

$$\omega_i = \frac{n_{[m]}}{n_{[m]z} - 1}, \quad \omega_{Y,i} = \frac{1 - e_{[m]z}}{e_{[m]z}}, \quad \omega_{\boldsymbol{X},i} = \frac{1}{e_{[m]z}(1-e_{[m]z})}, \quad i \in [m], \quad Z_i = z.$$

In this case,

$$(\hat{\boldsymbol{\gamma}}_{\text{ols},z})_S = \arg\min_{\boldsymbol{\gamma}_S} \sum_{m=1}^{M} \sum_{i\in[m]} \omega_i I(Z_i = z) \Big[\sqrt{\omega_{Y,i}} \big\{ Y_i(z) - \bar{Y}_{[m]z} \big\} - \sqrt{\omega_{\boldsymbol{X},i}} \big\{ (\boldsymbol{x}_i)_S - (\bar{\boldsymbol{x}}_{[m]z})_S \big\}^{\mathrm{T}} \boldsymbol{\gamma} \Big]^2.$$

In practice, S is usually unknown. Fortunately, we can directly estimate $\boldsymbol{\gamma}_{\text{proj}}$ by using the Lasso,

$$\hat{\boldsymbol{\gamma}}_{\text{lasso},z} = \arg\min_{\boldsymbol{\gamma}} \sum_{m=1}^{M} \sum_{i\in[m]} \omega_i I(Z_i = z) \Big\{ \sqrt{\omega_{Y,i}} \big\{ Y_i(z) - \bar{Y}_{[m]z} \big\} - \sqrt{\omega_{\boldsymbol{X},i}} \big\{ \boldsymbol{x}_i - \bar{\boldsymbol{x}}_{[m]z} \big\}^{\mathrm{T}} \boldsymbol{\gamma} \Big\}^2 + \lambda_z \|\boldsymbol{\gamma}\|_1.$$

We replace $\boldsymbol{\gamma}_{\text{proj}}$ with its estimator $\hat{\boldsymbol{\gamma}}_{\text{lasso}} = \hat{\boldsymbol{\gamma}}_{\text{lasso},1} + \hat{\boldsymbol{\gamma}}_{\text{lasso},0}$ and obtain the projection-originated Lasso-adjusted estimator as

$$\hat{\tau}_{\text{lasso2}} = \hat{\tau}_{\text{unadj}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{x}}^{\mathrm{T}} \hat{\boldsymbol{\gamma}}_{\text{lasso}}.$$

To investigate the asymptotic properties of $\hat{\tau}_{\text{lasso2}}$, we decompose the original potential outcomes in the following manner:

$$Y_i(z) = \bar{Y}_{[m]}(z) + (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_{[m]})^{\mathrm{T}} \boldsymbol{\gamma}_{\text{proj}} + \varepsilon_i^*(z), \quad i \in [m], \quad z = 0, 1.$$

Furthermore, to examine the convergence of $\hat{\boldsymbol{\gamma}}_{\text{lasso},z}$ to $\boldsymbol{\gamma}(z)$, we decompose the transformed potential outcomes into a linear combination of transformed covariates and an error term:

$$\sqrt{\omega_{Y,i}}\, Y_i(z) = \sqrt{\omega_{Y,i}}\, \bar{Y}_{[m]}(z) + \sqrt{\omega_{\boldsymbol{X},i}} (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_{[m]})^{\mathrm{T}} \boldsymbol{\gamma}(z) + \varepsilon_i^{\Delta}(z), \quad i \in [m], \quad z = 0, 1.$$

THEOREM 3. *If Conditions 1–2 hold with $\varepsilon_i(z)$ being replaced by both $\varepsilon_i^{\Delta}(z)$ and $\varepsilon_i^*(z)$ and $\boldsymbol{x}_i$ being replaced by $\sqrt{\omega_{\boldsymbol{X},i}}\boldsymbol{x}_i$, and Condition 3 holds with $\varepsilon_i(z)$ being replaced by $\varepsilon^*(z)$, and Conditions 4–5 hold with $\varepsilon_i(z)$ and $\boldsymbol{x}_i$ being replaced by $\varepsilon_i^{\Delta}(z)$ and $\sqrt{\omega_{\boldsymbol{X},i}}\boldsymbol{x}_i$, then*

$$\sqrt{n}(\hat{\tau}_{\text{lasso2}} - \tau) \xrightarrow{d} N(0, \sigma^2_{\text{lasso2}}),$$

where

$$\sigma^2_{\text{lasso2}} = \lim_{n\to\infty} \sum_{m=1}^{M} \pi_{[m]} \left\{ \frac{S^2_{[m]\varepsilon^*(1)}}{e_{[m]}} + \frac{S^2_{[m]\varepsilon^*(0)}}{1-e_{[m]}} - S^2_{[m]\{\varepsilon^*(1)-\varepsilon^*(0)\}} \right\}.$$

Furthermore,

$$\sigma^2_{\text{lasso2}} - \sigma^2_{\text{unadj}} = -\lim_{n\to\infty} \boldsymbol{\gamma}_{\text{proj}}^{\mathrm{T}} \left\{ \sum_{m=1}^{M} \pi_{[m]} \frac{S^2_{[m]\boldsymbol{X}}}{e_{[m]}(1-e_{[m]})} \right\} \boldsymbol{\gamma}_{\text{proj}} \leqslant 0.$$

REMARK 9. *Theorem 3 shows that $\hat{\tau}_{\text{lasso2}}$ is not only feasible but also has the same asymptotic distribution as $\tilde{\tau}_{\text{proj}}$ even for unequal propensity scores. In other words, $\hat{\tau}_{\text{lasso2}}$ has the smallest asymptotic variance among the estimators in $\{\hat{\tau}_{\text{unadj}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{x}}^{\mathrm{T}}\boldsymbol{\gamma} : \boldsymbol{\gamma} \in \mathbb{R}^p,\ \boldsymbol{\gamma}_{S^c} = 0\}$.*

Similar to the estimation of σ^2_{lasso}, we can conservatively estimate σ^2_{lasso2} by replacing the finite population quantities with the corresponding sample quantities. Precisely, let $\hat{s} = ||\hat{\boldsymbol{\gamma}}_{\text{lasso}}||_0$. The block-specific variance $S^2_{[m]\varepsilon^*(z)}$ can be estimated by the corresponding block-specific sample variance:

$$\hat{s}^2_{[m]\varepsilon^*(z)} = \frac{1}{n_{[m]z} - 1} \sum_{i\in[m]} I(Z_i = z) \left\{ Y_i(z) - \bar{Y}_{[m]z} - (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_{[m]z})^{\mathrm{T}} \hat{\boldsymbol{\gamma}}_{\text{lasso}} \right\}^2.$$

Then, σ^2_{lasso2} can be estimated as

$$\hat{\sigma}^2_{\text{lasso2}} = \frac{n}{n - \hat{s} - 1} \sum_{m=1}^{M} \pi_{[m]} \left\{ \frac{\hat{s}^2_{[m]\varepsilon^*(1)}}{e_{[m]}} + \frac{\hat{s}^2_{[m]\varepsilon^*(0)}}{1-e_{[m]}} \right\},$$

where the factor $n/\{n - \hat{s} - 1\}$ adjusts for the degrees of freedom.

THEOREM 4. *Suppose that the conditions associated with Theorem 3 and Condition 7 hold, then $\hat{\sigma}^2_{\text{lasso2}}$ converges in probability to*

$$\lim_{n\to\infty} \sum_{m=1}^{M} \pi_{[m]} \left\{ \frac{S^2_{[m]\varepsilon^*(1)}}{e_{[m]}} + \frac{S^2_{[m]\varepsilon^*(0)}}{1-e_{[m]}} \right\},$$

which is no less than σ^2_{lasso2} and no greater than the probability limit of $\hat{\sigma}^2_{\text{unadj}}$.

For the additive treatment effect within each block, i.e., $\tau_i = \tau_{[m]}$ for all $i \in [m]$, we have $S^2_{[m]\{\varepsilon^*(1)-\varepsilon^*(0)\}} = 0$. In this case, $\hat{\sigma}^2_{\text{lasso2}}$ is a consistent estimator of σ^2_{lasso2}.

Generally, $\hat{\sigma}^2_{\mathrm{lasso2}}$ is a conservative variance estimator. According to Theorems 3 and 4, we can construct a $1-\alpha$ confidence interval for τ:

$$\left[\hat{\tau}_{\mathrm{lasso2}} - q_{\alpha/2}\hat{\sigma}_{\mathrm{lasso2}}/\sqrt{n}, \hat{\tau}_{\mathrm{lasso2}} + q_{\alpha/2}\hat{\sigma}_{\mathrm{lasso2}}/\sqrt{n}\right].$$

The asymptotic coverage rate of the abovementioned confidence interval is greater than or equal to $1-\alpha$. Moreover, the length of this confidence interval is less than or equal to that based on $(\hat{\tau}_{\mathrm{unadj}}, \hat{\sigma}^2_{\mathrm{unadj}})$. Therefore, $\hat{\tau}_{\mathrm{lasso2}}$ improves, or, at least, does not degrade the estimation and inference efficiencies, regardless of equal or unequal propensity scores.

4. Rerandomization and high-dimensional regression adjustment

4.1. *Rerandomization*

Although blocking is widely used in practice, it can balance only discrete covariates. Rerandomization is a more general approach to balance both discrete and continuous covariates (Morgan and Rubin, 2012, 2015; Li et al., 2020). Rerandomization discards the treatment assignments that lead to covariate imbalances and accepts only those assignments that fulfil a prespecified balance criterion. Morgan and Rubin (2012) proposed the use of the Mahalanobis distance of the covariate means in the treatment and control groups to measure the covariate imbalances. In completely randomized experiments, we define $\hat{\tau}_{\boldsymbol{w}} = \bar{\boldsymbol{w}}_1 - \bar{\boldsymbol{w}}_0$ and the Mahalanobis distance as

$$\mathrm{Ma}(\boldsymbol{Z}, \boldsymbol{W}) = (\hat{\tau}_{\boldsymbol{w}})^{\mathrm{T}}\{\mathrm{cov}(\hat{\tau}_{\boldsymbol{w}})\}^{-1}\hat{\tau}_{\boldsymbol{w}}.$$

A treatment assignment is acceptable if and only if the corresponding Mahalanobis distance is less than or equal to a prespecified threshold $a > 0$, i.e., $\mathrm{Ma}(\boldsymbol{Z}, \boldsymbol{W}) \leqslant a$. We denote $\mathcal{M}_a = \{\boldsymbol{Z} : \mathrm{Ma}(\boldsymbol{Z}, \boldsymbol{W}) \leqslant a\}$ as the set of acceptable treatment assignments. Li et al. (2018) suggested choosing a suitable a to ensure that the probability of a treatment assignment satisfying the balance criterion equals a certain value, for example, $p_a = P\{\mathrm{Ma}(\boldsymbol{Z}, \boldsymbol{W}) \leqslant a\} = 0.001$. A general rerandomization procedure is as indicated in Algorithm 1. Under mild conditions, the average treatment effect estimator $\hat{\tau}_{\mathrm{unadj}}$ (with $M = 1$) subjected to rerandomization is consistent and asymptotically truncated normally distributed, with the asymptotic variance being no greater than that of $\hat{\tau}_{\mathrm{unadj}}$ under complete randomization, as reported by Morgan and Rubin (2012) and Li et al. (2018) and the references therein.

Algorithm 1 (Stratified) Rerandomization using the Mahalanobis distance

1. Collect covariates data $\boldsymbol{w}_i$, $i = 1, \ldots, n$;
2. (Re-)Randomize units to treatment and control groups by complete randomization (or stratified randomization) and obtain the treatment assignment vector $\boldsymbol{Z}$;
3. If $\mathrm{Ma}(\boldsymbol{Z}, \boldsymbol{W}) \leqslant a$, proceed to Step 4; otherwise, return to Step 2;
4. Conduct the physical experiment by using the treatment assignment $\boldsymbol{Z}$.

4.2. *Rerandomization plus Lasso*

Rerandomization is useful for balancing a fixed number of covariates $\boldsymbol{w}_i$ that are most relevant to the potential outcomes. However, the additional high-dimensional covariates

$\boldsymbol{x}_i$ may also be related to the potential outcomes and remain imbalanced in the case of rerandomization.

This phenomenon has motivated researchers to consider regression adjustment under rerandomization to adjust for the remaining covariate imbalances and further increase the efficiency. Li and Ding (2020) showed that rerandomization followed by regression adjustment using the OLS could increase the estimation and inference efficiencies when the analyzer uses both the covariates adopted in the design stage and additional covariates available only in the analysis stage. Nevertheless, this conclusion is true only for low-dimensional covariates. To extend this principle to high-dimensional settings, the proposed Lasso-adjusted estimators $\hat{\tau}_{\text{lasso}}$ and $\hat{\tau}_{\text{lasso2}}$ can be used. In the next section, we examine their asymptotic properties of these estimators in the case of rerandomization, as a special case of stratified rerandomization with $M = 1$.

5. Blocking, rerandomization, and high-dimensional regression adjustment

5.1. Stratified rerandomization (blocking plus rerandomization)

Both blocking and rerandomization are powerful methods to balance a fixed number of covariates. Scholars, such as Fisher and Rubin, have recommended combining these two methods in the design stage. Recently, Schultzberg and Johansson (2019) proposed a stratified rerandomization strategy in which stratified randomization is implemented instead of complete randomization in step 2 of Algorithm 1. However, this method is only applicable for the case of involving equal propensity scores, mainly because $\hat{\tau}_{\boldsymbol{w}}$ may be asymptotically biased for $\tau_{\boldsymbol{w}}$ when the propensity scores differ across blocks. To address this issue, Wang et al. (2020) modified the definition of $\hat{\tau}_{\boldsymbol{w}}$: $\hat{\tau}_{\boldsymbol{w}} = \sum_{m=1}^{M} \pi_{[m]}(\bar{\boldsymbol{w}}_{[m]1} - \bar{\boldsymbol{w}}_{[m]0})$, and showed that the asymptotic distribution of $\hat{\tau}_{\text{unadj}}$ in the modified stratified rerandomization strategy is a truncated normal distribution and its asymptotic variance, denoted as $\sigma^2_{\text{unadj}|\mathcal{M}}$, is less than or equal to that of $\hat{\tau}_{\text{unadj}}$ in the case of stratified randomization. Moreover, the asymptotic variance can be estimated using a conservative estimator $\hat{\sigma}^2_{\text{unadj}|\mathcal{M}}$, as indicated in the Supplementary Material. These conclusions hold in cases involving equal or unequal propensity scores.

The efficiency of $\hat{\tau}_{\text{unadj}}$ in the case of Wang et al.'s stratified rerandomization (referred to as stratified rerandomization in the following text) can be further increased by adjusting for the remaining imbalances in $\boldsymbol{x}_i$. In the following two sections, we examine the asymptotic properties of the Lasso-adjusted estimators, introduced in Section 3, in the case of stratified rerandomization.

5.2. Stratified rerandomization plus Lasso for equal propensity scores

To investigate the asymptotic properties of $\hat{\tau}_{\text{lasso}}$ in the case of stratified rerandomization, the decomposition specified in equation (2) must be performed. We assume that $\boldsymbol{W}$ is a subset of $\boldsymbol{X}_S$, i.e., $\{1, ..., k\} \subset S$. Moreover, we regard $\varepsilon_i(z)$ as the transformed potential outcomes and let $\hat{\tau}_{\varepsilon}$ be the weighted difference-in-means estimator for $\tau_{\varepsilon} = (1/n)\sum_{i=1}^{n}\{\varepsilon_i(1) - \varepsilon_i(0)\}$. Applying Proposition 2 reported by Wang et al. (2020)

to $\varepsilon_i(z)$, we obtain the covariance of $\sqrt{n}(\hat{\tau}_\varepsilon, \hat{\tau}_{\boldsymbol{w}}^{\mathrm{T}})^{\mathrm{T}}$ in the case of stratified randomization:

$$V\{\varepsilon, \boldsymbol{W}\} = \begin{pmatrix} V_{\varepsilon\varepsilon} & V_{\varepsilon \boldsymbol{W}} \\ V_{\boldsymbol{W}\varepsilon} & V_{\boldsymbol{W}\boldsymbol{W}} \end{pmatrix},$$

where

$$V_{\varepsilon\varepsilon} = \sigma^2_{\mathrm{lasso}}, \quad V_{\boldsymbol{W}\boldsymbol{W}} = \sum_{m=1}^{M} \pi_{[m]} \left\{ \frac{S^2_{[m]\boldsymbol{W}}}{e_{[m]}(1-e_{[m]})} \right\},$$

$$V_{\boldsymbol{W}\varepsilon} = V_{\varepsilon\boldsymbol{W}}^{\mathrm{T}} = \sum_{m=1}^{M} \pi_{[m]} \left\{ \frac{S_{[m]\boldsymbol{W}\varepsilon(1)}}{e_{[m]}} + \frac{S_{[m]\boldsymbol{W}\varepsilon(0)}}{1-e_{[m]}} \right\}.$$

The asymptotic distribution of $\hat{\tau}_{\mathrm{lasso}}$ in the case of stratified rerandomization depends on the squared multiple correlation between $\hat{\tau}_\varepsilon$ and $\hat{\tau}_{\boldsymbol{w}}$,

$$R^2_{\varepsilon,\boldsymbol{W}} = \lim_{n\to\infty} (V_{\varepsilon\boldsymbol{W}} V^{-1}_{\boldsymbol{W}\boldsymbol{W}} V_{\boldsymbol{W}\varepsilon})/V_{\varepsilon\varepsilon}.$$

Let $\varepsilon_0, D_1, \ldots, D_k$ be independent standard normal distributed random variables and let $L_{k,a} = \{D_1 \mid \sum_{i=1}^{k} D_i^2 \leqslant a\}$.

Condition 8. *The weighted covariances* $\sum_{m=1}^{M} \pi_{[m]} S^2_{[m]\boldsymbol{W}}/e_{[m]}$, $\sum_{m=1}^{M} \pi_{[m]} S^2_{[m]\boldsymbol{W}}/(1-e_{[m]})$, $\sum_{m=1}^{M} \pi_{[m]} S_{[m]\boldsymbol{W}\varepsilon(1)}/e_{[m]}$, *and* $\sum_{m=1}^{M} \pi_{[m]} S_{[m]\boldsymbol{W}\varepsilon(0)}/(1-e_{[m]})$ *tend to finite limits, and the limit of* $V\{\varepsilon, \boldsymbol{W}\}$ *is strictly positive definite.*

Theorem 5. *If Conditions 1–5, and 8 hold, then*

$$\{\sqrt{n}(\hat{\tau}_{\mathrm{lasso}} - \tau) \mid \mathcal{M}_a\} \xrightarrow{d} \sigma_{\mathrm{lasso}} \left(\sqrt{1-R^2_{\varepsilon,\boldsymbol{W}}}\,\varepsilon_0 + \sqrt{R^2_{\varepsilon,\boldsymbol{W}}}\,L_{k,a} \right).$$

Furthermore, if Condition 6 holds, then $R^2_{\varepsilon,\boldsymbol{W}} = 0$,

$$\{\sqrt{n}(\hat{\tau}_{\mathrm{lasso}} - \tau) \mid \mathcal{M}_a\} \xrightarrow{d} N(0, \sigma^2_{\mathrm{lasso}}),$$

and $\sigma^2_{\mathrm{lasso}} \leqslant \sigma^2_{\mathrm{unadj}|\mathcal{M}} \leqslant \sigma^2_{\mathrm{unadj}}$.

Theorem 5 implies that the asymptotic distribution of $\hat{\tau}_{\mathrm{lasso}}$ in the case of stratified rerandomization is a truncated normal distribution. For (asymptotically) equal propensity scores (Condition 6), the asymptotic distribution of $\hat{\tau}_{\mathrm{lasso}}$ in the case of stratified rerandomization is normal, with the asymptotic variance $\sigma^2_{\mathrm{lasso}}$ being no greater than those of $\hat{\tau}_{\mathrm{unadj}}$ in stratified randomization and stratified rerandomization. Thus, stratified rerandomization followed by the Lasso adjustment does not deteriorate the precision for equal propensity scores. In particular, when $M = 1$, Condition 6 holds trivially, and Theorem 5 extends the results of Li and Ding (2020) (Theorem 3 and Proposition 3) from a low-dimensional setting to a high-dimensional setting. In contrast, for (asymptotically) unequal propensity scores, $\hat{\tau}_{\mathrm{lasso}}$ may lead to a lower efficiency than that attained using $\hat{\tau}_{\mathrm{unadj}}$ in the case of stratified rerandomization. In the next section, we describe the second Lasso-adjusted estimator $\hat{\tau}_{\mathrm{lasso2}}$ to manage unequal propensity scores.

THEOREM 6. *If Conditions 1–8 hold, then in the case of stratified rerandomization, $\hat{\sigma}^2_{\text{lasso}}$ converges in probability to*

$$\lim_{n\to\infty} \sum_{m=1}^{M} \pi_{[m]} \left\{ \frac{S^2_{[m]\varepsilon(1)}}{e_{[m]}} + \frac{S^2_{[m]\varepsilon(0)}}{1-e_{[m]}} \right\} \geqslant \sigma^2_{\text{lasso}}.$$

Moreover, the following inequality holds in probability in the case of stratified rerandomization: $\hat{\sigma}^2_{\text{lasso}} \leqslant \hat{\sigma}^2_{\text{unadj}|\mathcal{M}} \leqslant \hat{\sigma}^2_{\text{unadj}}$.

When the individual treatment effect τ_i is constant within each block, i.e., $\tau_i = \tau_{[m]}$ for all $i \in [m]$, $\hat{\sigma}^2_{\text{lasso}}$ is a consistent estimator of σ^2_{lasso}. In the case involving equal propensity scores, we can construct an asymptotically conservative $1-\alpha$ confidence interval for τ, according to Theorems 5 and 6:

$$\left[\hat{\tau}_{\text{lasso}} - q_{\alpha/2}\hat{\sigma}_{\text{lasso}}/\sqrt{n}, \hat{\tau}_{\text{lasso}} + q_{\alpha/2}\hat{\sigma}_{\text{lasso}}/\sqrt{n}\right].$$

The length of this confidence interval is less than or equal to those based on $(\hat{\tau}_{\text{unadj}}, \hat{\sigma}^2_{\text{unadj}|\mathcal{M}})$ and $(\hat{\tau}_{\text{unadj}}, \hat{\sigma}^2_{\text{unadj}})$. Therefore, $\hat{\tau}_{\text{lasso}}$ does not degrade the estimation and inference efficiencies in the case of equal propensity scores, relative to that attained using $\hat{\tau}_{\text{unadj}}$ in both stratified randomization and stratified rerandomization scenarios.

5.3. Stratified rerandomization plus Lasso for unequal propensity scores

In the stratified rerandomization scenario, $\hat{\tau}_{\text{lasso}}$ may deteriorate the efficiency in cases involving unequal propensity scores. To solve this problem, we consider the projection estimator $\hat{\tau}_{\text{lasso2}}$ introduced in Section 3.3.

THEOREM 7. *If the conditions associated with Theorem 3 and Condition 8 hold, then*

$$\{\sqrt{n}(\hat{\tau}_{\text{lasso2}} - \tau) \mid \mathcal{M}_a\} \xrightarrow{d} N(0, \sigma^2_{\text{lasso2}}).$$

Furthermore, $\sigma^2_{\text{lasso2}} \leqslant \sigma^2_{\text{unadj}|\mathcal{M}} \leqslant \sigma^2_{\text{unadj}}$.

REMARK 10. *When the propensity scores are equal across blocks, Theorem 5 indicates that $\hat{\tau}_{\text{lasso}}$ has the same asymptotic distribution in the stratified randomization and stratified rerandomization scenarios. When the propensity scores differ across blocks, Theorem 7 implies that $\hat{\tau}_{\text{lasso2}}$ has the same asymptotic distribution in the stratified randomization and stratified rerandomization scenarios. Thus, the discussion in the Remarks 8 and 9, regarding the optimality of these two estimators, remains valid.*

Theorem 7 shows that the asymptotic distribution of $\hat{\tau}_{\text{lasso2}}$ in the stratified rerandomization case is normal, which can be viewed as a truncated normal with the threshold $a \to 0$. Moreover, the asymptotic variance of $\hat{\tau}_{\text{lasso2}}$ in the case of stratified rerandomization is no greater than those of $\hat{\tau}_{\text{unadj}}$ in the stratified randomization and stratified rerandomization scenarios, even for the case involving unequal propensity scores. Thus, the efficiency attained using $\hat{\tau}_{\text{lasso2}}$ is never lower than that achieved using $\hat{\tau}_{\text{unadj}}$.

Compared with that in stratified randomization, the efficiency in the stratified rerandomization scenario does not increase when $\hat{\tau}_{\text{lasso2}}$ is used in the analysis stage. Similar conclusions were derived by Li and Ding (2020), who examined the combination of rerandomization and low-dimensional regression adjustment. However, as discussed by Li and Ding (2020), stratified rerandomization is preferred because it can enable a more transparent analysis and avoid the bias associated with searching for a specific outcome model in the analysis stage (Cox, 2007; Rosenbaum, 2010; Lin, 2013).

THEOREM 8. *If the conditions associated with Theorem 3 and Conditions 7–8 hold, then in the stratified rerandomization scenario, $\hat{\sigma}^2_{\text{lasso2}}$ converges in probability to*

$$\lim_{n\to\infty}\sum_{m=1}^{M}\pi_{[m]}\left\{\frac{S^2_{[m]\varepsilon^*(1)}}{e_{[m]}}+\frac{S^2_{[m]\varepsilon^*(0)}}{1-e_{[m]}}\right\}\geqslant\sigma^2_{\text{lasso2}}.$$

Moreover, $\hat{\sigma}^2_{\text{lasso2}}\leqslant\hat{\sigma}^2_{\text{unadj}|\mathcal{M}}\leqslant\hat{\sigma}^2_{\text{unadj}}$ holds with the probability approaching to one in the stratified rerandomization scenario.

In the stratified rerandomization case, the variance estimator $\hat{\sigma}^2_{\text{lasso2}}$ is consistent if the unit-level treatment effect is constant within each block. Generally, the estimator is conservative. According to Theorems 7 and 8, we can construct a $1-\alpha$ confidence interval for τ:

$$\left[\hat{\tau}_{\text{lasso2}}-q_{\alpha/2}\hat{\sigma}_{\text{lasso2}}/\sqrt{n},\hat{\tau}_{\text{lasso2}}+q_{\alpha/2}\hat{\sigma}_{\text{lasso2}}/\sqrt{n}\right].$$

The asymptotic coverage rate of this confidence interval is greater than or equal to $1-\alpha$, and its length is less than or equal to those based on the estimated asymptotic distributions of $\hat{\tau}_{\text{unadj}}$ in the stratified randomization and stratified rerandomization cases. Therefore, $\hat{\tau}_{\text{lasso2}}$ is the most efficient estimator among all the considered estimators for unequal propensity scores.

6. Simulation

This section describes the simulation studies performed to examine the finite sample performance of the proposed methods. We set the sample size as $n=200$ and 500. We consider three scenarios of blocking: many small blocks (MS, with $n_{[m]}=10$ and $M=20$ or 50), a few large blocks (FL, with $n_{[m]}=100$ or 250 and $M=2$), and hybrid blocks (MS+FL, with $n^S_{[m]}=10$, $M^S=10$ or 20, $n^L_{[m]}=50$ or 150, and $M^L=2$, where the subscripts "S" and "L" denote small and large blocks, respectively). The potential outcomes are generated as

$$Y_i(z)=(B_i/M)^{2z+1}+\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}(z)+\varepsilon_i(z),\quad i=1,...,n,\quad z=0,1,$$

where $\boldsymbol{x}_i$ is generated from a p-dimensional multivariate normal distribution $N(0,\Sigma)$ with $\Sigma_{ij}=\rho^{|i-j|}$; the first s elements of $\boldsymbol{\beta}(z)$ are generated from the t_3 distribution, and the remaining elements are zero; and $\varepsilon_i(z)$ is generated from a normal distribution with a mean of zero and variance of σ^2 such that the signal-to-noise ratio equals 10. We set $p=400$, $s=10$, and $\rho=0.6$. The potential outcomes and covariates are generated once and then kept fixed.

We consider six designs: complete randomization (block = no, rerand = no), rerandomization (block = no, rerand = yes), and stratified randomization/rerandomization with equal/unequal propensity scores (block = eq/uneq, rerand = no/yes). We set $\boldsymbol{w}_i$ as the first $k = 4$ dimensions of $\boldsymbol{x}_i$ and $p_a = 0.001$ for rerandomization. The propensity scores $e_{[m]}$'s are equal to 0.5 or evenly spaced in value between 0.3 and 0.7. For each design, we consider the difference-in-means estimator (unadj) and two Lasso-adjusted estimators $\hat{\tau}_{\text{lasso}}$ and $\hat{\tau}_{\text{lasso2}}$ for equal/unequal propensity scores, respectively. We use the R package "glmnet" to fit the solution path of the Lasso. We choose the tuning parameter in the Lasso via 10-fold cross-validation (the number of the selected covariates is set to be less than $n/3$). We replicate the randomization/rerandomization 1000 times to evaluate the repeated sampling properties.

Figure 1 shows the distributions (violin plot) of different estimators. All of the distributions are symmetric around the true value of the average treatment effect. The distributions of the Lasso-adjusted estimators are more concentrated than those of the unadjusted estimators in both the randomization and rerandomization cases. Tables 1–3 present several summary statistics of different estimators. First, for all designs, the absolute value of the bias of each estimator is considerably smaller than its standard deviation (sd). Second, compared with the unadjusted estimator without rerandomization, the Lasso-adjusted estimator reduces the standard deviation by 40%–67%. Third, the empirical coverage probabilities (cp) of all estimators reach the nominal level. Fourth, compared with the unadjusted estimator without rerandomization, the Lasso-adjusted estimator decreases the mean confidence interval lengths (length) by 17%–73%. Finally, the Lasso-adjusted estimators in the stratified randomization and stratified rerandomization cases exhibit a comparable performance, as indicated by our theory. However, stratified rerandomization is preferable when a few important covariates needed to be balanced in the design stage. Thus, our final recommendation is to implement stratified rerandomization in the design stage and use the Lasso-adjusted estimator in the analysis stage.

7. Real data illustration

We use two real data sets to illustrate the performance of the proposed methods. The first data set pertains to a clinical trial that aimed at assessing the efficacy and safety of panitumumab combined with cisplatin and fluorouracil as first-line treatment for patients with recurrent or metastatic squamous-cell carcinoma of the head and neck (SCCHN) (Vermorken et al., 2013). Patients are stratified into $M = 8$ blocks based on previous treatment, primary tumour site, and performance status. The patients in each block are randomly assigned to the treatment and control groups with equal probability ($e_{[m]} = 1/2$). The outcome of interest is the logarithm of the progression-free survival days (time from randomization to disease progression or death). There are 30 baseline covariates for each patient, such as demographic variables and physiological indicators. We consider adjusting for the main effect, quadratic terms of the continuous covariates, and two-way interactions, resulting in a design matrix $\boldsymbol{X}$ with $p = 464$ columns (covariates) and $n = 428$ rows (observations). To evaluate the repeated sampling properties of different estimators, we must determine all the potential outcomes. For illustration purposes, we generate a synthetic data set. We use the Lasso to fit two sparse linear models for the treatment and control groups, respectively, and impute the potential outcomes by

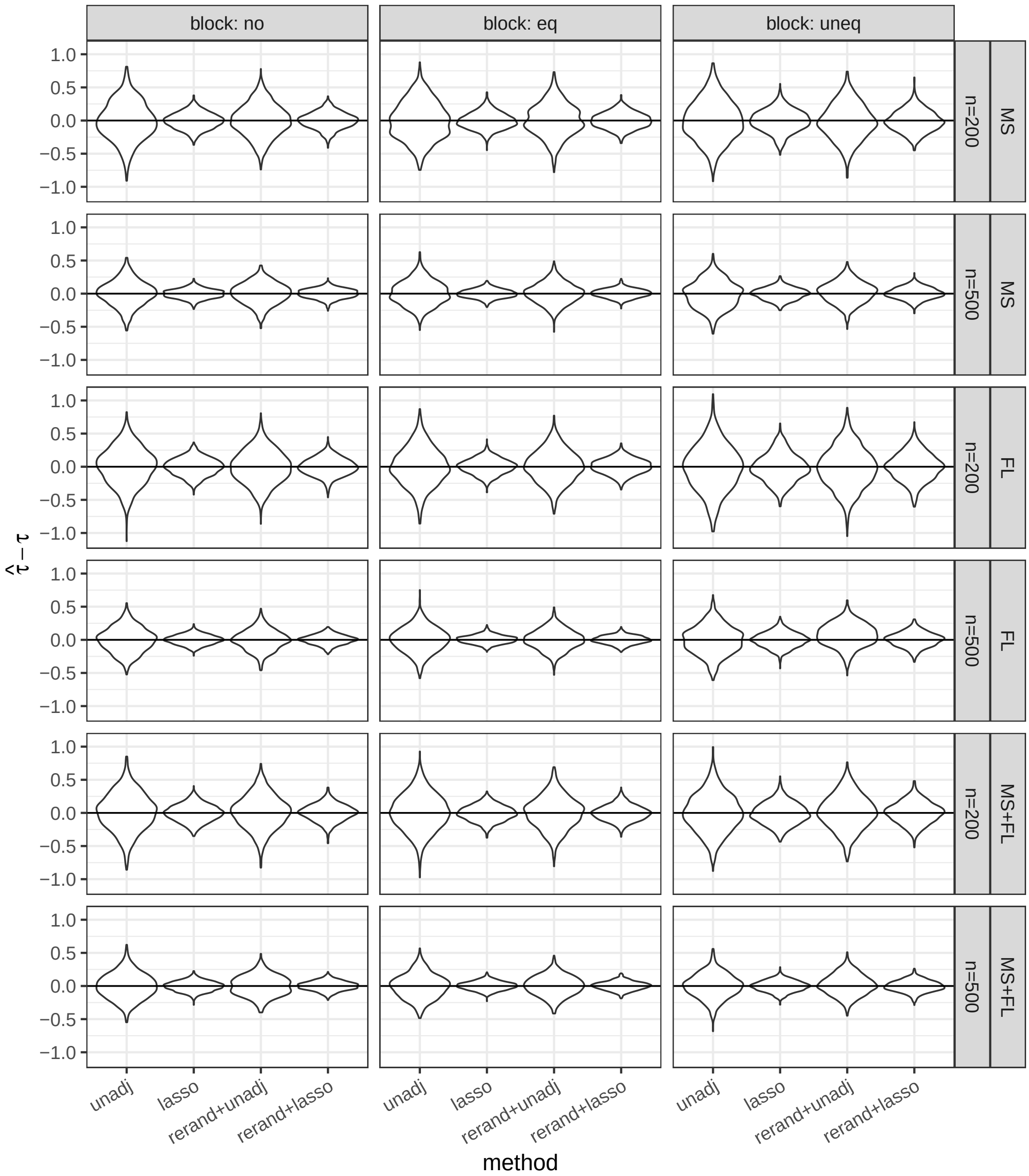

Fig. 1. Distributions of the average treatment effect estimators minus the true value of the average treatment effect in six cases.

Table 1. Performance of different estimators in the scenario of many small blocks.

n	block	rerand	est	bias	sd	sd%	rmse	cp	length	le%
200	no	no	unadj	1(1)	29(1)	-	29(1)	99(0)	152(0)	-
			lasso	0(0)	12(0)	58	12(0)	98(0)	58(0)	62
		yes	unadj	0(0)	24(1)	16	24(1)	100(0)	138(0)	9
			lasso	0(0)	12(0)	58	12(0)	98(0)	58(0)	62
	eq	no	unadj	0(1)	28(1)	-	28(1)	100(0)	152(0)	-
			lasso	0(0)	12(0)	56	12(0)	94(1)	48(0)	68
		yes	unadj	1(1)	24(1)	13	24(1)	99(0)	138(0)	10
			lasso	0(0)	12(0)	58	12(0)	96(1)	48(0)	68
	uneq	no	unadj	0(1)	30(1)	-	30(1)	99(0)	158(0)	-
			lasso	0(0)	16(0)	46	16(0)	100(0)	126(0)	21
		yes	unadj	2(1)	26(1)	13	26(1)	99(0)	140(0)	12
			lasso	1(0)	16(0)	47	16(0)	100(0)	126(0)	20
500	no	no	unadj	0(0)	19(0)	-	19(0)	99(0)	98(0)	-
			lasso	0(0)	7(0)	63	7(0)	98(0)	33(0)	66
		yes	unadj	0(0)	16(0)	16	16(0)	100(0)	88(0)	10
			lasso	0(0)	7(0)	62	7(0)	98(0)	33(0)	66
	eq	no	unadj	1(1)	18(0)	-	18(0)	99(0)	97(0)	-
			lasso	0(0)	7(0)	63	7(0)	96(1)	26(0)	73
		yes	unadj	0(0)	15(0)	15	15(0)	99(0)	87(0)	10
			lasso	0(0)	7(0)	64	7(0)	96(1)	26(0)	73
	uneq	no	unadj	1(1)	20(0)	-	20(0)	99(0)	101(0)	-
			lasso	0(0)	9(0)	56	9(0)	100(0)	79(0)	22
		yes	unadj	2(1)	16(0)	22	16(0)	100(0)	89(0)	12
			lasso	0(0)	9(0)	57	9(0)	100(0)	79(0)	22

Block indicates whether stratification is performed and whether the propensity scores of the blocks are equal; rerand indicates whether rerandomization is performed; bias indicates the absolute bias; sd is the standard deviation; sd% is the percentage decrease in the standard deviation relative to that for the unadjusted estimator without rerandomization; rmse is the root-mean-squared error; cp is the empirical coverage probability; length is the mean confidence interval length; le% is the percentage decrease in the confidence interval length relative to that associated with the unadjusted estimator without rerandomisation. The numbers in brackets are the corresponding standard errors estimated using the bootstrap with 500 replications. Bias, sd, rmse, cp, length, and their standard errors are multiplied by 100.

Table 2. Performance of different estimators in the scenario of a few large blocks.

n	block	rerand	est	bias	sd	sd%	rmse	cp	length	le%
200	no	no	unadj	1(1)	28(1)	-	28(1)	100(0)	153(0)	-
			lasso	0(0)	13(0)	53	13(0)	99(0)	60(0)	61
		yes	unadj	1(1)	25(1)	11	25(1)	100(0)	139(0)	9
			lasso	0(0)	13(0)	53	13(0)	98(0)	60(0)	61
	eq	no	unadj	0(1)	28(1)	-	28(1)	99(0)	151(0)	-
			lasso	0(0)	11(0)	59	11(0)	98(0)	54(0)	64
		yes	unadj	1(1)	24(0)	15	24(0)	100(0)	137(0)	9
			lasso	0(0)	11(0)	60	11(0)	98(0)	54(0)	64
	uneq	no	unadj	2(1)	34(1)	-	34(1)	98(0)	165(0)	-
			lasso	2(1)	21(0)	39	21(0)	100(0)	128(0)	22
		yes	unadj	2(1)	30(1)	12	30(1)	98(0)	146(0)	12
			lasso	0(0)	20(0)	40	20(0)	100(0)	129(0)	22
500	no	no	unadj	0(0)	18(0)	-	18(0)	100(0)	98(0)	-
			lasso	0(0)	7(0)	60	7(0)	99(0)	35(0)	64
		yes	unadj	0(0)	15(0)	18	15(0)	100(0)	88(0)	10
			lasso	0(0)	7(0)	61	7(0)	98(0)	35(0)	64
	eq	no	unadj	0(0)	19(0)	-	19(0)	99(0)	97(0)	-
			lasso	1(0)	6(0)	67	6(0)	98(0)	31(0)	68
		yes	unadj	0(0)	15(0)	20	15(0)	100(0)	87(0)	10
			lasso	0(0)	6(0)	67	6(0)	99(0)	31(0)	68
	uneq	no	unadj	1(1)	21(0)	-	21(0)	99(0)	107(0)	-
			lasso	0(0)	11(0)	47	11(0)	100(0)	79(0)	26
		yes	unadj	7(1)	17(0)	18	19(0)	99(0)	92(0)	14
			lasso	0(0)	11(0)	49	11(0)	100(0)	79(0)	26

Block indicates whether stratification is performed and whether the propensity scores of the blocks are equal; rerand indicates whether rerandomization is performed; bias indicates the absolute bias; sd is the standard deviation; sd% is the percentage decrease in the standard deviation relative to that for the unadjusted estimator without rerandomization; rmse is the root-mean-squared error; cp is the empirical coverage probability; length is the mean confidence interval length; le% is the percentage decrease in the confidence interval length relative to that associated with the unadjusted estimator without rerandomisation. The numbers in brackets are the corresponding standard errors estimated using the bootstrap with 500 replications. Bias, sd, rmse, cp, length, and their standard errors are multiplied by 100.

Table 3. Performance of different estimators in the scenario of hybrid blocks.

n	block	rerand	est	bias	sd	sd%	rmse	cp	length	le%
200	no	no	unadj	2(1)	28(1)	-	28(1)	99(0)	152(0)	-
			lasso	0(0)	13(0)	55	13(0)	98(0)	60(0)	61
		yes	unadj	1(1)	25(1)	11	25(1)	99(0)	139(0)	9
			lasso	0(0)	13(0)	55	13(0)	98(0)	60(0)	61
	eq	no	unadj	1(1)	27(1)	-	27(1)	100(0)	153(0)	-
			lasso	1(0)	12(0)	58	12(0)	97(1)	50(0)	67
		yes	unadj	0(0)	24(1)	12	24(1)	100(0)	138(0)	10
			lasso	0(0)	11(0)	58	11(0)	97(1)	50(0)	67
	uneq	no	unadj	0(1)	29(1)	-	29(1)	99(0)	153(0)	-
			lasso	1(0)	16(0)	44	16(0)	100(0)	127(0)	17
		yes	unadj	1(1)	26(1)	11	26(1)	100(0)	141(0)	8
			lasso	0(0)	16(0)	44	16(0)	100(0)	127(0)	17
500	no	no	unadj	0(0)	19(0)	-	19(0)	99(0)	98(0)	-
			lasso	0(0)	8(0)	60	8(0)	98(0)	34(0)	65
		yes	unadj	0(0)	15(0)	21	15(0)	100(0)	88(0)	10
			lasso	0(0)	7(0)	63	7(0)	99(0)	34(0)	65
	eq	no	unadj	0(0)	18(0)	-	18(0)	100(0)	96(0)	-
			lasso	0(0)	6(0)	64	6(0)	98(0)	29(0)	70
		yes	unadj	0(0)	15(0)	18	15(0)	100(0)	87(0)	10
			lasso	0(0)	6(0)	64	6(0)	97(1)	29(0)	70
	uneq	no	unadj	0(0)	18(0)	-	18(0)	99(0)	96(0)	-
			lasso	0(0)	8(0)	55	8(0)	100(0)	79(0)	18
		yes	unadj	2(0)	16(0)	14	16(0)	100(0)	89(0)	8
			lasso	0(0)	8(0)	53	8(0)	100(0)	79(0)	18

Block indicates whether stratification is performed and whether the propensity scores of the blocks are equal; rerand indicates whether rerandomization is performed; bias indicates the absolute bias; sd is the standard deviation; sd% is the percentage decrease in the standard deviation relative to that for the unadjusted estimator without rerandomization; rmse is the root-mean-squared error; cp is the empirical coverage probability; length is the mean confidence interval length; le% is the percentage decrease in the confidence interval length relative to that associated with the unadjusted estimator without rerandomization. The numbers in brackets are the corresponding standard errors estimated using the bootstrap with 500 replications. Bias, sd, rmse, cp, length, and their standard errors are multiplied by 100.

using the fitted models. We use the same blocking and propensity scores for each block as those in the original experiment and implement either stratified randomization or stratified rerandomization in the design stage. For rerandomization, we use the following covariates: age, sex, red blood cells, and white blood cells.

The second data set pertains to the "Opportunity Knocks" (OK) randomized experiment aimed at evaluating the effect of academic achievement awards on the academic performance of college students (Angrist et al., 2014). We focus on second-year college students. Based on the sex and discretized high school grades, the students are stratified into $M = 8$ blocks with sizes ranging from 42 to 90. In each block, only about 20 students are assigned to the treatment group (receiving incentives), and thus, the propensity scores are significantly different across blocks. We consider the average grades at the end of the year as the outcome. There are 23 baseline covariates, such as demographic variables and grade point average (GPA) in the previous year. We consider adjusting for the main effect, quadratic terms of the continuous covariates, and two-way interactions. The design matrix $\boldsymbol{X}$ has $p = 253$ columns (covariates) and $n = 506$ rows (observations). To evaluate the repeated sampling properties of different estimators, which depend on all of the potential outcomes, we match the treated units to the control units by using the 23 baseline covariates (Sekhon et al., 2011) to impute the unobserved potential outcomes. We use the same blocking and propensity scores for each block as those in the original experiment and implement either stratified randomization or stratified rerandomization in the design stage. For the rerandomization, we use the following covariates: age, high school grades, and GPA of the previous year.

For both data sets, we replicate the stratified randomization/rerandomization 1000 times to evaluate the repeated sampling properties of different estimators. Figure 2 and Table 4 show the results. Rerandomization is the preferable strategy, and the Lasso-adjusted estimator is superior to the unadjusted estimator. These conclusions are similar to those derived in the simulation studies.

8. Discussion

This study is aimed at enhancing the estimation and inference efficiencies of the average treatment effect in randomized experiments when many baseline covariates are available. We propose novel methods to combine blocking, rerandomization, and regression adjustment using the Lasso. Under mild conditions, we obtain the asymptotic distributions of the Lasso-adjusted average treatment effect estimators when blocking or rerandomization or both are implemented in the design stage. We demonstrate that the proposed Lasso-adjusted estimators enhance, or, at least, do not deteriorate the precision compared with that associated with the unadjusted estimator. Our results are randomization-based, allowing the outcome data generating model to be mis-specified. In addition, we define Neyman-type conservative variance estimators to construct asymptotically conservative confidence intervals or tests for the average treatment effect. Our final recommendation is to use blocking and rerandomization in the design stage to balance a subset of covariates that are most relevant to the potential outcomes and then implement regression adjustment using the Lasso in the analysis stage to adjust for the remaining covariate imbalances. Similar to the findings reported by Li and Ding (2020),

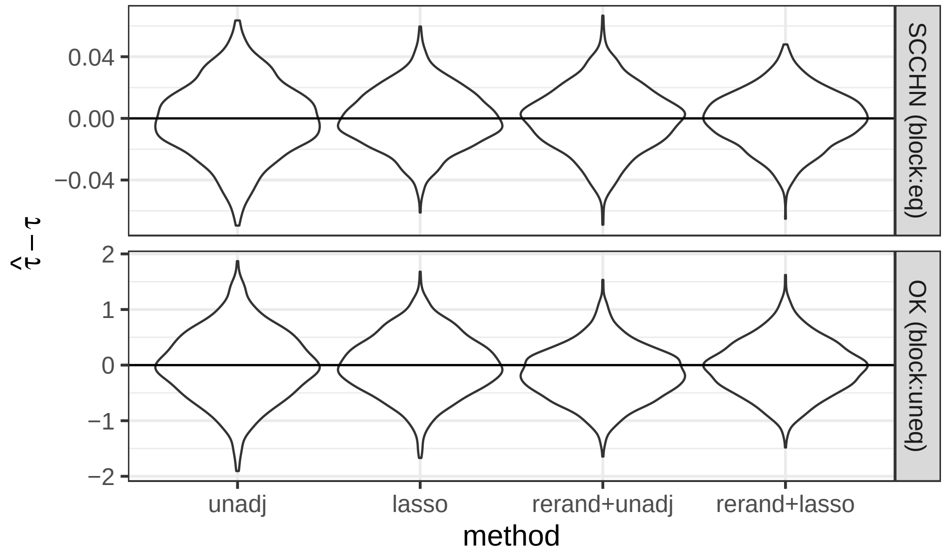

Fig. 2. Distributions of the average treatment effect estimators minus the true value of the average treatment effect for two real data sets.

Table 4. Performance of different estimators for two real data sets: SCCHN and OK.

rerand	est	bias	sd	sd%	rmse	cp	length	le%
				SCCHN				
no	unadj	0.1(0.06)	2.4(0.05)	-	2.4(0.05)	99(0.3)	12.3(0.01)	-
	lasso	0.02(0.04)	1.8(0.04)	23	1.8(0.04)	98(0.4)	8.8(0.02)	28
yes	unadj	0.07(0.05)	2(0.04)	15	2(0.04)	100(0.2)	11(0.01)	10
	lasso	0.04(0.04)	1.8(0.04)	26	1.8(0.04)	99(0.3)	8.8(0.02)	28
				OK				
no	unadj	0.4(1.2)	61(1.3)	-	61(1.3)	98(0.4)	294(0.4)	-
	lasso	0.2(1)	52(1.2)	14	52(1.2)	99(0.3)	256(0.5)	13
yes	unadj	15.6(1.5)	47(1.1)	22	49(1.1)	99(0.3)	263(0.4)	11
	lasso	4.1(1.5)	47(1)	22	47(1)	100(0.2)	257(0.4)	13

Rerand indicates whether rerandomization is performed; bias indicates the absolute bias; sd is the standard deviation; sd% is the percentage decrease in the standard deviation relative to the unadjusted estimator without rerandomization; rmse is the root-mean-squared error; cp is the empirical coverage probability; length is the mean confidence interval length; le% is the percentage decrease in the length relative to the unadjusted estimator without rerandomization. The numbers in brackets are the corresponding standard errors estimated using the bootstrap with 500 replications. Bias, sd, rmse, cp, length, and their standard errors are multiplied by 100.

when rerandomization or the combination of blocking and rerandomization is used in the design stage, the Lasso should consider all of the covariates used in the rerandomization to ensure efficiency gains.

To render the theory and methods more intuitive, we focus on inferring the average treatment effect for a binary treatment. Our analysis can be generalized to multiple value treatments, including factorial experiments as particular cases (Fisher, 1935; Yates, 1937; Dasgupta et al., 2015; Li et al., 2020). Moreover, it may be interesting to extend our results to other complicated settings, for example, binary outcomes based on penalized logistic regression (Freedman, 2008b; Zhang et al., 2008; Moore and van der Lann, 2009), regression adjustment using the Lasso in the event of noncompliance (Imbens and Angrist, 1994; Angrist and Imbens, 1995; Angrist et al., 1996), and the use of other machine learning methods such as random forest (Wager et al., 2016; Wu and Gagnon-Bartsch, 2018) in the analysis stage.

A limitation of our analysis is that the proposed Lasso-adjusted point and variance estimators require each block to have at least two treated and two control units. Thus, our results do not cover the case of randomized block experiments in which certain blocks are "small" in the sense that they consist of only one treated or one control unit (Pashley and Miratrix, 2020). The matched-pair design (Imai, 2008; Fogarty, 2018b) and finely stratified randomized experiments (Fogarty, 2018a) are special cases of this setting. It may be worth investigating methods to generalize the regression-adjusted estimators proposed by Fogarty (2018b) and Fogarty (2018a) to high-dimensional settings and use the variance estimator proposed by Pashley and Miratrix (2020) to handle such small blocks.

Acknowledgements

The second author was supported by the National Natural Science Foundation of China (Grant No. 12071242). The third author was supported by the National Natural Science Foundation of China (Grant No. 12001557); the Youth Talent Development Support Program (QYP202104), the Emerging Interdisciplinary Project, the Disciplinary Funding, and the School of Statistics and Mathematics in Central University of Finance and Economics.

References

Angrist, J. D. and Imbens, G. W. (1995) Two-stage least squares estimation of average causal effects in models with variable treatment intensity. *Journal of the American Statistical Association*, **90**, 431–442.

Angrist, J. D., Imbens, G. W. and Rubin, D. B. (1996) Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, **91**, 444–455.

Angrist, J. D., Oreopoulos, P. and Williams, T. (2014) When opportunity knocks, who answers? new evidence on college achievement awards. *Journal of Human Resources*, **49**, 572–610.

Athey, S. and Imbens, G. W. (2017) The state of applied econometrics: causality and policy evaluation. *Journal of Economic Perspectives*, **31**, 3–32.

Bloniarz, A., Liu, H., Zhang, C. H., Sekhon, J. S. and Yu, B. (2016) Lasso adjustments of treatment effect estimates in randomized experiments. *Proceedings of the National Academy of Sciences of the United States of America*, **113**, 7383–7390.

Bobkov, S. G. (2004) Concentration of normalized sums and a central limit theorem for noncorrelated random variables. *Annals of Probability*, **32**, 2884–2907.

Box, G. E. P., Hunter, J. S. and Hunter, W. G. (2005) *Statistics for Experimenters: Design, Innovation and Discovery.* New York: Wiley-Interscience.

Branson, Z., Dasgupta, T. and Rubin, D. B. (2016) Improving covariate balance in 2^k factorial designs via rerandomization with an application to a new york city department of education high school study. *The Annals of Applied Statistics*, **10**, 1958–1976.

Bugni, F. A., Canay, I. A. and Shaikh, A. M. (2018) Inference under covariate-adaptive randomization. *Journal of the American Statistical Association*, **113**, 1784–1796.

— (2019) Inference under covariate-adaptive randomization with multiple treatments. *Quantitative Economics*, **10**, 1747–1785.

Cox, D. R. (2007) Applied statistics: a review. *The Annals of Applied Statistics*, **1**, 1–16.

Dasgupta, T., Pillai, N. S. and Rubin, D. B. (2015) Causal inference from 2^K factorial designs by using potential outcomes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **77**, 727–753.

Efron, B. (1971) Forcing a sequential experiment to be balanced. *Biometrika*, **58**, 403–417.

Fisher, R. A. (1926) The arrangement of field experiments. *Journal of Ministry of Agriculture of Great Britain*, **33**, 503–513.

— (1935) *The Design of Experiments.* Oliver and Boyd, Edinburgh, 1st edn.

Fogarty, C. B. (2018a) On mitigating the analytical limitations of finely stratified experiments. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **80**, 1035–1056.

— (2018b) Regression-assisted inference for the average treatment effect in paired experiments. *Biometrika*, **105**, 994–1000.

Freedman, D. A. (2008a) On regression adjustments to experimental data. *Advances in Applied Mathematics*, **40**, 180–193.

— (2008b) Randomization does not justify logistic regression. *Statistical Science*, **23**, 237–249.

Green, S. B. and Byar, D. P. (1978) The effect of stratified randomization on size and power of statistical tests in clinical trials. *Journal of Chronic Diseases*, **31**, 445–454.

Higgins, M. J., Sävje, F. and Sekhon, J. S. (2015) Blocking estimators and inference under the Neyman–Rubin model. *arXiv preprint arXiv:1510.01103*.

Imai, K. (2008) Variance identification and efficiency analysis in randomized experiments under the matched-pair design. *Statistics in Medicine*, **27**, 4857–4873.

Imai, K., King, G. and Stuart, E. A. (2008) Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **171**, 481–502.

Imbens, G. W. (2011) Experimental design for unit and cluster randomid trials. *International Initiative for Impact Evaluation Paper*.

Imbens, G. W. and Angrist, J. D. (1994) Identification and estimation of local average treatment effects. *Econometrica*, **62**, 467–475.

Imbens, G. W. and Rubin, D. B. (2015) *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction.* Cambridge University Press.

Johansson, P. and Schultzberg, M. (2020) Rerandomization strategies for balancing covariates using pre-experimental longitudinal data. *Journal of Computational and Graphical Statistics*, **29**, 798–813.

Krieger, A. M., Azriel, D. and Kapelner, A. (2019) Nearly random designs with greatly improved balance. *Biometrika*, **106**, 695–701.

Lei, L. and Ding, P. (2020) Regression adjustment in completely randomized experiments with a diverging number of covariates. *Biometrika*, in press.

Li, X. and Ding, P. (2017) General forms of finite population central limit theorems with applications to causal inference. *Journal of the American Statistical Association*, **112**, 1759–1769.

— (2020) Rerandomization and regression adjustment. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **82**, 241–268.

Li, X., Ding, P. and Rubin, D. B. (2018) Asymptotic theory of rerandomization in treatment-control experiments. *Proceedings of the National Academy of Sciences of the United States of America*, **115**, 9157–9162.

— (2020) Rerandomization in 2^k factorial experiments. *The Annals of Statistics*, **48**, 43–63.

Lin, W. (2013) Agnostic notes on regression adjustments to experimental data: reexamining Freedman's critique. *The Annals of Applied Statistics*, **7**, 295–318.

Liu, H., Tu, F. and Ma, W. (2020) A general theory of regression adjustment for covariate-adaptive randomization: OLS, Lasso, and beyond. *arXiv preprint arXiv:2011.09734*.

Liu, H. and Yang, Y. (2020) Regression-adjusted average treatment effect estimators in stratified randomized experiments. *Biometrika*, **107**, 935–948.

Ma, W., Qin, Y., Li, Y. and Hu, F. (2020a) Statistical inference for covariate-adaptive randomization procedures. *Journal of the American Statistical Association*, **115**, 1488–1497.

Ma, W., Tu, F. and Liu, H. (2020b) Regression analysis for covariate-adaptive randomization: a robust and efficient inference perspective. *arXiv preprint arXiv:2009.02287*.

Massart, P. (1986) Rates of convergence in the central limit theorem for empirical processes. In *Geometrical and Statistical Aspects of Probability in Banach Spaces* (eds. X. Fernique, B. Heinkel, P.-A. Meyer and M. B. Marcus), vol. 1193, 73–109. Berlin, Heidelberg: Springer.

Miratrix, L. W., Sekhon, J. S. and Yu, B. (2013) Adjusting treatment effect estimates by post-stratification in randomized experiments. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **75**, 369–396.

Moore, K. L. and van der Lann, M. J. (2009) Covariate adjustment in randomized trials with binary outcomes: targeted maximum likelihood estimation. *Statistics in Medicine*, **28**, 39–64.

Morgan, K. L. and Rubin, D. B. (2012) Rerandomization to improve covariate balance in experiments. *The Annals of Statistics*, **40**, 1263–1282.

— (2015) Rerandomization to balance tiers of covariates. *Journal of the American Statistical Association*, **110**, 1412–1421.

Mukerjee, R., Dasgupta, T. and Rubin, D. B. (2018) Using standard tools from finite population sampling to improve causal inference for complex experiments. *Journal of the American Statistical Association*, **113**, 868–881.

Pashley, N. E. and Miratrix, L. W. (2020) Insights on variance estimation for blocked and matched pairs designs. *Journal of Educational and Behavioral Statistics*, in press.

Rosenbaum, P. R. (2010) *Design of Observational Studies.* New York: Springer.

Rosenberger, W. F. and Lachin, J. M. (2015) *Randomization in clinical trials: theory and practice.* John Wiley & Sons, 2nd edn.

Rosenberger, W. F. and Sverdlov, O. (2008) Handling covariates in the design of clinical trials. *Statistical Science*, **23**, 404–419.

Rubin, D. B. (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, **66**, 688–701.

— (1980) Randomization analysis of experimental data: the Fisher randomization test comment. *Journal of the American Statistical Association*, **75**, 591–593.

Schochet, P. Z. (2016) Statistical theory for the RCT-YES software: design-based causal inference for the RCTs (2nd ed.). *Tech. rep.*, Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Analytic Technical Assistance and Development.

Schultzberg, M. and Johansson, P. (2019) Re-randomization: a complement or substitute for stratification in randomized experiments? *Working Paper, Department of Statistics, Uppsala University, 2019:4.*

Sekhon, J. S. et al. (2011) Multivariate and propensity score matching software with automated balance optimization: the matching package for R. *Journal of Statistical Software*, **42**.

Splawa-Neyman, J., Dabrowska, D. M. and Speed, T. P. (1990) On the application of probability theory to agricultural experiments. *Statistical Science*, **5**, 465–472.

Tibshirani, R. J. (1996) Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B (Methodological)*, **58**, 267–288.

Tolstikhin, I. (2017) Concentration inequalities for samples without replacement. *Theory of Probability & Its Applications*, **61**, 462–481.

Vermorken, J. B., Stöhlmacher-Williams, J., Davidenko, I., Licitra, L., Winquist, E., Villanueva, C., Foa, P., Rottey, S., Skladowski, K., Tahara, M. et al. (2013) Cisplatin and fluorouracil with or without panitumumab in patients with recurrent or metastatic squamous-cell carcinoma of the head and neck (spectrum): an open-label phase 3 randomised trial. *The Lancet Oncology*, **14**, 697–710.

Wager, S., Du, W., Taylor, J. and Tibshirani, R. J. (2016) High-dimensional regression adjustments in randomized experiments. *Proceedings of the National Academy of Sciences of the United States of America*, **113**, 12673–12678.

Wang, B., Susukida, R., Mojtabai, R., Amin-Esmaeili, M. and Rosenblum, M. (2019) Model-robust inference for clinical trials that improve precision by stratified randomization and adjustment for additional baseline variables. *arXiv preprint arXiv:1910.13954.*

Wang, X., Wang, T. and Liu, H. (2020) Rerandomization in stratified randomized experiments. *arXiv preprint arXiv:2011.05195.*

Wilk, M. B. (1955) The randomization analysis of a generalized randomized block design. *Biometrika*, **42**, 70–9.

Wu, E. and Gagnon-Bartsch, J. A. (2018) The LOOP estimator: adjusting for covariates in randomized experiments. *Evaluation Review*, **42**, 458–488.

Yang, L. and Tsiatis, A. A. (2001) Efficiency study of estimators for a treatment effect in a pretest–posttest trial. *The American Statistician*, **55**, 314–321.

Yates, F. (1937) *The design and analysis of factorial experiments.* Technical Communication No. 35, Imperial Bureau of Soil Sciences, Harpenden.

Ye, T., Shao, J. and Zhao, Q. (2020) Principles for covariate adjustment in analyzing randomized clinical trials. *arXiv preprint arXiv:2009.11828.*

Ye, T., Yi, Y. and Shao, J. (2021) Inference on average treatment effect under minimization and other covariate-adaptive randomization methods. *Biometrika*, in press.

Zelen, M. (1974) The randomization and stratification of patients to clinical trials. *Journal of Chronic Diseases*, **27**, 365–375.

Zhang, M., Tsiatis, A. A. and Davidian, M. (2008) Improving efficiency of inferences in randomized clinical trials using auxiliary covariates. *Biometrics*, **55**, 707–715.

Zhang, Y., Wu, H. and Cheng, L. (2012) Some new deformation formulas about variance and covariance. In *2012 Proceedings of International Conference on Modelling, Identification and Control*, 987–992. IEEE.

Zhou, Q., Ernst, P. A., Morgan, K. L., Rubin, D. B. and Zhang, A. (2018) Sequential rerandomization. *Biometrika*, **105**, 745–752.

A. Some preliminary results

A.1. Bounds for sampling without replacement

The connection between randomized experiments and sampling surveys has been discussed in depth by many scholars (Lin, 2013; Li and Ding, 2017; Mukerjee et al., 2018; Lei and Ding, 2020). Both of them are based on a probability model of sampling without replacement from a finite population. We start by introducing Bobkov's inequality, which is a powerful tool to prove the concentration inequality for sampling without replacement, and it is also the key to prove our main results. In this section, we consider the completely randomized experiments, that is,

$$P(\boldsymbol{Z}=\boldsymbol{z})=\frac{n_1!n_0!}{n!}, \quad \sum_{i=1}^{n} I(z_i=1)=n_1, \quad z_i=0,1.$$

We denote the propensity score by $e_c=n_1/n$. The value space of $\boldsymbol{Z}$ is defined as

$$\mathcal{G}=\left\{\boldsymbol{z}=(z_1,\ldots,z_n)\in\{0,1\}^n : z_1+\cdots+z_n=n_1\right\}.$$

For every $\boldsymbol{z}\in\mathcal{G}$, we pick a pair of units (i,j) such that $z_i=1$ and $z_j=0$, and switch the value of z_i and z_j to obtain a "neighbour" of $\boldsymbol{z}$, denoted by $\boldsymbol{z}^{i,j}$. Clearly, for different (i,j), $\boldsymbol{z}$ has totally $n_1(n-n_1)$ neighbours. For every real-valued function f on $\mathcal{G}$, we define the discrete gradient $\nabla f(\boldsymbol{z})=(f(\boldsymbol{z})-f\left(\boldsymbol{z}^{i,j}\right))_{i,j}$, which is a $n_1(n-n_1)$ dimensional vector. We define the ℓ_2 norm of $\nabla f(\boldsymbol{z})$ through

$$\|\nabla f(\boldsymbol{z})\|_2^2=\sum_{i:z_i=1}\sum_{j:z_j=0}\left|f(\boldsymbol{z})-f\left(\boldsymbol{z}^{i,j}\right)\right|^2.$$

LEMMA 1 (BOBKOV (2004)). *For every real-valued function f on $\mathcal{G}$, if $\|\nabla f(\boldsymbol{z})\|_2\leqslant\sigma$ for all $z\in\mathcal{G}$, then*

$$E\exp\left[t\left\{f(\boldsymbol{Z})-Ef(\boldsymbol{Z})\right\}\right]\leqslant\exp\left\{\sigma^2t^2/(n+2)\right\}, \quad t\in\mathbb{R}.$$

Consider two sequences of real numbers $(a_1,\ldots,a_n)$ and $(b_1,\ldots,b_n)$, we denote

$$\bar{a}=\frac{1}{n}\sum_{i=1}^{n}a_i, \quad \bar{a}_1=\bar{a}_1(\boldsymbol{Z})=\frac{1}{n}\sum_{i=1}^{n}Z_ia_i,$$

$$S_{ab}=\frac{1}{n-1}\sum_{i=1}^{n}(a_i-\bar{a})(b_i-\bar{b}), \quad s_{ab}=s_{ab}(\boldsymbol{Z})=\frac{1}{n_1-1}\sum_{i=1}^{n}Z_i(a_i-\bar{a}_1)(b_i-\bar{b}_1).$$

The following result from Zhang et al. (2012) is useful to bound $\|\nabla f(\boldsymbol{z})\|_2^2$ in our proof.

LEMMA 2 (ZHANG ET AL. (2012)). *We have*

$$\sum_{i=1}^{n}(a_i-\bar{a})(b_i-\bar{b})=\frac{1}{n}\sum_{1\leqslant i<j\leqslant n}(a_i-a_j)(b_i-b_j).$$

Next, we apply Bobkov's inequality to derive the bounds for the sample mean and the sample covariance, respectively.

LEMMA 3. *For $t \in \mathbb{R}$, we have*

$$E \exp\{t(\bar{a}_1 - \bar{a})\} \leqslant \exp\{\sigma_{\text{mean}}^2 t^2/(n+2)\},$$

where $\sigma_{\text{mean}}^2 = e_c^{-2} n^{-1} \sum_{i=1}^n (a_i - \bar{a})^2$.

LEMMA 4. *For $t \in \mathbb{R}$, we have*

$$E \exp\{t(s_{ab} - S_{ab})\} \leqslant \exp\{\sigma_{\text{cov}}^2 t^2/(n+2)\},$$

where

$$\sigma_{\text{cov}}^2 = \left\{ \sqrt{\frac{1}{e_c^2 n} \sum_{i=1}^n (a_i - \bar{a})^2 (b_i - \bar{b})^2} + \sqrt{\frac{8}{e_c^3 n^2} \sum_{i=1}^n (a_i - \bar{a})^2 \sum_{i=1}^n (b_i - \bar{b})^2} \right\}^2.$$

REMARK 11. *Massart (1986) first established Lemma 3 with a better constant. His proof assumed that n/n_1 is an integer. Bloniarz et al. (2016) generalized Massart's result, allowing n/n_1 to be a non-integer. Tolstikhin (2017) proved Lemma 3 based on Bobkov's approach. The bound for the sample covariance in Lemma 4 is novel, to the best of our knowledge.*

A.2. Concentration inequalities for stratified randomization

We apply Lemmas 3 and 4 in each block to obtain concentration inequalities for the weighted sample mean and sample covariance under stratified random sampling without replacement.

PROPOSITION 1. *Consider a sequence of real numbers $(a_1, \ldots, a_n)$. For any $t > 0$,*

$$P\Big(\sum_{m=1}^M \pi_{[m]}(\bar{a}_{[m]1} - \bar{a}_{[m]}) \geqslant t\Big) \leqslant \exp\Big\{ -\frac{nt^2}{4\sigma_a^2}\Big\},$$

where $\sigma_a^2 = (1/n)\sum_{m=1}^M \sum_{i \in [m]} (a_i - \bar{a}_{[m]})^2 / e_{[m]}^2$.

PROOF (PROOF OF PROPOSITION 1). For any $\lambda > 0$ and $t > 0$, by Markov's inequality, we have

$$\begin{aligned} P\Big(\sum_{m=1}^M \pi_{[m]}(\bar{a}_{[m]1} - \bar{a}_{[m]}) \geqslant t\Big) &\leqslant \exp\{-\lambda t\} \cdot E \exp\Big\{\lambda \sum_{m=1}^M \pi_{[m]}(\bar{a}_{[m]1} - \bar{a}_{[m]})\Big\} \\ &= \exp\{-\lambda t\} \cdot \prod_{m=1}^M E \exp\Big\{\lambda \pi_{[m]}(\bar{a}_{[m]1} - \bar{a}_{[m]})\Big\}. \end{aligned}$$

By Lemma 3, we have

$$\prod_{m=1}^{M} E\exp\Big\{\lambda\pi_{[m]}(\bar{a}_{[m]1}-\bar{a}_{[m]})\Big\} \leqslant \prod_{m=1}^{M}\exp\Big\{\frac{\lambda^2\pi_{[m]}^2}{e_{[m]}^2 n_{[m]}^2}\sum_{i\in[m]}(a_i-\bar{a}_{[m]})^2\Big\}$$
$$= \exp\Big\{\frac{\lambda^2}{n^2}\sum_{m=1}^{M}\sum_{i\in[m]}(a_i-\bar{a}_{[m]})^2/e_{[m]}^2\Big\}$$
$$= \exp\Big\{\frac{\lambda^2}{n}\sigma_a^2\Big\}.$$

Thus,

$$P\Big(\sum_{m=1}^{M}\pi_{[m]}(\bar{a}_{[m]1}-\bar{a}_{[m]}) \geqslant t\Big) \leqslant \exp\Big\{-\lambda t + \frac{\lambda^2}{n}\sigma_a^2\Big\}.$$

The conclusion follows by taking $\lambda = nt/(2\sigma_a^2)$.

PROPOSITION 2. *Consider two sequences of real numbers $(a_1,\dots,a_n)$ and $(b_1,\dots,b_n)$. For any $t>0$,*

$$P\Big(\sum_{m=1}^{M}\pi_{[m]}(s_{[m]ab}-S_{[m]ab}) \geqslant t\Big) \leqslant \exp\Big\{-\frac{nt^2}{60(\kappa_a^4\kappa_b^4)^{1/2}}\Big\},$$

where $\kappa_a^4 = (1/n)\sum_{m=1}^{M}\sum_{i\in[m]}(a_i-\bar{a}_{[m]})^4/e_{[m]}^3$ *and* $\kappa_b^4 = (1/n)\sum_{m=1}^{M}\sum_{i\in[m]}(b_i-\bar{b}_{[m]})^4/e_{[m]}^3$.

PROOF (PROOF OF PROPOSITION 2). We denote

$$\sigma_{[m]\mathrm{cov}}^2 = \Bigg\{\sqrt{\frac{1}{e_{[m]}^2 n_{[m]}}\sum_{i\in[m]}(a_i-\bar{a}_{[m]})^2(b_i-\bar{b}_{[m]})^2} + \sqrt{\frac{8}{e_{[m]}^3 n_{[m]}^2}\sum_{i\in[m]}(a_i-\bar{a}_{[m]})^2\sum_{i\in[m]}(b_i-\bar{b}_{[m]})^2}\Bigg\}^2.$$

For any $\lambda>0$ and $t>0$, by Markov's inequality and Lemma 4, we have

$$P\Big(\sum_{m=1}^{M}\pi_{[m]}(s_{[m]ab}-S_{[m]ab}) \geqslant t\Big) \leqslant \exp\{-\lambda t\}\cdot E\exp\Big\{\lambda\sum_{m=1}^{M}\pi_{[m]}(s_{[m]ab}-S_{[m]ab})\Big\}$$
$$= \exp\{-\lambda t\}\cdot\prod_{m=1}^{M}E\exp\Big\{\lambda\pi_{[m]}(s_{[m]ab}-S_{[m]ab})\Big\}$$
$$\leqslant \exp\{-\lambda t\}\cdot\prod_{m=1}^{M}\exp\Big\{\frac{\lambda^2\pi_{[m]}^2}{n_{[m]}}\sigma_{[m]\mathrm{cov}}^2\Big\}$$
$$= \exp\Big\{-\lambda t + \frac{\lambda^2}{n}\sum_{m=1}^{M}\pi_{[m]}\sigma_{[m]\mathrm{cov}}^2\Big\}. \qquad (3)$$

By Minkowski's inequality, we have

$$\begin{aligned}\sum_{m=1}^{M} \pi_{[m]} \sigma_{[m]\mathrm{cov}}^2 = \sum_{m=1}^{M} \Bigg\{ & \sqrt{\frac{\pi_{[m]}}{e_{[m]}^2 n_{[m]}} \sum_{i \in [m]} (a_i - \bar{a}_{[m]})^2 (b_i - \bar{b}_{[m]})^2} \\ & + \sqrt{\frac{8\pi_{[m]}}{e_{[m]}^3 n_{[m]}^2} \sum_{i \in [m]} (a_i - \bar{a}_{[m]})^2 \sum_{i \in [m]} (b_i - \bar{b}_{[m]})^2} \Bigg\}^2 \\ \leqslant \Bigg\{ & \sqrt{\frac{1}{n} \sum_{m=1}^{M} \sum_{i \in [m]} (a_i - \bar{a}_{[m]})^2 (b_i - \bar{b}_{[m]})^2 / e_{[m]}^2} \\ & + \sqrt{\frac{8}{n} \sum_{m=1}^{M} \sum_{i \in [m]} (a_i - \bar{a}_{[m]})^2 \sum_{i \in [m]} (b_i - \bar{b}_{[m]})^2 / (e_{[m]}^3 n_{[m]})} \Bigg\}^2 . \end{aligned} \tag{4}$$

Then, we deal with the two terms in (4) separately. By Cauchy–Schwarz inequality,

$$\begin{aligned} & \frac{1}{n} \sum_{m=1}^{M} \sum_{i \in [m]} (a_i - \bar{a}_{[m]})^2 (b_i - \bar{b}_{[m]})^2 / e_{[m]}^2 \\ \leqslant & \Big\{ \frac{1}{n} \sum_{m=1}^{M} \sum_{i \in [m]} (a_i - \bar{a}_{[m]})^4 / e_{[m]}^2 \Big\}^{1/2} \Big\{ \frac{1}{n} \sum_{m=1}^{M} \sum_{i \in [m]} (b_i - \bar{b}_{[m]})^4 / e_{[m]}^2 \Big\}^{1/2} \\ \leqslant & \Big\{ \frac{1}{n} \sum_{m=1}^{M} \sum_{i \in [m]} (a_i - \bar{a}_{[m]})^4 / e_{[m]}^3 \Big\}^{1/2} \Big\{ \frac{1}{n} \sum_{m=1}^{M} \sum_{i \in [m]} (b_i - \bar{b}_{[m]})^4 / e_{[m]}^3 \Big\}^{1/2} . \end{aligned} \tag{5}$$

Applying Cauchy–Schwarz inequality twice, we have

$$\begin{aligned} & \frac{8}{n} \sum_{m=1}^{M} \Big[\sum_{i \in [m]} (a_i - \bar{a}_{[m]})^2 \sum_{i \in [m]} (b_i - \bar{b}_{[m]})^2 / (e_{[m]}^3 n_{[m]}) \Big] \\ \leqslant & \frac{8}{n} \Big\{ \sum_{m=1}^{M} \Big[\sum_{i \in [m]} (a_i - \bar{a}_{[m]})^2 \Big]^2 / (e_{[m]}^3 n_{[m]}) \Big\}^{1/2} \Big\{ \sum_{m=1}^{M} \Big[\sum_{i \in [m]} (b_i - \bar{b}_{[m]})^2 \Big]^2 / (e_{[m]}^3 n_{[m]}) \Big\}^{1/2} \\ \leqslant & 8 \Big\{ \frac{1}{n} \sum_{m=1}^{M} \sum_{i \in [m]} (a_i - \bar{a}_{[m]})^4 / e_{[m]}^3 \Big\}^{1/2} \Big\{ \frac{1}{n} \sum_{m=1}^{M} \sum_{i \in [m]} (b_i - \bar{b}_{[m]})^4 / e_{[m]}^3 \Big\}^{1/2} . \end{aligned} \tag{6}$$

Recall that

$$\kappa_a^4 = \frac{1}{n} \sum_{m=1}^{M} \sum_{i \in [m]} (a_i - \bar{a}_{[m]})^4 / e_{[m]}^3, \quad \kappa_b^4 = \frac{1}{n} \sum_{m=1}^{M} \sum_{i \in [m]} (b_i - \bar{b}_{[m]})^4 / e_{[m]}^3 .$$

Combining (4), (5), and (6), we have

$$\sum_{m=1}^{M} \pi_{[m]} \sigma_{[m]\mathrm{cov}}^2 \leqslant (1 + 2\sqrt{2})^2 (\kappa_a^4 \kappa_b^4)^{1/2} \leqslant 15 (\kappa_a^4 \kappa_b^4)^{1/2} . \tag{7}$$

Combining (3) and (7), we have

$$P\Big(\sum_{m=1}^{M}\pi_{[m]}(s_{[m]ab}-S_{[m]ab})\geqslant t\Big)\leqslant \exp\Big\{-\lambda t+\frac{15\lambda^2}{n}(\kappa_a^4\kappa_b^4)^{1/2}\Big\}.$$

The conclusion follows by taking $\lambda = nt/\{30(\kappa_a^4\kappa_b^4)^{1/2}\}$.

A.3. *Asymptotic theory for stratified randomization and stratified rerandomization*

In this section, we present useful results on the asymptotic distributions of $\hat{\tau}_{\text{unadj}}$ under stratified randomization (Liu and Yang, 2020) and stratified rerandomization (Wang et al., 2020), respectively. The maximum second moment condition used in this section is weaker than the bounded fourth moment condition in the main text.

CONDITION 9. *The maximum block-specific squared distance of the potential outcomes satisfies* $n^{-1}\max_{m=1,\dots,M}\max_{i\in[m]}\big\{Y_i(z)-\bar{Y}_{[m]}(z)\big\}^2\to 0$, *for* $z=0,1$.

CONDITION 10. *The weighted variances* $\sum_{m=1}^{M}\pi_{[m]}S^2_{[m]Y(1)}/e_{[m]}$, $\sum_{m=1}^{M}\pi_{[m]}S^2_{[m]Y(0)}/(1-e_{[m]})$, *and* $\sum_{m=1}^{M}\pi_{[m]}S^2_{[m]\{Y(1)-Y(0)\}}$ *tend to finite limits, positive for the first two, and the limit of* $\sum_{m=1}^{M}\pi_{[m]}\big[S^2_{[m]Y(1)}/e_{[m]}+S^2_{[m]Y(0)}/(1-e_{[m]})-S^2_{[m]\{Y(1)-Y(0)\}}\big]$ *is strictly positive.*

PROPOSITION 3 (LIU AND YANG (2020)). *If Conditions 1 and 9 hold, then*

$$\sum_{m=1}^{M}\pi_{[m]}\Big\{\frac{s^2_{[m]Y(z)}}{e_{[m]}}\Big\}-\sum_{m=1}^{M}\pi_{[m]}\Big\{\frac{S^2_{[m]Y(z)}}{e_{[m]}}\Big\}\xrightarrow{p}0,\quad z=0,1.$$

Furthermore, if Condition 10 holds, we have $\sqrt{n}(\hat{\tau}_{\text{unadj}}-\tau)\xrightarrow{d}N(0,\sigma^2_{\text{unadj}})$ *and*

$$\hat{\sigma}^2_{\text{unadj}}-\sigma^2_{\text{unadj}}-\sum_{m=1}^{M}\pi_{[m]}S^2_{[m]\{Y(1)-Y(0)\}}\xrightarrow{p}0.$$

According to Proposition 2 in Wang et al. (2020), the covariance of $\sqrt{n}(\hat{\tau}_{\text{unadj}},\hat{\tau}_{\boldsymbol{w}}^{\mathrm{T}})^{\mathrm{T}}$ under stratified randomization is

$$V\{Y,\boldsymbol{W}\}\;=\;\begin{pmatrix} V_{YY} & V_{Y\boldsymbol{W}}\\ V_{\boldsymbol{W}Y} & V_{\boldsymbol{W}\boldsymbol{W}}\end{pmatrix},$$

where

$$V_{YY}=\sigma^2_{\text{unadj}},\quad V_{\boldsymbol{W}\boldsymbol{W}}=\sum_{m=1}^{M}\pi_{[m]}\Big\{\frac{S^2_{[m]\boldsymbol{W}}}{e_{[m]}(1-e_{[m]})}\Big\},$$

$$V_{\boldsymbol{W}Y}=V_{Y\boldsymbol{W}}^{\mathrm{T}}=\sum_{m=1}^{M}\pi_{[m]}\Big\{\frac{S_{[m]\boldsymbol{W}Y(1)}}{e_{[m]}}+\frac{S_{[m]\boldsymbol{W}Y(0)}}{1-e_{[m]}}\Big\}.$$

The asymptotic distribution of $\hat{\tau}_{\text{unadj}}$ under stratified rerandomization depends on the squared multiple correlation between $\hat{\tau}_{\text{unadj}}$ and $\hat{\tau}_{\boldsymbol{w}}$,

$$R^2_{Y,\boldsymbol{W}} = \lim_{n\to\infty}(V_{Y\boldsymbol{W}}V^{-1}_{\boldsymbol{W}\boldsymbol{W}}V_{\boldsymbol{W}Y})/V_{YY}.$$

Since $V_{YY} = \sigma^2_{\text{unadj}}$, we can conservatively estimate V_{YY} by $\hat{\sigma}^2_{\text{unadj}}$. In addition, we can consistently estimate $V_{\boldsymbol{W}Y}$ by $\hat{V}_{\boldsymbol{W}Y} = \sum_{m=1}^{M}\pi_{[m]}\left\{s_{[m]\boldsymbol{W}Y(1)}/e_{[m]} + s_{[m]\boldsymbol{W}Y(0)}/(1-e_{[m]})\right\}$ and directly calculate $V_{\boldsymbol{W}\boldsymbol{W}}$ based on $\boldsymbol{w}_i$. Then, we can estimate $R^2_{Y,\boldsymbol{W}}$ by

$$\hat{R}^2_{Y,\boldsymbol{W}} = (\hat{V}^{\mathrm{T}}_{\boldsymbol{W}Y}V^{-1}_{\boldsymbol{W}\boldsymbol{W}}\hat{V}_{\boldsymbol{W}Y})/\hat{\sigma}^2_{\text{unadj}}.$$

Recall that ε_0, $D_1, \ldots, D_k$ are independent standard normal random variables and $L_{k,a} = \{D_1 \mid \sum_{i=1}^{k} D_i^2 \leqslant a\}$. Let $v_{k,a} = P(\chi^2_{k+2} \leqslant a)/P(\chi^2_k \leqslant a) \in (0,1)$. We can conservatively estimate the variance of $\hat{\tau}_{\text{unadj}}$ under stratified rerandomization by

$$\hat{\sigma}^2_{\text{unadj}|\mathcal{M}} = \hat{\sigma}^2_{\text{unadj}}\{1-(1-v_{k,a})\hat{R}^2_{Y,\boldsymbol{W}}\}.$$

CONDITION 11. *The maximum block-specific squared distance of the covariates $\boldsymbol{w}_i$ satisfies $n^{-1}\max_{m=1,\ldots,M}\max_{i\in[m]}||\boldsymbol{w}_i - \bar{\boldsymbol{w}}_{[m]}||^2_\infty \to 0$.*

CONDITION 12. *The weighted covariances $\sum_{m=1}^{M}\pi_{[m]}S^2_{[m]\boldsymbol{W}}/e_{[m]}$, $\sum_{m=1}^{M}\pi_{[m]}S^2_{[m]\boldsymbol{W}}/(1-e_{[m]})$, $\sum_{m=1}^{M}\pi_{[m]}S_{[m]\boldsymbol{W}Y(1)}/e_{[m]}$, and $\sum_{m=1}^{M}\pi_{[m]}S_{[m]\boldsymbol{W}Y(0)}/(1-e_{[m]})$ tend to finite limits, and the limit of $V\{Y,\boldsymbol{W}\}$ is strictly positive definite.*

PROPOSITION 4 (WANG ET AL. (2020)). *If Conditions 1, 9–12 hold, then $P(\mathcal{M}_a) \to P(\chi^2_k \leqslant a)$ and*

$$\{\sqrt{n}(\hat{\tau}_{\text{unadj}} - \tau) \mid \mathcal{M}_a\} \xrightarrow{d} \sigma_{\text{unadj}}\left\{\sqrt{1-R^2_{Y,\boldsymbol{W}}}\varepsilon_0 + \sqrt{R^2_{Y,\boldsymbol{W}}}L_{k,a}\right\}.$$

Moreover, the asymptotic variance of $\sqrt{n}(\hat{\tau}_{\text{unadj}} - \tau) \mid \mathcal{M}_a$ is

$$\sigma^2_{\text{unadj}|\mathcal{M}} = \sigma^2_{\text{unadj}}\{1-(1-v_{k,a})R^2_{Y,\boldsymbol{W}}\} \leqslant \sigma^2_{\text{unadj}},$$

and

$$\hat{\sigma}^2_{\text{unadj}|\mathcal{M}} \xrightarrow{p} \sigma^2_{\text{unadj}|\mathcal{M}} + \sum_{m=1}^{M}\pi_{[m]}S^2_{[m]\{Y(1)-Y(0)\}}.$$

Proposition 4 shows that the asymptotic distribution of $\hat{\tau}_{\text{unadj}}$ under stratified rerandomization is a truncated normal distribution, and its asymptotic variance is less than or equal to that of $\hat{\tau}_{\text{unadj}}$ under stratified randomization. Moreover, we can conservatively estimate the asymptotic variance.

B. Proof of main results

B.1. Proof of Theorem 1

PROOF. We first prove the asymptotic normality. By definition,

$$\hat{\tau}_{\text{lasso}} = \sum_{m=1}^{M}\pi_{[m]}\left[\left\{\bar{Y}_{[m]1} - (\bar{\boldsymbol{x}}_{[m]1} - \bar{\boldsymbol{x}}_{[m]})^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{\text{lasso},1}\right\} - \left\{\bar{Y}_{[m]0} - (\bar{\boldsymbol{x}}_{[m]0} - \bar{\boldsymbol{x}}_{[m]})^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{\text{lasso},0}\right\}\right].$$

According to the decomposition of potential outcomes,

$$Y_i(z) = \bar{Y}_{[m]}(z) + (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_{[m]})^{\mathrm{T}}\boldsymbol{\beta}(z) + \varepsilon_i(z), \quad i \in [m], \quad z = 0, 1,$$

we have

$$\bar{Y}_{[m]z} = \bar{Y}_{[m]}(z) + (\bar{\boldsymbol{x}}_{[m]z} - \bar{\boldsymbol{x}}_{[m]})^{\mathrm{T}}\boldsymbol{\beta}(z) + \bar{\varepsilon}_{[m]z}, \quad z = 0, 1.$$

Then,

$$\begin{aligned}
\hat{\tau}_{\text{lasso}} - \tau = & \sum_{m=1}^{M} \pi_{[m]} \Big[\Big\{ \bar{Y}_{[m]1} - \bar{Y}_{[m]}(1) - (\bar{\boldsymbol{x}}_{[m]1} - \bar{\boldsymbol{x}}_{[m]})^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{\text{lasso},1} \Big\} \\
& - \Big\{ \bar{Y}_{[m]0} - \bar{Y}_{[m]}(0) - (\bar{\boldsymbol{x}}_{[m]0} - \bar{\boldsymbol{x}}_{[m]})^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{\text{lasso},0} \Big\} \Big] \\
= & \sum_{m=1}^{M} \pi_{[m]} \big(\bar{\varepsilon}_{[m]1} - \bar{\varepsilon}_{[m]0} \big) + \sum_{m=1}^{M} \pi_{[m]} \big(\bar{\boldsymbol{x}}_{[m]1} - \bar{\boldsymbol{x}}_{[m]} \big)^{\mathrm{T}} \Big\{ \boldsymbol{\beta}(1) - \hat{\boldsymbol{\beta}}_{\text{lasso},1} \Big\} \\
& - \sum_{m=1}^{M} \pi_{[m]} \Big(\bar{\boldsymbol{x}}_{[m]0} - \bar{\boldsymbol{x}}_{[m]} \Big)^{\mathrm{T}} \Big\{ \boldsymbol{\beta}(0) - \hat{\boldsymbol{\beta}}_{\text{lasso},0} \Big\}.
\end{aligned}$$

Applying Proposition 3 to $\varepsilon_i(1)$ and $\varepsilon_i(0)$, we have

$$\sqrt{n} \sum_{m=1}^{M} \pi_{[m]} (\bar{\varepsilon}_{[m]1} - \bar{\varepsilon}_{[m]0}) \xrightarrow{d} N(0, \sigma^2_{\text{lasso}}).$$

It suffices for the asymptotic normality of $\hat{\tau}_{\text{lasso}}$ to show that

$$\sqrt{n} \sum_{m=1}^{M} \pi_{[m]} \big(\bar{\boldsymbol{x}}_{[m]1} - \bar{\boldsymbol{x}}_{[m]} \big)^{\mathrm{T}} \Big\{ \boldsymbol{\beta}(1) - \hat{\boldsymbol{\beta}}_{\text{lasso},1} \Big\} \xrightarrow{p} 0, \tag{8}$$

$$\sqrt{n} \sum_{m=1}^{M} \pi_{[m]} \big(\bar{\boldsymbol{x}}_{[m]0} - \bar{\boldsymbol{x}}_{[m]} \big)^{\mathrm{T}} \Big\{ \boldsymbol{\beta}(0) - \hat{\boldsymbol{\beta}}_{\text{lasso},0} \Big\} \xrightarrow{p} 0. \tag{9}$$

We will only prove (8), as the proof of (9) is similar. By Hölder inequality,

$$\Big| \sum_{m=1}^{M} \pi_{[m]} \big(\bar{\boldsymbol{x}}_{[m]1} - \bar{\boldsymbol{x}}_{[m]} \big)^{\mathrm{T}} \Big\{ \boldsymbol{\beta}(1) - \hat{\boldsymbol{\beta}}_{\text{lasso},1} \Big\} \Big| \leqslant \Big\| \sum_{m=1}^{M} \pi_{[m]} \big(\bar{\boldsymbol{x}}_{[m]1} - \bar{\boldsymbol{x}}_{[m]} \big) \Big\|_{\infty} \Big\| \boldsymbol{\beta}(1) - \hat{\boldsymbol{\beta}}_{\text{lasso},1} \Big\|_{1}.$$

To bound $\| \sum_{m=1}^{M} \pi_{[m]} \big(\bar{\boldsymbol{x}}_{[m]1} - \bar{\boldsymbol{x}}_{[m]} \big) \|_{\infty}$, we apply Proposition 1 and obtain the following Lemma.

Lemma 5. *If Conditions 1–2 hold, we have*

$$\Big\| \sum_{m=1}^{M} \pi_{[m]} \big(\bar{\boldsymbol{x}}_{[m]1} - \bar{\boldsymbol{x}}_{[m]} \big) \Big\|_{\infty} = O_p\left(\sqrt{\frac{\log p}{n}} \right).$$

To bound $\|\boldsymbol{\beta}(1) - \hat{\boldsymbol{\beta}}_{\text{lasso},1}\|_1$, we apply Proposition 2 and obtain the following two lemmas.

LEMMA 6. *Suppose that Conditions 1–2 and 5 hold, then, for the event*

$$\mathcal{L} = \left\{ \|\hat{\Sigma}_{\boldsymbol{X}\varepsilon(1)}\|_\infty \leqslant \eta\lambda_1 \right\}.$$

we have $P(\mathcal{L}) \to 1$.

LEMMA 7. *Suppose that Conditions 1–2 hold, then, for the event*

$$\mathcal{K} = \left\{ \|\hat{\Sigma}_{\boldsymbol{X}\boldsymbol{X}} - \Sigma_{\boldsymbol{X}\boldsymbol{X}}\|_\infty \leqslant \sqrt{180L/c^3}\sqrt{\log p/n} \right\},$$

we have $P(\mathcal{K}) \to 1$.

With the above two lemmas, we can be conditional on $\mathcal{L} \cap \mathcal{K}$ and establish the l_1 convergence rate of $\hat{\boldsymbol{\beta}}_{\text{lasso},1}$.

LEMMA 8. *If Conditions 1–2 and 4–5 hold, then* $\left\|\boldsymbol{\beta}(1) - \hat{\boldsymbol{\beta}}_{\text{lasso},1}\right\|_1 = O_p(s\lambda_1)$.

We will prove Lemmas 5–8 in Section C. For now, by Lemmas 5 and 8, we have

$$\left|\sqrt{n}\sum_{m=1}^{M} \pi_{[m]}\left(\bar{\boldsymbol{x}}_{[m]1} - \bar{\boldsymbol{x}}_{[m]}\right)^{\mathrm{T}}\left\{\boldsymbol{\beta}(1) - \hat{\boldsymbol{\beta}}_{\text{lasso},1}\right\}\right| = O_p\left(\sqrt{n}\cdot\sqrt{\frac{\log p}{n}}\cdot s\cdot\lambda_1\right) = o_p(1),$$

where the last equality is because of Condition 5.

Next, we compare the asymptotic variances of $\hat{\tau}_{\text{lasso}}$ and $\hat{\tau}_{\text{unadj}}$. By definition and the decomposition of potential outcomes, we have

$$\begin{aligned}
S^2_{[m]Y(1)} &= \frac{1}{n_{[m]}-1}\sum_{i\in[m]}\left\{Y_i(1) - \bar{Y}_{[m]}(1)\right\}^2 \\
&= \frac{1}{n_{[m]}-1}\sum_{i\in[m]}\left\{(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_{[m]})^{\mathrm{T}}\boldsymbol{\beta}(1) + \varepsilon_i(1)\right\}^2 \\
&= \boldsymbol{\beta}(1)^{\mathrm{T}}S^2_{[m]\boldsymbol{X}}\boldsymbol{\beta}(1) + S^2_{[m]\varepsilon(1)} + 2S^{\mathrm{T}}_{[m]\boldsymbol{X}\varepsilon(1)}\boldsymbol{\beta}(1). \qquad (10)
\end{aligned}$$

Similarly,

$$S^2_{[m]Y(0)} = \boldsymbol{\beta}(0)^{\mathrm{T}}S^2_{[m]\boldsymbol{X}}\boldsymbol{\beta}(0) + S^2_{[m]\varepsilon(0)} + 2S^{\mathrm{T}}_{[m]\boldsymbol{X}\varepsilon(0)}\boldsymbol{\beta}(0),$$

$$\begin{aligned}
S^2_{[m]\{Y(1)-Y(0)\}} =& \left\{\boldsymbol{\beta}(1) - \boldsymbol{\beta}(0)\right\}^{\mathrm{T}}S^2_{[m]\boldsymbol{X}}\left\{\boldsymbol{\beta}(1) - \boldsymbol{\beta}(0)\right\} + S^2_{[m]\{\varepsilon(1)-\varepsilon(0)\}} \\
&+ 2\left\{S_{[m]\boldsymbol{X}\varepsilon(1)} - S_{[m]\boldsymbol{X}\varepsilon(0)}\right\}^{\mathrm{T}}\left\{\boldsymbol{\beta}(1) - \boldsymbol{\beta}(0)\right\}.
\end{aligned}$$

Then, consider the mth block. With $\boldsymbol{\beta}_{[m]} = (1 - e_{[m]})\boldsymbol{\beta}(1) + e_{[m]}\boldsymbol{\beta}(0)$ and $\Delta^2_m = \boldsymbol{\beta}^{\mathrm{T}}_{[m]}S^2_{[m]\boldsymbol{X}}\boldsymbol{\beta}_m/\{e_{[m]}(1 - e_{[m]})\}$, we have

$$\begin{aligned}
&\left[\frac{S^2_{[m]Y(1)}}{e_{[m]}} + \frac{S^2_{[m]Y(0)}}{1 - e_{[m]}} - S^2_{[m]\{Y(1)-Y(0)\}}\right] - \left[\frac{S^2_{[m]\varepsilon(1)}}{e_{[m]}} + \frac{S^2_{[m]\varepsilon(0)}}{1 - e_{[m]}} - S^2_{[m]\{\varepsilon(1)-\varepsilon(0)\}}\right] \\
=&\frac{1}{e_{[m]}(1 - e_{[m]})}\boldsymbol{\beta}^{\mathrm{T}}_{[m]}S^2_{[m]\boldsymbol{X}}\boldsymbol{\beta}_m + \frac{2}{e_{[m]}}S^{\mathrm{T}}_{[m]\boldsymbol{X}\varepsilon(1)}\boldsymbol{\beta}_{[m]} + \frac{2}{1 - e_{[m]}}S^{\mathrm{T}}_{[m]\boldsymbol{X}\varepsilon(0)}\boldsymbol{\beta}_{[m]} \\
=&\Delta^2_m + 2\left\{\frac{S_{[m]\boldsymbol{X}\varepsilon(1)}}{e_{[m]}} + \frac{S_{[m]\boldsymbol{X}\varepsilon(0)}}{1 - e_{[m]}}\right\}^{\mathrm{T}}\boldsymbol{\beta}_{[m]}.
\end{aligned}$$

Thus,

$$\sigma^2_{\text{unadj}} - \sigma^2_{\text{lasso}} = \sum_{m=1}^{M} \pi_{[m]} \Delta_m^2 + 2 \sum_{m=1}^{M} \pi_{[m]} \Big\{ \frac{S^{\mathrm{T}}_{[m]\boldsymbol{X}\varepsilon(1)}}{e_{[m]}} \boldsymbol{\beta}_{[m]} + \frac{S^{\mathrm{T}}_{[m]\boldsymbol{X}\varepsilon(0)}}{1 - e_{[m]}} \boldsymbol{\beta}_{[m]} \Big\}.$$

It suffices to show that

$$\sum_{m=1}^{M} \pi_{[m]} \frac{S^{\mathrm{T}}_{[m]\boldsymbol{X}\varepsilon(1)}}{e_{[m]}} \boldsymbol{\beta}_{[m]} \to 0, \tag{11}$$

$$\sum_{m=1}^{M} \pi_{[m]} \frac{S^{\mathrm{T}}_{[m]\boldsymbol{X}\varepsilon(0)}}{1 - e_{[m]}} \boldsymbol{\beta}_{[m]} \to 0. \tag{12}$$

We only prove (11), as the proof of (12) is similar. Since $\{\boldsymbol{\beta}(z)\}_{S^c} = \mathbf{0}$ for $z = 0, 1$, then

$$(\boldsymbol{\beta}_{[m]})_{S^c} = \mathbf{0}, \quad \sum_{m=1}^{M} \pi_{[m]} \frac{S^{\mathrm{T}}_{[m]\boldsymbol{X}\varepsilon(1)}}{e_{[m]}} \boldsymbol{\beta}_{[m]} = \sum_{m=1}^{M} \pi_{[m]} \frac{S^{\mathrm{T}}_{[m]\boldsymbol{X}_S\varepsilon(1)}}{e_{[m]}} (\boldsymbol{\beta}_{[m]})_S.$$

By the definition of projection coefficients, the relevant covariates are orthogonal to the error terms in the sense that $\sum_{m=1}^{M} \pi_{[m]} S_{[m]\boldsymbol{X}_S\varepsilon(1)} = \mathbf{0}$. By simple algebra and Hölder inequality, we have

$$\begin{aligned}
\Big| \sum_{m=1}^{M} \pi_{[m]} \frac{S^{\mathrm{T}}_{[m]\boldsymbol{X}\varepsilon(1)}}{e_{[m]}} \boldsymbol{\beta}_{[m]} \Big| &= \Big| \sum_{m=1}^{M} \pi_{[m]} \frac{1 - e_{[m]}}{e_{[m]}} S^{\mathrm{T}}_{[m]\boldsymbol{X}_S\varepsilon(1)} \{\boldsymbol{\beta}(1)\}_S + \sum_{m=1}^{M} \pi_{[m]} S^{\mathrm{T}}_{[m]\boldsymbol{X}_S\varepsilon(1)} \{\boldsymbol{\beta}(0)\}_S \Big| \\
&= \Big| \sum_{m=1}^{M} \pi_{[m]} \frac{1 - e_{[m]}}{e_{[m]}} S^{\mathrm{T}}_{[m]\boldsymbol{X}_S\varepsilon(1)} \{\boldsymbol{\beta}(1)\}_S \Big| \\
&= \Big| \sum_{m=1}^{M} \pi_{[m]} \frac{1 - e_{[m]}}{e_{[m]}} S^{\mathrm{T}}_{[m]\boldsymbol{X}_S\varepsilon(1)} \{\boldsymbol{\beta}(1)\}_S \\
&\quad - \sum_{m=1}^{M} \pi_{[m]} \frac{1 - e^*}{e^*} S^{\mathrm{T}}_{[m]\boldsymbol{X}_S\varepsilon(1)} \{\boldsymbol{\beta}(1)\}_S \Big| \\
&\leqslant \max_{1 \leqslant m \leqslant M} \Big| \frac{1 - e_{[m]}}{e_{[m]}} - \frac{1 - e^*}{e^*} \Big| \cdot \sum_{m=1}^{M} \pi_{[m]} \big| S^{\mathrm{T}}_{[m]\boldsymbol{X}\varepsilon(1)} \boldsymbol{\beta}(1) \big|.
\end{aligned}$$

As $\max_{1 \leqslant m \leqslant M} |e_{[m]} - e^*| \to 0$ by Condition 6, we have

$$\max_{1 \leqslant m \leqslant M} \Big| \frac{1 - e_{[m]}}{e_{[m]}} - \frac{1 - e^*}{e^*} \Big| \to 0.$$

By the second moment condition of the potential outcomes and the fourth moment condition of the error terms (see Condition 2), we have

$$\begin{aligned}
&\sum_{m=1}^{M} \pi_{[m]} \left| S^{\mathrm{T}}_{[m]\boldsymbol{X}\varepsilon(1)} \boldsymbol{\beta}(1) \right| \\
=&\sum_{m=1}^{M} \pi_{[m]} \frac{1}{n_{[m]}-1} \Big| \sum_{i\in[m]} (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_{[m]})^{\mathrm{T}} \boldsymbol{\beta}(1) \varepsilon_i(1) \Big| \\
\leqslant&\frac{2}{n} \sum_{m=1}^{M} \sum_{i\in[m]} |(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_{[m]})^{\mathrm{T}} \boldsymbol{\beta}(1) \varepsilon_i(1)| \\
\leqslant&2 \left[\frac{1}{n} \sum_{m=1}^{M} \sum_{i\in[m]} \{(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_{[m]})^{\mathrm{T}} \boldsymbol{\beta}(1)\}^2 \cdot \frac{1}{n} \sum_{m=1}^{M} \sum_{i\in[m]} \varepsilon_i^2(1) \right]^{1/2} \\
\leqslant&4L^{1/2}. \qquad (13)
\end{aligned}$$

Thus,

$$\sum_{m=1}^{M} \pi_{[m]} \frac{S^{\mathrm{T}}_{[m]\boldsymbol{X}\varepsilon(1)}}{e_{[m]}} \boldsymbol{\beta}_{[m]} \to 0.$$

B.2. Proof of Theorem 2

To prove the result about the limit of $\hat{\sigma}^2_{\text{lasso}}$, it suffices to show that

$$\frac{n_1}{n_1 - \hat{s}(1) - 1} \sum_{m=1}^{M} \pi_{[m]} \frac{\hat{s}^2_{[m]\varepsilon(1)}}{e_{[m]}} - \sum_{m=1}^{M} \pi_{[m]} \frac{S^2_{[m]\varepsilon(1)}}{e_{[m]}} \xrightarrow{p} 0, \qquad (14)$$

$$\frac{n_0}{n_0 - \hat{s}(0) - 1} \sum_{m=1}^{M} \pi_{[m]} \frac{\hat{s}^2_{[m]\varepsilon(0)}}{1 - e_{[m]}} - \sum_{m=1}^{M} \pi_{[m]} \frac{S^2_{[m]\varepsilon(0)}}{1 - e_{[m]}} \xrightarrow{p} 0. \qquad (15)$$

We only prove (14), as the proof of (15) is similar. The following lemma bounds the number of the covariates selected by the lasso, $\hat{s}(z)$.

Lemma 9. *If Conditions 1–2, 4–5, and 7 hold, then there exists a constant C independent of n, such that the following holds with probability tending to one,*

$$\hat{s}(1) \leqslant Cs, \quad \hat{s}(0) \leqslant Cs.$$

We will prove Lemma 9 later. According to Lemma 9 and Conditions 1 and 5,

$$\frac{n_1}{n_1 - \hat{s}(1) - 1} \xrightarrow{p} 1.$$

Thus, it suffices for (14) to show that

$$\sum_{m=1}^{M} \pi_{[m]} \frac{\hat{s}^2_{[m]\varepsilon(1)}}{e_{[m]}} - \sum_{m=1}^{M} \pi_{[m]} \frac{S^2_{[m]\varepsilon(1)}}{e_{[m]}} \xrightarrow{p} 0. \qquad (16)$$

By the definition of $\hat{s}^2_{[m]\varepsilon(1)}$, we have

$$\hat{s}^2_{[m]\varepsilon(1)} = \frac{1}{n_{[m]1}-1}\sum_{i\in[m]} Z_i\big\{Y_i(1)-\bar{Y}_{[m]1}-(\boldsymbol{x}_i-\bar{\boldsymbol{x}}_{[m]1})^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{\text{lasso},1}\big\}^2.$$

Plugging in the decomposition of the potential outcomes,

$$Y_i(1)=\bar{Y}_{[m]}(1)+(\boldsymbol{x}_i-\bar{\boldsymbol{x}}_{[m]})^{\mathrm{T}}\boldsymbol{\beta}(1)+\varepsilon_i(1),\quad i\in[m],$$

we have

$$\begin{aligned}
\hat{s}^2_{[m]\varepsilon(1)} =& \frac{1}{n_{[m]1}-1}\sum_{i\in[m]} Z_i\big[(\boldsymbol{x}_i-\bar{\boldsymbol{x}}_{[m]1})^{\mathrm{T}}\big\{\boldsymbol{\beta}(1)-\hat{\boldsymbol{\beta}}_{\text{lasso},1}\big\}+\varepsilon_i(1)-\bar{\varepsilon}_{[m]1}\big]^2\\
=& \frac{1}{n_{[m]1}-1}\sum_{i\in[m]} Z_i\big[(\boldsymbol{x}_i-\bar{\boldsymbol{x}}_{[m]1})^{\mathrm{T}}\big\{\boldsymbol{\beta}(1)-\hat{\boldsymbol{\beta}}_{\text{lasso},1}\big\}\big]^2\\
&+\frac{1}{n_{[m]1}-1}\sum_{i\in[m]} Z_i\big\{\varepsilon_i(1)-\bar{\varepsilon}_{[m]1}\big\}^2\\
&+\frac{2}{n_{[m]1}-1}\sum_{i\in[m]} Z_i\big[(\boldsymbol{x}_i-\bar{\boldsymbol{x}}_{[m]1})^{\mathrm{T}}\big\{\boldsymbol{\beta}(1)-\hat{\boldsymbol{\beta}}_{\text{lasso},1}\big\}\big]\big\{\varepsilon_i(1)-\bar{\varepsilon}_{[m]1}\big\}\\
=&\big\{\boldsymbol{\beta}(1)-\hat{\boldsymbol{\beta}}_{\text{lasso},1}\big\}^{\mathrm{T}}s^2_{[m]\boldsymbol{X}}\big\{\boldsymbol{\beta}(1)-\hat{\boldsymbol{\beta}}_{\text{lasso},1}\big\}+s^2_{[m]\varepsilon(1)}+2s^{\mathrm{T}}_{[m]\boldsymbol{X}\varepsilon(1)}\big\{\boldsymbol{\beta}(1)-\hat{\boldsymbol{\beta}}_{\text{lasso},1}\big\},
\end{aligned}$$

where $s^2_{[m]\boldsymbol{X}}=s^2_{[m]\boldsymbol{X}(1)}$ stands for the sample covariance of $\boldsymbol{X}$ for the treatment group. In the following, we will use this simplified notation if there is no exceptional clarity. Then,

$$\begin{aligned}
\sum_{m=1}^{M}\pi_{[m]}\frac{\hat{s}^2_{[m]\varepsilon(1)}}{e_{[m]}} =& \sum_{m=1}^{M}\frac{\pi_{[m]}}{e_{[m]}}\cdot\big\{\boldsymbol{\beta}(1)-\hat{\boldsymbol{\beta}}_{\text{lasso},1}\big\}^{\mathrm{T}}s^2_{[m]\boldsymbol{X}}\big\{\boldsymbol{\beta}(1)-\hat{\boldsymbol{\beta}}_{\text{lasso},1}\big\}\\
&+\sum_{m=1}^{M}\frac{\pi_{[m]}}{e_{[m]}}\cdot s^2_{[m]\varepsilon(1)}+2\sum_{m=1}^{M}\frac{\pi_{[m]}}{e_{[m]}}\cdot s^{\mathrm{T}}_{[m]\boldsymbol{X}\varepsilon(1)}\big\{\boldsymbol{\beta}(1)-\hat{\boldsymbol{\beta}}_{\text{lasso},1}\big\}.
\end{aligned}\tag{17}$$

For the first term on the right hand of (17), by Condition 1, we have

$$\begin{aligned}
&\sum_{m=1}^{M}\frac{\pi_{[m]}}{e_{[m]}}\cdot\big\{\boldsymbol{\beta}(1)-\hat{\boldsymbol{\beta}}_{\text{lasso},1}\big\}^{\mathrm{T}}s^2_{[m]\boldsymbol{X}}\big\{\boldsymbol{\beta}(1)-\hat{\boldsymbol{\beta}}_{\text{lasso},1}\big\}\\
\leqslant&\frac{1}{c}\big\{\boldsymbol{\beta}(1)-\hat{\boldsymbol{\beta}}_{\text{lasso},1}\big\}^{\mathrm{T}}\hat{\Sigma}_{\boldsymbol{X}\boldsymbol{X}}\big\{\boldsymbol{\beta}(1)-\hat{\boldsymbol{\beta}}_{\text{lasso},1}\big\}\\
\leqslant&\frac{1}{c}\cdot\|\boldsymbol{\beta}(1)-\hat{\boldsymbol{\beta}}_{\text{lasso},1}\|_1^2\cdot\|\hat{\Sigma}_{\boldsymbol{X}\boldsymbol{X}}\|_\infty.
\end{aligned}$$

By the fourth moment condition of the covariates (see Condition 2), we have $\|\Sigma_{\boldsymbol{X}\boldsymbol{X}}\|_\infty\leqslant C$ for some constant C. Then, by Lemma 7 and Condition 5, we have

$$\|\hat{\Sigma}_{\boldsymbol{X}\boldsymbol{X}}\|_\infty=O_p(1).$$

According to Lemma 8 and Condition 5, we have

$$\|\boldsymbol{\beta}(1) - \hat{\boldsymbol{\beta}}_{\text{lasso},1}\|_1 = O_p(s\lambda_1) = o_p(1).$$

Therefore,

$$\sum_{m=1}^{M} \pi_{[m]} \frac{1}{e_{[m]}} \cdot \left\{\boldsymbol{\beta}(1) - \hat{\boldsymbol{\beta}}_{\text{lasso},1}\right\}^{\mathrm{T}} s^2_{[m]\boldsymbol{X}} \left\{\boldsymbol{\beta}(1) - \hat{\boldsymbol{\beta}}_{\text{lasso},1}\right\} \xrightarrow{p} 0.$$

For the second term on the right-hand side of (17), applying Proposition 3 to $\varepsilon_i(1)$, we have

$$\sum_{m=1}^{M} \pi_{[m]} \frac{s^2_{[m]\varepsilon(1)}}{e_{[m]}} - \sum_{m=1}^{M} \pi_{[m]} \frac{S^2_{[m]\varepsilon(1)}}{e_{[m]}} \xrightarrow{p} 0.$$

The third term on the right-hand side of (17) tends to zero in probability by Cauchy-Schwarz inequality. Therefore, (16) holds.

Next, we compare the limit of $\hat{\sigma}^2_{\text{lasso}}$ and σ^2_{lasso}. Combining (14) and (15), we have

$$\lim_{n\to\infty} (\hat{\sigma}^2_{\text{lasso}} - \sigma^2_{\text{lasso}}) = \lim_{n\to\infty} \sum_{m=1}^{M} \pi_{[m]} S^2_{[m]\{\varepsilon(1)-\varepsilon(0)\}} \geqslant 0.$$

Finally, we compare the limits of $\hat{\sigma}^2_{\text{lasso}}$ and $\hat{\sigma}^2_{\text{unadj}}$. According to Proposition 3, the probability limit of $\hat{\sigma}^2_{\text{unadj}}$ is the limit of

$$\sum_{m=1}^{M} \pi_{[m]} \left\{ \frac{S^2_{[m]Y(1)}}{e_{[m]}} + \frac{S^2_{[m]Y(0)}}{1 - e_{[m]}} \right\}.$$

Then the difference between the limits of $\hat{\sigma}^2_{\text{lasso}}$ and $\hat{\sigma}^2_{\text{unadj}}$ is the limit of

$$\sum_{m=1}^{M} \pi_{[m]} \frac{S^2_{[m]\varepsilon(1)} - S^2_{[m]Y(1)}}{e_{[m]}} + \sum_{m=1}^{M} \pi_{[m]} \frac{S^2_{[m]\varepsilon(0)} - S^2_{[m]Y(0)}}{1 - e_{[m]}}.$$

It suffices to show that

$$\lim_{n\to\infty} \sum_{m=1}^{M} \pi_{[m]} \frac{S^2_{[m]\varepsilon(1)} - S^2_{[m]Y(1)}}{e_{[m]}} \leqslant 0, \quad \lim_{n\to\infty} \sum_{m=1}^{M} \pi_{[m]} \frac{S^2_{[m]\varepsilon(0)} - S^2_{[m]Y(0)}}{1 - e_{[m]}} \leqslant 0.$$

We will only prove the first statement, since the proof of the second statement is similar. We have shown that (see equation (10))

$$S^2_{[m]Y(1)} = \boldsymbol{\beta}(1)^{\mathrm{T}} S^2_{[m]\boldsymbol{X}} \boldsymbol{\beta}(1) + S^2_{[m]\varepsilon(1)} + 2 S^{\mathrm{T}}_{[m]\boldsymbol{X}\varepsilon(1)} \boldsymbol{\beta}(1).$$

Then,

$$\sum_{m=1}^{M} \pi_{[m]} \frac{S^2_{[m]\varepsilon(1)} - S^2_{[m]Y(1)}}{e_{[m]}} = -\boldsymbol{\beta}(1)^{\mathrm{T}} \Big(\sum_{m=1}^{M} \pi_{[m]} \frac{S^2_{[m]\boldsymbol{X}}}{e_{[m]}} \Big) \boldsymbol{\beta}(1) - 2 \sum_{m=1}^{M} \pi_{[m]} \frac{S^{\mathrm{T}}_{[m]\boldsymbol{X}\varepsilon(1)}}{e_{[m]}} \boldsymbol{\beta}(1).$$

Similar to the proof of (11), we have

$$\sum_{m=1}^{M} \pi_{[m]} \frac{S^{\mathrm{T}}_{[m]\boldsymbol{X}\varepsilon(1)}}{e_{[m]}} \boldsymbol{\beta}(1) \to 0, \quad n \to \infty.$$

Therefore,

$$\lim_{n\to\infty} \sum_{m=1}^{M} \pi_{[m]} \frac{S^2_{[m]\varepsilon(1)} - S^2_{[m]Y(1)}}{e_{[m]}} \leqslant 0.$$

B.3. Proof of Theorem 3

We first prove the asymptotic normality. By definition, we have

$$\begin{aligned}
\hat{\tau}_{\mathrm{lasso2}} - \tau =& \hat{\tau}_{\mathrm{unadj}} - \tau - \hat{\tau}_{\boldsymbol{x}}^{\mathrm{T}} \hat{\gamma}_{\mathrm{lasso}} \\
=& \sum_{m=1}^{M} \pi_{[m]} \Big[\big\{ \bar{Y}_{[m]1} - \bar{Y}_{[m]}(1) - \big(\bar{\boldsymbol{x}}_{[m]1} - \bar{\boldsymbol{x}}_{[m]} \big)^{\mathrm{T}} \hat{\gamma}_{\mathrm{lasso}} \big\} \\
& \quad - \big\{ \bar{Y}_{[m]0} - \bar{Y}_{[m]}(0) - \big(\bar{\boldsymbol{x}}_{[m]0} - \bar{\boldsymbol{x}}_{[m]} \big)^{\mathrm{T}} \hat{\gamma}_{\mathrm{lasso}} \big\} \Big].
\end{aligned}$$

Recall the decomposition of the potential outcomes,

$$Y_i(z) = \bar{Y}_{[m]}(z) + \big(\bar{\boldsymbol{x}}_i - \bar{\boldsymbol{x}}_{[m]} \big)^{\mathrm{T}} \boldsymbol{\gamma}_{\mathrm{proj}} + \varepsilon_i^*(z), \quad i \in [m], \quad z = 0, 1, \tag{18}$$

we have

$$\begin{aligned}
\hat{\tau}_{\mathrm{lasso2}} - \tau =& \sum_{m=1}^{M} \pi_{[m]} \big(\bar{\varepsilon}^*_{[m]1} - \bar{\varepsilon}^*_{[m]0} \big) + \sum_{m=1}^{M} \pi_{[m]} \big(\bar{\boldsymbol{x}}_{[m]1} - \bar{\boldsymbol{x}}_{[m]} \big)^{\mathrm{T}} \big(\boldsymbol{\gamma}_{\mathrm{proj}} - \hat{\gamma}_{\mathrm{lasso}} \big) \\
& - \sum_{m=1}^{M} \pi_{[m]} \big(\bar{\boldsymbol{x}}_{[m]0} - \bar{\boldsymbol{x}}_{[m]} \big)^{\mathrm{T}} \big(\boldsymbol{\gamma}_{\mathrm{proj}} - \hat{\gamma}_{\mathrm{lasso}} \big).
\end{aligned}$$

Applying Proposition 3 to $\varepsilon_i^*(z)$, we have

$$\sqrt{n} \sum_{m=1}^{M} \pi_{[m]} \big(\bar{\varepsilon}^*_{[m]1} - \bar{\varepsilon}^*_{[m]0} \big) \xrightarrow{d} N(0, \sigma^2_{\mathrm{lasso2}}). \tag{19}$$

Since Conditions 1–2 and 4–5 hold for $\varepsilon_i^{\Delta}(z)$ and the transformed covariates $\sqrt{\omega_{\boldsymbol{X},i}}\, \boldsymbol{x}_i$, we apply Lemma 8 to $\hat{\gamma}_{\mathrm{lasso},z}$ and obtain

$$\big\| \boldsymbol{\gamma}(z) - \hat{\gamma}_{\mathrm{lasso},z} \big\|_1 = O_p(s\lambda_z), \quad z = 0, 1,$$

which leads to $\|\boldsymbol{\gamma}_{\mathrm{proj}} - \hat{\gamma}_{\mathrm{lasso}}\|_1 = O_p(s\lambda_1 + s\lambda_0)$. Then, by Hölder inequality,

$$\begin{aligned}
\Big| \sum_{m=1}^{M} \pi_{[m]} \big(\bar{\boldsymbol{x}}_{[m]1} - \bar{\boldsymbol{x}}_{[m]} \big)^{\mathrm{T}} \big(\boldsymbol{\gamma}_{\mathrm{proj}} - \hat{\gamma}_{\mathrm{lasso}} \big) \Big| &\leqslant \Big\| \sum_{m=1}^{M} \pi_{[m]} \big(\bar{\boldsymbol{x}}_{[m]1} - \bar{\boldsymbol{x}}_{[m]} \big) \Big\|_{\infty} \cdot \big\| \boldsymbol{\gamma}_{\mathrm{proj}} - \hat{\gamma}_{\mathrm{lasso}} \big\|_1 \\
&= O_p\Big(\sqrt{\frac{\log p}{n}} \Big) \cdot O_p(s\lambda_1 + s\lambda_0),
\end{aligned}$$

where the last equality is because of Lemma 5. Therefore,

$$\left|\sqrt{n}\sum_{m=1}^{M}\pi_{[m]}\big(\bar{\boldsymbol{x}}_{[m]1}-\bar{\boldsymbol{x}}_{[m]}\big)^{\mathrm{T}}\big(\boldsymbol{\gamma}_{\mathrm{proj}}-\hat{\boldsymbol{\gamma}}_{\mathrm{lasso}}\big)\right| = O_p\Big(\sqrt{n}\cdot\sqrt{\frac{\log p}{n}}\cdot s\cdot(\lambda_1+\lambda_0)\Big) = o_p(1), \quad (20)$$

where the last equality is because of Condition 5. Similarly,

$$\sqrt{n}\sum_{m=1}^{M}\pi_{[m]}\big(\bar{\boldsymbol{x}}_{[m]0}-\bar{\boldsymbol{x}}_{[m]}\big)^{\mathrm{T}}\big(\boldsymbol{\gamma}_{\mathrm{proj}}-\hat{\boldsymbol{\gamma}}_{\mathrm{lasso}}\big) \xrightarrow{p} 0. \quad (21)$$

Combining (19)–(21), we obtain the asymptotic normality of $\hat{\tau}_{\mathrm{lasso2}}$. Next, we compare the asymptotic variances $\sigma^2_{\mathrm{unadj}}$ and $\sigma^2_{\mathrm{lasso2}}$. By the decomposition in (18), we have

$$S^2_{[m]\varepsilon^*(1)} = S^2_{[m]Y(1)} + \boldsymbol{\gamma}_{\mathrm{proj}}^{\mathrm{T}} S^2_{[m]\boldsymbol{X}}\boldsymbol{\gamma}_{\mathrm{proj}} - 2S^{\mathrm{T}}_{[m]\boldsymbol{X}Y(1)}\boldsymbol{\gamma}_{\mathrm{proj}}. \quad (22)$$

Similarly, we have

$$S^2_{[m]\varepsilon^*(0)} = S^2_{[m]Y(0)} + \boldsymbol{\gamma}_{\mathrm{proj}}^{\mathrm{T}} S^2_{[m]\boldsymbol{X}}\boldsymbol{\gamma}_{\mathrm{proj}} - 2S^{\mathrm{T}}_{[m]\boldsymbol{X}Y(0)}\boldsymbol{\gamma}_{\mathrm{proj}}, \quad (23)$$

and

$$\begin{aligned}S^2_{[m]\{\varepsilon^*(1)-\varepsilon^*(0)\}} =& S^2_{[m]\{Y(1)-Y(0)\}} + (\boldsymbol{\gamma}_{\mathrm{proj}}-\boldsymbol{\gamma}_{\mathrm{proj}})^{\mathrm{T}} S^2_{[m]\boldsymbol{X}}(\boldsymbol{\gamma}_{\mathrm{proj}}-\boldsymbol{\gamma}_{\mathrm{proj}})\\ &- 2\{S_{[m]\boldsymbol{X}Y(1)} - S_{[m]\boldsymbol{X}Y(0)}\}^{\mathrm{T}}(\boldsymbol{\gamma}_{\mathrm{proj}}-\boldsymbol{\gamma}_{\mathrm{proj}})\\ =& S^2_{[m]\{Y(1)-Y(0)\}}.\end{aligned}$$

Thus,

$$\begin{aligned}\sigma^2_{\mathrm{unadj}} =& \lim_{n\to\infty}\sum_{m=1}^{M}\pi_{[m]}\Big[\frac{S^2_{[m]Y(1)}}{e_{[m]}} + \frac{S^2_{[m]Y(0)}}{1-e_{[m]}} - S^2_{[m]\{Y(1)-Y(0)\}}\Big]\\ =& \sigma^2_{\mathrm{lasso2}} - \lim_{n\to\infty}\Bigg[\boldsymbol{\gamma}_{\mathrm{proj}}^{\mathrm{T}}\Big\{\sum_{m=1}^{M}\pi_{[m]}\frac{S^2_{[m]\boldsymbol{X}}}{e_{[m]}(1-e_{[m]})}\Big\}\boldsymbol{\gamma}_{\mathrm{proj}}\\ &+ 2\sum_{m=1}^{M}\pi_{[m]}\Big\{\frac{S_{[m]\boldsymbol{X}Y(1)}}{e_{[m]}} + \frac{S_{[m]\boldsymbol{X}Y(0)}}{1-e_{[m]}}\Big\}^{\mathrm{T}}\boldsymbol{\gamma}_{\mathrm{proj}}\Bigg]\\ =& \sigma^2_{\mathrm{lasso2}} + \lim_{n\to\infty}\boldsymbol{\gamma}_{\mathrm{proj}}^{\mathrm{T}}\Big\{\sum_{m=1}^{M}\pi_{[m]}\frac{S^2_{[m]\boldsymbol{X}}}{e_{[m]}(1-e_{[m]})}\Big\}\boldsymbol{\gamma}_{\mathrm{proj}},\end{aligned}$$

where the last equality is because of the definition of $\boldsymbol{\gamma}_{\mathrm{proj}}$:

$$(\boldsymbol{\gamma}_{\mathrm{proj}})_S = \Big\{\sum_{m=1}^{M}\pi_{[m]}\frac{S^2_{[m]\boldsymbol{X}_S}}{e_{[m]}(1-e_{[m]})}\Big\}^{-1}\Big\{\sum_{m=1}^{M}\pi_{[m]}\frac{S_{[m]\boldsymbol{X}_SY(1)}}{e_{[m]}} + \sum_{m=1}^{M}\pi_{[m]}\frac{S_{[m]\boldsymbol{X}_SY(0)}}{1-e_{[m]}}\Big\},$$

and $(\boldsymbol{\gamma}_{\mathrm{proj}})_{S^c} = \mathbf{0}$.

B.4. Proof of Theorem 4

Since Conditions 1–2, 4–5, and 7 hold for $\varepsilon_i^{\Delta}(z)$ and the transformed covariates $\sqrt{\omega_{\boldsymbol{X},i}}\,\boldsymbol{x}_i$, we apply Lemma 9 to $\hat{\gamma}_{\text{lasso},z}$ and obtain

$$\left\|\hat{\gamma}_{\text{lasso},z}\right\|_0 \leqslant Cs, \quad z = 0, 1,$$

which leads to

$$\hat{s} = \left\|\hat{\gamma}_{\text{lasso},1} + \hat{\gamma}_{\text{lasso},0}\right\|_0 = O_p(s).$$

Thus, it suffices for proving the result about the probability limit of $\hat{\sigma}^2_{\text{lasso2}}$ to show that

$$\sum_{m=1}^{M} \pi_{[m]} \frac{\hat{s}^2_{[m]\varepsilon^*(1)}}{e_{[m]}} - \sum_{m=1}^{M} \pi_{[m]} \frac{S^2_{[m]\varepsilon^*(1)}}{e_{[m]}} \xrightarrow{p} 0, \quad \sum_{m=1}^{M} \pi_{[m]} \frac{\hat{s}^2_{[m]\varepsilon^*(0)}}{1-e_{[m]}} - \sum_{m=1}^{M} \pi_{[m]} \frac{S^2_{[m]\varepsilon^*(0)}}{1-e_{[m]}} \xrightarrow{p} 0.$$

Again, we will only prove the first statement. By definition and the decomposition (18),

$$\begin{aligned}\hat{s}^2_{\varepsilon^*(1)} &= \frac{1}{n_{[m]1}-1} \sum_{i\in[m]} Z_i\left\{Y_i(1) - \bar{Y}_{[m]1} - (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_{[m]1})^{\mathrm{T}}\hat{\gamma}_{\text{lasso}}\right\}^2 \\ &= s^2_{[m]\varepsilon^*(1)} + \left(\gamma_{\text{proj}} - \hat{\gamma}_{\text{lasso}}\right)^{\mathrm{T}} s^2_{[m]\boldsymbol{X}}\left(\gamma_{\text{proj}} - \hat{\gamma}_{\text{lasso}}\right) + 2 s^{\mathrm{T}}_{[m]\boldsymbol{X}\varepsilon^*(1)}\left(\gamma_{\text{proj}} - \hat{\gamma}_{\text{lasso}}\right).\end{aligned}$$

Therefore,

$$\begin{aligned}\sum_{m=1}^{M} \pi_{[m]} \frac{\hat{s}^2_{\varepsilon^*(1)}}{e_{[m]}} = \sum_{m=1}^{M} \pi_{[m]} \frac{s^2_{[m]\varepsilon^*(1)}}{e_{[m]}} &+ \left(\gamma_{\text{proj}} - \hat{\gamma}_{\text{lasso}}\right)^{\mathrm{T}} \left(\sum_{m=1}^{M} \pi_{[m]} \frac{s^2_{[m]\boldsymbol{X}}}{e_{[m]}}\right)\left(\gamma_{\text{proj}} - \hat{\gamma}_{\text{lasso}}\right) \\ &+ 2\sum_{m=1}^{M} \pi_{[m]} \frac{s^{\mathrm{T}}_{[m]\boldsymbol{X}\varepsilon^*(1)}}{e_{[m]}}\left(\gamma_{\text{proj}} - \hat{\gamma}_{\text{lasso}}\right).\end{aligned}$$

In the proof of Theorem 3, we have shown that $\|\gamma_{\text{proj}} - \hat{\gamma}_{\text{lasso}}\|_1 = O_p(s\lambda_1 + s\lambda_0)$. Using similar arguments as the proof for the limit of $\hat{\sigma}^2_{\text{lasso}}$ (Theorem 2), we have

$$\sum_{m=1}^{M} \pi_{[m]} \frac{\hat{s}^2_{[m]\varepsilon^*(1)}}{e_{[m]}} - \sum_{m=1}^{M} \pi_{[m]} \frac{S^2_{[m]\varepsilon^*(1)}}{e_{[m]}} \xrightarrow{p} 0.$$

Next, we compare the limits of $\hat{\sigma}^2_{\text{lasso2}}$ and σ^2_{lasso2}. By definition and the above proof, it is easy to see that

$$\lim_{n\to\infty}(\hat{\sigma}^2_{\text{lasso2}} - \sigma^2_{\text{lasso2}}) = \lim_{n\to\infty} \sum_{m=1}^{M} \pi_{[m]} S^2_{[m]\{\varepsilon^*(1)-\varepsilon^*(0)\}} \geqslant 0.$$

Finally, we compare the limits of $\hat{\sigma}^2_{\text{lasso2}}$ and $\hat{\sigma}^2_{\text{unadj}}$. According to Proposition 3, we have

$$\lim_{n\to\infty}(\hat{\sigma}^2_{\text{lasso2}} - \hat{\sigma}^2_{\text{unadj}}) = \lim_{n\to\infty}\left\{\sum_{m=1}^{M} \pi_{[m]} \frac{S^2_{[m]\varepsilon^*(1)} - S^2_{[m]Y(1)}}{e_{[m]}} + \sum_{m=1}^{M} \pi_{[m]} \frac{S^2_{[m]\varepsilon^*(0)} - S^2_{[m]Y(0)}}{1-e_{[m]}}\right\}.$$

We have shown in (22) and (23) that

$$S^2_{[m]\varepsilon^*(z)} = S^2_{[m]Y(z)} + \boldsymbol{\gamma}^{\mathrm{T}}_{\mathrm{proj}} S^2_{[m]\boldsymbol{X}} \boldsymbol{\gamma}_{\mathrm{proj}} - 2S^{\mathrm{T}}_{[m]\boldsymbol{X}Y(z)} \boldsymbol{\gamma}_{\mathrm{proj}}, \quad z = 0, 1.$$

Therefore,

$$\begin{aligned}
&\lim_{n\to\infty} (\hat{\sigma}^2_{\mathrm{lasso2}} - \hat{\sigma}^2_{\mathrm{unadj}}) \\
&= \lim_{n\to\infty} \left[\boldsymbol{\gamma}^{\mathrm{T}}_{\mathrm{proj}} \sum_{m=1}^{M} \pi_{[m]} \cdot \frac{S^2_{[m]\boldsymbol{X}}}{e_{[m]}(1-e_{[m]})} \boldsymbol{\gamma}_{\mathrm{proj}} - 2 \sum_{m=1}^{M} \pi_{[m]} \left\{ \frac{S_{[m]\boldsymbol{X}Y(1)}}{e_{[m]}} + \frac{S_{[m]\boldsymbol{X}Y(0)}}{1-e_{[m]}} \right\}^{\mathrm{T}} \boldsymbol{\gamma}_{\mathrm{proj}} \right] \\
&= - \lim_{n\to\infty} \boldsymbol{\gamma}^{\mathrm{T}}_{\mathrm{proj}} \left\{ \sum_{m=1}^{M} \pi_{[m]} \frac{S^2_{[m]\boldsymbol{X}}}{e_{[m]}(1-e_{[m]})} \right\} \boldsymbol{\gamma}_{\mathrm{proj}} \leqslant 0,
\end{aligned}$$

where the second equality is because of the definition of $\boldsymbol{\gamma}_{\mathrm{proj}}$.

B.5. Proof of Theorem 5

First, we prove the result on the asymptotic distribution of $\hat{\tau}_{\mathrm{lasso}}$ under stratified rerandomization. In the proof of Theorem 1, we have shown that

$$\begin{aligned}
\hat{\tau}_{\mathrm{lasso}} - \tau = &\sum_{m=1}^{M} \pi_{[m]} \left(\bar{\varepsilon}_{[m]1} - \bar{\varepsilon}_{[m]0} \right) + \sum_{m=1}^{M} \pi_{[m]} \left(\bar{\boldsymbol{x}}_{[m]1} - \bar{\boldsymbol{x}}_{[m]} \right)^{\mathrm{T}} \left\{ \boldsymbol{\beta}(1) - \hat{\boldsymbol{\beta}}_{\mathrm{lasso},1} \right\} \\
&- \sum_{m=1}^{M} \pi_{[m]} \left(\bar{\boldsymbol{x}}_{[m]0} - \bar{\boldsymbol{x}}_{[m]} \right)^{\mathrm{T}} \left\{ \boldsymbol{\beta}(0) - \hat{\boldsymbol{\beta}}_{\mathrm{lasso},0} \right\}.
\end{aligned}$$

Applying Proposition 4 to $\varepsilon_i(z)$, we have

$$\sqrt{n} \sum_{m=1}^{M} \pi_{[m]} \left(\bar{\varepsilon}_{[m]1} - \bar{\varepsilon}_{[m]0} \right) \mid \mathcal{M}_a \xrightarrow{d} \sigma_{\mathrm{lasso}} \left\{ \sqrt{1 - R^2_{\varepsilon,\boldsymbol{W}}}\, \varepsilon_0 + \sqrt{R^2_{\varepsilon,\boldsymbol{W}}}\, L_{k,a} \right\},$$

where

$$R^2_{\varepsilon,\boldsymbol{W}} = \lim_{n\to\infty} \mathrm{cov}(\hat{\tau}_\varepsilon, \hat{\tau}_{\boldsymbol{w}}) \mathrm{cov}(\hat{\tau}_{\boldsymbol{w}})^{-1} \mathrm{cov}(\hat{\tau}_{\boldsymbol{w}}, \hat{\tau}_\varepsilon) / \mathrm{var}(\hat{\tau}_\varepsilon),$$

$$\hat{\tau}_\varepsilon = \sum_{m=1}^{M} \pi_{[m]} \left(\bar{\varepsilon}_{[m]1} - \bar{\varepsilon}_{[m]0} \right).$$

It suffices for deriving the asymptotic distribution of $\hat{\tau}_{\mathrm{lasso}}$ to show that

$$\sqrt{n} \sum_{m=1}^{M} \pi_{[m]} \left(\bar{\boldsymbol{x}}_{[m]1} - \bar{\boldsymbol{x}}_{[m]} \right)^{\mathrm{T}} \left\{ \boldsymbol{\beta}_{\mathrm{lasso}}(1) - \hat{\boldsymbol{\beta}}_{\mathrm{lasso},1} \right\} \mid \mathcal{M}_a \xrightarrow{p} 0, \tag{24}$$

$$\sqrt{n} \sum_{m=1}^{M} \pi_{[m]} \left(\bar{\boldsymbol{x}}_{[m]0} - \bar{\boldsymbol{x}}_{[m]} \right)^{\mathrm{T}} \left\{ \boldsymbol{\beta}_{\mathrm{lasso}}(0) - \hat{\boldsymbol{\beta}}_{\mathrm{lasso},0} \right\} \mid \mathcal{M}_a \xrightarrow{p} 0. \tag{25}$$

We will only prove (24), as the proof of (25) is similar. According to Proposition 4,

$$P(\mathcal{M}_a) \longrightarrow P(\chi_k^2 < a) > 0.$$

Thus, it suffices for (24) to show that, under stratified randomization,

$$\sqrt{n}\sum_{m=1}^{M} \pi_{[m]}\left(\bar{\boldsymbol{x}}_{[m]1} - \bar{\boldsymbol{x}}_{[m]}\right)^{\mathrm{T}}\left\{\boldsymbol{\beta}_{\text{lasso}}(1) - \hat{\boldsymbol{\beta}}_{\text{lasso},1}\right\} \xrightarrow{p} 0,$$

which holds as shown in the proof of Theorem 1.

Second, we compare the asymptotic variances. According to Proposition 4, $\sigma^2_{\text{unadj}|\mathcal{M}} \leqslant \sigma^2_{\text{unadj}}$. We only need to prove that $R^2_{\varepsilon,\boldsymbol{W}} = 0$ and $\sigma^2_{\text{lasso}} \leqslant \sigma^2_{\text{unadj}|\mathcal{M}}$ if Condition 6 holds. By definition,

$$R^2_{\varepsilon,\boldsymbol{W}} = \lim_{n\to\infty}\left(V_{\varepsilon W}V_{WW}^{-1}V_{W\varepsilon}\right)/V_{\varepsilon\varepsilon}.$$

It suffices for $R^2_{\varepsilon,\boldsymbol{W}} = 0$ to show that, element-wise,

$$V_{\boldsymbol{X}\varepsilon} = \sum_{m=1}^{M}\pi_{[m]}\left\{\frac{S_{[m]\boldsymbol{W}\varepsilon(1)}}{e_{[m]}} + \frac{S_{[m]\boldsymbol{W}\varepsilon(0)}}{1-e_{[m]}}\right\} \to \mathbf{0}.$$

By definition of the projection error terms, $\varepsilon_i(z)$, we have

$$\sum_{m=1}^{M}\pi_{[m]}S_{[m]\boldsymbol{X}_S\varepsilon(1)} = \mathbf{0}.$$

Then we have $\sum_{m=1}^{M}\pi_{[m]}S_{[m]\boldsymbol{W}\varepsilon(1)} = \mathbf{0}$ since $\{1,2,\ldots,k\}\subset S$. Therefore,

$$\begin{aligned}\left\|\sum_{m=1}^{M}\pi_{[m]}\frac{S_{[m]\boldsymbol{W}\varepsilon(1)}}{e_{[m]}}\right\|_\infty &= \left\|\sum_{m=1}^{M}\pi_{[m]}\frac{S_{[m]\boldsymbol{W}\varepsilon(1)}}{e_{[m]}} - \sum_{m=1}^{M}\pi_{[m]}\frac{S_{[m]\boldsymbol{W}\varepsilon(1)}}{e^*}\right\|_\infty \\ &\leqslant \max_{m=1,\ldots,M}\left|\frac{1}{e_{[m]}} - \frac{1}{e^*}\right|\sum_{m=1}^{M}\pi_{[m]}\left\|S_{[m]\boldsymbol{W}\varepsilon(1)}\right\|_\infty \to 0,\end{aligned}$$

where the last convergence is because of Conditions 2 and 6. Similarly, element-wise,

$$\sum_{m=1}^{M}\pi_{[m]}\frac{S_{[m]\boldsymbol{W}\varepsilon(0)}}{1-e_{[m]}} \to \mathbf{0}.$$

Therefore, $R^2_{\varepsilon,\boldsymbol{W}} = 0$. Now we can compare σ^2_{lasso} and $\sigma^2_{\text{unadj}|\mathcal{M}}$. As $R^2_{\varepsilon,\boldsymbol{W}} = 0$, we have the asymptotic variance of $\sqrt{n}(\hat{\tau}_{\text{lasso}} - \tau) \mid \mathcal{M}_a$ is σ^2_{lasso}. According to Proposition 4, we have

$$\begin{aligned}\sigma^2_{\text{unadj}|\mathcal{M}} &= \sigma^2_{\text{unadj}}\left[1 - \left\{1 - v_{k,a}\right\}R^2_{Y,\boldsymbol{W}}\right] \\ &\geqslant \sigma^2_{\text{unadj}}(1 - R^2_{Y,\boldsymbol{W}}) \\ &= \lim_{n\to\infty}\left[\operatorname{var}\left(\hat{\tau}_{\text{unadj}}\right) - \operatorname{var}\left\{\operatorname{proj}\left(\hat{\tau}_{\text{unadj}} \mid \hat{\tau}_{\boldsymbol{w}}\right)\right\}\right] \\ &= \lim_{n\to\infty}\operatorname{var}\left\{\hat{\tau}_{\text{unadj}} - \operatorname{proj}\left(\hat{\tau}_{\text{unadj}} \mid \hat{\tau}_{\boldsymbol{w}}\right)\right\},\end{aligned}$$

where $\mathrm{proj}(\hat{\tau}_{\mathrm{unadj}} \mid \hat{\tau}_{\boldsymbol{w}})$ denote the projection (minimizing the variance) of $\hat{\tau}_{\mathrm{unadj}}$ onto $\hat{\tau}_{\boldsymbol{w}}$, and the last but one equality holds because (see Li and Ding (2020)):

$$R^2_{Y,\boldsymbol{W}} = \lim_{n\to\infty} \frac{\mathrm{var}\{\mathrm{proj}(\hat{\tau}_{\mathrm{unadj}} \mid \hat{\tau}_{\boldsymbol{w}})\}}{\mathrm{var}(\hat{\tau}_{\mathrm{unadj}})} = \lim_{n\to\infty} \frac{\mathrm{var}\{\mathrm{proj}(\hat{\tau}_{\mathrm{unadj}} \mid \hat{\tau}_{\boldsymbol{w}})\}}{\sigma^2_{\mathrm{unadj}}}.$$

By definition and simple algebra, we have

$$\sigma^2_{\mathrm{lasso}} = \lim_{n\to\infty} \mathrm{var}(\hat{\tau}_{\varepsilon}) = \lim_{n\to\infty} \mathrm{var}\Big\{\hat{\tau}_{\mathrm{unadj}} - \mathrm{proj}(\hat{\tau}_{\mathrm{unadj}} \mid \hat{\tau}_{\boldsymbol{x}_S})\Big\}.$$

Since $\{1,\dots,k\} \subset S$, we have $\sigma^2_{\mathrm{lasso}} \leqslant \sigma^2_{\mathrm{unadj}|\mathcal{M}}$.

B.6. Proof of Theorem 6

The proof for the probability limit of $\hat{\sigma}_{\mathrm{lasso}}$ can be obtained by conditional on $\mathcal{M}_a$, and following the proof of Theorem 2, so we omit it.

Next, we compare the estimated variances $\hat{\sigma}^2_{\mathrm{unadj}|\mathcal{M}}$ and $\hat{\sigma}^2_{\mathrm{unadj}}$. According to Proposition 3,

$$\hat{\sigma}^2_{\mathrm{unadj}} \xrightarrow{p} \sigma^2_{\mathrm{unadj}} + \lim_{n\to\infty} \sum_{m=1}^{M} \pi_{[m]} S^2_{[m]\{Y(1)-Y(0)\}}.$$

According to Proposition 4,

$$\begin{aligned}
\hat{\sigma}^2_{\mathrm{unadj}|\mathcal{M}} \quad &\xrightarrow{p} \quad \sigma^2_{\mathrm{unadj}|\mathcal{M}} + \lim_{n\to\infty} \sum_{m=1}^{M} \pi_{[m]} S^2_{[m]\{Y(1)-Y(0)\}} \\
&= \quad \sigma^2_{\mathrm{unadj}}\big[1 - \{1 - v_{k,a}\} R^2_{Y,\boldsymbol{W}}\big] + \lim_{n\to\infty} \sum_{m=1}^{M} \pi_{[m]} S^2_{[m]\{Y(1)-Y(0)\}} \\
&= \quad \sigma^2_{\mathrm{unadj}} + \lim_{n\to\infty} \sum_{m=1}^{M} \pi_{[m]} S^2_{[m]\{Y(1)-Y(0)\}} - \big\{1 - v_{k,a}\big\} \sigma^2_{\mathrm{unadj}} R^2_{Y,\boldsymbol{W}}.
\end{aligned}$$

Therefore, with probability tending to one, $\hat{\sigma}^2_{\mathrm{unadj}|\mathcal{M}} \leqslant \hat{\sigma}^2_{\mathrm{unadj}}$.

Finally, we compare $\hat{\sigma}^2_{\mathrm{lasso}}$ and $\hat{\sigma}^2_{\mathrm{unadj}|\mathcal{M}}$. We have

$$\begin{aligned}
\lim_{n\to\infty} \hat{\sigma}^2_{\mathrm{unadj}|\mathcal{M}} &= \sigma^2_{\mathrm{unadj}|\mathcal{M}} + \lim_{n\to\infty} \sum_{m=1}^{M} \pi_{[m]} S^2_{[m]\{Y(1)-Y(0)\}} \\
&\geqslant \sigma^2_{\mathrm{lasso}} + \lim_{n\to\infty} \sum_{m=1}^{M} \pi_{[m]} S^2_{[m]\{Y(1)-Y(0)\}} \\
&\geqslant \sigma^2_{\mathrm{lasso}} + \lim_{n\to\infty} \sum_{m=1}^{M} \pi_{[m]} S^2_{[m]\{\varepsilon(1)-\varepsilon(0)\}} \\
&= \lim_{n\to\infty} \hat{\sigma}^2_{\mathrm{lasso}},
\end{aligned}$$

where the second inequality is because of the projection of $Y_i(z)$,

$$Y_i(z) = \bar{Y}_{[m]}(z) + \big(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_{[m]}\big)^{\mathrm{T}} \boldsymbol{\beta}_{\text{lasso}}(z) + \varepsilon_i(z), \quad i \in [m], \quad z = 0, 1.$$

Thus, $\hat{\sigma}^2_{\text{lasso}} \leqslant \hat{\sigma}^2_{\text{unadj}|\mathcal{M}}$ holds with probability tending to one.

B.7. Proof of Theorem 7

First, we prove the result on the asymptotic distribution of $\hat{\tau}_{\text{lasso2}}$ under stratified rerandomization. In the proof of Theorem 3, we have shown that

$$\begin{aligned} \hat{\tau}_{\text{lasso2}} - \tau = &\sum_{m=1}^{M} \pi_{[m]} \big(\bar{\varepsilon}^*_{[m]1} - \bar{\varepsilon}^*_{[m]0}\big) + \sum_{m=1}^{M} \pi_{[m]} \big(\bar{\boldsymbol{x}}_{[m]1} - \bar{\boldsymbol{x}}_{[m]}\big)^{\mathrm{T}} \big(\boldsymbol{\gamma}_{\text{proj}} - \hat{\boldsymbol{\gamma}}_{\text{lasso}}\big) \\ &- \sum_{m=1}^{M} \pi_{[m]} \big(\bar{\boldsymbol{x}}_{[m]0} - \bar{\boldsymbol{x}}_{[m]}\big)^{\mathrm{T}} \big(\boldsymbol{\gamma}_{\text{proj}} - \hat{\boldsymbol{\gamma}}_{\text{lasso}}\big). \end{aligned}$$

Applying Proposition 4 to $\varepsilon^*_i(z)$, we have

$$\Big\{\sqrt{n} \sum_{m=1}^{M} \pi_{[m]} \big(\varepsilon^*_{[m]1} - \varepsilon^*_{[m]0}\big) \mid \mathcal{M}_a\Big\} \xrightarrow{d} \sigma_{\text{lasso2}} \Big(\sqrt{1 - R^2_{\varepsilon^*, \boldsymbol{W}}}\, \varepsilon_0 + \sqrt{R^2_{\varepsilon^*, \boldsymbol{W}}}\, L_{k,a}\Big),$$

where

$$R^2_{\varepsilon^*, \boldsymbol{W}} = \lim_{n\to\infty} \Big(V_{\varepsilon^* \boldsymbol{W}} V^{-1}_{\boldsymbol{W}\boldsymbol{W}} V_{\boldsymbol{W}\varepsilon^*}\Big) / V_{\varepsilon^*\varepsilon^*},$$

$$V_{\boldsymbol{W}\varepsilon^*} = \sum_{m=1}^{M} \pi_{[m]} \cdot \Big(\frac{S_{[m]\boldsymbol{W}\varepsilon^*(1)}}{e_{[m]}} + \frac{S_{[m]\boldsymbol{W}\varepsilon^*(0)}}{1 - e_{[m]}}\Big).$$

Recall the definition of $\boldsymbol{\gamma}_{\text{proj}}$:

$$(\boldsymbol{\gamma}_{\text{proj}})_S = \Bigg\{\sum_{m=1}^{M} \pi_{[m]} \frac{S^2_{[m]\boldsymbol{X}_S}}{e_{[m]}(1 - e_{[m]})}\Bigg\}^{-1} \Bigg\{\sum_{m=1}^{M} \pi_{[m]} \frac{S_{[m]\boldsymbol{X}_S Y(1)}}{e_{[m]}} + \sum_{m=1}^{M} \pi_{[m]} \frac{S_{[m]\boldsymbol{X}_S Y(0)}}{1 - e_{[m]}}\Bigg\},$$

$(\boldsymbol{\gamma}_{\text{proj}})_{S^c} = \mathbf{0}$, and the decomposition $Y_i(z) = \bar{Y}_{[m]}(z) + (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_{[m]})^{\mathrm{T}} \boldsymbol{\gamma}_{\text{proj}} + \varepsilon^*_i(z)$, we have

$$\sum_{m=1}^{M} \pi_{[m]} \Big\{\frac{S_{[m]\boldsymbol{X}_S\varepsilon^*(1)}}{e_{[m]}} + \frac{S_{[m]\boldsymbol{X}_S\varepsilon^*(0)}}{1 - e_{[m]}}\Big\} = \mathbf{0}.$$

Since $\{1, \ldots, k\} \subset S$, we have

$$\sum_{m=1}^{M} \pi_{[m]} \Big\{\frac{S_{[m]\boldsymbol{W}\varepsilon^*(1)}}{e_{[m]}} + \frac{S_{[m]\boldsymbol{W}\varepsilon^*(0)}}{1 - e_{[m]}}\Big\} = \mathbf{0}.$$

Therefore, $V_{\boldsymbol{W}\varepsilon^*} = 0$. Then, $R^2_{\varepsilon^*, \boldsymbol{W}} = 0$ and

$$\sqrt{n} \sum_{m=1}^{M} \pi_{[m]} \big(\bar{\varepsilon}^*_{[m]1} - \bar{\varepsilon}^*_{[m]0}\big) \mid \mathcal{M}_a \xrightarrow{d} N(0, \sigma^2_{\text{lasso2}}).$$

It suffices for the asymptotic normality of $\hat{\tau}_{\text{lasso2}}$ to show that,

$$\sqrt{n}\sum_{m=1}^{M}\pi_{[m]}\big(\bar{\boldsymbol{x}}_{[m]z}-\bar{\boldsymbol{x}}_{[m]}\big)^{\mathrm{T}}\big(\boldsymbol{\gamma}_{\text{proj}}-\hat{\boldsymbol{\gamma}}_{\text{lasso}}\big)\mid\mathcal{M}_a\xrightarrow{p}0,\quad z=0,1. \tag{26}$$

According to Proposition 4, $P(\mathcal{M}_a)\longrightarrow P(\chi_k^2<a)>0$. Thus, it suffices for (26) to show that, under stratified randomization,

$$\sum_{m=1}^{M}\pi_{[m]}\big(\bar{\boldsymbol{x}}_{[m]z}-\bar{\boldsymbol{x}}_{[m]}\big)^{\mathrm{T}}\big(\boldsymbol{\gamma}_{\text{proj}}-\hat{\boldsymbol{\gamma}}_{\text{lasso}}\big)\xrightarrow{p}0,\quad z=0,1,$$

which holds as shown in the proof of Theorem 3.

Next, we compare the asymptotic variances. According to Proposition 4, $\sigma^2_{\text{unadj}|\mathcal{M}}\leqslant\sigma^2_{\text{unadj}}$. Thus, it suffices to show that $\sigma^2_{\text{lasso2}}\leqslant\sigma^2_{\text{unadj}|\mathcal{M}}$. We have shown in the proof of Theorem 5 that

$$\begin{aligned}\sigma^2_{\text{unadj}|\mathcal{M}}&=\sigma^2_{\text{unadj}}\big[1-(1-v_{k,a})R^2_{Y,\boldsymbol{W}}\big]\\&\geqslant\sigma^2_{\text{unadj}}(1-R^2_{Y,\boldsymbol{W}})\\&=\lim_{n\to\infty}\text{var}\big\{\hat{\tau}_{\text{unadj}}-\text{proj}\big(\hat{\tau}_{\text{unadj}}\mid\hat{\tau}_{\boldsymbol{w}}\big)\big\}.\end{aligned}$$

By definition and the above proof for the asymptotic normality, we have

$$\begin{aligned}\sigma^2_{\text{lasso2}}&=\lim_{n\to\infty}\sum_{m=1}^{M}\pi_{[m]}\left\{\frac{S^2_{[m]\varepsilon^*(1)}}{e_{[m]}}+\frac{S^2_{[m]\varepsilon^*(0)}}{1-e_{[m]}}-S^2_{[m]\{\varepsilon^*(1)-\varepsilon^*(0)\}}\right\}\\&=\lim_{n\to\infty}E\Big\{\hat{\tau}_{\text{unadj}}-\tau-\hat{\tau}_{\boldsymbol{x}}^{\mathrm{T}}\boldsymbol{\gamma}_{\text{proj}}\Big\}^2\\&=\lim_{n\to\infty}\text{var}\Big\{\hat{\tau}_{\text{unadj}}-\text{proj}\Big(\hat{\tau}_{\text{unadj}}\mid\hat{\tau}_{\boldsymbol{x}_S}\Big)\Big\}\\&\leqslant\lim_{n\to\infty}\text{var}\big\{\hat{\tau}_{\text{unadj}}-\text{proj}\big(\hat{\tau}_{\text{unadj}}\mid\hat{\tau}_{\boldsymbol{w}}\big)\big\}\\&=\sigma^2_{\text{unadj}|\mathcal{M}},\end{aligned}$$

where the inequality is due to $\{1,\dots,k\}\subset S$.

B.8. *Proof of Theorem 8*

Similar to the proof of Theorem 4, we can show that, under stratified randomization,

$$\begin{aligned}\hat{\sigma}^2_{\text{lasso2}}&\xrightarrow{p}\lim_{n\to\infty}\sum_{m=1}^{M}\pi_{[m]}\left\{\frac{S^2_{[m]\varepsilon^*(1)}}{e_{[m]}}+\frac{S^2_{[m]\varepsilon^*(0)}}{1-e_{[m]}}\right\}\\&=\sigma^2_{\text{lasso2}}+\lim_{n\to\infty}\sum_{m=1}^{M}\pi_{[m]}\cdot S^2_{[m]\{\varepsilon^*(1)-\varepsilon^*(0)\}}.\end{aligned} \tag{27}$$

Since $P(\mathcal{M}_a)\to P(\chi_k^2\leqslant a)>0$, then (27) also holds under stratified rerandomization, i.e., conditional on $\mathcal{M}_a$. We have shown in Theorem 6 that $\hat{\sigma}^2_{\text{unadj}|\mathcal{M}}\leqslant\hat{\sigma}^2_{\text{unadj}}$

holds in probability under stratified rerandomization. Thus, we only need to show that, $\hat{\sigma}^2_{\text{lasso2}} \leqslant \hat{\sigma}^2_{\text{unadj}|\mathcal{M}}$ holds in probability under stratified rerandomization. Under stratified rerandomization, the probability limit of $\hat{\sigma}^2_{\text{unadj}|\mathcal{M}}$ satisfies:

$$\begin{aligned}\lim_{n\to\infty} \hat{\sigma}^2_{\text{unadj}|\mathcal{M}} &= \sigma^2_{\text{unadj}|\mathcal{M}} + \lim_{n\to\infty} \sum_{m=1}^{M} \pi_{[m]} S^2_{[m]\{Y(1)-Y(0)\}} \\ &\geqslant \sigma^2_{\text{lasso2}} + \lim_{n\to\infty} \sum_{m=1}^{M} \pi_{[m]} S^2_{[m]\{Y(1)-Y(0)\}}.\end{aligned}$$

In the proof of Theorem 3, we have shown that $S^2_{[m]\{\varepsilon^*(1)-\varepsilon^*(0)\}} = S^2_{[m]\{Y(1)-Y(0)\}}$. Therefore, we obtain $\hat{\sigma}^2_{\text{lasso2}} \leqslant \hat{\sigma}^2_{\text{unadj}|\mathcal{M}}$ holds in probability under stratified rerandomization.

C. Proof of Lemmas

C.1. Proof of Lemma 3

PROOF. By definition and simple calculus, we have

$$\|\nabla \bar{a}_1(\boldsymbol{z})\|_2^2 = \sum_{i:z_i=1}\sum_{j:z_j=0} \left|\bar{a}_1(\boldsymbol{z}) - \bar{a}_1(\boldsymbol{z}^{i,j})\right|^2 = \frac{1}{n_1^2}\sum_{i:z_i=1}\sum_{j:z_j=0} |a_i - a_j|^2. \tag{28}$$

Then, by Lemma 2, we have

$$\frac{1}{n_1^2}\sum_{i:z_i=1}\sum_{j:z_j=0} |a_i - a_j|^2 \leqslant \frac{1}{n_1^2}\sum_{1\leqslant i<j\leqslant n} |a_i - a_j|^2 = \frac{n}{n_1^2}\sum_{i=1}^{n}(a_i - \bar{a})^2 =: \sigma^2_{\text{mean}}. \tag{29}$$

Combining (28) and (29), Lemma 3 follows from Lemma 1.

C.2. Proof of Lemma 4

PROOF. We start by examining $s_{ab}(\boldsymbol{z}) - s_{ab}(\boldsymbol{z}^{i,j})$. By Lemma 2 and some simple calculus, we have

$$\begin{aligned}& s_{ab}(\boldsymbol{z}) - s_{ab}(\boldsymbol{z}^{i,j}) \\ =& \frac{1}{n_1(n_1-1)} \sum_{1\leqslant i'<j'\leqslant n} \{z_{i'}z_{j'}(a_{i'}-a_{j'})(b_{i'}-b_{j'}) - z^{i,j}_{i'}z^{i,j}_{j'}(a_{i'}-a_{j'})(b_{i'}-b_{j'})\} \\ =& \frac{1}{n_1(n_1-1)} \sum_{l\neq i} z_l\{(a_l-a_i)(b_l-b_i) - (a_l-a_j)(b_l-b_j)\} \\ =& \frac{1}{n_1(n_1-1)} \sum_{l\neq i} z_l\{a_ib_i - a_jb_j + (a_j-a_i)b_l + a_l(b_j-b_i)\} \\ =& \frac{1}{n_1(n_1-1)} \sum_{l\neq i} z_l\{(a_ib_i - a_i\bar{b} - \bar{a}b_i + \bar{a}\bar{b}) - (a_jb_j - a_j\bar{b} - \bar{a}b_j + \bar{a}\bar{b}) \\ & + (a_j-a_i)(b_l-\bar{b}) + (a_l-\bar{a})(b_j-b_i)\} \\ =& \frac{1}{n_1}(U_{ij}+V_{ij}),\end{aligned}$$

where

$$U_{ij} := (a_i - \bar{a})(b_i - \bar{b}) - (a_j - \bar{a})(b_j - \bar{b}),$$

$$V_{ij} := \frac{(a_j - a_i)}{n_1 - 1} \sum_{l \neq i} z_l (b_l - \bar{b}) + \frac{(b_j - b_i)}{n_1 - 1} \sum_{l \neq i} z_l (a_l - \bar{a}).$$

By Cauchy–Schwarz inequality, we have

$$\begin{aligned} V_{ij} \leqslant & |a_j - a_i| \sqrt{\frac{1}{n_1 - 1} \sum_{l=1}^{n} (b_l - \bar{b})^2} + |b_j - b_i| \sqrt{\frac{1}{n_1 - 1} \sum_{l=1}^{n} (a_l - \bar{a})^2} \\ \leqslant & |a_j - a_i| \sqrt{\frac{2}{e_c n} \sum_{l=1}^{n} (b_l - \bar{b})^2} + |b_j - b_i| \sqrt{\frac{2}{e_c n} \sum_{l=1}^{n} (a_l - \bar{a})^2}. \end{aligned} \tag{30}$$

Then we can bound $\|\nabla s_{ab}(\boldsymbol{z})\|_2^2$. By definition and Minkowski's inequality, we have

$$\begin{aligned} \|\nabla s_{ab}(\boldsymbol{z})\|_2^2 &= \sum_{i:z_i=1} \sum_{j:z_j=0} \left| s_{ab}(\boldsymbol{z}) - s_{ab}(\boldsymbol{z}^{i,j}) \right|^2 \\ &= \sum_{i:z_i=1} \sum_{j:z_j=0} |U_{ij}/n_1 + V_{ij}/n_1|^2 \\ &\leqslant \sum_{1 \leqslant i < j \leqslant n} |U_{ij}/n_1 + V_{ij}/n_1|^2 \\ &\leqslant (\sqrt{U} + \sqrt{V})^2, \end{aligned} \tag{31}$$

where

$$U := \frac{1}{n_1^2} \sum_{1 \leqslant i < j \leqslant n} U_{ij}^2, \quad V := \frac{1}{n_1^2} \sum_{1 \leqslant i < j \leqslant n} V_{ij}^2.$$

We bound U and V separately. By Lemma 2, we have

$$\begin{aligned} U =& \frac{1}{n_1^2} \sum_{1 \leqslant i < j \leqslant n} \{(a_i - \bar{a})(b_i - \bar{b}) - (a_j - \bar{a})(b_j - \bar{b})\}^2 \\ =& \frac{n}{n_1^2} \sum_{i=1}^{n} \left\{ (a_i - \bar{a})(b_i - \bar{b}) - \frac{1}{n} \sum_{j=1}^{n} (a_j - \bar{a})(b_j - \bar{b}) \right\}^2 \\ \leqslant & \frac{n}{n_1^2} \sum_{i=1}^{n} \{(a_i - \bar{a})(b_i - \bar{b})\}^2 \\ =& \frac{1}{e_c^2 n} \sum_{i=1}^{n} (a_i - \bar{a})^2 (b_i - \bar{b})^2. \end{aligned} \tag{32}$$

By (30), Minkowski's inequality, and Lemma 2, we have

$$
\begin{aligned}
V \leqslant & \frac{1}{n_1^2} \sum_{1\leqslant i<j\leqslant n} \left(|a_j - a_i| \sqrt{\frac{2}{e_c n}\sum_{i=1}^n (b_i - \bar{b})^2} + |b_j - b_i| \sqrt{\frac{2}{e_c n}\sum_{i=1}^n (a_i - \bar{a})^2} \right)^2 \\
\leqslant & \frac{1}{n_1^2} \left(\sqrt{\frac{2}{e_c n}\sum_{i=1}^n (b_i - \bar{b})^2 \sum_{1\leqslant i<j\leqslant n} (a_i - a_j)^2} + \sqrt{\frac{2}{e_c n}\sum_{i=1}^n (a_i - \bar{a})^2 \sum_{1\leqslant i<j\leqslant n} (b_i - b_j)^2} \right)^2 \\
= & \frac{8}{e_c^3 n^2} \sum_{i=1}^n (a_i - \bar{a})^2 \sum_{i=1}^n (b_i - \bar{b})^2. \qquad (33)
\end{aligned}
$$

Combining (31), (32), and (33), we have

$$
\|\nabla s_{ab}(\boldsymbol{z})\|_2^2 \leqslant \left\{ \sqrt{\frac{1}{e_c^2 n}\sum_{i=1}^n (a_i - \bar{a})^2 (b_i - \bar{b})^2} + \sqrt{\frac{8}{e_c^3 n^2}\sum_{i=1}^n (a_i - \bar{a})^2 \sum_{i=1}^n (b_i - \bar{b})^2} \right\}^2 =: \sigma_{\mathrm{cov}}^2.
$$

Then, the conclusion follows from Lemma 1.

C.3. Proof of Lemma 5

PROOF. For any $t > 0$, we have

$$
P\Big(\Big\| \sum_{m=1}^M \pi_{[m]} \big(\bar{\boldsymbol{x}}_{[m]1} - \bar{\boldsymbol{x}}_{[m]} \big) \Big\|_\infty \geqslant t \Big) \leqslant p \cdot \max_{1\leqslant j\leqslant p} P\Big(\Big| \sum_{m=1}^M \pi_{[m]} \big(\bar{x}_{[m]1,j} - \bar{x}_{[m],j} \big) \Big| \geqslant t \Big).
$$

Applying Proposition 1 to the jth covariate $\boldsymbol{X}_j$ (and $-\boldsymbol{X}_j$), we have

$$
P\Big(\Big\| \sum_{m=1}^M \pi_{[m]} \big(\bar{\boldsymbol{x}}_{[m]1} - \bar{\boldsymbol{x}}_{[m]} \big) \Big\|_\infty \geqslant t \Big) \leqslant 2p \cdot \max_{1\leqslant j\leqslant p} \exp\Big\{ -\frac{nt^2}{4\sigma_{x,j}^2} \Big\},
$$

where $\sigma_{x,j}^2 = (1/n)\sum_{m=1}^M \sum_{i\in[m]} (x_{ij} - \bar{x}_{[m],j})^2 / e_{[m]}^2$. According to Conditions 1–2 and Cauchy-Schwarz inequality, we have,

$$
\sigma_{x,j}^2 \leqslant \frac{1}{c^2} \cdot \frac{1}{n} \sum_{m=1}^M \sum_{i\in[m]} \Big(x_{ij} - \bar{x}_{[m],j} \Big)^2 \leqslant \frac{1}{c^2} \cdot \Big\{ \frac{1}{n} \sum_{m=1}^M \sum_{i\in[m]} \Big(x_{ij} - \bar{x}_{[m],j} \Big)^4 \Big\}^{1/2} \leqslant \frac{L^{1/2}}{c^2}.
$$

Therefore,

$$
P\Big(\Big\| \sum_{m=1}^M \pi_{[m]} \big(\bar{\boldsymbol{x}}_{[m]1} - \bar{\boldsymbol{x}}_{[m]} \big) \Big\|_\infty \geqslant t \Big) \leqslant 2\exp\Big(\log p - \frac{c^2 n t^2}{4L^{1/2}} \Big).
$$

Taking $t = \sqrt{8L^{1/2}/c^2} \cdot \sqrt{(\log p)/n}$ gives the result.

C.4. *Proof of Lemma 6*

PROOF. For any $t > 0$, by triangle inequality, we have

$$\begin{aligned} P(\|\hat{\Sigma}_{\boldsymbol{X}\varepsilon(1)}\|_\infty \geqslant t) &= P(\|\hat{\Sigma}_{\boldsymbol{X}\varepsilon(1)} - \Sigma_{\boldsymbol{X}\varepsilon(1)} + \Sigma_{\boldsymbol{X}\varepsilon(1)}\|_\infty \geqslant t) \\ &\leqslant P(\|\hat{\Sigma}_{\boldsymbol{X}\varepsilon(1)} - \Sigma_{\boldsymbol{X}\varepsilon(1)}\|_\infty \geqslant t - \|\Sigma_{\boldsymbol{X}\varepsilon(1)}\|_\infty) \\ &\leqslant P(\|\hat{\Sigma}_{\boldsymbol{X}\varepsilon(1)} - \Sigma_{\boldsymbol{X}\varepsilon(1)}\|_\infty \geqslant t - \delta_n) \\ &\leqslant p \cdot \max_{1 \leqslant j \leqslant p} P(|\hat{\Sigma}_{X_j\varepsilon(1)} - \Sigma_{X_j\varepsilon(1)}| \geqslant t - \delta_n). \end{aligned}$$

Applying Proposition 2 to $a_i = x_{ij}$ and $b_i = \varepsilon_i(1)$, we have

$$P\left(\left|\hat{\Sigma}_{X_j\varepsilon(1)} - \Sigma_{X_j\varepsilon(1)}\right| \geqslant t - \delta_n\right) \leqslant 2\exp\left\{-\frac{n(t-\delta_n)^2}{60(\kappa_a^4\kappa_b^4)^{1/2}}\right\} \leqslant 2\exp\left\{-\frac{c^3 n(t-\delta_n)^2}{60L}\right\},$$

where the last inequality is due to Conditions 1 and 2. Thus, we have

$$P(\|\hat{\Sigma}_{\boldsymbol{X}\varepsilon(1)}\|_\infty \geqslant t) \leqslant 2p\exp\left\{-\frac{c^3 n(t-\delta_n)^2}{60L}\right\}.$$

Then,

$$P(\|\hat{\Sigma}_{\boldsymbol{X}\varepsilon(1)}\|_\infty \geqslant t) \to 0, \quad \forall t \geqslant \sqrt{\frac{120L\log p}{c^3 n}} + \delta_n.$$

By Condition 5 on λ_1, i.e., $\eta\lambda_1 \geqslant \sqrt{120L\log p/(c^3 n)} + \delta_n$, we have

$$P\left(\|\hat{\Sigma}_{\boldsymbol{X}\varepsilon(1)}\|_\infty \leqslant \eta\lambda_1\right) \to 1.$$

C.5. *Proof of Lemma 7*

PROOF. For any $t > 0$,

$$P(\|\hat{\Sigma}_{\boldsymbol{XX}} - \Sigma_{\boldsymbol{XX}}\|_\infty \geqslant t) \leqslant p^2 \cdot \max_{1 \leqslant j,l \leqslant p} P(|\hat{\Sigma}_{X_jX_l} - \Sigma_{X_jX_l}| \geqslant t).$$

Applying Proposition 2 to $a_i = x_{ij}$ and $b_i = x_{il}$, we have

$$P\left(\left|\hat{\Sigma}_{X_jX_l} - \Sigma_{X_jX_l}\right| \geqslant t\right) \leqslant 2\exp\left\{-\frac{nt^2}{60(\kappa_a^4\kappa_b^4)^{1/2}}\right\} \leqslant 2\exp\left\{-\frac{c^3 nt^2}{60L}\right\},$$

where the last inequality is due to Conditions 1 and 2. Thus, we have

$$P(\|\hat{\Sigma}_{\boldsymbol{XX}} - \Sigma_{\boldsymbol{XX}}\|_\infty \geqslant t) \leqslant 2p^2\exp\left\{-\frac{c^3 nt^2}{60L}\right\}.$$

Taking

$$t = \sqrt{\frac{180L\log p}{c^3 n}},$$

we have

$$P\left(\|\hat{\Sigma}_{\boldsymbol{XX}} - \Sigma_{\boldsymbol{XX}}\|_\infty \leqslant \sqrt{180L/c^3}\sqrt{(\log p)/n}\right) \to 1.$$

C.6. Proof of Lemma 8

Proof. The proof is similar to the proof of Lemma S3 in Bloniarz et al. (2016). The difference is that (1) we use the weighted l_2 loss function for the Lasso, and (2) we use stratified randomization, which needs new concentration inequalities to handle. Recall the notations as follows:

$$\Sigma_{\boldsymbol{XX}} = \sum_{m=1}^{M} \pi_{[m]} S^2_{[m]\boldsymbol{X}}, \quad \Sigma_{\boldsymbol{X}\varepsilon(1)} = \sum_{m=1}^{M} \pi_{[m]} S_{[m]\boldsymbol{X}\varepsilon(1)}, \quad \hat{\Sigma}_{\boldsymbol{X}\varepsilon(1)} = \sum_{m=1}^{M} \pi_{[m]} s_{[m]\boldsymbol{X}\varepsilon(1)},$$
$$\hat{\Sigma}_{\boldsymbol{XX}} = \sum_{m=1}^{M} \pi_{[m]} s^2_{[m]\boldsymbol{X}} = \sum_{m=1}^{M} \pi_{[m]} \cdot \frac{1}{n_{[m]1}-1} \sum_{i\in[m]} Z_i (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_{[m]1})^{\mathrm{T}} (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_{[m]1}).$$

We further denote $\boldsymbol{h} = \hat{\boldsymbol{\beta}}_{\text{lasso},1} - \boldsymbol{\beta}(1)$. According to the KKT condition, we have

$$\sum_{m=1}^{M} \pi_{[m]} \cdot \frac{1}{n_{[m]1}-1} \sum_{i\in[m]} Z_i \big(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_{[m]1}\big) \Big\{ Y_i(1) - \bar{Y}_{[m]1} - \big(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_{[m]1}\big)^{\mathrm{T}} \hat{\boldsymbol{\beta}}_{\text{lasso},1} \Big\} = \lambda_1 \kappa, \quad (34)$$

where κ is the sub-gradient of $\|\boldsymbol{\beta}\|_1$ taking value at $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}_{\text{lasso},1}$, i.e.,

$$\kappa \in \partial \|\boldsymbol{\beta}\|_1 \Big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}_{\text{lasso},1}} \quad \text{with} \quad \begin{cases} \kappa_j = \operatorname{sign}((\hat{\boldsymbol{\beta}}_{\text{lasso},1})_j) \text{ for } j \text{ such that } (\hat{\boldsymbol{\beta}}_{\text{lasso},1})_j \neq 0 \\ \kappa_j \in [-1,1], \quad \text{otherwise}. \end{cases}$$

By the decomposition of $Y_i(1)$, we have

$$Y_i(1) - \bar{Y}_{[m]1} = \big(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_{[m]1}\big)^{\mathrm{T}} \boldsymbol{\beta}(1) + \varepsilon_i(1) - \bar{\varepsilon}_{[m]1}.$$

Then (34) becomes

$$\hat{\Sigma}_{\boldsymbol{XX}} \{\boldsymbol{\beta}(1) - \hat{\boldsymbol{\beta}}_{\text{lasso},1}\} + \hat{\Sigma}_{\boldsymbol{X}\varepsilon(1)} = \lambda_1 \kappa. \quad (35)$$

Multiplying both sides of (35) by $-\boldsymbol{h}^{\mathrm{T}} = \{\boldsymbol{\beta}(1) - \hat{\boldsymbol{\beta}}_{\text{lasso},1}\}^{\mathrm{T}}$, we have

$$\boldsymbol{h}^{\mathrm{T}} \hat{\Sigma}_{\boldsymbol{XX}} \boldsymbol{h} - \boldsymbol{h}^{\mathrm{T}} \hat{\Sigma}_{\boldsymbol{X}\varepsilon(1)} = \lambda_1 (-\boldsymbol{h})^{\mathrm{T}} \kappa \leqslant \lambda_1 \big(\|\boldsymbol{\beta}(1)\|_1 - \|\hat{\boldsymbol{\beta}}_{\text{lasso},1}\|_1 \big),$$

where the last inequality is because

$$\{\boldsymbol{\beta}(1)\}^{\mathrm{T}} \kappa \leqslant \|\boldsymbol{\beta}(1)\|_1 \|\kappa\|_\infty \leqslant \|\boldsymbol{\beta}(1)\|_1 \quad \text{and} \quad \hat{\boldsymbol{\beta}}^{\mathrm{T}}_{\text{lasso},1} \kappa = \|\hat{\boldsymbol{\beta}}_{\text{lasso},1}\|_1.$$

Rearranging and by Hölder inequality, we have

$$\boldsymbol{h}^{\mathrm{T}} \hat{\Sigma}_{\boldsymbol{XX}} \boldsymbol{h} \leqslant \lambda_1 \big\{ \|\boldsymbol{\beta}(1)\|_1 - \|\hat{\boldsymbol{\beta}}_{\text{lasso},1}\|_1 \big\} + \|\boldsymbol{h}\|_1 \cdot \|\hat{\Sigma}_{\boldsymbol{X}\varepsilon(1)}\|_\infty. \quad (36)$$

By the definition of $\boldsymbol{h}$ and several applications of the triangle inequality, we have

$$\begin{aligned} \|\boldsymbol{\beta}(1)\|_1 - \|\hat{\boldsymbol{\beta}}_{\text{lasso},1}\|_1 &= \|\{\boldsymbol{\beta}(1)\}_S\|_1 - \|\{\hat{\boldsymbol{\beta}}_{\text{lasso},1}\}_S\|_1 + \|\{\boldsymbol{\beta}(1)\}_{S^c}\|_1 - \|\{\hat{\boldsymbol{\beta}}_{\text{lasso},1}\}_{S^c}\|_1 \\ &\leqslant \|\boldsymbol{h}_S\|_1 + \|\{\boldsymbol{\beta}(1)\}_{S^c}\|_1 - \{\|\boldsymbol{h}_{S^c}\|_1 - \|\{\boldsymbol{\beta}(1)\}_{S^c}\|_1\} \\ &= \|\boldsymbol{h}_S\|_1 - \|\boldsymbol{h}_{S^c}\|_1 + 2\|\{\boldsymbol{\beta}(1)\}_{S^c}\|_1. \end{aligned}$$

Therefore, conditional on $\mathcal{L}$ with $P(\mathcal{L}) \to 1$ (Lemma 6), we have

$$\begin{aligned} 0 \leqslant \boldsymbol{h}^{\mathrm{T}}\hat{\Sigma}_{\boldsymbol{X}\boldsymbol{X}}\boldsymbol{h} &\leqslant \lambda_1\Big(\|\boldsymbol{h}_S\|_1 - \|\boldsymbol{h}_{S^c}\|_1 + 2\|\{\boldsymbol{\beta}(1)\}_{S^c}\|_1 + \eta\|\boldsymbol{h}\|_1\Big) \\ &\leqslant \lambda_1\Big\{(\eta-1)\|\boldsymbol{h}_{S^c}\|_1 + (1+\eta)\|\boldsymbol{h}_S\|_1 + 2\|\{\boldsymbol{\beta}(1)\}_{S^c}\|_1\Big\}. \end{aligned}$$

Then,

$$(1-\eta)\|\boldsymbol{h}_{S^c}\|_1 \leqslant (1+\eta)\|\boldsymbol{h}_S\|_1 + 2\|\{\boldsymbol{\beta}(1)\}_{S^c}\|_1 \leqslant (1+\eta)\|\boldsymbol{h}_S\|_1 + 2s\lambda_1. \tag{37}$$

Recall that $\xi > 1$ and $0 < \eta < (\xi-1)/(\xi+1) < 1$, thus $(1-\eta)\xi - (1+\eta) > 0$. To proceed, considering the following two cases:

(1) If $\|\boldsymbol{h}_S\|_1 \leqslant 2s\lambda_1/\{(1-\eta)\xi-(1+\eta)\}$, then by (37):

$$\begin{aligned} \|\boldsymbol{h}\|_1 =& \|\boldsymbol{h}_S\|_1 + \|\boldsymbol{h}_{S^c}\|_1 \\ \leqslant& \|\boldsymbol{h}_S\|_1 + \frac{1+\eta}{1-\eta}\|\boldsymbol{h}_S\|_1 + \frac{2s\lambda_1}{1-\eta} \\ =& \frac{1}{1-\eta}\Big\{2\|\boldsymbol{h}_S\|_1 + 2s\lambda_1\Big\} \\ \leqslant& \frac{2s\lambda_1}{1-\eta}\Big\{\frac{2}{(1-\eta)\xi-(1+\eta)} + 1\Big\}. \end{aligned}$$

Then,

$$\|\boldsymbol{h}\|_1 = \|\hat{\boldsymbol{\beta}}_{\text{lasso},1} - \boldsymbol{\beta}(1)\|_1 = O_p(s\lambda_1).$$

(2) If $2s\lambda_1 < \{(1-\eta)\xi-(1+\eta)\}\|\boldsymbol{h}_S\|_1$, then by (37), we have

$$\|\boldsymbol{h}_{S^c}\|_1 \leqslant \frac{1+\eta}{1-\eta}\|\boldsymbol{h}_S\|_1 + \frac{(1-\eta)\xi-(1+\eta)}{1-\eta}\|\boldsymbol{h}_S\|_1 = \xi\|\boldsymbol{h}_S\|_1.$$

According to Condition 4, we have

$$\|\boldsymbol{h}\|_1 = \|\boldsymbol{h}_S\|_1 + \|\boldsymbol{h}_{S^c}\|_1 \leqslant (1+\xi)\|\boldsymbol{h}_S\|_1 \leqslant (1+\xi)\cdot C\cdot s\cdot\|\Sigma_{\boldsymbol{X}\boldsymbol{X}}\boldsymbol{h}\|_\infty. \tag{38}$$

Taking the l_∞-norm on both sides of KKT condition (35), we have, conditional on the event $\mathcal{L}$,

$$\|\hat{\Sigma}_{\boldsymbol{X}\boldsymbol{X}}\boldsymbol{h}\|_\infty \leqslant \lambda_1 + \|\hat{\Sigma}_{\boldsymbol{X}\varepsilon(1)}\|_\infty \leqslant (1+\eta)\lambda_1. \tag{39}$$

By triangle inequality and Hölder inequality, and conditional on the event $\mathcal{L}\cap\mathcal{K}$, we have

$$\begin{aligned} s\|\Sigma_{\boldsymbol{X}\boldsymbol{X}}\boldsymbol{h}\|_\infty &\leqslant s\|\hat{\Sigma}_{\boldsymbol{X}\boldsymbol{X}} - \Sigma_{\boldsymbol{X}\boldsymbol{X}}\|_\infty\|\boldsymbol{h}\|_1 + s\|\hat{\Sigma}_{\boldsymbol{X}\boldsymbol{X}}\boldsymbol{h}\|_\infty \\ &\leqslant s\sqrt{180L/c^3}\sqrt{(\log p)/n}\|\boldsymbol{h}\|_1 + s\|\hat{\Sigma}_{\boldsymbol{X}\boldsymbol{X}}\boldsymbol{h}\|_\infty \\ &\leqslant s\sqrt{180L/c^3}\sqrt{(\log p)/n}\|\boldsymbol{h}\|_1 + s(1+\eta)\lambda_1 \\ &= o(1)\|\boldsymbol{h}\|_1 + s(1+\eta)\lambda_1, \end{aligned}$$

where the third inequality is because of inequality (39) and the last equality is because of Condition 5. Combining above inequality and (38), we obtain

$$\|\boldsymbol{h}\|_1 = \|\hat{\boldsymbol{\beta}}_{\text{lasso},1} - \boldsymbol{\beta}(1)\|_1 = O_p(s\lambda_1).$$

C.7. Proof of Lemma 9

We will only prove that $\hat{s}(1) \leqslant \widetilde{C}s$ holds in probability for some constant $\widetilde{C}$; the proof for $\hat{s}(0) \leqslant \widetilde{C}s$ is similar. First, we bound $\|\hat{\Sigma}_{\boldsymbol{X}\boldsymbol{X}}\{\boldsymbol{\beta}(1) - \hat{\boldsymbol{\beta}}_{\text{lasso},1}\}\|_2^2$ from below. Specifically, we examine the jth element of $\hat{\Sigma}_{\boldsymbol{X}\boldsymbol{X}}\{\boldsymbol{\beta}(1) - \hat{\boldsymbol{\beta}}_{\text{lasso},1}\}$, that is, $\boldsymbol{e}_j^{\mathrm{T}}\hat{\Sigma}_{\boldsymbol{X}\boldsymbol{X}}\{\boldsymbol{\beta}(1) - \hat{\boldsymbol{\beta}}_{\text{lasso},1}\}$, where $\boldsymbol{e}_j$ is a p-dimensional vector with one on its jth entry and zero on other entries. According to KKT condition (35), we have shown that

$$\hat{\Sigma}_{\boldsymbol{X}\boldsymbol{X}}\{\boldsymbol{\beta}(1) - \hat{\boldsymbol{\beta}}_{\text{lasso},1}\} + \hat{\Sigma}_{\boldsymbol{X}\varepsilon(1)} = \lambda_1 \kappa.$$

For any $j \in \{1, 2, \ldots, p\}$ such that $\big(\hat{\boldsymbol{\beta}}_{\text{lasso},1}\big)_j \neq 0$, we have

$$\left|\boldsymbol{e}_j^{\mathrm{T}}\hat{\Sigma}_{\boldsymbol{X}\boldsymbol{X}}\{\boldsymbol{\beta}(1) - \hat{\boldsymbol{\beta}}_{\text{lasso},1}\} + \boldsymbol{e}_j^{\mathrm{T}}\hat{\Sigma}_{\boldsymbol{X}\varepsilon(1)}\right| = \lambda_1.$$

Conditional on $\mathcal{L} = \left\{\|\hat{\Sigma}_{\boldsymbol{X}\varepsilon(1)}\|_\infty \leqslant \eta\lambda_1\right\}$ with $P(\mathcal{L}) \to 1$ (Lemma 6), and by triangle inequality, we have,

$$\left|\boldsymbol{e}_j^{\mathrm{T}}\hat{\Sigma}_{\boldsymbol{X}\boldsymbol{X}}\{\boldsymbol{\beta}(1) - \hat{\boldsymbol{\beta}}_{\text{lasso},1}\}\right| \geqslant \lambda_1 - \left|\boldsymbol{e}_j^{\mathrm{T}}\hat{\Sigma}_{\boldsymbol{X}\varepsilon(1)}\right| \geqslant \lambda_1 - \|\hat{\Sigma}_{\boldsymbol{X}\varepsilon(1)}\|_\infty \geqslant (1-\eta)\lambda_1.$$

Then, summing up the square of the elements of $\hat{\Sigma}_{\boldsymbol{X}\boldsymbol{X}}\{\boldsymbol{\beta}(1) - \hat{\boldsymbol{\beta}}_{\text{lasso},1}\}$, we have

$$\begin{aligned}
\|\hat{\Sigma}_{\boldsymbol{X}\boldsymbol{X}}\{\boldsymbol{\beta}(1) - \hat{\boldsymbol{\beta}}_{\text{lasso},1}\}\|_2^2 =& \sum_{j=1}^{p}\left|\boldsymbol{e}_j^{\mathrm{T}}\hat{\Sigma}_{\boldsymbol{X}\boldsymbol{X}}\{\boldsymbol{\beta}(1) - \hat{\boldsymbol{\beta}}_{\text{lasso},1}\}\right|^2 \\
\geqslant& \sum_{j:(\hat{\boldsymbol{\beta}}_{\text{lasso},1})_j \neq 0}\left|\boldsymbol{e}_j^{\mathrm{T}}\hat{\Sigma}_{\boldsymbol{X}\boldsymbol{X}}\{\boldsymbol{\beta}(1) - \hat{\boldsymbol{\beta}}_{\text{lasso},1}\}\right|^2 \\
\geqslant& (1-\eta)^2\lambda_1^2\hat{s}(1). \qquad (40)
\end{aligned}$$

Second, we bound $\|\hat{\Sigma}_{\boldsymbol{X}\boldsymbol{X}}\{\boldsymbol{\beta}(1) - \hat{\boldsymbol{\beta}}_{\text{lasso},1}\}\|_2^2$ from above:

$$\begin{aligned}
\|\hat{\Sigma}_{\boldsymbol{X}\boldsymbol{X}}\{\boldsymbol{\beta}(1) - \hat{\boldsymbol{\beta}}_{\text{lasso},1}\}\|_2^2 \leqslant& \Lambda_{\max}\big(\hat{\Sigma}_{\boldsymbol{X}\boldsymbol{X}}\big) \cdot \left\|\hat{\Sigma}_{\boldsymbol{X}\boldsymbol{X}}^{1/2}\{\boldsymbol{\beta}(1) - \hat{\boldsymbol{\beta}}_{\text{lasso},1}\}\right\|_2^2 \\
=& \Lambda_{\max}\big(\hat{\Sigma}_{\boldsymbol{X}\boldsymbol{X}}\big) \cdot \{\boldsymbol{\beta}(1) - \hat{\boldsymbol{\beta}}_{\text{lasso},1}\}^{\mathrm{T}}\hat{\Sigma}_{\boldsymbol{X}\boldsymbol{X}}\{\boldsymbol{\beta}(1) - \hat{\boldsymbol{\beta}}_{\text{lasso},1}\} \\
=& \Lambda_{\max}\big(\hat{\Sigma}_{\boldsymbol{X}\boldsymbol{X}}\big) \cdot \boldsymbol{h}^{\mathrm{T}}\hat{\Sigma}_{\boldsymbol{X}\boldsymbol{X}}\boldsymbol{h}. \qquad (41)
\end{aligned}$$

We deal with $\boldsymbol{h}^{\mathrm{T}}\hat{\Sigma}_{\boldsymbol{X}\boldsymbol{X}}\boldsymbol{h}$ and $\Lambda_{\max}(\hat{\Sigma}_{\boldsymbol{X}\boldsymbol{X}})$ separately. To bound $\boldsymbol{h}^{\mathrm{T}}\hat{\Sigma}_{\boldsymbol{X}\boldsymbol{X}}\boldsymbol{h}$, we have shown that (see (36) in the proof of Lemma 8),

$$\boldsymbol{h}^{\mathrm{T}}\hat{\Sigma}_{\boldsymbol{X}\boldsymbol{X}}\boldsymbol{h} \leqslant \lambda_1\{\|\boldsymbol{\beta}(1)\|_1 - \|\hat{\boldsymbol{\beta}}_{\text{lasso},1}\|_1\} + \|\boldsymbol{h}\|_1\|\hat{\Sigma}_{\boldsymbol{X}\varepsilon(1)}\|_\infty.$$

Conditional on $\mathcal{L} = \left\{\|\hat{\Sigma}_{\boldsymbol{X}\varepsilon(1)}\|_\infty \leqslant \eta\lambda_1\right\}$ with $P(\mathcal{L}) \to 1$ (Lemma 6), and by triangle inequality, we have

$$\boldsymbol{h}^{\mathrm{T}}\hat{\Sigma}_{\boldsymbol{X}\boldsymbol{X}}\boldsymbol{h} \leqslant \lambda_1(1+\eta)\|\boldsymbol{h}\|_1.$$

According to the proof of Lemma 8, with probability tending to one, there exists a constant C, such that

$$\|\boldsymbol{h}\|_1 = \left\|\boldsymbol{\beta}(1) - \hat{\boldsymbol{\beta}}_{\text{lasso},1}\right\|_1 \leqslant C\lambda_1 s.$$

Therefore, we have

$$\boldsymbol{h}^{\mathrm{T}}\hat{\Sigma}_{\boldsymbol{X}\boldsymbol{X}}\boldsymbol{h} \leqslant C(1+\eta)\lambda_1^2 s. \tag{42}$$

To bound $\Lambda_{\max}(\hat{\Sigma}_{\boldsymbol{X}\boldsymbol{X}})$, we expand its expression to

$$\Lambda_{\max}(\hat{\Sigma}_{\boldsymbol{X}\boldsymbol{X}}) = \max_{\boldsymbol{u}:\|\boldsymbol{u}\|_2=1} \boldsymbol{u}^{\mathrm{T}}\hat{\Sigma}_{\boldsymbol{X}\boldsymbol{X}}\boldsymbol{u} = \max_{\boldsymbol{u}:\|\boldsymbol{u}\|_2=1} \sum_{m=1}^{M} \pi_{[m]}\boldsymbol{u}^{\mathrm{T}} s_{[m]\boldsymbol{X}}^2 \boldsymbol{u}. \tag{43}$$

By decomposing $s_{[m]\boldsymbol{X}}^2$, we have

$$\begin{aligned}
&(n_{[m]1}-1)\boldsymbol{u}^{\mathrm{T}} s_{[m]\boldsymbol{X}}^2 \boldsymbol{u} \\
=& \sum_{i\in[m]} \boldsymbol{u}^{\mathrm{T}} Z_i(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_{[m]})(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_{[m]})^{\mathrm{T}}\boldsymbol{u} - n_{[m]1}\boldsymbol{u}^{\mathrm{T}}(\bar{\boldsymbol{x}}_{[m]1} - \bar{\boldsymbol{x}}_{[m]})(\bar{\boldsymbol{x}}_{[m]1} - \bar{\boldsymbol{x}}_{[m]})^{\mathrm{T}}\boldsymbol{u} \\
\leqslant& \sum_{i\in[m]} \boldsymbol{u}^{\mathrm{T}} Z_i(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_{[m]})(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_{[m]})^{\mathrm{T}}\boldsymbol{u} \\
\leqslant& \sum_{i\in[m]} \boldsymbol{u}^{\mathrm{T}} (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_{[m]})(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_{[m]})^{\mathrm{T}}\boldsymbol{u} \\
=&(n_{[m]}-1)\boldsymbol{u}^{\mathrm{T}} S_{[m]\boldsymbol{X}}^2 \boldsymbol{u}.
\end{aligned}$$

By Condition 1 and $n_{[m]1} \geqslant 2$, we have

$$\boldsymbol{u}^{\mathrm{T}} s_{[m]\boldsymbol{X}}^2 \boldsymbol{u} \leqslant \frac{(n_{[m]}-1)}{(n_{[m]1}-1)}\boldsymbol{u}^{\mathrm{T}} S_{[m]\boldsymbol{X}}^2 \boldsymbol{u} \leqslant \frac{2n_{[m]}}{n_{[m]1}}\boldsymbol{u}^{\mathrm{T}} S_{[m]\boldsymbol{X}}^2 \boldsymbol{u} \leqslant \frac{2}{c}\boldsymbol{u}^{\mathrm{T}} S_{[m]\boldsymbol{X}}^2 \boldsymbol{u}.$$

Plugging above inequality into (43), by Condition 7, we have

$$\Lambda_{\max}(\hat{\Sigma}_{\boldsymbol{X}\boldsymbol{X}}) \leqslant \frac{2}{c} \max_{\boldsymbol{u}:\|\boldsymbol{u}\|_2=1} \boldsymbol{u}^{\mathrm{T}} \sum_{m=1}^{M} \pi_{[m]} S_{[m]\boldsymbol{X}}^2 \boldsymbol{u} \leqslant \frac{2}{c}\Lambda_{\max}(\Sigma_{\boldsymbol{X}\boldsymbol{X}}) \leqslant \frac{2C}{c}. \tag{44}$$

Combining (41), (42), and (44), the following inequality holds in probability:

$$\|\hat{\Sigma}_{\boldsymbol{X}\boldsymbol{X}}\{\boldsymbol{\beta}(1) - \hat{\boldsymbol{\beta}}_{\text{lasso},1}\}\|_2^2 \leqslant 2C^2(1+\eta)\lambda_1^2 s/c. \tag{45}$$

Finally, combining (40) and (45), we have, with probability tending to one,

$$\hat{s}(1) \leqslant \frac{2C^2(1+\eta)}{c(1-\eta)^2} s =: \widetilde{C}s.$$